ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2023 March: 29(1): 27~40 http://dx.doi.org/10.13088/jiis.2023.29.1.027

언어모델을 활용한 콘텐츠 메타 데이터 기반 유사 콘텐츠 추천 모델

김동환
티빙
(donghwan.kim9@cj.net)

스마트 기기의 보급률 증가와 더불어 코로나의 영향으로 스마트 기기를 통한 미디어 콘텐츠의 소비가 크게 늘어나고 있다. 이러한 추세와 더불어 OTT 플랫폼을 통한 미디어 콘텐츠의 시청과 콘텐츠의 양이 늘어나고 있어서 해당 플랫폼에서의 콘텐츠 추천이 중요해지고 있다. 콘텐츠 기반 추천 관련 기존 연구들은 콘텐츠의 특징을 가리키는 메타 데이터를 활용하는 경우가 대부분이었고 콘텐츠 자체의 내용적인 메타 데이터를 활용하는 경우는 부족한 상황이다. 이에 따라 본 논문은 콘텐츠의 내용적인 부분을 설명하는 제목과 시놉시스를 포함한 다양한 텍스트 데이터를 바탕으로 유사한 콘텐츠를 추천하고자 하였다. 텍스트 데이터를 학습하기 위한 모델은 한국어 언어모델 중에 성능이 우수한 KLUE-RoBERTa-large를 활용하였다. 학습 데이터는 콘텐츠 제목, 시놉시스, 복합 장르, 감독, 배우, 해시 태그 정보를 포함하는 2만여건의 콘텐츠 메타 데이터를 사용하였으며 정형 데이터로 구분되어 있는 여러 텍스트 피처를 입력하기 위해 해당 피처를 가리키는 스페셜 토큰으로 텍스트 피처들을 이어붙여서 언어모델에 입력하였다. 콘텐츠들 간에 3자 비교를 하는 방식과 테스트셋 레이블링에 다중 검수를 적용하여 모델의 유사도 분류 능력을 점검하는 테스트셋의 상대성과 객관성을 도모하였다. 콘텐츠 메타 텍스트 데이터에 대한 임베딩을 파인튜닝 학습하기 위해 장르 분류와 해시태그 분류 예측 태스크로 실험하였다. 결과적으로 해시태그 분류 모델이 유사도 테스트셋 기준으로 90%이상의 정확도를 보였고 기본 언어모델 대비 9% 이상 향상되었다. 해시태그 분류 학습을 통해 언어모델의 유사 콘텐츠 분류 능력이 향상됨을 알 수 있었고 콘텐츠 기반 필터링을 위한 언어모델의 활용 가치를 보여주었다.

주제어 : 콘텐츠 기반 필터링, 추천 시스템, 자연어처리, 언어모델, RoBERTa

논문접수일 : 2022년 11월 15일 논문수정일 : 2022년 11월 15일 게재확정일 : 2022년 12월 2일
원고유형 : 학술대회 Fast Track 교신저자 : 김동환

1. 서론

빅데이터 시대 정보 기술의 발전과 함께 비정형 데이터가 방대하게 쌓이고 있다. 이에 따라 전통적인 정형 데이터뿐만 아니라 비정형 데이터 기반의 인공지능 기술을 통해 비즈니스에 활용하는 기술 대한 관심이 늘어나고 많은 발전을 이루어왔다. 더욱이 스마트 기기의 보급률 증가와 더불어 코로나의 영향으로 스마트 기기를 통한 미디어 콘텐츠의 소비가 크게 늘어나고 있다(신지형, 김윤화, 2021). 이런 추세에서 OTT 플랫폼을 통한 미디어 콘텐츠의 시청이 늘어나고 있어서 해당 플랫폼에서의 콘텐츠 추천이 중요해지고 있다(김지현 등, 2021). OTT 플랫폼 내 콘텐츠 추천의 정확도가 사용자의 플랫폼 만족도와 신뢰도에 긍정적인 영향을 주어 사용자 유지율에 도움을 줄 수 있는 것이다(김헌, 2021). 본 연구는 콘텐츠의 텍스트 메타 데이터를 자연어

처리 언어모델을 통해 OTT 플랫폼 내 유사 콘텐츠 추천에 활용하는 방안을 연구하였다.

본 논문의 구성은 다음과 같다. 2장에서는 언어모델과 콘텐츠 기반 필터링 관련 기본 이론들을 서술하며 관련된 선행 연구들을 살펴보고자 한다. 3장에서는 콘텐츠 텍스트 메타 데이터 기반으로 유사 콘텐츠 추천 시스템 모델의 기술적인 방법론을 설명한다. 4장은 제안한 학습 모델의 성능 고도화 실험을 수행하여 유사 콘텐츠 추천 모델의 일반화 성능을 분석한다. 마지막으로 5장은 본 연구의 향후 연구 방향에 대하여 제시한다.

2. 관련 연구

2.1. RoBERTa

RoBERTa모델의 전신 모델인 BERT 언어모델은 대규모의 텍스트 데이터를 비지도학습으로 사전 학습한 언어모델로서 자연어처리 발전에 큰 영향을 끼쳤다. BERT는 Transformer 기반 모델로서 Transformer의 인코더 디코더 블록 중 인코더 블록을 여러 계층으로 쌓아서 만든 모델이다(Vaswani et al., 2017; Devlin et al., 2018). 사전 학습은 랜덤으로 마스킹된 단어를 예측하는 MLM(masked language model)과 두 개의 문장이 문맥적으로 연달아 등장 가능한 문장인지 예측하는 NSP(next sentence prediction) 두 가지 방식으로 수행한다. 양방향으로 텍스트를 학습하게 되어 더 좋은 의미 표상 정보를 얻을 수 있다.

위의 BERT 모델의 성능을 강화시키기 위해서 학습 데이터를 추가하고 하이퍼파라미터 및 훈련 기법을 조정하여 학습한 모델이 RoBERTa다(Liu et al., 2019). 사전 학습 방법 중에 하나인 NSP 방법은 제외하고 MLM 방법만으로 학습하였으며 더 큰 학습 데이터 및 더 긴 시퀀스로 더 오래 학습을 수행하고 동적 마스킹을 적용하여 보다 정교한 의미표상 정보를 얻도록 개선하였다. 결과적으로 BERT를 포함한 이전 모델들의 GLUE 벤치마크 성능을 앞질렀다. 본 연구에서는 RoBERTa의 한국어 말뭉치 기준으로 사전 학습한 대표적인 모델인 KLUE-RoBERTa-large를 사용하였다(Park et al., 2021). KLUE 자체 벤치마크 관련하여 대부분의 자연어 처리 다운 스트림 태스크에서 이전 결과보다 우수한 성능 보였기 때문이다.

2.2. 텍스트 데이터를 활용한 콘텐츠 기반 협업 필터링

추천 시스템에는 크게 협업 필터링과 콘텐츠 기반 필터링 두 가지 방법론이 있다. 협업 필터링은 사용자와 아이템 간 상호작용 데이터를 기반으로 추천해주는 방법론이다. 반면에 콘텐츠 기반 필터링은 사용자가 선호하거나 구매한 아이템과 유사한 아이템을 추천해주는 방법론이다. 콘텐츠 기반 필터링의 경우 아이템 자체의 특성을 기반으로 추천해주기에 사용자-아이템 상호작용 데이터가 없어도 추천이 가능한 장점이 있다. 이로 인해 평가나 구매 이력이 없는 신규 사용자나 아이템에 대한 추천이 어려운 콜드 스타트의 문제를 어느 정도 해소할 수 있다.

콘텐츠 기반 필터링 관련 연구는 다음과 같다. 영화와 평점 정보 메타 데이터를 바탕으로 순환 신경망 기반의 추천 모델을 가중치 결합 기법을 활용하여 개선한 연구가 있었다(권명하 등, 2018). 사용자의 영화 평점과 영화 장르 행렬을 내적하여

얻은 결과 행렬을 기반으로 콘텐츠 간의 유사도를 구하여 추천영화를 결정하는 연구가 있었다(Reddy et al., 2019). 콘텐츠의 이미지, 오디오, 태그, 장르 등의 메타 데이터를 바탕으로 아이템 기반 최근접 이웃 모델을 활용하여 유사한 영화를 추천하고 애플리케이션까지 구현하는 연구도 있었다(Deldjoo et al., 2019). 이와 같이 기존 연구들은 콘텐츠의 특징을 가리키는 메타 데이터를 활용하는 경우가 대부분이었고 콘텐츠 자체의 내용적인 메타 데이터를 활용하는 경우는 부족한 상황이다.

한편 텍스트 기반 추천시스템 관련 연구는 다음과 같다. 전통적으로는 단어 출현 빈도에 따라 텍스트를 분석하는 BoW(Bag of Words)와 TFIDF(Term Frequency-Inverse Document Frequency) 방식의 임베딩을 추천시스템에 적용하였다(Yin et al., 2018). BERT 언어모델을 활용하여 음식 분야의 리뷰 텍스트 기반으로 평점을 예측하여 추천시스템에 활용한 연구도 있었다(박호연, 김경재, 2021). 해당 연구에서는 BERT 모델이 기존의 협업 필터링 기반 모델이나 순환 신경망 기반 모델보다 성능이 우월하였다.

이외에도 사용자 리뷰 텍스트를 BERT 등의 언어모델을 통해 분석하여 추천시스템에 활용한 연구가 다수 있었다. BERT 언어모델을 활용하여 속성카테고리 기반 감성분류 방법론을 추천시스템에 활용한 연구가 있었다(이유린 등, 2021). 속성카테고리 기반 감성분류 방법론은 텍스트의 벡터를 추출할 때 보편적인 방법인 CLS 토큰만 활용하는 것이 아니라 속성 카테고리에 해당하는 중요 토큰들에 가중치를 주어 감성 분류의 성능 향상을 도모한 방법론이다(박현정, 신경식, 2020). BERT 언어모델을 활용하여 사용자 리뷰를 바탕으로 평점을 예측하도록 학습한 뒤 이를 추천시스템에 활용한 연구도 있었다(박호연, 김경재, 2021). 비슷하게 KoBERT 언어모델을 활용하여 사용자 리뷰를 바탕으로 감성 평점을 예측하도록 학습한 뒤 해당 감성 피처와 함께 다른 영화 속성 피처들을 추천 시스템에 활용한 연구도 있었다(홍태호 등, 2022).

본 논문은 콘텐츠의 내용적인 부분을 설명하는 제목과 시놉시스를 포함한 다양한 텍스트 데이터를 바탕으로 유사한 콘텐츠를 추천하고자 자연어처리의 언어모델인 KLUE-RoBERTa-large을 적용하는 연구를 수행하고자 한다.

3. 연구 방법

3.1. 연구 데이터

사내에서 보유하고 있는 2만여건의 콘텐츠 메타 데이터 중에 콘텐츠 식별자 코드를 포함하여 텍스트 피처를 사용하였다. 텍스트 피처는 <표 1>에서와 같이 콘텐츠 제목, 시놉시스, 복합 장르, 감독, 배우, 해시태그 정보다. 복합 장르는 대분류 장르와 소분류 장르 모두 활용하였다. 예를 들면, ‘액션/SF’라는 대분류 장르의 소분류 장르에는 ‘액션’, ‘판타지’, ‘SF’, ‘어드벤처’, ‘전쟁’, ‘무협’ 등이 있다. 감독과 배우는 너무 많은 경우 최대 5명 이하의 이름만 활용하였다. 해시 태그 정보는 주제, 감정, 목적 등의 내용을 가리키는 태그 정보다.

언어모델의 입력으로 들어가는 텍스트 데이터는 이어져 있는 시퀀스형 데이터여서 콘텐츠 메타 데이터와 같이 정형 데이터로 구분되어 있는 텍스트 데이터는 언어모델에 직접적으로 입력할 수 없다. 보통은 구분되는 텍스트 피처 간에

〈표 1〉 콘텐츠 메타 데이터 예시

코드	제목	장르	감독	배우	해시 태그	시놉시스
M000xxxx	하모니	드라마, 음악	강대규	김윤진, 장영남, ...	결핍된 자들의 상호보완, 감동적인, 마음이 따뜻해지는, …	여자교도소를 배경으로, 사연 많은 수감자들이 교도소 내 합창단으로 …

[SEP] 구분자를 사용하여 입력 시퀀스 텍스트 데이터를 만든다. 하지만 현재 콘텐츠 메타 데이터는 [SEP] 토큰을 사용하기에는 많은 종류의 텍스트 피처가 있어서 [SEP] 토큰만으로 구분한다면 모델 입장에서 해당 구분자의 순서만으로 다양한 피처들을 구분지어 인식할지가 불확실하다. 관련해서 두 엔티티 간의 관계를 추론하는 R-BERT모델은 타겟 텍스트의 앞뒤로 스페셜 토큰을 삽입하는 방식을 사용하였다 (Wu & He, 2019). 해당 방법을 참고하여 아래와 같이 제목과 시놉시스를 제외한 텍스트 피처 앞뒤로 해당 피처를 가리키는 스페셜 토큰을 두어 학습하였다. 물론 해당 스페셜 토큰도 언어모델의 어휘사전에 추가하였다.

1) 스페셜 토큰 예시: 타이틀[SEP]시놉시스 [GENRE]장르1 장르2[/GENRE][DIR]감독 [/DIR][ATR]배우1 배우2[/ATR][TAG]태그1 태그2[/TAG]

3.2. 학습 태스크 설계

콘텐츠를 좋아하는 취향의 중요한 기준을 예측하도록 학습한 모델이라면 해당 기준으로 유사한 콘텐츠를 잘 구분해낼 것이다. 관련해서 <표 2>에서와 같이 대체로 장르, 해시태그, 감독, 배우가 콘텐츠를 보는 취향을 상당히 반영한다. 감독과 배우는 예측 대상으로 분류하기에는 분류 종류 수도 많고 케이스 별 데이터가 많지 않아서 일반화된 의미표상을 학습하지 못할 것이다. 반대로 해시태그와 장르는 상대적으로 반복되는 단어들이 많아서 해당 기준으로 예측하여 학습하도록 설계하였다. 장르는 정해진 카테고리 내에서 개별 데이터 마다 등장하며, 해시 태그도 주요 명사 토큰들은 기학습 과정에서부터 많이 학습되기에 일반화된 의미표상을 학습할 것이다. 이에 따라 언어모델로 하여금 장르나 해시 태그를 예측하는 훈련을 하면 장르나 해시 태그 측면에서 예측값이 동일한 콘텐츠 간에는 유사하게 임베딩될 것이다. 이로써 해당 모델은 중요한 취향 기준으로 유사한 콘텐츠를 잘 분류할 수 있게 된다.

〈표 2〉 좋아하는 영화의 기준 예시

좋아하는 영화 표현 예시	기준 구분
액션 영화 좋아한다.	장르(액션)
일본 영화 좋아한다.	해시태그(#일본배경)
봉준호 감독 영화 좋아한다.	감독(봉준호)
감동적인 영화 보고 싶다.	해시태그(#감동적인)
류승룡 배우 영화는 믿고 본다.	배우(류승룡)

장르 분류는 텍스트 분류 방식으로 해시 태그 분류는 MLM 방식으로 학습을 수행하였다. 장르는 카테고리 내 분류 수가 어느정도 제한되어 있으므로 일반적인 텍스트 분류로 학습하였으며 콘텐츠마다 소속 장르가 여러 개이기 때문에 다중 레이블 방식으로 예측하였다. 반면 해시 태그

〈표 3〉 학습 태스크별 입력과 타겟 구조

예측 방식	입력	타겟
장르 분류	타이틀[SEP]시놉시스[DIR]감독[/DIR][ATR]배우1 배우2[/ATR][TAG]태그1 태그2[/TAG]	장르1, 장르4
해시 태그 분류	타이틀[SEP]시놉시스[GENRE]장르1 장르2[/GENRE][DIR]감독[/DIR][ATR]배우1 배우2[/ATR][TAG]태그1 [MASK][/TAG]	[MASK]=태그2

분류는 해시 태그 종류 수가 정해져 있지 않고 자유롭게 추가될 여지를 고려하여 시퀀스 내에서 해시 태그에 대해 마스킹된 토큰을 예측하는 방식으로 학습하였다. 마스킹 대상 토큰은 해시 태그 영역([TAG]~[/TAG])내에서 “#”으로 시작하지 않는 토큰들 중 랜덤으로 1개를 마스킹하였다. RoBERTa 모델의 BPE 토크나이저 특성 상 토큰이 “#”으로 시작하는 토큰은 앞 토큰에 의존적이거나 문법적인 의미의 토큰이기에 상대적으로 체언과 용언과 같이 핵심의미를 담은 토큰은 “#”으로 시작하지 않은 토큰이라 판단하였기 때문이다. 이에 따라 해시 태그 MLM 학습 과정에서 모델은 해시 태그 영역에 있는 토큰을 다른 텍스트 피처에서 얻을 수 있는 문맥정보를 바탕으로 추론하게 된다. 해시태그가 여러 개인 경우 해시태그 마스크 토큰을 두 군데 이상 두면 어느 위치에 어떤 해시태그 토큰인지 모델이 알 수 없으므로 마스크 토큰은 랜덤으로 해시 태그 한 개만 적용하였다. 해시 태그 토큰을 잘 맞출 수 있도록 언어모델의 의미표상이 더욱 정교해지고 콘텐츠 간의 텍스트 메타 유사도를 더욱 정확하게 계산할 수 있게 된다. 각 학습 태스크별 입력과 타겟값 구조는 <표 3>과 같다.

3.3. 테스트셋 설계 및 구축

유사도는 절대적인 개념이 아니라 상대적인 개념이기에 테스트셋도 상대적인 개념을 잘 반영해야 한다. 절대적인 개념이 아니라는 것의 의미는 두 콘텐츠 간에 비교할 때 유사한지 여부를 얘기할 수 없다는 의미다. 상대적인 개념을 반영하기 위해서 3개 콘텐츠 간의 3자 비교를 통해 더 유사한 콘텐츠인지 여부를 분류하는 방식을 설계하였다. 기준 콘텐츠와 더 유사한 콘텐츠의 유사도가 기준 콘텐츠와 덜 유사한 콘텐츠의 유사도 스코어보다 높을 때 맞추는 방식의 정확도를 구하는 것이다. 더욱이 상대적인 개념을 보완하며 테스트셋의 정확도 품질을 도모하기 위해 3명이상의 다중 검수를 수행하여 검수자가 동일하게 판단한 174건의 객관적인 케이스만 테스트셋으로 활용했다. 유사도를 판단하는 사람마다 어떤 사람은 장르에, 어떤 사람은 해시 태그 내용에, 혹은 시놉시스 내용에 주관적인 가중치를 두어 판단할 수 있기 때문이다.

유사도 비교 기준도 상대적으로 쉬운 기준과 어려운 기준 모두 구축하였다. 모델에 따라 어려운 기준의 성능향상과 쉬운 기준의 성능향상이 반드시 비례하지는 않을 수 있기 때문에 다방면으로 보기 위함이다. 구체적으로 아래와 같은 기준으로 <그림 1>의 예시와 같이 설계하였다.

1) 유사도 테스트셋 콘텐츠 항목 기준: 아래 기준 외에 제목과 시놉시스 내용은 기본적으로 고려

콘텐츠	유사콘텐츠	덜유사콘텐츠	상이한 콘텐츠
M000124405	M000320433	M000284439	M000333734
도리화가	소리꾼	황진이	알렉산더 : 절대영웅의탄생
사극 , 드라마 , 음악	사극 , 드라마 , 음악	사극	모험
이종필	조정래	배창호	이고르 카리오노프
류승룡,수지,김남길,노영학,송새벽,이동휘,안재홍,황채원,김태훈,진선규,김소진	이봉근,이유리,김하연,박철민,김동완 ,김민준,정무성,임성철,김병춘,정인기,한인수,김강현,오지혜,박재민,손숙,조상기	장미희,안성기,신일룡	스베트라나 바쿠리나,안톤 팜부쉬니
주인공의 고군분투 , 고난을 딛고 일어서는 주인공의 자질 ,흥미로운, 마음을 끄는 ,실화를 바탕으로 한 작품이 보고 싶을 때,원대한 꿈을 갖고 싶을 때, 감동적인 ,코로나 시국에 집콕할 때	고난을 딛고 일어서는 주인공의 자질 , 주인공의 고군분투 , 마음을 끄는 , 감동적인 ,역경을 딛고 일어난 서사가 보고싶을 때,보고싶은 영화라서,안타까운,기분전환,인과응보와 권선징악의 메세지	배신,가슴아픈,한국배경,이별,관계,남녀,미녀,90년이전	영웅,왕자,중세배경,역사,왕위,전쟁,2000년대
어릴 적 부모를 잃은 채선은 우연히 듣게 된 재효의 아름다운 소리에 이끌려 소리꾼의 꿈을 품는다. 여자가 판소리를 할 수 없었던 시대에 꿈을 향해 나아가는 과정을 그린 영화	납치된 아내를 찾기 위해 여정에 오르는 소리꾼 학규를 필두로 그와 함께하는 이들의 이야기를 그린 영화. 이 작품은 심청전과 춘향전을 소재로 만든 한국형 뮤지컬 영화이다.	황진사의 딸로 재색을 겸비한 진이는 혼례전날 그녀를 짝사랑하던 갖바치의 자살로 일방적으로 파혼을 당하고 갖바치의 죽음에 충격을 받고 일개 기녀로 변신한다. 기녀로 명성을 떨치던 진이는 벽계수를 만나 서로 사랑하게 되지만 그가 명나라 사신으로 발탁되어 그녀의 품을 떠나버리자 배신감에 사로잡혀 유랑길에 오른다. 그후 경제적으로 무능한 선비 이생과 만나 삶을 같이하지만 그가 전락 사당패들에게 진이 자신을 팔아 넘기려고 함을 알고 스스로 사당패를 따라나선다.	젊은 왕자 알렉산더는 동쪽의 호드와 서쪽의 독일 기사단 그리고 스웨덴, 동서쪽을 방어해야 했다. 한편 알렉산더는 일부 귀족 계층이 음모를 꾸미고 그들의 무역을 도와줄 스웨덴과 독일을 위해 노브고로드를 배신할 준비가 되어 있다는 것을 알게 된다. 그리고 정체 모를 누군가가 알렉산더의 결혼식에서 그를 독살하려는 사건이 일어나면서 그의 가장 친한 친구가 의심을 받게 된다. 알렉산더에게는 침략자들로부터 그의 사람들을 보호하고 진짜 독살범이 누구인지 찾는 것 외에는 다른 방법이 없는 것 같은데

〈그림 1〉 유사도 테스트셋 예시

a. 유사한 콘텐츠: 기준 콘텐츠와 장르와 해시태그에 공통점 있음

b. 덜 유사한 콘텐츠: 기준 콘텐츠와 장르만 공통점이 있고 해시태그에는 상대적으로 없음

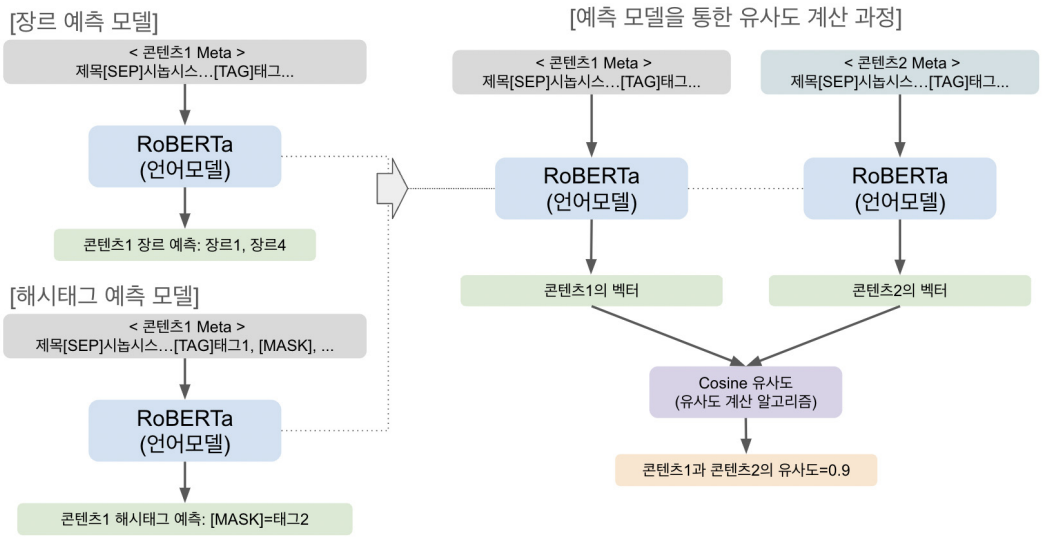
c. 상이한 콘텐츠: 기준 콘텐츠와 장르 및 해시태그 모두 공통점이 없음

2) 유사도 테스트셋 분류 정확도 종류: 분류 정확도2가 상대적으로 쉬운 기준

a. 분류 정확도1: 기준 콘텐츠와 유사한 콘텐츠 간의 유사도가 기준 콘텐츠와 덜 유사한 콘텐츠 간의 유사도보다 높으면 맞춘 것으로 계산한 정확도

b. 분류 정확도2: 기준 콘텐츠와 유사한 콘텐츠 간의 유사도가 기준 콘텐츠와 상이한 콘텐츠 간의 유사도보다 높으면 맞춘 것으로 계산한 정확도

장르 혹은 해시 태그 분류 예측을 수행한 모델을 바탕으로 유사도를 계산하는 과정은 <그림 2>와 같다. 예측 태스크를 수행하면서 모델은 콘텐츠 간의 유사 관계를 더욱 잘 식별할 수 있도록 더욱 정교한 의미표상을 제공하도록 학습이 된다. 이어서 유사도 계산 단계에서 예측 태스크를 추론할 때 사용하는 헤드 레이어를 제외하고 모델 자체의 마지막 은닉계층 임베딩 값의 평균을 해당 콘텐츠 메타의 임베딩 값으로 사용한다. 각각의 콘텐츠 메타의 임베딩 값 간에 코사인 유사도 함수를 사용하고 최종 유사도 스코어를 계산한다. 이런 식으로 계산하여 실무적으로는 모든 콘텐츠 간의 유사도를 계산할 수 있고 특정 콘텐츠와 가장 유사한 콘텐츠의 상대적인 목록을 획득할 수 있게 된다.



〈그림 2〉 학습한 예측 모델을 통한 유사도 계산 과정

4. 연구 결과

4.1. 장르 분류와 해시태그 분류 모델의 성능 비교

태스크 별 예측 모델 성능은 다음과 같다. 장르 분류에서 사용한 정확도 지표는 모델이 가장 확률 높게 예측한 장르가 해당 콘텐츠가 소속한 여러 장르 중에 하나라도 해당되면 맞춘 것을 기준으로 계산하였다. 태스크 별 학습 시 데이터 분할은 학습셋, 검증셋, 테스트셋 비율이 90%, 5%, 5%를 차지하도록 분할하였다. 사내 데이터 기준으로 장르 분류가 대분류는 8가지 소분류는 24가지가 존재하는데, 대분류 예측의 경우 약 70%, 소분류 예측의 경우 약 66%의 정확도를 보였다. 해시태그 분류에서 사용한 정확도 지표는 모델이 마스킹된 토큰을 맞춘 것을 기준으로 계산하였다. 해시태그 분류의 경우 랜덤 마스킹에 따라 약 47%~60%의 정확도를 보였다. 하지만 장르 분류이나 해시태그 분류의 태스크 성능은 중요하지 않다. 장르나 해시태그 분류를 하는 것 자체가 목적이 아니라 얼마나 유사한 것끼리 임베딩 차원에서 유사하게 위치시키느냐가 중요하기 때문이다.

이에 따라 유사도 모델 테스트셋 성능 기준으로 태스크 별 예측 모델 성능을 비교해보았다. 기본 언어모델도 대량의 말뭉치에서 사전 학습을 한 모델이기에 어느정도 유사도 분류 성능을 갖고 있다. 기본 언어모델 대비 태스크 별 파인튜닝 학습 모델의 성능을 비교했다. 테스트셋 정확도1과 정확도2의 기준으로 대분류 장르 분류 모델은 각각 1.95% 하락, 7.77% 상승을 보였다. 소분류 장르 분류 모델은 각각 3.89% 하락, 9.71% 상승을 보였고 해시태그 분류 모델은 각각 6.79% 상승, 6.8% 상승을 보였다.

해시 태그 모델이 두 정확도 모두 향상이 있는 것과 달리 장르 분류 모델은 정확도2만 향상이

있으므로 장르 측면에서만 유사도 식별 능력이 있다고 할 수 있다. 장르 분류 모델 학습과정에서 공통된 장르의 콘텐츠 간에 해시태그를 기반으로 더 유사한 콘텐츠를 구별해내는 능력은 훈련한 적이 없으므로 정확도1의 향상이 없을 수 있다. 반면에 해시태그 모델은 해시 태그 정보를 기준으로 학습함에도 장르 측면에서의 유사도 식별 능력도 좋아졌다고 할 수 있다. 이에 따라 최종 모델은 해시태그 모델을 사용하였다.

4.2. 성능 고도화 실험

4.2.1. 벡터 획득 방법 변경 실험

입력 텍스트의 유사도를 비교할 때 사용할 벡터를 얻는 방식을 변경해보면서 테스트셋 정확도 성능을 비교해보았다. 입력 텍스트의 유사도를 비교할 때 사용할 벡터를 얻는 방식은 다음 세가지가 있다.

1) Pooler Output: 언어모델의 [CLS] 토큰의 마지막 은닉층 출력 벡터를 입력 텍스트 벡터로 간주
2) Last Hidden States 평균: 언어모델의 모든 단어의 마지막 은닉층 출력 벡터에 대해서 평균 풀링을 수행한 벡터를 입력 텍스트 벡터로 간주
3) Last Hidden States 최대값: 언어모델의 모든 단어의 마지막 은닉층 출력 벡터에 대해서 맥스 풀링을 수행한 벡터를 입력 텍스트 벡터로 간주

장르 분류이나 해시태그 분류의 태스크 훈련을 수행하지 않은 기본 언어모델도 어느 정도의 테스트셋 성능이 나오므로 기본 언어모델 기준으로 위 3가지 방식의 성능을 비교했다. 기본 언어모델은 영화 감독과 배우 이름 피처에 대해 기학습 과정에서 학습한 적이 없으므로 제목, 시놉시스, 장르 및 태그데이터만 입력 데이터로 사용해서 테스트했다. Last Hidden States 평균값으로 설정했을 때가 성능이 제일 좋았다. 테스트셋 정확도1과 정확도2의 기준으로 Last Hidden States 최대값으로 설정했을 때 각각 5.2%, 7.91% 하락하였고 Pooler Output인 경우 11.65%, 13.59% 하락하였다. 결과적으로 최종 모델은 Last Hidden States 평균값을 사용하였다.

4.2.2. 피처 별 가중치 임베딩 실험

특정 피처 위치에 해당하는 마지막 은닉상태 값들에 가중치를 주면 유사도 정확도 스코어가 향상되는지 실험해보았다. 다음의 해시태그 분류 모델을 바탕으로 다음의 방식대로 테스트셋 성능을 비교해보았다.

1) 해시태그 값들에 가중치: [TAG]~[/TAG] 토큰 값들 위치에 있는 마지막 은닉층 벡터값들을 2배 가중치
2) 장르 값들에 가중치: [GENRE]~[/GENRE] 토큰 값들 위치에 있는 마지막 은닉층 벡터값들을 2배 가중치
3) 타이틀 및 시놉시스 값들에 가중치: “제목 [SEP]시놉시스” 토큰 값들 위치에 있는 마지막 은닉층 벡터값들을 2배 가중치
4) 장르 및 해시태그 값들에 가중치: [TAG]~[/TAG]와 [GENRE]~[/GENRE] 토큰 값들 위치에 있는 마지막 은닉층 벡터값들을 2배 가중치

장르 값들에 가중치를 부여한 모델이 테스트셋 성능이 제일 좋았다. 테스트셋 정확도1과 정

확도2의 기준으로 해시태그 값들에 가중치를 부여한 모델이 각각 1.94%, 6.8% 하락하였다. 또한 타이틀 및 시놉시스 값들에 가중치를 부여한 모델은 각각 4.85%, 1.94% 하락하였고 장르 및 해시태그 값들에 가중치를 부여한 모델은 각각 0.97%, 4.86% 하락하였다. 결과적으로 최종 모델은 장르 값들에 가중치를 부여한 모델을 사용하였다.

4.2.3. 어휘 사전 추가 실험

메타 텍스트의 빈출 단어들을 기본 언어모델의 어휘사전에 추가하여 학습할 경우 성능 향상이 되는지 실험해보았다. 빈출 단어가 어휘사전에 추가될 경우 분절되지 않고 단일 토큰으로 모델에서 인식하게 된다. 다음과 같은 기준으로 어휘사전에 단어들을 추가하였다.

1) 빈출 단일어절 해시태그: 99%의 콘텐츠들이 해당 목록 중 하나라도 갖고 있는 목록이며, 이중 단일어절이 아닌 경우는 제외
2) 해시태그에서 추출한 명사목록: MECAB 형태소 분석기를 이용하여 해시태그 텍스트들을 형태소 분석 후 얻은 명사 목록
3) 장르 단어 목록: 대분류, 소분류 장르 목록

어휘 사전에 빈출단어를 추가한 모델이 그렇지 않은 모델에 비해 테스트셋 정확도1, 정확도2의 기준으로 각각 4.6%, 2.3% 하락하였다. 기본 언어모델의 어휘사전에 가능한 추가하지 않고 기존에 언어모델이 사전 학습했던 토큰들에 대한 임베딩 정보를 최대한 활용하는 것이 모델 성능에 최적인 것이다. 신규 추가된 단어 토큰들에 대해 일반화된 의미표상을 얻기에는 데이터가 충분하지 않은 것으로 판단된다.

결과적으로 최종 모델은 텍스트 피처간 구분자인 스페셜 토큰 및 대분류 장르에 대한 단어들만 어휘사전에 추가하였다. 대분류 장르 단어를 추가한 이유는 해당 단어들을 추가했을 때 단일 토큰으로 인식함으로써 인식할 수 있는 시퀀스 길이가 늘어나며 유사도 분류 성능 손실이 없었기 때문이다.

4.2.4. 장르 및 해시태그 동시 예측 모델 실험

기존의 해시태그 MLM 학습에서 해시 태그뿐만 아니라 장르에도 랜덤으로 마스킹을 씌워서 훈련할 경우 성능 향상이 되는지 실험하였다. 아래의 입력 데이터 예시와 같이 장르 토큰 중에서도 랜덤으로 1개의 장르 토큰에 마스킹을 씌우는 것이다.

1) 해시태그 및 장르 분류 입력 데이터: 제목 [SEP]시놉시스[GENRE]장르1, [MASK] [/GENRE]...[TAG]태그1, [MASK]

장르까지 MLM을 적용한 모델이 그렇지 않은 모델에 비해 테스트셋 정확도1, 정확도2의 기준으로 각각 1.73%, 0.57% 하락하였다. 결과적으로 해시태그만 MLM을 적용한 모델을 최종모델로 사용하였다.

4.3. 최종 모델 일반화 성능

해시태그 분류 최종 모델의 일반화 성능을 파악하기 위해서 데이터 분할의 비율을 조율하고 여러 차례의 테스트셋 성능을 구하여 분석하였다. 해시태그 모델 학습 시 훈련셋, 검증셋, 테스트셋 관련해서 모델의 훈련 주기가 1주일이라면 1주일 안에 모델이 만나게 될 신규 콘텐츠 개수만큼 테스트셋 규모를 설정하였다. 예상되는 개수가 100여건이여서 테스트셋, 검증셋을 각각

〈표 4〉 최종 모델 유사도 테스트셋 일반화 성능

모델 종류	테스트셋 정확도1	테스트셋 정확도2
기본 언어모델	78.16%	88.51%
해시태그 분류 최종 모델 1차	**90.23%**	**98.28%**
해시태그 분류 최종 모델 2차	**91.38%**	**97.70%**
해시태그 분류 최종 모델 3차	**91.38%**	**97.70%**

100건을 할당하고 나머지 데이터를 모두 훈련셋으로 할당하였다. 위와 같은 랜덤 데이터 분할을 여러 차례 수행하여 테스트셋 성능을 분석하였다. 한번의 테스트셋 성능 보다는 여러 번의 테스트셋 성능 비교하면 보다 정교한 일반화 성능을 파악할 수 있기 때문이다. <표 4>에서는 3차례의 랜덤 샘플링으로 훈련셋, 검증셋, 테스트셋을 분할하여 각각의 해시태그 학습 성능을 비교하였다.

결과적으로 해시태그 분류 최종 모델은 유사도 테스트셋 정확도1과 정확도2의 기준에서 각각 90.23%~91.38%와 97.7%~98.28%의 성능을 보였다. 해시태그 분류 학습하기 전의 기본 언어모델 성능인 78.16%와 88.51%에 비해 각각 대략 12%, 9% 향상되었다. 해시태그 분류 모델의 학습 방법이 콘텐츠 메타 내용을 바탕으로 유사 콘텐츠를 구별하는 능력을 유의미하게 향상시키는 것이다.

5. 결론

5.1. 연구 결과 정리

본 연구의 연구 결과는 다음과 같이 정리할 수 있다. 첫째, 정형 데이터 형식의 텍스트 데이터를 바탕으로 언어모델 기반의 학습 방법을 고안하였다. 피처 별로 스페셜 토큰을 기준으로 구별하여 입력하는 방안을 적용하여 파인튜닝 훈련을 수행하였다. 파인튜닝 학습 태스크 관련해서는 장르 분류 모델과 달리 장르와 해시태그 및 시놉시스 등의 내용을 종합적으로 고려하여 유사 콘텐츠 구별 능력이 좋은 해시태그 모델을 선택하였다.

둘째, 언어 모델의 유사도를 분류하는 능력을 점검하는 테스트셋을 설계하였다. 유사도의 상대적인 개념을 고려하여 3개 콘텐츠 간의 3자 비교를 하는 방식의 테스트셋을 구현하였고 테스트셋의 객관성을 위하여 다중 검수를 수행하였다. 유사도 비교 기준도 유사 콘텐츠 간에 좀 더 유사한 콘텐츠를 분류하는 어려운 기준과 상이한 콘텐츠를 분류하는 쉬운 기준 모두 실험함으로써 모델의 유사도 분류 능력을 정교하게 파악하고자 하였다.

셋째, 다양한 고도화 방안을 실험하면서 결과적으로 해시태그 MLM 태스크를 통해서 언어모델의 유사 콘텐츠를 구별하는 능력을 향상시킬 수 있음을 확인하였다. 유사도 계산의 기준이 되는 텍스트 벡터 획득 방법은 테스트셋 성능이 제일 좋은 마지막 은닉층의 벡터의 평균값으로 사용하였다. 피처 가중치 관련해서는 장르 값들에 해당하는 마지막 은닉층 벡터들에 가중치를 주는 방식이 테스트셋 성능이 제일 좋았다. 또한 신규 어휘사전을 가능한 추가하지 않고 언어모델 기학습 과정에서 학습했던 토큰들을 최대한

활용하는 것이 테스트셋 성능이 좋았다. 더욱이 장르와 해시태그 분류을 동시에 실행하지 않고 해시태그만 예측하여 훈련하는 경우가 유사도 테스트셋 성능이 뛰어났다.

5.2. 향후 연구 방향

본 연구에 이어서 향후 연구 방향은 다음과 같다. 첫째, 해시태그 MLM 학습한 모델을 시멘틱 검색에 활용해볼 수 있다. 해시태그 분류 모델을 통해 언어모델은 콘텐츠 메타 텍스트 별로 보다 정교한 임베딩을 얻을 수 있게 된다. 이를 통해 예를 들면 “감동적인 영화”와 같은 시멘틱 검색어를 입력했을 때 해당 검색어의 벡터와 콘텐츠 메타 텍스트 벡터 간의 유사도를 구할 수 있게 된다. 자주 등장하는 시멘틱 검색어의 주요 유사한 콘텐츠들의 스코어값을 사전에 구해서 검색엔진을 위한 활용 방안을 연구해볼 수 있다.

둘째, 개인화 혹은 사용자 그룹화 유사 콘텐츠 추천을 연구할 수 있다. 본 연구의 해시태그 분류 모델을 통한 유사 콘텐츠 구별 모델은 사용자 별로 일관되게 콘텐츠 별 유사 콘텐츠를 보여준다. 하지만 사용자 별로 좋아하는 장르나 태그 정보가 다를 수 있다. 이러한 사용자 개인의 취향을 반영하여 상위 유사 콘텐츠 목록을 재조정한다면 추천 품질 향상에 기여할 것이다. 사용자 개인화 뿐만 아니라 비슷한 사용자 간의 그룹별로 추천 결과를 다르게 보여주는 방안도 고려해볼 수 있다.

참고문헌(References)

[국내 문헌]

권명하, 공성언, & 최용석. (2018). 임베딩을 활용한 순환 신경망 기반 추천 모델의 성능 향상 기법. 정보과학회논문지, 45(7), 659-666.

김지현, 하희정, 김서희, & 정영욱. (2021). OTT 서비스 콘텐츠 추천 사용자 경험 분석-넷플릭스 사례를 중심으로. Journal of Integrated Design Research, 20(2), 73-87.

김헌. (2021). OTT 서비스 콘텐츠 추천 시스템 수용 저항에 영향을 미치는 요인: 넷플릭스 이용자를 중심으로. 방송통신연구, 9-46.

신지형, 김윤화. (2021). KISDI STAT REPORT 2020년 한국미디어패널 조사결과주요 내용. 정보통신정책연구원ICT데이터사이언스연구본부, 21-01호.

이유린, 윤서빈, & 안현철. (2021). 속성 카테고리 기반 감성분석을 활용한 추천시스템. 한국지능정보시스템학회 학술대회논문집, 34-35.

박현정, & 신경식. (2020). BERT 를 활용한 속성 기반 감성분석: 속성카테고리 감성분류 모델 개발. 지능정보연구, 26(4), 1-25.

박호연, 김경재. (2021). BERT 기반 감성분석을 이용한 추천시스템. 지능정보연구, 27(2), 1-15.

박호연, & 김경재. (2021). BERT 기반 감성분석을 이용한 추천시스템. 지능정보연구, 27(2), 1-15.

홍태호, 홍준우, 김은미, & 김민수. (2022). 영화 리뷰의 상품 속성과 고객 속성을 통합한 지능형 추천시스템. 지능정보연구, 28(2), 1-18

[국외 문헌]

Deldjoo, Y., Schedl, M., & Elahi, M. (2019, September). Movie genome recommender: a novel recommender system based on multimedia content. In 2019 International Conference on Content-Based Multimedia Indexing (CBMI) (pp. 1-4). IEEE.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional

transformers for language understanding. arXiv preprint arXiv:1810.04805.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Pack, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park, K., Shin, J., Kim, S., Park, L., Oh, A., Ha, J., & Cho, K. (2021). KLUE: Korean Language Understanding Evaluation. arXiv

Reddy, S. R. S., Nalluri, S., Kunisetti, S., Ashok, S., & Venkatesh, B. (2019). Content-based movie recommendation system using genre correlation. In Smart Intelligent Computing and Applications (pp. 391-397). Springer, Singapore.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wu, S., & He, Y. (2019, November). Enriching pre-trained language model with entity information for relation classification. In Proceedings of the 28th ACM international conference on information and knowledge management (pp. 2361-2364).

Yin, H., Wang, W., Chen, L., Du, X., Nguyen, Q. V. H., & Huang, Z. (2018). Mobi-SAGE-RS: A sparse additive generative model-based mobile application recommender system. Knowledge-Based Systems, 157, 68-80.

Abstract

Similar Contents Recommendation Model Based On Contents Meta Data Using Language Model

Donghwan Kim*

With the increase in the spread of smart devices and the impact of COVID-19, the consumption of media contents through smart devices has significantly increased. Along with this trend, the amount of media contents viewed through OTT platforms is increasing, that makes contents recommendations on these platforms more important. Previous contents-based recommendation researches have mostly utilized metadata that describes the characteristics of the contents, with a shortage of researches that utilize the contents' own descriptive metadata. In this paper, various text data including titles and synopses that describe the contents were used to recommend similar contents. KLUE-RoBERTa-large, a Korean language model with excellent performance, was used to train the model on the text data. A dataset of over 20,000 contents metadata including titles, synopses, composite genres, directors, actors, and hash tags information was used as training data. To enter the various text features into the language model, the features were concatenated using special tokens that indicate each feature. The test set was designed to promote the relative and objective nature of the model's similarity classification ability by using the three contents comparison method and applying multiple inspections to label the test set. Genres classification and hash tag classification prediction tasks were used to fine-tune the embeddings for the contents meta text data. As a result, the hash tag classification model showed an accuracy of over 90% based on the similarity test set, which was more than 9% better than the baseline language model. Through hash tag classification training, it was found that the language model's ability to classify similar contents was improved, which demonstrated the value of using a language model for the contents-based filtering.

Key Words : Content-based filtering, recommendation systems, natural language processing, language model, RoBERTa.

Received : November 15, 2022 Revised : November 15, 2022 Accepted : December 2, 2022
Corresponding Author : Donghwan Kim

* Corresponding author: Donghwan Kim
TVING
20F, 34 Sangamsan-ro, Mapo-gu, Seoul, Korea
E-mail: donghwan.kim9@cj.net

저 자 소 개

김동환
현재 TVING이라는 OTT 서비스 회사에서 데이터 사이언티스트로 재직중이다. 영국 셰필드 대학교에 데이터 과학 석사학위를 취득하였으며, 주요 관심 분야는 추천시스템, 자연어처리 등이다.