ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 29, No. 11, pp. 509-517, 2023. 11
https://doi.org/10.5626/KTCP.2023.29.11.509

트랜스포머 기반 한국어 요약을 위한 메타모픽 테스트 케이스 생성

(Generating Metamorphic Test Cases for Transformer-based Korean Summary)

이 인 균 † 강 동 수 ††
(Inkyoun Lee) (Dongsu Kang)

요 약 최근 ChatGPT와 같은 AI 기반 소프트웨어가 인기이다. 소프트웨어의 품질 보증을 위한 테스팅에 대한 관심도 높아지고 있다. 본 연구는 소프트웨어 테스팅을 위한 트랜스포머 기반 한국어 요약을 위해 메타모픽 관계를 이용한 테스트 케이스를 제안한다. 먼저, 국방일보를 활용한 테스트셋을 일정한 규칙을 이용하여 변형시킨 후 T5 모델에 입력하고, 그에 따른 출력 결과와 기존의 출력 결과가 메타모픽 관계를 만족하는지 확인하고, 문서요약 성능지표인 Rouge와 Rdass를 이용하여 모델의 성능을 평가한다. 실험 결과 이름 혹은 명사를 변형하는 MR_1과 해당 명사/동사를 동의어로 변형하는 MR_5와 MR_6이 82%로 메타모픽 관계를 만족하였고, T5 요약 성능은 기존 연구와 비교한 결과 타 모델에 비해 13% 향상되었다. 이후, 기존 연구에서 다루지 않았던 Rouge-u, Rouge-su, Rdass 지표를 활용하여, 한국어 요약을 평가하는 성능지표의 종류를 확장하였다.

키워드: 한국어 자연어 처리, T5, 메타모픽 테스팅, 테스트 케이스 생성, 문서요약 성능지표

Abstract Recently, AI-based software such as ChatGPT has become popular. Consequently, interest in quality assurance testing of software is increasing. This study proposes a test case using a metamorphic relationship for a transformer-based Korean summary for software testing. First, the test set using the Defense Daily is transformed using certain rules and then entered into the T5 model. After inputting the transformed test set, we checked whether the output result according to the input and the existing output result satisfy the metamorphic relationship. We then evaluated the performance of the model using the document summary performance metrics Rouge and Rdass. The experimental results showed that MR_1, which transforms names or nouns, and MR_5 and MR_6, which transform nouns/verbs into synonyms, satisfy the metamorphic relationship with 82%. In addition, the summarization performance of the T5 improved by 13% compared to the models in the previous study. After that, We used Rouge-u, Rouge-su, and Rdass scores. These are the scores that were not covered in the previous studies. Through these scores, the types of performance scores that evaluate the Korean summaries were expanded.

Keywords: Korean natural language processing, T5, metamorphic testing, test case generation, document summary performance metrics

† 학생회원 : 국방대학교 컴퓨터공학 학생
iklee1990@naver.com
†† 종신회원 : 국방대학교 컴퓨터공학/사이버전 교수
(Korea Nat'l Defense Univ.)
greatkoko@hotmail.com
(Corresponding author임)

논문접수 : 2023년 6월 7일
(Received 7 June 2023)
논문수정 : 2023년 8월 28일
(Revised 28 August 2023)
심사완료 : 2023년 9월 3일
(Accepted 3 September 2023)

정보과학회 컴퓨팅의 실제 논문지 제29권 제11호(2023. 11)

1. 서 론

AI 기술은 급속하게 발전하고 있으며, 다양한 AI가 우리의 삶과 관련된 모든 분야에 도입되어 사람들이 더 편리하게 생활할 수 있는 사회적 기반이 되고 있다. 그러나 AI가 오류를 일으키게 되면 우리의 생활이나 재산, 나아가 생명에 이르기까지 손해를 끼치는 경우도 발생할 수 있다. 이로 인해 리스크를 줄여 오류가 최소화된 AI를 사용하고자 하는 수요도 증가하고 있다.

본 연구는 최근 관심이 높아진 ChatGPT(Chat Generative Pretrained Transformer)와 같은 트랜스포머 기반 언어 모델을 활용하여 자연어 처리 분야 중 필요성이 높고, 활용 범위 또한 넓은 문서 요약 분야에 한국어를 접목시켜 AI 소프트웨어를 실제 사용하기 전에 테스팅하여 AI의 오류 발생 가능성을 찾아내는 방법을 제안한다.

문서 요약은 주관성의 개입 가능성과 사람이 직접 수행하더라도 수행하는 사람에 따라 다른 결과가 나올 수 있는 어렵고, 시간이 많이 소모되는 작업으로 특히 요약 분야는 자연어 처리 분야 중 처리하기 가장 어려운 작업으로 알려져 있다[1]. 본 논문에서는 뛰어난 성능을 보여주고 있는 T5(Text-To-Text Transfer Transformer)를 활용하여 한국어 요약을 위한 메타모픽 테스트 케이스를 생성하고, 사전훈련 언어 모델을 통해 생성된 요약문과 사람이 미리 작성해 놓은 요약문을 비교하여 한국어 요약 모델을 평가한다.

현재 국·내외에서 연구중인 메타모픽 테스팅 관련 기존 연구는 이미지 분류 모델을 활용한 연구가 대부분이다. 또한 자연어 처리에 메타모픽 테스팅을 적용한 기존 연구의 언어는 고립어에 속하는 영어와 중국어이다. 본 논문은 위와 같은 기존의 방법들을 유사 응용한 것이 아닌 모국어로서 교착어에 속하는 한국어를 대상으로 언어모델 기반의 인공지능 소프트웨어를 테스팅 하는 방법을 새롭게 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 인공지능 SW 테스팅, 딥러닝 기반 인공지능 언어모델, 기존 관련 연구에 대해 소개하고, 3장에서는 메타모픽 관계를 이용한 테스트 케이스 생성을 제안한다. 4장에서는 T5를 이용한 한국어 요약과 국방일보 신문기사를 편집한 자체 제작 텍스트 테스트셋을 이용하여 생성한 실제 모델에 대해 제안한 테스트 케이스를 적용한다. 이후 문서 요약 성능지표를 통해 한국어 요약의 성능을 평가하고, 5장에서 결론을 맺는다.

2. 관련 연구

2.1 인공지능 SW 테스팅

기존의 소프트웨어 테스팅은 테스팅을 위한 입력 데이터를 테스트 대상 소프트웨어에 입력하여 테스팅을 수행하였다. 테스팅 수행을 통해 기대했던 바와 동일한 결과가 출력되면 성공, 기대했던 바와 다른 데이터가 출력되면 실패라고 판정한다.

하지만, 인공지능 기반 시스템의 테스팅 방법은 다양한 예측 불가한 결과 때문에 메타모픽 테스팅 방법이 제안되었다. 메타모픽 테스팅(Metamorphic Testing)은 개별 테스트의 예상 결과를 알 수 없는 인공지능 시스템의 경우에 복수의 실행 결과 간의 관계를 메타모픽 관계(MR, Metamorphic Relation)로 정의하고 테스팅하는 방법이다[2,3].

메타모픽 관계를 언어 모델에 적용하면 하나의 단어만 바꾸어도 전체 요약에 영향을 주게 된다. 예를 들어 표 1은 한국어 텍스트를 요약해 주는 사이트(https://smodin.io/ko)로 콩쥐를 팥쥐로 이름 하나만 변경하여 실험한 결과 출력값이 달라지게 된다. 즉, 메타모픽 테스팅은 입력/출력값 사이에 존재하는 메타모픽 관계를 이용하여 AI 기반 소프트웨어를 테스팅 하게 된다.

표 1 언어 모델에서의 메타모픽 관계 예시
Table 1 Example of metamorphic relation in a language model

Input	콩쥐가 일찍 모친을 여의고 계모를 얻었는데 계모에게는 팥쥐라는 딸이 있었다. 계모는 **콩쥐(→ 팥쥐)**에게만 힘든 집안일을 다 시키니 콩쥐의 고생이 이만저만 아니었다. 하루는 팥쥐 모녀가 나라의 잔치에 가면서 강피를 찌어놓고, 밑 빠진 독에 물을 채워놓으라고 하였다. 콩쥐가 독 앞에서 울고 있으니 두꺼비가 나와 깨진 독을 등으로 막아 물을 채울 수 있게 해주었고, 새들이 날아와 강피를 쪼아 찌어주었다. -중략-. 하루는 구슬에서 콩쥐의 혼이 나와 세자에게 시신의 위치를 아뢰니, 세자가 시신을 찾아 구슬로 바르니 콩쥐가 다시 회생하게 되었다. 결국 세자는 콩쥐와 행복하게 살게 되었고 팥쥐와 계모에게 큰 벌을 내렸다.
Output (콩쥐)	콩쥐가 일찍 모친을 여의고 계모를 얻었는데 계모에게는 팥쥐라는 딸이 있었다. 계모는 콩쥐에게만 힘든 집안일을 다 시키니 콩쥐의 고생이 이만저만 아니었다. **하루는 콩쥐 모녀가 나라의 잔치에 가면서 강피를 찌어놓고, 밑 빠진 독에 물을 채워놓으라고 하였다.**
Output (팥쥐)	콩쥐가 일찍 모친을 여의고 계모를 얻었는데 계모에게는 팥쥐라는 딸이 있었다. 계모는 콩쥐에게만 힘든 집안일을 다 시키니 콩쥐의 고생이 이만저만 아니었다. **하루는 구슬에서 콩쥐의 혼이 나와 세자에게 시신의 위치를 아뢰니, 세자가 시신을 찾아 구슬로 바르니 콩쥐가 다시 회생하게 되었다.**

인공지능 언어 모델에 메타모픽 테스팅을 적용한 연구는 중국어 자연어 처리 시스템 평가[4]가 있다. 문자

의 수가 적은 영어와 달리 중국어는 많은 수의 한자로 구성되어 있다. 이러한 이유로 영어를 활용한 평가 방법인 문자 삭제, 문자 교환, 시제, 대·소문자 및 단·복수 활용과 같은 방법을 중국어에는 적용할 수 없다. 또한 중국어는 아주 작은 변화에도 의미에 큰 변화를 가져올 수 있어 문법 구조 또한 영어와 상당한 차이점을 가지고 있다. 이를 바탕으로 해당 연구에서는 3가지 태스크로 구분하여 text similarity, text summarization, text classification에 메타모픽 관계를 활용한 테스트 케이스를 적용하여 실험을 진행하였다. 실험결과 메타모픽 테스팅 방법은 중국어 자연어 처리의 성능을 평가하는데 적합한 평가 방법임을 주장하고 있다.

2.2 딥러닝 기반 인공지능 언어모델

딥러닝 인공지능 언어모델은 활발한 연구를 통해 현재까지도 계속해서 발전되고 있다. 인공지능 언어모델은 언어이해 모델, 언어생성 모델, 언어 이해 및 생성 모델

표 2 딥러닝 기반 인공지능 언어 모델 유형별 특성
Table 2 Characteristics of deep learning-based artificial intelligence language model types

Sort	Language Understanding Model	Language Generation Model	Language Understanding and Generation Model
Concept	A model that pre-learns the context of words from large amounts of data to understand the grammar and meaning of words contained in input sentences	A model that predicts the next best word for a given word column by pre- learning large amounts of data	A model that uses a language comprehension model and a language generation model to generate output sentences by understanding input sentences
Example Model	Bi-LSTM (1997, Schuster & Paliwal)	GPT-1 (2018, OpenAI)	MarianMT (2016, University of Edinburgh)
	BERT (2018, Google)	GPT-2 (2019, OpenAI)	CTRL (2019, OpenAI)
	RoBERTa (2019, Facebook)	BART (2019, Facebook)	T5 (2020, Google)
	ALBERT (2019, Google)	GPT-3 (2020, OpenAI)	Megatron (2021, MS / nVidia)
	SpanBERT (2019, Allen Institute for AI & University of Washington)	KoGPT (2021, KAKAO)	HyperClova (2021, Naver)
	Bard (2023, Google)	GPT-3.5 (2022, OpenAI)	
		BLOOM (2022, BigScience)	
		GPT-4 (2023, OpenAI)	

표 3 T5의 텍스트-투-텍스트 프레임워크
Table 3 Text-to-text framework in T5

Sort	Input	Output
Summary	“요약: 콩쥐가 일찍 모친을 여의고 계모를 얻었는데 계모에게는 팥쥐라는 딸이 있었다. 계모는 콩쥐에게만 힘든 집안일을 다 시키니 콩쥐의 고생이 이만저만 아니었다. 하루는 팥쥐 모녀가 나라의 잔치에 가면서...”	“세자는 콩쥐와 행복하게 살게 되었고 팥쥐와 계모에게 큰 벌을 내렸다.”
Measure similarity between sentences	“문장 간 유사도 측정 : 나는 너를 좋아한다. / 나는 너가 마음에 든다.”	“4.5, 의미 유사”
Accuracy assesment of sentences	“문장의 정확성 평가: 이 논문은 날아다니는 것을 좋아한다.”	“말이 되지 않는 문장”
Translation	“영어를 한국어로 번역: This is good.”	“이것은 좋다.”

로 구분되며 표 2와 같은 특징을 가지고 있다.

인공지능 언어모델은 성과가 가시적으로 나타나는 분야로 2003년에 뉴럴 언어모델(Neural Language Models)이 제안되었고, 2013년에는 단순한 언어모델인 워드 임베딩(Word Embeddings) 모델 Word2Vec가 제안되었다. 이후 전이 학습(Transfer Learning) 개념과 함께 2017년에 구글의 트랜스포머가 발표[5]되고, GPT 시리즈가 차례대로 등장하면서 사전 학습 기반의 딥러닝 인공지능 언어모델(Pretrained Models)이 발전하고 있다. 그중 T5는 구글 AI 연구팀이 개발한 트랜스포머 기반 언어 모델의 하나로 정답 추출과 문서 요약 등의 여러 태스크들을 통합한 모델이다. T5는 각 태스크를 Text-to-Text 문제로 치환하는 방법으로 나타내어 언어 이해, 언어 생성 태스크 등에서 좋은 성능을 보여주고 있다. 표 3은 구글에서 쓴 논문의 T5 텍스트-투-텍스트 프레임 워크[6]를 한국어 자연어 처리에 적용한 것으로 T5는 요약, 유사도 측정, 정확성 평가, 번역 등의 작업을 수행한다.

언어모델을 지식 저장 도구로 활용하는 추세에 맞추어 T5는 위와 같이 다양한 자연어 처리 작업에 활용될 수 있기 때문에 한가지 모델로 여러 작업을 수행할 수 있다는 측면에서 다중 작업 학습이 가능하다. 또한 T5는 사전학습 단계에서 방대한 양의 한국어 데이터를 학습하여 한국어의 문법, 어휘, 문맥 등을 습득하는 단계를 거쳐 한국어의 특성과 구조를 학습하였다.

이를 통해 T5는 우리가 입력한 한국어 문장의 문맥과 의미를 확인하고, 앞에서 언급한 과정을 통해 문장의 중요한 정보를 추출하여 요약과 같은 태스크를 효과적으로 수행할 수 있다. 추가로 T5는 Google에서 대중성을 확보하여 관련 연구, 개발 등에 대한 다양한 자료와 커뮤니티가 형성되어 있어 본 논문에서는 T5를 선정하였다.

2.3 기존 한국어 요약 연구

자연어 처리는 통상 복잡하고 까다로운 분야로 알려져 있다. 그런데 수많은 언어 중 한국어를 처리하는 과정은 유독 어렵게 느껴진다고 한다. 왜냐하면 한국어는 어간에 접사가 붙어 의미와 문법적 기능이 변화하는 교착어에 속하기 때문이다. 교착어에 속하는 한국어만의 특징은 표 4와 같다.

기존 한국어 요약 연구는 다음과 같다. 먼저, 타임라인 기반의 하나의 사건에 대한 뉴스 스트림 요약[7]으로 평가지표는 BLEU, Rouge-1, 2, L을 이용하였다. 다음은 한국어 요약을 위한 참고자료 및 문서인식 의미 평가 방법[8]으로, Rouge-1, 2, L과 RDASS 지표를 활용하였다. 다음은 트랜스포머 기반 인코더-디코더 제목 생성 모델[9]을 이용한 것으로, 평가지표는 Rouge-1, 2, L을 이용하였다. 이후 BART와 GPT 모델에 학습한 사전훈련 언어모델을 사용하여 문서요약 생성 성능을 비교[10]하였고, 문서생성 요약 성능지표로는 Rouge-1, 2, L을 활용하였다. 마지막으로 한국어 생성요약 모델인 KoBART 기반 리뷰 원문과 요약문을 활용한 연구[11]가 있다.

표 4 교착어에 속하는 한국어만의 특징
Table 4 Characteristics of the Korean language that belongs to the agglutinative language

Characteristics	Example / Description
Meaning due to addition of affixes	사과(어간) +를(접사) 일 때 사과는 목적어가 되지만, 사과(어간) +가(접사) 일 때는 주어가 되어 문법적 기능이 달라짐
Flexible word order	나는 축구를 하러 간다. / 축구를 하러 나는 간다. 와 같이 한국어는 단어의 순서를 바꾸어도 전체 맥락을 이해하는데 전혀 문제가 되지 않음
An ambiguous spacing rule	한국어는 띄어쓰기를 지키지 않아도 문장의 맥락을 이해하는데 큰 무리가 없음
No difference between a written review and a question, no subject	점심 먹었어. / 점심 먹었어? 와 같이 동일한 문장에 마침표 대신 물음표를 붙이고, 주어에 대한 정보를 생략하더라도 문장이 완성됨

3. 테스트 케이스 생성

본 장에서는 트랜스포머 기반 한국어 요약을 위한 메타모픽 테스트 케이스 생성을 제안한다. 먼저, 메타모픽 관계를 이용한 테스트 케이스 생성에 대해서 제안하고, 기존의 테스트셋을 활용하여 새로운 테스트 케이스 생성을 위한 변형규칙을 제안한다. 그림 1은 IDEF0(Icam DEFinition for Function Modeling) 표기법으로 메타모픽 테스팅 프로세스를 도식화한 것이다.

3.1 메타모픽 관계를 이용한 테스트 케이스 생성

메타모픽 관계를 이용하여 소스 테스트 케이스 기반으로 팔로우업 테스트 케이스를 생성하기 때문에 두 테스트 케이스의 사전조건과 수행 절차는 동일하다.

메타모픽 관계 MR(Metamorphic Relation)은 주어진 테스트셋 T를 T'로 변형한 후 메타모픽 테스팅을 적용하기 위해 T와 T'사이에 일정한 관계가 성립되도록 정의한다. 이는 명사, 동사 등에 변형을 가하더라도 T5의 요약 결과는 유사해야 하는 메타모픽 관계를 정의한다. 이 메타모픽 관계를 수식(t는 입력에 이용되는 테스트 텍스트, T는 테스트셋, P는 구현하려는 프로그램, f는 변형규칙, T'는 변형규칙을 적용하여 생성한 테스트셋, P(T)는 T5의 요약 결과)으로 나타내면 식 (1)과 같다.

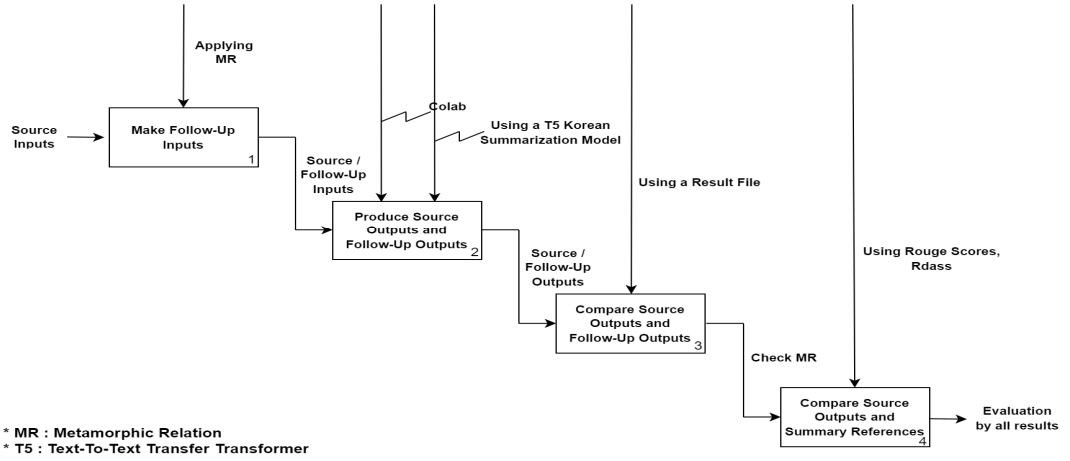

그림 1 메타모픽 테스팅 절차
Fig. 1 The metamorphic testing process

$$MR_n = \frac{\forall t \in T,\ P(T) = Summarization\ of\ T5\ Model\ (T)}{f(T) = T' \rightarrow P(T) \fallingdotseq P(T')} \quad (1)$$

위 메타모픽 관계를 만족하는 MR_n(n: 1 ~ 7, 각 테스트 케이스)과 각 테스트 케이스에 대한 예시는 다음과 같다.

MR_1 : 신문기사 텍스트 내 이름 혹은 명사 변형
- 예시) 이종섭(→ 이종균 / MR_1 적용) 장관은 어려운 안보상황에서 각자 소임을 다해 준 직원들에게 격려와 고마움을 표한 뒤 “2023년에도 대한민국의 자유·평화·번영을 강력한 힘으로 뒷받침할 수 있는 ‘튼튼한 국방, 과학기술 강군’ 건설에 최선을 다할 것”을 주문했다.

MR_2 : 신문기사 텍스트 내 국가명 변형
- 예시) 양국은 또 북한, 중국, 러시아(→ 시리아 / MR_2 적용) 등 증대하는 안보 위협에 대응하기 위해 군사 훈련을 강화하고 연합 방위 태세를 업그레이드할 방침이다.

MR_3 : 신문기사 텍스트 내 직업명 변형
- 예시) 항구, 해협 등 연해에서 선박의 입·출항로를 안 내하는 도선사(→ 항해사 / MR_3 적용)는 임금이 높아 2위에 올랐다.

MR_4 : 신문기사 텍스트 내 구두점 변형
- 예시) 국군의무학교는 3일 “(“ → ‘ / MR_4 적용)파병준비단 장병을 대상으로 이달 19~20일 유엔 PKO 임무단의 의료역량을 강화하기 위해 유엔 인증 표준 응급 처치 교육인 ‘Buddy First Aid(BFA·전투원 응급 처치)’를 첫 시행한다” (” → ’ / MR_4 적용)고 밝혔다.

MR_5 : 신문기사 텍스트 내 해당 명사를 동의어 명사로 변형
- 예시) 무열혁신 4.0의 첫 중점과제인(→ 중요과제인 / MR_5 적용) 군사대비태세 분야를 이끌고 있는 장종중(준장) 작전처장은 “군사대비태세 완비는 ‘절대 양보하거나 물러설 수 없다’는 인식 아래 추진해야 하는 무열혁신 4.0의 핵심 분야”라고 강조했다.

MR_6 : 신문기사 텍스트 내 해당 동사를 동의어 동사로 변형
- 예시) 대화생방테러특수임무대(CRST)가 화생방 작용제 탐지와 식별을 실시했다.(→ 시행했다. / MR_6 적용)

MR_7 : 신문기사 텍스트 내 문장 순서 변형
- 예시) 생성형 인공지능(AI) 챗GPT 개발사인 오픈AI가 14일(현지시간) 더 ‘똑똑해진’ 인공지능 툴을 공개했다. 오픈AI는 이날 대규모 AI 언어모델(LLM)인 GPT-4를 출시했다고 밝혔다. 전 세계적으로 인기를 끌고 있는 챗GPT에 적용된 GPT-3.5의 업그레이드 버전이다.(→ 두 번째 문장으로 이동 / MR_7 적용) 오픈AI는 GPT-4 모델이 많은 전문적인 시험에서 ‘인간 수준의 능력’을 보여줬다고 설명했다. 미국 모의 변호사 시험에서는 90번째, 대학 입학 자격시험인 SAT읽기와 수학시험에서는 각각 93번째와 89번째의 백분위수를 기록했다고 이 회사는 강조했다. SAT 등 주요 시험에서 상위

10%에 해당한다는 것이다. 오픈AI는 “평소 대화에서는 GPT-3.5와 차이가 크게 나지 않을 수 있다”면서도 “GPT-4는 훨씬 더 신뢰할 수 있고, 창의적이며 더 미묘한 명령을 처리할 수 있다”고 평가했다. 또 이전 모델보다 틀린 답이나 주제를 벗어난 답은 적다며 많은 표준화된 시험에서 인간보다 더 좋은 성적을 낼 것이라고 덧붙였다.

4. 실험 및 평가

4.1 실 험

트레이닝셋은 한국지능정보사회진흥원(NIA)의 인공지능 학습용 데이터 구축사업(AI-Hub)을 통해 수집된 문서요약 텍스트 중 신문기사 텍스트를 활용하였고, 테스트 셋은 국방일보 신문기사를 활용하여 자체 제작 하였다. 트레이닝 셋과 테스트 셋의 내용은 표 5와 같다.

표 5 트레이닝 셋과 테스트 셋
Table 5 Training sets and test sets

Data Type	Data Form	Quantity
Training Set	Providing AI-Hub Newspaper article texts	300,000 original datas / 600,000 summaries
Test Set	Self-produced Newspaper article texts (Edited by the National Defense Daily)	700 original datas / 700 summaries

실험순서는 먼저 Source Inputs에 메타모픽 관계 MR_n(n : 1 ~ 7, 각 테스트 케이스)을 적용한다. 이후 생성된 Follow-Up Inputs와 Source Inputs를 트레이닝 셋을 활용하여 사전 학습이 완료된 T5에 입력한다. 실험이 완료된 후 자동으로 생성된 결과 파일을 이용하여 Source Outputs와 Follow-Up Outputs를 비교하고, 이를 통해 메타모픽 관계 만족여부를 판단한다. 이 과정을 진행한 후 Rouge와 Rdass를 사전에 사람이 만든 요약문과 모델이 생성한 요약문에 적용하여 결과를 도출/검증한다.

4.2 모델 검증 및 기존연구 비교

T5 한국어 요약은 입력한 텍스트에서 핵심 정보를 추출하고 요약문을 생성하는데 중점을 두는 프로세스이다. 입력한 텍스트를 통해 T5는 문장의 각 항목별 관련성, 일관성 등의 요소를 고려하여 핵심 문장을 식별하여 최종적으로 요약문을 생성한다. 이를 바탕으로 실험 진행은 표 6과 같이 진행되었고, 표 6은 메타모픽 관계 네 번째인 구두점 변형 후 실험한 결과 메타모픽 관계를 만족하였다.

표 6 실험 진행 경과
Table 6 Experiment progress

Original
해병대가 국방과학기술을 활용한 미래 전력 구축에 박차를 가하기 위해 한국국방기술학회와 손을 맞잡았다. 해병대사령부는 지난 17일 부대 대회의실에서 한국국방기술학회와 업무협약을 맺고, 국방과학기술 분야 정보 교류 및 공동연구체계를 구축하기로 했다. 행사에는 김계환(중장) 해병대사령관과 박영욱 한국국방기술학회 이사장 등 두 기관 주요 직위자들이 참석했다. 두 기관이 서명한 업무협약서에는 해병대 미래 전력 건설을 위한 국방과학기술 개발동향 등 정보 교류, 국방과학기술 발전을 위한 지식·정보·노하우 상호 활용 및 교류·협력, 국방과학기술 교류를 위한 각종 행사 협력 등의 내용이 담겼다. 박 이사장은 행사를 마친 뒤 ‘과학기술, 미래 국방과 만나다’라는 주제의 강연에서 미래 국방과학기술을 소개했다. 김 사령관은 “(” → 삭제 / MR_4 적용)해병대는 4차 산업혁명 시대에 발맞춰 국방과학기술을 접목한 인공지능(AI) 기반 유·무인 복합체계 적용과 스마트 부대 운용을 위해 미래 혁신에 노력을 집중하고 있다”(” → 삭제 / MR_4 적용)며 “이번 업무협약 체결은 해병대의 미래 전력 건설과 비전 구상에 큰 도움이 될 것”이라고 강조했다.
Original Summary Results
해병대사령부는 지난 17일 부대 대회의실에서 한국국방기술학회와 업무협약을 맺고, 국방과학기술 분야 정보 교류 및 공동연구체계를 구축하기로 했으며, 박 이사장은 행사를 마친 뒤 과학기술, 미래 국방과 만나다라는 주제의 강연에서 미래 국방과학기술을 소개했다.
Test Case Application Summary Results
해병대사령부는 지난 17일 부대 대회의실에서 한국국방기술학회와 업무협약을 맺고, 국방과학기술 분야 정보 교류 및 공동연구체계를 구축하기로 했으며, 박 이사장은 행사를 마친 뒤 과학기술, 미래 국방과 만나다라는 주제의 강연에서 미래 국방과학기술을 소개했다.

표 7은 메타모픽 관계 다섯 번째인 해당 명사를 동의어 명사로 변형 후 실험한 결과 변형한 부분을 제외하고 문장의 형태나 틀이 전혀 훼손되지 않아 메타모픽 관계를 만족하는 케이스로 선정하였다.

그림 2는 메타모픽 관계를 적용한 실험 결과로 텍스트 내 해당 명사와 동사를 동의어 명사와 동사로 변형한 MR_5와 MR_6이 각각 90%와 82%의 확률로 메타모픽 관계를 만족 하였고, 텍스트 내 문장 순서를 변형하는 MR_7이 18% 만족하는 것으로 나타나 대부분 메타모픽 관계를 만족하지 않아 테스트에 실패(Test Fail)한 것으로 판단하였다. 이름 혹은 명사를 변형하여 실험한 MR_1도 75%로 만족하였고, 국가명과 직업명, 구두점을 변형한 MR_2 ~ MR_4는 메타모픽 관계를 만족하지 않는 경우가 많았다.

표 7 메타모픽 관계를 만족하는 사례
Table 7 Case that satisfies a metamorphic relation

Original
육군에는 전차·자주포, 공군에는 전투기가 있듯 해군에는 함정이라는 대표적 무기체계가 있습니다. 전차·전투기와 비교되는 함정의 특징은 '거대하다'는 점일 것입니다. 함형별로 크기에 차이는 있지만, 해군 주력 함정인 충무공이순신급 구축함은 길이 150m에 무게는 4400톤(경하)에 달합니다. 거대한 덩치를 자랑하는 함정은 다른 무기체계보다 출동을 위해 많은 사전작업(→ 사전점검 / MR_5 적용)이 필요합니다. 출항 전에 주요 무장과 각종 전자장비가 잘 작동하는지 꼼꼼히 살펴야 합니다. 또 승조원들이 승함을 완료했는지 인원 확인도 출항 전에 이뤄집니다. 안전작업도 필수죠. 그래서 모든 함정은 정해진 절차에 따라 촘촘히 출항 준비를 합니다. 준비가 잘못되거나 빠트린 요소가 있다면 전투력에 악영향을 미치기 때문이죠.
Original Summary Results
해군 주력 함정인 충무공이순신급 구축함은 길이 150m에 무게는 4400톤(경하)에 달하며 다른 무기체계보다 출동을 위해 많은 사전작업이 필요하기 때문에 정해진 절차에 따라 촘촘히 출항 준비를 합니다.
Test Case Application Summary Results
해군 주력 함정인 충무공이순신급 구축함은 길이 150m에 무게는 4400톤(경하)에 달하며 다른 무기체계보다 출동을 위해 많은 사전점검이 필요하기 때문에 모든 함정은 정해진 절차에 따라 촘촘히 출항 준비를 합니다.

그림 2 실험 결과
Fig. 2 Experimental results

다음으로 사전훈련 언어 모델을 통해 생성된 요약문과 사람이 미리 작성해 놓은 참조 요약문을 비교하여 한국어 요약의 성능을 평가하였다.

평가지표로는 문서 요약 성능 평가에 많이 사용되는 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)와 RDASS(Referenceless DAsymmetric Sentence Similarity)를 적용하여 평가하였다.

먼저, 문서요약 성능지표로 많이 사용되는 Rouge는 n-gram(연속된 n개의 단어를 의미)을 이용해 참조 요약문과 생성된 요약문을 비교하여 얼마나 겹치는지를 확인하여 수치로 나타내는 요약모델 성능 평가 척도이다.[12] Rouge는 Rouge-1, Rouge-2 등의 Rouge-N이 있다. Rouge-N은 n-gram 기반의 recall(모델이 요약한 결과와 사람이 요약한 결과 간의 공유 비율을 계산하여, 모델이 사람이 요약한 결과를 얼마나 잘 추출하고 있는지 측정하는 지표)을 측정하는 지표이다. Rouge-L은 모델이 요약한 결과와 사람이 요약한 결과 간 가장 긴 공통부분의 길이를 측정하여 모델의 요약 성능을 평가하는 지표이다. Rouge-SU는 Skip-Bigram(문장 내에서 두 단어 사이에 하나 이상의 단어가 존재하는 경우)과 Unigram(텍스트를 분석할 때 단어를 가장 작은 단위로 쪼개어 처리하는 방법)을 모두 고려하는 recall 지표이다. 마지막으로 Rouge-U는 Unigram을 기반으로 하는 recall 지표로 모델이 요약한 결과와 사람이 요약한 결과 간에 공유되는 단어의 수를 계산한다.

Rdass는 모델이 요약한 결과와 사람이 요약한 결과 간의 문장 유사도를 계산하는 것으로, Rouge와 달리 Rdass는 사람이 요약한 결과를 참조하지 않고도 계산이 가능하다.

표 8은 T5 한국어 요약 성능 평가 결과로 Rouge-1, 2, L 성능지표를 활용하였다. 이 외에 기존 연구에서는 다루지 않았지만 위에서 언급한 것을 바탕으로 Rouge-U, SU와 Rdass 성능지표도 활용하였다. 통상 한국어 요약 성능 평가지표는 대부분의 기존 연구에서 Rouge-1, 2,

표 8 T5 한국어 요약 성능 평가 결과
Table 8 T5 Korean summarization performance evaluation results

Sort	Rouge-1	Rouge-2	Rouge-L	Rouge-U	Rouge-SU	Rdass
MR_1	0.4	0.31	0.55	0.56	0.36	0.46
MR_2	0.02	0	0.03	0.03	0	0.01
MR_3	0.36	0.29	0.35	0.38	0.21	0.29
MR_4	0.4	0.31	0.53	0.53	0.24	0.39
MR_5	0.45	0.36	0.51	0.57	0.39	0.48
MR_6	0.41	0.32	0.55	0.55	0.4	0.48
MR_7	0.41	0.32	0.59	0.61	0.36	0.49

표 9 기존 연구와 비교
Table 9 Comparison of existing studies

Method	Used Dataset	Used Model	Used Score			
			Rouge1	Rouge2	Rouge-L	Rdass
Stream [7]	2,000 newspaper articles	Bi-LSTM	0.32	0.1	0.22	-
Reference [8]	3,000,000 newspaper articles	BERT	0.14	0	0.14	0.71
Title [9]	25,564 thesis abstract data	T5	0.26	0.12	0.26	-
Document [10]	183,000 news articles, etc.	BART	0.46	0.35	0.4	-
Loss Function [11]	5,500 review data	BART	0.24	0.15	0.24	-
Proposed Method	300,000 news articles, etc.	T5	0.35	0.27	0.44	0.37

L이 사용되는데 본 논문에서 Rouge-U, SU와 Rdass 성능지표를 제시한 것은 한국어 요약 성능 평가지표로는 Rouge-U, SU와 Rdass가 보다 합리적으로 비교할 수 있는 Metric으로 판단하였기 때문이다.

한국어를 영어와 비교하였을 때, 표 4처럼 한국어는 언어의 특성상 순서보다는 조사가 중요하기 때문에 순서가 아닌 조합으로 성능을 판단할 수 있는 Rouge-U, SU가 Rouge-1, 2, L에 비해 더 적절한 평가지표로 판단하였다. 예를 들어, 사람이 요약한 결과 문장은 “이강인이 공을 찼다.”이고, 모델이 요약한 결과 문장이 “찼다 축구공을 이강인이”라면, 어순이 바뀌었기 때문에 순서가 중요한 Rouge-1, 2, L의 결과치는 낮아지게 된다. 이를 보완하기 위해 Rouge-SU, U는 앞에서 언급한 Skip-Bigram과 Unigram을 기반으로 계산하여 기존 지표에 비해 한국어 요약 성능 평가지표로 적합하다고 할 수 있다. Rdass 역시 같은 이유와 더불어 문장의 유사도로 판단하는 지표이기에 의미론적으로 성능을 분석하기 위해서는 단어의 순서를 비교하는 것이 적절한 지표가 되지 않을 수도 있다고 판단하여 Rouge-U, SU와 Rdass 지표를 이용하여 T5의 요약 성능을 평가하였다.

표 9는 Rouge와 Rdass를 성능지표로 이용한 기존 연구와 본 연구의 결과를 비교한 것이다.

표 9를 통해 기존 연구와 본 연구의 결과를 각 성능지표인 Rouge-1, 2, L 수치의 평균값으로 비교해 보면 본 논문의 Proposed Method가 Stream[7]에 비해 14%, Reference[8]에 비해 26%, Title[9]에 비해 14%, Loss Function[11]에 비해 14% 높았다.

그러나, Document[10]은 본 논문의 Proposed Method보다 평균값이 5% 높아 BART 모델의 요약 성능도 뛰어난 것으로 확인하였다.

이를 종합하면, 본 논문의 T5 모델 성능이 기존 연구의 모델 성능에 비해 약 13% 향상되었다.

마지막으로, Rdass 지표는 Reference[8]의 수치가 본 논문의 수치에 비해 34%로 크게 높아 모델이 요약한 결과와 사람이 요약한 결과의 문장 유사도 측면에서는 T5가 BERT에 비하여 성능이 상대적으로 떨어진다는 것을 확인할 수 있었다.

5. 결론 및 향후연구

트랜스포머 기반의 T5를 활용한 한국어 요약은 특정 테스트셋에 대한 요약 결과와 실제 환경에서의 광범위한 입력에 대한 요약 결과가 유사함을 보장해야 한다. 그러나 기존의 평가지표만으로는 이에 대한 보장이 어렵다. 이에 따라 본 연구에서는 한국어 요약을 위한 메타모픽 테스트 케이스 생성을 제안하였다.

제안은 메타모픽 관계를 통해 변형시킨 테스트셋을 모델에 입력하여 얻어진 출력 결과를 기존의 출력 결과와 비교하여 메타모픽 관계를 준수하는지 확인하는 방법이다. 제안의 효과를 확인하기 위해 T5, AI-Hub에서 제공하는 문서요약 텍스트 트레이닝셋, 국방일보를 편집한 자체 제작 테스트셋을 이용하여, 실제 모델을 생성하고 테스트를 적용하여 연구를 수행하였다. 도출한 결과 이름 혹은 명사를 변형한 MR_1은 75%, 해당 명사/동사를 동의어로 변형하는 MR_5와 MR_6은 각각 90%와 82%의 높은 확률을 보였다. 반면에, 국가명과 직업명, 구두점을 변형한 MR_2~MR_4는 메타모픽 관계를 만족하지 않는 경우가 많았다.

본 연구를 통해 메타모픽 테스팅의 프로세스를 지속적으로 개선하고 최적화하여 인공지능 소프트웨어를 테스팅 할 때 다양한 측면에서 결함을 감지하고 메타모픽 관계와 같은 더 나은 대체 방법을 찾는 등의 연구가 지속되어야 할 것이다.

References

[1] Hosuon Yoo, Kunhui Lee, Seunghoon Na, “Sequence-to-sequence Models for Korean Dialogue Summarization”, *Proc of the Korea Computer Congress 2022*, pp. 350-352, 2022. (in Korean)

[2] Dongsu Kang, "Bridging Fuzz Testing and Metamorphic Testing for Classification of Machine Learning," *40th IEEE International Conference on Consumer Electronics (ICCE),* Las Vegas, USA, pp. 1-2, 2022.
[3] Byeongwoo Na, Dongsu Kang, "A Method for Generating Metamorphic Test Cases for CNN Image Classification Models", *KIISE Transactions on Computing Practives,* Vol. 28, No. 1, pp. 33-41, 2022. (in Korean)
[4] Lingzi Jin, Zuohua Ding, Huihui Zhou, "Evaluation of Chinese Natural Language Processing System Based on Metamorphic Testing", MDPI(Multi-disciplinary Digital Publishing Institute) Mathematics, 2022.
[5] 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.
[6] 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.
[7] Ian Jung, Su jeong Choi, Seyoung Park, "News Stream Summarization for an Event based on Timeline", *Journal of KIISE*, Vol. 46, No. 11, pp. 1140-1148, 2019. (in Korean)
[8] Dongyub Lee, Myeongcheol Shin, Taesun Whang, Seungwoo Cho, Byeongil Ko, Daniel Lee, Eunggyun Kim, Jaechoon Jo, "Reference and Document Aware Semantic Evaluation Methods for Korean Language Summarization", *International Conference on Computational Linguistics (COLING)*, pp. 5604-5616, 2020.
[9] Sujin Seong, Jeongwon Cha, "Transformer Encoder-Decoder based Title Generation Model with Word Loss and Repetition Penalty", *KIISE Transactions on Computing Practices,* Vol. 27, No. 4, pp. 210-215, 2021. (in Korean)
[10] Minchae Song, Kyungshik Shin, "A Study of Pre-trained Language Models for Korean Language Generation", *Journal of Intelligence and Information Systems,* pp. 309-328, 2022. (in Korean)
[11] Dahoon Gu, Byungwon On, Dongwon Jeong, "Relevance and Redundancy-based Loss Function of KoBART Model for Improvement of the Factual Inconsistency Problem in Abstractive Summarization", *Journal of KIIT*, Vol. 20, No. 12, pp. 25-36, 2022. (in Korean)
[12] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries", Text summarization branches out, 2004.

이 인 균
2014년 명지대학교 컴퓨터공학과 졸업(학사). 2022년~현재 국방대학교 컴퓨터공학과 석사과정. 관심분야는 한국어 자연어 처리, 딥러닝, 인공지능 SW 테스팅

강 동 수
정보과학회 컴퓨팅의 실제 논문지 제 29 권 제 7 호 참조