딥러닝 기반 한국 표준 산업분류 자동분류 모델 비교

우찬균*, 임희석**
*고려대학교 컴퓨터정보통신대학원 빅데이터융합학과
**고려대학교 컴퓨터학과 교수
ckwoo@korea.ac.kr, limhseok@korea.ac.kr

Comparison of Korean Standard Industrial Classification Automatic Classification Model on Deep Learning

Chan Kyun Woo*, Heui Seok Lim**
*Dept. of Big Data Convergence, Korea University Graduate School of Computer and Information Technology
**Dept. of Computer Science and Engineering, Korea University

요 약

통계청에서는 지역별고용조사, 인구총조사 등 다양한 조사를 실시하고 있다. 이러한 조사에서는 응답자의 사업체명, 사업체가 주로 하는 일, 응답자가 한 일, 부서 및 직책 정보 등을 조사해서 조사되어진 자료를 토대로 한국 표준 산업분류 형태로 코드를 부여해 주고 있다. 각 조사에서는 자연어 형태로 입력을 받아서 자료처리 기간에 코딩작업을 하는 조사가 있고, 조사원이 입력을 하면서 자동코딩시스템을 이용해서 산업분류 코드를 입력하는 방식도 있다. 본 연구에서는 전자의 방법을 자동화하는 것에 초점을 두었다. 딥러닝 알고리즘을 이용해서 기존에 코드부여가 완료된 자료를 가지고 실험을 해본 결과 조사된 모든 항목을 사용했을 때에는 CNN이 81.36%로 가장 좋은 성능을 보였고, 항목을 2가지로 (사업체가 주로 하는 일/응답자가 한 일) 줄였을 경우 전체적으로 더 좋은 성능을 보였다.그 중에 CNN-LSTM이 85.91%로 가장 좋은 성능을 보였다.

keywords : KSIC, Autocode, Deep Learning, CNN, LSTM, CNN-LSTM

1. 서론

통계청에서는 지역별고용조사, 인구총조사 등 다양한 조사를 실시하고 있다. 이러한 조사에서 응답자의 사업체명, 사업체가 주로 하는 일, 응답자가 한 일, 근무부서와 직책 정보를 받아서 산업이나 직업을 분류한다. 한국 표준 산업분류는 총 5개의 단계로 분류가 되어있는데 대분류 21개, 중분류 77개, 소분류 232개, 세분류 495개, 세세분류 1,196개로 매우 다양하게 분류가 되어 있다. 분류가 매우 다양하기 때문에 조사원이나 응답자가 관련된 분류에 정확한 지식을 가지고 있지 않다면 잘못된 분류정보를 선택할 수 있고, 너무 복잡한 내용으로 조사표를 구성할 경우 무응답 비율이 많아질 수 있다.

각 조사에서는 이러한 정보를 자연어로 입력을 받아서 자료처리 기간에 코드를 부여해 주는 방법으로 코드를 부여해 주는 조사가 있는 반면, 자동코딩시스템의 도움을 받아서 응답자의 정보를 자동코딩시스템 정보를 토대로 코드를 부여해 주는 조사도 있다. 두 방법 모두 1차적으로 사례사전 및 색인DB를 이용해서 자동으로 코드를 분류한다. 이렇게 분류를 했을 때 모든 조사사료가 자동으로 분류되지는 않고 일부 데이터만 분류가 되고 또는 순위를 매겨서 해당 조사자료는 어떠한 코드에 가장 적합한지 추천을 해 준다. 말 그대로 현재의 자동코딩시스템은 코드부여 작업을 도와주는 시스템이다.

세계의 통계청은 이러한 조사자료의 자동분류에 대한 고민을 함께하고 있다. 특히 최근에는 UNECE 통계현대화 그룹에서 머신러닝을 공식 통계생산[1]에 이용하기 위해서 머신러닝 그룹을 운영하고 있다. 각 나라의 현재 머신러닝을 이용하는 현황을 공유하기도 하고 여러 머신러닝 테스트나 논문을 공유하면서 공식통계에 머신러닝을 이용하기 위해서 노력 하고 있다. 특히 캐나다나 미국의 같은 경우 딥러닝을 이용한 실험에서 아주 좋은 성능을 보이고 있고 실제 조사에서도 머신러닝이나 딥러닝을 이용하고 있다. 본 연구에서는 한국 통계청 조사자료를 딥러닝 알고리즘을 이용해서 한국 표준 산업분류를 자동으로 분류하는 실험을 해보았다.

2. 자동코딩 방법

통계청에서는 산업·직업분류를 정확하고 빠르게 분류하는 것을 도와주기 위해서 자동코딩시스템을 운영하고 있다. 자동코딩시스템은 크게 2가지로 나뉜다. 첫번째로 사례사전관리시스템이다. 이 시스템은 지역별고용조사, 전국사업체조사 등 실제 코드 부여가 완료된 조사 자료를 가지고 사람이 규칙을 만들어서 관리하는 시스템이다. 매년 상·하반기를 나누어서 자료를 업데이트 하고 있다. 정확도가 매우 높고 코딩 결과는 1:1로 결과가 나온다. 두번째로는 색인DB가 있다. 색인DB는 사례사전관리시스템과 동일하게 동일한 조사에서 학습데이터를 가져 오지만 코드를 부여하는 방식은 다르다. 색인DB는 조사자료에서 색인어를 추출해서 해당 DB에 저장해 놓고 코드를 부여해야 하는 자료가 들어오면 DB의 정보를 통해서 코드를 부여해 주는 방식이다. 코드는 사용자가 정하는 범위에 따라 1순위부터 10순위 이하까지 설정이 가능하다. 1순위는 입력한 자료와 가장 가까운 코드를 보여주는 방식이다.

일반적으로 사례사전관리시스템에서는 전체데이터가 모두 코드부여가 되지 않기 때문에 사례사전관리시스템과 색인DB를 같이 사용하게 된다. 사례사전에서 나온 코드와 색인DB에서 나온코드를 비교해서 내검원이나 조사원이 최종 코드를 부여해 주는 방식으로 산업·직업코드를 부여해 주고 있다.

3. 모델 및 평가

3-1. 학습데이터

학습데이터는 통계청 조사에서 산업 분류코드가 대분류로 부여된 자료를 사용했다. 총 14,831건의 자료를 사용했다. 학습데이터로 10,000건, 평가로 1,864건, 테스트로 2,967건을 사용했다.

학습데이터는 KoNLPy 라이브러리 okt 객체를 사용해서 어간 단위의 형태소 토크나이징을 했다. 각 데이터의 길이가 다르기 때문에 길이를 8단어로 통일했다. 8단어보다 긴 데이터는 뒷부분을 자르고 짧을 경우는 0 값으로 패딩처리 했다. 그리고 의미 없는 단어 O, o, ㈜ 등의 단어를 불용어 처리 했다.

3-2. 모델구성 및 평가 방법

모델은 학습데이터가 각 딥러닝 알고리즘에 어떻게 동작하는지 실험하기 위해서 CNN (Convolution Neural Network), LSTM (Long Short-Term Memory units), CNN-LSTM 3가지의 모델을 가지고 비교를 했다.

합성곱 신경망(CNN)은 1개의 합성곱 신경망 + 맥스풀링 층을 사용하였다. 배치크기는 256, 필터크기는 3, optimization은 adam을 사용했고 overfitting을 막기 위해서 dropout 은 0.2 로 설정했다. 평가는 accuracy를 비교했다.

순환 신경망 모델(LSTM)은 128 메모리 셀을 가진 LSTM 레이어 1개, Dense 레이어 1개로 구성했다. 합성곱 신경망(CNN)과 동일하게 optimization은 adam을 사용했고 평가는 accuracy를 비교했다.

마지막으로 순환 컨볼루션 신경망 모델 (CNN-LSTM)은 컨볼루션 레이어에서 나온 feature vector들을 MaxPooling을 통해서 1/4로 줄인 다음에 순환 신경망 모델(LSTM) 입력으로 사용하도록 구성했다. 위와 동일하게 optimization은 adam을 사용했고 평가는 accuracy를 비교했다.

3-3. Feature 구성

첫 번째 실험 **(Case 1)**은 사업체명, 사업체가 주로 하는 일, 응답자가 한일, 근무부서와 직책 정보 모든 데이터를 feature 해서 각 모델을 돌려 보았다.

두 번째로는 **(Case 2)** 각 항목을 구분해 주고 학습시키는 것이 성능이 좋은 선행연구[2]를 활용해서 각 항목 앞에 구분자 A,B,C,D를 넣어서 학습을 시켰다.

마지막으로는 **(Case 3)** 지금까지 한국 표준 산업분류 코딩작업을 해 본 경험으로 코드 분류에 주로 영향을 주는 요소는 대부분 사업체가 주로 하는 일, 응답자가 한 일 두 항목이 영향을 주고 있다는 판단에 의해서 딥러닝 모델도 두 항목만을 가지고 학습을 시켜 보았다.

3-4. 하이퍼파라미터

각 모델을 구성하고 가장 좋은 성능의 모델을 비교해 보기위해서 파라미터값을 조절 하는데 epochs 20, batch size 128 일 때 가장 좋은 성능을 보였다.

4. 실험결과

3가지 Case를 비교해 본 결과 **Case 3이 전체적으로 가장 좋은 성능을 보였다.** 그 중에 **CNN-LSTM 모델이 85.91%**로 가장 좋은 성능을 보였다. 3가지 모델을 비교 했을 때에는 모두 비슷한 성능을 보였지만 대체적으로 CNN이 좀 더 좋은 성능을 보였다.

Case	CNN	LSTM	CNN-LSTM
Case 1	81.36%	79.23%	79.51%
Case 2	81.05%	78.66%	80.24%
Case 3	85.40%	83.48%	85.91%

표 1 실험 결과

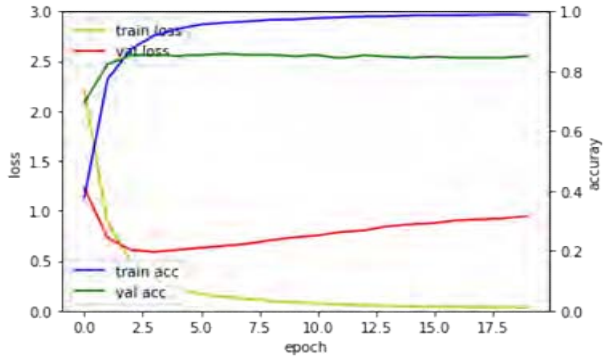

그림 1 CNN

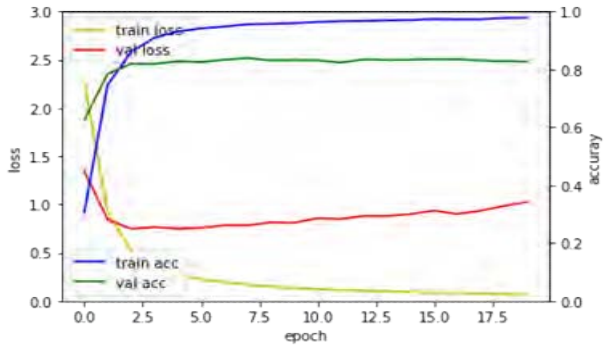

그림 2 LSTM

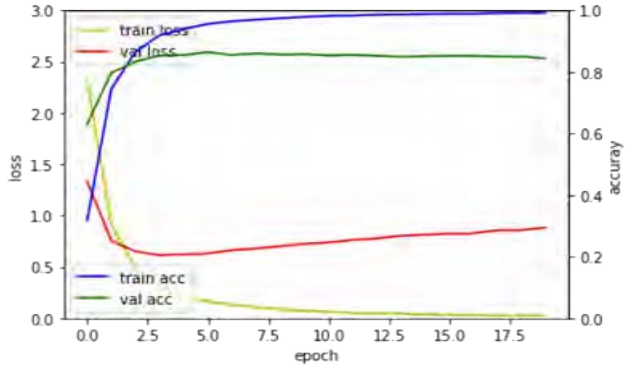

그림 3 CNN-LSTM

5. 결론 및 향후 연구

실험 결과 조사 되어진 모든 자료를 사용하는 것보다는 분류를 잘할 수 있는 항목을 선택해서 학습하는 것이 성능이 더 좋은 것을 확인했다. 그리고 Feature가 많을 때에는 CNN이 좋은 성능을 보였고 Feature가 적을 때에는 CNN-LSTM이 좋은 성능을 보였다.

향후 연구에서는 이번 연구를 토대로 조사된 항목 중 어떠한 항목을 선택하는 것이 산업분류 자동분류 성능을 높일 수 있는지 비교해 보고, 한국어 단어의 임베딩 방법을 달리해서 성능을 비교해 볼 예정이다.

참고문헌

[1] UNECE Machine Learning Team "The use of machine learning in official statistics' November 2018

[2] 김병수, CNN을 활용한 화계 계정코드 분류, 2019 한국정보기술학회·한국디지털콘텐츠학회 하계 공동학술대회 논문집,2019,583p