Annotating and Modeling Fine-grained Factuality in Summarization

Tanya Goyal and Greg Durrett
NAACL 2021
Second-tier French club Nimes has landed in trouble after its former president was found guilty of trying to fix matches and arrested [...] French league disciplinary commission said on Tuesday that Jean-Marc Conrad tried to fix four matches [...] Seven games involving Nimes were investigated after the arrest last November. [...] French football has been hit with its first match-fixing scandal. Seven games involving Nimes were arrested last November.

- Fluent and grammatical text.
- Combines information from different parts of the input.
- World knowledge e.g. Nimes is a football club.
- Often hallucinates/ misinterprets information in the source.
Can we identify factual errors?

[...] Seven games involving Nimes were investigated after Conrad was arrested last November [...] + Seven games involving Nimes were arrested last November. (non-factual)
Second-tier French club Nimes has landed in trouble after its former president was found guilty of trying to fix matches and arrested [...] French league disciplinary commission said on Tuesday that Jean-Marc Conrad tried to fix four matches [...] Seven games involving Nimes were investigated after the arrest last November. [...] Seven games involving Nimes were investigated last November.
Overview

Evaluate Synthetic Factuality Datasets

Do synthetic datasets target the errors from summarization models?

- Seven games were being investigated.
- Nine games were being investigated.

No, synthetic datasets handle a limited set of error types.

Evaluate Modeling Formulations for Factuality

What granularity of factuality models are needed?

- summary-level
  - Nine games were being arrested.
- fine-grained
  - Nine games were being arrested.

Fine-grained works better, error localisation helps train better models!
Evaluating Synthetic Training Datasets

- Define a taxonomy of errors.
- Manually categorise errors in CNN/DM and XSUM model-generated summaries and synthetic datasets.
- Compare error distributions.
Evaluating Synthetic Training Datasets

- Define a taxonomy of errors.

Second-tier French club Nimes has landed in trouble after its former president was found guilty of trying to fix matches and arrested [...] French league disciplinary commission said on Tuesday that Jean-Marc Conrad tried to fix four matches [...] Seven games involving Nimes were investigated after the arrest last November. [...]
Evaluating Synthetic Training Datasets

- Manually categorise errors in CNN/DM and XSUM model-generated summaries and synthetic datasets.

- Artifically corrupt reference summary (Entity/No./Pronoun etc. Swap)

- Paraphrase and select low-prob. option as non-factual instance.

- Seven games involving Nimes were investigated last October.

- Last November, Nimes of games were investigated.

Seven games involving Nimes were investigated last November.

Entity-Centric Synthetic Data [1,2,3]

Generation-Centric Synthetic Data [4]

[1] Kryściński et al., EMNLP 2020
[2] Zhao et al., EMNLP Findings 2020
[3] Cao et al., EMNLP 2020
[4] Goyal et al., EMNLP Findings 2020
Evaluating Synthetic Training Datasets

- Manually categorise errors in CNN/DM and XSUM model-generated summaries and synthetic datasets.

**Entity-centric (Ent-C)**
- 50 corrupted summaries for Ent-C

**Generation-centric (Gen-C)**
- 50 corrupted summaries for Gen-C
- 100 generated summaries for XSUM
- 50 generated summaries for CNNDM

**CNN/DM XSUM**

**SOTA Summarization Model**
Error Analysis

- **Compare Error Distributions.**

  - **CNNNDM**
    - Extrinsic: 0.02
    - Intrinsic: 0.93
    - Others: 0.05

  - **XSum**
    - Extrinsic: 0.62
    - Intrinsic: 0.33
    - Others: 0.05

  - **ENT-C**
    - Extrinsic: 0.17
    - Intrinsic: 0.66
    - Others: 0.17

  - **GEN-C**
    - Extrinsic: 0.14
    - Intrinsic: 0.51
    - Others: 0.35

- **CNNNDM models primarily make intrinsic errors, XSUM makes extrinsic errors.**

- **Synthetic datasets target a different error distribution compared to real generation errors.**
Error Analysis

- Compare Error Distributions.

**CNNDM**

| Type    | Score |
|---------|-------|
| Extrinsic | 0.02  |
| Intrinsic | 0.93  |
| Others   | 0.05  |

**ENT-C**

| Type    | Score |
|---------|-------|
| Extrinsic | 0.17  |
| Intrinsic | 0.66  |
| Others   | 0.17  |

**CNNDM Intrinsic**

| Type    | Score |
|---------|-------|
| Entity  | 0.21  |
| Event   | 0.50  |
| NP      | 0.22  |

**ENT-C Intrinsic**

| Type    | Score |
|---------|-------|
| **mostly pronoun swap errors** | 0.00  |
| Entity  | 0.40  |
| Event   | 0.26  |

Different!
How do models on synthetic data perform?

- Train sentence-level models on synthetic datasets (50k ex).

- Test Data: Human-annotated test set for XSUM and CNNDM [1,2].

[1] Kryściński et al., EMNLP2020
[2] Maynez et al., ACL2020
Results

- Metric: Label Balanced Classification Accuracy

| CNNDM          | Training Data | Accuracy |
|----------------|---------------|----------|
|                | Ent-C         | 72.3     |
|                | Gen-C         | 64.4     |

| XSUM           | Training Data | Accuracy |
|----------------|---------------|----------|
|                | Ent-C         | 50.9     |
|                | Gen-C         | 54.2     |

- Close to majority label performance!

Do synthetic datasets target the errors from summarization models? **No**, synthetic datasets handle a limited set of error types.

(Fortunately, we have human-annotated data! More on this later)
Overview

Evaluate Synthetic Factuality Datasets

Do synthetic datasets target the errors from summarization models?

No, synthetic datasets handle a limited set of error types.

Seven games were being investigated.

Nine games were being investigated.

Evaluate Modeling Formulations for Factuality

What granularity of factuality models are needed?

summary-level annotations

Nine games were being arrested.

fine-grained annotations

Nine games were being arrested.

Fine-grained works better, error localisation helps train better models!
Evaluate Modeling Formulations for Factuality

Compare two kinds of models:

- Summary-level annotations ( Nine games were being arrested. )
- Fine-grained annotations ( Nine games were being arrested. )
Evaluating Factuality in Generation with Dependency-level Entailment
Goyal and Durrett, Findings of EMNLP2020

Seven games involving Nimes were investigated after Conrad was arrested last November.

For each arc, is the relationship defined by that dependency arc entailed by the input?

Arc-level entailment decisions are independent, helps localization!
DAE model

- Concat input and output and encode.
- Parse the output to obtain dependency arcs.
- For each dependency arc, compute arc representation.
- Predict arc level entailment.

Seven games involving Nimes were investigated after Conrad was arrested last November.

Input: Seven games involving Nimes were investigated after Conrad was arrested last November.

Pre-trained Encoder Model

[SEP] Seven games involving Nimes were arrested.

Generated Summary
Training

What do we need?

- We use human-annotated training dataset with span highlighting of non-factual parts. [1]

An 18th century coin believed to be worth more than #1m has been discovered.

[input, summary] pairs with arc-level factuality labels.

[1] Maynez et al., ACL2020
Results: XSUM

- Small human annotated training data provides better supervision than large synthetic datasets.
- Fine-grained factuality modeling and annotations outperform sentence-level counterpart.

| Training Data | Accuracy |
|---------------|----------|
| Ent-C         | 50.9     |
| Gen-C         | 54.2     |
| Sent-level    | 65.6     |
| DAE           | 78.7     |

Human-annotated Training Data
Improving Summarization Models

Error localization (via DAE) can help de-noise noisy summarization training data like XSUM!

- Train models by maximizing the log likelihood of “correct” words only.

| Model       | Avg. score |
|-------------|------------|
| Baseline    | 0.37       |
| DAE-based   | 0.46       |

Reference summary is technically true but is unsupported by the input.

A medieval coin believed to be worth more than $1m has [...]
Takeaways

- Existing synthetic datasets are not aligned with actual generation errors of summarization models, especially in challenging domains like XSUM.
- Fine-grained human annotation data can lead to better factuality models, as well as enable training of more factual summarization models!

Thank you!