역번역을 활용한 AI생성 텍스트 탐지 회피 전략에 관한 실험적 연구

임승우, 김지희
동국대학교 컴퓨터AI학부
Department of Computer Science and Artificial Intelligence. Dongguk University

daniel73919@gmail.com , jihie.kim@dgu.edu

목 차

요 약

대규모 언어 모델(LLM)이 발전하면서 AI 생성 텍스트를 구분하는 기술이 중요해졌지만, '역번역(Back-Translation)'과 같은 회피 기술은 탐지 모델의 신뢰도를 크게 위협한다. 본 연구는 역번역 공격의 성공률을 좌우하는 번역 엔진, 피봇 언어의 문법적 거리, 번역 경로의 구조적 복잡성 등 핵심 변수를 체계적으로 분석하는 후속 실험을 진행하였다. 공개 데이터셋인 ESPERANTO와 RoBERTa 기반 탐지 모델을 사용한 결과, 표준 품질의 번역기(Google Translate)가 고품질 번역기(DeepL)보다 탐지 회피에 훨씬 효과적임을 확인했다. 특히, 한국어처럼 문법 구조가 상이한 언어로 역번역했을 때, 오히려 AI 탐지 점수가 증가하는 '적대적 정제(Adversarial Purification)'라는 특이 현상을 발견하였다. 또한, 공격의 성공 여부는 번역 경로의 복잡성이나 반복 횟수가 아닌, 가장 강력한 첫 변환 단계에서 결정된다는 사실을 확인했다. 본 연구의 결과는 현재 탐지 모델의 취약점을 명확히 보여주는 한편, ESPERANTO 논문에서 제안된 MESAS와 같이 역번역 공격에 강건한 모델의 필요성을 다시 한번 강조하며, 향후 발전된 탐지 모델 개발을 위한 기초 자료를 제공한다.

1. 서 론

대규모 언어 모델(LLM)의 빠른 발전으로 AI 가 만드는 텍스트의 품질은 사람이 쓴 글과 구분하기 어려울 정도가 되었다. 이로 인해 학계의 정직성을 지키고 가짜 뉴스를 막기 위한 AI 텍스트 탐지 기술이 주목받고 있지만, 탐지 시스템을 속이려는 회피 기술 또한 계속해서 발전하고 있다. 그중 '역번역'은 AI 가 쓴 글을 다른 언어로 번역했다가 다시 원래 언어로 되돌리는 간단한 방법만으로 텍스트의 통계적 특징을 바꿔 탐지를 무력화하는 효과적인 공격 방법이다.

선행 연구인 ESPERANTO 는 이러한 역번역 공격의 유효성을 입증하는 동시에, 역번역된 텍스트에 대해 탐지율 감소가 1.85%에 불과한 강력한 방어 모델 MESAS(Modified ESAS)를 대응책으로 제시한 바 있다. 하지만 ESPERANTO 연구는 방어 모델 제시에 집중한 반면, 공격 자체를 구성하는 다양한 변수(번역 엔진, 언어적 특성, 경로 복잡성 등)가 회피율에 미치는 영향에 대해서는 깊게 다루지 않았다. 본 연구는 바로 이 지점에 착안하여, 역번역 공격의 메커니즘을 더 깊이 이해하고 현재 탐지 모델들이 가진 구체적인 취약점을 규명함으로써, 향후 MESAS와 같이 더 강건한 탐지 모델을 개발하는 데 기여하고자 한다

1.1. 데이터셋 및 탐지 모델

실험에는 공개 데이터셋인 ESPERANTO [1]의 뉴스 기사 샘플을 사용했으며, 탐지 모델로는 Hugging Face에 공개된 RoBERTa 기반의 roberta-base-openai-detector 를 활용하였다. 탐지 점수는 0과 1 사이의 확률값으로, 1에 가까울수록 AI 생성 텍스트로 판단함을 의미한다.

1.2. 독립 변인 (Independent Variables)

본 실험에서는 세 가지 핵심 독립 변인을 설정하여 탐지 회피율에 미치는 영향을 측정했다.

가. 번역 엔진 품질: 번역기 품질에 따른 차이를 보기 위해 고품질 번역기(DeepL), 표준 품질 번역기(Google Translate), 그리고 오픈소스 번역기(LibreTranslate)를 사용했다.

나. 피봇 언어의 계통적 거리: 영어와 문법 구조가 비슷한 인도유럽어족의 독일어(DE)와, 구조가 완전히 다른 고립어인 한국어(KO)를 피봇 언어로 사용하여 언어 간 거리가 미치는 영향을 확인했다.

다. 번역 경로의 구조: 번역 과정의 구조적 차이를 분석하기 위해 두 가지 하위 실험을 진행했다.

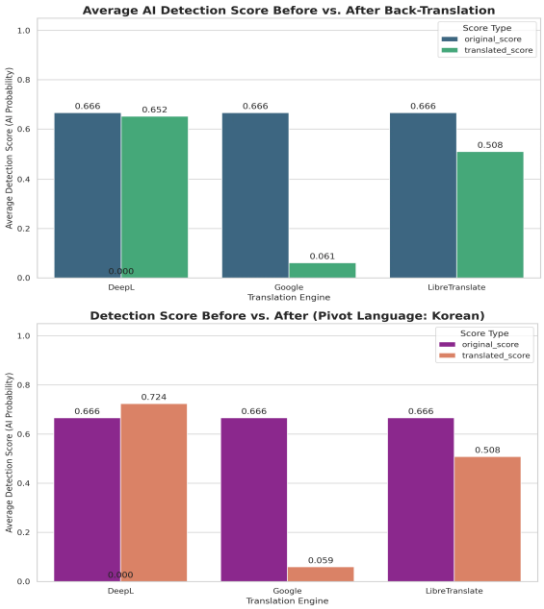
1. 반복 횟수(Iteration Depth): 단일 피봇 언어(KO)를 1 회, 3 회, 5 회 반복하여 번역 횟수가 미치는 영향을 분석했다.

2. 경로 복잡성(Path Complexity): 단일 언어를 거치는 경로와, 계통적으로 가깝거나 먼 여러 언어를 순차적으로 거치는 '연쇄 번역' 경로의 회피 성능을 비교했다.

3. 실험 결과 및 분석

3.1. 번역 엔진 및 언어 거리에 따른 효과

첫 번째 실험에서는 피봇 언어로 독일어(DE)를 사용하여 번역 엔진별 성능을 비교했다. 그 결과, 표준 품질의 Google Translate 가 탐지 점수를 90.8% 감소시키며 가장 높은 회피 성능을 보였다.



(그림 1,2) 피봇 언어 독일어(DE) 사용 시 엔진별 탐지 점수 변화, 한국어 사용시 변화

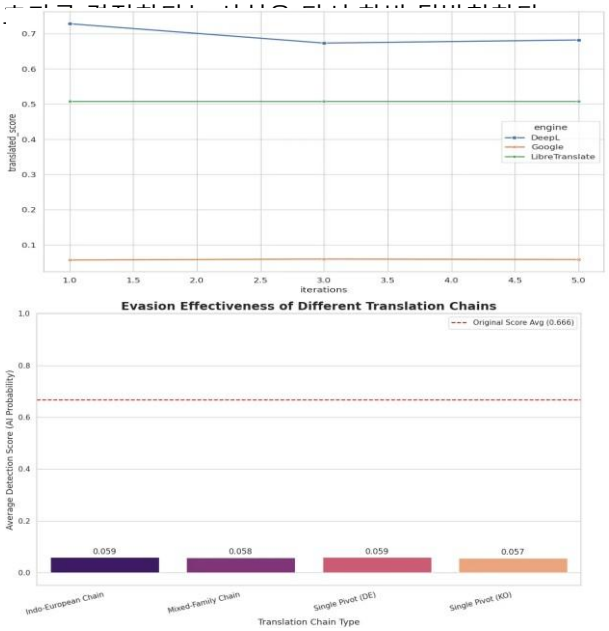
두 번째 실험에서는 피봇 언어를 계통적으로 먼 한국어(KO)로 변경하였다. 그 결과는 (그림 1)과 같이 매우 직관에 반하는 양상을 보였다. Google Translate 는 여전히 강력한 회피 성능을 보였으나, 고품질 엔진인 DeepL 은 오히려 탐지 점수가 **8.7% 증가**하는 현상이 발생했다. 이는 DeepL 의 완벽에 가까운 번역 과정이 AI 텍스트의 기계적 특징을 오히려 더 강화시키는 **'적대적 정제(Adversarial Purification)'** 현상으로 분석된다. 이 결과는 번역 품질이 높다고 해서 회피 성능이 비례하여 증가하지 않음을 시사한다.

3.2. 번역 경로 구조에 따른 효과

번역 경로의 구조(반복 횟수, 경로 복잡성)가 미치는 영향을 분석한 결과, 두 실험 모두에서 '최초 단계의 우위성(Primacy of the First Step)'이라는 일관된 결론이 도출되었다.

(그림 3)은 한국어를 피봇 언어로 하여 번역을 1, 3, 5 회 반복했을 때의 탐지 점수 변화를 보여준다. 가장 효과적인 Google Translate 의 경우, 첫 번째 반복에서 이미 탐지 점수가 0.059 로 바닥에 도달했으며 추가적인 반복은 거의 영향을 주지 못했다. 이는 공격의 효과가 첫 단계에 집중되어 있으며, 반복 횟수를 늘리는 것은 유의미한 이점을 제공하지 못함을 의미한다.

또한, 여러 언어를 순차적으로 거치는 '연쇄 번역'과 단일 언어를 거치는 경로의 성능을 비교한 결과, (그림 4)과 같이 모든 경로에서 탐지 점수는 약 0.058 수준으로 거의 동일하게 나타났다. 이는 경로의 복잡성이나 길이가 아닌, 경로에 포함된 가장 강력한 단일 변환 단계가 전체 공격의



(그림 3) 번역 반복 횟수에 따른 엔진별 탐지 점수 변화 & (그림 4) 번역 경로 복잡성에 따른 탐지 점수 비교

4. 결론

연구 결과, (1) 표준 품질의 번역 엔진이 가장 효과적인 회피 도구이며, (2) 고품질 엔진은 계통적으로 먼 언어와 함께 사용될 때 오히려 텍스트를 더 탐지하기 쉽게 만드는 '적대적 정제' 현상을 유발할 수 있음을 발견했다.
(3) 또한, 공격의 성공은 번역 경로의 복잡성이나 반복 횟수가 아닌, 가장 강력한 단일 변환 단계에서 대부분 결정된다는 사실을 입증했다. 이러한 발견은 현재의 탐지 모델들이 기계 번역 과정에서 발생하는 통계적 섭동(perturbation)에 매우 취약하다는 점을 명확히 보여준다. 향후 연구에서는 더 다양한 탐지 모델을 대상으로 본 실험을 확장하고, 회피 과정에서의 의미 보존도를 정량적으로 측정하여 공격의 실효성을 종합적으로 평가할 필요가 있다

참고 문헌

[1] Ayoobi, N., et al., "ESPERANTO: A Benchmark for Evaluating Modern Detectors of AI-Generated Text via Multi-Language Back-Translation