Reference and Document Aware Semantic Evaluation Methods for Korean Language Summarization

Dongyub Lee, Myeongcheol Shin, Taesun Whang, Seungwoo Cho,

Byeongil Ko, Daniel Lee, Eunggyun Kim, Jaechoon Jo

Overview

- The task of text summarization is to generate a reference summary that conveys all the salient information of an original document.
- Most summarization models are evaluated using recall-oriented understudy for gisting evaluation (ROUGE), which measures n-gram overlaps between generated and reference summaries.
- However, as the ROUGE score is computed based on an n-gram overlap, the score can be low even if two words have the same semantic meaning.
- This tendency is particularly prevalent in Korean, which is an agglutinative language that combines various morphemes into a word to express several meanings and grammatical functions, unlike English.
- To overcome this, we propose evaluation metrics that reflect semantic meanings of a reference summary and the original document, Reference and Document Aware Semantic Score (RDASS).

Methodology

- (1) We leverage a pre-trained SBERT to construct summary and document representations.
- (2) Cosine similarity between generated summary (v_p) and document (v_d) representations is calculated to obtain the semantic similarity score,

$$\text{s(p,d)} = \cos(v_p, v_d) = \frac{v_p^T \cdot v_d}{\|v_p\| \|v_d\|}.$$

- (3) Likewise, we calculate generated summary (v_p) and reference summary (v_r) as,

$$\text{s(p,r)} = \cos(v_p, v_r) = \frac{v_p^T \cdot v_r}{\|v_p\| \|v_r\|}.$$

- (4) Given a reference and source document, the reference-document-aware semantic score (RDASS) of the generated summary is defined by averaging s(p,d) and s(p,r).
- We also experimented with a sum, max and min operation between s(p,d) and s(p,r) but averaging the two scores reports highest correlation with human judgment.
- We also propose a fine-tuning method for SBERT that uses the abstractive summarization model. We refer to the fine-tuned SBERT with abstractive summarization model as "FWA-SBERT."

Correlation with Human Judgment

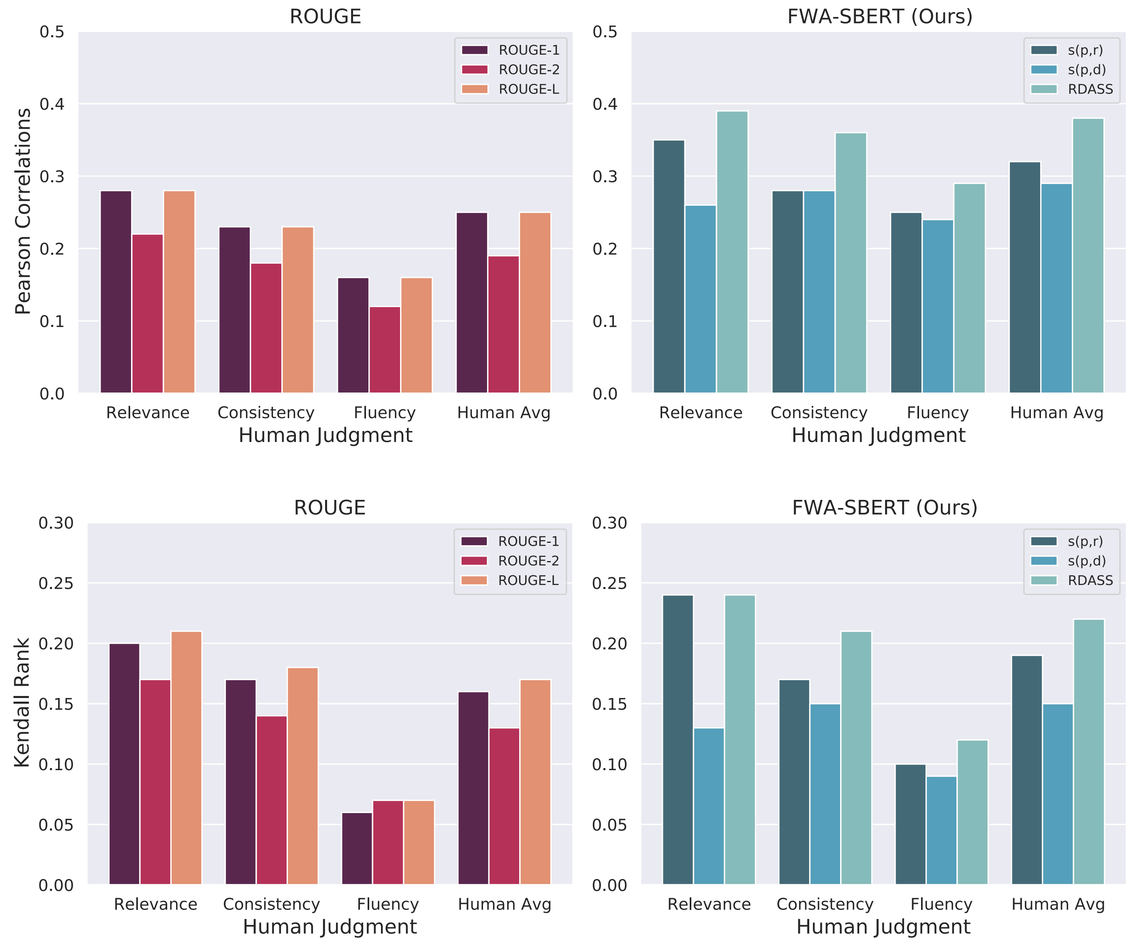

Figure: Pearson correlations and Kendall rank of the proposed evaluation metrics with human judgment.

- Relevance represents the degree of appropriateness of the document, consistency represents the degree of factualness, and fluency represents the degree of the quality of generated summary. Additionally, human avg represents the average value of the scores for the three indicators.
- Given a document, reference summary, and generated summary, each annotator scored in the range of 1 to 5 points for the evaluation indicator (i.e., relevance, consistency, fluency).

Correlation Comparisons

	ROUGE-1	ROUGE-2	ROUGE-L	$s(p,r)$	$s(p,d)$	*RDASS*
ROUGE-1	1.00	0.84	0.99	0.64	0.16	0.54
ROUGE-2		1.00	0.85	0.52	0.09	0.45
ROUGE-L			1.00	0.63	0.17	0.53
$s(p,r)$				1.00	0.32	0.77
$s(p,d)$					1.00	0.69
RDASS						1.00

Qualitative Analysis

Article
리오넬 메시(30 · fc바르셀로나)가 자신의 서른 번째 생일을 가족과 함께 오붓하게 보냈다. 지난 24일 만 서른 살이 된 메시는 자신의 인스타그램에 집에서 가족들과 함께 보낸 생일상을 찍은 사진을 올렸다. 메시는 오랜 그의 여자친구이자, 이제 아내가 되는 안토넬라 로쿠조(29), 아들 티아고가 함께 다정하게 사진을 찍었다. Lionel Messi (30 fc Barcelona) spent his thirtieth birthday with his family. Messi, who turned thirty on the 24th, posted a picture of his birthday on Instagram with his family at home. Messi was tenderly photographed by his longtime girlfriend, Antonella Rokujo (29), and his son, Thiago.
Reference Summary
메시가 30번째 생일 함께한 이는 아내와 아들 Messi's 30th birthday with his wife and son.
Generated Summary
메시 30번째 생일, 가족과 함께 오붓하게 보내 On the 30th birthday of Messi, he had a good time with his family. **Rouge(1 / 2 / L)**: 0.14 / 0.00 / 0.14 ***RDASS***: 0.81 **Human Evaluation (relevance / consistency / fluency)**: 4.4 / 4.2 / 4.2