국내 학술지 출현 학과정보 데이터셋 구축 및 자동분류

김병규*, 류범종*, 심형섭O
*한국과학기술정보연구원 데이터기반문제해결연구단,
O한국과학기술정보연구원 개방형데이터융합연구단
e-mail: {bk.kim, ybj}@kisti.re.kr*, hsshim@kisti.re.krO

Dataset construction and Automatic classification of Department information appearing in Domestic journals

Byungkyu Kim*, Beom-Jong You*, Hyoung-Seop ShimO
*Dept. of data-centric problem solving research, KISTI,
ODept. of Open Data Convergence Research, KISTI

● 요 약 ●

과학기술 문헌을 활용한 계량정보분석에서 학과정보의 활용은 매유 유용하다. 본 논문에서는 한국과학기술인용색인데이터베이스에 등재된 국내 학술지 논문에 출현하는 대학기관 소속 저자의 학과정보를 추출하고 데이터 정제 및 학과유형 분류 처리를 통해 학과정보 데이터셋을 구축하였다. 학과정보 데이터셋을 학습데이터와 검증데이터로 이용하여 딥러닝 기반의 자동분류 모델을 구현하였으며, 모델 성능 평가 결과는 한글 학과정보 기준 98.6%와 영문 학과정보 기준 97.6%의 정확률로 측정되었다. 향후 과학기술 분야별 지적관계 분석 및 논문 주제분류 등에 학과정보 자동분류 처리기의 활용이 기대된다.

키워드: KSCD(Korea Science Citation Database), 학과유형(Department Type), 딥러닝(Deep learning)

I. Introduction

연구자의 학과정보는 연구자 식별 및 논문의 주제분류를 지원하고 과학기술 분야별 지적관계 분석을 위해 다양하게 활용될 수 있는 중요한 정보자원이다. 하지만 계량분석연구에서 해당 정보를 사용하기 위해서는 데이터의 수집부터 학과정보의 추출 및 데이터 정제 등의 전처리 작업, 학과유형에 대한 정확한 분류 처리까지의 단계별 데이터 처리 과정들을 통해 실제 데이터를 제작해야 하는 어려움이 존재한다. 이에 본 논문에서는 재사용이 가능하며 자동분류 처리의 기반이 되는 학과정보 데이터셋을 구축하였다. 이를 위해 한국과학기술정보연구원(이하 KISTI)의 한국과학기술인용색인데이터베이스(이하 KSCD)와 대학교육협의회(이하 대교협)의 대학기관 학과유형 분류체계를 활용하였다[1][2]. 또한 데이터셋을 기반으로 학과정보를 자동분류하기 위하여 딥러닝 기반의 모델을 개발하였다. 학과정보 데이터셋은 한글 학과명과 영문 학과명 기준으로 각각 제작하였으며 모델 개발과 실제 학습과 평가도 한글과 영문으로 이원화하여 구현하였다.

II. Materials and Methods

국내 학술지에 출현하는 학과정보 데이터셋 구축을 위하여 KISTI가 2002년부터 개발해온 KSCD를 활용하였다. KSCD에 등재된 국내 학술지 500종의 2015년부터 2017년까지의 논문 83,633건에서 대학기관 저자 기준으로 159,083개의 저자소속정보를 사용하였으며, 기관명을 제거 등 데이터 정제를 위한 전처리 작업을 실시하고 학과유형 분류를 수작업으로 진행하였다. 학과유형 분류 시 대교협에서 제공하는 국내 교육편제단위 표준분류체계를 기준으로 삼았다.

딥러닝 기반의 학과정보 자동분류 모델은 LSTM(Long Short-Term Memory models)을 이용하여 구현하였다. LSTM은 기존 RNN((Recurrent Neural Networks) 보다 성능이 개선된 순환신경망 이다. 학과명 한글 데이터의 형태소 분석을 위해 Mecab을 사용하였으며 불필요한 단어를 불용어 목록으로 정리하여 해당 단어는 제외되도록 하였다.

모델의 구성은 기울기 소실 문제를 해결할 수 있는 ReLU함수와 출력계층에서 활성화 함수는 다중 클래스 분류를 위한 SoftMax 함수를 사용하였으며, 손실함수는 다중분류를 위한 categorical _crossentropy를 이용하였다. 학습과 테스트 데이터는 7:3 비율로 나누어 학습을 수행하였으며 학습 횟수는 30번으로 하고 과적합 방지를 위해 조기 종료 기법을 이용하였다. 연구 수행 체계는 아래

Fig. 1과 같다.

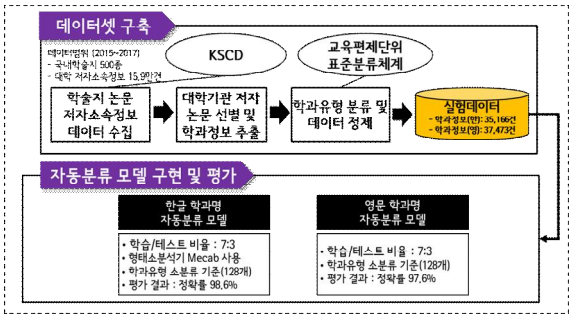

Fig. 1. Research Process Overview

III. Result

1) 학과정보 데이터셋 구축

2장의 Fig.1의 방법을 통해 데이터셋을 제작하였으며, 학과정보 데이터(국문 38,618건, 영문 44,970건)에서 출현 수가 많은 분류 선택 및 출현횟수 3회 이상의 조건으로 데이터를 필터링하여 최종 실험데이터(한글 35,166건, 영문 37,473건)를 구축하였다. 국문 학과정보 데이터를 학과유형 중분류 기준으로 10순위까지 살펴보면, 건설(19.8%), 의료(11.0%), 전기·전자·컴퓨터(10.6%), 보건(8.4%), 사회과학(7.1%), 간호(6.6%), 경영·경제(5.0%), 기계(4.4%), 화공·고분자·에너지(4.2%), 수학·물리·천문·지구(4.0%)으로 파악되었다.

2) 학과정보 자동분류 모델 구현

학과정보 자동분류 모델은 데이터셋의 언어별로 한글버전과 영문버전으로 이원화하여 구현하였다. 모델 성능 평가 결과는 한글과 영문 각각 98.6%, 97.6%의 정확률로 측정되었다. 실험데이터에 출현한 학과유형은 소분류 기준으로 128개로써, 3장 1절에서 살펴본 바와 같이 유형별로 데이터셋 구성에서 편차가 크다. 학습데이터 수가 작은 학과유형의 경우 모델의 예측 정확도가 상대적으로 낮을 가능성이 높기 때문에 데이터셋이 보완과 이를 고려한 모델의 개선이 필요하다.

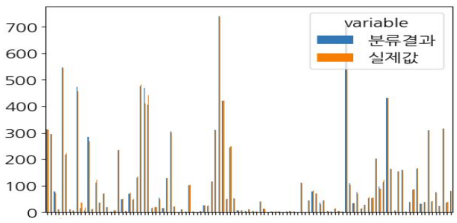

Fig. 2. Classification Result by Dept. Type

IV. Conclusions

본 논문에서는 학과정보 분석 활용을 위해 데이터의 수집부터 학과정보의 추출 및 데이터 정제 등의 전처리 작업, 학과유형에 대한 정확한 분류 처리를 통해 데이터셋을 구축하였다. 또한 데이터셋을 학습 및 테스트 데이터로 사용하여 딥러닝 기반의 자동분류 모델을 구현하였다. 모델 성능 평가 결과는 한글 학과정보 기준 98.6%와 영문 학과정보 기준 97.6%의 정확률로 측정되었다. 향후 본 연구결과를 다양한 계량정보 분석에 활용할 계획이다.

ACKNOWLEDGEMENT

이 논문은 한국과학기술정보연구원 주요사업의 지원을 받아 수행된 연구임

REFERENCES

[1] H. Choi, B. Kim, Y. Jung, and S. Choi, "Korean scholarly information analysis based on Korea Science Citation Database (KSCD)," Collnet Journal of Scientometrics and Information Management, vol. 7, No. 1, pp. 1-33, Jun. 2013.

[2] Korean Council for University Education, The standard classification of university education units, https://www.data.go.kr/data/15014632/fileData.do