spaCy를 이용한 TV 드라마 대본의 주인공 및 배경 분석

유민경○, 장수지, 배병철
홍익대학교 게임학부
{mingyeongyu8, klop100418}@gmail.com, byuc@hongik.ac.kr

An Analysis of Main Characters and Settings of a TV Drama Script Using spaCy

Minkyeong Yu○, Suji Jang, Byung-Chull Bae
School of Games, Hongik Universtiy

요 약

본 논문은 드라마 <또 오해영>의 스크립트를 대상으로 spaCy툴을 이용하여 각본 내 캐릭터, 장소, 시간을 중심으로 상호 관계성을 파악하는 연구 방법을 기술한다. 이를 위해 캐릭터의 등장 빈도를 중심으로 주요 캐릭터들을 분류하고, 캐릭터별 근접 인덱스에서 나타나는 장소와 시간과의 관계를 파악하였다. 분석 결과, 자주 등장하는 캐릭터일수록 드라마의 배경을 더 잘 파악할 수 있었고, 분석한 NER(Named Entity Recognition) 간의 관계를 통해 직업을 추출한 실험에서 80%의 정확도를 보였다. 본 논문에서는 NER 분석을 통해 배경, 장소, 시간 등의 정보를 추출하여 드라마 스크립트 내 주요 캐릭터들간의 관계를 파악하는 연구 방법을 제시한다.

1. 서 론

문화의 발전과 함께 장르를 불문하고 다양한 문서 데이터들이 계속해서 생성되고 있다. 이러한 문서를 대상으로 조사, 비평, 분석, 연구 등이 꾸준히 이루어지고 있으며, 딥러닝 등과 같은 다양한 알고리즘을 통해 문서 분석이 가능하다[1], 특히, '스토리'를 대상으로 한 연구의 경우 해석에 있어 주관 혹은 사전 지식을 활용한 맥락 이해가 필요하다는 특징으로 인해 타 문서들에 비해 연구가 상대적으로 적은 성향을 띠고 있다[2]. 또한, 스토리의 특성상 자유로운 배경 전환과 다양한 캐릭터의 등장이 이루어지므로 학습을 위한 자료 축적량이 충분하지 않아 의미 있는 데이터를 얻기가 어렵다. 그러나 스토리는 작품을 이끌어나가는 중심축이고 작자와 캐릭터의 세계관을 형성하기 때문에 매우 중요한 연구 주제이다.

최근 딥러닝 기술을 이용한 텍스트 분석 연구는 감정 분류[3], 이벤트 추출[4] 등 다양한 분야에서 연구되고 있지만 특히 맥락 정보 추출에 대한 연구가 활발히 진행중이다. 정확한 맥락 정보 추출을 하기 위해서는 언어적 맥락과 상황적 맥락을 두루 고려해야 한다.

본 논문은 spaCy툴을 이용하여 TV 드라마 <또 오해영 에피소드 1>에 대한 맥락 이해의 기반이 되는 등장인물과 시간, 장소 데이터 간의 의미 있는 데이터를 추출하여 등장인물 간의 관계를 분석하고자 한다.

2. 관련 연구

2.1 자연어 처리 연구

자연어 처리 기능을 제공하는 여러 라이브러리를 이용해 많은 연구가 진행되고 있으며, 본 연구에서 사용하는 spaCy는 파이썬 기반의 자연어 처리를 위한 오픈 소스 라이브러리이다. 인공지능을 이용하여 자연어 처리를 할 때, 각각의 라이브러리가 지원하는 개체명 인식기를 사용해 개체명 인식(NER; Named Entity Recognition)을 처리한다. 개체명 인식(NER)이란 코퍼스로부터 단어가 어떤 유형인지 식별하는 것을 말하며, 다양한 분야에서 응용 가능하다 (예:의도 인식 기반의 챗봇 제작 [5]).

자연어 처리에서 spaCy와 함께 많이 사용되는 라이브러리로서 nltk (natural language tool kit [6])가 있으나, 기존 nltk에서는 제공해주지 않던 Python에서의 Dependency parsing 기능을 제공해주고 속도 측면에서도 빠른 spaCy의 사용이 증가하고 있다. 본 연구에서는 사용의 편리성 등을 고려하여 spaCy를 기반으로 분석을 진행한다.

2.2 맥락 분석

문맥 구성에 있어 기초가 되는 것은 바로 '단어'이다[7]. 그러므로 각각의 단어를 개체명으로 인식하게 함으로써 상호 간의 관계를 찾을 수 있다. 서술어로 사용되는 동사와 형용사는 감정을 분석하는 데 용이하며, 인물이나 사물을 지칭하는 명사구는 대화 맥락을 이해하는 데 효과적이다. 문맥 정보 추출을 위해 명사구 기반의 지식그래프를 이용한 연구 [8]가 있었으며, 단어 벡터를 이용한 자연어처리 기반을 적용한 사례 [9]도 있다. 이러한 연구들은 인간의 별도의 지도가 없이도 기계가 스스로 스토리를 이해하게 하는 것을 목적으로 하고 있다고 할 수 있다.

3. 연구 방법

3.1 데이터 셋

본 연구는 스토리, 주인공 및 배경 분석을 위해 TV 드

라마 <또 오해영 에피소드 1>의 영어 스크립트를 이용하였다. 드라마의 특성상 등장인물당 한 번의 발화 길이가 2문장에서 4문장으로 구성되며 대명사 및 상대적인 시간을 나타내는 시간 직시(Time Dexis) 표현 등이 다수 등장한다. 시간 직시 표현은 주요 등장 인물과 시간 및 공간적 배경 정보 등 문락 분석에 활용될 수 있는 중요한 정보를 포함하고 있다. 예를 들어 "철수가 내일 그곳에 간다."란 시간 직시 표현이 있다. 그 후 설정해 둔 범위 내에서 "철수가 10월 12일 학교에 간다."란 비직시적 표현이 재등장하면, 서로간의 문맥 정보를 비교해 '내일 = 10월 12일', '그곳 = 학교'란 데이터를 얻을 수 있다.

다수의 시간 직시 표현으로 일정 부분 추가적인 데이터 학습 과정이 필요하나, TV에 시각적으로 보여야 하므로 자주 등장하는 만큼 주요 인물, 시공간적 배경일 가능성이 높으므로 본 연구에 활용하기 용이하다.

3.2 spaCy툴을 이용한 주요 인물, 시공간 추출 방법

문서의 핵심어를 추출하기 위해 문서 내에 등장한 단어의 빈도수를 활용하는 것은 대표적인 핵심어 추출 방법 중 하나이다. 어떤 단어가 특정 문서 내에서 차지하는 중요도를 나타내는 TF-ID에서도 단어의 빈도수(Term frequency)를 고려한다. 본 연구는 이와 같은 방식으로 핵심어를 추출하며, 그림 1의 3단계 과정을 거쳐 정보를 추출한다. 1단계에서는 대문자와 소문자를 구별하지 않기 위해 모든 문자를 소문자로 통일하고, spaCy에서 정의해 둔 stopword 및 정의되지 않은 일부 구두점이나 특수문자를 제거했다. spaCy는 영어 데이터셋을 중점적으로 다루도록 만들어졌기 때문에 한국의 location data를 수작업을 통해 LOC로 학습시키고, sat, sun과 같은 일부 표현을 TIME으로 추가 학습시켰다. 마지막 단계에서는 앞의 과정을 통해 학습된 데이터를 이용하여 캐릭터별 등장 빈도를 분석하고, 빈도가 높은 순으로 10명을 추출하여 그들과 관련된 장소와 시간을 추출하였다.

드라마는 여러 문장에 걸쳐 한 주제의 대화가 종결되므로 인접한 단어 사이의 거리를 특수문자, 공백을 포함한 500자 내로 설정하였다. 이러한 과정을 마친 후, <spaCy기반의 캐릭터, 장소, 시간 추출의 정확도>를 분석하고, 캐릭터별로 자주 등장하는 장소, 시간 순으로 표를 작성하여 그들과의 관계를 파악하였다.

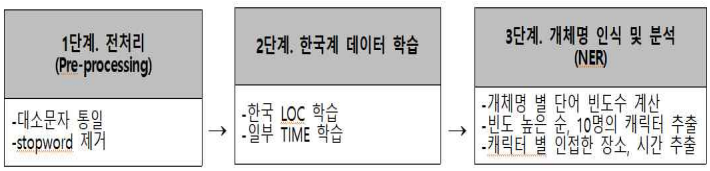

그림 1. SpaCy를 이용한 주요 인물, 시공간 추출 3단계

3.3 결과

연구 결과 캐릭터, 장소, 시간별 spaCy의 정확도는 캐릭터: 66.13%, 장소: 82.14%, 시간: 86.54%의 정확도를 보였다 (표1 참조). 특히 캐릭터에서 66%라는 비교적 낮은 수치를 보였는데, 분류된 내용의 분석 결과, 감탄사나 줄임말, 신조어를 제대로 분류하지 못하는 것으로 나타났다. 특히, "Ohhhhhhh", "Aja-ja-ja-ja"와 같은 감탄사에서 취약함이 두드러졌다.

캐릭터마다 이름과 인접한 장소를 추출한 결과, 각각 10~40개의 장소가 등장했다. 이는 단순히 언급만 한 장소도 포함하는 결과로, 드라마 특성상 화면 간의 전환이 빈번함을 고려하더라도 상당히 많은 장소가 등장함을 알 수 있다. 장소만으로는 단순히 캐릭터가 이 장소에 가본 적이 있거나 있는 상태임을 짐작만 할 수 있으므로, 민주언론시민연합 방송모니터위원회가 2019년 10개 방송사 드라마 110편의 주요 등장인물 447명의 직업을 분석해 발표한 자료를 이용해 딕셔너리를 수작업으로 만들어 이용하였다. 캐릭터 '이름-장소-직업'이 인접되어 있을 경우 캐릭터의 직업을 추측하는 방식이었으며, 이에 대한 정확도는 80%로 상당히 높은 결과를 보였다. 'Anna'의 경우 '△'로 직업 일치 여부를 나타냈는데, 드라마에서 Anna라는 인물은 convenience store의 manager 또는 staff로 등장하는 것이 맞지만 직업 딕셔너리에 정의된 staff와 의미에서 약간의 차이가 있기 때문에 이와 같이 평가하였다. 또한, Dokyung, Haeyoung 캐릭터에서 자주 등장하는 장소로 house, WORK_OF_ART가 있었는데 드라마의 주인공이 공통적으로 자주 등장하는 것을 바탕으로 두 캐릭터 간의 접점 혹은 중심배경으로 추측할 수 있었다. 실제로 드라마의 내용을 보면 하우스 쉐어링(house sharing)을 통해 두 인물이 가까워짐을 알 수 있다.

표 1. 캐릭터, 장소, 시간별 spaCy 정확도

	캐릭터	장소	시간
정확도(%)	66.13	82.14	86.54

표 2. 추출된 직업의 일치 여부

	Haeyeong	Sugyeong	Pretty Haeyeong	Jinsang
추출된 직업	representative	director	TF team leader	lawyer
일치 여부	x	o	o	o

	Taejin	Gyeongsu	Sangseok	Dokyung	Hoon	Anna
추출된 직업	entrepreneur	x	engineer	director	staff	staff
일치 여부	o	x	o	o	o	△

드라마는 앞서 언급하였듯이 상대적인 시간표현이 많으므로 today, yesterday 등과 같은 단어만 가지고서는 의미 있는 데이터를 추출하기 어려웠다. 따라서, 장소-시간-캐릭터가 인접해 있는 경우를 별도로 추출하였다. 이때, 시간 사이의 관계성 파악을 위해 추출된 TIME별로 캐릭터의 등장횟수를 표현하였는데, 그 결과 주인공 해영(Haeyoung)과 도경(Dokyung)의 경우 오늘(today) 또는 내일 (tomorrow)과 같은 현재시제와 가까운 시간대의 단어에 자주 등장했다. 이는 주인공의 행동을 중심으로 현재 상황이 진행되어가는 드라마의 특징과 부합한 특징을 보인다. 또한, Haeyoung의 경우 며칠 전에 (a few days ago)와 같은 과거를 회상하는 단어가 초반에 다수 등장했는데, 드라마 스크립트의 내용을 확인한 결과 해영의 결혼이 무산되고 과거 일을 자주 회상했기 때문이다.

4. 결론 및 향후 과제

본 연구는 TV 드라마 대본을 기반으로, 자연어처리 라이브러인 spaCy를 이용해 개체명 분석을 수행하였다, 분석 결과, 캐릭터에 대한 분류는 66.13%이며, 장소에 대한 분류는 82.14%, 그리고 시간에 대한 분류는 86.54%의 정확도를 얻었으며, ORG와 PERSON에 대해 상대적으로 낮은 정확성을 보였다. spaCy는 영어를 중점적으로 다루기 때문에 한글 언어 지원에 대한 지속적인 업데이트가 필요할 것으로 보이며, 이 점이 개선되면 보다 정확한 결과를 얻을 수 있을 것이다. 본 연구에서는 명사만을 중심으로 관계를 추측했으나, 동사구와 형용사구 등 전체적 의존성 파악을 추가하는 것이 향후 스토리 맥락 이해에 좀 더 도움이 될 것이라고 예상한다.

본 논문에서는 비록 명사구만 분석했으나, 기존의 문장 내의 관계를 통해 맥락 이해를 하려던 시도를 넘어, 드라마 대본에서의 NER 분석을 통해 캐릭터와 배경 간의 관계성을 추측하고자 한 것에 의의가 있다고 할 수 있다. 향후 연구로는 동사구와 형용사구 사이의 관계를 통해 전체적인 의존성을 파악하고, 다양한 언어 모델과 알고리즘을 추가적으로 적용할 계획이다.

Acknowledgement

This work was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science of Korea(NRF) funded by the Ministry of Science and ICT(2017R1A2B4010499) and Institute for Information & communications Technology Promotion(IITP) grantfunded by the Korea government (MSIT)(No.2017-0-01772, Development of QA systems for video Story Understanding to pass the Video Turing Test).

참고 문헌

[1] P Parvathi, T S Jyothis,“Identifying Relevant Text from Text Document Using Deep Learning”, ICCSDET, 2018

[2] 이수경, 박규병, “딥러닝이 탐구하지 못한 언어와 5가지 태스크”, 카카오AI리포트, 2018

[3] Qi Wang, Lei Sun, Zheng Chen,“Sentiment Analysis of Reviews Based on Deep Learning Model”, ICIS, 2019

[4] Guandan Chen, Qingchao Kong, Wenji Mao,“Online event detection and tracking in social media based on neural similarity metric learning”, ISI, 2017

[5] 김용기, 이창희, 이정민,“쇼핑몰 지능형 챗봇의 자연어 처리를 위한 패션쇼핑 개체명 인식 사전 구축”, 한국경영과학회, 2018

[6] Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. 2009. O’Reilly Media

[7] 고창수, “자연어 처리의 현황과 전망”, 우리말학회, 2012

[8] 유소엽, 정옥란, “BERT와 지식 그래프를 이용한 한국어 문맥 정보 추출 시스템”, 한국 인터넷 정보학회, 21권, 3호, 123-131, 2020

[9] Monisha Kanakaraj, Ram Mohana Reddy Guddeti, “NLP based sentiment analysis on Twitter data using ensemble classifiers”, IEEE, 2015