Annotating and Modeling Fine-grained Factuality in Summarization

Tanya Goyal and Greg Durrett

NAACL 2021

News summarization

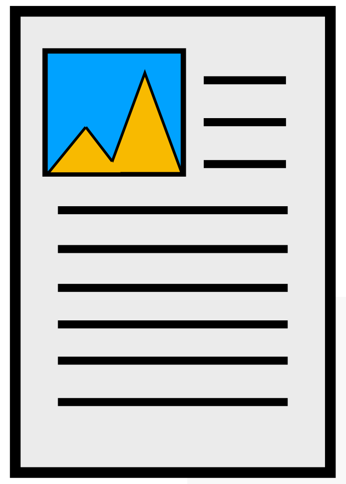
Second-tier French club Nimes has landed in trouble after its former president was found guilty of trying to fix matches and arrested […] French league disciplinary commission said on Tuesday that Jean-Marc Conrad tried to fix four matches […] Seven games involving Nimes were investigated after the arrest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

French football has been hit with its first match-fixing scandal.

Seven games involving Nimes were ~~arrested~~ investigated last November.

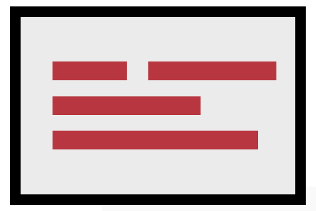
- ✓ Fluent and grammatical text.
- ✓ Combines information from different parts of the input.
- ✓ World knowledge e.g. *Nimes is a football club.*
- ✗ Often hallucinates/ misinterprets information in the source.

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)

Prior Work: Synthetic training datasets

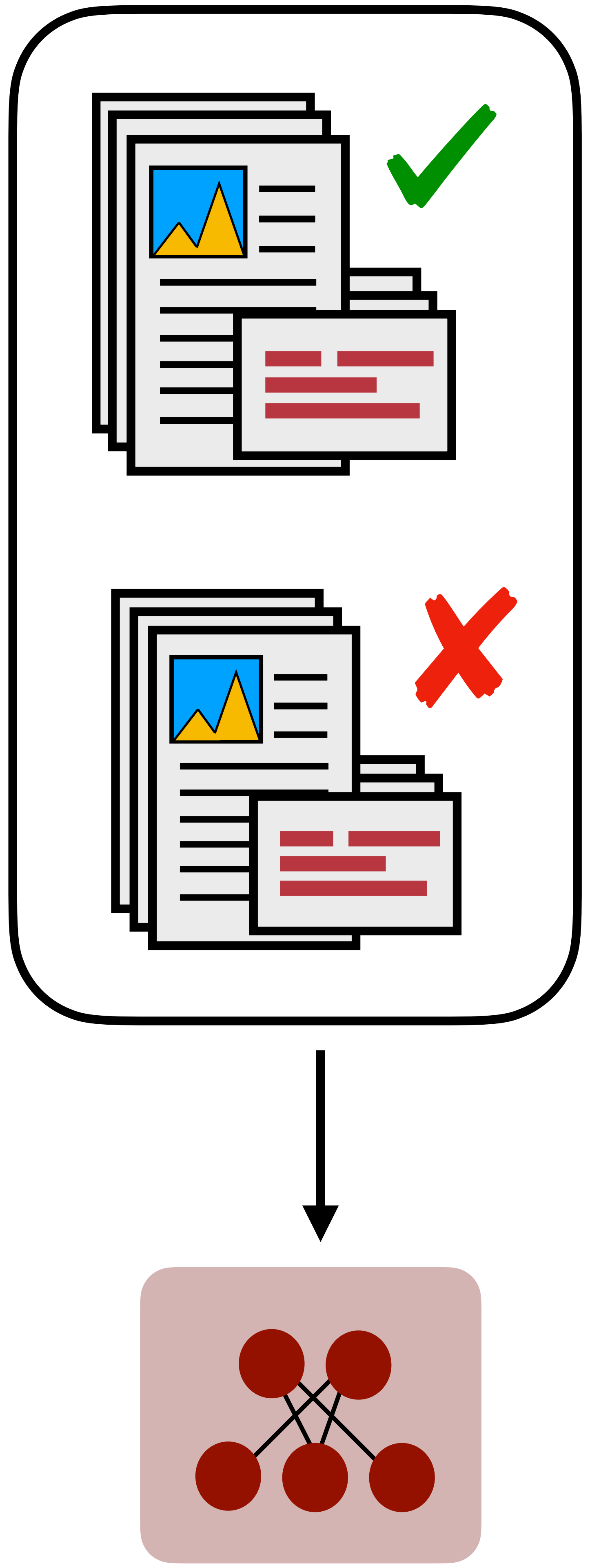
Second-tier French club Nimes has landed in trouble after its former president was found guilty of trying to fix matches and arrested […] French league disciplinary commission said on Tuesday that Jean-Marc Conrad tried to fix four matches […] Seven games involving Nimes were investigated after the arrest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

Seven games involving Nimes were investigated last November.

Artificial Corruption

Nine games involving Nimes were investigated last November.

Entity-swap | Noise Injection | Negation | …



Factuality Model

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.

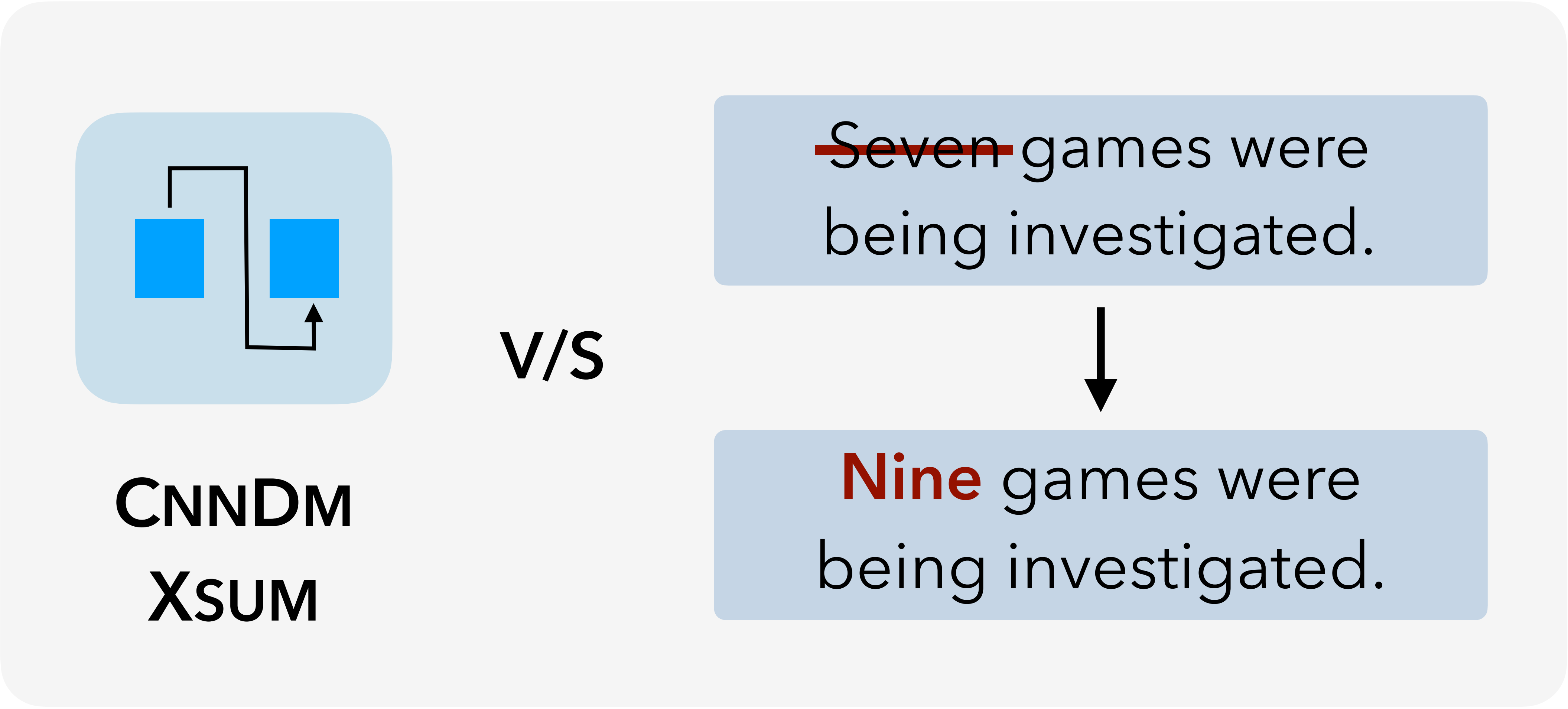
Evaluate Modeling Formulations for Factuality

What granularity of factuality models are needed?

summary-level ✗		fine-grained
Nine games were being arrested.	V/S	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

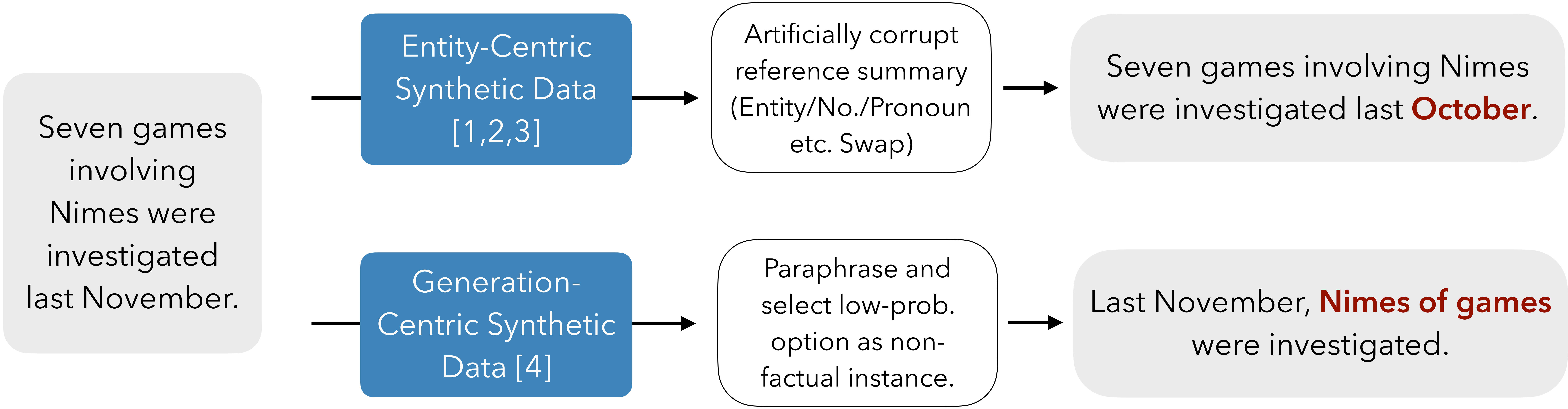
- **Define a taxonomy of errors.**

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. […]

	Extrinsic New information introduced	**Intrinsic** Information in the article misinterpreted.
Entity Related		**Conrad Marc** was arrested last November for …
Event Related	… trying to fix matches **by bribing players** …	
Noun-Phrase Related		
Others (Noise/Grammar)	Nimes **has has** landed in trouble after its former president …	

Evaluating Synthetic Training Datasets

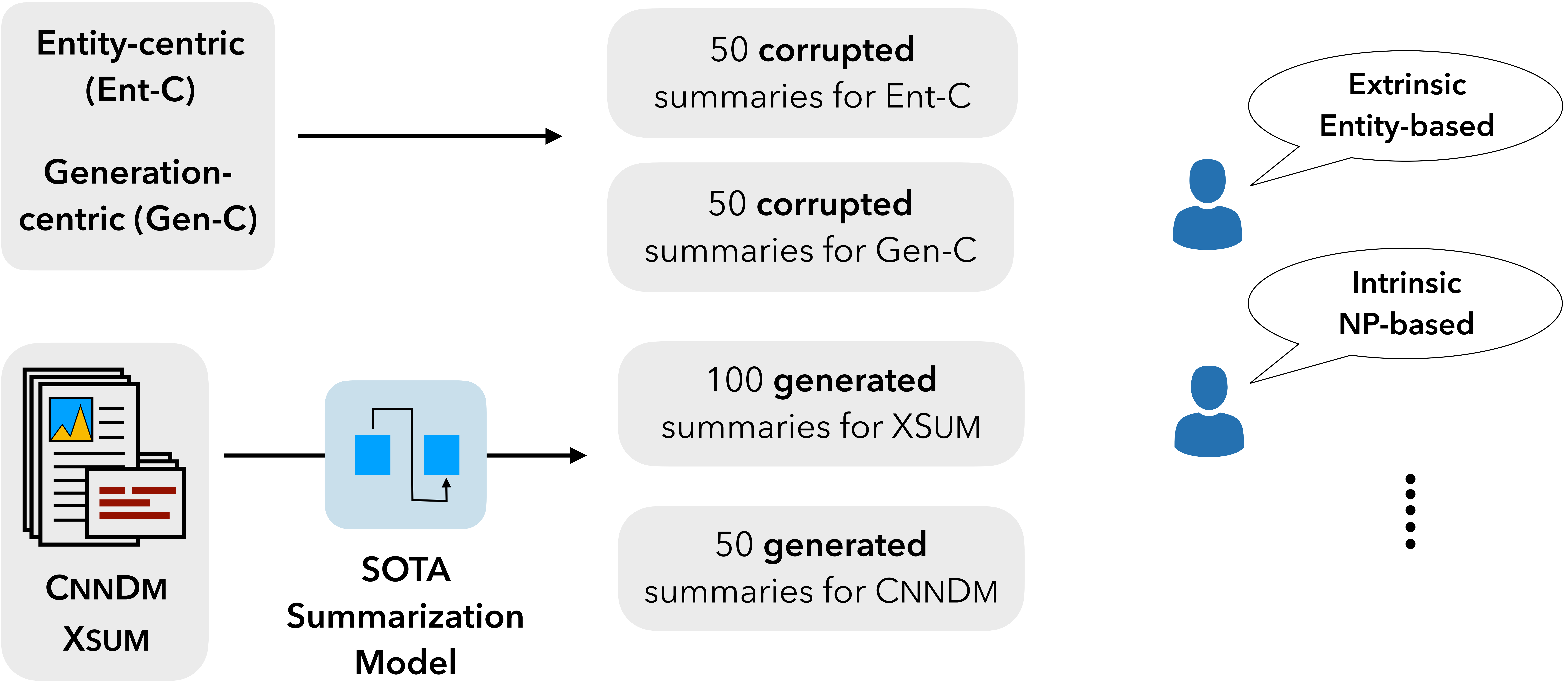
- **Manually categorise errors in CNN/DM and XSUM model-generated summaries and synthetic datasets.**



[1] Kryściński et al., EMNLP2020
[2] Zhao et al., EMNLP Findings 2020
[3] Cao et al., EMNLP 2020
[4] Goyal et al., EMNLP Findings 2020

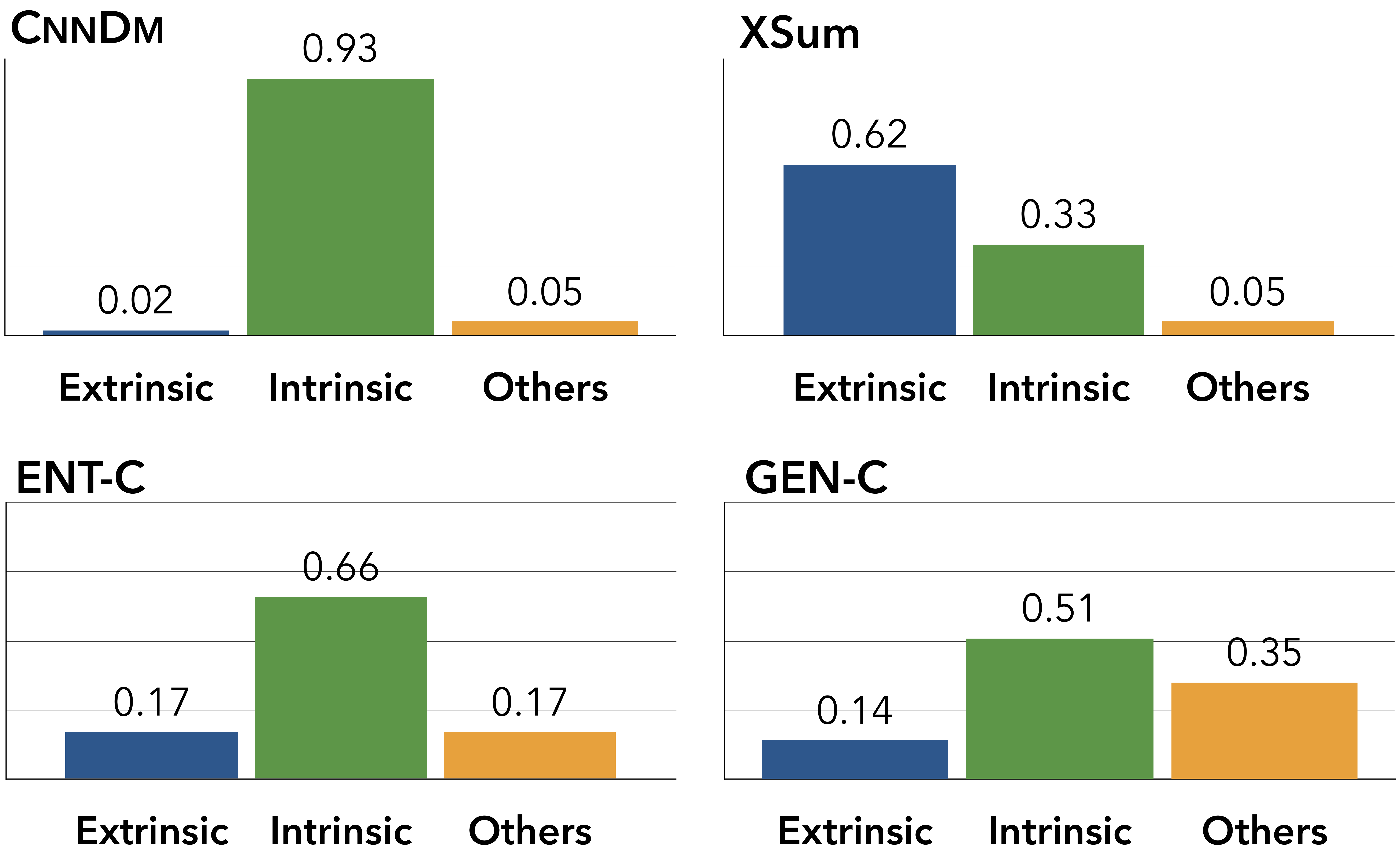
Evaluating Synthetic Training Datasets

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



Error Analysis

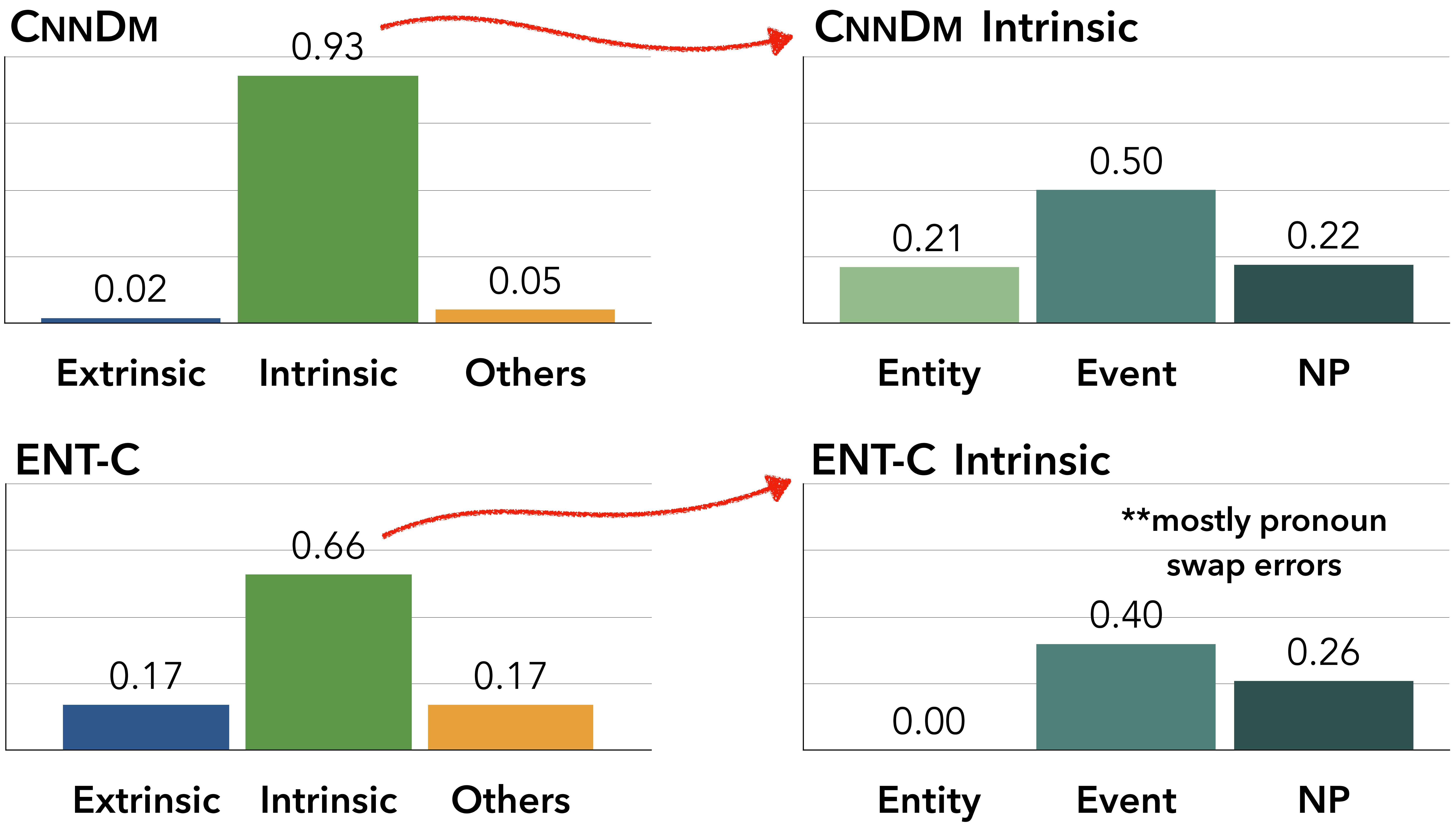
- **Compare Error Distributions.**



- CNNDM models primarily make intrinsic errors, XSUM makes extrinsic errors.
- Synthetic datasets target a different error distribution compared to real generation errors.

Error Analysis

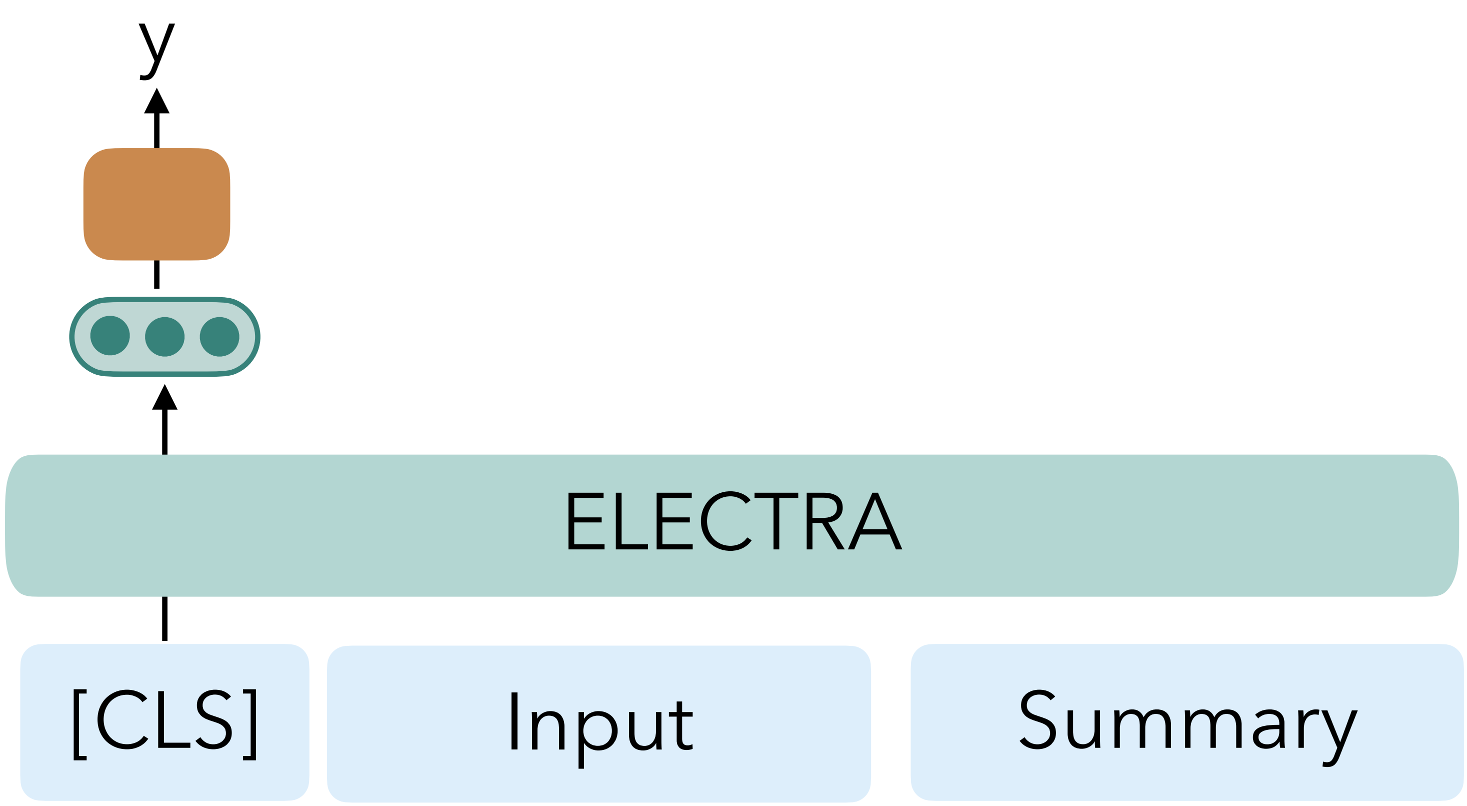
- **Compare Error Distributions.**



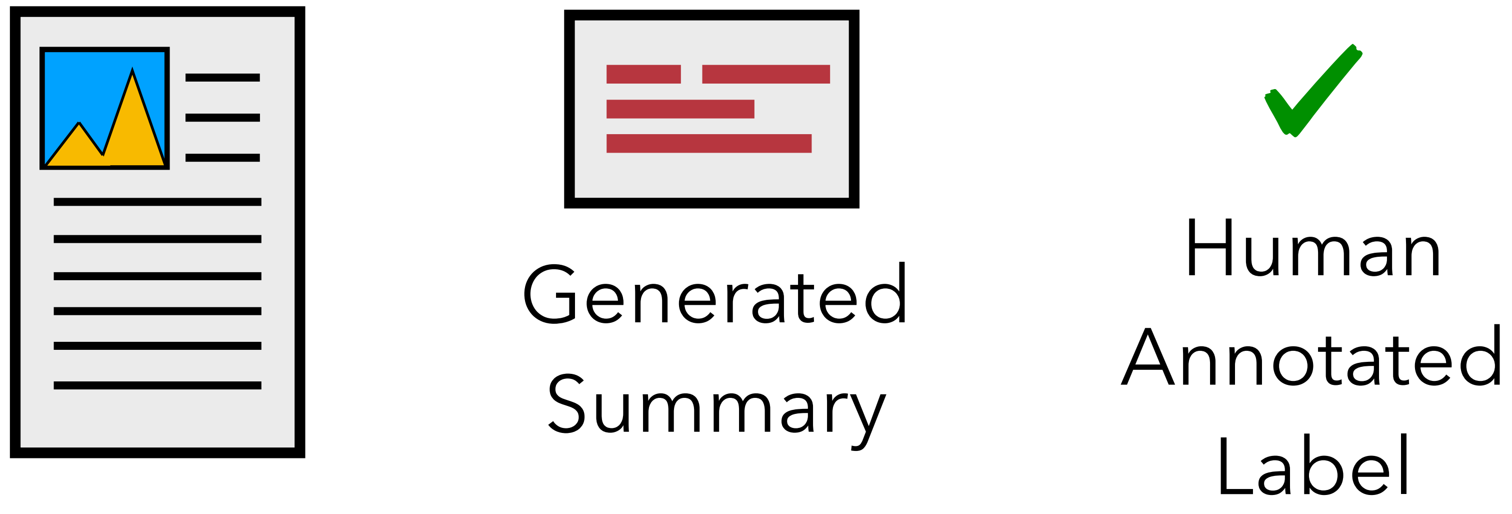
Different!

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?

~~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 annotations ✗		fine-grained annotations
Nine games were being arrested.	V/S	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.

v/s

fine-grained annotations

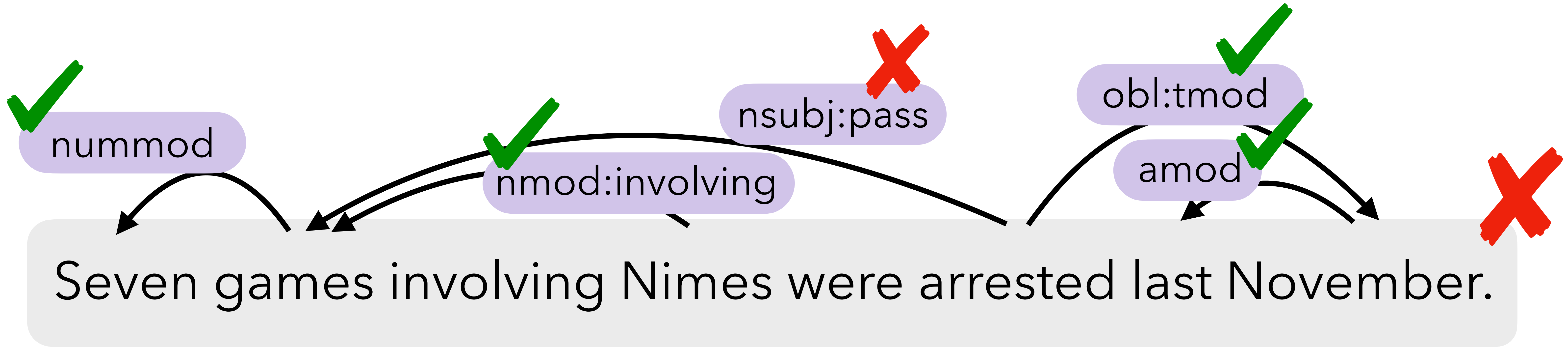
Nine games were being arrested.

Error Localisation Model: DAE

Evaluating Factuality in Generation with Dependency-level Entailment
Goyal and Durrett, Findings of EMNLP2020

Localizes error at the dependency arc level

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

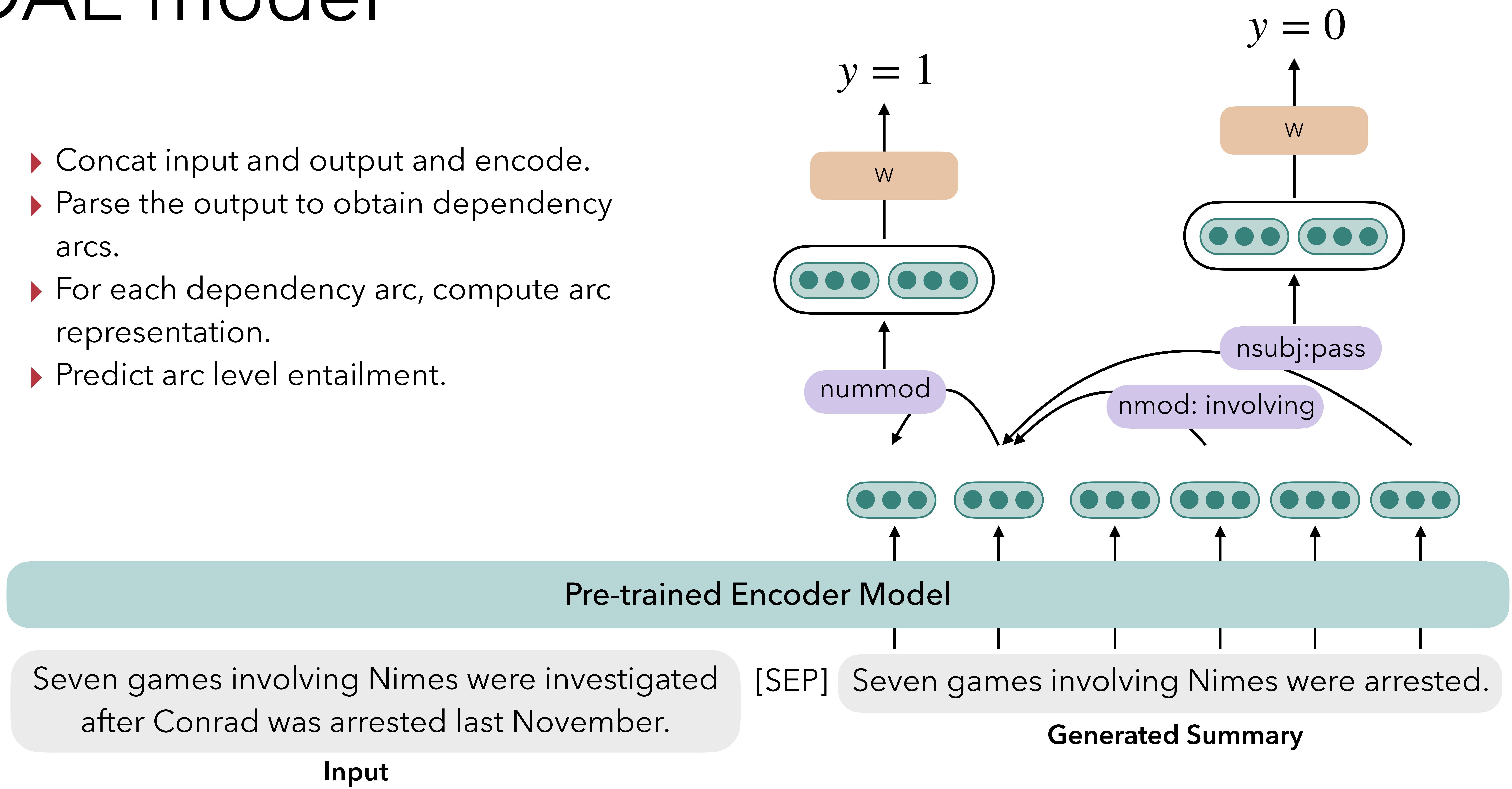


For each arc, is the relationship defined by that dependency arc entailed by the input?

Arc-level entailment decisions are independent, helps localization!

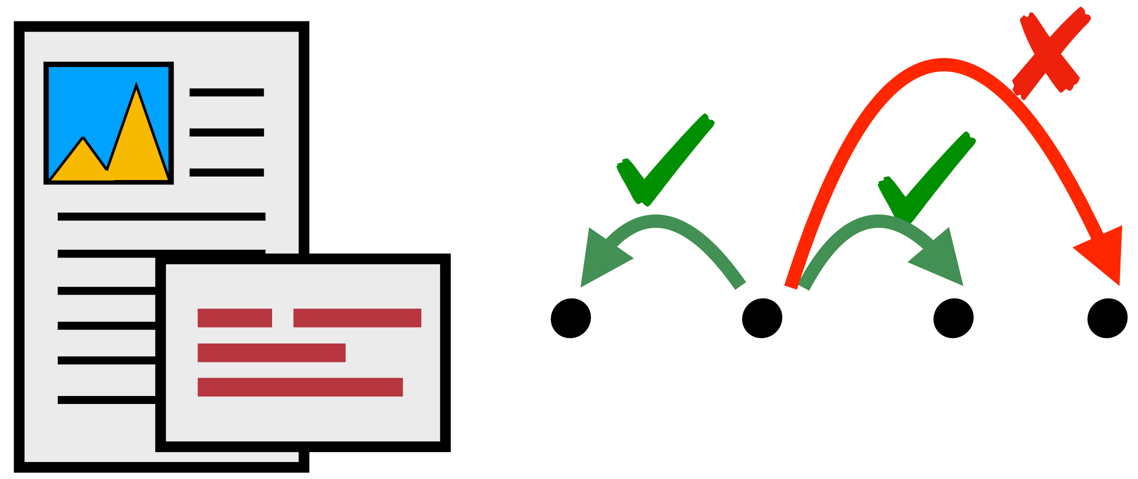
DAE model

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



Training

What do we need?



(input, summary) pairs with
arc-level factuality labels.

- 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.

[1] Maynez et al., ACL2020

Results: XSUM

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

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

Improving Summarization Models

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

A medieval coin believed to be worth more than #1m has […]

Reference Summary

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

Can be identified by DAE

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

Model	Avg. score
Baseline	0.37
DAE-based	0.46

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!