

**SK네트웍스 Family AI과정 3기  
 모델링 및 평가 테스트 계획 및 결과 보고서**



**□ 개요**

* 산출물 단계 : 모델링 및 평가
* 평가 산출물 : 테스트 계획 및 결과 보고서
* 제출 일자 : 2024.12.24 (화)
* 깃허브 경로 : <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN03-FINAL-1Team>
* 작성 팀원 : 서민정, 진윤화

1. **논문 요약 모델**

| **개요** | * 본 실험에서는 세 가지 논문의 레퍼런스 요약을 기반으로 LLM의 요약 결과를 비교하였다. Eng-eng 요약 이후 해당 요약 내용을 LLM의 system prompt에 포함시켜 원하는 형태의, 파싱 가능한 output을 추출하는 것이 목적이다. |
| --- | --- |
| **테스트 대상** | * AYA(LLM) + BART (eng-eng 요약) * AYA(LLM) + PRIMERA (eng-eng 요약) * LlaMA(LLM) + BART (eng-eng 요약) * LlaMA(LLM) + PRIMERA (eng-eng 요약) * OpenAI(LLM) + BART (eng-eng 요약) * OpenAI(LLM) + PRIMERA (eng-eng 요약) |
| **테스트 환경** | * Colab GPU (A100) |
| **내용** | * 각 모델에서 eng-eng 요약을 수행한 후, 해당 결과를 시스템 프롬프트에 입력하여 LLM이 논문의 키워드, 요약, 핵심 방법론, 핵심 활용 기술 및 설명, 실험 내용, 실험 결과를 JSON 형태로 추출하는 작업을 진행하였다. |
| **결과** | * PRIMERA (eng-eng 요약)   요약 시 동일한 단어를 반복하거나 후반 내용에 치중하는 경향을 보였다.   * AYA (LLM)   요청된 Key-Value의 형태는 제대로 지켰으나, 논문의 주요 내용을 잘 포함하지 못했다. (요약의 정확성과 깊이 부족)   * LLaMA (LLM)   자연스러운 형태의 결과를 생성했으나, 요청한 키값을 일부 누락하는 경우가 존재하였다. (모델의 신뢰성을 하락)   * OpenAI(LLM) + BART(eng-eng 요약) 조합   본 조합은 논문의 전반적인 내용을 잘 반영하고, 요청한 키-밸류 형태의 값을 지속적으로 정확하게 생성하는 성과를 보였다. 따라서 본 실험에서는 해당 조합이 가장 우수한 성능을 보여주었다.  **OpenAI + BART 조합**이 논문의 핵심 내용을 충실히 반영하고, 요청한 Key-Value형태의 값들을 지속적으로 정확하게 생성하는 우수한 성능을 보였다. PRIMERA와 AYA, LLaMA 모델들은 각각 반복적인 내용, 정보 누락 등의 한계를 보였다.    [그림 1. 논문 요약 기능의 실험 결과 및 판단의 근거] |

1. **논문 검색용 임베딩 모델**

| **개요** | * 본 실험은 세 가지 논문을 기반으로 검색 성능을 비교하였다. OpenAI에게 논문을 입력하여 해당 논문에서 유발할 수 있는 궁금증을 추출하고, 이를 바탕으로 키워드를 추출한 뒤, 키워드와 논문 간의 유사도를 기반으로 검색 시스템을 평가하였다. 본 실험의 목표는 해당 유사도가 높은 임베딩 모델을 찾는 것이다. |
| --- | --- |
| **테스트 대상** | * 데이터의 변화   + 기존 Abstract 기반 검색 시스템 성능 (SPECTER)   + 개선된 다양한 데이터 기반 검색 시스템 성능 (SPECTER) * 모델의 변화   + SPECTER 기반 검색 시스템 성능   + SciBERT 기반 검색 시스템 성능 |
| **테스트 환경** | * Colab GPU (A100) |
| **내용** | * 레퍼런스 키워드와 구축한 데이터 간의 **cosine similarity**를 계산하고, 데이터 변화와 모델 변화 간의 유사도 차이를 평가하여 최적의 데이터와 모델을 선택한다. |
| **결과** | * 데이터의 변화   기존의 Abstract 기반 검색 시스템보다 개선된 다양한 데이터 기반 검색 시스템이 동일 모델 기준으로 10% 이상 성능 향상을 보였다.   * 모델의 변화   개선된 데이터 기반 검색 시스템을 기준으로, SciBERT는 SPECTER에 비해 5% 이상 높은 유사도를 나타냈다.  개선된 검색 시스템에서 **SciBERT**는 기존 데이터 대비 15% 이상의 유사도 향상을 보여, 보다 효과적인 성능을 확인할 수 있었다.    [그림 2. 논문 검색 기능의 실험 결과 및 판단의 근거] |

1. **선행 논문 유사도 모델**

| **개요** | * LLM 텍스트 임베딩 모델 성능 평가를 위한 MTEP 벤치마크 테스트 결과 보고 * 선행 논문 추천 기능을 구현하기 위해 원 논문과 선행 논문의 요약을 text embedding을 통해 벡터로 구현하고, cosine similarity를 구함. |
| --- | --- |
| **MTEB 벤치마크** | * text embedding에 사용될 모델의 성능을 평가하기 위해 MTEB의 STS16 task를 활용함. * MTEB 벤치마크를 통해 snowflake-arctic-embed-l-v2.0, all-mpnet-base-v2, nomic-embed-text-v1.5 모델들을 테스트함. |
| **MTEB 벤치마크 결과** | * 각 모델별 main score는 다음과 같음   + snowflake-arctic-embed-l-v2.0: 0.834902   + all-mpnet-base-v2: 0.800299606531299   + nomic-embed-text-v1.5: 0.834902 * 위의 결과를 통해 snowflake-arctic-embed-l-v2.0, nomic-embed-text-v1.5 모델을 선택함. |
| **실제 텍스트 적용** | * snowflake-arctic-embed-l-v2.0, nomic-embed-text-v1.5 모델을 실제 텍스트에 적용한 결과, nomic-embed-text-v1.5모델은 한국어 모델에 적절한 성능을 보이지 못하여 snowflake-arctic-embed-l-v2.0을 최종 선택함. |
| **결과** | * MTEB 벤치마크를 통해 snowflake-arctic-embed-l-v2.0, nomic-embed-text-v1.5 모델을 선택하였으나, 한국어 지원 이슈로 인하여 snowflake-arctic-embed-l-v2.0을 최종 선택함. |