1216 수상작 리뷰

월간 데이콘 법원 판결 예측 AI 경진대회

제공된 데이터를 이용해 사건에서 첫 번째 당사자가 승리할 가능성을 예측하는 텍스트 분류 문제

데이터 구조

- **train.csv**
  - `ID`: 사건 샘플 ID
  - `first_party`: 사건의 첫 번째 당사자
  - `second_party`: 사건의 두 번째 당사자
  - `facts`: 사건의 주요 내용
  - `first_party_winner`: 첫 번째 당사자가 승리했는지 여부 (0: 패배, 1: 승리)
- **test.csv**
  - `ID`, `first_party`, `second_party`, `facts`: 테스트 데이터로 제공된 사건 정보
- **sample_submission.csv**
  - 제출 형식으로 사건 ID와 첫 번째 당사자 승리 여부 예측 결과를 포함

코드 흐름

1. **Data Augmentation**
   - 데이터 불균형 문제 해결을 위해 `first_party`와 `second_party`의 순서를 뒤바꾸는 방식으로 augmentation 수행
   - 사건의 승소 여부(`first_party_winner`)에 따라 각 당사자의 역할 바꿔 샘플 수 늘림
2. **Residual Multilayer Perceptron 설계**
   - 텍스트 임베딩을 활용한 추가 네트워크 구성

- Backbone 네트워크를 freeze한 상태에서 residual connection 기반 MLP 설계해 추가 학습
- Dropout, BatchNorm 등으로 정규화된 구조를 사용해 과적합 방지

3. **Cosine Similarity 기반 Classifier**
   - `first_party`, `second_party`, `facts` 각각의 임베딩 벡터를 활용해 승소 가능성 평가
   - Cosine similarity를 이용해 당사자와 사건 내용 간의 관계를 계산하여 승자 예측
   - 두 당사자의 similarity를 비교하여 최종 logits 출력
4. **Ensemble Model 구성**
   - **BERT 계열 모델**과 **LLM (e.g., Vicuna-13B)**를 조합하여 예측 결과 생성
     - BERT 기반 모델 : 데이터 전처리 및 텍스트 분류 수행
     - LLM : few-shot learning을 통해 테스트 데이터와 유사한 판결 사례를 선별하여 활용

차별점 및 배울 점

- **데이터 증강 방식**:

  단순히 샘플 수를 늘리는 것이 아니라 `first_party`와 `second_party`의 순서를 바꿔 새로운 샘플을 생성하는 방식으로, 데이터 불균형 문제를 효과적으로 해결.

- **Residual Connection 활용**:
  - Backbone 네트워크 freeze : 이미 학습된 임베딩을 그대로 사용해 추가 학습 시간을 줄일 수 있음
  - residual connection 기반 MLP : 정보 손실을 줄이면서 기존 임베딩에 새로운 패턴을 더 잘 반영할 수 있도록 설계됨

  → 따라서 Backbone 네트워크를 freeze한 상태에서 residual connection 기반 MLP 설계를 활용하면 과적합을 방지하고 중요한 특징만 학습할 수 있음

- **Cosine Similarity 기반 접근**:

  Cosine similarity : 는 두 벡터 간의 패턴을 비교해 유사도 측정하는 방법

  당사자(`first_party`, `second_party`)와 사건 내용(`facts`)이 서로 연관된 정보를 가질 가능성이 높은 법률 데이터의 특수성을 고려해 해당 방법을 선택함.

해당 모델은 당사자와 사건 내용 간의 유사도 계산을 통해 사건의 맥락에서 어느 당사자가 승리할 가능성이 높은지 평가하는 법률 데이터 맞춤형 접근법이라 볼 수 있음

- **Ensemble 전략**:

  대형 언어 모델(LLM)과 BERT 계열 모델을 조합하여 서로 다른 접근 방식을 통합해 성능을 높임.