문화유산 기반 말뭉치 구축 시스템에 관한 연구

박찬우*, 송제호**
*한국전자통신연구원
**전북대학교 융합기술공학부
e-mail:gamer@etri.re.kr

A Study on the Cultural Heritage-based Corpus Building System

Chan-Woo Park*, Je-Ho Song**
*Electronics and Telecommunications Research Institute
**Dept. of Convergence Technology Engineering, Jeonbuk National University

요 약

본 논문에서는 텍스트 기반의 문화유산에서 의미있는 단어를 자동으로 분류하기 위해서는 해당 문화유산의 상세설명에서 각 단어의 의미를 파악하고 관계를 정의해줘야 한다. 하지만, 국립중앙박물관에서 운영하는 e-뮤지엄 사이트 기준 200만건이 넘는 문화유산 데이터의 관계를 사람이 일일이 정의하는 것은 현실적으로 인력, 시간이 많이 소요되기 때문에 쉽지 않은 것이 현실이다. 본 논문에서는 각 문화유산간의 관계 연결을 딥러닝을 활용할 것을 염두에 두고, 지도학습용 데이터 생성을 직관적으로 하기 위한 어노테이션 시스템을 제안한다.

1. 서론

일반적인 언어 도메인에서 사용되는 단어는 문화유산이라는 도메인에서 사용되는 단어가 다르기 때문에, 개체명 인식(NER: Named Entity Recognition, 이하 NER)의 성능이 제대로 나오지 않는다. 딥러닝의 특성상 말뭉치(자연언어 연구를 위해 특정 목적을 가지고 언어의 표본을 추출한 집합)의 질과 양에 따라 구축되는 시스템의 결과값에 큰 영향을 끼치게 된다[1].

기계학습의 전처리 단계에서 데이터의 노이즈 혹은 서로 모순되는 내용을 태깅하게 되면 데이터의 일관성을 잃게 되어, 딥러닝 학습에 의한 자동분류가 제대로 되지 않는 문제가 있다. 이처럼 태깅시스템 구축의 기획 단계에서 '어떻게 하면 전통문화라는 도메인에서 질과 양을 충족하는 말뭉치를 생성할 수 있을까?' 에 대한 문제가 있었다. 개체명 인식의 사전학습을 위한 말뭉치가 중요한 이유는 개체명이 인명, 지명, 기관명 등과 같은 고유한 개체를 나타내는 고유명사나 명사를 뜻하기 때문이다. 이처럼 소스 데이터의 정확성을 높이기 위해 일반인에 비해 전통문화의 도메인을 잘 이해하고, 해석할 수 있는 한국전통문화대학교의 이종욱 교수팀(문화유산 산업학과)과 협업하여 단어의 속성 및 관계에 대한 부분을 정의하고 태깅 작업을 진행함으로써 데이터의 전처리단계에서 데이터의 정확성을 확보할 수 있게 되었다[2-3].

2. 본론

2.1 문화유산 데이터 획득

공공데이터 포털을 통해 '문화유산' 키워드로 검색된 데이터들 중 국립중앙박물관에서 제공하는 이뮤지엄 데이터가 본 논문에서 필요로하는 데이터였다. 이뮤지엄에서 제공하는 「Open API이용가이드」라는 웹 페이지를 살펴보면, '[이뮤지엄]기술문서양식_170413.docx' 를 다운받을 수 있다. 해당 기술문서에서 '소장품 상세조회' 를 통해 다양한 정보를 확인할 수 있다. Call back URL에 serviceKey와 id를 입력해주면 원하는 유물정보의 획득이 가능하다. 여기서 ServiceKey는 공공데이터 포털에서 발급받은 인증키를 입력하고, id는 이뮤지엄에서 사용하는 단일 소장품의 고유id값을 말한다. 열거한 방법으로 요청메시지를 이뮤지엄 서버로 보내면 하기와 같은 정보를 받을 수 있고, 이를 바탕으로 MariaDB를 생성하였다.

2.2 문화유산 데이터 속성 정의

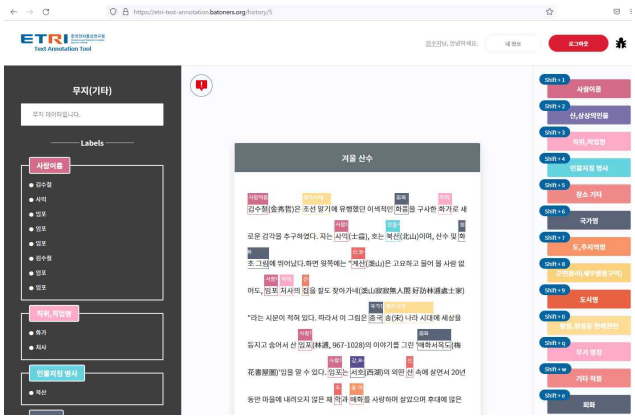
자연물, 인공물, 서체/그림체, 인물, 형용사 등으로 속성값을 정의하였고, 이를 기반으로 말뭉치 데이터를 구축을 위한 태깅 시스템을 개발하였다[4-5].

개발 과정에서 국립국어원에서 발표한 '개체명 분석 말뭉치 구축' 이라는 것을 접하게 되었고, 이를 활용하여 기계학

습의 결과를 향상시킬 방법을 모색하게 되었다. 국립국어원에서 무료제공하는 '모두의 말뭉치'는 신문, 일상 대화 데이터 300만 어절(문어 200만 어절, 구어 100만 어절)이 학습된 말뭉치이다. 이는 15개 대분류와 141개 세분류로 구분되어 있다. 기 학습된 말뭉치 데이터에 더해 '전통문화' 데이터를 태깅해 더하면 성능향상을 기대할 수 있다. 사전학습된 데이터를 활용해 성능향상이 이루어진 연구들 역시 많이 있기 때문에, 이 방법을 활용하기로 하였다[6].

2.3 어노테이션 시스템 구현

e-뮤지엄에서 크롤링한 문화유산의 텍스트 데이터 중, 1,048,575건의 대해 db dump를 완료하였고, 이 과정에서 작업의 편의성을 높이기 위해 의/식/주 등 17개의 사전 정의된(용도분류코드 2단계) 분류에 의해 나눠 놓았다. 데이터 태깅은 저장된 데이터를 불러와 해당 단어를 드래그한 후, 사전 정의된 92가지 세분류 중 하나를 우측에서 선택할 수 있게끔 만들었으며, 태깅된 데이터는 좌측에서 따로 세분류별로 확인이 가능하도록 만들었다.



[그림 1] 어노테이션 시스템 스크린샷

2.4 문화유산 데이터 태깅 및 NER 결과

개체명 인식(NER)은 문장으로부터 개체로 인식 가능한 정보를 모델이 해석하여 특정 개체가 적합한 태그를 분류(Classification)하는 연구이다. 대표적인 한국어 모델로는 논문의 맨 앞에서 명시한 KoBERT외에도 HanBERT, KoELECTRA와 같은 모델들이 있다.

모델 성능의 경우 Hyper-parameters 중 각 Learning rate에 대한 random seed를 5개씩 적용한 결과로 한 모델당 20번씩 학습한 평균 수치를 기재하였다. 추가로 한자 관련 정보를 제외한 전처리를 다양한 모델에 적용하여서도 성능 실험을 진행하였다. 한자를 포함한 데이터 성능과 큰 차이가 없었으나, 이는 전체 데이터로 재실험을 진행할 예정이다. 앞서 내용을 토대로 성능 측정을 한 결과는 표 1과 같다.

아래 결과에서 알 수 있듯이, 모두의 말뭉치를 기반으로 한 NRE 모델에 비해 낮은 성능이 나왔지만, 전체 데이터셋이 아닌 샘플 데이터셋으로 학습한 결과이고, 아직 태깅중인 데이터들도 일부 섞여 있기 때문에 추후 전체 태깅 데이터로 학습하면 성능이 향상될 것으로 기대된다.

[표 1] NER 성능 결과

Results	Precision	Recall	F1-score
KoBERT	67.24%	59.26%	63%
KoELECTRA	66.10%	60.63%	63.25%

3. 결론

본 논문에서는 텍스트 기반 문화유산의 어노테이션 시스템 구축에 관한 연구를 진행하였다. 일반적으로 사람들이 사용하는 단어(평문+신문사설)들로 이루어진 말뭉치와 '문화유산' 도메인에서 사용하는 단어들과는 상당한 괴리감이 있어, 기존의 말뭉치를 기반으로 하는 NER로는 '문화유산' 도메인에서 필요한 개체명이 인식되지 않는 문제점을 발견했고, 이를 극복하기 위해서는 새로운 말뭉치가 필요하다는 점을 확인하였다.

참고문헌

[1] Chi Hoon Lee, Yeon Ji Lee, Dong Hee Lee, "A Study of Fine Tuning Pre-Trained Korean BERT for Question Answering Performance Development", Journal of Information Technology Service, Korea, pp 83-91, 2020

[2] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz,"A Bayesian approach to filtering junk e-mail," AAAI'98 Workshop on Learning for Text Categorization.,1998.

[3] P. J. Kim, "A Study on automatic assignment of descriptors using machine learning," Journal of the Korean Society for Information Management, Vol.23 No.1, pp.279-299, 2006

[4] CIDC CRM(Conceptual Reference Model), http://www.cidoc-crm.org/

[5] Ju Hee Suh(2016), "A study on constructing semantic structure of historic site information by implementing KCHDM : focusing on achasan Koguryo historic site", KAIST

[6] National Institute of the Korean Language, Guidelines for constructing entity name analysis corpus 2019, "everyone's corpus", https://corpus.korean.go.kr/