The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.12, pp.691~699
eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.12.691

A Study on Automatic Metrics for Korean Text Abstractive Summarization

Sehwi Yoon† · Youhyun Shin††

ABSTRACT

This study aims to analyze and validate automatic evaluation metrics for Korean abstractive summarization. The unique linguistic characteristics of each language require evaluation metrics designed for them, underscoring the importance of research focused on Korean. Research on summarization and its meta-evaluation is extremely limited, especially for Korean. Therefore, by validating reliable automatic evaluation metrics using Korean summarization data, this study contributes to future research on Korean models in the fields of natural language generation. Human evaluation, widely regarded as the most reliable metric, is time-consuming and costly. Thus, research into automatic evaluation metrics holds significant importance for efficiency. In this study, summaries from three models—T5, KoBART,and GPT-3.5 Turbo—were evaluated based on their fluency, consistency, and relevance using 10 Korean documents and their corresponding reference summaries. Correlation coefficients were calculated between human evaluations and automatic metrics for fluency, consistency, and relevance. The results showed that for T5 summaries, the correlation coefficients for consistency and relevance were 0.33 and 0.26, respectively, while for KoBART summaries, the coefficients for fluency and relevance were 0.33 and 0.40, respectively. BERTScore demonstrated the highest correlation, indicating its effectiveness for Korean summaries. Meanwhile, GPT-3.5 Turbo summaries showed significant correlations of 0.23 and 0.17 in consistency and relevance using HaRiM+, a metric developed to detect hallucinations in recent work. Additionally, the correlation analysis by document type revealed that T5 summaries showed high correlations with the BLEU metric for briefing and meeting minutes, KoBART summaries and GPT-3.5 Turbo summaries both demonstrated high correlations with BERTScore for narrative and editorial documents, respectively. These findings emphasize the importance of selecting evaluation metrics tailored to specific document types. Therefore, this study provides a basis for selecting appropriate evaluation metrics tailored to the objectives of specific tasks in future Korean summarization research.

Keywords : Natural Language Processing, Abstractive Summarization, Automatic Metric, Transformer, Large Language Model

한국어 생성 요약 성능 평가 지표 분석 연구

윤 세 휘† · 신 유 현††

요 약

본 연구는 한국어 생성 요약의 자동 평가 지표를 분석하고 검증하는 것을 목표로 한다. 언어마다 고유한 특성이 다르므로 각 언어에 적합한 평가 지표의 필요성에 따라 한국어에 특화된 연구가 요구된다. 현재 한국어를 대상으로 한 생성 요약 및 메타 평가 연구는 다른 언어에 비해 훨씬 부족한 상황이다. 따라서 한국어 생성 요약 데이터를 활용하여 평가 기준 및 문서 유형에 따른 신뢰성 있는 자동 평가 지표를 검증함으로써 향후 생성 요약 및 자연어 생성 분야의 한국어 모델 연구에 이바지하고자 한다. 요약 모델 평가 시 공신력 있는 지표로 여겨지는 인간 평가(Human Evaluation)는 시간과 비용이 많이 소요되므로 자동 평가 지표 연구는 효율성 측면에서도 중요한 의의가 있다. 10가지 한국어 문서와 참조 요약문, 세 가지 모델(T5, KoBART, GPT-3.5 Turbo) 생성 요약문을 대상으로 유창성, 일관성, 관련성 기준으로 인간 평가와 자동 평가 지표의 상관계수를 산출하였다. 평가 기준별 상관 분석 결과, T5 요약문에서는 일관성, 관련성에서 각각 0.33, 0.26, KoBART 요약문에서는 유창성, 관련성에서 0.33, 0.40의 상관계수와 함께 BERTScore가 제일 높은 상관관계를 보여 한국어 요약문 평가에 효과적인 지표임을 확인하였다. 한편, 대규모 언어모델인 GPT-3.5 Turbo 요약문은 환각 가능성 감지를 위해 개발된 평가 지표 HaRiM+가 일관성, 관련성 측면에서 0.23, 0.17의 유의미한 상관관계를 보였다. 또한, 문서 유형별 상관 분석 결과, T5 요약문은 보도자료와 회의록에서 BLEU 지표와 높은 상관관계를 나타냈고, KoBART 요약문은 나레이션 문서에서, GPT-3.5 Turbo 요약문은 사설 문서에서 BERTScore와 높은 상관관계를 보였다. 이러한 결과는 특정 문서 유형에 적합한 평가 지표를 선택하는 것이 중요하다는 점을 강조한다. 이와 같이, 본 연구는 향후 한국어 요약 연구에서 목표 과제의 목적에 따라 적합한 평가 지표를 선택하는 근거로서 활용할 수 있다.

키워드 : 자연어 처리, 생성 요약, 자동 평가 지표, 트랜스포머, 대규모 언어모델

※ 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2024-00352711).

† 비 회 원 : 인천대학교 컴퓨터공학과 석사과정
†† 정 회 원 : 인천대학교 컴퓨터공학부 부교수

Manuscript Received : October 8, 2024
First Revision : November 26, 2024
Accepted : November 28, 2024

* Corresponding Author : Youhyun Shin(yhshin@inu.ac.kr)

1. 서 론

인공지능 기술을 활용한 텍스트 요약 모델은 크게 두 가지 방법, 추출 요약 기법과 생성 요약 기법으로 구분된다[1]. 추출 요약은 본문에서 중요한 문장이나 구절을 추출하여 요약문을 구성하는 기법으로, 예로 TextRank[2]를 들 수 있다. 생성 요약은 본문의 내용을 재구성하여 새로운 표현을 사용해 요약문을 작성하는 방식이다. 생성 요약은 본문을 재구성하기 때문에 더 복잡하고 시간이 많이 소요될 수 있으나, 복잡한 문서의 경우 요약문이 더 자연스럽다는 장점이 있다. 요약 모델 연구는 뉴스 기사부터 고객 상담[3]까지 다양한 도메인에 응용 가능한 분야지만, 특히 의료[4], 학문[5], 법률[6], 금융[7]과 같이 전문성이 큰 도메인일수록 그 필요성과 효과가 증가한다. 따라서, 본 연구에서는 한국어 문서에 대한 포괄적인 분석을 목표로 하는 만큼 추출 요약보다 생성 요약이 다양한 문서 유형에 대해 더 유연하고 자연스러운 요약문을 생성할 수 있다는 점을 고려하여, 생성 요약 모델만을 사용하여 성능 평가 지표를 분석한다.

기존 텍스트 요약 연구에서는 요약문 평가에 ROUGE[8]와 BERTScore[9]와 같은 자동 평가 지표가 사용된다. ROUGE는 n-gram 일치율을 기반으로 요약 성능을 평가하며, BERTScore는 사전학습된 언어모델을 기반으로 문맥적 유사성을 측정하는 평가 지표이다. 해당 지표들은 정량적 평가로 빠르게 점수 산출이 가능하지만, 요약문이 실질적으로 얼마나 유창하고 자연스러운지를 판단하기 위해 요약 연구에서는 인간 평가(Human Evaluation)와 같이 정성적 평가를 추가로 진행하여 결과의 신뢰도를 높이는 것이 보편적이다. 인간 평가 방법은 사람이 직접 요약문을 평가하는 방법으로, 유창성(Fluency), 일관성(Consistency), 관련성(Relevance) 등의 측면에서 요약문의 품질을 평가하여 점수를 매긴다[10]. 이는 자동 평가 지표가 놓칠 수 있는 세부적인 문맥적 요소를 고려하기 때문에 신뢰도 높은 평가 방법으로 인정받고 있다. 다만 인간 평가 방법은 매번 기준과 평가자가 다르므로 연구 간 일관된 결과를 얻기 어렵고 시간과 비용이 많이 소요된다는 한계가 있다. 따라서 요약 모델에 사용되는 자동 평가 지표들이 인간 평가와 얼마나 유사하게 동작하는지 상관계수를 산출하여 신뢰도를 검증하는 메타 평가 연구가 점차 증가하는 추세이다[11-13]. 그러나 현재 한국어를 대상으로 한 요약 모델에 대한 자동 평가 지표 검증 연구는 매우 부족한 상황이다. 대부분의 연구가 영어 데이터셋을 중심으로 이루어져 있어 한국어에 적합한 평가 지표가 충분히 검증되지 않은 상태이다. 이러한 연구 배경을 토대로, 본 연구는 한국어 데이터셋을 사용하여 인간 평가를 직접 수행하였다. 이를 바탕으로 자동 평가 지표 점수와 인간 평가 점수 간의 상관계수를 분석하여, 한국어 요약 모델에 적합한 평가 지표를 검증하였다.

본 연구에서 한국어 데이터셋으로 포괄적인 성능 평가 지표 검증을 위해 다양한 문서 유형이 포함된 AI 허브의 '요약문 및 레포트 생성 데이터'[14]를 사용한다. 이 데이터셋은 다양한 문서 유형(보도자료, 사설, 회의록 등)을 포함하고 있어, 서로 다른 주제와 형식의 한국어 문서에 대한 요약문을 폭넓게 평가할 수 있다. 또한, 참조 요약문과 비교 분석을 위해 KoBART 모델[15]과 T5 모델[16], GPT-3.5 Turbo 모델[17]로 생성한 요약문을 평가 대상에 추가하였다.

인간 평가는 문서 유형별로 선정된 본문에 대한 4가지 요약문(참조 요약문, KoBART, T5, GPT-3.5 Turbo)을 평가하는 방식으로 진행되었다. 컴퓨터공학 전공 대학원생 3인이 평가자로서 참여하여 본문과 요약문을 읽고 각 요약문을 유창성, 일관성, 관련성 기준에 따라 평가하였다. 유창성은 문법적 정확성과 자연스러움을, 일관성은 원문과 요약문 간의 논리적 일치를, 관련성은 요약문이 원문의 핵심 내용을 포함하고 있는지를 평가하는 기준이다. 평가 기준에 대한 구체적인 설정과 세부 내용은 2장에서 다룬다.

위 과정을 통해 얻은 인간 평가 점수를 사용하여 자동 평가 점수와의 상관계수를 산출하는 실험을 진행하였다. 평가 기준별로 상관계수를 산출하여 각 자동 평가 지표가 요약문의 어떤 요소에서 인간 평가를 잘 모사하는지 분석하였다. 또한, 문서 유형별 상관계수를 산출하여 다양한 형태의 한국어 문서에서 각 문서 유형에 적합한 자동 평가 지표가 무엇인지 검증하였다.

본 연구의 기여점은 다음과 같다. 첫째, 본 연구는 한국어 생성 요약의 메타 평가를 수행하여 기존에 연구가 미비했던 한국어를 대상으로 생성 요약 성능 평가 지표의 적합성을 분석하였다. 둘째, 다양한 문서 유형으로 구성된 데이터셋을 사용하여 한국어 생성 요약에 대한 평가가 더 포괄적이며, 향후 응용 가능한 결과를 도출하였다. 셋째, 한국어 요약문에 대한 인간 평가를 직접 시행하여 자동 평가 지표와의 상관 분석 결과의 신뢰성을 높였다. 마지막으로, 인간 평가와 자동 평가 지표 간 상관관계를 분석하여 한국어 요약 평가에서 적합한 자동 평가 지표를 검증하고 그 유효성을 입증하였다.

2. 이론적 배경

2.1 요약 자동 평가 지표

인간 평가(Human Evaluation)는 사람이 직접 본문과 요약문을 읽고 유창성(Fluency), 일관성(Consistency), 관련성(Relevance) 등의 사전에 설정된 평가 기준에 따라 요약문을 평가하는 방법이다. 그러나 이 방법은 전문 인력을 비롯하여 많은 시간과 비용이 필요하며, 평가자의 일관성을 유지하는 데 어려움이 있어, 효율적이고 객관적인 평가 방법의 필요성이 제기된다. 이러한 이유로 인간 평가를 모사하며, 보다 효율적이고 정확한 요약 성능을 평가할 수 있는 자동 평가 지표에 관한 연구가 활발히 이루어지고 있다[11-13].

요약 자동 평가 지표는 일반적으로 내재적 평가 지표, 외재적 평가 지표로 분류된다[18]. 내재적 평가 지표는 원본 문서

와 요약문 간의 유사성을 기반으로 요약문의 품질을 평가하는 방법이다. 이는 요약문의 관련성, 유창성, 문법적 정확성 등을 측정하는 데 유용하며, 대표적인 자동 평가 지표로는 ROUGE, BLEU[19], BERTScore가 있다. 한편, 외재적 평가 지표는 요약문이 실제 응용에서 얼마나 유용한지를 평가하는 방법으로, 요약문이 특정 작업을 수행하는 데 얼마나 도움이 되는지를 측정한다. 예를 들어, 요약문을 읽고 질문에 답하는 능력을 평가하는 방식이 있다[20]. 이 평가 방법은 요약문의 효용성을 간접적으로 평가하는 데 유용하다.

본 연구에서는 요약 평가 지표로 ROUGE, BLEU, BERTScore, BLEURT[21], HaRiM+[22] 5가지 자동 평가 지표를 사용한다. ROUGE와 BLEU는 전통적인 n-gram 기반 지표로, 요약문과 원문의 일치율을 평가하며, BERTScore와 BLEURT는 사전학습된 언어모델을 활용하여 문맥적 유사성을 반영하므로 더 정교한 평가가 가능하다. HaRiM+는 최신 연구에서 대규모 언어모델의 환각(hallucination) 문제를 감지하기 위해 제안된 평가 지표이다.

1) ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)[8]는 텍스트 요약 성능을 평가하기 위한 대표적인 자동 평가 지표로, 재현율(Recall)을 기반으로 원본 텍스트와 요약문 간의 유사성을 측정한다. 2004년 Chin-Yew Lin에 의해 개발된 ROUGE는 참조 요약문과 자동 생성 요약문 간의 n-gram 일치율을 계산하여 평가하는 방식으로, 대규모 요약 연구에서 널리 사용되는 표준적인 성능 평가 방법이다. 주로 1-gram, 2-gram 그리고 최장 공통부분 순서(Longest Common Subsequence, LCS)를 기반으로 산출한 ROUGE-1, ROUGE-2, ROUGE-L을 사용한다. ROUGE는 빠르고 효율적인 평가가 가능하지만, 의미적 유사성을 충분히 반영하지 못한다는 한계가 있어 최근에는 BERTScore[9]와 같은 의미 기반 평가 지표와 함께 사용된다.

2) BLEU

BLEU (Bilingual Evaluation Understudy)[19]는 원래 기계번역 성능을 평가하기 위해 개발된 n-gram 기반 자동 평가 지표로, 요약문과 원본 텍스트 간의 n-gram 유사성을 측정한다. BLEU는 주로 1-gram부터 4-gram까지의 n-gram을 사용하며, 짧은 요약문이 과도하게 높은 점수를 받지 않도록 길이 패널티(Brevity Penalty)를 적용하여 점수를 보정한다. 한편, BLEU 또한 n-gram기반 지표이므로 문맥적 의미를 충분히 반영하지 못하는 한계를 가지고 있다.

3) BERTScore

BERTScore[9]는 사전 학습된 BERT (Bidirectional Encoder Representations from Transformers)[23] 모델을 활용한 평가 지표로, 원본 텍스트와 요약문 간의 의미적 유사성을 측정한다. 각 단어의 임베딩 벡터를 계산한 후, 두 텍스트 사이의 코사인 유사도를 통해 유사성을 평가한다. BERTScore는 BLEU나 ROUGE와 같은 전통적인 n-gram 기반 지표와 달리 의미를 반영할 수 있다는 장점이 있어, 요약 성능 평가에 널리 사용되고 있다.

BERTScore는 요약 성능 평가에서 정밀도(Precision), 재현율(Recall), F1 점수를 제공한다. 정밀도는 모델이 생성한 요약문이 원본 문서와 얼마나 정확하게 일치하는지를 나타내며, 재현율은 원본 문서의 중요한 내용을 요약문이 얼마나 잘 포착하는지를 평가한다. F1 점수는 정밀도와 재현율의 균형을 평가하여, 요약문의 전반적인 품질을 종합적으로 판단한다.

본 연구에서는 다국어 데이터로 사전학습된 BERT 모델[1]을 사용한 평가 지표는 BERTScore, 한국어 뉴스 댓글로 사전학습된 BERT 모델[2]을 사용한 평가 지표는 KoBERTScore로 표기하였다.

4) BLEURT

BLEURT (Bilingual Evaluation Understudy with Representations from Transformers)[21]는 사전 학습된 BERT 모델을 활용하여 문맥적 유사성과 의미적 차이를 평가하는 자동 평가 지표이다. BLEURT는 대규모의 인간 평가 데이터를 기반으로 학습되었으며, 단순한 n-gram 일치가 아닌 문맥적 유사성과 의미적 차이를 평가할 수 있다. 이 지표는 요약문이 원문과 얼마나 유사한지, 또는 어느 정도로 의미를 정확히 반영하는지를 판단하는 데 유용하다.

5) HaRiM+

HaRiM+[22]는 생성된 요약의 품질을 평가하기 위해 환각(hallucination) 가능성을 측정하는 기준을 사용하는 참조 없는 평가 지표이다. 환각 가능성이란, 요약문이 원문에 없는 정보를 생성할 가능성을 의미한다. 이 평가 지표는 추가 모델 훈련이나 모듈 없이 토큰 가능성에 기반하여 환각 가능성을 계산한다. HaRiM+는 FRANK[24], QAGS[25], SummEval[26] 세 가지 요약 벤치마크에서 인간 평가 점수와 높은 상관관계를 보여 환각 가능성을 평가하는 데 유용한 도구임을 입증하였다.

2.2 상관계수

상관계수(Correlation Coefficient)는 두 변수 간의 상관관계의 정도를 나타내는 통계적 지표로, -1에서 1 사이의 값을 가진다. 값이 1에 가까울수록 두 변수는 강한 양의 상관관계를, -1에 가까울수록 강한 음의 상관관계를 가지는 것을 의미한다. 0에 가까운 값은 상관관계가 거의 없음을 나타낸다. 각 상관계수는 p-value가 0.05 미만일 때 유의미한 상관관계를 가진다고 해석한다. 본 연구에서는 Pearson, Spearman, Kendall 세 가

1) https://huggingface.co/google-bert/bert-base-multilingual-cased
2) https://huggingface.co/beomi/kcbert-base

지 상관계수를 사용하여 자동 평가 지표와 인간 평가 간의 상관관계를 분석한다.

1) Pearson 상관계수

Pearson 상관계수는 두 변수 간의 선형 상관관계를 측정하는 지표로, 두 변수의 공분산을 각 변수의 표준편차로 나누어 계산한다. Pearson 상관계수는 다음과 같은 공식을 따른다.

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}} \quad (1)$$

여기서 X_i와 Y_i는 각 변수의 값이며, $\overline{X}$와 $\overline{Y}$는 각 변수의 평균을 나타낸다.

2) Spearman 상관계수

Spearman 상관계수는 두 변수 간의 순위 상관관계를 측정하는 지표이다. 데이터의 순위 차이를 통해 상관관계를 계산하며, 순위 차이가 작을수록 두 변수 간의 관계가 강하다고 본다. Spearman 상관계수는 다음과 같은 공식을 따른다.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (2)$$

여기서 d_i는 각 관측값에 대한 두 변수의 순위 차이 (X_i의 순위와 Y_i의 순위 간 차이), n은 데이터 쌍의 총 개수(표본 크기)를 의미한다.

3) Kendall 상관계수

Kendall 상관계수는 두 변수 간의 순위 일치도를 측정하는 지표로, 순위가 일치하는 쌍(Concordant pairs)과 일치하지 않는 쌍(Discordant pairs)의 수를 비교하여 계산한다. Kendall 상관계수는 다음과 같은 공식을 따른다.

$$\tau = \frac{C - D}{\sqrt{(C + D + T_1)(C + D + T_2)}} \quad (3)$$

여기서 C는 순위가 일치하는 쌍의 수, D는 순위가 일치하지 않는 쌍의 수, T_1과 T_2는 각 변수에서 동률인 쌍의 수이다.

2.3 한국어 요약 모델

요약 태스크에는 인코더-디코더 모델이 효과적임은 실험적으로 증명되어 널리 알려진 사실이다[27]. 인코더는 전체 본문의 내용을 이해하고 중요한 정보를 학습하고, 디코더는 이를 바탕으로 요약문을 생성하는 방식으로 동작한다. 이러한 구조는 복잡한 문맥을 요약하는 데 뛰어난 성능을 발휘한다. 특히, 한국어로 사전 학습된 인코더-디코더 모델은 제한적이므로 본 논문에서는 공개된 한국어 인코더-디코더 모델 중 성능이 뛰어난 T5 모델[3] 과 KoBART 모델[4] 그리고 최근 다양한 분야에서 뛰어난 성능을 보이는 대규모 언어 모델(Large Language Model, LLM)인 GPT-3.5 Turbo 모델을 요약 모델로 사용하였다.

3. 실 험

3.1 데이터셋

본 연구는 한국어 생성 요약의 자동 평가 지표 성능을 검증하기 위해 AI 허브에서 제공한 '요약문 및 레포트 생성 데이터'[13]를 활용하였다. 이 데이터셋은 다양한 문서 유형(보도자료, 사설, 역사・문화재, 문학, 회의록, 나레이션, 뉴스기사, 보고서, 간행문, 연설문)에 대해 본문 텍스트와 사람이 작성한 참조 요약문, 추출 요약문이 포함되어 있다.

본 연구에서 사용된 데이터셋은 회의록, 문학, 연설, 나레이션, 보도자료와 같이 서로 다른 성격을 지닌 다양한 문서 유형을 포함하고 있다. Table 1에서 일부 문서 유형별 데이터에서 추출한 예시를 보인다. 각 문서 유형은 각 문서가 작성된 배경이 다르며, 본문 구조와 정보 전달 방식에서 차이를 보여, 요약 모델이 다양한 형태의 텍스트를 어떻게 처리하는지를 평가할 수 있는 기준을 제공한다. 예를 들어, 회의록과 같은 대화 데이터는 발화 간의 관계를 유지해야 하며, 문학과 같은 인물에 중점을 둔 텍스트 데이터는 감정적 요소를 적절히 요약해야 한다. 연설 및 보도자료는 명확한 정보 전달에 중점을 두고, 나레이션은 상황의 전개와 맥락을 중심으로 요약되어야 한다. 또한, 회의록과 연설은 다른 문서들과 달리 대화체라는 언어적 특성을 고려하여 요약해야 한다.

Table 1. Data Examples by Document Type

Document Type	Data Example
minute	윤후덕 위원] "그것도 부족하다고 생각하시지요 김광재 이사장님?" 한국철도시설공단이사장 김광재] "예 그동안의 투자비를 감안한다면 아마 한 30~40년 되어야 투자비 회수가 되는 수준으로 알고 있습니다." (생략)
literature	이 생각 저 생각에 설어지면 품에 지닌 사진을 몇 번이고 몇 번이고 꺼내보았다. 사진을 들여다보면 그는 재 없이 한바탕 울고야 말았다. 그러나 눈물이 마를 만하면 그는 또다시 사진을 꺼내 보았다. 이 지옥에 들어온지 삼년 동안 그 사진만이 그의 유일한 동무였고 위안이었다. (생략)
speech	안녕하십니까? (생략) 2019년도 농식품 수출 동향 및 수출사업 추진내역에 대해서 설명드리겠습니다.
narration	(생략) 낭만이 있는 나라 세인트루시아. 이 나라는 훌륭한 야외결혼식 장소와 멋진 신혼여행지로도 유명하다. (생략) 보트를 타고 카리브해로 나가보기로 했다.
briefing	〈보도 주요 내용(뉴스1, 2020. 4. 21)〉 국회사무처는 제21대 국회 개원을 앞두고, 정보기기의 내구연한과 실소요를 고려하여 과거보다 예산을 대폭 절감 집행하고 있습니다. (생략)

3) https://huggingface.co/paust/pko-t5-base
4) https://huggingface.co/gogamza/kobart-base-v2

Table 2. Criteria for Human Evaluation

	Point 1	Point 2	Point 3	Point 4	Point 5
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Fluency	The sentence is grammatically flawless, natural, easy to understand, and logical.				
Consistency	The summary follows the original text in terms of meaning, topic, subject, action, human relationships, etc.				
Relevance	The summary captures the keywords and main ideas from the original text.				

이처럼 다양한 문서 유형에서 생성된 요약문을 단일한 평가 기준으로 평가하기에는 복잡성이 크다. 따라서 본 연구는 각 문서 유형에 적합한 평가 메트릭을 제안하고, 이를 통해 한국어 요약 연구에서 더욱 신뢰성 있는 자동 평가 지표를 검증할 기초를 마련하고자 한다.

3.2 인간 평가

본 연구의 주요 목표는 한국어 생성 요약의 자동 평가 지표와 인간 평가 간 상관관계를 분석하는 것이다. 이를 위해 다양한 유형의 문서에서 생성된 요약문을 대상으로 인간 평가를 진행하고, 그 결과를 자동 평가 지표(ROUGE, BLEU, BERTScore 등)와 비교하였다. 이 실험을 통해 자동 평가 지표의 신뢰도와 유효성을 분석하고자 한다.

실험에 사용한 AI허브의 '요약문 및 레포트 생성 데이터셋'에는 보도자료, 문학, 회의록, 나레이션, 연설문 등의 문서 유형이 포함되어 있다. 각 문서 유형에서 10개의 문서를 무작위로 추출하여 총 100개의 문서를 사용하였으며, 각 문서에 대해 참조 요약문과 KoBART, T5, GPT-3.5 Turbo 모델로 생성한 요약문을 평가 대상으로 선정하였다. 최종적으로 각 문서에 대해 총 400개의 요약문이 평가되었다.

인간 평가는 유창성(Fluency), 일관성(Consistency), 관련성(Relevance)이라는 세 가지 기준으로 진행되었다. 각 평가 기준은 Table 2와 같이 정의된다. 유창성은 요약문이 문법적으로 올바르고 자연스럽게 작성되었는지를 평가하며, 일관성은 요약문이 원문과 사실적으로 일치하는지를 평가한다. 관련성은 요약문이 원문의 핵심 내용을 잘 반영하고 있는지를 평가한다. 선정한 평가 기준을 기반으로 리커트 5점 척도로 요약문을 평가한다. 평가자로서 세 명의 컴퓨터공학 전공 대학원생이 참여하였으며, 각 평가자는 요약문 출처(참조 요약문, KoBART, T5, GPT-3.5 Turbo)를 알지 못한 상태에서 평가를 진행하였다. 모든 평가자에게 Table 2를 제공하여 사전에 통일된 기준을 통해 평가의 일관성을 유지하였다.

Table 3. Average Scores of Human Evaluation by criteria for each summary

	reference	T5	KoBART	GPT-3.5 Turbo
Fluency	4.82	4.71	3.97	4.64
Consistency	4.83	4.73	4.14	4.63
Relevance	4.07	4.03	3.83	4.39

Table 3에 따르면 인간 평가 결과, 네 가지 요약문 중 참조 요약문이 모든 평가 기준에서 가장 높은 평가를 받았다. 나머지 세 가지 생성 요약문 중에서 유창성, 일관성에서는 T5 모델이, 관련성에서는 GPT-3.5 Turbo 모델이 가장 높은 점수를 받았다. KoBART 모델은 모든 평가 기준에서 가장 낮은 점수를 받았다.

Table 4에서는 참조 요약문에 대한 문서 유형별 인간 평가 점수 결과를 보인다. 회의록, 간행문, 뉴스기사는 유창성, 일관성, 관련성 측면에서 전반적으로 좋은 평가를 받았으며, 특히 회의록은 주제 안건의 명확성 덕분에 제일 높은 평가를 받았다. 반면, 보도자료는 4.18점으로 가장 낮은 평균 점수를 보였다. 특히 관련성이 3.30으로 매우 낮게 나왔다. 이러한 결과는 문서 유형에 따라 요약 성능이 달라질 수 있음을 시사하며, 요약 모델이 다양한 문서 유형에 적합하게 최적화될 필요가 있음을 보여준다.

3.3 인간 평가-자동 평가 지표 상관 분석

1) 평가 기준별 상관 분석

본 연구에서는 T5 및 KoBART, GPT-3.5 Turbo 모델이 생성한 요약문에 대해 인간 평가와 자동 평가 지표 간 상관관계를 분석하였다. Table 5, Table 6, Table 7은 각각 T5, KoBART, GPT-3.5 Turbo 모델 요약문의 인간 평가 점수와 자동 평가 지표 간 상관분석 결과이다. 각 표에서는 세 가지 평가 기준별 상관계수를 구분하여 제시함으로써, 자동 평가 지표가 각 기준(유창성, 일관성, 관련성)에서 어떻게 다르게 작동하는지 명확히 파악하고자 했다. 본 상관 분석 실험에서는 Pearson 상관계수(r), Spearman 상관계수(ρ), Kendall 상관계수(τ)를 사용하여 상관 분석을 수행했다. 단, p-value가 0.05 미만인 상관계수 값은 밑줄 한 개, p-value가 0.01 미만인 상관계수 값은 밑줄 두 개로 표시하였다.

Table 5에서 알 수 있듯이, T5 모델의 상관 분석 결과, 유창성

Table 4. Human Evaluation by document type for reference summaries

	briefing	edit	his・cul	literature	minute	narration	news	paper	public	speech
Fluency	4.40	4.60	4.80	4.93	4.97	4.97	4.83	4.87	4.93	4.87
Consistency	4.83	4.70	4.63	4.83	5.00	4.77	4.83	4.80	4.97	4.90
Relevance	3.30	4.17	4.00	4.07	4.40	3.87	4.40	4.07	4.23	4.17
AVG.	4.18	4.49	4.48	4.61	4.79	4.53	4.69	4.58	4.71	4.64

Table 5. Correlation Coefficients between Human Scores and Evaluation Metrics for T5 Summaries

	Fluency			Consistency			Relevance		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
ROUGE-1	0.08	0.18	0.13	0.23	0.29	0.23	0.13	0.23	0.18
ROUGE-2	0.07	0.18	0.15	0.25	0.32	0.26	0.12	**0.254**	0.20
ROUGE-L	0.09	0.17	0.13	0.23	0.31	0.24	0.10	0.21	0.16
BLEU-1	-0.07	0.01	0.01	0.05	0.16	0.13	0.04	0.08	0.06
BLEU-2	-0.02	0.04	0.03	0.02	0.13	0.10	-0.05	-0.02	-0.02
BLEU-3	-0.00	0.06	0.05	-0.01	0.10	0.07	-0.11	-0.10	-0.07
BLEU-4	0.01	0.07	0.05	-0.03	0.05	0.04	-0.15	-0.16	-0.11
BERT $Score_p$	0.08	0.16	0.13	0.21	0.26	0.21	0.06	0.17	0.13
BERT $Score_r$	0.11	0.18	0.14	0.26	**0.33**	0.26	0.17	0.255	0.19
BERT $Score_{f1}$	0.10	0.18	0.14	0.24	0.30	0.24	0.12	0.22	0.19
KoBERT $Score_{f1}$	0.12	0.19	0.15	0.26	0.30	0.23	0.15	0.22	0.17
HaRiM+	0.02	0.13	0.10	-0.06	0.02	0.02	-0.06	-0.10	-0.07
BLEURT	-0.10	-0.04	-0.03	-0.09	-0.06	-0.05	0.12	0.10	0.07

Table 6. Correlation Coefficients between Human Scores and Evaluation Metrics for KoBART Summaries

	Fluency			Consistency			Relevance		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
ROUGE-1	0.28	0.23	0.17	0.25	0.24	0.17	0.28	0.33	0.24
ROUGE-2	0.20	0.15	0.11	0.21	0.20	0.14	0.22	0.28	0.20
ROUGE-L	0.23	0.19	0.14	0.26	0.24	0.17	0.27	0.30	0.21
BLEU-1	0.26	0.24	0.17	0.10	0.04	0.03	-0.10	-0.07	-0.05
BLEU-2	0.22	0.19	0.14	0.11	0.05	0.04	-0.17	-0.11	-0.08
BLEU-3	0.20	0.14	0.10	0.12	0.07	0.05	-0.20	-0.14	-0.10
BLEU-4	0.18	0.12	0.09	0.13	0.09	0.06	-0.22	-0.15	-0.11
BERT $Score_p$	**0.33**	0.28	0.21	0.36	0.34	0.25	0.18	0.23	0.17
BERT $Score_r$	0.22	0.16	0.12	0.28	0.28	0.20	**0.40**	0.39	0.29
BERT $Score_{f1}$	0.30	0.22	0.16	0.35	0.34	0.25	0.32	0.34	0.25
KoBERT $Score_{f1}$	0.19	0.16	0.12	**0.38**	0.36	0.27	0.28	0.31	0.23
HaRiM+	0.12	0.10	0.08	0.12	0.11	0.09	-0.12	-0.02	-0.01
BLEURT	-0.16	-0.17	-0.12	0.04	0.02	0.02	0.04	0.00	0.00

에서는 어떠한 자동 평가 지표도 인간 평가와 유의미한 상관관계를 보이지 않았다. 반면, 일관성과 관련성에서 일부 자동 평가 지표는 인간 평가와 유의미한 상관관계를 보였다. 일관성 기준에서는 ROUGE-1, ROUGE-2, ROUGE-L과 $BERTScore_p$, $BERTScore_r$, $BERTScore_{f1}$, $KoBERTScore_{f1}$이 유의미한 상관관계를 나타냈으며, 이 중에서도 $BERTScore_r$이 0.33으로 가장 높은 상관계수를 보였다. 관련성 기준에서는 ROUGE-1, ROUGE-2, ROUGE-L과 $BERTScore_r$, $BERTScore_{f1}$, $KoBERTScore_{f1}$이 유의미한 상관관계를 보였으며, $BERTScore_r$이 0.255로 가장 높은 상관계수를 보였다.

Table 6의 KoBART 모델의 상관 분석 결과에서는 유창성, 일관성, 관련성 모두에서 유의미한 상관관계를 보였다. 유창성 기준에서는 ROUGE-1, ROUGE-L, BLEU-1, BLEU-2와 $BERTScore_p$, $BERTScore_r$, $BERTScore_{f1}$이 유의미한 상관관계를 나타냈으며, 이 중 $BERTScore_p$가 0.33으로 가장 높은 상관계수를 보였다. 일관성 기준에서는 ROUGE-1, ROUGE-2, ROUGE-L과 $BERTScore_p$, $BERTScore_r$, $BERTScore_{f1}$, $KoBERTScore_{f1}$이 유의미한 상관관계를 보였으며, $KoBERTScore_{f1}$이 0.38로 가장 높은 상관계수를 보였다. 관련성 기준에서도 일관성 기준과 동일한 자동 평가 지표와 유의미한 상관관계를 보였으며, $BERTScore_r$가 0.40으로 가장 높은 상관계수를 보였다.

한편, Table 7의 GPT-3.5 Turbo 모델의 상관 분석 결과, 유창성 기준에서는 인간 평가와 유의미한 상관관계를 보인 자동 평가 지표가 없었다. 또한 T5 모델과 KoBART 모델에 비해 전반적으로 자동 평가 지표 점수와 낮은 상관계수를 보이는 경향이 있었다. 그러나 일관성과 관련성 기준에서 유일하게 HaRiM+가 유의미한 상관관계를 보였다. 일관성에서는 0.21, 관련성에서는 0.23으로 가장 높은 상관계수를 보였다.

Table 7. Correlation Coefficients between Human Scores and Evaluation Metrics for GPT-3.5 Turbo Summaries

	Fluency			Consistency			Relevance		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
ROUGE-1	0.11	0.15	0.12	0.01	0.05	0.04	0.15	0.21	0.16
ROUGE-2	0.16	0.10	0.08	0.13	0.14	0.10	0.16	0.18	0.13
ROUGE-L	0.17	0.14	0.10	0.07	0.05	0.04	0.16	0.20	0.15
BLEU-1	0.02	-0.00	0.00	-0.06	-0.01	-0.01	-0.09	-0.08	-0.06
BLEU-2	0.05	0.03	0.02	-0.06	-0.04	-0.03	-0.13	-0.16	-0.12
BLEU-3	0.06	0.06	0.05	-0.05	-0.05	-0.04	-0.14	-0.18	-0.13
BLEU-4	0.06	0.08	0.06	-0.05	-0.03	-0.02	-0.15	-0.18	-0.13
BERT $Score_p$	-0.03	-0.02	-0.01	0.03	0.09	0.07	0.07	0.09	0.07
BERT $Score_r$	0.06	0.06	0.05	0.02	0.06	0.04	0.15	0.17	0.12
BERT $Score_{f1}$	0.02	0.01	0.00	0.03	0.04	0.04	0.12	0.14	0.11
KoBERT $Score_{f1}$	0.09	0.07	0.06	0.02	0.06	0.06	0.11	0.13	0.09
HaRiM+	0.12	0.05	0.04	0.17	**0.21**	0.16	0.19	**0.23**	0.17
BLEURT	-0.07	0.01	0.00	-0.02	-0.02	-0.01	-0.07	-0.13	-0.09

2) 문서 유형별 상관 분석

3.3.2에서는 T5, KoBART, GPT-3.5 Turbo 세 가지 요약 모델에 대한 문서 유형별 세 가지 상관계수(Pearson, Spearman, Kendall)를 히트맵을 통해 시각화하고 분석하였다. 상관계수를 통해 인간 평가 점수와 자동 평가 지표 간의 일치 정도를 측정함으로써 각 문서 유형별로 적합한 자동 평가 지표를 파악하고자 하였다. 더 나아가, 이러한 결과는 목표 작업에 적합한 평가 지표를 제시하는 데 기여할 수 있다.

문서 유형별 상관계수를 산출한 결과, 각 문서 유형에 따라 T5, KoBART, GPT-3.5 Turbo 모델의 상관계수가 다르게 나타났다. Fig. 1에서 T5 모델은 보도자료에서 BLEU-1에서 0.73, BLEU-2에서 0.68, BLEU-3에서 0.63의 높은 상관관계를 기록하고, 보고서에서는 ROUGE-1에서 0.64의 높은 상관관계를 기록하였다. 또한, 역사·문화재 문서에서는 BLEURT에서 0.71로 높은 상관관계를 나타냈으며, 회의록에서는 BLEU에서 0.75로 가장 높은 상관관계를 보였다. 이 문서들은 모두 정보가 명확하고 구조화된 특성을 지닌 문서들로, 이러한 문서 특징이 상관관계에서 반영된 것으로 해석할 수 있다. 반면, 나레이션과 뉴스기사에서는 전반적으로 매우 낮은 상관관계를 보여, 서술 방식과 정보 전달 방식이 다소 다양하거나 비일관적일 수 있어, 이러한 문서 유형에서는 자동 평가 지표가 인간 평가와 일관된 상관관계를 유지하기 어려운 것으로 나타났다.

Fig. 2에서 KoBART 모델은 회의록에서 $BERTScore_p$에서 0.82, $BERTScore_{f1}$에서 0.67, BLEU-3에서 0.69, BLEU-4에서 0.72로 높은 상관관계를 보였고, 나레이션 문서에서도 ROUGE-1에서 0.78, ROUGE-2에서 0.66, ROUGE-3에서 0.80, $BERTScore_p$에서 0.74, $BERTScore_r$에서 0.78, $BERTScore_{f1}$에서 0.80으로 매우 높은 상관관계를 기록하였으며, 문학 문서에서도 ROUGE-2에서 0.71로 높은 상관관계를 나타냈다. 이 문서들은 모두 내용의 흐름과 일관성이 중요한 서술적 특성을 지니고 있어, 이러한 문서 특성이 상관관계에 반영된 것으로 보인다. 반면, 보고서와 연설문에서는 전반적으로 낮은 상관관계를 보였으며, 특히 BLEU 지표에서 매우 낮은 상관관계를 기록하였다.

Fig. 3에서 GPT-3.5 Turbo 모델에서는 유일하게 사설 문서에서만 유의미한 상관관계를 기록하였다. $BERTScore_p$에서 0.75, $BERTScore_r$에서 0.67, $BERTScore_{f1}$에서 0.74, BLEURT에서 0.68로 사설 문서의 요약에서 비교적 높은 상관관계를 보였다. GPT-3.5 Turbo 모델은 세 요약 모델 중에서 전반적으로 인간 평가와 가장 낮은 상관관계를 보였으며, 이는 GPT-3.5 Turbo 모델이 대규모 언어모델임에도 불구하고 한국어 데이터에 대한 추가적인 훈련 없이 제로샷 방식을 채택했기 때문에, 자동 평가 지표와의 상관관계가 낮게 나타났을 가능성이 있다. 특히, 대규모 언어모델은 창의적이고 유연한 텍스트 생성을 목표로 하므로 n-gram 일치율이나 문맥적 유사성을 평가하는 기존 자동 지표와 일관된 상관관계를 보이기 어려울 수 있다.

4. 결 론

요약 모델 연구에서는 자동 평가 지표의 한계로 인해 인간 평가를 추가로 수행하는 것이 보편적이다. 그러나 인간 평가는 큰 비용이 필요하므로 새로운 자동 평가 지표의 제안 및 기존 자동 평가 지표들을 평가하는 메타 평가 관련 연구가 활발히 진행되고 있다. 한편, 한국어 요약 모델에 한해서는 관련 연구가 미비한 상황이다. 이에 따라 본 연구는 한국어 생성 요약 평가 지표에 대한 포괄적 분석으로, 향후 한국어 자연어 처리 연구에 응용 가능한 기초 자료를 제공하고자 하였다.

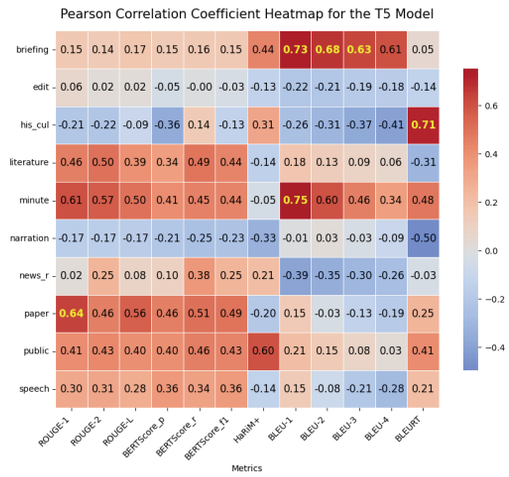


Fig. 1. Heatmap of Pearson Correlation Coefficients by document type for the T5 summaries

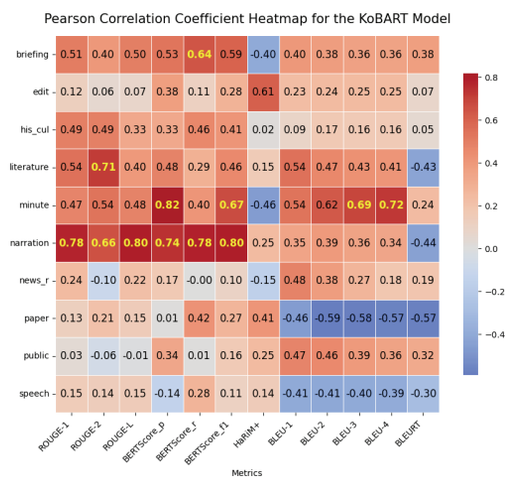


Fig. 2. Heatmap of Pearson Correlation Coefficients by document type for the KoBART summaries

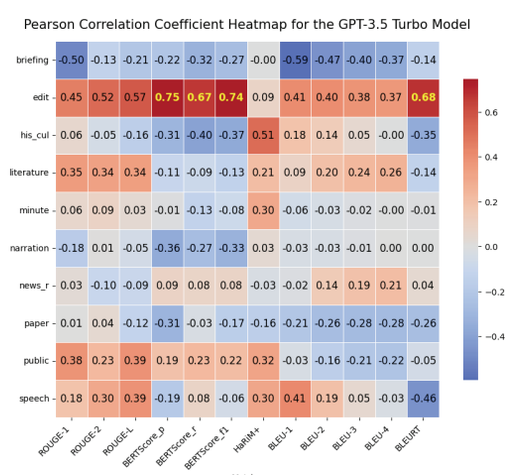


Fig. 3. Heatmap of Pearson Correlation Coefficients by document type for GPT-3.5 Turbo summaries

실험은 크게 평가 기준별 상관 분석, 문서 유형별 상관 분석으로 나누어 진행하였다. 평가 기준별 상관 분석을 통해 각 자동 평가 지표가 모델별 요약문을 얼마나 효과적으로 평가하는지 확인할 수 있었다. T5와 KoBART 모델의 경우, BERT Score가 유창성 및 일관성 기준에서 가장 높은 상관관계를 나타냈으며, 이는 BERTScore가 해당 모델이 생성한 요약문을 문맥적 유사성 측면에서 잘 평가하고 있음을 의미한다. 반면, GPT-3.5 Turbo 모델에서는 일관성과 관련성 기준에서 HaRiM+ 지표와 높은 상관관계를 보였으며, 이는 환각 가능성 감지를 목적으로 한 HaRiM+가 이 모델의 요약문을 평가하는 데 적합하다는 것을 시사한다. 이는 각 모델의 요약문 특성에 맞는 평가 지표를 선택하는 것이 중요하다는 것을 의미한다.

문서 유형별 상관 분석 결과는 각 평가 지표가 특정 문서 유형에서 생성된 요약문을 어떻게 평가하는지를 보여준다. T5 모델의 경우, 정보가 명확하고 구조화된 문서 유형인 보도자료와 회의록에서 BLEU 지표와 높은 상관관계를 보였는데, 이는 BLEU가 이러한 문서의 요약문 평가에 적합하다는 것을 시사한다. KoBART 모델은 서사적 요소가 강한 나레이션 문서에서 BERTScore와 높은 상관관계를 나타냈으며, 이는 BERTScore가 나레이션과 같이 문맥적 일관성이 중요한 문서에서 신뢰성 있는 평가를 제공할 수 있음을 보여준다. 반면, GPT-3.5 Turbo 모델은 대부분의 문서 유형에서 낮은 상관관계를 보였으나, 사설 문서에서는 HaRiM+ 지표와의 상관관계가 높아, 이 메트릭이 사설 문서 요약 평가에서 유용할 가능성을 시사한다. 이러한 결과는 특정 메트릭이 특정 문서 유형에서 더 효과적으로 요약문을 평가할 수 있음을 보여주며, 각 문서 유형에 맞는 적절한 평가 지표를 선택하는 것이 요약 평가에서 중요하다는 점을 강조한다.

다시 말해, 본 연구는 해결하고자 하는 요약 문제에 따라 문서 특징과 원하는 요약문이 다르므로 상황에 맞는 평가 지표를 정하는 근거를 제공한다는 점에서 의의가 있다. 예를 들면, 대규모 언어모델을 사용하여 요약문을 생성하는 경우 환각 문제를 잡아내는 것이 중요하므로 GPT-3.5 Turbo 요약문의 일관성 평가를 잘 하는 HaRiM+를 사용해 평가할 수 있다. 또한 모든 평가 기준에서 전반적으로 ROUGE보다 BERTScore가 상관계수가 높다는 점을 확인함으로써 한국어 요약문 평가에 BERTScore가 적합함을 보인 것에 의의가 있다.

반면, 본 연구는 몇 가지 한계점이 존재한다. 첫째, 대규모 언어모델 요약문의 경우, 제로샷(zero-shot) 생성 요약문이므로 한국어 데이터로 사전학습된 언어모델에서 생성한 다른 요약문들의 점수 비교에 불리할 수 있다. 둘째, 인간 평가가 컴퓨터공학 전공 대학원생 3명에 의해 수행되었기 때문에 평가자 수가 제한적이기 때문에 평가 결과의 일반화에 한계가 있을 수 있고, 언어 전문가가 아니라는 점에서 평가 결과에 신뢰성에 대한 우려가 있을 수 있다. 이러한 한계에도 불구하고 본 연구는 한국어 생성 요약의 자동 평가 지표에 대한 중요한 통찰을 제공한다. 향후 연구에서는 더 다양한 데이터셋과 평가 모델을 활용하여 연구를 확장하고, 평가자 수를 늘려 인간 평가의 신뢰성을 높이는 방향으로 진행될 필요가 있다.

References

[1] D. Yadav, J. Desai, and A. K. Yadav, “Automatic Text Summarization Methods: A Comprehensive Review,” *arXiv preprint arXiv:2204.01849*, 2022.

[2] R. Mihalcea and P. Tarau, “Textrank: Bringing order into text,” in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (EMNLP), Barcelona, Spain, pp.404-411, 2004.

[3] M.-H. Choi, T.-Y. Kim, J.-S. Shin, and S.-H. Koh, “Performance Comparison between Language Generation Models for Summarization of Customer Consultation,” in *Proceedings of Symposium of the Korean Institute of communications and Information Sciences (KICS)*, Gangwon, Korea, pp.932-933, 2023.

[4] J. Lee and H. Lee, “Entity-aware Medical Document Summarization using Large Language Model,” *The Institute of Electronics and Information Engineers (IEIE) Academic Presentation Paper Collection*, Jeju, Korea, pp.2820-2822, 2024.

[5] J.-W. Lee, T.-H. Kim, D.-G. Shin, and W.-S. Jo, “Korean Information Summary System for National R&D Projcet Information Summary,” *The Korea Institute of information and Communication Engineering (KIICE) Academic Presentation Paper Collection*, Jeju, Korea, pp.72-74, 2022.

[6] E. Kim and H. Lim. “Comparative study of legal document summary method based on pre-trained model.” in *Proceedings of Annual Conference of Korea Information Processing Society* (ACK), Vol.28, No.2, pp.614-617, 2021.

[7] J. K. Bae, “A Study on the Construction of Financial-Specific Language Model Applicable to the Financial Institutions,” *Journal of the Korea Industrial Information Systems Research*, Vol.29, No.3, pp.79-87, 2024.

[8] C. Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” *in Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, pp.74-81, 2004.

[9] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” *in Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020.

[10] C. van der Lee, A. Gatt, E. van Miltenburg, and E. Krahmer, “Human evaluation of automatically generated text: Current trends and best practice guidelines,” *Computer*

Speech & Language, Vol.67, pp.101151, 2021.

[11] M. Gao and X. Wan, "DialSummEval: Revisiting Summarization Evaluation for Dialogues," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp.5693-5709, 2022.

[12] S. Gabriel, A. Celikyilmaz, R. Jha, Y. Choi, and J. Gao, "Go Figure! A Meta Evaluation of Factuality in Summarization," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp.478-487, 2021.

[13] O. Honovich et al., "TRUE: Re-evaluating Factual Consistency Evaluation," in *Proceedings of the DialDoc Workshop at ACL* 2022, Online, pp.161-175, 2022.

[14] AI Hub, "요약문 및 레포트 생성 데이터," [Internet], https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=582

[15] SKT-AI, "KoBART: Korean Generative Pretrained Transformer," [Internet], https://github.com/SKT-AI/Ko BART.

[16] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, Vol.21, No.140, pp.1-67, 2020.

[17] OpenAI, "GPT-3.5 Turbo Documentation," [Internet], https://platform.openai.com/docs/models/gpt-3-5-turbo.

[18] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, Vol.165, pp.113679, 2021.

[19] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA, pp.311-318, 2002.

[20] M. Eyal, T. Baumel, and M. Elhadad, "Question Answering as an Automatic Evaluation Metric for News Article Summarization," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT), Minneapolis, MN, USA, pp.3938-3948, 2019.

[21] T. Sellam, D. Das, and A. P. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.7881-7892, 2020.

[22] S. J. Park and J. S. Lee, "HaRiM+: Evaluating Summary Quality with Hallucination Risk," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp.895-924, 2022.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT), Minneapolis, MN, USA, pp.4171-4186, 2019.

[24] A. Pagnoni, V. Balachandran, and Y. Tsvetkov, "Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp.4812-4829, 2021.

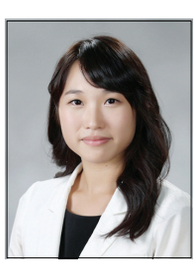
[25] S. Wang, P. Gao, Y. Zhang, and J. Li, "Asking and answering questions to evaluate the factual consistency of summaries," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.5008-5020, 2020.

[26] A. R. Fabbri, W. Kryscinski, B. McCann, C. Xiong, R. Socher, and D. R. Radev, "SummEval: Re-evaluating summarization evaluation," *Transactions of the Association for Computational Linguistics*, Vol.9, pp.391-409, 2021.

[27] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS 2014)*, Montreal, Canada, pp.3104-3112, 2014.

윤 세 휘

https://orcid.org/0009-0004-0192-6742
e-mail : inysher@inu.ac.kr
2022년 인천대학교 컴퓨터공학부(학사)
2022년~현 재 인천대학교 컴퓨터공학과 석사과정
관심분야 : 자연어 처리, 인공지능



신 유 현

https://orcid.org/0000-0001-7013-9057
e-mail : yhshin@inu.ac.kr
2019년 서울대학교 컴퓨터공학부(박사)
2020년~현 재 인천대학교 컴퓨터공학부 부교수
관심분야 : 자연어 처리, 음성 언어 이해, 인공지능