# Evaluating the Factual Correctness for Abstractive Summarization

Yuhui Zhang*
yuhuiz@stanford.edu

## The Problem

- Text summarization: extractive, abstractive.
- Applications: news, laws, clinical, biomedical.
- However, **30%**[1] of summaries generated by abstractive models contain factual inconsistencies.

This is a **critical** issue for neural abstractive summarization.

**How can we evaluate the factual correctness?**

## Abstractive Summarization

Most recent works about abstractive summarization are based on sequence-to-sequence (seq2seq) architecture:

***Seq2Seq***: Basic seq2seq architecture.
***Pointer-Generator***[2]: Allow to copy from source text.
***ML***[3]: Attend over source and target text separately.
***ML+RL***[3]: Training with reinforcement learning.

Summaries are generated and sampled from CNN/DM dataset using these models.[4]

## Factual Score

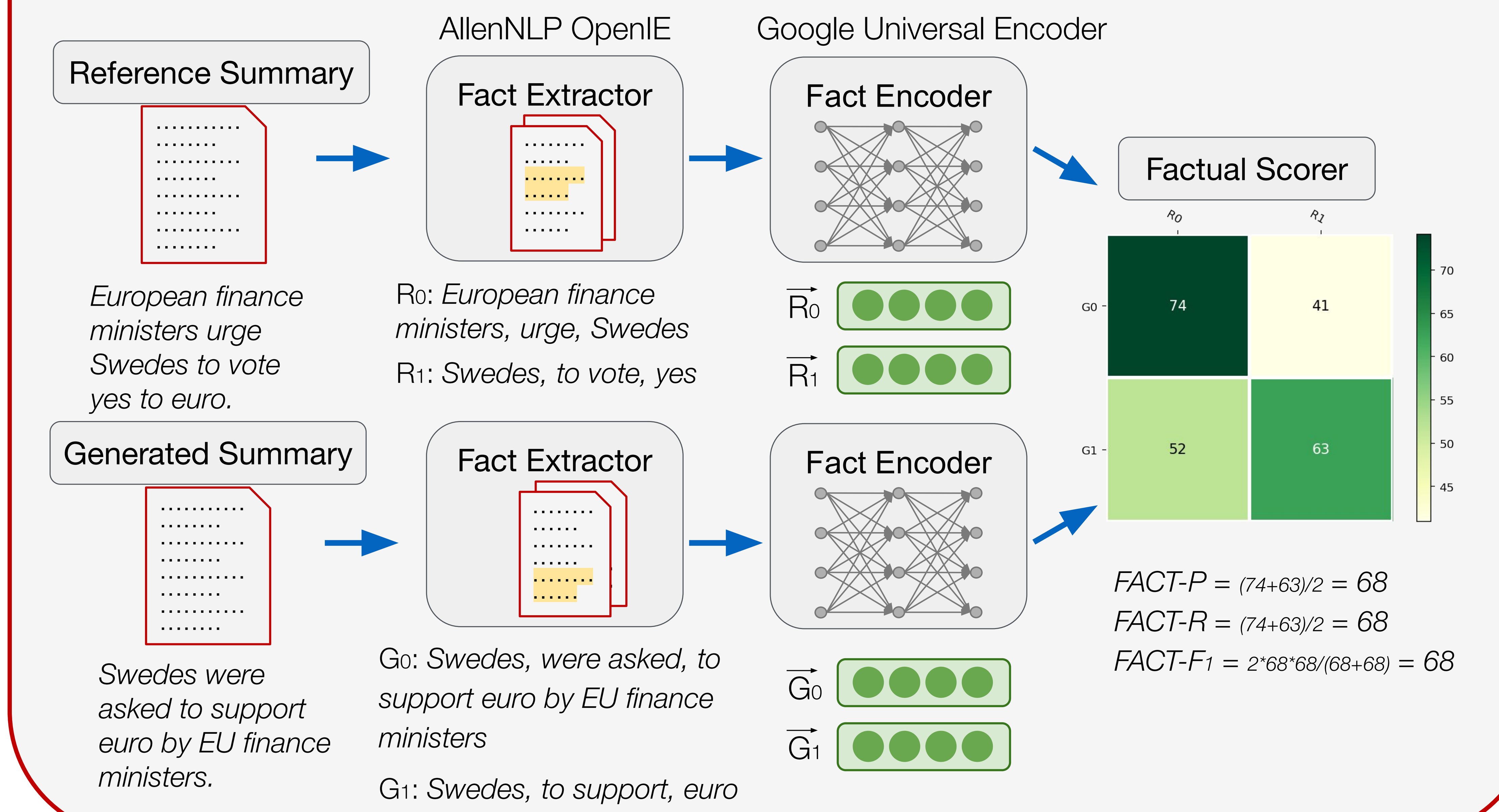
***Fact Extractor***: we use AllenNLP open information extraction (OpenIE) toolkit to extract facts from text. Each fact is a triple (argument, predicate, argument).

***Fact Encoder***: We concatenate the fact triple and use Google universal sentence encoder to generate fact embedding.

***Factual Scorer***: We use cosine-similarity to estimate the relevance of each fact pair, and then compute precision, recall and F1 by averaging across facts from generated summary and facts from reference summary.

## Factual Score

The overview of **factual score** computation:

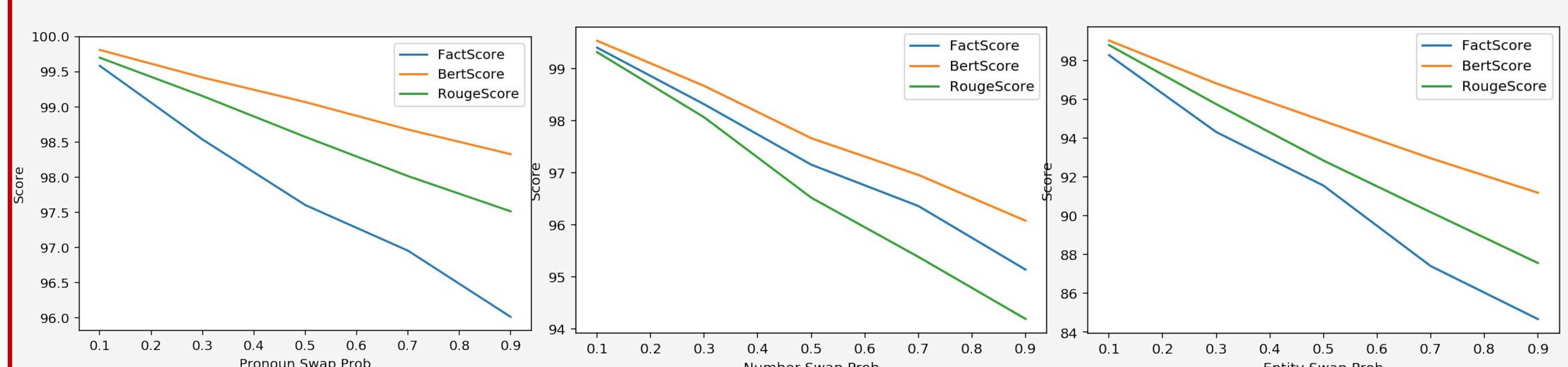


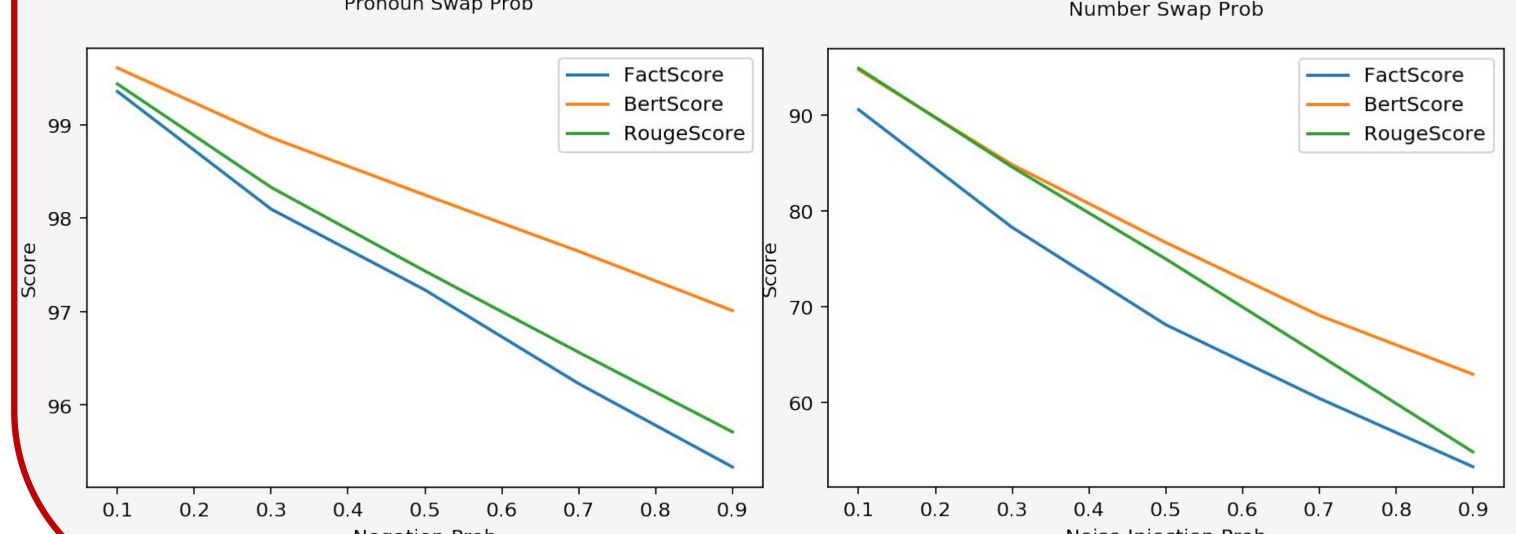

## Falsity Attack

We manually generate false examples with 5 simple text transformations:

Truth: Andrew is a professor at Stanford, and he teaches CS 230 for many years.
Falsity: Andrew is not a professor at Berkeley, and she teaches CS 231 for many years years.

negation | entity swap | pronoun swap | number swap | noise injection




**X-axis: transmation prob**
**Y-axis: factual score**

**Metric Sensitivity:**
FACT > ROUGE > BERT

**Falsity Sensitivity:**
Noise Injection > Entity Swap > Negation > Number Swap > Pronoun Swap

## Results

**Evaluations of abstractive summarization with…**

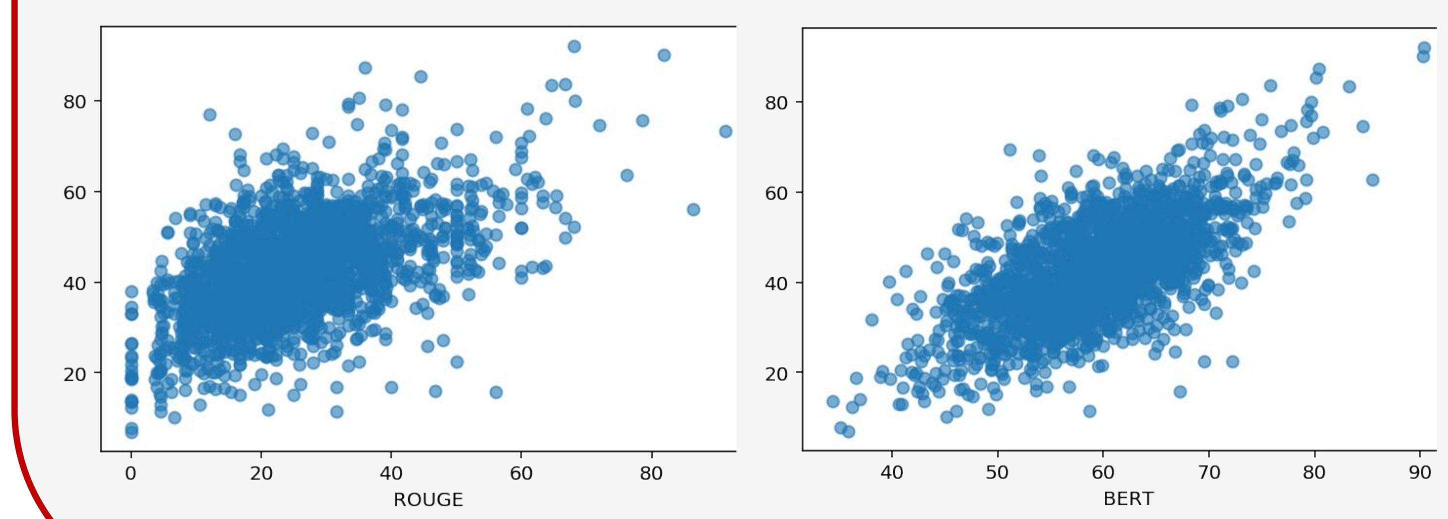
- ***ROUGE-L Score*** (n-gram hard-match evaluation)
- ***BERT Score***[5] (token soft-match evaluation)
- ***Factual Score*** (factual correctness evaluation)

| System | ROUGE | BERT | FACT |
|---|---|---|---|
| Seq2seq | 19.94 | 55.01 | 39.61 |
| Pointer-Generator | 27.62 | 60.20 | 43.49 |
| ML | 26.57 | 60.35 | 42.83 |
| ML+RL | 28.63 | 61.72 | 45.13 |

*Factual score is consistent with human evaluation: ML+RL > Pointer-Generator >≈ ML > Seq2seq*

**Relation of factual score with …**

- ***ROUGE-L Score***
- ***BERT Score*** *(more strongly correlated)*




## Discussion and Future Work

- Encoder is much more sensitive to noun phrases than number, pronoun and negation → Design better **fact encoder** architecture.
- OpenIE outputs contain duplicated facts and noisy facts → Try different ways to **denoise** OpenIE outputs.
- **Reinforcement learning** on factual score.

* Research project with Yuhao Zhang and Christopher D Manning.

[1] Kryscinski, Wojciech, et al. *Neural Text Summarization: A Critical Evaluation.* In EMNLP-IJCNLP (2019).
[2] See, Abigail, Peter J. Liu, and Christopher D. Manning. *Get To The Point: Summarization with Pointer-Generator Networks.* In ACL (2017).
[3] Paulus, Romain, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In ICLR (2018).
[4] Chaganty, Arun, Stephen Mussmann, and Percy Liang. *The price of debiasing automatic metrics in natural language evaluation.* in ACL (2018).
[5] Zhang, Tianyi, et al. *BERTScore: Evaluating Text Generation with BERT.* arXiv:1904.09675.