International Journal of Advanced Culture Technology Vol.12 No.4 408-417 (2024)
DOI https://doi.org/10.17703/IJACT.2024.12.4.408

IJACT 24-12-48

Test Cases Generation and Transformer Language Models Performance Comparison for Korean Natural Language Processing-based AI SW

[1]In-Kyoun Lee, [2]Min-Seong Jin, [3]Gwang-Woon Lee, [4]Gun-Woo Park

[1]Lecturer, Korea Armed Forces Nursing Academy
[2]Freelancer, Kongju National University Graduate School of Education, Korea
[3]Lecturer, Korea Armed Forces Nursing Academy
[4]Professor, Department of Computer Engineering, DaeJeon University, Korea
[1]iklee1990@naver.com, [2]jinmin323@naver.com, [3]kh1touch@naver.com, [4]pgw4050@dju.kr

Abstract

Since the emergence of ChatGPT, transformer-based language models have become highly popular. This study utilizes a transformer-based approach to measure Korean sentence similarity for software testing. By doing so, we propose test cases using metamorphic relationships. The performance of the transformer models is then compared using similarity measures. First, we create a test set by transforming sentences from the Defense Daily according to specific rules. We then input these transformed sentences into the RoBERTa, Electra, and T5 models. We check whether the similarity measure between the original sentence and its variant satisfies the metamorphic relationship. The performance of each model is then compared using a similarity measure. In our experiments, the RoBERTa model satisfied metamorphic relations in MR5 and MR6, which involved transforming nouns and verbs into synonyms, and in MR7, which involved altering sentence order, with accuracy rates of 80%, 85%, and 88%, respectively. All tests passed except MR7 (77%) for the Electra model and MR1 (73%) for the T5 model. Finally, we compared the performance of each model. In the comparison, the Electra model outperformed the T5 model (99.96%) and the RoBERTa model (99.62%) with an accuracy of 99.97%.

Keywords: *Transformer Language Model, Korean Natural Language Processing, Metamorphic Testing, Test Case Generation*

1. INTRODUCTION

The Korean military is promoting the application of AI in all fields to foster an AI-based high-tech powerhouse. In addition, recently, AI has also been applied to healthcare, and relevant fields have been developing rapidly.[1] Examples include services such as those that analyze conversations with patients in real-time to create clinical records; those that detects urgent situations for the elderly living alone; and an AI speaker that helps people with severe disabilities live independently. This field of AI in the military and healthcare is expected to have great synergy in the future by combining disaster nursing and military nursing.

Manuscript received: October 17, 2024 / revised: November 21, 2024 / accepted: December 1, 2024
Corresponding Author: iklee1990@naver.com
Tel:+82-10-6253-2744
Lecturer, Korea Armed Forces Nursing Academy

healthcare is expected to have great synergy in the future by combining disaster nursing and military nursing. However, any error can cause unimaginable damage. This is why it is crucial to conduct thorough testing before deploying AI-based systems.

In this study, we apply the transformer language models RoBERTa (A Robustly Optimized BERT Pretraining Approach), Electra (Efficiently Learning an Envoder that Classifies Token Replacements Accurately), and T5 (Text-To-Text Transfer Transformer) to Korean sentence similarity measurement. After applying them, test cases are designed to find possible errors in AI by testing the AI-based software before it is directly used by users. After that, the performance of each model will be compared.

This paper is organized as follows. Chapter 2 introduces SW testing of AI-based systems, the Transformer language model, and existing research on Korean natural language processing. Chapter 3 suggests test cases for testing AI software. In Chapter 4, we will apply the proposed test cases using RoBERTa, Electra, T5 Korean sentence similarity measurement models, and a self-created test set compiled from Defense Daily newspaper articles. Later, we compare the performance of each model and conclude in Chapter 5.

2. RELATED RESEARCH

2.1 SW Testing of AI-based System

Traditional software testing predicts that when you input a value of A, it will output a value of B. It turned out that it produced the expected result and a failure if it produced an unexpected result.

However, in an AI-based system, inputting a value of A may result in various unpredictable outcomes, such as C or D. These potential errors propose a metamorphic testing method. Metamorphic testing is a method of testing an AI-based system whose output value is difficult to predict by defining a metamorphic relationship between the execution result of an initial input value A and the execution result of some variant of A' as a metamorphic relationship.[2-4]

For instance, let's use the math property sin(x) = sin(π - x). First, set the equation: sin(7) = sin(π - 7). In this case, the exact values of sin(7) and sin(π - 7) are unknown. However, if sin(7) has a value of 0.6569, then by the math above, sin(π - 7) should also have a value of 0.6569. Based on this, we define sin(7) = sin(π - 7) as a metamorphic relation. Define the first input value, sin(7), as the Source Test Case and the second input value, sin(π - 7), as the Follow-Up Test Case.

In this way, if the result value of the Source Test Case and the result value of the Follow-Up Test Case are not the same, the metamorphic relationship is not satisfied, and the test is defined as a Test Fail.

Table 1 shows the application of the method described above to a transformer-based Korean sentence similarity measurement model. You can see that the text similarity before and after transforming the sentence order in the text is not 100%, indicating that changing the sentence order alone affects the similarity.

Table 1. Performance Comparison of Korean Sentence Similarity Measurement Models with Test Case

Input	President Yoon Seok-yul on Friday announced his intention to build the world's largest new 'advanced system semiconductor cluster' in the Seoul metropolitan area based on a massive private investment of 300 trillion won. While presiding over the '14th Emergency Economic and Civil Affairs Conference' held at the

<table>
<tr><td></td><td colspan="2">Cheong Wa
Dae State Guesthouse on the same day, President Yoon announced the establishment of the semiconductor cluster, saying, “We must ensure that private investments totaling more than 550 trillion won in six high-tech industries, including semiconductors, planned by 2026, are promptly realized, and the government must provide location, R&D, manpower, and tax support.” -emphasis added.
At the meeting, 14 candidate sites for national advanced industrial complexes were announced. President Yoon Seok-yeol announced on the 15th that he will build the world's largest new 'advanced system semiconductor cluster' in the Seoul metropolitan area based on a large-scale private investment of 300 trillion won.
President Yoon presided over the '14th Emergency Economic and Civil Affairs Conference' held at the Cheong Wa Dae State Guesthouse on the same day, saying, “We must ensure that private investments totaling more than 550 trillion won in six high-tech industries, including semiconductors, planned by 2026 are quickly realized, and the government must provide location, R&D, human resources, and tax support.”
-emphasis added.
At the meeting, 14 candidate sites for national advanced industrial complexes were announced. <b>Daejeon (nano-semiconductor, aerospace), Cheonan (future mobility, semiconductor), Cheongju (railroad), Hongseong (hydrogen, future vehicles, secondary batteries), Gwangju (core components for future vehicles), Goheung (space launch vehicle), and Iksan (food tech) in Honam, Wanju (hydrogen storage and utilization manufacturing), Changwon (defense, nuclear power), Daegu (future cars and robots), Andong (biomedicine), Gyeongju (small module nuclear power plant), Uljin (hydrogen utilized from nuclear power plants), Gangneung (natural product bio), Gangwon-do (natural product bio), etc. (→ Move to the second sentence)</b>
President Yoon said that “speed is the key,” and that the government will make every effort to ensure that the national high-tech industrial complexes announced today will be promoted quickly.</td></tr>
<tr><td>RoBERTa</td><td>Electra</td><td>T5</td></tr>
<tr><td>97.89%</td><td>99.5%</td><td>99.95%</td></tr>
</table>

A related study applies metamorphic testing to a Chinese sentence similarity measurement model.[5] In this study, we applied metamorphic test cases to sentence similarity, sentence summarization, and sentence classification. Their experiments confirm that the metamorphic testing method is suitable for processing Chinese natural language.

2.2 Transformer Language Model

In 2017, Google announced Transformer, along with the concept of Transfer Learning, a way to shorten learning time and maximize efficiency by applying prior learning to new areas.[6] In 2018, Google released the BERT (Bidirectional Encoder Representations from Transformers) language model, which leverages Transformers to learn both the preceding and following information of each text to better understand the meaning of the text, enabling a variety of natural language processing such as answering questions, summarizing sentences, and measuring similarity between sentences.[7]

RoBERTa and Electra are language models that utilize BERT, while T5 is a Transformer-based language model. RoBERTa is a language model released by Facebook in July 2019 that proposes a way to optimally learn BERT.[8] Unlike BERT, RoBERTa uses a Masked Language Model (a model that masks and predicts randomly occurring words rather than sequentially occurring words in each sentence) to enhance the model's capabilities.

Unlike BERT and RoBERTa, which used the Masked Language Model as the language model announced by Google in May 2020, Electra uses the Replace Token Detection method (where the Generator transforms the real input token into a fake token, and the Discriminator guesses whether the fake token is original or replaced).[9] T5 is a language model announced by Google in October 2020 that uses a text-to-text translation method and has shown good performance in areas such as translation, sentence accuracy assessment, sentence similarity measurements, summarization.[10]

2.3 Existing Korean Natural Language Processing Research

Existing studies applying metamorphic testing to Korean natural language processing include summarization and translation.[2-4] Summarization is challenging because it is relatively subjective and can yield different results depending on who is doing it.[11] Translation, while much improved, still faces many challenges before it can be perfected, such as changes in the result caused by altering a single word.

In previous studies combining Korean summarization and Korean-English translation with the concept of metamorphic testing, test cases that transform nouns and verbs into synonyms commonly satisfied the metamorphic relationship. However, test cases that varied sentence order and punctuation often did not satisfy the metamorphic relationship. Through this, we argue that the test cases selected in existing research are effective for the software testing of AI-based systems.

3. Test Case Generation

In Chapter 3, we propose to generate test cases for Korean sentence similarity using the Transformer language models RoBERTa, Electra, and T5. An example of each test case is shown below.

MR1 : Replace name or noun in newspaper text

Example) Minister of National Defense Lee Jong-seop met with entry-level executives to hear about their difficulties in service and conveyed his willingness to actively work to improve service conditions. The meeting was attended by officials from the Ministry of Defense, including Kim Sung-joon (→ Kim Sung-jin / MR1 application), head of the Personnel and Welfare Division, and other key military officials, including the Chief of Personnel Staff of the Army, Navy, and Air Force Headquarters and the Chief of Staff of the Marine Corps, as well as more than 60 junior officers who are serving in the field. -emphasis added. In addition, the ministry said it plans to consult with relevant ministries to remodel old cadre dormitories and improve single-occupancy rooms so that entry-level cadres can play a leading role in fostering a combat-type strong military.

MR2 : Replace country name in newspaper text

Example) Deepening and Expanding Defense Cooperation and the Leap Forward of the U.S.-ROK

Alliance introduced the achievements and directions of the U.S.-ROK military alliance, which is evolving into a global comprehensive strategic alliance, including deepening cooperation, expanding defense exchanges, and international peacekeeping operations. -Medium-. It also covered the direction of defense exchanges and cooperation with neighboring countries such as China (→ Thailand / MR2 application) and Russia, as well as with partner countries in Southeast Asia, Oceania, Southwest Asia, Central Asia, the Middle East, and Europe.

MR3 : Replace occupation name in newspaper text

Example) The center specifically supports the establishment and management of the Ministry of National Defense's 'Defense Data Construction Roadmap'. In addition, it consults on defense data construction projects of the Ministry of National Defense, each military, and related organizations, and provides technical support for defense data utilization. “In the mid- to long-term, we will secure and train data engineers (→ programmers / MR3 applications) to develop into a dedicated organization that supports data analysis and artificial intelligence (AI) development for scientific decision-making of our military at all stages from defense data collection to disposal.” Yoon's mission is driven by the power of data. -emphasis added. “I will lead the center to become the premier defense data organization and earn the trust of our military.”

MR4 : Replace punctuations in newspaper text

Example) The Korean Language Proficiency Test, one of the requirements for applying for the executive position, will be changed to a point system. Previously, candidates had to pass a written Korean history examination with a certain score or substitute a test score to pass the first round of screening. -emphasis added. An Air Force official explained, “(” → deleted / MR4 application)The entire Air Force, including the department in charge, is considering how to improve the cadre selection system to foster excellent talents in a fair and objective way.“ (” → deleted / MR4 application).

MR5 : Replace noun with synonym in newspaper text

Example) The 32nd Division held an intense “Joint Military-Police Counterterrorism Comprehensive Exercise” at the Namsejong Reserve Military Training Center in Geumnam-myeon, Sejong-si, Sejong-do, on March 24 to eliminate the possibility of terrorism in the region. The drill focused on eliminating the possibility of terrorism in the region as North Korea's recent provocations continue(→persist / MR5 application). -emphasis added-. “The military and police, who are at the forefront of regional defense to protect the safety and lives of the people, trained together to establish an integrated defense and combat readiness,” said Lt. Col. Yoon Sang-soon, commander of the riot police.

MR6 : Replace verb with synonym in newspaper text

Example) The Marine Corps 6th Brigade successfully concluded a comprehensive island defense exercise aimed at “establishing a decisive posture for the absolute defense of the Northwest Territories. -emphasis added. “It was a great opportunity to strengthen our combat capabilities and readiness necessary to establish a decisive posture,” said Moon Tae-sung (Capt.), 9th Company commander, ”and I will contribute (→ serve / MR6 application) to creating a unit that perfectly executes its assigned mission even in complex situations with the enemy.”

MR7 : Change the order of sentence in newspaper text

Example) The Ministry of National Defense (MND) announced on Friday that it has revised the 'Life Cycle Management Work Order,' which mainly stipulates the role of each agency in managing parts to ensure stable operation and utilization of weapon systems and cost analysis and management for each stage of the life cycle. The Life Cycle Management Work Order was enacted in February 2021 to provide basic procedures and guidelines for the life cycle management of weapon systems and power support systems in accordance with the Munitions Management Act, the Defense Business Act, and the Act on the Development and Support of the Defense Industry. It was partially revised in May last year to strengthen follow-up military support. The revision established basic policies, procedures, and duties for each organization by establishing weapon system parts management and revising life cycle cost analysis. First, the revision stipulated all aspects of parts management to be linked to 'identification of improvement needs, commercialization, and application of improved parts' in consideration of reliability, end-of-life management, and localization. To ensure weapon system operability, the principles, roles, and work procedures for parts management in the acquisition and operation and maintenance phases were also specified.(→ Move to the fourth sentence / MR7 application) The basis for the establishment and operation of the parts management portal was established, and plans and procedures for commercialization of identified improvement needs were also established. -omitted-. In particular, for major weapon systems, the ROKDA will conduct an in-depth review of O&M costs to improve the affordability of acquisition alternatives and the rationality of financial planning.

4. EXPERIMENT AND ANALYSIS

4.1 Experiment

The experiment was conducted using RoBERTa, Electra, and T5 based Korean sentence similarity measurement pretrained models. The test set consists of 700 original data and 700 variant data generated by applying each test case, using articles from the Defense Daily newspaper.

Figure 1 illustrates the metamorphic testing process in IDEF0 (Icam DEFinition for Function Modeling) notation.

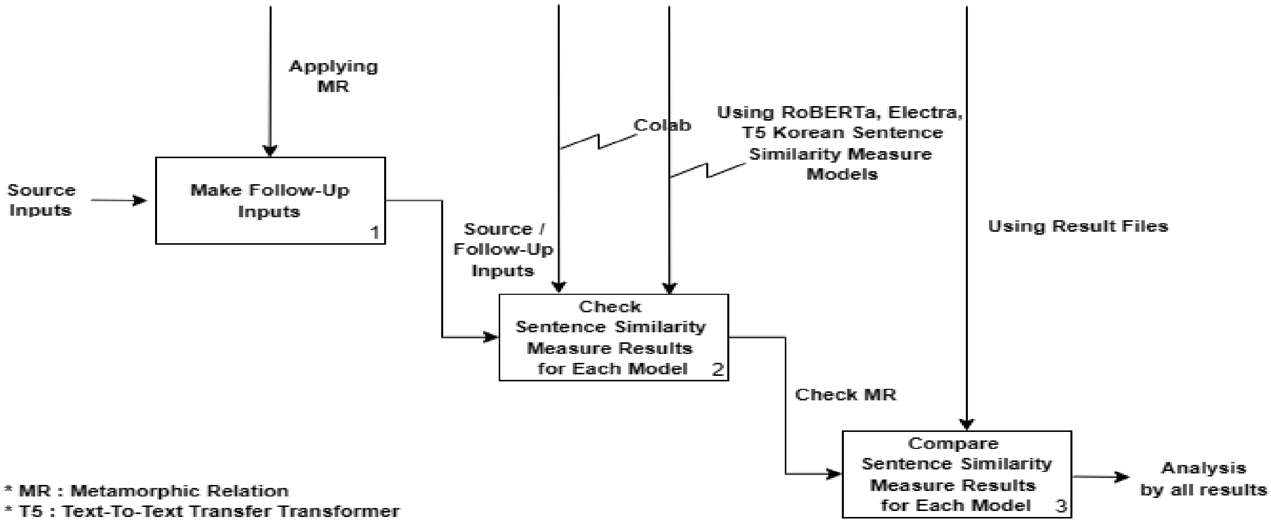
Applying MR
Colab
Using RoBERTa, Electra, T5 Korean Sentence Similarity Measure Models
Source Inputs
Make Follow-Up Inputs
1
Source / Follow-Up Inputs
Using Result Files
Check Sentence Similarity Measure Results for Each Model
2
Check MR
Compare Sentence Similarity Measure Results for Each Model
3
Analysis by all results
* MR : Metamorphic Relation
* T5 : Text-To-Text Transfer Transformer

Figure 1. The metamorphic testing process

First, we apply each test case to the first input, Source Inputs, to transform it. Follow-Up Inputs, where the data generated by the transformation becomes the second input. Enter each of the configured inputs (Source Inputs, Follow-Up Inputs) into the RoBERTa, Electra, and T5 Korean sentence similarity models using Google Colab. Check the output to determine if the sentence similarity measure for each model satisfies the metamorphic relationship or if the test fails. The performance of each model is then compared to draw conclusions.

4.2 Result Analysis / Model Performance Comparison

Korean Sentence Similarity Measurement based on RoBERTa, Electra, and T5 We measured the sentence similarity between 100 original data and 100 modified data for each test case (MR1 ~ MR7) in the pretrained model. After the measurement, average the similarity measure results for each test case in each model. Only if the metamorphic relationship was above average was the test considered successful.

Figure 2 shows the results of the RoBERTa model, with 73% of the tests failing for MR1, which modifies names or nouns in the text of newspaper articles, 76% for MR2 and MR3, which modify country names and occupations, and 73% for MR4, which modifies punctuation in the text of newspaper articles. In contrast, MR5 through MR7 tested successfully with an average of 84%.

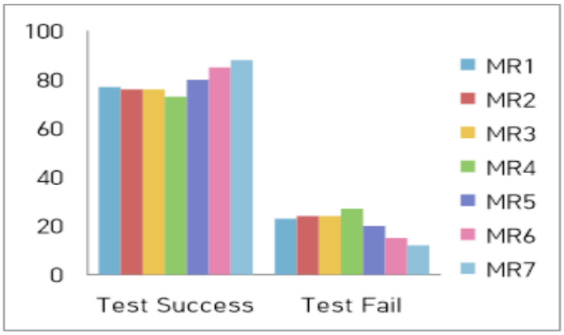

Figure 2. Experimental result with RoBERTa Model

Figure 3 shows that MR7, which transforms the order of sentences in the text of a newspaper article, failed 77% of the time when tested with the Electra model. In contrast, MR1 through MR6 were tested successfully with an average of 86%. In particular, MR5 and MR6, which transform the corresponding nouns/verbs in the newspaper text into synonymous nouns/verbs, often satisfied the metamorphic relationship 94% and 92% of the time, respectively.

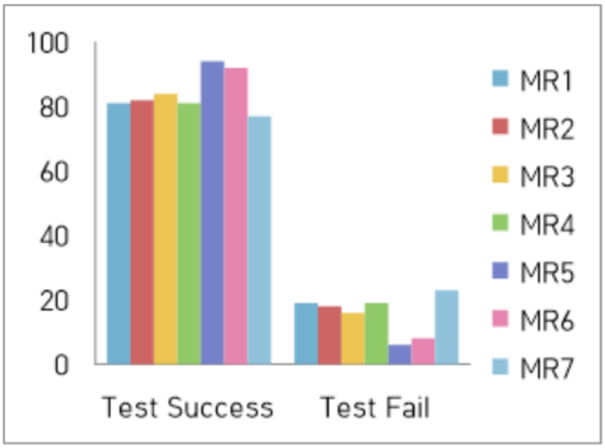

Figure 3. Experimental result with Electra Model

Figure 4 shows the results of the T5 model, where MR1, which transforms names or nouns in the text of newspaper articles, failed the test with a 73% success rate. In contrast, MR2 through MR7 tested successfully with an average success rate of 87%. In particular, MR5 and MR6 passed the test with high probabilities of 92% and 94%, respectively, which is similar to the inter-experimental results with the Electra model.

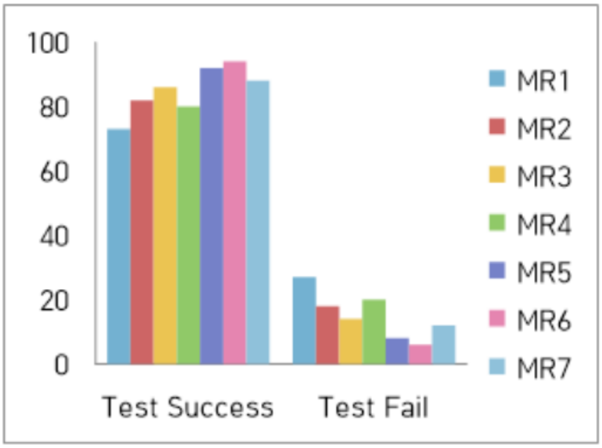

Figure 4. Experimental result with T5

We then compared the performance of each model by averaging all the sentence similarity results for each model. Figure 5 shows the sentence similarity measurements for each model.

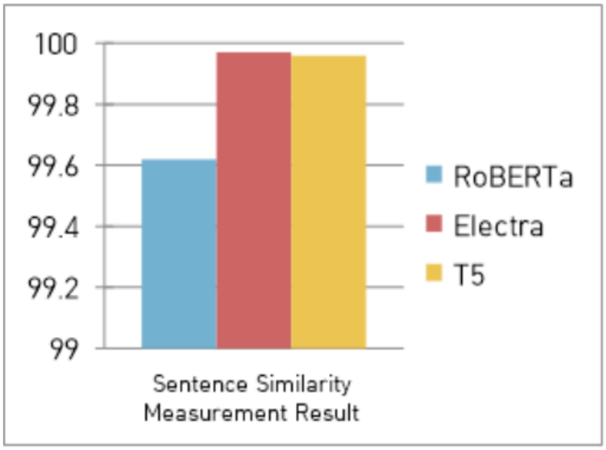

Figure 5. Sentence Similarity Measurement Result for Each Models

We see that the Electra model is the best of the three models, outperforming the RoBERTa model by 0.35% and the T5 model by 0.01%.

5. CONCLUSION

Prior to this study, the researcher developed metamorphic test cases for Transformer-based Korean summarization and Korean-English translation. We applied these test cases, presented at the 2024 IPACT Conference, to measure Korean sentence similarity[12]. Building on these results, we employed a metamorphic testing method to evaluate Korean sentence similarity using pre-trained Transformer models, specifically RoBERTa, Electra, and T5.

The findings indicate that MR5 and MR6, which transform nouns/verbs in newspaper text into synonyms, successfully satisfied the metamorphic relationship with an average accuracy of 90% in Korean sentence similarity, mirroring the results in Korean summarization and Korean-English translation. Conversely, the RoBERTa model showed that MR1 through MR4 often failed to meet the metamorphic criteria, averaging 76%. The Electra model demonstrated lower performance, with MR7 satisfying the criteria 77% of the time and MR1 meeting the criteria 73% of the time, similar to the T5 model.

This study confirms that the proposed testing method is an effective evaluation tool for software testing of AI-based systems utilizing Korean natural language processing. The method holds significant potential for future applications, such as testing and evaluating AI systems used in military voice recognition or medical AI.

REFERENCES

[1] Sungyeon Yoon et al., “Trends in Deep Learning- based Medical Optical Character Recognition”, The Journal of the Convergence on Culture Technology (JCCT), Vol.10, No. 2, pp.453-458, 2024.

[2] Inkyoun Lee, Dongsu Kang, “Generating Metamorphic Test Cases for Transformer-based Korean Summary”, KIISE Transactions on Computing Practives, Vol. 29, No. 11, pp.509-517, 2023.

[3] Inkyoun Lee, Dongsu Kang, “Generating Metamorphic Test Cases using Transformer Language Model”, National Defense Conference for Graduate Students organized by the Korea National Defense University, pp.521-533, 2023.

[4] Inkyoun Lee, “Generating Metamorphic Test Cases for Transformer-based Korean Natural Language Processing”, master's thesis at the Korea National Defense University, pp.1-51, 2024.

[5] Lingzi Jin, Zuohua Ding, Huihui Zhou, “Evaluation of Chinese Natural Language Processing System Based on Metamorphic Testing”, MDPI(Multidisciplinary Digital Publishing Institute) Mathematics, 2022.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, “Attention is all you need”, Proceedings of the 31st International Conference on Neural Information Processing Systems, pp.6000-6010, 2017.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, In North American Association for Computational Linguistics (NAACL), 2019.

[8] Y.Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv preprint arXiv:1907.11692, 2019.

[9] Clark, K., Luong, M. T., Le, Q. V. and Manning, C. D.,“ELECTRA: pre-training text encoders as discriminators rather than generators, Proc. of 8th International Conference on Learning Representations, pp.1-14, 2020.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqu Zhou, Wei Li, Peter J. Liu, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”, Journal of Machine Learning Research, 21, pp.1-67, 2020.

[11] Sungyeon Yoon, Minseo Park, “Media-based Analysis of Gasoline Inventory with Korean Text Summarization”, The Journal of the Convergence on Culture Technology (JCCT), Vol.9, No. 5, pp.509-515, 2023.

[12] Inkyoun Lee, Gwangun Lee, Gunwoo Park, “Test Cases Generation and Model Performance Comparison for Transformer-based Korean Sentence Similarity”, 2024 IPACT CONFERENCE, Vol.8, No. 1, pp.9-14, 2024.