KoBERT를 활용한 한국 드라마 대본 대사 성별 구분

이세희○, 선금규

KAIST 문화기술대학원, 아모레퍼시픽

neptuney@kaist.ac.kr, keumkyu_sun@amorepacific.com

Gender classification of Korean drama script lines using KoBERT

Se-Hui Yi○, Gum-Kyu Sun

KAIST GSCT, Amore pacific

요 약

최근 글로벌 OTT 서비스에서 한국드라마가 세계적 인기를 얻음에 따라 드라마 콘텐츠의 가치가 높아지고 있다. 드라마 대본은 드라마 제작에 있어서 핵심이 되는 데이터로, 특히 대사에는 인물의 특성이 잘 나타나 있다. 본 논문에서는 KoBERT 모델을 활용해 드라마 대사에서 인물의 특성 중 하나인 성별을 구분하고 실험 결과를 제시한다. KoBERT 모델로 대사의 성별을 분류한 뒤, 콘텐츠 분석과 인공지능 창작 측면에서의 활용 가능성에 대해 논의한다.

주제어: 자연어처리, 드라마대본, KoBERT, 텍스트분석

1. 서론

최근 넷플릭스 등의 글로벌 OTT 서비스에서 한국 드라마가 세계적인 인기를 얻음에 따라, 드라마 제작에 핵심이 되는 대본의 중요성이 커지고 있다. 대본은 대사와 지문으로 구성되어 있는 텍스트 데이터로, 특히 대사에는 남녀노소 다양한 캐릭터들의 특성이 고스란히 반영되어 있다. 본 연구에서는 다양한 인물의 특성 중 가장 확실하게 구분할 수 있는 "성별"을 KoBERT 모델을 활용하여 분류할 수 있는지 실험하고 결과를 제시한다. 이후 본 연구의 한계점과 보완할 방법을 논의하고 향후 대본, 소설 등의 콘텐츠 분석과 창작에 활용할 수 있는 연구 방향을 모색한다.

2. 관련 연구

KoBERT 모델이 나온 이후, 콘텐츠 관련 다양한 한글 데이터에서 감정 분석 관련 연구가 이루어지고 있다[1]. 감정은 상당히 주관적인 영역으로써 분석 시 중립, 분노, 혐오, 공포, 행복, 슬픔 등으로 분류하여 해당 항목 수치의 높고 낮음을 통해 감정을 판단한다. 해당 데이터를 남긴 사람에게 직접 감정을 물어보지 못하므로 추정만 할 수 있다는 한계점이 있다. 하지만 드라마 인물 성별의 경우 주관적 영역 없이 사실을 기반으로 태깅할 수 있기 때문에 더욱 명확한 결과가 도출될 것으로 예상했다.

3. 실험 설정

3.1. 데이터 특성

본 논문에서 KoBERT[1] 모델을 활용하여 분석한 드라마 대본은 ai 허브[2]에서 제공하는 데이터셋 중 "방송 콘텐츠 대본 요약[3]" 데이터의 원본 데이터이다. 요약문을 도출하기 전의 데이터 중, 현대드라마에 해당하는 원천 데이터를 사용하였다. 해당 데이터에 드라마는 17개로 쌈 마이웨이, 여우누이던, 굿바이 솔로, 동안미녀, 강이 되어 만나리 등이 있다.

본 논문에서는 드라마 대본 안에서 성별 구분이 가능한 대사만을 골라 KoBERT 모델을 적용하였다. 따라서 상대방의 이름을 부르는 등의 너무 짧은 대사(10글자 미만), 문장부호만으로 이뤄진 대사, 여러 인물들이 동시에 외치는 대사는 제외하였다.

대사 분석에 활용된 문장 수와 문장 별 평균 글자 수, 드라마 별 인물 수는 표1과 같다. 분석에 활용된 대사의 전체 평균 길이는 34.99글자이며, 최소값은 11글자. 최대값은 306글자이다.

표1. 분석 데이터 통계

드라마명	문장 수	평균 글자 수	인물 수
강이 되어 만나리	1194	29.53	72
굿바이 솔로	223	36.49	26

[1] https://github.com/SKTBrain/KoBERT

[2] https://www.aihub.or.kr/

[3] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=591

내남자의 비밀	3379	38.63	122
동안미녀	1650	31.70	81
루비반지	13599	35.37	199
미스 몬테크리스토	994	42.35	68
빨강구두	965	34.83	61
뻐꾸기 둥지	15006	36.42	206
쌈마이웨이	496	31.06	52
어여쁜당신	2590	32.49	62
얼렁뚱땅 흥신소	81	27.34	14
여름아 부탁해	7088	27.21	91
여우누이뎐	138	21.50	28
천상의약속	14572	38.21	165
최고다 이순신	10764	34.11	133
트로트의 연인	131	21.62	20
행복한여자	3801	33.85	91

성별 분류를 위한 학습 데이터는 네이버, 구글 검색 데이터를 수집하여 그림1에서처럼 드라마 별 인물 데이터에 성별을 정리하였다.

그림1. 인물 별 성별 정리 예시

	drama_name	name	Gender	gender_group
0	내남자의비밀	해림	여	1
1	내남자의비밀	미령	여	1
2	내남자의비밀	재욱	남	0
3	최고다이순신	일도	남	0
4	최고다이순신	미령	여	1
...	...	...	...	...
1678	내남자의비밀	장단옥	여	1
1679	내남자의비밀	단옥	여	1

3.2. 데이터 전 처리

데이터에 태깅된 성별을 남자는 0, 여자는 1로 라벨링하였다. 또 각 데이터가 KoBERT에 입력값으로 들어갈 수 있도록 토큰화 및 패딩의 전처리를 시행하였다.

3.3. 파라미터 설정

모델의 파라미터는 batch size = 64, learning rate = 5e-5, epoch=5, dropout=0.5로 설정하였다. 크로스 엔트로피를 loss로 설정하여 모델은 epoch가 증가할수록 loss를 감소시키고 accuracy를 향상시킨다. 모델 성능을 평가하기 위하여 데이터를 랜덤하게 섞고 train과 test를 8:2 비율로 나누었다. 나머지 파라미터는 default 값을 적용하였다.

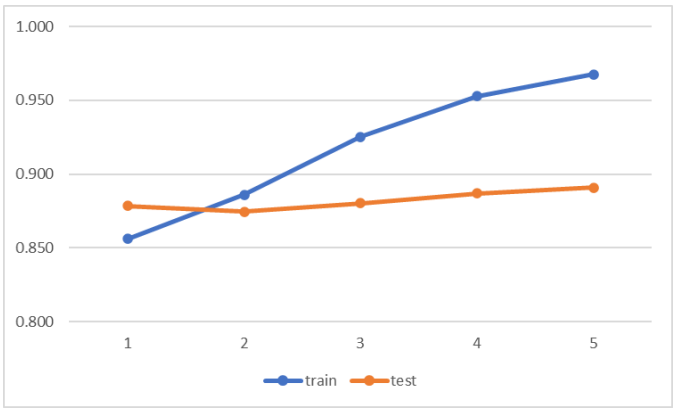
4. 실험 결과

Epoch를 추가하며 train데이터 셋으로 모델을 생성하고 test 데이터셋으로 평가하였다. 표2와 그림2에서 보이듯 Epoch를 증가시킬수록 train과 test의 성능이 향상하는 것을 확인할 수 있었다. 특히 세 번째 epoch 에서는 accuracy가 0.925이상으로 좋은 성능으로 학습되는 것을 확인할 수 있었으며, test 데이터도 해당 모델에서 accuracy 0.88로 좋은 결과를 낼 수 있었다. 하지만 train의 epoch가 증가할수록 train과 test의 성능의 차이가 커지는 것을 볼 수 있었고, 오버 피팅을 방지하기 위해서 epoch을 5회 이상으로 진행하지 않았다.

표2. 실험 결과값

epoch	train	test
1	0.856	0.879
2	0.886	0.875
3	0.925	0.880
4	0.953	0.887
5	0.968	0.891

그림2. 트레인값과 테스트값 추이



5. 결론 및 논의

본 논문에서는 한국 드라마 17편의 대본에 해당되는 대사를 KoBERT모델을 적용해 성별을 구분할 수 있는지 실험하였다. 그 결과 train, test 모두 좋은 성능으로 대사를 통해 성별을 분류할 수 있었다.

다만 17편의 드라마 대본은 16부작 미니시리즈와 100부작 이상의 장편 드라마가 혼재되어 있어 분량이 일정치 않다는 한계가 있다. 이를 보완하려면 16부작으로 구성된 미니시리즈 대본 데이터셋을 구축하거나, 50편 이상의 장편드라마 데이터셋을 구축하는 일이 필요하다. 또는 장르 별로 나누어 드라마대본 데이터셋을 구축한다면 장르별 대사의 특성을 도출할 수 있을 것이다.

또한 대본 데이터의 수가 크지 않아 한국 드라마 대사의 보편적 특성을 도출하기에는 어려움이 있다. 향후, 더 많은 편수의 드라마 대본 데이터셋이 구축될 경우 연구 신뢰도를 높일 수 있다.

위에 언급한 데이터 구축이 완성된다면 후속 연구에서는 성별 뿐만 아니라, 나이, 성격, 직업 등 다양한 인물의 특성을 분석하여 콘텐츠 분석 및 인공지능 창작 방법론에 응용할 수 있다.

본 연구로 17개 드라마에 대한 인물 성별 데이터도 구축되었으므로, 성별에 따른 인물의 감정 등을 분석하는 데에도 기여할 수 있을 것이다.

이번 실험의 파라미터 중 대사 별 최대 적용 글자수가 64로 제한되어 있고 글자수를 기준으로 대사를 분석했지만, 형태소 단위로 나누어 분류한다면 본 실험의 결과를 보완할 수 있을 것이다.

감사의 말

이 연구는 과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축된 “방송 콘텐츠 대본 요약”을 활용하여 수행된 연구입니다. 본 연구에 활용된 데이터는 AI허브(aihub.or.kr)에서 다운로드 받으실 수 있습니다.

참고문헌

[1] 최다은, 김효민, 이혜린, 황유림 “KoBERT 기반 Youtube 자막 감정 분석 연구” ASK 학술발표대회 논문집(29권 1호) pp. 513-516, 2022

[2] 노윤석, 곽창욱, 김선중, 박성배, 이상조, "토픽 모델을 이용한 방송 대본 분석 사례 연구" 제27회 한글 및 한국어 정보처리 학술대회 논문집pp.228-230,2015

[3] 최선주, 박명관, 김유희, “KoBERT와 KR-BERT의 은닉층별 통사 및 의미 처리 성능 평가” 제33회 한글 및 한국어 정보처리 학술대회 논문집 pp.340-345, 2021

[4] 나승훈, et al. Deep Biaffine Attention을 이용한 한국어 의존 파싱, KCC, pp. 584-586, 2017