

하이브리드 임베딩 가중치 조절을 통한 스토리 기반 문서 검색 성능 최적화

이윤선¹, 이채은¹, 이채민¹, 하지원², 김윤희¹

¹숙명여자대학교 소프트웨어학부 컴퓨터과학전공

²서울대학교 컴퓨터공학부

{2116810, leechaeen, 2210403, yulan}@sookmyung.ac.kr

jwh0245@snu.ac.kr

Optimizing Story-Based Document Retrieval using Hybrid Embedding Weight Tuning

Yoonsun Lee¹, Chae Eun Lee¹, Chae Min Lee¹, Jiwon Ha², Yoonhee Kim¹

¹Department of Computer Science, Sookmyung Women's Univ.

²Department of Computer Science and Engineering, Seoul National University

요약

기존 Hybrid 임베딩 연구는 문체와 의미 임베딩을 고정 비율로만 조합하고, 가중치 변화에 따른 검색 성능 영향은 충분히 검토하지 않았다. 본 연구는 문체적 특성이 뚜렷한 문학 문서를 대상으로, 의미 임베딩(text-embedding-ada-002)과 문체 임베딩(text-embedding-3-small)을 다양한 비율로 조합한 하이브리드 임베딩을 설계하고, 검색 성능을 정량적으로 분석한다. 실험에는 Project Gutenberg 의 영어 문학 원문 20 편을 활용하였으며, 평가 지표로 Precision@k, Cosine Similarity, Diversity Score 를 사용한다. 이를 통해 문체-의미 균형에 따른 최적 가중치 조합을 도출한다.

I. 서론

문학 작품과 같은 장문의 스토리 기반 문서를 검색하거나 생성할 때, 문서에 내재한 의미와 문체 모두를 정밀하게 반영하는 것이 중요하다. 기존의 단일 임베딩 모델은 문체와 의미를 동시에 분석하기 어렵다는 한계가 있다. 이를 보완하기 위해 Hybrid 임베딩 기법이 제안되었으며 [9], 저자들의 기존 연구 [1]에서는 의미와 문체 임베딩을 조합해 문학 문서를 분석하였다. 그러나 해당 연구는 문체-의미 임베딩의 가중치를 5:5 로 고정하고, 가중치 비율 변화에 따른 성능 영향을 정량적으로 검토하지 않았다.

본 연구는 문체적 특성이 강한 문학 작품을 중심으로, 두 임베딩의 조합 비율(α)을 변화시키며 검색 성능을 분석하고, 최적 가중치 구성을 도출하고자 한다. 이를 통해 의미-문체 조합의 영향력을 파악하고, 문서 분석 성능 개선에 기여하고자 한다.

II. 관련 연구

2.1. 단일 임베딩 모델의 한계 - 대부분의 검색·생성 시스템은 BERT [6]나 Ada [7] 등 단일 언어 모델에서 추출한 임베딩 벡터로 질의와 문서를 표현한다. 이러한 방식은 간결하지만 문체와 의미를 동시에 정밀하게 포착하기 어렵다는 지적이 있다 [3]. 특히 문학 작품이나 법률 판례처럼 문체의 중요도가 높은 도메인에서는 스타일-내용 간 불일치로 인해 검색 정확도와 생성 품질이 저하되는 현상이 빈번히 나타난다.

2.2. 문체-의미 혼합 임베딩 연구 동향 - CLIP 등 일부 멀티모달 모델은 의미 표현과 문체적 다양성을 함께 포

착하는 구조를 제안하였다 [2][4]. 이를 연구는 공통적으로 문체와 의미 신호를 별도로 학습한 뒤, 가중합(Weighted Sum)이나 대조 정렬(Contrastive Alignment)을 통해 재결합함으로써 단일 임베딩 대비 표현력이 향상된다는 점을 강조한다.

2.3. 연구 공백 및 기여 - 선행 연구는 문체-의미 가중치 비율(α)이 성능에 미치는 영향을 정량적으로 비교하지 않았고, 문학처럼 문체 변동성이 큰 환경에서도 충분히 검증되지 않았다.

본 연구는 이를 보완하기 위해 문체 특화(text-embedding-3-small)와 의미 특화(text-embedding-ada-002) 임베딩을 조합한 하이브리드 구조를 설계하고 [7], 다양한 α 값에서의 성능을 비교해 최적 가중치를 도출한다.

III. 연구 방법

3.1. 하이브리드 임베딩 구조

본 연구에서 제안하는 하이브리드 임베딩(E_{hyb})은 다음과 같이 정의된다:

$$E_{hyb} = \alpha E_{style} + (1 - \alpha) E_{semantic}$$

여기서 $E_{semantic}$ 은 OpenAI 의 의미 기반 임베딩 모델인 text-embedding-ada-002 로부터 생성되며, E_{style} 은 문체적 표현을 반영할 수 있는 text-embedding-3-small 모델로부터 생성된다. 두 모델은 동일한 출력 차원(1536)을 가지므로 정규화 없이 가중 평균이 가능하다. $\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.8\}$ 의 다양한 비율을 실험하여 문체

와 의미 간 가중치 변화가 검색 성능에 미치는 영향을 분석하였다 (그림 1 참조).

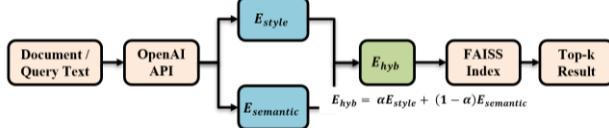


그림 1 하이브리드 임베딩 기반 문서 검색 시스템 구조

3.2. 전처리 파이프라인 - 입력 데이터는 Project Gutenberg에서 수집한 영어 문학 작품 원문 20 편으로 구성되며 [5], 각 작품은 약 200 단어 단위로 청크 분할된다. 이후 각 청크에 대해 OpenAI API를 사용하여 두 임베딩 모델 각각으로 임베딩을 생성하고, 청크별 임베딩의 평균값을 산출하여 문서 단위 벡터를 구성하였다. 이 벡터들은 향후 검색 실험에서 하이브리드 가중치 적용의 기반이 된다.

IV. 실험 설정 및 평가 지표

4.1. 데이터셋 구성 - 실험에는 서로 다른 작가의 문체가 반영된 Project Gutenberg의 영어 문학 원문 20 건을 사용하였다 [5]. 각 작가는 2~3 편의 소설을 기준으로 구성되었으며, 각 작품은 수천 단어 이상의 긴 서사형 텍스트로 구성되어 있어 문체 정보 및 의미 정보 모두를 임베딩 실험에 활용하기에 적합하다.

4.2. 임베딩 및 검색 시스템 구성 - 각 문서 및 질의에 대해 두 임베딩 모델의 벡터를 생성하고, 설정된 α 값에 따라 가중 평균하여 하이브리드 벡터를 구성하였다. 검색은 FAISS 라이브러리 [8]를 이용한 코사인 유사도 기반 최근접 이웃 검색 방식으로 수행되며, 각 설정별로 동일 질의 집합에 대해 평가하였다.

4.3. 평가 지표

- **Precision@k:** 검색 결과 상위 k 개 문서 중 동일 작가의 다른 작품이 포함된 비율로, 문체 유사성 기반 검색 성능을 나타낸다.
- **Cosine Similarity:** 질의 임베딩과 가장 유사한 검색 문서 간 유사도를 평균하여, 검색된 내용의 의미 일관성을 측정한다.
- **Diversity Score:** 검색 상위 k 개 벡터 간 평균 유사도를 역산하여 계산하며, 결과의 표현 다양성과 중복도를 평가한다.

V. 실험 결과 및 분석

본 연구에서는 문체 임베딩과 의미 임베딩 간 가중치 비율을 조절하는 하이퍼파라미터 α 값을 0.2부터 0.8 까지 변화시키며 추천 성능을 정량적으로 비교하였다. 실험에는 **Precision@5**, **Average Cosine Similarity**, **Diversity Score**의 세 가지 평가지표를 사용하였으며, 각각은 추천 정확도, 의미 유사도, 결과 다양성을 측정한다.

실험 결과, **Precision@5**는 $\alpha = 0.5\sim0.7$ 구간에서 최고 성능을 보였고, α 가 너무 낮거나 높을 경우 0.32로 소폭 하락하였다. 이는 임베딩 간 가중치 균형이 추천 정확도에 중요한 영향을 준다는 점을 시사한다 (그림 2 참조).

한편, **Diversity Score**는 α 증가에 따라 급격히 감소하는 양상을 보였다. $\alpha = 0.2$ 에서는 0.1995로 시작하여, $\alpha = 0.8$ 에서는 0.0408 까지 감소하였다. 이는 의미 임베딩 중심의 추천 방식이 결과 다양성 측면에서는 한계를 가짐을 의미한다 (그림 2 및 그림 3 참조).

마지막으로, **Average Cosine Similarity**는 전 구간에서 1.0000으로 유지되어 의미 임베딩 기반의 유사도 계산은 일정하게 유지됨을 확인하였다.

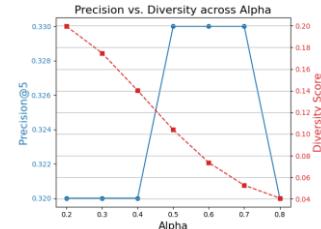


그림 2 α 값에 따른 Precision@5 와 Diversity Score 의 관계

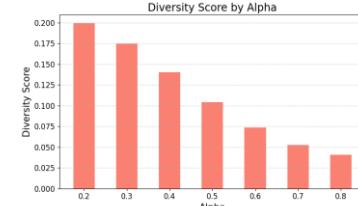


그림 3 α 값에 따른 Diversity Score 변화

본 실험은 α 조절을 통해 추천의 정밀도와 표현 다양성 간의 trade-off 조절이 가능하다는 점을 실증적으로 보여준다.

VI. 결론

본 연구는 문체와 의미 임베딩의 가중치 조절이 문학 기반 문서 추천 성능에 미치는 영향을 정량적으로 분석했다. 실험 결과, 두 임베딩을 5:5 또는 7:3 비율로 조합할 때 Precision@5 가 가장 높았고, 의미 임베딩 비중이 높을수록 정밀도, 문체 임베딩 비중이 높을수록 표현 다양성이 향상되는 경향을 보였다.

하이브리드 가중치 조절은 긴 문서 도메인에 적용 가능하며, 향후 자동화 및 동적 조정 기법으로 확장될 수 있다.

ACKNOWLEDGMENT

이 논문은 2025년도 정부 재원(과학기술정보통신부 여대학원생 공학연구팀제 지원사업)으로 과학기술정보통신부와 한국여성과학기술인육성재단의 지원을 받아 수행되었습니다.

참 고 문 헌

- [1] C. E. Lee, Y. S. Lee, and Y. Kim, "Hybrid Embedding and LLM RAG-Based Virtual System for Literary Evaluation," in Proc. KICS Symp., 2025.
- [2] A. Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," in Proc. ICML, vol. 139, 2021.
- [3] J. Ngiam et al., "Multimodal Deep Learning," in Proc. ICML, 2011.
- [4] A. Conneau et al., "Word Translation without Parallel Data," in Proc. ICLR, 2018.
- [5] Project Gutenberg, <https://gutenberg.org/>, Accessed Apr. 2025.
- [6] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019.
- [7] OpenAI, "New and Improved Embedding Model," <https://openai.com/index/new-and-improved-embedding-model/>, Accessed Apr. 2025.
- [8] M. Douze et al., "The Faiss Library," arXiv:2401.08281, 2024.
- [9] Z. Qiu et al., "A Hybrid Semantic Representation Method Based on Fusion Conceptual Knowledge and Weighted Word Embeddings for English Texts," Information, 15(11), 2024.