NAACL

# Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking

**Hwanhee Lee**[1], Kang Min Yoo[2], Joonsuk Park[2,3], Hwaran Lee[2], Kyomin Jung[1]

[1]Seoul National University, [2]NAVER AI Lab, [3]University of Richmond



# Factual Consistency Evaluation for Abstractive Summaries

- **Factual consistency**: whether the *generated content* is *factually consistent* with the source. (i.e. article)
- Although the generated text is fluent, if there is a minor factual error, the summary is totally wrong.

**Article**: Scientists from harvard medical school have discovered a way of turning stem cells into killing machines to fight brain cancer. In experiments on mice, the stem cells were genetically engineered to (…)

**Reference Summary:** Scientists in the us have developed a stem cell therapy for brain tumours.

**Summary 1:** Scientists in the us have developed a stem cell therapy for killing normal cells.

**Factuality:** ***inconsistent*** **BLEU-4:** 0.758 **ROUGE-L:** 0.820

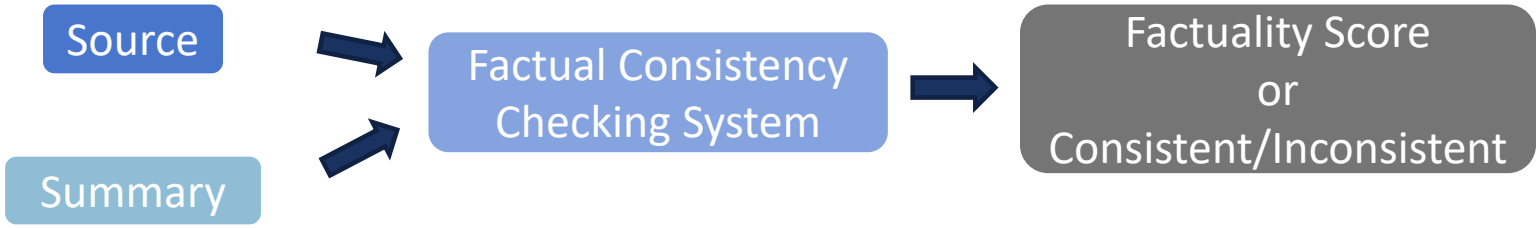
**Summary 2:** Scientists from harvard have discovered a new therapy for tumours in the brain.

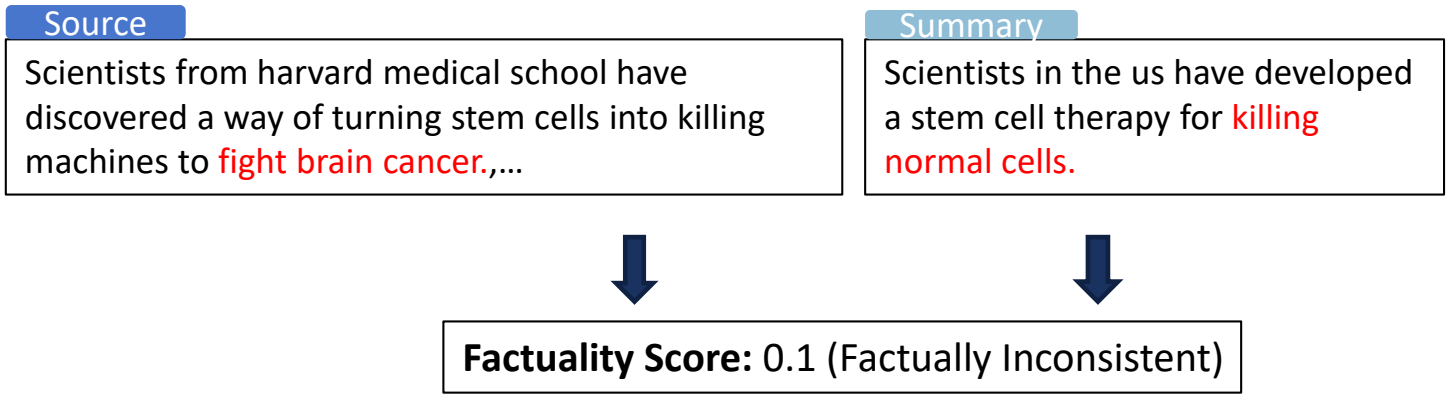
**Factuality:** ***consistent*** **BLEU-4:** 0.000 **ROUGE-L:** 0.462

*We cannot trust BLEU, ROUGE!*

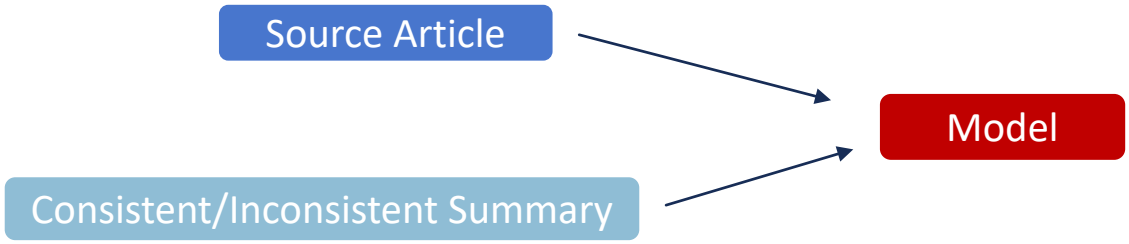
# Research Goal

- Developing a factual consistency checking system for abstractive summaries that focuses on **"factual consistency"** with the source



# Previous Works: Weakly Supervised – 1) FactCC

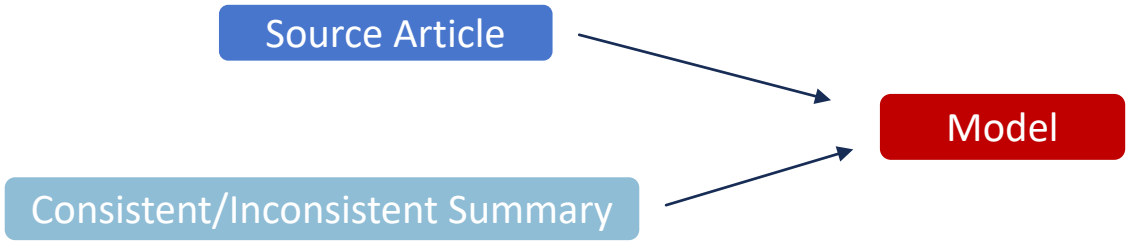



- Corrupting reference summary using rule-based substitution to generate inconsistent summaries

| Transformation | Original sentence | Transformed sentence |
|---|---|---|
| Paraphrasing | Sheriff Lee Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney's office in support of bribery charges against three city officials. | Two weeks after the US Attorney's Office issued photos to support bribery allegations against three municipal officials, Lee Baca has now decided to recall about 200 badges issued by his department to local politicians. |
| Sentence negation | Snow was predicted later in the weekend for Atlanta and areas even further south. | Snow wasn't predicted later in the weekend for Atlanta and areas even further south. |
| Pronoun swap | It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets. | It comes after your estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets. |
| Entity swap | Charlton coach Guy Luzon had said on Monday: 'Alou Diarra is training with us.' | Charlton coach Bordeaux had said on Monday: 'Alou Diarra is training with us.' |
| Number swap | He says he wants to pay off the $12.6million lien so he can sell the house and be done with it. | He says he wants to pay off the $3.45million lien so he can sell the house and be done done with it. |
| Noise injection | Snow was predicted later in the weekend for Atlanta and areas even further south. | Snow was was predicted later in the weekend for Atlanta and areas ~~even~~ further south. |

Kryściński et al., Evaluating the Factual Consistency of Abstractive Text Summarization, EMNLP 2020

# Previous Works: Weakly Supervised – 2) DocNLI




- Corrupting reference summary using Mask-and-Fill with pre-trained language models to generate inconsistent summaries.

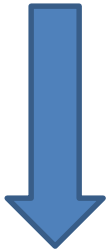
| | | |
|---|---|---|
| | real summ. | The Serious Fraud Office has reportedly dropped a criminal investigation into three businessmen who had been accused of conspiring to make corrupt payments to secure contracts in Iraq. The SFO launched an investigation into Petrofac in May 2017 as part of a wider probe into Monaco-based oil consultancy Unaoil. |
| fake summaries | word repl. | The Serious financial Office has reportedly launched a criminal investigation into three businessmen who had been accused of conspiring to make corrupt payments to oil contracts in Iraq. The SFO launched an investigation into corruption in May 2017 as part of a wider investigation into Monaco-based financial consultancy firms. |
| | entity repl. | Unaoil has reportedly dropped a criminal investigation into three businessmen who had been accused of conspiring to make corrupt payments to secure contracts in Monaco . The SFO launched an investigation into Monaco in May 2017 as part of a wider probe into Petrofac-based oil consultancy The Serious Fraud Office. |
| | sent repl. | The Serious Fraud Office has reportedly dropped a criminal investigation into three businessmen who had been accused of conspiring to make corrupt payments to secure contracts in Iraq. A spokesman for the SFO said it was "unable to confirm or deny" that an inquiry had taken place. |

DocNLI: A Large-scale Dataset for Document-level Natural Language Inference, Findings of ACL 2021

# Limitation of Simple Substitution

**Article:** Guus Hiddink, the Russia and Chelsea coach, has had much to smile about in his 22-year managerial career. „. . . , Enjoying success around the world – at different levels with different players in different cultures – has made Guus Hiddink one of the most admired bosses around. „. . . , Hiddink's resume includes stints in other high-pressure jobs such as Fenerbahce, Valencia and Real Madrid. „. . . , But the straight-speaking Dutchman is loyal to the project he has in charge of the Russian national side and insists he will leave Chelsea at the end of the season regardless.

**Reference Summary:** Born in 1946, Hiddink has become one of the best managers in the world . Dutchman has enjoyed huge success at club and international level. He's currently coach of Russia and is in charge of Chelsea until end of the season.



Fill in the mask

**Generated Inconsistent Summary**

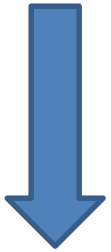
Born in 1946, *Dutchman* has become one of *the most respected politicians* in the world. Dutchman is enjoyed *success at the Olympics and World Cup*. He's currently the *President of Russia* and is in charge of *the country* until the end of the season.

- To easy to be discerned
- Irrelevant to article

# Using Masked Article

**Article:** Guus Hiddink, the Russia and Chelsea coach, has had much to smile about in his 22-year managerial career. „. . . , Enjoying success around the world – at different levels with different players in different cultures – has made Guus Hiddink one of the most admired bosses around. „. . . , Hiddink's resume includes stints in other high-pressure jobs such as Fenerbahce, Valencia and Real Madrid. „. . . , But the straight-speaking Dutchman is loyal to the project he has in charge of the Russian national side and insists he will leave Chelsea at the end of the season regardless.

**Reference Summary:** Born in 1946, Hiddink has become one of the best managers in the world . Dutchman has enjoyed huge success at club and international level. He's currently coach of Russia and is in charge of Chelsea until end of the season.



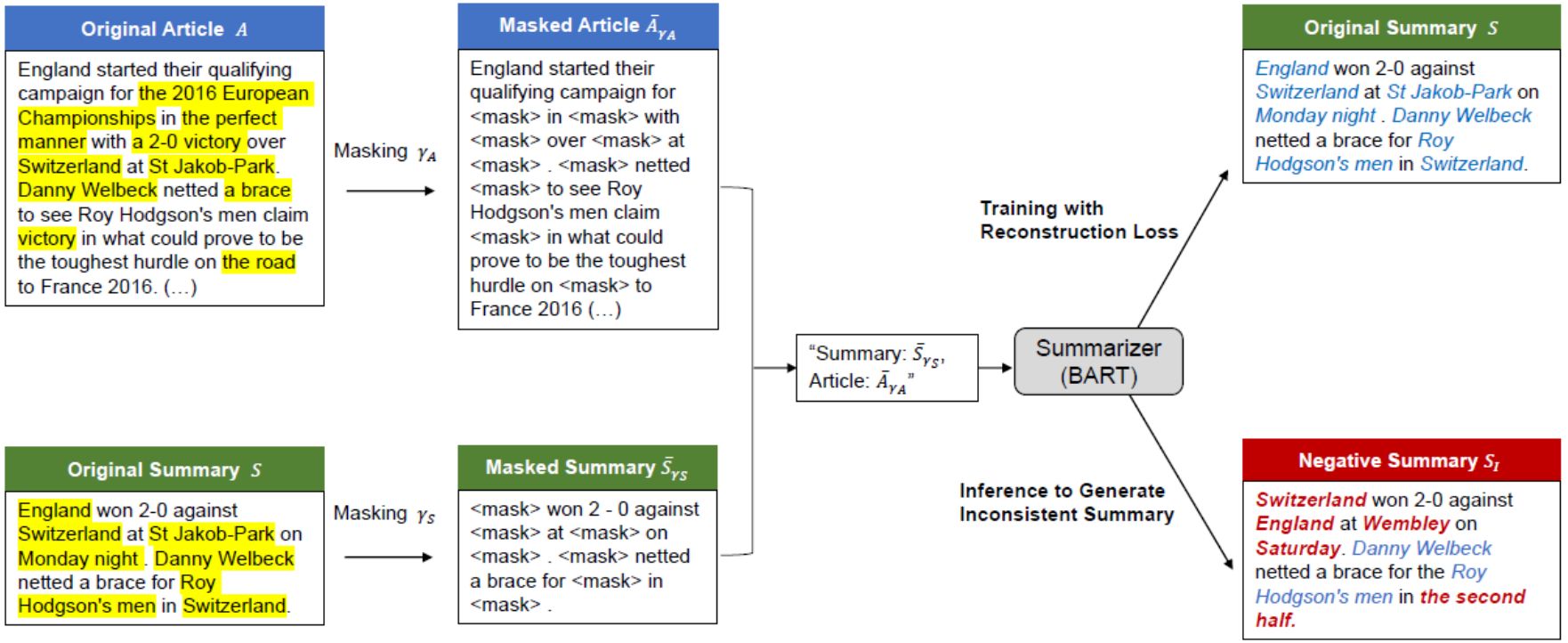
Fill in the mask

**Generated Inconsistent Summary**

Born in 1946, *Hiddink* has become one of *the most admired managers* in the world. Dutchman has enjoyed *successful spells* at *Chelsea and Real Madrid*. He's currently *manager of Russia* and is in charge of *the country* until the end of the season.

- Relevant to article
- More natural, but still inconsistent

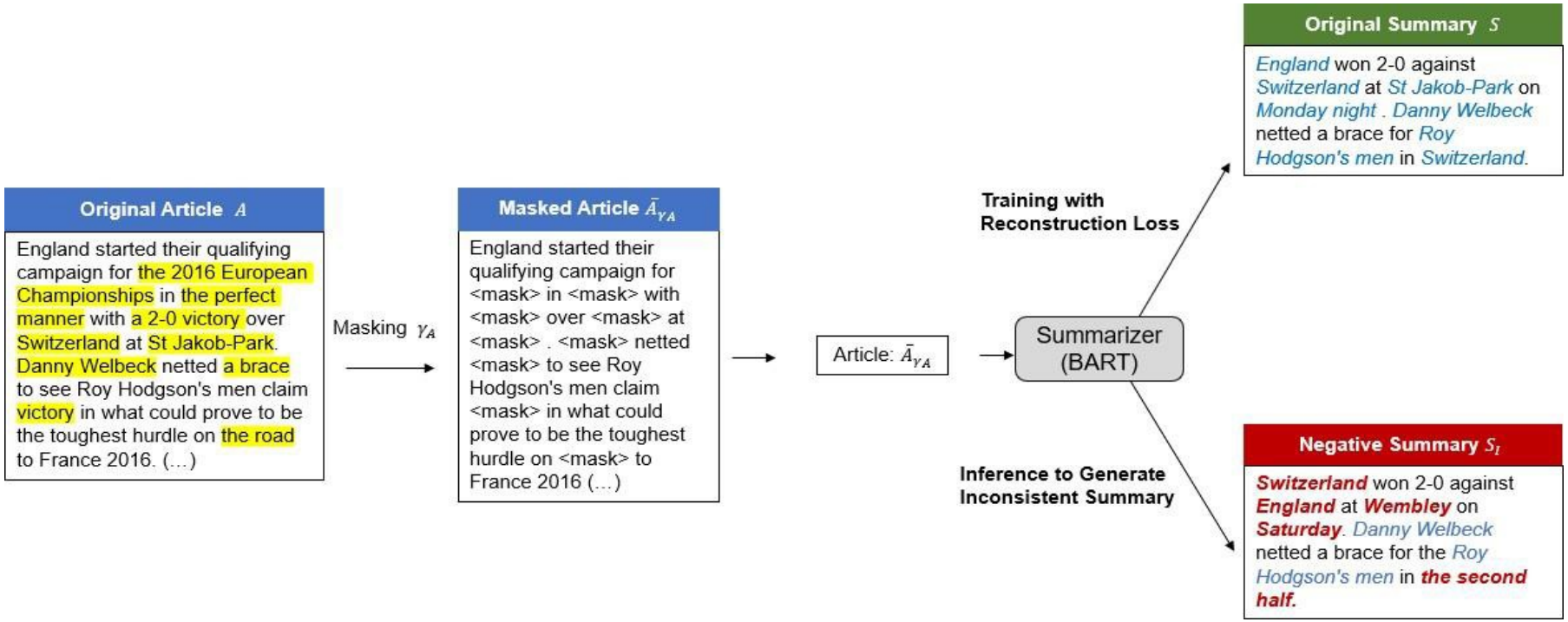
# Proposed Method 1: Mask-and-Fill with Masked Article (MFMA)




**Goal**: Generating inconsistent summary.

1) Masking keyphrases in both the summary and the article.
2) Filling the masks in the summary using the masked summary and the masked article.

# Proposed Method 2: Masked Summarization (MSM)




**Goal**: Generating inconsistent summary.

1) Masking keyphrases in the summary.
2) Generating a summary using the masked summary.

# Experiments: Classification Accuracy

- Macro-F1 score and class-balanced accuracy of the human annotated factual consistency for the benchmark datasets based on CNN/DM and XSUM
  - CNN/DM

| Dataset | FactCC-Test | | SummEval | | QAGS-CNN/DM | | FRANK-CNN/DM | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Metric** | **F1** | **BA** | **F1** | **BA** | **F1** | **BA** | **F1** | **BA** | **F1** | **BA** |
| ***Baselines*** | | | | | | | | | | |
| **FactCC** | 71.0 | 71.3 | 65.1 | 68.2 | 69.3 | 69.6 | 64.1 | 63.9 | 67.4 | 68.2 |
| **DocNLI** | 67.2 | 71.0 | **71.5** | **71.3** | 62.4 | 66.2 | 66.0 | 66.0 | 66.8 | 68.6 |
| **MNLI** | 55.0 | 56.0 | 51.7 | 51.7 | 48.6 | 53.4 | 50.4 | 53.3 | 51.4 | 53.6 |
| **FEVER** | 57.9 | 56.2 | 52.6 | 53.6 | 39.4 | 53.3 | 49.8 | 55.6 | 49.9 | 54.7 |
| **MF** | 59.9 | 64.1 | 68.2 | 67.5 | 47.6 | 56.9 | 62.4 | 62.7 | 59.5 | 62.8 |
| ***Ours*** | | | | | | | | | | |
| **MFMA** | **79.7** | **84.5** | 71.3 | 69.6 | **70.5** | **72.3** | **69.5** | 69.2 | **72.8** | **73.9** |
| **MSM** | 70.6 | 72.7 | 66.8 | 68.2 | 67.6 | 68.7 | 69.6 | **69.3** | 68.6 | 69.7 |

  - XSum

| Dataset | XSumHall | | QAGS-XSum | | FRANK-XSum | | Average | |
|---|---|---|---|---|---|---|---|---|
| **Metric** | **F1** | **BA** | **F1** | **BA** | **F1** | **BA** | **F1** | **BA** |
| ***Baselines*** | | | | | | | | |
| **FactCC** | 52.1 | **61.8** | 63.6 | 63.7 | 50.7 | 58.0 | 55.5 | 61.2 |
| **DocNLI** | 55.1 | **56.4** | 65.3 | 66.0 | **60.3** | **63.4** | 60.2 | **61.9** |
| **MNLI** | 33.3 | 52.1 | 45.2 | 51.1 | 28.8 | 50.6 | 35.8 | 51.3 |
| **FEVER** | 53.1 | 55.5 | 62.2 | 63.7 | 54.9 | 63.5 | 56.7 | 60.9 |
| **MF** | 53.6 | 53.3 | 54.6 | 54.9 | 55.7 | 55.3 | 54.6 | 54.5 |
| ***Ours*** | | | | | | | | |
| **MFMA** | **55.5** | 56.0 | **66.6** | **67.0** | 59.6 | 59.6 | **60.6** | 60.9 |
| **MSM** | 52.6 | 53.9 | 50.8 | 55.5 | 50.8 | 51.3 | 51.4 | 53.6 |

- Our proposed method outperforms the previous methods on 5 of 7 benchmarks.

# Experiments: Correlation with Human Judgments

- Summary level Pearson Correlation(r) and Spearman's Correlation(ρ) between various automatic metrics and human judgments of factual consistency for the model generated summaries.

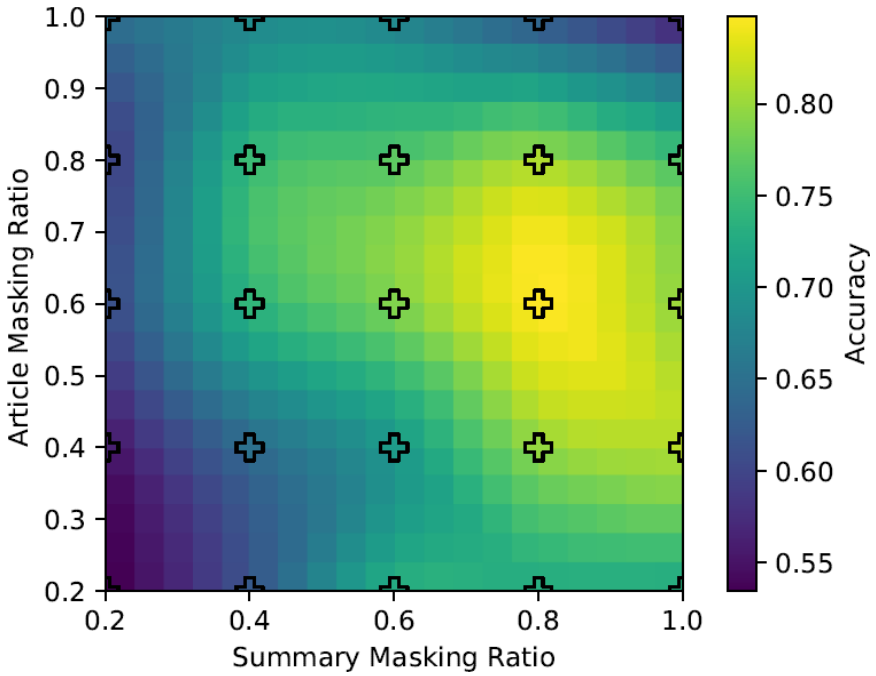
- We use the confidence of consistency label for entailment based metrics.

| Dataset | SummEval | | QAGS-CNN/DM | | QAGS-XSum | | FRANK-CNN/DM | | FRANK-XSum | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | $r$ | $\rho$ | $r$ | $\rho$ | $r$ | $\rho$ | $r$ | $\rho$ | $r$ | $\rho$ |
| *Baselines* | | | | | | | | | | |
| ROUGE-L | 0.16 | 0.14 | 0.29 | 0.24 | 0.13 | 0.13 | 0.16 | 0.13 | 0.16 | 0.13 |
| BLEU-4 | 0.11 | 0.12 | 0.18 | 0.23 | 0.03 | 0.03 | 0.16 | 0.17 | 0.11 | 0.14 |
| METEOR | 0.18 | 0.16 | 0.26 | 0.25 | 0.11 | 0.12 | 0.29 | 0.28 | 0.18 | 0.16 |
| BERTScore | 0.16 | 0.14 | 0.37 | 0.36 | 0.11 | 0.13 | 0.33 | 0.30 | 0.19 | 0.17 |
| QuestEval | 0.35 | 0.30 | 0.42 | 0.36 | 0.20 | 0.20 | 0.46 | 0.41 | 0.19 | 0.18 |
| CoCo | 0.42 | 0.36 | **0.67** | 0.57 | 0.20 | 0.18 | 0.50 | 0.45 | 0.14 | 0.12 |
| FactCC | 0.38 | 0.36 | 0.45 | 0.48 | 0.30 | 0.30 | 0.32 | 0.36 | 0.09 | 0.08 |
| DocNLI | 0.51 | **0.41** | 0.60 | 0.59 | 0.36 | 0.35 | 0.49 | **0.49** | **0.25** | **0.21** |
| MNLI | 0.11 | 0.13 | 0.19 | 0.22 | 0.08 | 0.10 | 0.15 | 0.16 | 0.02 | 0.03 |
| FEVER | 0.33 | 0.32 | 0.40 | 0.34 | **0.38** | **0.41** | 0.38 | 0.43 | 0.20 | 0.19 |
| MF | 0.44 | 0.35 | 0.43 | 0.30 | 0.10 | 0.10 | 0.40 | 0.39 | 0.10 | 0.13 |
| *Ours* | | | | | | | | | | |
| MFMA | **0.52** | 0.38 | 0.62 | **0.65** | 0.37 | 0.38 | **0.52** | 0.45 | 0.16 | 0.17 |
| MSM | 0.43 | 0.36 | 0.50 | 0.48 | 0.20 | 0.22 | 0.51 | 0.48 | 0.05 | 0.09 |

- We show that our method has higher or competitive correlation with human judgments than previous methods.

# Experiments: Masked Ratio

➢ Performance among Masked Ratio for Mask-and-Fill with Masked Article.




- We experiment with each of the five combinations of article mask ratio and summary mask ratio, and then plot the interpolated results.

- We can infer that there is an optimal masking ratio combination.

# Closing Remarks

- We proposed an effective generation method of factually inconsistent summaries, called MFMA (Mask-and-Fill with Masked Article).

- We masked some keyphrases of both the source text and corresponding reference summaries, then let a summarization model generate plausible but factually inconsistent summaries by inferring the masked contents.

- Experiments on seven benchmark datasets demonstrated that factual consistency classifiers trained using these inconsistent summaries generally outperformed existing models in various benchmark datasets.

**Code**: https://github.com/hwanheelee1993/MFMA