# FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization

2020 ACL

**2022.02.17 장준원**

IMLL
Industrial Machine Learning Lab

# Content

1. Introduction (Faithfulness?)
2. The Abstractive-Faithfulness Tradeoff
3. FEQA : Faithfulness Evaluation with Question Answering
4. Experiments
5. Related Work
6. Conclusion

IMLL
Industrial Machine Learning Lab

# Introduction

## ❖ Faithfulness (충실성)

- is the concept of unfailingly remaining loyal to someone or something

## ❖ Faithfulness in Summarization?

- Generate content **consistent** with the source document
- Recent studies have shown that 30% of generated summaries contain unfaithful information (inconsistent)
- Example

| |
|---|
| **Source.** The world's oldest person has died a few weeks after celebrating her 117th birthday. Born on March 5, 1898, the great-grandmother had lived through two world wars, the invention of the television and the first successful powered aeroplane flight by the wright brothers... |
| **Output sentence.** The world 's oldest person has died on March 5, 1898. |

# Introduction

- Current models are limited by a trade-off between abstractiveness and faithfulness
- Diverse set of existing automatic evaluation metrics such as ROUGE, BERTScore, and learned entailment models'correlations with human scores of faithfulness drop significantly on highly abstractive summaries
- Proposed automatically generated QA pairs framework to represent information in the summary and validate it against the source.
- FEQA has significantly higher correlation with human scores of faithfulness and is the only metric that correlates with human scores on highly abstractive summaries from XSum dataset.

# The Abstractive-Faithfulness Tradeoff

❖ **Authors shows the tradeoff relationship between Abstractive and Faithfulness (요약성이 강할수록 본문내 1개의 문장(혹은 문장의 일부를) 그대로 가져오지 않는다) answering the following questions**

- How to quantify abstractiveness of a summary? (Abstractiveness의 정량화)
- Is abstractiveness encouraged more by the data or the model? (데이터 때문? 모델 때문?)
- How does being abstractive affect faithfulness? (trade-off 관계 명시)

# The Abstractive-Faithfulness Tradeoff

## ❖ Characterizing Abstractiveness of a Summary (Abstractiveness 정량화)

- Abstractive summary : **rephrasing** important content into **brief** statement
- A **more abstractive summary** sentence aggregates content over a **larger chunk of source text**= consequently it must **copy fewer words** to maintain brevity.
- Introduce 5 terms to measure abstractiveness (*** defines level of abstractiveness)**
  - Sentence Extraction (src 문장 그대로 추출) *
  - Span Extraction (src 문장 내 span 그대로 추출) **

    Src - *"the plane was coming back from the NCAA final, according to spokesman John Twork"*

    Summary - "the plane was coming back from the NCAA final"
  - Word Extraction (src 문장 내 token 그대로 추출) ***

    Src - *"Capybara Joejoe who lives in Las Vegas has almost 60,000 followers on Instagram"*

    Summary - *"Capybara Joejoe has almost 60,000 followers"*
  - Perfect Funsion_k (k개의 src 문장 내 token 그대로 추출) ****

    Src - *"Capybara Joejoe lives in Las vegas." and "He has almost 60,000 followers on Instagram."*

    Summary - *"Capybara Joejoe has almost 60,000 followers*
  - Novel-n-gram (src에서 등장하지 않은 n-gram) ****

# The Abstractive-Faithfulness Tradeoff

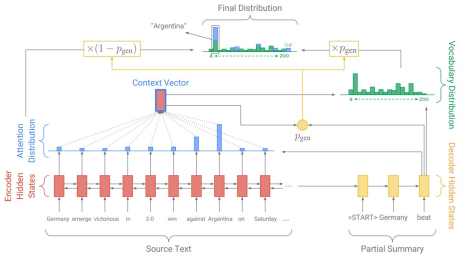
## ❖ Is abstractiveness from the model or the data?

- abstractiveness scores (5 terms in previous slide) for both the **reference summaries** and summaries generated (**model summaries**) from a diverse set of models on two datasets.

  - Datasets
    - CNN/DM (about 3 summary sentences)
    - XSUM (1 summary sentence = highly abstractive)

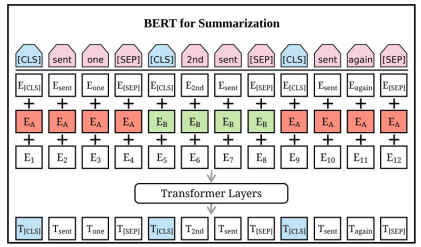
| | CNN/DM | XSum |
|---|---|---|
| # Training Documents | 287,227 | 204,045 |
| # Validation Documents | 13,368 | 11,332 |
| # Test Documents | 11,490 | 11,334 |
| Document: avg # of tokens | 781.00 | 431.07 |
| Document: avg # of sents. | 40.00 | 33.00 |
| Summary: avg # tokens | 56.00 | 23.26 |
| Summary: avg # of sents. | 3.75 | 1.00 |

| Systems | Extractor | Encoder | Decoder |
|---|---|---|---|
| PGC | – | LSTM | LSTM+copy |
| FASTRL | sentences | LSTM | LSTM+copy |
| BOTTOMUP | words | LSTM | LSTM+copy |
| TCONV | – | CNN+topic | CNN |
| BERTSUM | – | BERT-based | Transformer |




  - Models
    - PGC
      copy mechanism (Pointer-Generator) with coverage mechanism (alleviate repeated generation)
    - FASTRL
      Extract salient sentences > condense extracted sentences with Pointer-Generator
    - BOTTMUP
      selects words from the src > Pointer-Generator (constrained selected word)
    - TCON
      CNN conditioned on the topics of the article
    - BERTSUM
      fine-tune BERT Encoder with Extractive summary > Jointly train decoder (abstract summarizer) with Encoder

$$\mathbf{x} = \{x_1, x_2, x_3, x_4\}$$
$$\mathcal{C} = \{0,0,0,0\} \xrightarrow{(x_2 x_3, y_m y_{m+1})} \mathcal{C} = \{0,1,1,0\}$$



# The Abstractive-Faithfulness Tradeoff

## ❖ Is abstractiveness from the model or the data? > Dataset

- Results

| Dataset | Model | Extraction | | | Perfect fusion | | Novel $n$-grams | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Sentence | Span | Word | $k=2$ | $k\geq 2$ | $n=1$ | $n=2$ | $n=3$ |
| CNN/DM | Ref | 1.39 | 2.14 | 9.27 | 12.92 | 14.87 | 12.40 | 51.03 | 71.22 |
| | PGC | 35.45 | 34.18 | 15.45 | 10.90 | 1.61 | 0.62 | 3.33 | 7.42 |
| | FASTRL | 8.94 | 40.06 | 39.64 | 4.22 | 0.84 | 0.82 | 10.89 | 20.74 |
| | BOTTOMUP | 7.65 | 17.98 | 36.75 | 21.86 | 6.77 | 0.86 | 11.44 | 22.40 |
| | BERTSUM | – | 13.73 | 53.40 | 16.18 | 4.39 | 5.23 | 14.55 | 23.09 |
| XSum | Ref | – | – | – | 0.87 | 0.77 | 39.20 | 84.98 | 96.05 |
| | PGC | – | – | – | 0.41 | 3.47 | 30.08 | 74.27 | 91.27 |
| | TCONV | – | – | – | 0.35 | 2.31 | 34.07 | 80.62 | 95.12 |
| | BERTSUM | – | – | – | 0.33 | 3.15 | 28.93 | 75.85 | 91.41 |

1. CNN/DM is more extractive than Xsum (none of the summary sentences in XSum are formed by copying from a single source sentence)
2. training data has a larger influence on the abstractiveness of model outputs.

   Content is more rephrased in. Xsum (src 한문장 내의 token set을 그대로 가져다 summary를 만들지 못함)

   Both dataset, models fail to generate abstractive summary as reference (see novel n-grams) > need inducive bias (진정한 NLU 능력이 필요함... 데이터 더 넣기??)
3. different models have different ways of doing extraction

   PCG : Copy

   FASTRL, BOTTOMUP : Condense (see previous slide)

# The Abstractive-Faithfulness Tradeoff

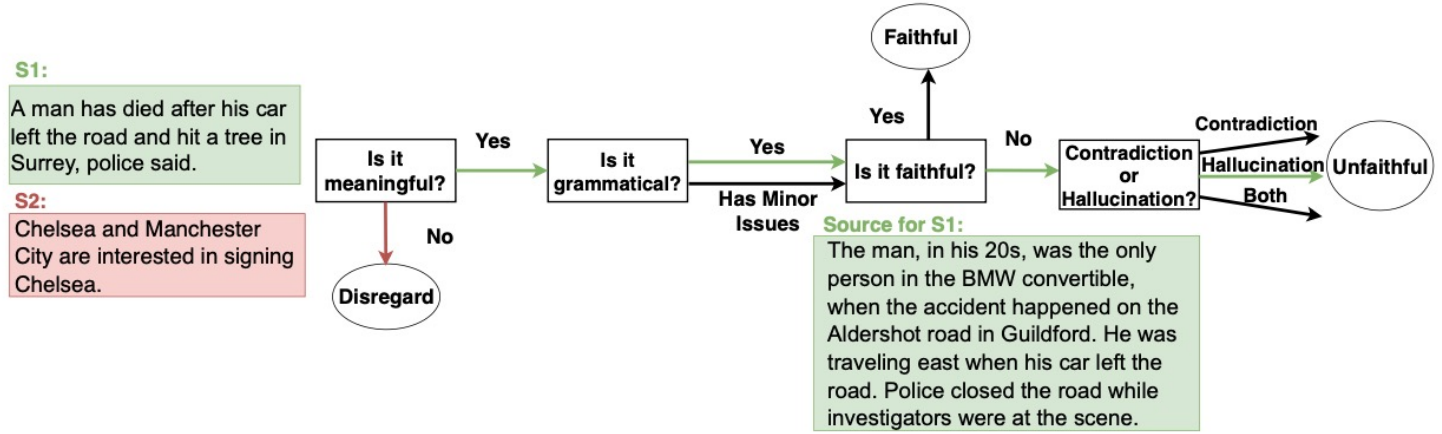
## ❖ Annotating Summary Faithfulness

- Perform Human annotations on the output summary of each models
- excluding output sentences that are either an exact copy or a substring of one of the source sentences

(sentence, span, word extraction in previous slides)

- Human annotation process (standardize to verify the inter-annotator agreement)

1. Meaningful (말이 됨?) > 2. Grammatical? (문법적?) > 3.Faithful? (src내용을 반영?) > 4. UnFaithful 이유?



# The Abstractive-Faithfulness Tradeoff

## ❖ Annotating Summary Faithfulness

- Results (1000 samples from each model output)

| Dataset | Model | Grammaticality | | | Faithfulness | | |
|---|---|---|---|---|---|---|---|
| | | Score | Agreement | Abstractiveness | Score | Agreement | Abstractiveness |
| CNN/DM | PGC | 93.34 | 94.04 | 10.05 | 70.05 | 77.28 | 13.35 |
| | FASTRL | 83.06 | 88.05 | 44.46 | 68.27 | 77.45 | 49.74 |
| | BOTTOMUP | 85.83 | 89.19 | 29.62 | 64.17 | 76.04 | 42.36 |
| | BERTSUM | 97.53 | 97.65 | 29.44 | 95.03 | 95.14 | 39.16 |
| XSum | PGC | 65.85 | 81.03 | 91.10 | 40.33 | 71.63 | 97.06 |
| | TCONV | 70.85 | 85.03 | 94.94 | 38.96 | 69.90 | 98.81 |
| | BERTSUM | 90.44 | 91.80 | 91.50 | 60.54 | 70.00 | 97.60 |

**Score** : % of annotators that selected "meaningful" and "faithful" for grammaticality and faithfulness annotation tasks

**Agreement** : % of annotate the majority class (inter-annotator agreement) | **Abstractiveness** : percentage of novel trigrams

1. Grammaticality

   all models are scored high on grammaticality with high inter-annotator agreement

   the grammaticality scores drop significantly in Xsum

2. Faithfulness

   CNN/DM have significantly higher faithfulness scores than XSum

   human agreement on faithfulness is also lower for abstractive summaries from Xsum (Abstractive-Faithfulness Tradeoff 확인)

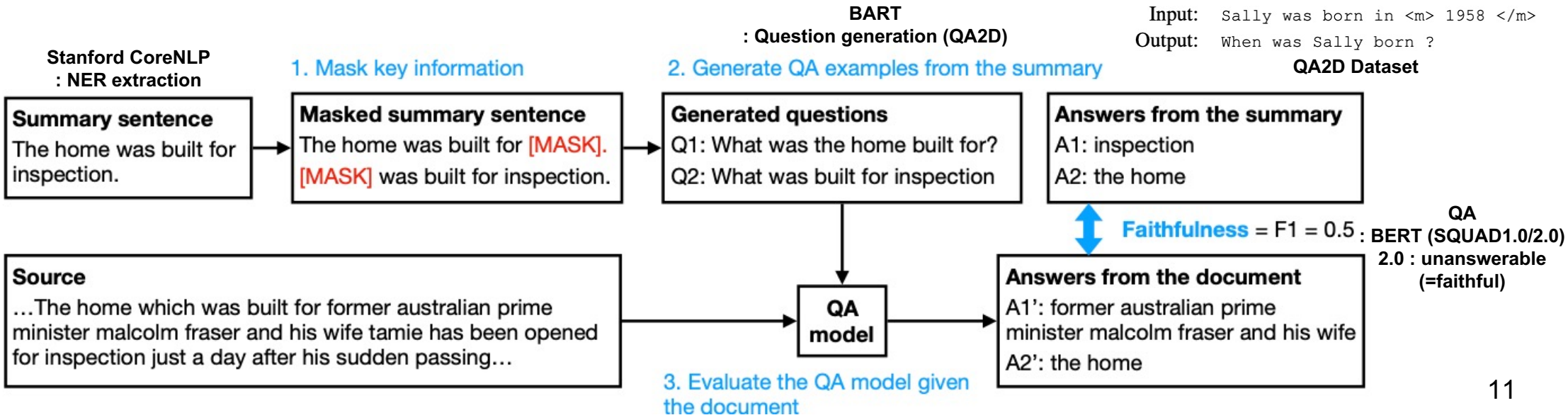
# FEQA : Faithfulness Evaluation with Question Answering

## ❖ FEQA

- Key challenge
  - to faithfulness evaluation is to verify highly abstractive sentences against the source document.

(abtractivness가 높은 summary일수록 Faithfulness가 낮은 경향을 보이므로)

- Proposed Metric Structure
  - Faithful한 summary로 Question을 만든다면, Source를 가지고 정답을 정확히 추론할 수 있음



# Experiment

## ❖ Rouge Score

- Reference Summary의 n-gram 중 Model Summary의 n-gram과 일치하는 비율 (Recall)

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

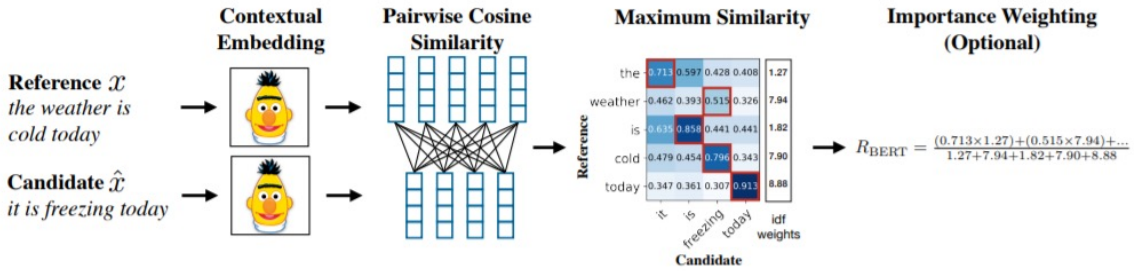
## ❖ BLEU Score

- Model Summary의 n-gram 중 Reference Summary의 n-gram과 일치하는 비율 (Precision)

$$p_n = \frac{\sum_{n\text{-}gram \in Candidate} Count_{clip}(n\text{-}gram)}{\sum_{n\text{-}gram \in Candidate} Count(n\text{-}gram)} \qquad BLEU = exp(\sum_{n=1}^{N} w_n \log p_n) \qquad BLEU = BP \times exp(\sum_{n=1}^{N} w_n \log p_n)$$

## ❖ BERT Score

- Bert Embedding과 Cosine Similarity를 활용해 reference summary와 model summary의 유사도를 계산하는 방법



# Experiment

## ❖ Baseline Metrics (each sentence in src <-> model generated summary)

- Word overlap-based metrics (avg)
  - Rouge (1,2,L)
  - BLEU (only report BLEU-4 as no other variation have significant correlation)
- Embedding-based metrics (max)
  - BERTSC (BERTScore)
- Model-based metrics
  - RE (Relation Extraction)

  (precision for the relation triplets extracted from the summary sentence and the source document)
  - ENT (Entailment)

  (summary sentence is entailed by the source)

| | | |
|---|---|---|
| $\mathbf{P}^a$ | A senior is waiting at the window of a restaurant that serves sandwiches. | Relationship |
| $\mathbf{H}^b$ | A person waits to be served his food. | Entailment |
| | A man is looking to order a grilled cheese sandwich. | Neutral |
| | A man is waiting in line for the bus. | Contradiction |
| $^a\mathbf{P}$, Premise. $^b\mathbf{H}$, Hypothesis. | | |

**Entailment**

# Experiment

## ❖ Result

- Metric Comparison (인간과 가장 유사하게 평가한 metric은?)

| Metric | CNN/DM P | CNN/DM S | XSum P | XSum S |
|---|---|---|---|---|
| Word overlap-based | | | | |
| R-1 | 12.02** | 15.86** | −2.57 | 0.07 |
| R-2 | 13.25** | 15.99** | −5.78 | −8.47 |
| R-L | 12.58** | 16.49** | −6.37 | −9.68 |
| B-4 | 12.09** | 11.68** | −6.76 | −10.02 |
| Embedding-based | | | | |
| BERTSc | 11.07* | 10.70* | 10.06 | 10.69 |
| Model-based | | | | |
| RE | 8.58* | 5.52 | 1.62 | 2.32 |
| ENT | 2.80 | 3.65 | −5.62 | −3.85 |
| FEQA | **32.01**** | **28.23**** | **26.31**** | **21.34**** |

1. FEQA가 extractive setting, abstractive setting 모두에서 인간과 가장 높은 상관관계를 기록
2. Word-overlap(Rouge)는 extractive setting일수록 점수가 높음
3. Entailment 기반의 metric은 성능이 나쁨

# Experiment

## ❖ Result

- Content selection and faithfulness (Rouge로 Faithfulness를 평가할 수 있는가?)
  - Correlation between Reference Summary <-> Model Summary (vs human annotation)

| | CNN/DM | | XSum | |
|---|---|---|---|---|
| Metric | P | S | P | S |
| ROUGE-1 | 15.31** | 14.92** | 5.44 | 5.79 |
| ROUGE-2 | 15.10** | 16.39** | 8.25 | 6.79 |
| ROUGE-L | 13.33** | 13.35** | 4.61 | 3.97 |

  - Example (high Rouge but unfaithful summary)

| **Reference** | **Output Sentence** |
|---|---|
| ... University of Nebraska researcher has revealed why stress is bad for you. Limited periods of stress are good, as they release cortisol... | University of Nebraska researcher has revealed why stress is bad for you, stimulating your body to produce an important hormone called cortisol. |
| ...Indian air force and Nepalese army medical team launch rescue mission to bring injured people to hospitals in Kathmandu. Forshani Tamang's family carried her for four hours to reach help after she was wounded when their home was destroyed... | Indian air crew and Nepalese army medical team were killed in Nepal's Sindhupalchok quake. |

# Experiment

## ❖ Result

### ▪ Analysis and limitations of QA-based evaluation

- Example (Summary가 Faithful할수록 OA(output answer)=SA(source answer)

| | Source | Output Sentence | Question | OA | SA |
|---|---|---|---|---|---|
| Unfaithful | ...However, Winger Ross Wallace (knee) and right-back Steven Reid (calf) could return for the Barclays premier league contest... | **Dean Marney** and Steven Reid could return for the Barclays Premier League match. | Who and Steven Reid could return for the premier league match? | Dean Mar-ney | Ross Wal-lace |
| Faithful | ...Miss Bruck, 22, from maybe has not been seen since the early hours of October 26, 2014. She has not been seen for six months... | Miss Bruck, 22, from maybe has not been seen for **six months.** | How long has Miss Bruck, 22 from not been seen for? | six months | six months |

- unanswerable questions을 제대로 판별하지 못하거나, source에 답이 있는데 unanswerable로 대답하는게 QA metric의 한계 (해당 metric은 QA system에 dependent함)
- QA system이 exact match 기반으로 faithfulness를 평가하는 것도 한계로 지적
  - “Donald Trump” vs. “the President of the United States Donald Trump” > 틀렸다고 평가함

# Related Work

## ❖ Problems in current neural generation models

- problems with repetition and generic responses have received lots of attention
- semantic errors in model outputs (in text generation area)
  - adequacy in MT
  - faithfulness in Summarization
  - consistency in Dialogue

## ❖ Automated evaluation for NLG.

- Automated NLG evaluation is challenging as it often requires deep understanding(NLU) of the text
- Widely used overlap-based (ROUGE) do not correlate well with human judgments

# Conclusion

- current models suffer from an inherent(=data) trade-off between abstractiveness and faithfulness
- current models tend to concatenate unrelated spans and hallucinate details when generating more abstractive sentences.
- A new inductive bias or additional supervision is needed for learning reliable models.

(Need NLU)

- While our QA-based metric correlates better with human judgment and is useful for model development, it is limited by the quality of the QA model

# Thank you

IMLL
Industrial Machine Learning Lab

# Reference


- See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.
- Tu, Z., Lu, Z., Liu, Y., Liu, X., & Li, H. (2016). Modeling coverage for neural machine translation. arXiv preprint arXiv:1601.04811.
- Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. arXiv preprint arXiv:1908.08345.