토픽 모델을 이용한 방송 대본 분석 사례 연구

노윤석[12], 곽창욱[2], 김선중[1], 박성배[2], 이상조[2]
한국전자통신연구원 방송통신미디어연구소 스마트미디어플랫폼연구실[1]
경북대학교 IT대학 컴퓨터학부[2]
{ysnoh, cukwak, sbpark, sjlee}@sejong.knu.ac.kr[2], kimsj@etri.re.kr

A case study of a broadcast script by using topic model

Yunseok Noh[12], Chang-Uk Kwak[2], Sun-Joong Kim[1], Seong-Bae Park[2], Sang-Jo Lee[2]
ETRI[1], Kyungpook National University[2]

요 약

방송 대본은 방송 콘텐츠에 대해 얻을 수 있는 가장 주요한 텍스트 데이터 중에 하나이다. 본 논문에서는 토픽 모델을 통해 방송 대본 분석을 수행하고 그 결과를 제시한다. 방송 대본을 토픽 모델로 학습하기위해 대본의 장면 단위로 문서를 구성하여 학습하여 대본의 장면을 분석하고 등장인물 단위로 문서를 구성하여 등장인물을 분석하여 그 특징을 살펴본다. 토픽 모델을 사용하여 방송 대본을 분석하는 과정에서 방송 대본이 가지는 특징을 분석하고 그로부터 향후 연구방향에 대해 논의한다.

주제어: 방송대본, 토픽모델, 텍스트 분석

1. 서론

방송 대본은 방송 콘텐츠의 제작에 있어서 기본이 되는 글로써 방송 대본 분석을 통해 해당 방송의 내용을 파악하고 정보를 제공하는데 이용할 수 있다. 최근에는 방송 콘텐츠가 단순히 실시간 방영에 그치지 않고 인터넷, IPTV 등의 다양한 매체를 통한 VOD 서비스 및 콘텐츠 검색, 정보 제공 서비스 등으로 확대 활용되고 있다. 이런 환경에서 콘텐츠 분석 및 정보 제공을 위해 이용할 수 있는 텍스트 데이터로서 방송 대본은 중요한 위치를 차지한다. 본 논문에서는 1회차 분량의 드라마 대본에 대한 구체적인 사례 연구를 수행하였다. 토픽 모델[1]은 텍스트 데이터를 분석하기위한 효과적인 도구이며, 본 연구에서는 토픽 모델을 통해 방송 대본을 분석한 결과를 제시하고 앞으로 연구 방향을 모색한다.

2. 실험 설정

본 논문에서 토픽 모델을 통해 분석한 방송 대본은 SBS 드라마 ‘ 풍문으로 들었소’ 6화 (이하 풍문 6화)이다. 드라마 1회 분량의 대본을 토픽 모델로 효과적으로 분석하기 위해 두 가지 방법으로 대본 데이터를 구성하였다. 드라마 대본은 일반적으로 장면을 나타내는 식별자에 의해 장면 단위로 구분할 수 있으며 각 장면은 하나 혹은 두 개 정도의 이야기를 포함하고 있어 의미적으로 적절한 문서 단위가 될 수 있다. 따라서 각 장면이 하나의 문서가 되는 문서 집합을 토픽 모델 학습을 위한 데이터로 구성할 수 있다. 분석 대상이 된 풍문 6화 대본은 총 72개의 장면 문서로 구성된다.

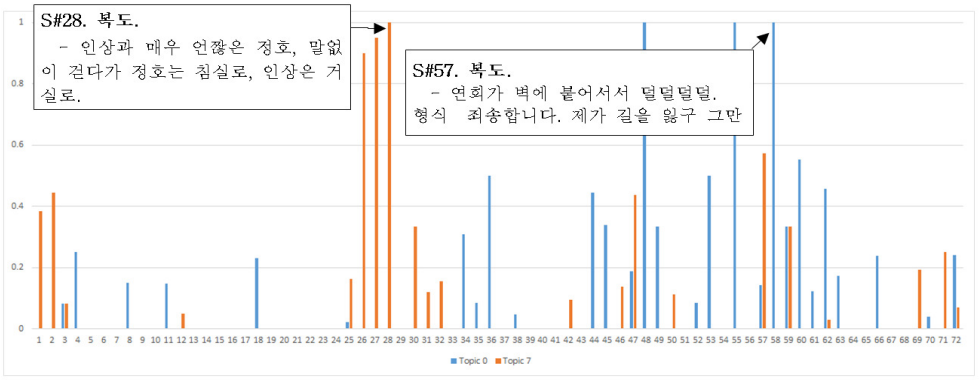
등장인물 분석을 위해 각 등장인물 별로 대사를 모아 의사 문서 (pseudo-document)를 구성하는 방식을 채택하였다. 드라마는 몇 가지 사건으로 구성되며 각 등장인물은 제각기 서로 다른 사건에 관여한다. 드라마 대본에서는 각 등장인물이 대사를 통해 드라마의 사건과 관계를 맺게 된다. 따라서 등장인물 문서의 토픽을 살펴봄으로써 각 등장인물의 특징과 드라마 내 사건과의 관련성, 등장인물 간의 관계 등을 분석할 수 있을 것이다.

표 1 데이터 통계

	문서 수	단어 수	어휘 수	문서별 평균 단어 수
장면 문서	72	1,763	786	24.48
등장인물 문서	21	960	555	45.71

장면 문서의 경우 극도로 짧은 문서가 다수 존재하며 하나의 장면으로 취급될 수 있는 장면들이 방송 대본의 특성상 여러 장면으로 분할되어있는 경우가 많다. 반면 등장인물 문서의 경우 상대적으로 짧은 문서가 많이 생기지는 않으나 전체 문서 수가 크게 적으며, 풍문 6화의 경우 21개의 등장인물 문서가 만들어진다. 본 논문에서는 효과적으로 토픽 모델을 학습하기 위해 형태소 분석 후 명사만을 사용하였고 불용어 (stop words)를 제거하였다. 장면 문서의 경우 각 대사에 대한 등장인물 정보 역시 제거하였다. 표 1은 데이터 전처리를 수행한 후의 풍문 6화의 데이터 통계이다.

본 논문에서는 방송 대본의 토픽 학습을 위해 비모수 모델(nonparametric model)인 Hierarchical Dirichlet Process (HDP) [2]를 사용하였다. 토픽 모델은, 문서 데이터의 경우, 의미적으로 관련이 있는 단어들을 같은 토픽으로 군집화하여 표현하고 개개의 문서를 토픽들의 확률 분포로 나타내는 베이지언 확률 모델이다. 따라서 주어진 문서의 토픽 분포는 해당 문서를 의미적으로 표현한 것으로 이해할 수 있으며, 이러한 의미 정보를 통해 데이터를 분석하고 여러 응용문제에 활용하는 것이 가능하다. 일반적으로 토픽 모델은 학습 시 데이터에서 추출할 토픽의 개수를 사용자 매개변수로 입력받는데 학습할 데이터에 대해 최적의 토픽 개수를 사전에 아는 것이 매



Topic 0	Topic 7
진애	아기
형식	정호
인상	소리
웃음	인상
연희	침대
정호	혜옥
죄송	거실
마루	울음
부부	침실
진짓상	오르골

그림 1 지문 토픽의 분포

우 어렵다. 이 문제를 해결하는 모델이 비모수 토픽 모델이며 본 논문에서 사용한 HDP는 가장 대표적인 비모수 토픽 모델 중 하나이다.

3. 방송대본의 토픽 분석

장면 문서로 토픽을 풍문 6화를 학습한 결과 62개의 토픽이 형성되었다. 먼저 생각해볼 것은 데이터 필터링을 거친 72개의 문서로부터 62개의 토픽이 학습되었다는 것이다. 일반적인 경우보다 데이터의 양에 비해 많은 토픽이 학습되었으며, 학습된 토픽을 관찰한 결과 크게 지문 토픽, 장면 특화 토픽, 줄거리 토픽으로 분류할 수 있었다. 첫 번째로 전체 대본의 약 30%를 차지하는 지문을 나타내는 토픽이 형성되었다. 드라마 대본에서 지문은 주로 해당 장면의 장소 묘사와 등장인물들의 행동 묘사를 포함한다. 등장인물의 이름과 장소명이 지문 토픽에서 높은 확률을 가지는 것이 이러한 대본의 특성을 반영한다. 대부분의 장면 문서가 하나 이상의 지문을 포함하고 있으며 짧은 지문으로만 구성된 장면 역시 다수 존재한다. 그림 1은 풍문 6화에 나타난 두 지문 관련 토픽의 장면 문서에 대한 분포를 나타낸 것이다. 대다수의 장면에 지문 토픽이 분포하는 것을 확인 할 수 있으며 몇몇 장면에서는 지문만 나타나는 것 역시 알 수 있다. 그림 오른쪽은 지문 토픽에 나타나는 단어 상위 10개를 보인 것으로 등장인물 이름과 장소명 등이 주로 나타나는 것을 볼 수 있다. 이를 통해 지문이 주로 장면의 장소와 등장인물 행동 묘사에 주력한다는 것을 유추할 수 있다.

두 번째 유형은 장면 특화 토픽이다. 등장인물의 대사가 주 내용을 이루는 방송 대본의 특성상 여러 장면에 반복적으로 나타나는 단어가 일반적인 문서에 비해 상대적으로 적다. 풍문 6화의 경우 전체 786개 어휘 중 560개 어휘가 1개 장면에서만 나타났다. 이러한 특징이 반영된 것이 장면 특화 토픽으로, 장면 특화 토픽은 1개의 장면에서만 나타나며 해당 장면에 나타나는 단어들로만 구성된다. 이 장면 특화 토픽이 학습 문서 수에 비해 많은 토픽이 생성되는 원인이라 할 수 있다.

마지막 유형은 토픽 모델을 학습했을 때 일반적으로 관찰할 수 있는 토픽이다. 이러한 토픽들은 방송 대본에

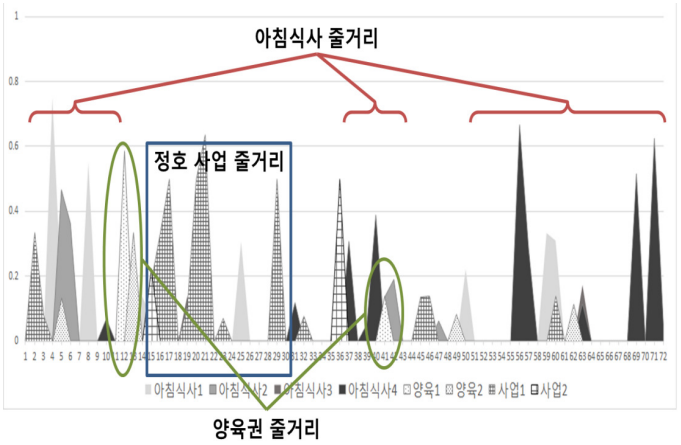


그림 2 줄거리 토픽의 분포

내포된 여러 이야기 줄거리들과 직간접적으로 관련된 토픽이다. 풍문 6화의 경우 '양가 부모님의 아침 식사', '아기 양육권', '정호의 사업' 이라는 큰 세 가지 줄거리로 이야기가 진행된다. 표 2는 풍문 6화를 구성하는 세 가지 주요 줄거리를 잘 나타내는 토픽 8개를 보인 것이다. 주요 줄거리와의 관련성은 드라마를 본 사람이 직접 평가한 것으로, 드라마를 직접 보지 않았더라도 일견 동의할 수 있을 정도로 토픽이 줄거리를 잘 묘사함을 확인할 수 있다. 줄거리 토픽을 통해 드라마의 줄거리 이해 및 이야기 구성 방식 등을 분석할 수 있으며 이는 다음 장에서 다시 설명한다.

3.1 방송 대본의 줄거리 구성

풍문 6화는 앞서 언급한대로 크게 세 가지의 줄거리로 대본이 구성되어 있다. 학습된 62개의 토픽 중 각 줄거리와 밀접한 연관이 있는 대표적인 토픽 8개를 선택하여 풍문 6화의 줄거리 구성을 살펴보았다.

그림 2는 풍문 6화의 주요 줄거리 토픽의 분포를 나타낸 것이다. 단색 계열은 '아침 식사' 줄거리, 점 패턴 계열은 '양육권' 줄거리, 체크 패턴 계열은 '사업' 줄거리를 나타낸다. 그림을 통해 드라마 초반부와 후반부에 '아침 식사' 이야기가 진행되고 드라마 중반에 '정호의 사업' 이야기가 진행되며 양육권 관련 줄거리

표 2 주요 줄거리 토픽의 상위 10 단어

	아침식사1	아침식사2	아침식사3	아침식사4	양육1	양육2	사업1	사업2
1	정순	사람	식사	밥상	소송	서재	정호	우두
2	선숙	연희	화장	식전	양육	육아	양비서	대산그룹
3	집사	전화	준비	술잔	파악	자리	어깨	무리
4	왜건	식사	마련	진영	옆구리	유머	안내	긍정
5	식당	아침	신고	경지	조건	격식	백대헌	플로우
6	주방	걸레	혼인	한말씀	청탁	밤참	태우	sns
7	조림	월요일	통화	전자	술값	거울	탁자	후반
8	갈치	각서	할머니	찻상	넥타이	조언	애기	
9	목례	통화	오찬	마루	핸드폰	자판	목례	
10	경태	담당	귀걸	전통	여건	부탁	대산	

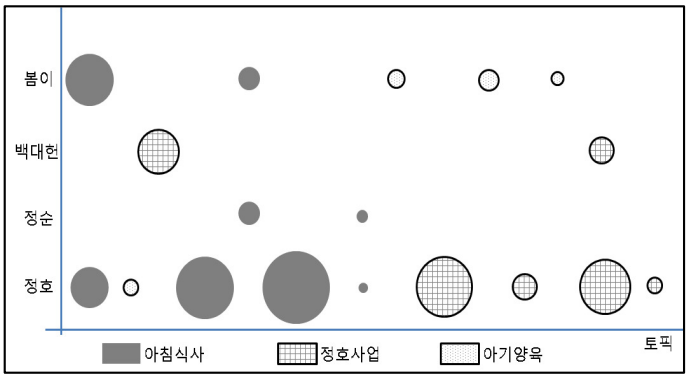


그림 3 등장인물들의 주요 줄거리 토픽 분포

가 중간 중간 나타나는 것을 확인할 수 있다. 이런 형태의 이야기 진행은 신문이나 백과사전, 기술문서 등에서 나타나는 일자 진행과는 다르며, 문학에서 나타나는 전형적인 비선형적 이야기 진행 형태라 볼 수 있다.

3.2 등장인물 분석

토픽 관점에서 등장인물을 분석하기위해 등장인물의 대사만을 남긴 장면 문서를 학습하고 등장인물 의사 문서의 토픽을 추론하는 과정을 수행하였다. 그림 3은 주요 등장인물인 봄이와 정호, 보조 등장인물인 정순과 백대헌의 주요 토픽을 나타낸 것이다. 각 가로축은 개별 토픽을 나타내고 원의 크기는 등장인물과 토픽의 관련성 정도를 나타낸다. 그림을 통해 먼저 확인할 수 있는 것은 주요 등장인물과 보조 인물 간의 역할 차이를 관여하는 토픽의 수로 확인할 수 있다는 것이다. 주요 등장인물 중 한 명인 봄이의 경우 '양육권'과 '아침 식사' 줄거리 관련 토픽들로 표현되는 반면 보조 인물 중 한 명인 정순의 경우 '아침 식사' 줄거리 관련 토픽들로만 표현된다. 마찬가지로 정호는 '아침 식사'와 '정호의 사업', '양육권' 관련 토픽이 고루 나타나 드라마의 주요 줄거리에 모두 관여하는 명실상부 주 등장인물임을 유추할 수 있다. 그러나 백대헌의 경우 '정호의 사업' 줄거리 관련 토픽에 국한되는 등장인물이다. 그마저도 정호에 비해 더 적은 수의 '정호의 사업' 줄거리 토픽과 관련을 맺고 있는 것이 드러나 풍문 6화에서 보조적 인물로 기능하는 것을 알 수 있다.

4. 결론 및 논의

본 논문에서는 방송 대본을 토픽 모델을 통해 분석하고 그 결과에 대해 논의하였다. 장면 문서와 등장인물 문서, 두 가지 다른 방법으로 문서를 구성한 후 토픽 모델을 학습하였으며 그 차이를 비교하였다. 토픽 모델을 통해 방송 대본의 형태와 줄거리 구성을 분석하였으며 등장인물의 토픽 분석 역시 수행하였다. 그러나 방송 대본의 특성상 대본 그 자체를 토픽 모델을 통해 분석하기에는 명확한 한계가 따른다. 우선 일반적으로 사람이 떠올리는 '토픽'은 추상화를 거친 줄거리나 주제 등에서 비롯되지만 대본의 경우 대부분이 대사와 지문으로 구성되어있다. 그 결과 추출된 토픽이 쉽게 이해할 수 있는 개념적인 형태를 갖추지 못하는 경우가 생기고 이는 토픽 모델의 품질에 직간접적으로 영향을 미칠 수 있다. 이를 보완하기 위해 방송 대본 외부의 데이터를 활용하기위한 고민이 필요하다. 대사와 지문으로 이루어진 대본보다는 더 추상화된 내용을 담고 있는 외부데이터의 활용은 학습된 토픽이 좀 더 사람이 해석하기 용이한 형태로 되도록 도움을 줄 것이며 전반적으로 양이 많지 않은 방송 대본 데이터를 양적으로도 보완할 수 있을 것이다.

감사의 글

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.B0125-15-1002, 개방형 미디어 생태계 구축을 위한 시맨틱 클러스터 기반 시청상황 적응형 스마트방송 기술 개발)

참고문헌

[1] D. M. Blei, "Probabilistic topic models." *Communications of the ACM* 55,4, 2012.

[2] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes." *The American statistical association*, 2006.