한국어 하이브리드 텍스트의 인간-AI 작성 경계 탐지: CrossEncoder 기반 직접 학습 접근법

(Detecting Human-AI Boundaries in Korean Hybrid Texts: A CrossEncoder-based Direct Learning Approach)

이도현*, 김미수**, 손남례***

(Do Hyun Lee*, Misoo Kim*, Nam Rye Son**)

요 약

최근 ChatGPT와 같은 생성형 인공지능의 급속한 발전은 텍스트 생성 방식에 근본적인 변화를 가져왔다. 특히 인간과 AI가 협업하여 작성하는 하이브리드 텍스트가 증가함에 따라, 기존의 단순한 이진 분류 접근법의 한계가 드러나고 문장 단위의 경계 탐지 기술 필요성이 부각되고 있다. 본 연구는 한국어 자기소개서에서 인간이 작성한 부분과 AI가 생성한 부분 간의 경계를 자동으로 탐지하는 문제를 다룬다. 이를 위해 기존의 Triplet Loss 기반 간접 학습 방식의 구조적 한계를 분석하고, 분포 기반 Adaptive Threshold 전략을 결합한 CrossEncoder 기반 직접 경계 탐지 모델을 제안한다.

실험은 9,047건의 한국어 하이브리드 자기소개서로 새롭게 구축된 데이터셋을 활용하여 수행되었다. 제안하는 CrossEncoder 모델은 Fixed Top-K 기준에서 F1-score 0.597을 달성하여 Triplet Loss 대비 약 27% 향상된 성능을 보였다. 또한 Adaptive Threshold 방식을 도입했을 때 전반적으로 탐지 정밀도와 실용성이 크게 개선되었으며, 영어 데이터셋에서는 CrossEncoder가 가장 높은 성능(0.736)을 기록하였다. 다만 한국어 데이터셋에서는 Adaptive Threshold 적용 시 TriBERT가 CrossEncoder보다 근소하게 우세한 결과를 보였다. 이러한 결과는 CrossEncoder 접근이 한국어 환경에서 경계 탐지에 효과적임을 보여주는 동시에, Adaptive Threshold 방식이 모델 구조와 무관하게 성능 향상에 기여한다는 점을 시사한다.

■ 중심어 : AI 텍스트 탐지 ; 하이브리드 텍스트 ; 문장 경계 탐지 ; 생성형 AI ; 대규모언어모델

Abstract

The rapid advancement of generative artificial intelligence, exemplified by systems such as ChatGPT, has fundamentally transformed the way text is produced. With the increasing prevalence of hybrid texts collaboratively authored by humans and AI, the limitations of traditional binary classification approaches have become evident, underscoring the need for sentence-level boundary detection techniques. This study addresses the problem of automatically detecting boundaries between human-written and AI-generated segments in Korean personal statements. We first analyze the structural limitations of the existing Triplet Loss-based indirect learning method and then propose a CrossEncoder-based direct boundary detection model combined with a distribution-based Adaptive Threshold strategy.

Experiments were conducted on a newly constructed dataset of 9,047 Korean hybrid personal statements. Under the Fixed Top-K evaluation, the proposed CrossEncoder model achieved an F1-score of 0.597, representing a 27% improvement over the Triplet Loss baseline. Moreover, the introduction of Adaptive Thresholding substantially improved detection precision and overall practicality across both models, with CrossEncoder attaining the highest performance (0.736) on the English dataset. However, on the Korean dataset, TriBERT slightly outperformed CrossEncoder when Adaptive Thresholding was applied. These findings demonstrate that while CrossEncoder provides an effective approach for boundary detection in Korean hybrid texts, Adaptive Thresholding itself serves as a robust enhancement mechanism regardless of model choice.

■ keywords : AI Text Detection ; Hybrid Text ; Sentence Boundary Detection ; Generative AI ; Large Language Model

Ⅰ. 서 론

대규모 언어 모델(LLM)은 특정 영역에서 인간을 능가할 정도로 고도화되었으며, 2022년 11월

* 준회원, 전남대학교 인공지능학부
** 정회원, 전남대학교 인공지능학부
*** 정회원, 전남대학교 소프트웨어중심대학사업단

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 소프트웨어중심대학사업의 연구결과로 수행되었습니다. (2021-0-01409)
접수일자 : 2025년 08월 22일 게재확정일 : 2025년 09월 23일
교신저자 : 손남례 e-mail : nrson72@gmail.com

ChatGPT를 시작으로 Claude, Gemini 등 다양한 생성형 AI가 일상화되면서 텍스트 생성 기술의 패러다임이 근본적으로 변화하였다. 이러한 기술 진보는 문서 작성의 효율성을 획기적으로 향상시키며, 창의적 글쓰기부터 업무 보고서 작성에 이르기까지 폭넓은 활용을 이끌고 있다. 그러나 이러한 발전은 텍스트의 진정성과 저작권, 공정성에 대한 새로운 문제를 야기하고 있다. 특히 자기소개서와 같은 평가 목적의 글에서 AI 활용은 채용의 신뢰성과 공정성을 저해할 수 있으며, 교육 현장에서도 과제나 보고서 작성에 AI가 과도하게 개입되는 현상이 나타나고 있다[1][2]. 실제 사용자들은 전체 텍스트를 AI가 작성하기보다는 일부 표현을 보완하거나 특정 문장을 대체하는 방식으로 AI를 활용한다. 즉, 인간과 AI가 협업하여 작성한 "하이브리드 텍스트"가 일반화되고 있다. 이러한 하이브리드 작성 방식은 기존의 이진 분류(AI vs. Human) 기반 탐지 방식으로는 효과적인 탐지가 어렵기 때문에, 문장 단위의 경계를 식별할 수 있는 보다 정교한 탐지 기술이 요구된다.

텍스트 내 경계(boundary) 탐지는 단순한 기술적 문제를 넘어, AI 시대의 평가 시스템이 어떻게 진화해야 하는가에 대한 근본적인 질문을 제기한다. 텍스트 중 어떤 부분이 인간의 결과물이고, 어떤 부분이 AI의 기여인지를 구분할 수 있다면 평가자는 보다 공정하고 정밀한 판단을 내릴 수 있다. 또한 이는 AI 활용에 대한 규제 및 가이드라인 수립에도 실질적 근거를 제공할 수 있다. 그러나 현재까지 대부분의 AI 텍스트 탐지 연구는 영어를 중심으로 이루어져 있으며, 한국어와 같은 교착어에 적합한 연구는 매우 제한적이다. 한국어는 조사, 어미 변화, 높임법 등 복잡한 구조를 갖고 있어 영어 기반 탐지기의 정확도가 낮고 적용이 어렵다. 따라서 한국어의 언어적 특수성을 반영한 탐지 기술이 필요하다.

따라서 본 연구는 한국어 자기소개서 문서를 대상으로 인간과 AI 간의 작성 경계를 문장 단위로 탐지하는 기술을 제안한다. 이를 위해 다양한 하이브리드 작성 패턴을 반영한 데이터셋을 직접 구축하고, 기존 Triplet Loss 기반 접근의 한계를 분석한 뒤 CrossEncoder 기반 문장쌍 분류 방식을 제안한다. 또한 정답 경계 수에 맞춰 유연하게 작동하는 Adaptive Threshold 방식을 도입함으로써 실제 적용 가능성을 높이고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 AI텍스트 탐지기술과 TriBERT에 대하여 설명한다. 3장에서는 한국어 데이터 셋 구축 방법 및 하이브리드 텍스트 생성결과 및 분포에 대하여서 서술한다. 4장에서는 기존 TriBERT 문제점을 분석하고 CrossEncoder 기반 문장쌍 분류 방식을 제안한다. 5장에서는 실험 및 결과를 분석한다. 마지막으로 6장에서는 결론 및 향후연구를 기술한다.

Ⅱ. 관련 연구

1. AI 텍스트 탐지 기술

AI 텍스트 탐지 기술은 최근 대규모 언어모델(LLM)의 발전과 함께 지속적인 진화를 거쳐 왔으며, 크게 세 가지 연구 흐름으로 구분할 수 있다.

첫째, 표면적 통계 지표 기반 탐지이다. 초기 연구들은 문장 길이, 어휘 다양성, n-gram 빈도와 같은 표면적 통계 특징을 활용한 이진 분류 방식을 주로 채택하였다. 예를 들어, Grechul 외 [3]는 tf-idf 벡터화와 로지스틱 회귀, 랜덤 포레스트를 결합하여 F1-score 0.9에 달하는 높은 성능을 보고하였다. 그러나 이러한 접근은 최신 LLM이 인간과 유사한 고급 표현을 생성하는 시점에서는 식별력이 급격히 저하되는 한계를 가진다.

둘째, 의미 기반 탐지로의 전환이다. Mitchell 외 [4]가 제안한 DetectGPT는 텍스트 확률 곡률(probability curvature) 개념을 활용하여, 인간이 작성한 문장은 언어 모델의 국소적 최대 확률

영역에 위치하고, AI 생성 텍스트는 보다 평탄한 확률 지형에 위치한다는 가정에 기반하였다. DetectGPT는 별도 학습 없이도 zero-shot 탐지가 가능하다는 장점을 지니지만, 단일 모델에 의존하고 문서 단위 탐지에 집중되어 있어 문장 경계 단위의 탐지에는 적용이 어렵다.

셋째, 적대적 학습을 통한 강건성 강화이다. Hu 외 [5]는 RADAR 모델을 통해 패러프레이징, 문체 변형 등 다양한 회피 전략에 내성을 확보할 수 있는 탐지 프레임워크를 제안하였다. 해당 접근은 탐지기의 적대적 강건성을 강화하였으나, 여전히 문서 전체를 대상으로 한 이진 분류 방식에 머물러 있어, 실제 환경에서 흔히 나타나는 인간 - AI 혼합(hybrid) 텍스트의 문장 단위 경계 탐지에는 한계를 지닌다.

따라서 기존 연구들은 AI 텍스트 탐지의 정밀도와 강건성을 점진적으로 향상시켜 왔으나 대부분 문서 전체를 하나의 단위로 처리하는 이진 분류 관점에 국한되어 있다. 그러나 실제 온라인 환경에서는 인간과 AI가 혼합하여 작성한 하이브리드 텍스트가 빈번히 존재하므로, 문장 단위의 정밀한 경계 탐지가 필요하다. 본 연구는 이러한 한계를 극복하고, 한국어 텍스트의 언어적 특성을 반영한 정교한 문장 단위 경계 탐지 접근을 제안한다.

2. TriBERT기반 경계 탐지

자동 경계 탐지 방법론 중 하나로 Zeng 외 [6]가 제안한 TriBERT는 교육 분야 영어 에세이를 대상으로 개발된 모델이다. TriBERT는 Triplet Loss 기반 메트릭 학습(metric learning)을 활용하여 동일 작성자의 문장은 임베딩 공간에서 가깝게, 상이한 작성자의 문장은 멀게 배치되도록 학습한다. 구체적으로, anchor - positive(같은 작성자) - negative(다른 작성자) 삼중 문장을 입력받아, 유클리드 거리 함수를 이용해 상대적 거리를 학습하며 수식은 식 (1)과 같다.

$$L_{triplet} = \max(0, d(a,p) - d(a,n) + margin) \quad (1)$$

식(1)에서 d는 유클리드 거리 함수이며, a,p,n은 각각 anchor 문장, 같은 작성자의 positive 문장, 다른 작성자의 negative 문장을 의미한다.

학습이 완료되면 각 문서를 문장 단위로 분할하고 인접 문장군을 묶어 평균 벡터(프로토타입)를 생성한다. 이후 인접 프로토타입 간의 거리를 계산하여, 거리 값이 큰 지점을 작성자 경계 후보로 탐지한다. 실제 영어 데이터셋에서 TriBERT는 유의미한 성능을 기록하였으며, 구현이 명확하고 재현 가능성이 높다는 점에서 본 연구의 베이스라인으로 설정하였다.

그러나 TriBERT 접근법은 다음과 같은 구조적 한계를 가진다. 첫째, 학습 목적과 추론 목표 간 불일치이다. 즉, Triplet Loss는 작성자 간 문장 구분을 잘 학습하도록 설계되었으나, 실제 과제는 경계 위치 예측이다. 따라서 작성자 구분 성능이 높다고 해서 경계 탐지 정확도가 보장되지는 않는다.

둘째, 지도 학습 신호의 비효율성이다. 즉, Triplet Loss는 문장 간 유사도 차이만 활용하고, 데이터셋에 포함된 정확한 경계 레이블을 직접적으로 활용하지 못한다. 이로 인해 경계 탐지라는 목적과 불일치하는 학습 구조적 제약이 발생한다.

셋째, 거리 기반 예측의 불안정성이다. 즉, 본 연구의 사전 분석에 따르면, 실제 경계 지점에서 평균 임베딩 거리가 비경계 지점보다 약 48% 크다는 정량적 차이는 확인되었다. 그러나 표준편차가 매우 커 단순 임계값 기반 탐지는 안정적인 성능을 확보하기 어렵다.

마지막으로 Fixed Top-K 평가 전략의 제약이다. 즉, TriBERT는 사전에 정해진 Top-K 방식으로 일정 개수의 경계만을 예측한다. 그러나 실제 문서는 작성자 수와 경계 수가 다양하므로, 이러한 제약은 오탐율을 높이고 실제 응용 환경과 괴리를 발생시킨다.

이와 같은 한계는 특히 한국어 자기소개서와

같이 다양한 문체와 서술 방식이 혼재된 텍스트에서 더욱 두드러지게 나타난다. 따라서 본 연구는 TriBERT의 한계를 보완하기 위해 CrossEncoder 기반의 직접 학습 방식과 Adaptive 평가 전략을 제안한다. CrossEncoder는 문장 쌍 단위로 경계 여부를 직접 학습하여 목적-추론 불일치를 해소할 수 있으며, Adaptive 전략은 문서별 변동성을 반영하여 보다 유연하고 정밀한 경계 탐지를 가능하게 한다.

Ⅲ. 한국어 데이터셋 구축

1. 문제정의 및 구축 전략

본 연구는 문장 수준에서 인간 작성 문장(Human, H)과 AI 생성 문장(Machine, M)의 전환 경계를 탐지하는 문제를 다룬다. 하나의 문서 D는 문장 시퀀스 $\{S_1, S_2, ..., S_n\}$으로 구성되며, 각 문장 S_i는 작성 주체에 따라 $l_i \in \{H, M\}$로 라벨링된다. 문장 간 작성자가 바뀌는 지점 $l_i \neq l_{i+1}$을 경계점 b_j로 정의하고, 이러한 경계점들의 집합 $B = \{b_1, b_2, ..., b_m\}$을 정확히 예측하는 함수 $f : D \rightarrow B$를 학습하는 것이 본 과제의 핵심이다.

기존에는 한국어 하이브리드 텍스트 경계 탐지에 활용 가능한 공개 데이터가 부재하였다. 이에 본 연구는 생성형 AI가 본격적으로 활용되기 이전인 2022년 11월을 기준 시점으로 설정하고, 국내 취업 포털(잡코리아)에서 29,916건의 자기소개서를 수집하였다. 이후 세 가지 필터링 과정을 통해 고품질의 한국어 데이터셋을 구축하였다.

첫째, 100단어 이상으로 실질적인 자기소개 내용을 포함하는 문서만을 선별하였다. 선별 이유는 너무 짧은 텍스트에서 문체나 표현의 특징을 파악하기 어렵고, 경계 탐지를 위한 충분한 문맥 정보를 제공하지 못하기 때문이다. 둘째, 프라이버시 보호 및 모델 편향 방지를 위해 마스킹된 엔터티가 포함된 문서를 제거하였다. 즉, 'O'로 가려진 엔터티들은 모델이 인간 작성 텍스트의 특징으로 잘못 학습할 가능성이 있어 데이터의 순수성을 해칠 수 있다고 판단했다. 셋째, '강점', '약점', 'N자 이내로 작성하시오' 등의 서식 지시문과 같은 비본문적 요소를 제거하였다.

이러한 정제 과정을 거쳐 최종적으로 9,047건의 순수한 인간 작성 자기소개서 데이터셋을 구축하였으며, 이는 이후 하이브리드 텍스트 생성을 위한 원천 데이터로 활용되었다.

2. 하이브리드 텍스트 생성 방법 및 구성

본 연구의 목적은 실제 사용자들이 생성형 AI를 활용하는 다양한 글쓰기 전략을 정밀하게 반영한 학습용 데이터를 구축하는 데 있다. 이를 위해 수집된 9,047건의 인간 작성 자기소개서를 기반으로, 문서 내 일부 문장을 삭제한 뒤 GPT-4.1-mini 모델을 활용하여 해당 부분을 대체 생성하는 방식으로 AI 문장을 삽입하였다. 이 과정을 통해 하나의 문서 내에 인간 작성 문장과 AI 생성 문장이 혼합된 현실적인 하이브리드 텍스트를 인위적으로 재현하였다.

실제 사용자들은 자기소개서 작성 시 전면적인 AI 의존보다는 서론 구성, 중간 전환 표현, 결론 요약과 같이 특정 구간에서 부분적으로 AI를 활용하는 경향이 있다. 이러한 실제 활용 패턴을 반영하기 위해 본 연구는 총 6개의 태스크(Task 1~6)를 설계하였다. 각 태스크는 하이브리드 텍스트의 구조, 작성자 전환 위치, 그리고 경계의 개수에 따라 분류된다.

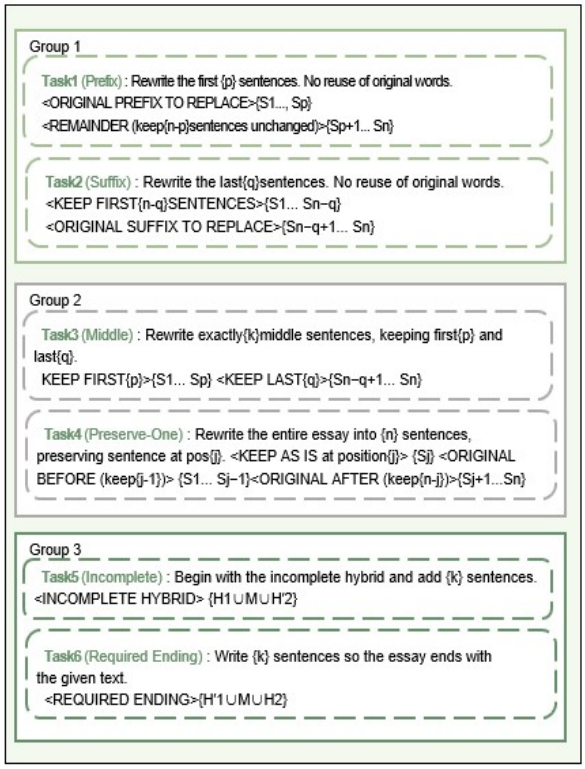

그림 1. 하이브리드 텍스트 생성 프롬프트
(Hybrid text generation prompt)

그림 (1)은 하이브리드 텍스트 생성을 위해 고안된 여섯 가지 태스크를 Group 1~3으로 구분하여 시각적으로 제시한 개념도이다. 그룹별 특징은 다음과 같다.

Group 1 (Prefix & Suffix): Task 1과 Task 2는 문서의 앞부분(도입부) 또는 뒷부분(결론부)을 AI가 재작성하는 구조이다. 사용자가 자기소개서의 시작이나 마무리를 보완하기 위해 AI를 활용하는 전형적인 패턴을 모델링하며, 각 태스크는 1개의 경계를 포함한다.

Group 2 (Middle & Preserve-One): Task 3과 Task 4는 문서의 중간 문장을 AI가 재작성하거나, 특정 핵심 문장은 인간이 직접 작성하고 도입 및 결말을 AI가 보완하는 구조이다. 이는 인간이 핵심 내용을 담당하고, 주변부의 연결과 정리를 AI에 맡기는 복합적 활용 패턴을 반영하며, 각 태스크는 2개의 경계를 포함한다.

Group 3 (Incomplete & Required): Task 5와 Task 6은 Task 3의 변형으로, 문서의 앞이나 뒤를 추가로 AI가 보완하는 구조이다. 사용자가 초안을 부분적으로 작성한 뒤 반복적으로 AI에 보완을 요청하는 시나리오를 반영하며, 각 태스크는 3개의 경계를 포함한다.

각 태스크에는 GPT-4.1-mini 모델을 활용한 프롬프트를 통해 구현되었으며, 생성된 문장은 문맥의 논리적 연결성과 표현의 자연스러움을 보장하기 위해 사후 검수와 후처리 과정을 거쳤다. 이를 통해 AI 생성 문장이 인간 작성 문장과 유기적으로 결합된 고품질 하이브리드 텍스트를 생성하였다.

3. 하이브리드 텍스트 생성 결과 및 분포

본 연구는 정제된 9,047개의 자기소개서를 기반으로 표(1)와 같이 6가지 태스크 유형별로 균등하게 하이브리드 텍스트를 생성하였다. 표 (1)의 각 태스크는 서로 다른 경계 구조와 작성자 전환 양식, 경계 수, 데이터 셋을 포함하고 있다. 이와 같이 전체 데이터셋은 태스크별로 약 1,500건씩 균등하게 구성되었으며, 다양한 구조적 유형과 경계 수를 포함하고 있다. 이를 통해 모델은 단순한 이진 분류가 아닌, 실제 문서 작성에

표 1. 하이브리드 데이터셋(Composition of the Hybrid Dataset)

Task	Description	Hyper Test Structure	#Boundary	Number of Datasets
1	자기소개서 뒷부분 재작성	H -> M	1	1,509
2	자기소개서 앞부분 재작성	M -> H	2	1,509
3	자기소개서 중간 부분 재작성	H -> M -> H	2	1,508
4	자기소개서 앞뒤 부분을 재작성. 중간 문장만 원본 그대로 유지	M -> H -> M	2	1,507
5	Task 3 진행 후 뒷부분 일부 재작성	H -> M -> H -> M	3	1,507
6	Task 3 진행 후 앞부분 일부 재작성	M -> H -> M -> H	3	1,507

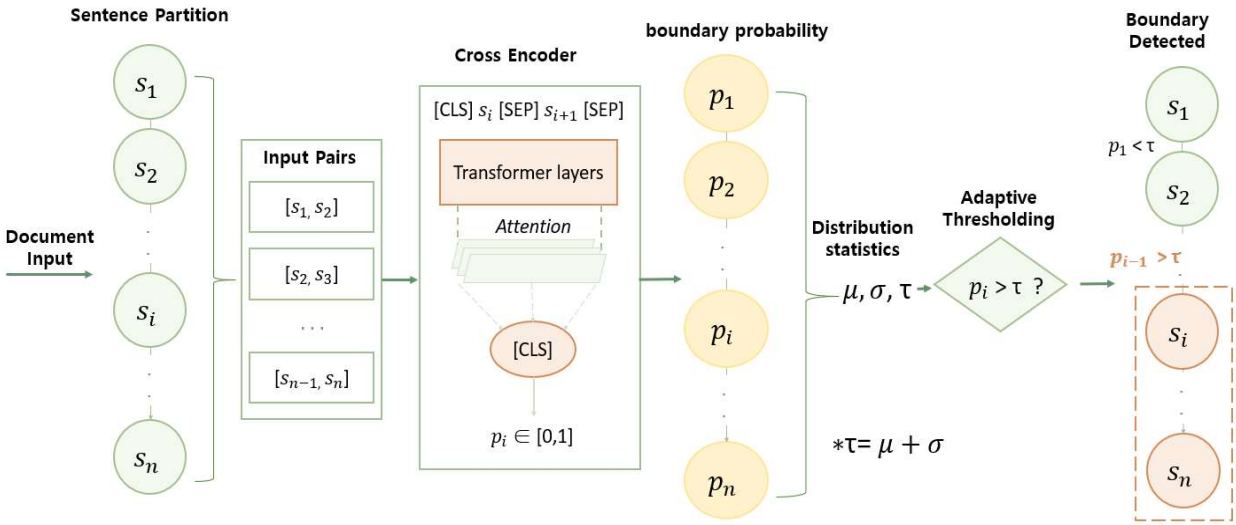

그림 2. 제안한 CrossEncoder 기반 경계 검출 방법(The proposed CrossEncoder-based boundary detection method)

서 발생할 수 있는 다양한 작성자 전환 양상을 학습하고 일반화할 수 있도록 설계하였다.

IV. 제안한 방법

1. Cross Encoder 기반 모델

기존 TriBERT의 한계를 극복하기 위해 본 연구에서는 그림 (2)와 같이 CrossEncoder 기반 경계 검출 방법을 제안한다. 즉, 문서 입력 단계에서 문장을 분할하고, 문장 쌍을 생성한 후 CrossEncoder에 입력하여 경계 확률을 산출하고, 마지막으로 Adaptive Threshold 모듈을 통해 최종 경계 여부를 판별한다.

CrossEncoder는 BERT와 같은 트랜스포머 기반 모델을 활용하여 두 입력 시퀀스 간의 관계를 직접 모델링한다. 본 연구에서는 인접한 문장 쌍 (S_i, S_{i+1})을 [CLS]S_i [SEP] S_{i+1} [SEP] 형태로 결합하여 모델에 입력한다. 이때 self-attention 메커니즘은 두 문장 간의 미묘한 상호작용을 포착할 수 있으며, 문체의 일관성, 어휘 선택 패턴의 변화, 문장 구조의 불연속성, 논리적 연결성의 단절 등 경계 탐지에 중요한 다양한 신호를 통합적으로 고려한다. 최종적으로 [CLS] 토큰의 임베딩을 기반으로 이진 분류를 수행하여 해당 지점이 경계인지 여부를 예측한다.

이 접근법의 장점은 실제 경계 정보를 직접적으로 활용한다는 점이다. 훈련 데이터에서 모든 인접 문장 쌍에 대해 경계 여부를 명시적으로 라벨링하고, 모델은 “이 두 문장 사이가 경계인가?”라는 질문에 직접 답하도록 학습된다. 이를 통해 학습 목표와 추론 목표 간 불일치를 제거하고, 예측의 정확도를 높일 수 있다. 또한 CrossEncoder는 문장 간 미세한 변화를 효과적으로 반영할 수 있다. 한국어의 경우 조사 변화, 어미 활용의 미묘한 차이, 높임법의 일관성 등이 중요한 단서가 되는데, 트랜스포머의 어텐션 메커니즘은 이러한 언어적 특징을 정교하게 학습할 수 있어 한국어 하이브리드 텍스트의 경계 탐지에 적합하다.

2. 분포 기반 Adaptive Threshold 설정

기존 Fixed Top-K 방식은 문서 내 실제 경계 수와 관계없이 항상 일정 개수(K)의 경계를 예측하도록 설계되어 있다. 이로 인해 실제 경계 수가 K보다 적을 경우 불필요한 지점을 오탐하거나, K보다 많을 경우 일부 경계를 놓치는 문제가 발생한다. 예를 들어, 실제 경계가 하나인 문서에서 Fixed Top-K(K=3)를 적용하면 상위 3

개의 점수를 모두 경계로 예측해야 하므로 Precision이 최대 1/3으로 제한되고, 결과적으로 F1-score도 구조적으로 0.5를 넘기 어렵다. 실제 실험에서도 Fixed Top-K를 적용한 경우 전체 샘플의 41.2%가 F1-score 0.5에 머무르는 등 실용성의 한계가 확인되었다.

이러한 문제를 해결하기 위해 본 연구에서는 Adaptive Threshold 방식을 제안한다. 이 방식은 문서별 예측 점수 분포를 활용하여, 문서 특성에 따라 임계값을 동적으로 설정한다. 임계값은 다음식(2)과 같이 정의된다.

$$Threadhold = \mu + \sigma \quad (2)$$

식(2)에서 μ는 해당 문서의 모든 인접 문장 쌍에 대한 모델의 예측 점수(경계 확률)의 평균이고, σ는 표준편차이다. 모델 출력이 이 임계값을 초과하는 경우에만 해당 지점을 경계로 판단한다.

예를 들어, 어떤 문서의 예측 점수가 [0.12, 0.85, 0.34, 0.29]일 경우, Fixed Top-K(K=3)는 0.85, 0.34, 0.29를 모두 경계로 탐지한다. 반면 Adaptive Threshold 방식에서는 평균 μ=0.40, 표준편차 σ=0.23를 계산하여 임계값 0.63를 설정하고, 0.85만을 경계로 예측한다. 이를 통해 불필요한 오탐을 줄이고 실제 상황을 더 정밀하게 반영할 수 있다.

V. 실험 및 결과

1. 실험 설계

본 연구는 자체 구축한 한국어 자기소개서 데이터셋을 중심으로 실험을 진행하였다. 더불어, 제안 방법론의 일반화 가능성을 검증하고 기존 연구 [6]의 TriBERT 모델과 비교하기 위하여, Dugan 외 [7]이 구축한 영어 하이브리드 에세이 데이터셋에도 동일한 방법론을 적용하였다. 해당 영어 에세이 데이터셋은 총 17,136개의 하이브리드 에세이로 구성되어 있으며, 인간 작성 부분과 AI 생성 부분의 경계 정보가 명확히 라벨링되어 있다. 이를 통해 한국어 환경에서의 성능 개선뿐 아니라 언어 독립적인 방법론의 유효성 또한 검증하고자 하였다.

2. 임베딩 모델 선정

TriBERT 원 논문에서는 영어 임베딩 모델인 all-mpnet-base-v2를 사용하여 우수한 성능을 보고하였다. 그러나 이를 한국어 데이터셋에 직접 적용한 초기 실험에서 F1-score는 0.271에 불과하였다. 이는 영어와 한국어의 언어적 특성 차이로 인해 영어 기반 임베딩 모델이 한국어 텍스트의 문맥적·의미적 특성을 충분히 반영하지 못했기 때문으로 해석된다.

이에 따라 한국어 특화 문장 임베딩 모델을 탐색하고 동일 조건에서 비교 실험을 수행하였다. KoSimCSE-BERT, ko-sbert-nli, KoBERT 등을 검증한 결과, KoSimCSE-BERT가 F1-score 0.468로 가장 우수한 성능을 보였다. 이는 영어 기반 임베딩 대비 약 73% 성능 향상으로, 경계 탐지 태스크에서 언어별 특화 모델의 중요성을

표2. 성능분석 비교(Comparative Analysis of Performance)

데이터 셋	방법	Precision	Recall	F1-score
영어	Fixed + TriBert	0.455	0.830	0.564
	Fixed + CrossEncoder	0.508	0.876	0.616
	Adaptive + TriBert	0.685	0.771	0.685
	Adaptive + CrossEncoder	0.750	0.778	**0.736**
한국어	Fixed + TriBert	0.380	0.697	0.468
	Fixed + CrossEncoder	0.509	0.819	0.597
	Adaptive + TriBert	0.668	0.745	0.664
	Adaptive + CrossEncoder	0.651	0.632	**0.641**

뚜렷하게 보여준다. 특히 KoSimCSE-BERT는 대조 학습(contrastive learning) 기반으로 학습되어 문장 간 의미적 유사도를 정밀하게 포착할 수 있어 경계 탐지 작업에 특히 효과적인 것으로 판단된다.

3. 실험결과 및 분석

표 (2)는 제안한 CrossEncoder 기반 방법론과 기존 Triplet Loss 기반 방법론을 Fixed Top-K 및 Adaptive Threshold 두 가지 평가 방식으로 적용하여 영어 및 한국어 데이터셋에서 비교한 결과이다. 본 연구에서 보고한 F1-score는 각 문서 단위에서 산출한 Precision과 Recall을 기반으로 F1-score를 계산한 뒤, 전체 문서에 대해 매크로 평균(macro-averaged F1-score)을 적용한 값이다. 따라서 표에 제시된 Precision 및 Recall 값의 단순 조화평균과는 직접적으로 일치하지 않을 수 있다. 이는 문서 간 성능 분포의 편차를 반영하기 위함이며, 결과의 재현성을 위해 산출 기준을 명시한다.

CrossEncoder 기반 모델은 Fixed Top-K 기준에서 Triplet Loss 대비 전반적으로 우수한 성능을 보였다. 영어 데이터셋의 경우 CrossEncoder는 F1-score 0.616을 기록하여 Triplet Loss(0.564) 대비 약 9% 향상된 결과를 보였다. 한국어 데이터셋에서는 그 차이가 더 크게 나타나, CrossEncoder가 0.597을 기록하며 Triplet Loss(0.468) 대비 약 27% 향상되었다. 이는 CrossEncoder의 어텐션 메커니즘이 한국어의 복잡한 문장 구조와 문체 변화를 보다 정교하게 포착할 수 있음을 시사한다.

또한 탐지 방식 변화의 효과도 두드러졌다. Adaptive Threshold 방식을 적용했을 때 모든 조건에서 성능이 크게 개선되었다. 영어 데이터셋에서 Triplet Loss는 F1-score가 0.564에서 0.685로 약 21% 향상되었고, CrossEncoder는 0.616에서 0.736으로 약 19% 개선되었다. 특히 Precision의 개선 폭이 컸는데, Triplet Loss의 경우 0.455에서 0.685로 약 51% 증가하였다. 이는 Adaptive Threshold 방식이 불필요한 오탐을 줄여 경계 탐지의 신뢰도를 높이는 핵심적인 개선임을 보여주며, 실제 응용 가능성을 크게 확장시킨다.

한편, 특정 조건에서는 예외적인 현상도 관찰되었다. 한국어 데이터셋에서 CrossEncoder가 Fixed Top-K 기준에서는 Triplet Loss보다 우수했지만, Adaptive Threshold 적용 시 Triplet Loss(F1-score 0.664)가 CrossEncoder(0.641)보다 근소하게 높은 성능을 보였다. 이는 CrossEncoder가 경계를 직접 분류하는 반면, Triplet Loss는 문장 간 유사도 학습을 통해 간접적으로 경계를 추론하기 때문에 데이터 분포 특성상 Adaptive Threshold와의 결합에서 더 유리하게 작용했을 가능성이 있다. 이러한 차이는 향후 데이터셋 확장 및 모델 구조 분석을 통해 보다 심층적으로 검증할 필요가 있다.

VI. 결론 10및 향후연구

본 연구는 한국어 자기소개서 내 인간 및 AI 생성 텍스트의 경계를 자동으로 탐지하는 것을 목표로 하였다. 이를 위해 한국어의 언어적 특성을 반영하고 6가지 하이브리드 패턴을 포함하는 9,047건의 자기소개서 데이터셋을 구축하였다.

제안한 CrossEncoder 기반 경계 탐지 모델은 기존 Triplet Loss 기반 접근법의 한계를 보완하며, Fixed Top-K 평가 기준에서 성능을 유의미하게 향상시켰다. 특히 한국어 데이터셋에서 F1-score 0.597을 달성하여 Triplet Loss(0.468) 대비 약 27% 개선된 결과를 보였다. 이는 문장 간 미묘한 언어적·문체적 변화를 직접 학습할 수 있는 CrossEncoder 구조가 경계 탐지에 효과적임을 보여준다. 또한 분포 기반 Adaptive Threshold 방식을 도입함으로써 Fixed Top-K의 구조적 한계를 극복하고, 탐지 정확도와 실용성을 동시에 향상시켰다.

본 연구는 다음과 같이 몇 가지 한계를 지닌다. 첫째, 하이브리드 텍스트 생성을 위해 GPT-4.1-mini 단일 모델만을 사용하였기 때문에 실제 환경에서 다양한 AI 모델이 혼재되는 상황을 충분히 반영하지 못하였다. 둘째, 자기소개서라는 특정 도메인과 문장 단위 경계 탐지에 국한되어 있어, 다른 장르나 더 세밀한 단위(예: 구나 어절)로의 일반화 가능성은 추가 검증이 필요하다. 셋째, 분포 기반 Adaptive Threshold 방식의 경우 문서 길이에 따라 임계값 산출에 편향이 발생할 가능성이 있으며, 이에 대한 체계적인 검증이 추가적으로 필요하다.

향후 연구에서는 다양한 AI 모델과 장르를 포함하는 데이터셋 확장을 통해 실제 환경을 보다 충실히 반영하고, 문단·구·어절 등 다양한 단위에서의 경계 탐지 기법을 개발 것이다. 또한 문서 길이에 따른 Adaptive Threshold의 민감도를 정량적으로 분석하고, 필요 시 길이 보정(length normalization) 기법을 도입하여 강건성을 강화할 예정이다.

REFERENCES

[1] 무하유, "2024년 AI 자기소개서 트렌드 리포트," https://www.etnews.com/20250225000303 (accessed Feb., 01, 2025).

[2] 고용노동부, 한국고용정보원, "2023년 하반기 기업 채용동향조사 결과," 고용노동부 보도자료, https://www.moel.go.kr/news/enews/report/enewsView.do?news_seq=16352 (accessed Dec., 01, 2023).

[3] C.S. Grecu and M.E. Breaban, "A Dual-Approach for AI-Generated Text Detection," *Proc. of the 26th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC),* pp. 206-214, IEEE, 2024.

[4] E. Mitchell, Y. Lee, A. Khazatsky, C.D. Manning, and C. Finn, "DetectGPT: zero-shot machine-generated text detection using probability curvature," *Proc. of the 40th International Conference on Machine Learning (ICML'23),* vol. 202, pp. 24950-24962, JMLR.org, 2023.

[5] X. Hu, P.-Y. Chen, and T.-Y. Ho, "RADAR: robust AI-text detection via adversarial learning," *Proc. of the 37th International Conference on Neural Information Processing Systems (NIPS '23),* pp. 15077-15095, Curran Associates Inc., 2023.

[6] Z. Zeng, L. Sha, Y. Li, K. Yang, D. Gašević, and G. Chen, "Towards automatic boundary detection for human-AI collaborative hybrid essay in education," *Proc. of the AAAI Conference on Artificial Intelligence,* vol. 38, no. 20, pp. 22502-22510, 2024.

[7] L. Dugan, D. Ippolito, A. Kirubarajan, S. Shi, and C. Callison-Burch, "Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text," *Proc. of the AAAI Conference on Artificial Intelligence,* vol. 37, no. 11, pp. 12763-12771, 2023.

저 자 소 개

이도현(준회원)

2025년 전남대학교 인공지능학부 학사 졸업.

<주관심분야 : 자연어처리, 머신러닝, 딥러닝, 소프트웨어공학, 대형언어모델>

김미수(정회원)

2013년 고려대학교 경영학부 학사 졸업.

2021년 성균관대학교 전자전기컴퓨터공학과 박사 졸업.

현재 전남대학교 인공지능학부 조교수

<주관심분야 : 소프트웨어 디버깅 자동화, 자연어처리>

손남례(정회원)

2005년 전남대 전산학과 박사 졸업.

2011년~2017년 한국전자통신연구원 연구원

2017년~2021년 호남대 정보통신공학과 교수

2021년~현재 전남대 소프트웨어중심대학사업단 교수

<주관심분야:빅데이터분석솔루션, 전력IT, 딥러닝 등>