외국어 생성 차단을 위한 LLM 디코딩 제어 기법

이하은, 이건희*
에이치디씨랩스

{haeun.lee, Gunhee_Lee}@hdc-labs.com

LLM Decoding Control Method for Blocking Foreign Language Generation

Haeun Lee, Gunhee Lee*
HDC LABS.

요 약

본 연구는 대형 언어 모델(LLM) 기반 한국어 서비스 환경에서 빈번하게 발생하는 외국어 토큰 생성 문제를 완화하기 위해, 디코딩 단계에서 외국어 생성을 제어하는 기법을 제안한다. 제안하는 방법은 LogitsProcessor 모듈을 활용하여 디코딩 과정에서 외국어 토큰 후보를 사전에 제거함으로써, 외국어 토큰의 최초 생성을 차단하도록 설계되었다. 외국어 판별은 Unicode 범위 기반 방식으로 수행되며, 다양한 외국어 문자 범위를 포괄적으로 고려한다. Qwen3-32B 모델을 대상으로 한 실험 결과, 제안 기법은 외국어 토큰 혼입률을 약 79% 감소시켰으며, 연쇄적 외국어 생성을 효과적으로 억제하는 것으로 나타났다. 이를 통해 본 연구는 모델 재학습이나 추가 데이터 없이도 한국어 기반 LLM 서비스에서 외국어 혼입 문제를 완화할 수 있는 실용적인 디코딩 제어 방안을 제시한다.

Ⅰ. 서 론

최근 대형 언어 모델(Large Language Model, LLM)은 다국어 코퍼스를 기반으로 학습되어 단일 언어 환경에서도 다양한 외국어 토큰을 생성할 수 있는 능력을 보인다. 이러한 특성은 범용성 측면에서는 장점으로 작용하나, 특정 언어 기반 서비스 환경에서는 의도하지 않은 외국어 출력으로 인해 사용자 경험과 신뢰도를 저해하는 요인이 된다. 특히 한국어 기반 서비스에서는 중국어, 일본어 등의 외국어가 응답에 혼입되는 현상이 품질 문제로 직결될 수 있다.

기존 연구에서는 프롬프트 엔지니어링을 통해 한국어 사용을 지시하거나[1], 생성된 출력에서 외국어를 사후 제거하는 후처리 방식을 활용해 왔다[2]. 그러나 이러한 접근법은 모델 내부의 토큰 선택 확률을 직접 제어하지 못하며, 외국어 토큰이 이미 생성된 이후에 대응한다는 구조적 한계를 가진다. 또한 파인튜닝이나 강화학습 기반 방법[3][4]은 추가적인 학습 데이터와 계산 비용을 요구한다는 부담이 있다.

본 연구에서는 특히 Qwen 계열 언어 모델에서 중국어 생성 문제가 상대적으로 두드러진다는 점에 주목한다. Qwen 계열 모델은 중국어 데이터를 대규모로 포함한 학습 특성으로 인해 한국어 질의에 대해서도 중국어 한자나 발음 표기(pinyin)를 확률적으로 생성하는 경향을 보인다. 또한 LLM 의 확률적 생성 특성상 한 번 외국어 토큰이 생성되면 이후 토큰들도 동일 언어 분포를 따라 연쇄적으로 생성되는 경향이 뚜렷하다. 즉 초기 한두 개의 외국어 토큰이 생성되는 순간, 이후 디코딩 과정에서 해당 언어 토큰의 조건부 확률이 급격히 상승하며 외국어 출력이 지속되는 문제가 발생한다. 이는 외국어 출력을 사후적으로 제거하는 방식이 근본적인 해결책이 되기 어려우며 외국어 토큰의 최초 생성을 사전에 차단하는 디코딩 단계 제어의 중요성을 시사한다.

이에 본 논문에서는 외국어 생성 문제를 출력 후 정제 문제가 아닌 디코딩 과정에서의 토큰 선택 제어 문제로 재정의하고, LogitsProcessor 를 활용한 디코딩 단계 외국어 차단 기법을 제안한다. 제안하는 방법은 Unicode 범위 기반 외국어 판별을 통해 외국어 토큰 후보를 사전에 제거함으로써, 모델이 외국어 토큰으로 확률 질량을 이동시키기 이전에 생성을 억제하도록 설계되었다. Qwen3-32B 모델을 대상으로 한 실험 결과, 제안한 기법은 중국어 한자 및 발음 기호의 직접적인 출력을 현저히 감소시켰으며 응답 전반이 한국어 중심으로 유지되는 효과를 보였다.

본 연구의 주요 기여는 다음과 같다.

1. Qwen 계열 모델의 외국어 생성 특성과 연쇄적 생성 패턴을 분석하고, 디코딩 단계 제어의 필요성을 실증적으로 제시한다.
2. LogitsProcessor 를 활용한 디코딩 단계 외국어 차단 기법을 제안하며, 모델 재학습 없이 적용 가능한 실용적 해결책을 제시한다.
3. 한국어 기반 프롬프트에 대한 실험을 통해 제안 기법이 외국어 혼입을 효과적으로 억제함을 입증한다.

Ⅱ. 본론

1. 실험 환경 및 데이터 구성

본 연구에서는 텍스트 전용 대규모 언어 모델인 Qwen3 32B 를 실험 대상으로 사용하였다. Qwen3 는 다국어 데이터로 학습된 모델로, 특히 중국어와 영어에 강점을 보이며 한국어 질의에 대해서도 중국어 토큰을 혼합 생성하는 경향이 관찰된다.

외국어 차단 효과를 검증하기 위해, 한국어로 작성되었으나 외국어 관련 설명을 요구하는 프롬프트를 구성하였다. 예를 들어 중국어 표현, 중국의 언어 환경, 중국어 단어의 의미 등을 묻는 질의를 사용하였으며, 이는 모델이 설명 과정에서 중국어 예시나 한자를 직접 출력할 가능성이 높아 차단 로직의 효과를 관찰하기에 적합하다.

2. 제안 기법: 디코딩 단계 외국어 차단

제안하는 프레임워크는 LLM 의 디코딩 과정에서 외국어 토큰의 생성을 사전에 차단하여 한국어 중심의 응답을 유도한다. 주요 구성 요소는 다음과 같다.

(1) Unicode 기반 외국어 토큰 판별 모듈

외국어 판별은 Unicode 범위 기반 방식을 사용하였다. 중국어 한자(CJK Unified Ideographs 및 확장 영역), 일본어, 키릴 문자, 태국 문자, 그리고 발음기호(pinyin, IPA, 결합 악센트 등)에 해당하는 Unicode 범위를 사전에 정의하고, 토큰 디코딩 결과에 해당 문자가 포함될 경우 외국어 토큰으로 분류하였다. 특히 Qwen 계열 모델의 토크나이저는 서브워드 단위로 텍스트를 분할하므로, 각 토큰 후보를 디코딩하여 실제 문자열을 확인한 후 외국어 여부를 판정하였다.

(2) LogitsProcessor 기반 확률 제어 모듈

디코딩은 샘플링 기반 생성 방식을 사용하였으며, 외국어 차단 로직은 HuggingFace 의 LogitsProcessor 인터페이스를 활용하여 구현하였다. LogitsProcessor 는 각 디코딩 스텝에서 다음 토큰의 확률 분포(logits)를 조작할 수 있는 인터페이스로, 본 연구에서는 이를 통해 외국어 토큰의 logits 값을 $-\infty$로 설정하여 디코딩 과정에서 선택되지 않도록 제어하였다. 이 방식은 외국어 토큰이 실제로 생성되기 이전 단계에서 후보를 제거함으로써, 외국어 생성의 연쇄적 확산을 방지하는 데 기여한다.

3. 성능 평가

제안한 기법을 실제 한국어 질의 환경에 적용하여 성능을 검증하였다. 외국어 차단 로직을 적용하지 않은 경우와 적용한 경우를 비교하여, 외국어 토큰 생성 빈도와 응답 품질의 변화를 측정하였다.

외국어 차단 로직을 적용하지 않은 경우, 모든 프롬프트에서 중국어 한자와 발음 표기(pinyin)가 빈번하게 혼합된 응답이 관찰되었다. 특히 설명형 프롬프트에서는 한국어 문장 중간에 외국어 토큰이 삽입되며, 한 번 외국어 토큰이 생성된 이후에는 해당 언어로 응답이 지속되는 경향이 두드러졌다. 이러한 결과는 Qwen 계열 모델이 중국어 토큰에 대해 상대적으로 높은 사전 확률을 갖고 있으며, 디코딩 과정에서 초기 외국어 토큰이 생성될 경우 이후 토큰 선택에 강한 영향을 미친다는 점을 보여준다.

반면 제안한 외국어 차단 로직을 적용한 경우, 중국어 한자 및 발음 기호의 출력 빈도는 현저히 감소하였다. 외국어 토큰이 디코딩 후보 단계에서 제거됨에 따라, 모델은 한국어 토큰 중심으로 응답을 구성하였으며 응답 전반의 언어 일관성이 크게 향상되었다. 특히 외국어 토큰의 최초 생성을 차단함으로써 이후 디코딩 단계에서 외국어로 확률 질량이 이동하는 현상이 효과적으로 억제됨을 확인하였다.

정량적 분석을 위해 총 30 개의 한국어 프롬프트에 대해 5 회씩 샘플링하여 총 150 개의 응답을 생성하였다. 그 결과, 외국어 차단 로직 미적용 시 외국어 혼입률은 평균 23.4%였으나, 적용 후에는 4.9%로 약 79% 감소하였다. 또한 외국어 토큰이 연속적으로 생성되는 연쇄 생성률은 12.7%에서 0.2%로 크게 감소하였으며, 응답당 평균 외국어 토큰 수 역시 3.8 개에서 0.3 개로 줄어들었다. 이러한 결과는 제안한 기법이 외국어 토큰의 생성 빈도뿐 아니라 연쇄적 확산까지 효과적으로 억제함을 보여준다.

표1. 외국어 토큰 생성 억제 효과 비교

	차단 로직 미적용	차단 로직 적용
외국어 혼입률 (%)	23.4	4.9
연쇄 생성률 (%)	12.7	0.2
평균 외국어 토큰 수	3.8	0.3

Ⅲ. 결론

본 논문에서는 Qwen3-32B 모델을 대상으로 외국어 생성 문제를 디코딩 단계에서 제어하는 기법을 제안하였다. Qwen 계열 모델에서 두드러지는 중국어 생성 경향과 외국어 생성의 연쇄적 특성을 분석하고, 외국어 토큰의 최초 생성을 차단하는 디코딩 제어가 효과적인 해결책임을 실험적으로 확인하였다.

실험 결과 제안한 기법은 외국어 토큰 혼입률을 약 79% 감소시키고 연쇄적 외국어 생성을 유의미하게 억제하는 것으로 나타났다. 본 기법은 모델 재학습이나 추가 데이터 없이 적용 가능하다는 점에서, 한국어 기반 LLM 서비스 환경에서 외국어 혼입으로 인한 사용자 경험 저하를 완화하는 실용적인 대안이 될 수 있다.

향후 연구에서는 토큰 결합 기반 언어 판별 기법의 고도화와 맥락 기반 선택적 차단 로직을 통해 보다 정교한 디코딩 제어가 가능할 것으로 기대된다. 또한 다양한 LLM 아키텍처와 언어 환경에 대한 확장 연구를 통해 제안 기법의 범용성을 검증하고자 한다.

참 고 문 헌

[1] Mann, Ben, et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 1.3 (2020): 3.

[2] Qiu, Ruichen, et al. "A Survey on Unlearning in Large Language Models." arXiv preprint arXiv:2510.25117 (2025).

[3] Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in neural information processing systems 35 (2022): 27730-27744.

[4] Bai, Yuntao, et al. "Constitutional ai: Harmlessness from ai feedback." arXiv preprint arXiv:2212.08073 (2022).