KeyBERTScore: Evaluation Metric for BERT-based Text Summarization

Kim YoungRok, Sang-Chul Kim
Kookmin University
goldhanwool@kookmin.ac.kr, sckim7@kookmin.ac.kr

***Abstract*—This paper proposes an evaluation metric for the qualitative assessment of text summarization using SBERT-enhanced KeyBERT for keyword similarity measurement. The proposed metric utilizes the KoBART model, which has undergone transfer learning on datasets of academic papers and legal texts, to summarize article texts. KeyBERT is then employed to extract keywords from both the original and summarized texts, and the similarity between these keywords is calculated. This process evaluates how well the summary captures the main keywords of the document.**

Keywords—Summarization, KoBART, KeyBERT, Language Model Evaluation

I. INTRODUCTION

The text summarization task involves extracting key information from an original document, and it plays a vital role in filtering information and compressing knowledge selectively. In this context, evaluating the quality of a summary is a significant challenge in the text summarization field.

Text summarization evaluation methods are divided into quantitative and qualitative assessments. Quantitative evaluation methods can measuring grammatical consistency, spelling, and the structural aspects of language through word continuity between the original document and the summary quickly and automatically. In contrast, qualitative methods focus on determining the extent to which the summarized text encompasses the important elements of the original content.

A major issue with qualitative evaluation methods, which are traditionally conducted by humans, is their subjectivity and inconsistency. The evaluator's personal opinions or levels of understanding can influence the results significantly, which can reduce reliability. In addition, qualitative evaluation methods are time consuming and labor intensive, especially for large datasets. [1]

To address these issues, this study proposes a qualitative evaluation method that utilizes language models. Large-scale pretrained language models have emerged with the continuing advancements in AI technologies. These models can understand complex linguistic structures and meanings; thus, they possess significant potential for text summarization evaluation.

The advantages of using language models for qualitative evaluation are manifold. First, objectivity and consistency are enhanced. The model is not influenced by subjective human opinions and can evaluate summaries based on consistent criteria, e.g., importance, abstraction, and factual accuracy. Second, they facilitate the low-cost evaluation of large datasets. An automated evaluation process enables quick and efficient assessment of extensive datasets. Third, they can be applied to various aspects of summary quality, e.g., fact checking, sentiment analysis, and topic inclusion, by integrating different language models to establish useful evaluation standards.

Evaluations by language models provide objective metrics about how well a summary reflects the essential information of the original text, thereby enabling objective comparison of summaries and contributing to the improvement of summary quality through automated feedback.

This study explores the practicality and effectiveness of using language models for qualitative evaluation and proposes a text summarization evaluation method.

II. RELATED WORK

According to a previous study [2], machine-generated summaries from current large-scale language learning models are preferred by users over human-written reference summaries. In addition to utilizing language models for text summarization, such models are being used to evaluate the quality of summaries.

A. BERT Score

The BERT score, which is language model evaluation metric for qualitative assessment, utilizes BERT to evaluate semantic coherence between sentences and words, moving beyond traditional N-gram–based word matching assessments. This approach was proposed to address the limitations of previous evaluation methods, e.g., BLEU and ROUGE, which focus solely on word matching and frequently miss the nuances of varied expressions in different contexts [3].

The BERT score utilizes the embedding component of the BERT model to obtain vector representations of the predicted summary and the reference summary. Then, it calculates the cosine similarity to express semantic similarity quantitatively.

B. GPT Score

The recently proposed GPT score utilizes the GPT model to evaluate summaries generated by language models, e.g., BART. The GPT score measures how well a generated summary S reflects the content of a given document D. It quantitatively assesses the extent to which the original document's content is accurately represented in the summary, providing a probabilistic measurement of the summary's quality.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

III. PROPOSAL

The proposed method first analyzes academic paper and legal document datasets from AI-HUB's public data to understand the characteristics and distribution of the data. This analysis informs the composition and size of the dataset required for training. Then, the structure and features of the KoBART [4] model are defined. To facilitate effective training of the model, a training framework is established using Pytorch Ignite and the Huggingface library. Summary inference is then performed on the article dataset using the fine-tuned KoBART model. The quality of the summary text is evaluated using the KeyBERT library, which utilizes sentence BERT (SBERT), which is a BERT-based language model. This process allows for qualitative analysis of the model's summarization capabilities, provides a comprehensive evaluation of the model's overall performance and summarization abilities, and identifies areas of improvement.

A. Proposed Model

The language model selected to assess text summarization quality is KoBART, which is pretrained on Korean data based on the BART architecture. BART is a Transformer-based model that is widely used for various natural language processing tasks, e.g., text summarization and translation. Having learned the complex vocabulary and grammatical structures of the Korean language, KoBART is optimized to process Korean text, thereby making it a crucial model for Korean text summarization tasks.

The KoBART model is divided into three models for inference on news texts, i.e., the base model without fine tuning, the paper model, which has undergone transfer learning on a dataset of academic papers, and the law model, which has been fine-tuned with a dataset of legal texts.

KeyBERT is then used for evaluation. KeyBERT uses the pretrained SBERT language model for sentence pair embeddings to identify important keywords and sentences between the document and the corresponding summary.

B. EmbedRank and Maximal Marginal Relevance Keyword Extraction Algorithms

After extracting keyword candidates, the principles of the EmbedRank [5] algorithm are applied, where the SBERT model is employed to embed both the original text and the keyword candidates. Then, the embedded document and keyword tokens are measured in terms of their similarity and relevance through cosine similarity in vector space. Here, relevance is ranked, and keywords are extracted in order of their importance.

$$\text{MMR} := \underset{C_i \in C \setminus K}{\arg\max} \Big[\lambda \cdot \widetilde{cos_{sim}}(C_i, \text{doc}) - (1-\lambda) \max_{C_j \in K} \widetilde{cos_{sim}}(C_i, C_j) \Big],$$

To select the final keywords based on their ranking, an algorithm maximal marginal relevance (MMR) algorithm is employed to ensure that not only duplicated or semantically similar keywords are selected. This approach allows for the selection of words or items that are highly relevant to the document but are also distinct from each other. Here, the most relevant (argmax) or least relevant keywords are determined by adjusting λ based on the cosine similarity between the keyword candidates.

C. KeyBERT Score Evaluation

The KeyBERT architecture is shown in Figure 1. The cosine similarity is calculated to evaluate the similarity between the keywords extracted from the original text and the summary, which allows for a quantitative assessment of how closely related the summary is to the content of the original text.

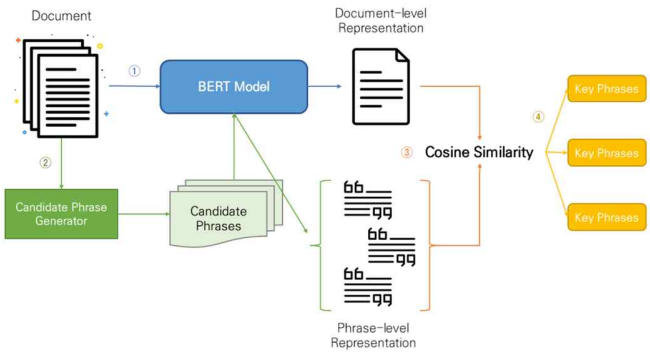

Fig. 1. KeyBERT architecture

The proposed structure involves fine-tuning a Korean pretrained KoBART model on two datasets to evaluate the model's summarization performance (Figure 2).

The fine-tuned models are saved and then loaded to perform summary inference on news text data. The inferred summary data and the original news text are then analyzed using a BERT model to measure vector cosine similarity, from which central keywords are extracted. After the extraction process, the final score (i.e., the KeyBERTScore) is calculated based on the similarity between these keywords.

The KeyBERTScore is distinctive in that it does not require human reference summaries of the original text, which sets it apart from other evaluation methods. However, this approach relies on the performance and understanding of the language model.

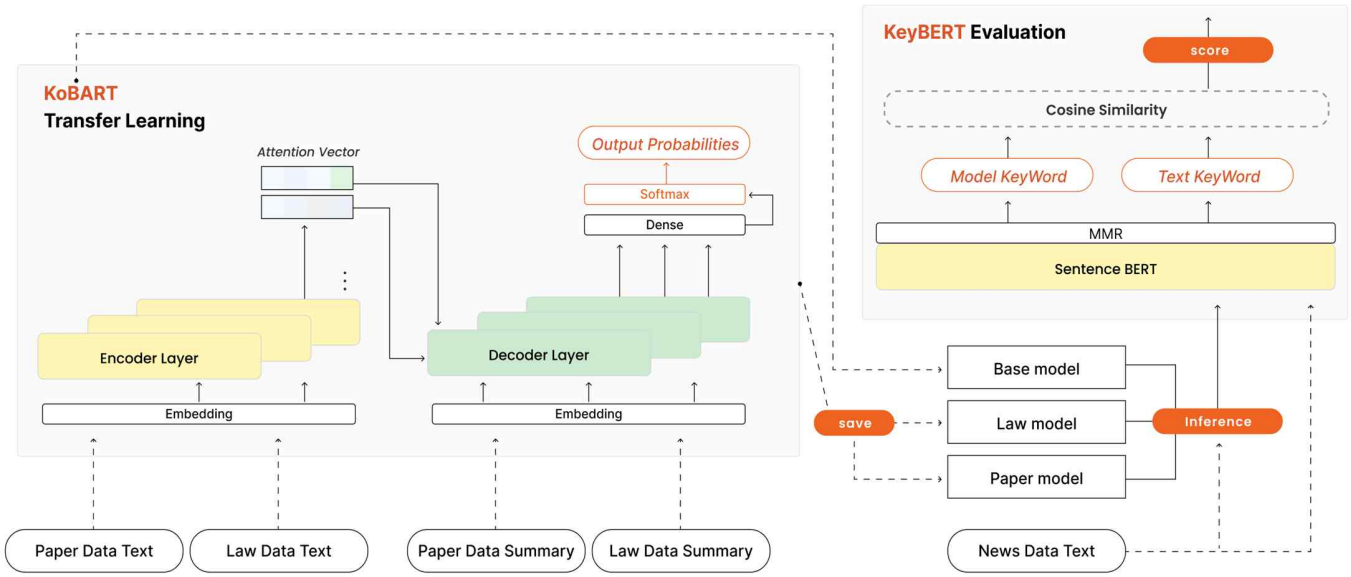

Fig. 2. Proposed structure using KoBART and KeyBERT

IV. Experiment

The pretrained model used was KoBart-summarization with Python version 3.9, a batch size of 8, and 10 epochs. Here, a single GTX 3,060 GPU with 12 GB of memory was utilized. Note that the training parameters were saved as logs and checkpoints at the end of each epoch.

During model inference, the length of the summary characters was set between 64 and 204 tokens. The evaluation metrics included BLEU, Rouge-1, Rouge-2, Rouge-L, and the BertScore. A total of 1,000 test data samples were used for inference.

TABLE I. ROUGE and BLEU scores of summaries predicted by KoBART

	BLEU and ROUGH Score		
	Base	***Law***	***Paper***
BLEU	0.69208550	0.64711780	0.65845206
Rouge-1 r	0.34305188	0.26454838	0.39081399
Rouge-1 p	0.25962700	0.17879586	0.26809042
Rouge-1 f	0.28741977	0.20651198	0.31069127
Rouge-2 r	0.17526117	0.13598923	0.20887659
Rouge-2 p	0.13384375	0.09411904	0.14831841
Rouge-2 f	0.14726886	0.10745742	0.16918689
Rouge-L r	0.30057269	0.22670833	0.36607101
Rouge-L p	0.22840808	0.15379100	0.25207688
Rouge-L f	0.25244125	0.17748785	0.29155951

Fig. 3. ROUGE and BLEU scores for summaries predicted by KoBART

As shown in Figure 3, the ROUGE scores and N-gram word occurrences suggest that the paper model (trained on academic data) outperforms the other models.

TABLE II. BertScore of summaries predicted by KoBART

	BertScore		
	Precision	***Recall***	***F1 Score***
Base model	0.7617	0.7777	0.7693
Law model	0.7494	0.7633	0.7366
Paper model	0.7556	0.7825	0.7685

Fig. 4. BertScore score of summaries predicted by KoBART

According to the BERTscore results in Figure 4, the baseline model appears to perform slightly better than the other models. Note that the evaluation dataset contains article summary datasets; thus, the baseline model, which is grounded in related fields, may be better at capturing the meanings of words.

TABLE III. KeyBertScore of summaries predicted by KoBART

	MSS	***MMR***
Base model	26.4	26.4
Paper model	30.6	26.4
Law model	22.6	14.6
Reference	14.8	10.2

Fig. 5. KeyBERTScore of summaries predicted by KoBART

Figure 5 shows the KeyBERTScore for the summaries predicted by KoBART. In Table III, the main sentence score (MSS) measures the similarity of the N-gram keywords extracted from the original text and the summary using cosine similarity. The MMR assesses the similarity in terms of the appearance of similar N-gram words between the original and summary texts, with a focus on reducing keyword duplication and increasing diversity.

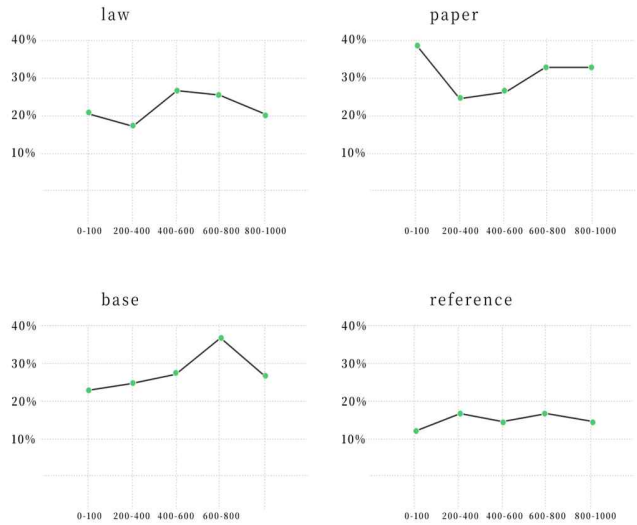

Fig. 6. KeyBERTScore graph for all models

Note that the reference summaries labeled by humans obtained low scores, which indicates that it was challenging for the model to extract the key keywords from the original text. The graph shown in Figure 6 represents an analysis conducted on randomly selected summaries: 10 from each of the five intervals out of a total of 1,000 model summaries. This graph shows the results of calculating the similarity between the keywords extracted from the summaries by the base and paper models using KeyBERT compared to the keywords of the original text.

The results shown in Figure 5 demonstrate that both the base and paper models maintained the quality of the summary within the 25% to 40% range. The law model (trained on the legal dataset) demonstrates relatively lower performance compared to the other transfer learning models. The paper and base models exhibit stable performance in both the quantitative evaluation (ROUGE score) and qualitative evaluation using the BERT language model. This suggests that the inference performance can vary depending on the characteristics of the dataset.

The original texts and the corresponding summaries inferred by each model are shown below.

TABLE IV. ACTUAL SUMMARIES PREDICTED BY KOBART

원본 텍스트
강원도 일대를 덮친 대형 산불을 계기로 소방공무원의 국가직 전환이 급물살을 타고 있다. 소방공무원의 국가직 전환으로 처우 개선뿐 아니라 소방인력과 장비의 지역 간 격차를 해소해야 한다는 여론도 높아지고 있다. 지난 5 일 '소방공무원을 국가직으로 전환해 주세요'란 제목의 청와대 국민 청원에는 22 만여 명이 동의했다. 문재인 대통령은 9 일 국무회의에서 소방공무원 국가직 전환 관련 법안을 신속히 처리할 것을 국회에 촉구했다. 문 대통령은 이날 "정치적 쟁점이 크게 있는 법안이 아닌 만큼 소방공무원 국가직 전환 관련 법안이 신속하게 처리돼 올해 7 월부터 차질 없이 시행될 수 있도록 국회의 협조를 요청한다"고 말했다. 소방공무원 국가직 전환은 2017 년 대선 당시 문 대통령이 내세운 공약 중 하나다. 지난해 말 국회 행정안전위원회에서 소방법 등 관련 법에 대한 논의가 이뤄졌지만 여야가 이견을 좁히지 못해 소위원회 문턱을 넘지 못했다. 현재 추진 중인 소방공무원 국가직 전환은 2022 년까지 소방공무원 2 만 명 충원을 위한 재원을 중앙정부가 부담하고, 대형 진화용 헬기 등 고가 소방장비 확보, 소방공무원 수당을 개선하기 위한 재원 확보 등을 골자로 한다. 소방·안전에 대한 국가 재정지원 강화를 통해 시·도별 부족 인력 격차를 줄이기 위해서다. 단 소방에 대한 시·도지사의 인사권과 지휘·통설권은 유지한다. 더불어민주당은 소방공무원의 국가직 전환을 위한 법안들을 4 월 국회 내 처리하겠다고 약속하고, 야당에 협조를 요청했다. 권미혁 민주당 의원은 이날 국회에서 열린 행정안전위원회 현안 보고에서 "소방관의 국가직 전환 관련 법이 통과될 기회가 있었는데 자유한국당의 원내지도부 지시로 의결 직전 무산됐다"며 "소방 서비스 향상과 신속한 재난 대응을 위해 소방기본법, 소방공무원법, 지방공무원법, 국가공무원정원법 등을 조속히 심사하길 부탁한다"고 당부했다. 야당은 소방공무원의 국가직 전환에 앞서 관계부처 간 조율 미흡을 보완할 필요가 있다고 반박했다. 이채익 한국당 의원은 "소방직의 국가직화를 반대하는 것은 아니다"며 "소방청과 재정당국 등 관계부처 간 이견 조율이 미흡해 업무를 어떻게 배분할지 논의되지 않은 부분이 있었다"고 지적했다.
Base model 추론 요약문
소방공 공무원의 국가직 전환으로 처우 개선뿐 아니라 소방인력과 장비의 지역 간 격차를 격차를 해소해야 한다는 여론도 높아지고 있는 소방공무원을 국가직으로 전환해 주세요'란 제목의 청와대 국민 청원에는 22 만여 명이 동의했다. 문재인 대통령은 9 일 국무회의에서 소방공무원 국가직 전환 관련 법안을 신속히 처리할 것을 국회에 촉구했다.
Paper model 추론 요약문
강원도 일대를 덮친 대형 산불을 계기로 소방공무원의 국가직 전환이 급물살을 타고 있다. 지난 5 일 '소방공무원을 국가직으로 전환해 주세요'란 제목의 청와대 국민 청원에는 22 만여 명이 동의했다. 문재인 대통령은 9 일 국무회의에서 소방공무원 국가직전환 관련 법안을 신속히 처리할 것을 국회에 촉구했다.
Law model 추론 요약문
소방공공무원의 국가직 전환으로 처우 개선뿐 아니라 소방인력과 장비의 지역 간 격차를 해소해야 한다는 여론도 높아지고 있는바, 소방공무원을 국가직으로 전환해 주세요'란 제목의 청와대 국민 청원에는 22 만여 명이 동의한바, 위 청원들을 신속히 처리할 것을 국회에 촉구한다.
Reference
문 대통령은 강원도의 대형 산불을 계기로 소방공무원 국가직 전환 관련 법안의 신속 처리로 7 월 시행 되길 국회에 요청했고 야당은 소방공무원 국가직 전환에 앞서 관계 부처간 조율하여 보완이 필요하다고 반박하였다.

Fig. 7. Actual summaries predicted by KoBART

TABLE V. ACTUAL SUMMARIES PREDICTED BY KOBART

	KeyWord	***Score***
Text	'국민 청원', '국가 전환 관련', '논의 부분 지적	1
Law model	'국가 전환 제목', '여론 소방 공무원', '국민 청원 동의'	0.31
Base model	'문재인 대통령 국무회의', '방공 공무원 국가', '국민 청원'	0.23
Reference	'신속 처리 시행', '대통령 강원도 대형', '공무원 국가'	0.06
Paper model	'국민 청원', '국가 전환 관련', '논의 부분 지적'	0.22

	KeyWord	***Score***
Text	'산청군 신청 약초', '추진 올해 대한민국', '자산 민간 자생'	1
Law model	'테마 기반 구축', '주제 농촌 불어', '농림축산식품부 추진 농촌'	0.05
Base model	'함양군 산청군', '항노화 사업', '산청군 한방약 복합'	0.13
Reference	'산업 일자리 자립', '조직 활용', '산청군 한방약 복합'	0.06
Paper model	'사업 농촌', '신청 전국', '산청군 한방약 복합'	0.14

	KeyWord	***Score***
text	'힐링 라이프 실현', '생활 인프라 안식처', '일과 균형 일상'	1
law-model	'공간 쾌적', '교육 교통 생활', '충북 여중 고등학교'	0.06
base-model	'일상 행복', '주거 환경', '추구 라이프'	0.15
reference	'자연 최첨단 시스템', '생활 교통 인프라', '일대 분양 주택'	0.12
paper-model	'일상 행복 추구', '생활 인프라 안식처', '선호 주거 환경'	0.28

Fig. 8. Detailed KeyBERT extracted keywords

As shown in Figure 7, analyzing the keywords extracted by KeyBERT reveals that the content of the original text can be predicted even with N-gram keywords. This indicates that an evaluation method using keyword extraction can be useful. In addition, cases where the main keywords of the original text were not identified were also observed.

V. Conclusion

This paper has explored the use of language models to evaluate text summarization quality compared to traditional methods, e.g., ROUGE and BLEU. We have demonstrated the potential for automating qualitative human evaluations of summary quality. Building on previous research, we proposed the KeyBERTScore, which is obtained by comparing the original text and the summary by extracting keywords, which is an essential element in capturing the theme of a summary. The proposed KeyBERTScore can be useful in numerically evaluating whether a summary contains significant content. However, given the limitations of language models, further research is required to explore text summarization evaluation methods that utilize language models from various perspectives.

Acknowledgment

This work was supported by Korea Research Institute for defense Technology Planning and advancement(KRIT) grant funded by the Korea government(DAPA(Defense Acquisition Program Administration)) (KRIT-CT-23-041, LiDAR/RADAR Supported Edge AI-based Highly Reliable IR/UV FSO/OCC Specialized Laboratory, 2024). This research was supported by the MIST(Ministry of Science, ICT), Korea, under the National Program for Excellence in SW), supervised by the IITP(Institute of Information & communications Technology Planning & Evaluation) in 2022"(2022-0-00964).

References

[1] Huang, Y., Feng, X., Feng, X., & Qin, B. (2023). The factual inconsistency problem in abstractive text summarization: A survey. arXiv:2104.14839v3.

[2] Liu, Y., Shi, K., He, K. S., Ye, L., Fabbri, A. R., Liu, P., Radev, D., & Cohan, A. (2023). On learning to summarize with large language models as references. arXiv:2305.00000v2.

[3] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In International Conference on Learning Representations (ICLR 2020).

[4] Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., & Jaggi, M. (2018). Simple unsupervised keyphrase extraction using sentence embeddings. arXiv preprint arXiv:1801.04470v3.

[5] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461. Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conference Magnetics Japan, p. 301, 1982].