NAVER CONNECT | boostcamp AI Tech

NLP 기초 프로젝트

주제 분류 프로젝트
(Data-Centric)

NLP 15조

이정민(팀장), 김진재, 박규태, 윤선웅, 임한택

1. 개요

이 문서는 2024년 10월 28일 (월)부터 2024년 11월 7일 (목)까지 진행한 Data-Centric 프로젝트에서 NLP 15조가 수행한 결과의 Wrap-Up 보고서이다.

1.1. 개인별 역할

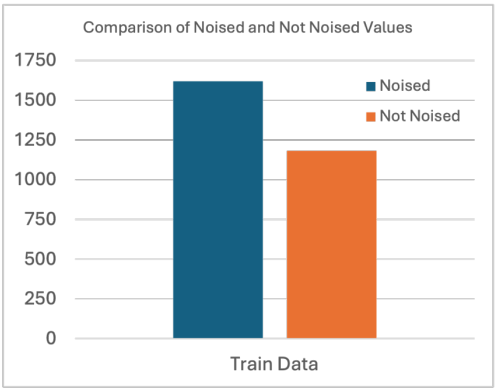
팀 내에서 개인별로 다음과 같은 역할을 담당하였다.

팀원	역할
김진재	클러스터링, 데이터 증강, LLM 오라벨 수정
박규태	EDA, 데이터 노이즈 제거, 데이터 증강, LLM 오라벨 수정
윤선웅	클러스터링, 데이터 증강, LLM 오라벨 수정
이정민	클러스터링, 데이터 노이즈 제거, 명사 추출 제거
임한택	데이터 노이즈 제거, 데이터 증강, 데이터 역번역

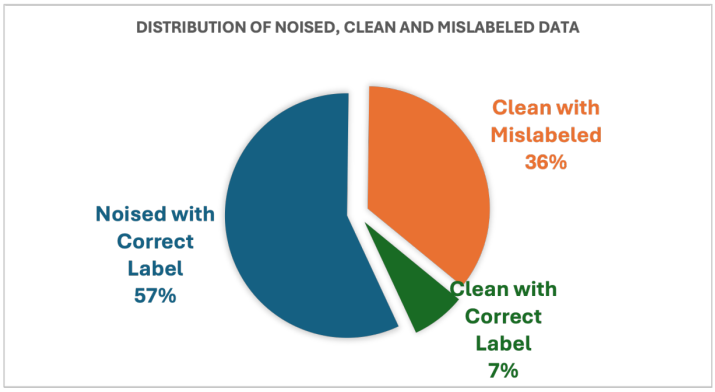
2. 데이터 분석

2.1. 노이즈, 비노이즈 데이터 개수 분석

원본 데이터셋 분석 결과 데이터셋 총 개수는 2800개, 노이즈 데이터 개수는 1600개 (전체의 57.14%), 비노이즈 데이터 개수는 1200개 (전체의 42.86%)로 파악하였다. 대회에서 제공된 정보에 따르면 노이즈가 낀 데이터는 모두 라벨링이 정상적으로 되어있고, 노이즈가 없는 데이터는 200개만을 제외한 나머지가 오라벨되어 있다. 데이터셋의 절반 이상이 노이즈를 포함하고 있어 학습 시 모델의 성능 저하를 유발할 가능성이 높은 구조였다. 따라서 신중한 노이즈 제거와 정제 작업이 우선시되어야 한다고 판단하였다.



< 학습 데이터 내 노이즈 여부 비교 >



< 노이즈, 오라벨, 정상 데이터 비율 >

3. 노이즈 데이터 처리

3.1 LLM을 통한 노이즈 제거

1600개의 Noise 데이터는 오라벨되어 있지 않기에 이를 활용하기 위해 Noise 선별 작업 수행을 하였다. 선별 작업에는 LLM을 활용하였는데, LGAI-EXAONE/EXAONE-3.0-7.8B-Instruct 모델에 Prompt Engineering을 거친 질문에서 생성된 답변을 활용하였다. 최종적으로 대회에서는 두 가지의 Noise 분류 버전이 활용되었다.

- ◆ 2800개 Train Data 중 1689개 Noise로 분류한 Version 1 데이터셋 (Temperature 0.9)
- ◆ 2800개 Train Data 중 1619개 Noise로 분류한 Version 2 데이터셋 (Temperature 0.1)

두 버전을 비교해보면 Version 2가 Version 1에 비해 더 간결한 프롬프트를 사용하였고, Temperature를 0.1로 감소시켜 모델이 더 정확한 판정을 하도록 유도하였다. 이를 통해 70개 가량의 오라벨을 추가적으로 찾을 수 있었다. 2.1.에서 언급한 대회 규칙에 의해 노이즈 데이터의 추정치는 1600개에 가까울수록 필터링 성능이 좋다고 판단하였다. 노이즈로 판별된 문장은 이후, 원래 문장으로 복원할 것을 요청하는 프롬프트를 통해 노이즈를 제거하였다.

3.2. LLM 모델 선정

이번 대회에서 주로 사용된 LLM과 선정 이유는 아래 표와 같다.

모델명	선정 이유
LGAI-EXAONE/EXAONE-3.0-7.8B-Instruct	KoMT 벤치마크에서 높은 점수 기록
CohereForAI/aya-expanse-8b	m-Arenahard 벤치마크에서 높은 Win-Rate 달성
rtzr/ko-gemma-2-9b-it	Horangi Leaderboard에서 높은 AVG 점수 기록

3.3. 명사 기반 중복 데이터 제거

데이터 내 증강된 텍스트에 포함된 명사가 반복되는 경우가 많아 불필요한 중복 데이터를 줄이기 위해 Okt 라이브러리를 활용하여 각 문장에서 주요 명사를 추출하였다. 이를 통해 데이터의 중복성을 줄이고 독립성을 강화하여 모델의 성능을 높일 수 있었다.

모든 데이터셋의 명사를 추출하여, 명사가 총 등장한 횟수, 등장한 고유한 라벨의 개수, 각 라벨 당 등장한 횟수를 3개의 Factor로 고려하였다. 특정 명사 토큰이 여러 번 존재하거나, 정확성이 낮은 토큰이 다수 포함된 데이터를 제거하고 정확성이 높은 토큰이 있는 데이터를 남기는 등의 전반적인 데이터 품질 개선 실험을 시도하였다. 실험 결과로 불확실한 명사 토큰이 1회 이상 들어있는 Text를 제거하는 것이 가장 큰 효과를 보였다. 중복 데이터를 제거한 결과, 성능 저하는 미미하지만 데이터셋 양을 크게 감소시켰다. 이를 통해 효과적인 중복 제거의 가능성을 확인하였다.

	Before	After
Total datasets	15,588	10,411 (-34%)
F1 Score (Public)	0.8515	0.8492 (-0.2%P)

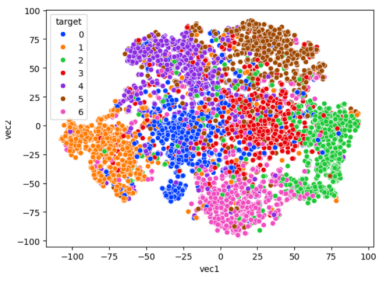
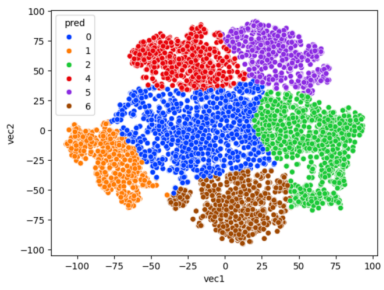
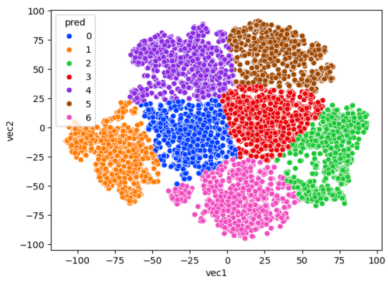
4. 오라벨 데이터 처리

4.1. Clustering

원본 데이터셋에서 Version 2를 기준으로 1619개가 아닌 나머지 1181개를 오라벨 데이터로 추정했고 각각 GMM(Gaussian Mixture Model), K-Means로 라벨링을 다시 진행했다. 당시 SOTA 데이터셋으로 klue/bert-base를 Fine-tuning하여 만든 ssunbear/bert-base-finetuned-ynat(F1 : 0.8315)을 임베딩 모델로 사용하였다. Clustering 결과 F1은 0.7806에서 0.8143로 향상했다. 아래 표는 오라벨 데이터의 Clustering 결과를 GMM과 K-Means로 시각화하였다.

4.2. LLM을 이용한 증강

교정된 오라벨 데이터 1181개를 각각 EXAONE-3.0-7.8B-Instruct와 aya-expanse-8b로 프롬프팅을 통해 총 3543개로 증강하였고 해당 F1 점수는 0.8143에서 0.8283로 향상했다. 아래는 데이터셋 프롬프팅 Instruction 예시와 모델을 통해 증강된 데이터의 예시이다.

오라벨 데이터 분포	GMM (Gaussian Mixture Model)	K-Means Clustering
		
원본 F1: 0.7806	적용 X 라벨 7개 -> 라벨 6개	적용 O F1: 0.8143

< 클러스터링 방법론별 시각화 >

[Instruction]
다음 뉴스 기사 제목과 동일한 카테고리에서 나올 기사 제목 1개를 생성 하시오.
답변은 오직 제목 1개 만을 포함해야 하며 불가능한 경우는 없다.

갤노트8 주말 27만대 개통…시장은 불법 보조금 얼룩

[LGAI-EXAONE/EXAONE-3.0-7.8B-Instruct]
갤럭시 S23 5G 출시 첫 주말, 15만대 이상 판매… 사전예약 초과 속 재고 부족 우려"

[CohereForAI/aya-expanse-8b]
아이폰X 출시 후 갤노트8 판매량과 불법 보조금 경쟁 심화

< LLM을 이용한 데이터 증강 예시 >

5. 역번역

LLM 증강 시 생성된 토큰의 표현이 제한적이라는 판단 하에서 토큰의 다양성을 높이고자 역번역을 수행하였다. 이때, 기본 역번역인 한국어-영어-한국어 순으로 역번역을 수행하여 단어의 다양성을 확보하였다. 순환 역번역과 피봇 역번역 기법도 검토하였으나, DeepL Free API의 사용량 제한으로 인해 기본 역번역을 채택하게 되었다. 역번역 결과 기존의 비슷한 LLM의 토큰을 어려운 토큰으로 대체하여 Context를 깊게 학습함으로써, F1 점수를 0.07%P 향상시킬 수 있었다.

한국어
"부동산 소유권 사칭, 30억원 챙긴 빌런 잡혀"

↓ EXAONE-3.0-7.8B-Instruct

English
"Real Estate Title Fraud: Criminals Busted for 3 Billion in Gains"

↓ DeepL Free API

한국어
부동산 소유권 사기: 30억 달러의 이득을 챙긴 범죄자 적발

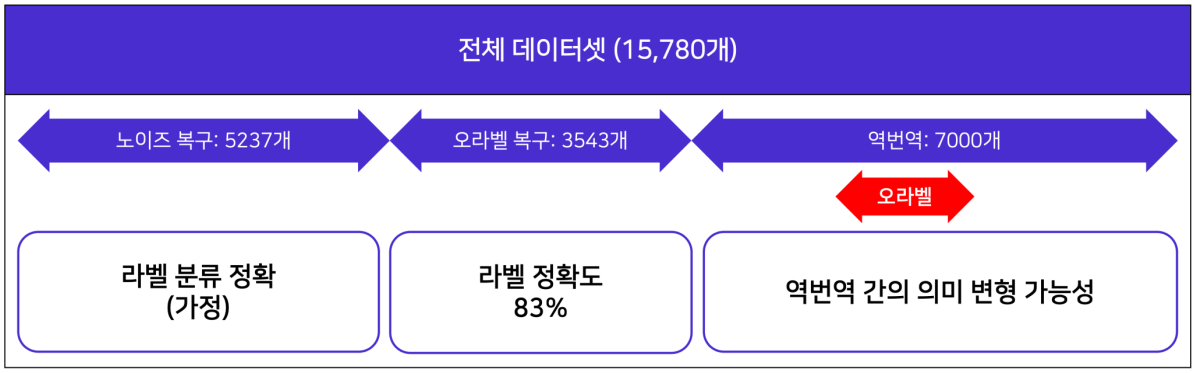
< 역번역 과정 도식 >

Public	Before	After
Accuracy Score	0.8607	0.8611
F1 Score	0.8564	0.8571 (+0.07%P)

< 역번역 결과 >

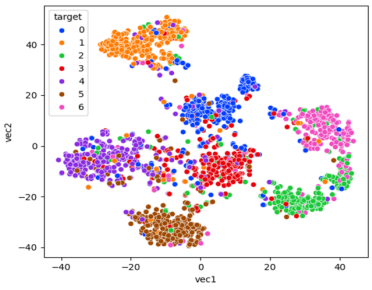
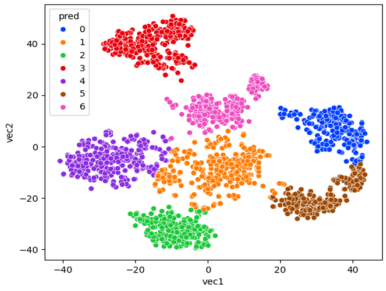
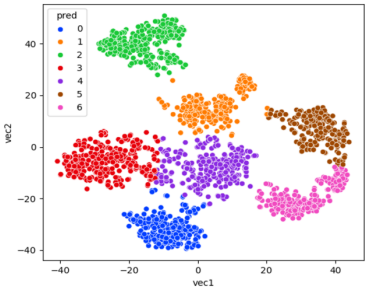
6. Reclustering

클러스터링을 기반으로 F1 0.8571점을 달성하였고, 역번역 7000개 데이터 셋에서 역번역 간의 의미 변형 가능성을 고려하여 reclustering하였다. 15,780개 데이터 셋으로 klue/bert-base를 직접 finetuning하여 만든 ssunbear/bert-base-finetuned-ynat-v2(F1: 0.8571)를 임베딩 모델로 활용하였다.



< 최종 활용 데이터셋 도식 >

클러스터링 모델로 역번역 데이터셋 1000~2000개 단위로 여러번 오라벨 탐지를 진행하였다. 가장 많은 오라벨을 배출한 10000-12000 행에서 448개를 보정하였고, F1 점수 0.8584의 SOTA를 달성하였다. 아래는 GMM과 K-Means를 활용하여 오라벨 된 448개의 데이터 셋의 reclustering을 시각화한 표다.

오라벨 데이터 분포	GMM (Gaussian Mixture Model)	K-Means Clustering
		
역번역 데이터셋의 오라벨 분포 확인	435개 변형, 적용 X	448개 변형, 적용 O F1: 0.8584

< Reclustering 결과 시각화 >

7. 결과

Data-Centric 프로젝트 결과 최종 1등으로 Private 점수 기준 Accuracy: 0.8623, F1: 0.8584을 SOTA로 기록하였다. 이러한 성과는 주로 프롬프트 기반 노이즈 제거 및 증강, 중복 단어 제거, 클러스터링, 역번역 등 다양한 데이터 중심 기법을 성공적으로 적용한 결과라고 생각한다.

8. 팀 회고

본 프로젝트에서 우리 팀은 데이터 품질 개선에 초점을 맞춘 다각적 접근으로 Public과 Private 리더보드 1위를 달성했다. 특히 LLM을 활용한 노이즈 제거와 데이터 증강 및 재라벨링, 클러스터링 기반의 오라벨 보정, 역번역을 통한 데이터 다양성 확보 등 여러 방법론을 체계적으로 실험하였고 검증했다는 점이 주효했다.

다만 초기 단계에서 다양한 실험과 방법론 테스트에 많은 시간이 소요되었고, 여러 처리 단계에서 발생한 데이터셋의 버전 관리에 어려움이 있었다. 이는 향후 프로젝트에서 체계적인 실험 계획 수립과 데이터 버전 관리 시스템 구축의 필요성을 시사한다.

결과적으로 이번 프로젝트를 통해 모델 개선뿐만 아니라 데이터 품질 관리의 중요성을 실감할 수 있었다. 팀원들의 전문성을 살린 효율적인 역할 분담과 지속적인 커뮤니케이션이 SOTA 달성이라는 성과로 이어졌으며, 이러한 경험은 향후 유사 프로젝트 수행 시 중요한 참고 자료가 될 것으로 기대된다.