KoboNet: 한국어 AI 생성 텍스트 탐지를 위한 딥러닝 모델*

장원익[01], 장현종[02], 허의남†
[01,02]경희대학원 인공지능학과
†경희대학원 컴퓨터공학과
ericegg0104@khu.ac.kr, lezelamu@naver.com, johnhuh@khu.ac.kr

KoboNet: A Deep Learning Model for Detecting AI-Generated Korean Text

Wonik Jang[01], Hyunjong Jang[02], Eui-Nam Huh†
[01,02]Department of Artificial Intelligence, Kyung Hee University
†Department of Computer Engineering, Kyung Hee University
ericegg0104@khu.ac.kr, lezelamu@naver.com, johnhuh@khu.ac.kr

요 약

본 연구는 최근 대형 언어 모델(LLM, Large Language Model)의 보편화로 인해 증가하고 있는 인공지능 생성 텍스트에 대한 탐지 수요에 대응하기 위해 기존 기술과 연구를 기반으로 새로운 한국어 Ai 생성 텍스트 탐지 모델인 KoboNet을 소개한다. 이 모델은 텍스트의 음운 분석 피쳐와 품사 분석 피쳐를 사용한 모델로 경량성, 일반화 가능성, 단어 변경 등의 후처리를 통한 탐지 회피 공격에 대한 대응성을 목표로 하였으며 7천 개의 문서를 이용한 실험을 통해 즉시 현장에 적용 가능할 정도의 성능을 입증하였다. 본 연구는 텍스트 내의 특징을 단어 빈도 분석, 어휘 난이도 피쳐와 같은 방법이 아닌 글의 전체적 문체 분석을 통해 AI를 구분해 내는 탐지 방법이 효과적임을 시사하며 향후 다양한 언어와 서비스 분야에 확장될 수 있는 기반을 제공한다.

1. 서 론

최근 대형 언어 모델(Large Language Model, LLM)의 발전으로 인해 인간이 이해할 수 있을 정도로 자연스러운 텍스트를 자동으로 생성하는 AI 시스템이 광범위하게 활용되고 있다. 동시에 이로 인한 사회적 혼란과 정보 신뢰성 저하에 대한 우려가 제기되고 있다. 특히 생성형 AI를 이용한 조작 콘텐츠나 허위 정보의 유포는 일반 사용자뿐 아니라 언론, 학계 등에도 심각한 영향을 미칠 수 있다.[1]

기존의 AI 텍스트 탐지 기술은 LLM을 사용한 탐지 방법과 머신러닝 기반 탐지 방법을 사용한다. LLM을 이용한 탐지 기술은 높은 성능을 보이나 일반화 성능을 높이기 위해 많은 데이터를 필요로 하고 학습과 추론에 많은 계산 자원이 소모된다. 머신러닝 기반 탐지 방법은 적은 자원을 사용하면서도 준수한 성능을 보인다.[2] 그러나 이 방법은 탐지를 위해 '어휘 다양성', '문장 길이', '어휘 복잡성', '문장 구조 복잡성'과 같은 지표를 사용하기 때문에 단어의 난이도와 어휘 복잡성을 뚜렷하게 구분하기 어려운 한국어에서의 탐지율이 높지 못하다.

본 논문에서는 AI 텍스트에 대한 언어학적 차이 분석 연구를 기반으로 텍스트에서의 품사 사용 비율과 발음 사용 비율 분석을 통해 언어 지문을 추출하여 AI 생성 텍스트를 구분하는 연구를 진행한다. 이 연구에서는 실험을 통해 본 연구가 제안하는 Kobonet 모델의 생성형 언어 모델에 대한 일반화 능력과 경량성, 후처리를 통한 탐지 회피 대응 능력을 검증한다.

2. 관련 연구

대형 언어 모델의 급속한 발전은 자연어 생성 능력의 획기적인 향상을 이끌어냈으며 이에 따라 AI 생성 텍스트의 사용도 폭넓게 확산되고 있다. 2023년 연구에 따르면 이러한 AI 텍스트는 교육, 언론, 과학 분야 등 다양한 영역에 영향을 미치고 있으며, 정보의 신뢰성 및 윤리적 책임 문제로 인해 생성 텍스트 탐지 기술의 중요성이 점차 강조되고 있다. [1]

대형 언어 모델이 매우 고도화된 한국어 텍스트 생성 능력을 갖추었지만 대부분의 영어권 언어를 기반으로 한 AI 생성 텍스트 탐지 모델은 한국어를 대상으로 매우 저조한 능력을 보인다. [3] 한국어 텍스트에서의 AI 생성 텍스트 탐지를 주제로 하는 다른 연구에서는 기존 머신러닝 기반 분류 방법에서 자주 사용하는 n-gram, 단어 간격 분석 등의 피쳐에 쉼표 사용 패턴 특징을 추가로 사용한 분류기를 고안하여 한국어에서 AI의 문체 특징은 특정 단어가 아닌 품사 분석을 통해 드러날 수 있음을 제안한다. AI 생성 텍스트와 인간 작성 텍스트의 차이를 언어학적으로 분석한 연구[4]는 우리가 텍스트의 어떤 특징을 피쳐로 사용하는 것이 좋은지에 대한 강력한 힌트를 말한다. 이 논문의 연구에 따르면 AI와 사람은 음운적 분석에서 각자 선호하는 발음이 통계적인 차이를 가진다. 논문은 이를 AI가 특정 발음을 선호할 가능성이 있다고 서술한다. 다른 분석에서 AI와 사람이 작성한 글은 가독성 점수에서 많은 차이를 보이는데 그 원인이 인간 작성자가 어려운 발음이 많이 나오거나 유사한 품사가 반복되는 것, 동일한 단어를 여러 번 사용하는 걸 피하는 특성에서 기인하였을 가능성이 있다. 따라서 이 연구의 결과에 기반하여 음운적 특성 및 품사 분석 결과를 피쳐로 사용하는 탐지 모델이 성능을 보이리라는 충분한 근거를 시사한다.

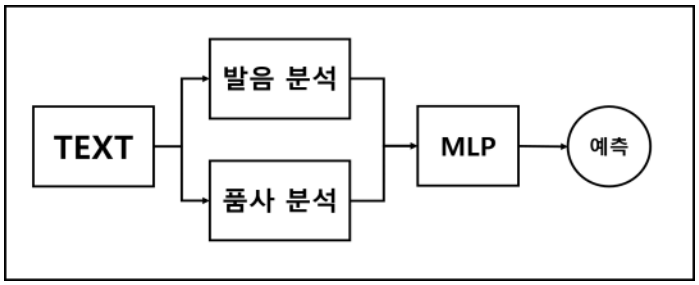
3. 제안 방법

3.1. KoboNet 모델 구조

본 모델은 전처리된 특징 벡터를 입력으로 받아 총 3개의 은닉층(hidden layer)을 거치는 MLP 구조로 구성된다. 각 은닉층은 128차원의 출력을 가지며, ReLU 활성화 함수와 Dropout(0.3)이

* 이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00155911, 기여율 50%)과 정부 (과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2023-00258649, 기여율 50%).
† 교신저자 : 허의남

적용된다. 출력층은 1차원으로 구성되며, 이진 분류를 위해 sigmoid 함수를 사용한다. 해당 출력은 모델이 AI 생성 텍스트임을 판단한 확률로도 사용된다.



[그림 1] KoboNet 모델 구조도

3.2. 발음 정보 추출

발음 정보 추출은 IPA 발음 데이터를 기반으로 나누어진다. 한국어 텍스트를 KoG2P[5] 코드를 이용한 알고리즘에 기반하여 IPA 발음 기호로 변환한 후 발음 기호들의 등장 횟수를 발음 방법에 따라 다양한 기준으로 분류한다.[표 1] 이에 따라 분류된 데이터를 전체 토큰 수로 나누어 비율 데이터로 가공한다.

[표 1] 발음 방법에 따른 IPA 기호 분류

분류	IPA 기호 분류
조음 방법	파열음: p, p^h, p̪, t, t^h, t̪, k, k^h, k̪ 파찰음: tɕ, $tɕ^h$, tɕ͈ 마찰음: s, s̪, h 비음: m, n, ŋ 설측음: l 탄음: ɾ
조음 위치	양순음: p, p^h, p̪, m 치조음: t, t^h, t̪, s, s̪, n, l, ɾ 연구개음: k, k^h, k̪, ŋ 경구개음: j, tɕ, $tɕ^h$, tɕ͈ 성문음: h
혀 높이	고모음: i, u, ɯ 중-고모음: e, o 중-저모음: ʌ, ɛ 저모음: a
혀 위치	전설모음: i, e, ɛ 중앙모음: ʌ, a 후설모음: u, o, ɯ
원순모음	원순모음: o, u, w, we, wa, wʌ 비원순모음: i, e, ɛ, a, ʌ, ɯ, j, ja, jo, ju, je, jʌ
이중모음	이중모음: ja, jo, ju, je, jʌ, wa, we, wʌ, ɯi
단모음	단모음: i, u, e, o, ʌ, ɯ, ɛ, a

3.3. 형태소-품사 정보 추출

형태소 분석 후 품사 정보를 추출하는 과정은 koNLpy[6]의 komoran 모듈을 활용하였다. komoran 모듈의 분석을 통해 추출된 품사를 Universal POS Tagset[7] 기준에 따라 10개 범주로 분류한 결과를 전체 토큰 수로 나누어 비율 데이터로 가공한다.

3.4. 학습 데이터 구성

본 실험에 사용되는 데이터는 AI가 생성한 텍스트와 사람이 생성한 텍스트 데이터로 구성된다. 본 연구에서는 각 클래스(사람 작성, AI 생성)가 동일한 비율로 구성된 총 5,400개의 훈련 데이터셋과, 각각 800개의 샘플로 이루어진 두 개의 테스트셋을 활용하였다. 훈련 및 기본 테스트셋에 포함된 AI 생성 텍스트는 GPT-3.5-turbo 모델을 통해 생성되었으며, 고난이도 테스트셋에는 GPT-4 모델로 생성한 텍스트를 사용하였다.

사람 작성 텍스트는 AI-Hub에서 제공하는 "요약문 및 레포트 생성 데이터셋"[8] 중 뉴스 기사 본문 데이터를 기반으로 수집하였다. 이를 통해 제안 모델의 AI 탐지 성능뿐 아니라, 다양한 모델에 대한 일반화 가능성도 함께 평가하였다.

4. 실험 결과 및 분석

4.1. 실험 환경

모델 학습 및 평가 실험은 Python 기반의 PyTorch 프레임워크를 이용하여 수행되었으며, 주요 설정은 다음과 같다. 학습은 5분 이내에 30 epoch 동안 수행되었으며, optimizer로는 Adam을, 손실 함수로는 "Binary Cross Entropy with Logits Loss"를사용하였다. 학습률은 0.001로 고정하였고, batch size는 32로 설정하였다. 전처리된 입력 벡터의 차원 수는 총 32차원(품사 분포 + 발음 자질)이다.

4.2. 실험 결과

본 연구에서는 각 800개의 행을 가진 테스트셋을 대상으로 모델의 성능을 점검하였다. 성능 평가지표로 Accuracy, Precision, Recall, F1-Score를 사용하였으며 테스트셋 내 라벨에 따른 지표를 **[표 2]**에 결과를 정리하였다. KoboNet 모델은 AI 텍스트 탐지에 대하여 높은 정밀도를, 인간 생성 텍스트에 대해서는 높은 재현율을 나타내었다. 이는 본 연구에서 제안하는 모델이 높은 정확도로 신중하게 AI 텍스트를 판별하는 경향을 가진다는 것을 나타낸다. 또한 GPT4 모델이 생성한 텍스트를 학습하지 않았음에도 우수한 성능을 유지하여, KoboNet 모델 및 알고리즘이 트랜스포머 모델의 특징을 잘 이용하여 다양한 LLM 모델 및 최신 모델 등의 생성 텍스트 구분에도 효과적일 가능성을 시사한다.

[표 2] 테스트셋에 대한 모델 성적 결과

테스트셋	GPT 3.5-Turbo Human Text	GPT 3.5-Turbo AI Text
Precision	0.8683	0.9686
Recall	0.9725	0.8518
F1-Score	0.9175	0.9064
Total Accuracy	0.9123	
	GPT 4 Human Text	**GPT 4 AI Text**
Precision	0.8553	0.9703
Recall	0.9750	0.8321
F1-Score	0.9112	0.8959
Total Accuracy	0.9042	

본 연구에서는 타 텍스트 탐지 모델과의 비교를 위해 KatFish 데이터셋을 이용한 성능 점검 실험을 함께 진행하였다. 데이터셋의 essay 항목에 대한 성능 실험을 진행하였으며 이 결과를 **[표 3]**에 나타내었다. KoboNet 모델은 기존 실험 데이터로만 학습된 Not trained 모델과 KatFish 데이터셋을 학습한 Trained 모델

을 사용하였다. 약 700개의 문서를 통한 테스트에서 KoboNet은 한 번도 학습하지 않은 형식의 데이터에 대해서도 높은 성능을 보여주었다. 이에 그치지 않고 간단한 학습을 함으로써 KoboNet의 알고리즘이 텍스트의 형식에 종속되지 않음을 보였다.

[표 3] KatFish 벤치마크 성적 결과

모델	AUROC
DetectGPT	45.80
Exaone 3.5	81.27
KatFishNet	82.99
KoboNet(Ours)-Not trained	83.47
KoboNet(Ours)-Trained	95.64

● **KoboNet 결과 외 정보는 KatFishNet 논문[3]에서 인용됨**

본 연구에서는 타 탐지 기술의 약점이 되기 쉬운 단어 치환을 통한 탐지 회피에 대한 KoboNet의 대응력을 실험하였다. 이를 위해 실험에 사용되는 문서의 단어 빈도 분석을 하여 AI 작성 텍스트에서 많이 사용되지만 인간 작성 텍스트에서는 많이 사용되지 않는 단어들을 선정하고 이를 일정 확률에 따라 자연스러운 단어로 치환하거나 제거하는 후처리 규칙을 작성하였다. "이를 위해", "또한", "-있습니다."와 같은 직관적으로 AI가 작성했을 것이라고 느껴지는 단어들이 많이 포함되었다. 이 규칙에 쉼표를 제거하는 규칙을 더해 후처리 코드를 작성하였고 이를 문서에 적용하였다.

이 후처리 코드는 GPT Killer[9]에 AI가 작성한 문서 중 무작위 20개 문서를 이용한 대조 실험에서 GPT Killer 모델의 확신도를 평균 74%에서 40%로 낮아지게 하였다. 후처리된 테스트셋에 대한 KoboNet의 대응 능력 실험에서 KoboNet은 강건한 대응 능력을 유지하였다. **[표 4]**

[표 4] 단어 치환 탐지 회피에 대한 모델 성적 결과 (%)

	기본 문서	후처리 문서
GPT-turbo 3.5 Human	91.75	92.37
GPT-turbo 3.5 AI	92.25	92.00
GPT-4 Human	91.12	91.37
GPT-4 AI	91.12	91.37

4.3. 분석 및 고찰

본 논문은 AI가 작성한 텍스트의 탐지 방법을 언어학적 측면에서의 차이점을 이용하여 수행하고자 하는 새로운 시도를 수행하였다. KoboNet 알고리즘은 작성자의 문체 특징을 사용된 단어의 난이도, 길이, 빈도와 같은 구체적인 지표가 아닌 AI가 고려하지 못하는 발음의 편리성, 문장 구성에서의 자연스러움 등을 언어학 통계 지표를 통해 사용함으로써 분류를 수행한다. KoboNet 알고리즘은 이 특징을 통해 영어 언어권에서 주로 효과적이었던 AI 텍스트 탐지 기술의 영역을 한국어 영역까지 효과적으로 넓힐 수 있음을 증명한다. 실험을 통해 KoboNet은 cpu에서도 매우 짧은 시간 내로 학습 및 추론이 가능함을 보여주었으며, 한국어에서의 AI 작성 여부 탐지에 대한 능력을 입증하였다. 또한 학습하지 않은 AI가 생성한 문서에 대한 탐지력을 증명한 실험을 통해 추후 등장할 매우 다양한 고급 언어 모델에 대해 일반화가 가능함을 시사한다. 이에 그치지 않고 기존 탐지 기술의 약점으로 여겨졌던 단어 치환, 쉼표 제거, 일부에 대한 수정을 통해 탐지를 회피하려는 공격에 대해 KoboNet은 그 알고리즘으로 인한 매우 높은 대응력을 보인다. 위와 같은 특성에 근거하여 본 연구가 향후 인터넷 공간에서의 대규모 데이터에 대한 AI 생성 텍스트 검사, 데이터 진위성 판단, 가짜뉴스 탐지, LLM 모델의 훈련 평가 분야에 기여할 수 있을 것으로 기대된다.

5. 참고문헌

[1] X. Liu, Y. Wang, and J. Li, A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions, arXiv:2309.09035, 2023.

[2] Yongye Su and Yuqing Wu, Robust Detection of LLM-Generated Text: A Comparative Analysis, arXiv:2411.06248v1, 2024.

[3] 박신우, 황지원, 최재웅, "KatFishNet: 한국어 생성 텍스트 탐지를 위한 언어 기반 접근," 한국정보과학회 학술대회 논문집, 2023.

[4] Georgios P. Georgiou, Differentiating between human-written and AI-generated texts using linguistic features automatically extracted from an online computational tool, arxiv:2407.03646, 2024.

[5] 조예진. "Korean Grapheme-to-Phoneme Analyzer (KoG2P)" GitHub repository. 2017.

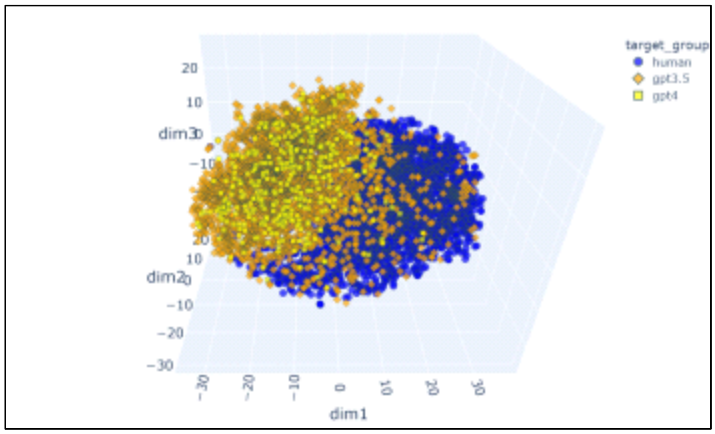
[6] 박정민 외, "KoNLPy: 파이썬 한국어 자연어처리 패키지," 제26회 한글 및 한국어 정보처리 학술대회, 2014.

[7] S. Petrov, D. Das, and R. McDonald, "A Universal Part-of-Speech Tagset," in Proceedings of the International Conference on Language Resources and Evaluation (LREC), 2012.

[8] AI Hub, 요약문 및 레포트 생성 데이터, 2021.

[9] 무하유, GPT Killer, https://www.copykiller.com

6. 부록



[그림 2] 전처리 과정을 거친 데이터의 분포 시각화

실험에 사용한 문서에 대해 KoboNet의 전처리 과정을 통해 추철된 피쳐를 t-SNE 차원축소하여 시각화하였다. 분류기 이전에 입력되는 피쳐들이 이미 일정한 정도의 분류가 되어있는 것을 통해 KoboNet 연구의 분류 기준의 타당성을 증명한다.

[표 5] 후처리 규칙 예시

기준 단어	수정 예시
이에 대한 것으로	그것으로, 예시로
기대된다	여겨진다, 주목받는다
이를 통해, 이를 위해, 또는	<제거>

[표 5]는 후처리 규칙에 대한 일부 예시이다. 실험에서는 위와 같은 변경 규칙을 15가지 이상 지정한 코드를 사용하였다. 이는 문서들에 대한 단어 빈도분석에 근거하였다.