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

Task

- We develop a factual consistency checking model for abstractive summaries.

Source Article

Scientists from harvard medical school have discovered a way of turning stem cells into killing machines to fight brain cancer.,…

Generated Summary

Scientists in the us have developed a stem cell therapy for killing normal cells.

Factuality Score: 0.1 ➡ *Factually Inconsistent*

- We generate synthetic inconsistent summaries and train a classifier to distinguish them with the reference summaries.

Motivation

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 has made Guus Hiddink one of the most admired bosses around. ,. . . , 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. He's currently coach of Russia and is in charge of Chelsea until end of the season.

Mask-and-fill Summary with BART:

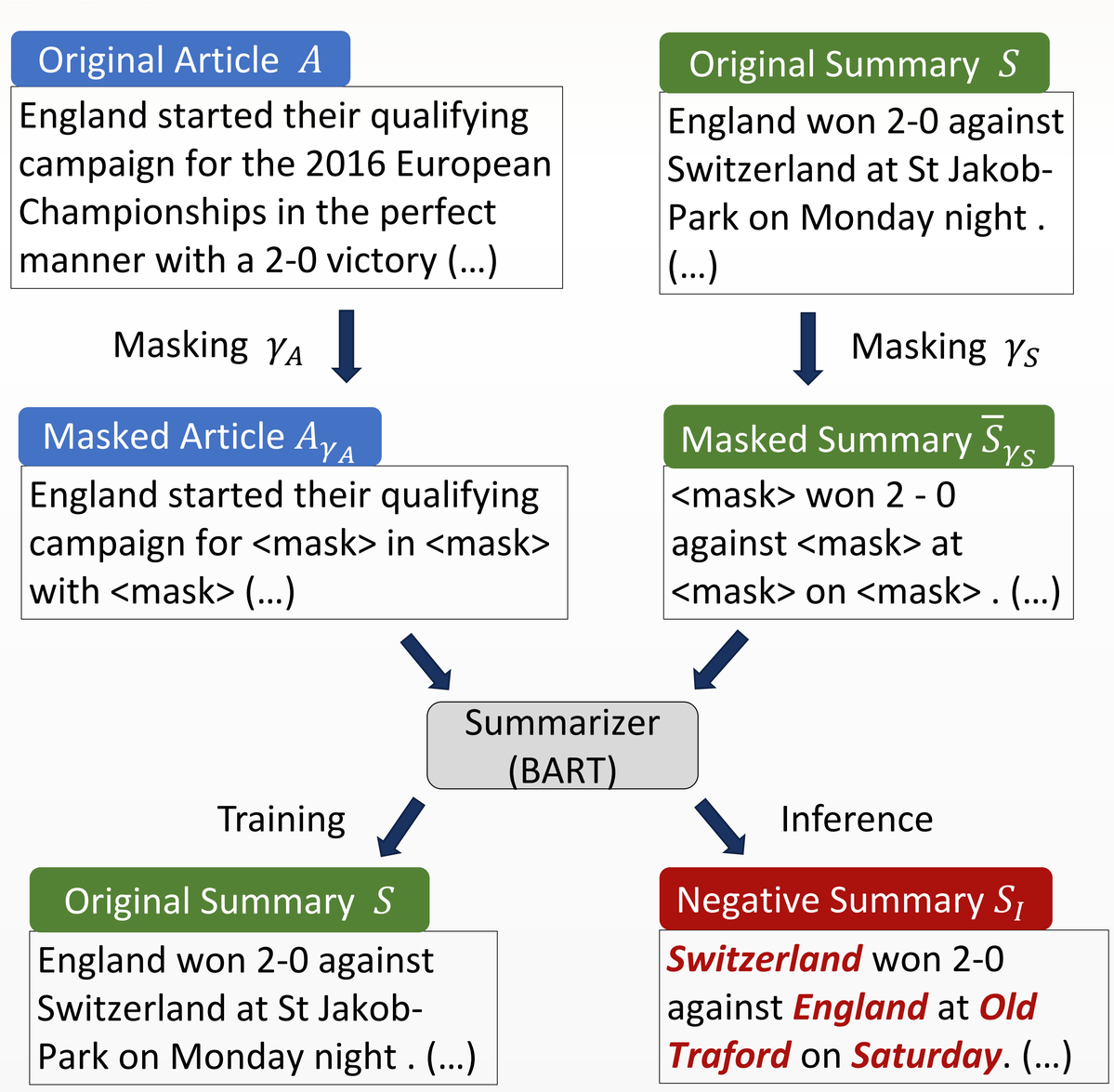
Born in 1946, *Dutchman* has become one of *the most respected politicians* in the world. He's currently the *President of Russia* and is in charge of *the country* until the end of the season.

- Simply replacing keywords in the reference summaries using the pre-trained language models results in negative summaries that significantly diverge from the source texts.

Our Contributions

- We propose a novel negative summary generation method for training factual consistency classifiers for abstractive summaries.
- We show the efficacy of our method on seven benchmark datasets using classification performance and correlation with human judgment.
- We analyze the characteristics, such as affinity and diversity, of the negative summaries generated using our method.

Mask-and-Fill with Masked Article

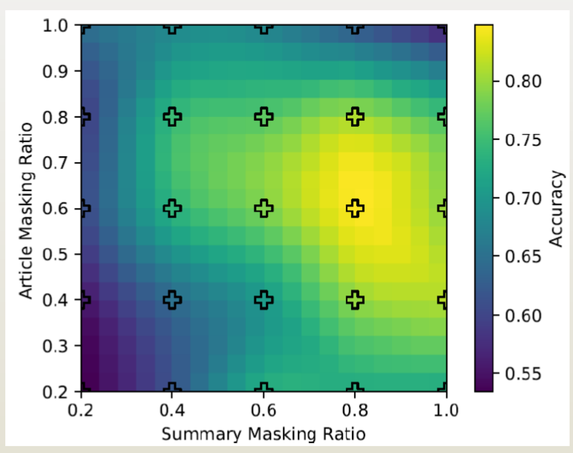


Experimental Results

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

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 (macro-F1).



- We find that too high masking ratio decreases performance by sacrificing affinity. On the other hand, too lower masking ratio leads to generate consistent summaries and degrades the performance.