Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles

Shan Jiang, Simon Baumgartner, Abe Ittycheriah, Cong Yu
Background: fact-checks
Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles

Background: fact-checks

Factoring Fact-Checks

Fact-checking is the act of checking factual information in non-fictional text in order to determine the veracity and correctness of the factual statements in the text.
Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles

Background: fact-checks

Factoring Fact-Checks

Fact-checking = the act of checking facts.
An article that does fact-checking is called “fact-checks”.
Background: fact-checks

No, ‘newspaper’ isn’t an acronym for ‘north, east, west, south, past and present event report’

Has the word “newspaper” really been an acronym all this time? That’s what one viral Facebook post claims. According to the post, which has gotten over 2,400 shares in 24 hours, “newspaper” is an acronym for “North, East, West, South, Past and Present Report.”

……
The word paper alone has origins in the Latin word “papyrus,” the stalks used to make paper, and the Greek word “papyros”. This claim is a repurposed hoax. We rate it Pants on Fire!
Background: factors

No, ‘newspaper’ isn’t an acronym for ‘north, east, west, south, past and present event report’

Has the word “newspaper” really been an acronym all this time? That’s what one viral Facebook post claims. According to the post, which has gotten over 2,400 shares in 24 hours, “newspaper” is an acronym for “North, East, West, South, Past and Present Report.”

Claim

……

The word paper alone has origins in the Latin word “papyrus,” the stalks used to make paper, and the Greek word “papyros”. This claim is a repurposed hoax. We rate it Pants on Fire!
Background: factors

No, ‘newspaper’ isn’t an acronym for ‘north, east, west, south, past and present event report’

Has the word “newspaper” really been an acronym all this time? That’s what one viral Facebook post claims. According to the post, which has gotten over 2,400 shares in 24 hours, “newspaper” is an acronym for “North, East, West, South, Past and Present Report.”

The word paper alone has origins in the Latin word “papyrus,” the stalks used to make paper, and the Greek word “papyros.” This claim is a repurposed hoax. We rate it Pants on Fire!
Background: **factors**

Has the word “newspaper” really been an acronym all this time?
That’s what one viral Facebook post claims. **Claimant**

According to the post, which has gotten over 2,400 shares in 24 hours, “newspaper” is an acronym for “North, East, West, South, Past and Present Report.” **Claim**

......

The word paper alone has origins in the Latin word “papyrus,” the stalks used to make paper, and the Greek word “papyros”.
This claim is a repurposed hoax. We rate it **Pants on Fire!** **Verdict**
Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles

Background: factors

Claim: “newspaper” is an acronym for “North, East, West, South, Past and Present Report.”

Claimant: viral Facebook post

Verdict: Pants on Fire!
Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles

S. Jiang et al.

Background: application

newspaper acronym

No, 'newspaper' isn't an acronym for 'north, east, west, south ...
https://www.politifact.com › statements › sep › facebook-posts › no-newsp...

Claim: Says the word newspaper stands for "north, east, west, south, past and present event report."

Claimed by: Facebook posts

Fact check by PolitiFact: Pants on Fire
Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles

S. Jiang et al.

Background: application

No, ‘newspaper’ is not an acronym of ‘North, East, West ...’
https://africacheck.org/fbcheck/no-newspaper-is-not-an-acronym-of-north-east-west...

Claim: ‘Newspaper’ is an acronym of ‘North, East, West, South, Past and Present Events Report’

False · Fact checked by ResultPartUpdater
Background: motivation

Question: How to get these factors?
Background: motivation

Fact-check markup tool:

https://toolbox.google.com/factcheck
Background: motivation

ClaimReview markup:

https://schema.org/ClaimReview
Background: problem

Problem:

It takes time!

As of July 2019, < 50% fact-checkers use it. [1]

Background: proposal

Proposal:

Automatically extracting factors from fact-checks.

(factoring fact-checks)
Steps:

• Explore fact-check data for patterns of factors.
• Experiment with information extraction models.
Steps: data

Steps:

- Explore fact-check data for patterns of factors.
- Experiment with information extraction models.
Data: source

- Fact-check dataset from DataCommons. [2]

Data: source

- Fact-check dataset from DataCommons. [2]
- 6,216 fact-checks (English).

Data: source

- Fact-check dataset from DataCommons. [2]
- 6,216 fact-checks (English).
- Reported factors (claim, claimant, verdict).

**Data:** who are the fact-checkers?

- politifact.com: 3915
- factcheck.org: 856
- washingtonpost.com: 627
- factly.in: 250
- africacheck.org: 220
- Others: 348

**Well-known**

**Under-represented**
Data: who are the fact-checkers?

Fact-checker

- politifact.com: 3915
- factcheck.org: 856
- washingtonpost.com: 627
- factly.in: 250
- africacheck.org: 220

Others: 348

Well-known

Under-represented

# of fact-checks
Data: who are the fact-checkers?

Useful later for experiments.
Data: can factors be found in the fact-checks?

- Exact string matching.
- Out of 6,216 fact-checks, 80% of claimants, 76% of verdicts, and 32% of claims can be matched.
Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles

Data: can factors be found in the fact-checks?

- Exact string matching.
- Out of 6,216 fact-checks, 80% of claimants, 76% of verdicts, and 32% of claims can be matched.

Paraphrasing.

Claim in article: “newspaper” is an acronym for “North, East, West, South, Past and Present Report.”

Reported claim: says the word newspaper stands for “north, east, west, south, past and present report.”
Data: can factors be found in the fact-checks?

- **Exact string matching.**

- **Out of 6,216 fact-checks, 80% of claimants, 76% of verdicts, and 32% of claims can be matched.**

- **At least 2/3 of overlap.**

- **Minimum window substring matching.** [3]

---

Data: can factors be found in the fact-checks?

79% of claims.
Data: can factors be found in the fact-checks?

79% of claims.  80% of claimants.
Data: can factors be found in the fact-checks?

79% of claims.  80% of claimants.  80% of verdicts.
Data: where are the factors in the fact-check?

- Relative position.
- Position / length of the fact-check.

  0 = the head of the fact-check.

  1 = the tail of the fact-check.
Data: where are the factors in the fact-check?

- Relative position.
- Position / length of the fact-check.
  
  0 = the head of the fact-check.
  
  1 = the tail of the fact-check.

- Separate well-known and under-represented fact-checkers
Data: where are the factors in the fact-check?

Claims:

Well-known:
head and tail

Under-represented:
head only
Data: where are the factors in the fact-check?

Claims:
Well-known: head and tail
Under-represented: head only

Claimants:
Well-known: head only
Under-represented: head only
Data: where are the factors in the fact-check?

Claims:
Well-known: head and tail
Under-represented: head only

Claimants:
Well-known: head only
Under-represented: head only

Verdicts:
Well-known: tail only
Under-represented: head only
Data: where are the factors in the fact-check?

Useful later for experiments.

Claims:
Well-known:
head and tail
Under-represented:
head only

Claimants:
Well-known:
head only
Under-represented:
head only

Verdicts:
Well-known:
tail only
Under-represented:
head only
Steps: experiments

Steps:

• Explore fact-check data for patterns of factors.
• Experiment with information extraction models.
Task: intuition

- The factor *per se*.

**Claim**: factual statement, numbers, statistics, etc.

**Claimant**: person, organization, etc.

**Verdict**: true, false, pants on fire, Pinocchio, etc.
Task: intuition

• The factor *per se*.

Claim: factual statement, numbers, statistics, etc.
Claimant: person, organization, etc.
Verdict: true, false, pants on fire, Pinocchio, etc.

• Surrounding context of the factor.

Claim: someone said/claimed (...)
Claimant: (someone) said/claimed ...
Verdict: we rate it (...), a (false) rumor claims ...
Task: formulation

- Sequence tagging task.
- Input: fact-check (sequence of tokens).
- Output: equal-length sequence of labels.
Task: formulation

- Sequence tagging task.
- Input: fact-check (sequence of tokens).
- Output: equal-length sequence of labels.

John Doe made a false claim that the earth is flat.
Task: problem

Problem:

Factors can be paraphrased.

Need to generate ground-truth token-level labels.
Task: ground-truth

Generating ground-truth labels w/ rule-based taggers.

Claim: The earth is flat.
Claimant: John Doe
Verdict: False

Fluent tagger:
John  M  Doe made a false claim that the earth is actually 100% flat.

Concise tagger:
John  M  Doe made a false claim that the earth is actually 100% flat.
Experiments: baseline

- **Claim**: ClaimBuster, top “check-worthiness”. [4]
Experiments: baseline

- **Claim**: ClaimBuster, top “check-worthiness”. [4]
- **Claimant**: entity tagging + majority.

Experiments: baseline

- **Claim**: ClaimBuster, top “check-worthiness”. [4]
- **Claimant**: entity tagging + majority.
- **Verdict**: majority.

Experiments: BERT

Replace last layer w/ tagging + cross entropy loss.

[CLS] A post says that D.A.R.E. ... this false.
Experiments: problem

- BERT has default maximum sequence length: 512.
- Feed to it paragraph by paragraph.
Experiments: problem

• BERT has default maximum sequence length: 512.
• Feed to it paragraph by paragraph.
• Model only uses information of the input *per se*.
• Add external information.
Experiments: previous observation

Most factors are in heads and tails of fact-checks.
Experiments: modification

Replace [CLS] w/ paragraph position [HEAD]/[BODY]/[TAIL].
Experiments: previous observation

Power-law distribution of fact-checkers.
Experiments: data splitting

80% train, 10% dev, 10% test

# of fact-checks

Fact-checker

- politifact.com 3915
- factcheck.org 856
- washingtonpost.com
- factly.in 250
- africacheck.org 220
- Others 348
- Under-represented
Experiments: data splitting

- politifact.com: 3915
- factcheck.org: 856
- factly.in: 250
- africachek.org: 220
- Others: 348

80% train, 10% dev, 10% test

Only for test (generalization)
Experiments: evaluation

• ROUGE (F1, precision, recall)

• Tight score: if not tagged, ROUGE = 0.

• Loose score: only count if tagged.

• In a cell: tight score (loose score)
## Results: overall performance

<table>
<thead>
<tr>
<th>Lead token</th>
<th>Tagger</th>
<th><strong>Claim</strong> ROUGE-1</th>
<th><strong>Claimant</strong> ROUGE-1</th>
<th><strong>Verdict</strong> ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>.183 (1.183)</td>
<td>.300 (3.300)</td>
<td>.141 (1.141)</td>
</tr>
<tr>
<td>[CLS]</td>
<td>Fluent</td>
<td>.636 (8.53)</td>
<td>.669 (8.97)</td>
<td>.633 (8.50)</td>
</tr>
<tr>
<td></td>
<td>Concise</td>
<td>.592 (8.64)</td>
<td>.615 (8.97)</td>
<td>.596 (8.70)</td>
</tr>
<tr>
<td>Paragraph</td>
<td>Fluent</td>
<td>.638 (8.54)</td>
<td>.674 (9.02)</td>
<td>.637 (8.53)</td>
</tr>
<tr>
<td>position</td>
<td>Concise</td>
<td>.646 (8.66)</td>
<td>.664 (8.89)</td>
<td>.652 (8.73)</td>
</tr>
</tbody>
</table>
**Results:** overall performance

<table>
<thead>
<tr>
<th>Lead token</th>
<th>Tagger</th>
<th><strong>Claim</strong> ROUGE-1</th>
<th><strong>Claimant</strong> ROUGE-1</th>
<th><strong>Verdict</strong> ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>.183</td>
<td>.300 (.300)</td>
<td>.141 (.141)</td>
</tr>
<tr>
<td>[CLS]</td>
<td>Fluent</td>
<td>.636</td>
<td>.669 (.897)</td>
<td>.633 (.850)</td>
</tr>
<tr>
<td></td>
<td>Concise</td>
<td>.592</td>
<td>.615 (.897)</td>
<td>.596 (.870)</td>
</tr>
<tr>
<td>Paragraph</td>
<td>Fluent</td>
<td>.638</td>
<td>.674 (.902)</td>
<td>.637 (.853)</td>
</tr>
<tr>
<td>position</td>
<td>Concise</td>
<td>.646</td>
<td>.664 (.889)</td>
<td>.652 (.873)</td>
</tr>
</tbody>
</table>

- Poor performance of baseline methods.
Results: overall performance

<table>
<thead>
<tr>
<th>Lead token</th>
<th>Tagger</th>
<th>Claim ROUGE-1</th>
<th>Claimant ROUGE-1</th>
<th>Verdict ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>.183 (.183)</td>
<td>.300 (.300)</td>
<td>.141 (.141)</td>
</tr>
<tr>
<td>[CLS]</td>
<td>Fluent</td>
<td>.636 (.853)</td>
<td>.669 (.897)</td>
<td>.633 (.850)</td>
</tr>
<tr>
<td></td>
<td>Concise</td>
<td>.592 (.864)</td>
<td>.615 (.897)</td>
<td>.596 (.870)</td>
</tr>
<tr>
<td>Paragraph</td>
<td>Fluent</td>
<td>.638 (.854)</td>
<td>.674 (.902)</td>
<td>.637 (.853)</td>
</tr>
<tr>
<td>position</td>
<td>Concise</td>
<td>.646 (.866)</td>
<td>.664 (.889)</td>
<td>.652 (.873)</td>
</tr>
</tbody>
</table>

- Poor performance of baseline methods.
- Improved performance w/ vanilla BERT.
### Results: overall performance

| Lead token | Tagger     | ![Claim ROUGE-1](f1|precision|recall) | ![Claimant ROUGE-1](f1|precision|recall) | ![Verdict ROUGE-1](f1|precision|recall) |
|------------|-----------|---------------------------------|---------------------------------|---------------------------------|
| Baseline   |           | .183 (.183) | .300 (.300) | .141 (.141) | .237 (.237) | .181 (.181) | .352 (.352) | .660 (.660) | .638 (.638) | .702 (.704) |
| [CLS]      | Fluent    | .636 (.853) | .669 (.897) | .633 (.850) | .769 (.894) | .803 (.934) | .759 (.883) | .931 (.975) | .934 (.979) | .930 (.974) |
|            | Concise   | .592 (.864) | .615 (.897) | .596 (.870) | .784 (.907) | .789 (.913) | .783 (.906) | .938 (.971) | .940 (.973) | .938 (.970) |
| Paragraph  | Fluent    | .638 (.854) | .674 (.902) | .637 (.853) | .794 (.889) | .821 (.919) | .789 (.884) | .940 (.978) | .942 (.980) | .939 (.978) |
| position   | Concise   | .646 (.866) | .664 (.889) | .652 (.873) | .839 (.928) | .852 (.943) | .834 (.923) | .941 (.975) | .944 (.979) | .940 (.974) |

- Poor performance of baseline methods.
- Improved performance w/ vanilla BERT.
- Further improved performance w/ paragraph tokens.
### Results: overall performance

<table>
<thead>
<tr>
<th>Lead token</th>
<th>Tagger</th>
<th>Claim ROUGE-1</th>
<th></th>
<th></th>
<th>Claimant ROUGE-1</th>
<th></th>
<th></th>
<th>Verdict ROUGE-1</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>.183 (.183)</td>
<td>.300 (.300)</td>
<td>.141 (.141)</td>
<td>.237 (.237)</td>
<td>.181 (.181)</td>
<td>.352 (.352)</td>
<td>.660 (.660)</td>
<td>.638 (.638)</td>
<td>.702 (.704)</td>
</tr>
<tr>
<td>[CLS]</td>
<td>Fluent</td>
<td>.636 (.853)</td>
<td>.669 (.897)</td>
<td>.633 (.850)</td>
<td>.769 (.94)</td>
<td>.803 (.934)</td>
<td>.759 (.883)</td>
<td>.931 (.975)</td>
<td>.934 (.979)</td>
<td>.930 (.974)</td>
</tr>
<tr>
<td></td>
<td>Concise</td>
<td>.592 (.864)</td>
<td>.615 (.897)</td>
<td>.596 (.870)</td>
<td>.784 (.907)</td>
<td>.789 (.913)</td>
<td>.783 (.906)</td>
<td>.938 (.971)</td>
<td>.940 (.973)</td>
<td>.938 (.970)</td>
</tr>
<tr>
<td>Paragraph</td>
<td>Fluent</td>
<td>.638 (.854)</td>
<td>.674 (.902)</td>
<td>.637 (.853)</td>
<td>.794 (.889)</td>
<td>.821 (.919)</td>
<td>.789 (.884)</td>
<td>.940 (.978)</td>
<td>.942 (.980)</td>
<td>.939 (.978)</td>
</tr>
<tr>
<td>position</td>
<td>Concise</td>
<td>.646 (.866)</td>
<td>.664 (.889)</td>
<td>.652 (.873)</td>
<td>.839 (.928)</td>
<td>.852 (.943)</td>
<td>.834 (.923)</td>
<td>.941 (.975)</td>
<td>.944 (.979)</td>
<td>.940 (.974)</td>
</tr>
</tbody>
</table>

- **Claim**: ~0.65 (~0.85).
Results: overall performance

<table>
<thead>
<tr>
<th>Lead token</th>
<th>Tagger</th>
<th><strong>Claim</strong> ROUGE-1</th>
<th><strong>Claimant</strong> ROUGE-1</th>
<th><strong>Verdict</strong> ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>.183</td>
<td>.300</td>
<td>.141</td>
</tr>
<tr>
<td>[CLS]</td>
<td>Fluent</td>
<td>.636</td>
<td>.669</td>
<td>.633</td>
</tr>
<tr>
<td></td>
<td>Concise</td>
<td>.592</td>
<td>.615</td>
<td>.596</td>
</tr>
<tr>
<td>Paragraph</td>
<td>Fluent</td>
<td>.638</td>
<td>.674</td>
<td>.637</td>
</tr>
<tr>
<td>position</td>
<td>Concise</td>
<td>.646</td>
<td>.664</td>
<td>.652</td>
</tr>
</tbody>
</table>

- **Claim**: ~0.65 (~0.85).
- **Claimant**: ~0.8 (~0.9).
Results: overall performance

<table>
<thead>
<tr>
<th>Lead token</th>
<th>Tagger</th>
<th>Claim ROUGE-1</th>
<th>Claimant ROUGE-1</th>
<th>Verdict ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>.183 (.183)</td>
<td>.300 (.300)</td>
<td>.141 (.141)</td>
</tr>
<tr>
<td>[CLS]</td>
<td>Fluent</td>
<td>.636 (.853)</td>
<td>.669 (.897)</td>
<td>.633 (.850)</td>
</tr>
<tr>
<td></td>
<td>Concise</td>
<td>.592 (.864)</td>
<td>.615 (.897)</td>
<td>.596 (.870)</td>
</tr>
<tr>
<td>Paragraph</td>
<td>Fluent</td>
<td>.638 (.854)</td>
<td>.674 (.902)</td>
<td>.637 (.853)</td>
</tr>
<tr>
<td>position</td>
<td>Concise</td>
<td>.646 (.866)</td>
<td>.664 (.889)</td>
<td>.652 (.873)</td>
</tr>
</tbody>
</table>

- **Claim**: ~0.65 (~0.85).
- **Claimant**: ~0.8 (~0.9).
- **Verdict**: ~0.94 (~0.97).
Results: generalization

<table>
<thead>
<tr>
<th>Lead token</th>
<th>Tagger</th>
<th>Claim ROUGE-1</th>
<th>Claimant ROUGE-1</th>
<th>Verdict ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>.175</td>
<td>.372</td>
<td>.122</td>
</tr>
<tr>
<td>[CLS]</td>
<td></td>
<td>.444</td>
<td>.483</td>
<td>.443</td>
</tr>
<tr>
<td>Concise</td>
<td></td>
<td>.467</td>
<td>.522</td>
<td>.597</td>
</tr>
<tr>
<td>position</td>
<td>Concise</td>
<td>.527</td>
<td>.532</td>
<td>.559</td>
</tr>
</tbody>
</table>

- Deteriorated performance for under-represented fact-checkers.
Results: generalization

<table>
<thead>
<tr>
<th>Lead token</th>
<th>Tagger</th>
<th>Claim ROUGE-1</th>
<th>Claimant ROUGE-1</th>
<th>Verdict ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>.175 (.175)</td>
<td>.372 (.372)</td>
<td>.122 (.122)</td>
</tr>
<tr>
<td>[CLS]</td>
<td>Fluent</td>
<td>.444 (.725)</td>
<td>.483 (.788)</td>
<td>.443 (.724)</td>
</tr>
<tr>
<td></td>
<td>Concise</td>
<td>.386 (.713)</td>
<td>.406 (.748)</td>
<td>.406 (.749)</td>
</tr>
<tr>
<td>Paragraph</td>
<td>Fluent</td>
<td>.519 (.728)</td>
<td>.566 (.794)</td>
<td>.517 (.725)</td>
</tr>
<tr>
<td>position</td>
<td>Concise</td>
<td>.527 (.738)</td>
<td>.532 (.744)</td>
<td>.559 (.781)</td>
</tr>
</tbody>
</table>

- Deteriorated performance for under-represented fact-checkers.
- BERT w/ paragraph tokens still performs the best.
### Results: generalization

<table>
<thead>
<tr>
<th>Lead token</th>
<th>Tagger</th>
<th>Claim ROUGE-1</th>
<th>Claimant ROUGE-1</th>
<th>Verdict ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>.175 (.175)</td>
<td>.372 (.372)</td>
<td>.122 (.122)</td>
</tr>
<tr>
<td>[CLS]</td>
<td>Fluent</td>
<td>.444 (.725)</td>
<td>.483 (.788)</td>
<td>.443 (.724)</td>
</tr>
<tr>
<td></td>
<td>Concise</td>
<td>.386 (.713)</td>
<td>.406 (.748)</td>
<td>.406 (.749)</td>
</tr>
<tr>
<td>Paragraph position</td>
<td>Fluent</td>
<td>.519 (.728)</td>
<td>.566 (.794)</td>
<td>.517 (.725)</td>
</tr>
<tr>
<td></td>
<td>Concise</td>
<td>.527 (.738)</td>
<td>.532 (.744)</td>
<td>.559 (.781)</td>
</tr>
</tbody>
</table>

- **Claim**: ~0.5 (~0.7) from ~0.65 (~0.85).
- **Claimant**: ~0.4 (~0.7) from ~0.8 (~0.9).
- **Verdict**: ~0.4 (~0.8) from ~0.94 (~0.97).
### Results: improving generalization

<table>
<thead>
<tr>
<th>Train set</th>
<th>Tagger</th>
<th><strong>Claim</strong> ROUGE-1</th>
<th><strong>Claimant</strong> ROUGE-1</th>
<th><strong>Verdict</strong> ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Well-known ones only</td>
<td>Fluent</td>
<td>.519 (.728)</td>
<td>.566 (.794)</td>
<td>.517 (.725)</td>
</tr>
<tr>
<td></td>
<td>Concise</td>
<td>.527 (.738)</td>
<td>.532 (.744)</td>
<td>.559 (.781)</td>
</tr>
<tr>
<td>under-represented mixed</td>
<td>Fluent</td>
<td>.495 (.761)</td>
<td>.540 (.830)</td>
<td>.489 (.752)</td>
</tr>
<tr>
<td></td>
<td>Concise</td>
<td>.519 (.782)</td>
<td>.544 (.819)</td>
<td>.536 (.807)</td>
</tr>
</tbody>
</table>

- Mix half of under-represented fact-checkers to train.
### Results: improving generalization

<table>
<thead>
<tr>
<th>Train set</th>
<th>Tagger</th>
<th>Claim ROUGE-1</th>
<th></th>
<th>Claimant ROUGE-1</th>
<th></th>
<th>Verdict ROUGE-1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Well-known ones only</td>
<td>Fluent</td>
<td>.519</td>
<td>.566</td>
<td>.517</td>
<td>.377</td>
<td>.510</td>
<td>.342</td>
</tr>
<tr>
<td></td>
<td>Concise</td>
<td>.527</td>
<td>.532</td>
<td>.559</td>
<td>.462</td>
<td>.549</td>
<td>.436</td>
</tr>
<tr>
<td>under-represented</td>
<td>Fluent</td>
<td>.495</td>
<td>.540</td>
<td>.489</td>
<td>.550</td>
<td>.639</td>
<td>.528</td>
</tr>
<tr>
<td>mixed</td>
<td>Concise</td>
<td>.519</td>
<td>.544</td>
<td>.536</td>
<td>.575</td>
<td>.599</td>
<td>.581</td>
</tr>
</tbody>
</table>

- Mix half under-represented fact-checkers to train.
- Improved performance for **claimant** and **verdict**.
- Similar results for tagging **claim**.
Results: error analysis

• Not tagging: unseen patterns.
  e.g., long and unseen factors with explanations.
Results: error analysis

- Not tagging: unseen patterns.
  e.g., long and unseen factors with explanations.
- Wrongly tagging: confusing patterns.
  e.g., “(someone) claimed (...) on (date)” in a fact-check has a high likelihood of tagging as claim.
Results: error analysis

• Not tagging: unseen patterns.
e.g., long and unseen factors with explanations.

• Wrongly tagging: confusing patterns.
e.g., “(someone) claimed (...) on (date)” in a fact-check has a high likelihood of tagging as claim.

• Partially tagging: unusual patterns.
e.g., “the 45th and current president of the United States Donald Trump” as the claimant, our model tend to tag only “Donald Trump”.
Application: pre-population

Pre-population the fact-check markup tool:

• Enter article URL.
• Pre-populating factors.
• Check, revise, submit.
Conclusion: takeaways

• Proposed *factoring fact-checks*. 
Conclusion: takeaways

• Proposed *factoring fact-checks*.

• Observations from data exploration.
Conclusion: takeaways

- Proposed *factoring fact-checks*.
- Observations from data exploration.
- Applicable performance for well-known fact-checkers.
Conclusion: takeaways

• Proposed *factoring fact-checks*.
• Observations from data exploration.
• Applicable performance for well-known fact-checkers.
• Promising direction for under-represented fact-checkers.
Thank you!

Please send questions to: sjiang@ccs.neu.edu