Improving Faithfulness of Abstractive Summarization Models

Tanay Dixit[1], Fei Wang[2], Ehsan Qasemi[2] , Muhao Chen[2]

1: Indian Institute of Technology Madras, 2: USC Information Science Institute

USC Viterbi School of Engineering Information Sciences Institute

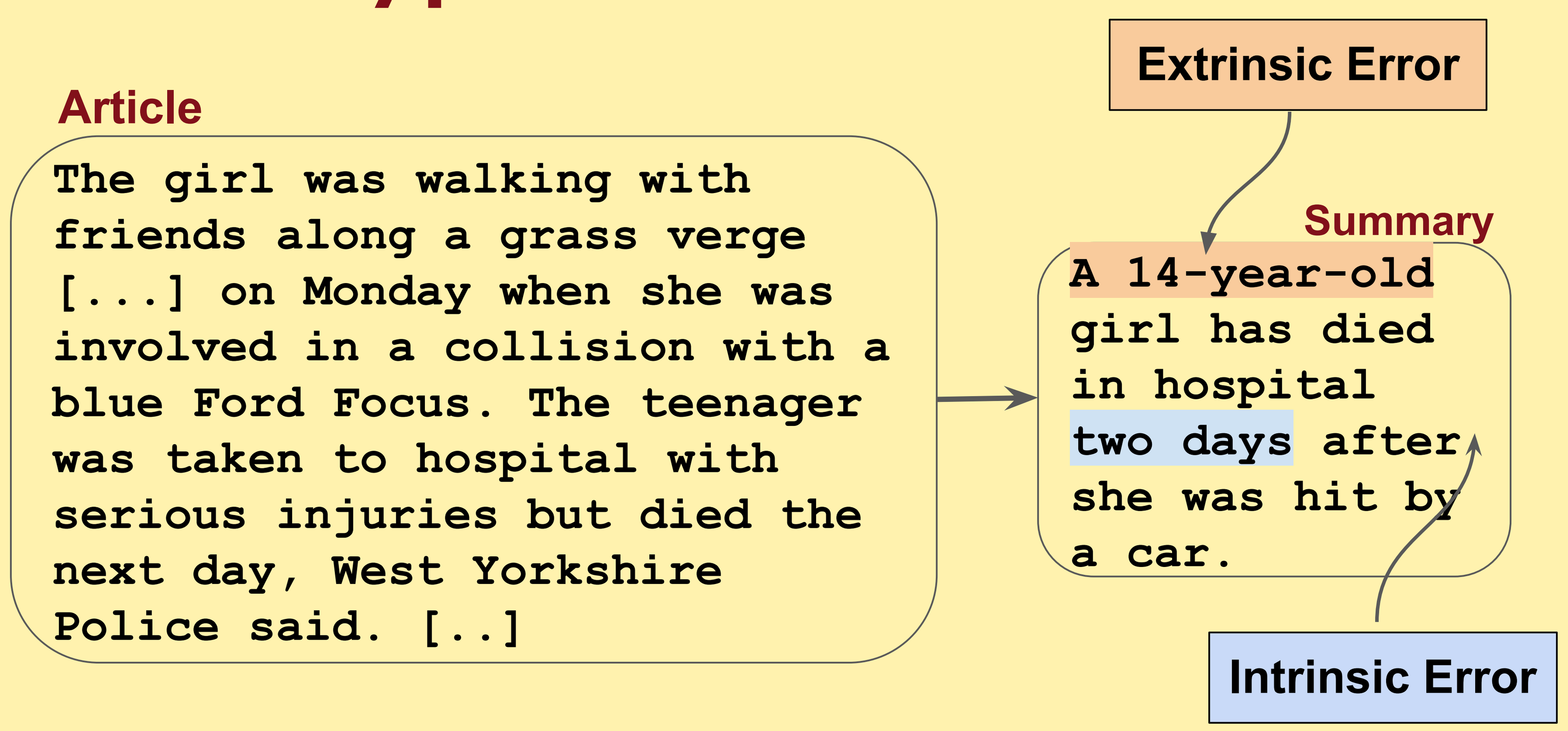
Work performed under REU Site program supported by NSF grant #2051101

Problem Statement

A recent study (Kryscinski et al. (2020)) showed that approx **30%** of the summaries generated my state of the art models were **factuality inconsistent**.

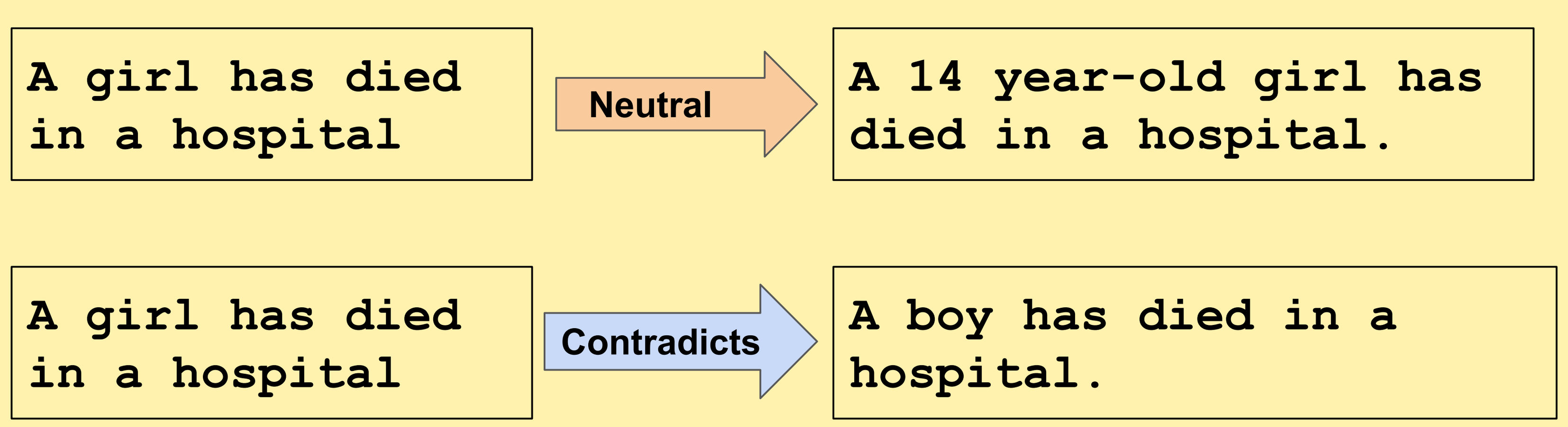
Need? Improving model factuality is key in its wide spread use on various platforms, as factuality is the most critical component

Error Types



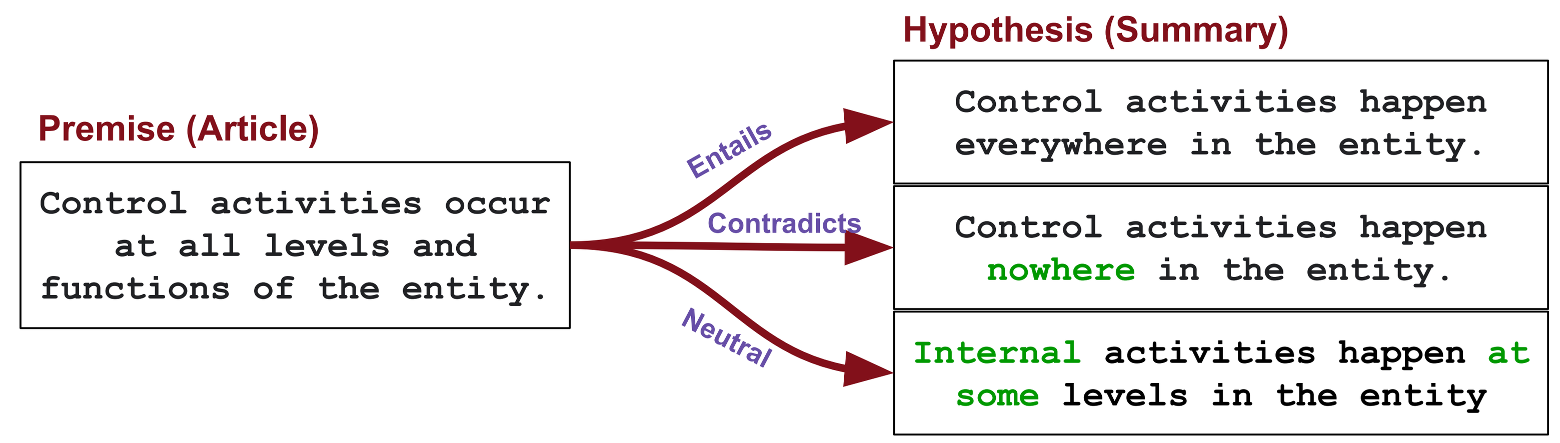
In this work we focus on reducing the extrinsic errors made by summarization models. To achieve this we introduce a **new auxiliary loss** as the standard MLE loss is not capable of capturing extrinsic errors.

Connecting NLI & Summarization



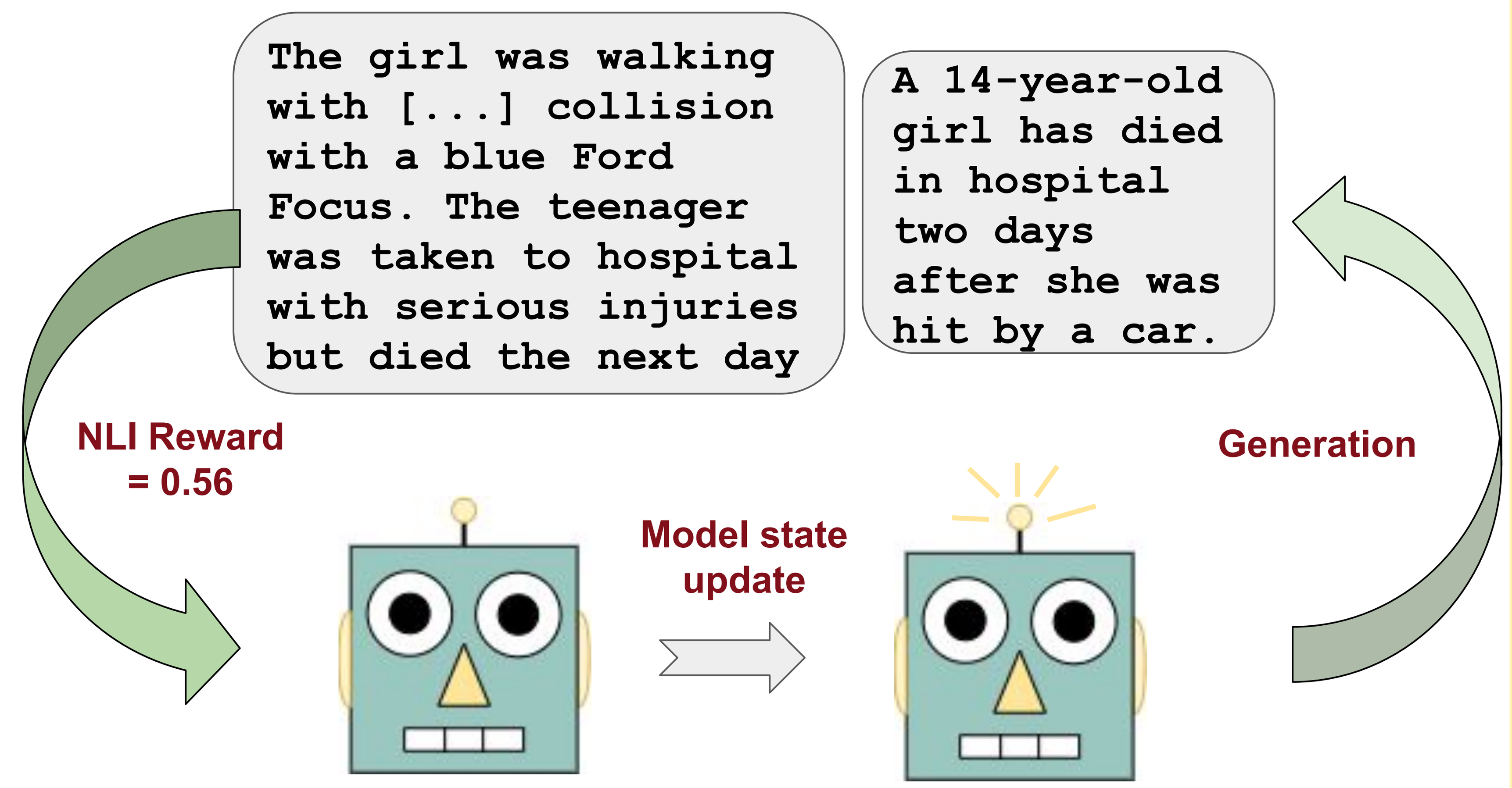
Classifying errors as a Natural Language Inference (NLI) task

NLI Reward



We train a document level NLI model using synthetic data [3].Given the article, the reward model is trained to classify the summary as either faithful or not.

Approach



REINFORCE Algorithm

$$\text{loss} = (r(w)^s - r(w)^b) * \nabla_\theta log\ p_\theta(w^s)$$

$$r(w)^s = NLI(\text{article, sampled summary})$$

$$r(w)^b = NLI(\text{article, baseline summary})$$

Reward Analysis

Reward Model	AggreFact-Xsum (full)	AggreFact-Xsum (SoTA)
SummaC	64.35	56.10
MNLI + Falsesum	67.40	58.44
ANLI + Falsesum	73.40	63.78

Results

Model	Rouge-L	FactCC	QEval
PEGASUS	39.07	25.48	32.50
CLIFF	38.18	25.18	33.21
FaithPEGA (ours)	38.34	26.34	33.20

Conclusions/ Future Works

- Using document level reward signals can be better for tasks where the gold summaries itself have errors
- Prior works have focused on reducing factuality errors but have not dug deeper to understand which category of errors are reducing. Having this knowledge can help the community in developing faithful systems.
- Since our reward model was trained on news related corpus we could only experiment with news datasets but it would be interesting to try our approach out on other domain like email/ medical reports summarization

[1] Evaluating the Factual Consistency of Abstractive Text Summarization",Kryscinski et al EMNLP 2020
[2] Policy Gradient Methods for Reinforcement Learning with Function Approximation Sutton et al Neurips 1999
[3] Falsesum: Generating Document-level NLI Examples for Recognizing Factual Inconsistency in Summarization Utama et al arxiv 2022
[4] SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization Laban et al 2021 TACL