Journal of the Korea Convergence Society
Vol. 12. No. 4, pp. 23-30, 2021

ISSN 2233-4890 / eISSN 2713-6353
https://doi.org/10.15207/JKCS.2021.12.4.023

딥러닝 기법을 활용한 산업/직업 자동코딩 시스템

임정우[1], 문현석[1], 이찬희[1], 우찬균[2], 임희석[3*]
[1]고려대학교 컴퓨터학과 석박사통합과정, [2]통계청 조사시스템관리과 전산주무관, [3]고려대학교 컴퓨터학과 교수

An Automated Industry and Occupation Coding System using Deep Learning

Jungwoo Lim[1], Hyeonseok Moon[1], Chanhee Lee[1], Chankyun Woo[2], Heuiseok Lim[3*]
[1]Master & Ph.D Integrated Student, Department of Computer Science and Engineering, Korea University
[2]Computing Assistant Deputy Director, Survey System Management Division
[3]Professor, Department of Computer Science and Engineering, Korea University

요 약 본 산업/직업 자동코딩 시스템은 조사 대상자들이 응답한 방대한 양의 산업/직업을 설명하는 자연어 데이터에 통계 분류 코드를 자동으로 부여하는 시스템이다. 본 연구는 기존의 정보검색 기반의 산업/직업 자동코딩시스템과 다르게 딥러닝을 이용하여 색인 DB가 필요하지 않고 분류 수준에 상관없이 코드를 부여할 수 있는 시스템을 제안한다. 또한, 자연어 처리에 특화된 딥러닝 기법인 KoBERT를 적용한 제안 모델은 인구주택총조사 산업/직업 코드 분류, 그리고 사업체기초조사 산업 코드 분류에서 각각 95.65%, 91.45%, 97.66%의 Top 10 정확도를 보인다. 제안한 모델 실험 후 향후 개선 가능성을 데이터/모델링 관점으로 분석한다.

주제어 : 통계 분류, 융합, 산업/직업 자동코딩, 딥러닝, Bi-LSTM, KoBERT

Abstract An Automated Industry and Occupation Coding System assigns statistical classification code to the enormous amount of natural language data collected from people who write about their industry and occupation. Unlike previous studies that applied information retrieval, we propose a system that does not need an index database and gives proper code regardless of the level of classification. Also, we show our model, which utilized KoBERT that achieves high performance in natural language downstream tasks with deep learning, outperforms baseline. Our method achieves 95.65%, 91.51%, and 97.66% in Occupation/Industry Code Classification of Population and Housing Census, and Industry Code Classification of Census on Basic Characteristics of Establishments. Moreover, we also demonstrate future improvements through error analysis in the respect of data and modeling.

Key Words : Statistic Code Convergence, Classification, Automated Industry/Occupation Coding, Deep learning, Bi-LSTM, KoBERT

*This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2020-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques).
*Corresponding Author : Heuiseok Lim(limhseok@korea.ac.kr)
Received February 5, 2021 Revised March 3, 2021
Accepted April 20, 2021 Published April 28, 2021

1. 서론

1.1 서론

통계청에서는 통계 데이터의 정확성과 비교가능성을 위하여 표준산업분류, 표준직업분류 등 다양한 통계표준분류를 관리하고 있다[1]. 통계청은 이러한 통계표준분류를 이용하여 매 5(0년, 5년)년 전체 가구를 조사하는 인구주택총조사와 사업체 기초 통계조사[1]를 시행한다. 인구주택총조사는 국가 기본 통계조사로서, 대한민국 영토 내의 모든 인구주택을 조사하는 것이고[2], 사업체 기초통계조사는 전국의 지역별 사업체의 규모 및 분포를 파악하는 조사이다[3]. 인구주택총조사의 표본조사에서는 조사대상의 산업 및 직업에 대한 답변을 기반으로 산업코드와 직업코드를 부여한다. 사업체기초통계조사에서는 사업체의 기본사항에 대한 답변을 기준으로 산업분류코드를 부여하게 된다. 통계청은 이렇게 자연어 형태로 조사되어진 조사자료를 산업/직업분류로 코딩을 하기 위해서 자료처리 기간에 내검원을 채용해서 고용하고 코딩하는 방식으로 분류 업무를 진행하고 있는데 이는 다양한 문제점을 야기한다[4].

여러 문제점 중 하나는 시간과 비용이다. 조사원들은 채용된 후 별도의 교육을 받은 후 조사대상의 답변에 알맞은 통계표준분류 코드를 코딩하는 작업을 수행하게 된다. 이때 조사 때마다 매번 조사원들을 채용하고 교육을 진행해야 하므로 많은 시간과 비용이 소요된다. 또 다른 문제점은 수동코딩 과정에서 여러 조사원 및 내검원들이 많은 코딩작업을 나누어 수행하기 때문에 코딩의 일관성이 떨어질 수 있다는 것이다. 더 나아가, 조사원 및 내검원들은 매번 단기간 조사를 위해 모집되었기 때문에, 통계 분류에 대한 전문성도 통계표준분류 전문가들에 비해 비교적 떨어지는 것도 하나의 문제점이다. 이는 이후 내검원의 지식과 성향에 따라 같은 데이터를 다르게 판단하는 결과를 초래하기도 한다[5].

이러한 문제점들을 해결하기 위하여 기존에는 지식 검색기법을 이용하여 자동코딩을 하는 방식들이 2000년대부터 제기되어왔는데[1,4,6], 이러한 방법들은 여전히 대량의 데이터에 대해서도 색인 DB를 구축해야 한다는 불편함과 수준별로만 코드를 분류할 수 있다는 점에서 그 한계가 있다. 이를 극복하기 위해 본 연구에서는 색인DB를 따로 구축하지 않아도 되고, 수준에 상관없이 코드를 분류할 수 있는 딥러닝 기반 모델을 제안한다. 또한, 본 연구에서 제안하는 딥러닝 기반 산업/직업 자동코딩 모델은 대량의 데이터에 대해서도 무리 없이 자동코딩이 가능하고, 자연어 기반 하위 태스크들에 높은 성능을 보이는 트랜스포머[7] 구조를 이용하기 때문에 이전보다 더욱 효율적이고 효과적인 자동코딩 시스템이 될 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 산업/직업 자동코딩 및 분류에 대한 국외, 국내 연구들을 살펴볼 것이다. 3장에서는 본 연구에서 제안하는 딥러닝 기반 산업/직업 분류 모델의 구조를 소개할 것이다. 4장에서는 딥러닝 기반 산업/직업 분류 모델을 이용하여 실제 인구주택총조사, 사업체 기초조사에서의 성능을 분석하고 오류 분석결과도 설명한다. 5장에서는 본 연구의 결론 및 향후 연구를 소개한다.

2. 관련 연구

기존의 자동 산업/직업 자동코딩 및 분류를 위해서 여러 연구가 진행되어왔는데, 본 연구에서는 국외/국내로 나누어 설명하고자 한다.

2.1 산업/직업 자동코딩을 위한 국외 연구

국외에서는 미국 지역사회조사(American Community Survey)에서 수집한 개인의 데이터를 수집하여, 개인이 속한 산업과 직업을 자동으로 코딩하려는 연구가 진행되었다[8]. 이때 수집한 개인의 데이터는 자연어로 수집되었으며, 답변에 적절한 산업/직업 코드를 부여한다. 이 연구에서는 먼저 데이터 사전을 통해서 특정 산업/직업 코드와 단어들 간의 어떤 관계가 있는지 정의하고, 입력받은 자연어 데이터에 맞는 산업/직업 코드의 후보군을 구성한다. 그다음, 로지스틱 회귀를 이용하여 후보군 중 어떤 후보 코드가 정답일지 예측한다. 또 다른 자동코딩 연구로는 Wood et al.[9]의 연구가 있다. 해당 연구에서는 유료 EverString DB 혹은 웹에서 크롤링 및 스크래핑한 데이터셋을 이용하여 수집한 산업 관련 자연어 문서가 어떤 북미산업분류코드(North American Industry Classification System)를 지칭하는지 예측하는 모델을 만든 연구이다. 이를 위해 딥러닝 모델을 사용하였지만, 자연어 질의에 한계가 있는 완전연결 네트워크 (Fully Connected Network) [10] 모델구조를 사용하였다는 한계가 존재한다.

1) 현재는 '전국사업체조사' 로 명칭이 바뀌었다. (2008.10 기준)

2.2 산업/직업 자동코딩을 위한 국내 연구

국내의 자동산업분류를 위한 대표 연구사례로는 우찬균의 '딥러닝 기반 한국 표준 산업분류 자동분류 모델 비교'와 이재성의 '한국표준산업분류를 기준으로 한 문서의 자동분류 모델에 관한 연구' 가 있다 [5,11]. '딥러닝 기반 한국 표준 산업분류 자동분류 모델 비교'에서는 한국 통계청 조사자료와 CNN+LSTM 알고리즘을 이용하여 한국 표준 산업분류 자동분류를 수행하였지만, 대분류에 대해서만 분류를 진행했다는 한계점이 존재한다. '한국표준산업분류를 기준으로 한 문서의 자동 분류 모델에 관한 연구'에서는 기업의 사업 성격을 나타내는 한국표준산업분류(KSIC)를 기준으로 정보를 분류하는 방법을 제안하였다. 하지만 이 연구에서는 한국 통계청 조사자료가 아닌 한국표준산업분류에 관련 문서들로 산업분류 모델을 구성하였다는 점에서 차이가 있다. 본 연구에서는 이러한 기존의 모델 및 데이터 차원에서 존재하는 한계점을 자연어 기반 모델과 통계 분류 데이터 극복하고자 하였다.

3. 딥러닝 기반 산업/직업 분류 모델

3.1 Bi-LSTM

자연어처리에 이용되는 딥러닝 기법을 적용한 산업 및 직업 분류코드 분류모델은 아직 제안된 바가 없다. 이에 본 연구에서는 BERT기반 분류모델의 효용성과 유능함을 증명하기 위하여 Bidirectional Long Shot Term Memory(Bi-LSTM) [12] 기반의 모델을 설계하여 그 성능의 비교 대상으로 삼는다. Bi-LSTM은 입력으로 들어온 문장의 순방향과 역방향 문맥을 동시에 파악할 수 있다는 점에서 문장 이해에 강점을 보인다. 특히, 한 방향으로만 문장을 해석하는 LSTM의 한계점을 극복하고 양방향적인 문맥을 파악할 수 있게 되어 딥러닝 기반의 여러 자연어 처리 작업에 뛰어난 성능을 보인다 [13-15].

LSTM은 토큰 단위로 분절된 문장 $[x_0, x_1, \cdots, x_{n-1}]$을 입력으로 받아 순차적으로 정보를 처리하여, $[y_0, y_1, \cdots, y_{n-1}]$라는 결과물을 출력한다. Bi-LSTM은 기존 문장을 입력으로 받는 순방향 LSTM과 더불어, 순서를 뒤집은 문장 $[x_{n-1}, x_{n-2}, \cdots, x_0]$을 입력으로 받아 $[y_{n-1}, y_{n-2}, \cdots, y_0]$를 출력하는 역방향 LSTM을 추가적으로 활용한다. 순차적으로 입력을 해석하는 모델의 구조는, 문장의 길이가 길어질수록 앞에서 처리한 문장의 정보가 옅어지는 문제(Gradient vanishing problem)이 존재하기 때문에, LSTM에서는 이 문제를 해결하기 위하여 망각게이트(forget gate)를 도입한다.

LSTM을 통해 입력 문장을 처리하는 과정은 다음과 같다.

$$
\begin{aligned}
i^{(t)} &= \sigma_g(W^{(i)}x^{(t)} + U^{(i)}h^{(t-1)} + b^{(i)}) \\
f^{(t)} &= \sigma(W^{(f)}x^{(t)} + U^{(f)}h^{(t-1)} + b^{(f)}) \\
o^{(t)} &= \sigma(W^{(o)}x^{(t)} + U^{(o)}h^{(t-1)} + b^{(o)}) \\
\tilde{c}^{(t)} &= \sigma_c(W^{(c)}x^{(t)} + U^{(c)}h^{(t-1)} + b^{(c)}) \\
c^{(t)} &= f^{(t)} \odot \tilde{c}^{(t-1)} + i^{(t)} \odot \tilde{c}^{(t)} \\
h^{(t)} &= o^{(t)} \odot \sigma_h(c^{(t)})
\end{aligned}
$$

여기서 i, f, o, c 는 각각 입력게이트(input gate), 망각게이트(forget gate), 출력게이트(output gate), 상태(cell state) 를 의미하고, W, U 는 각각 LSTM을 통해 학습되는 가중치 행렬(weight matrix), 그리고 b는 편향값(bias) 를 의미한다. ⊙ 는 행렬의 각 원소별 곱을 의미하며, σ는 활성화함수(activation function)를 의미하고, 본 연구에서는 Tanh함수를 활용하였다.

산업/직업 자동코딩은 자연어 문장을 입력으로 받아 해당 문장의 분류군을 찾는 목적이 있기에, 분류군을 찾아내기 위하여 마지막 분류 층을 추가한다. 이 분류는 즉 순방향 LSTM에서 $[x_0, x_1, \cdots, x_{n-1}]$ 을 입력으로 받아 출력하는 y_{n-1}과, 역방향 LSTM에서 $[x_{n-1}, x_{n-2}, \cdots, x_0]$을 입력으로 받아 출력하는 y'_0를 이어 붙인 결과를 가중치 벡터를 통해 학습시키는 선형 분류기(Linear Classifier) 에 통과시키는 과정이라고 볼 수 있다. 이 과정을 통해 얻은 벡터값에 Softmax함수를 적용시킴으로써 입력 문장이 각 분류군으로 분류될 확률값을 획득하고 이를 통해 분류군을 예측한다.

3.2 KoBERT

BERT[16]가 등장하기 전의 자연어 처리연구는 문장을 순방향, 혹은 역방향의 방향대로만 처리하였기에 깊은 문맥적 의미를 파악하는 데에 한계가 존재하였다. BERT는 가중 곱 어텐션(scaled dot product attention)을 기반으로, 다중헤드 어텐션(multi head attention)과 자가 참조 어텐션(self attention) [7] 을 활용하여 정해진 방향대로 문장을 해석하는 문제를 해결하였다. BERT는 등장과 동시에 감정분석이나 자연어추론 등, 문장의 이해를 요구하는 대부분의 자연어처리 분야에서 가장 좋은

성능을 보였고, 현재까지도 자연어처리의 많은 분야에서 활용되고 있다[17-19]. 한국에서는 SKT-brain에서 한국어위키, 한국어 뉴스 2,500만 문장을 통해 학습한 KoBERT[2]가 등장하여 자연어처리 많은 분야에서 뛰어난 성능을 보여주고 있다.

BERT에서는 입력 문장의 의미를 양 방향적으로 파악하기 위해 같은 문장을 세 개의 입력 Q, K, V으로 복사한다. 그 후, Q, K, V에 대한 가중 곱 어텐션 값을 구함으로써 입력 문장의 의미를 파악하는 과정을 거친다. 이때, 더 깊은 의미파악을 위하여 다중헤드 어텐션 구조를 도입한다. 이를 수식으로 나타내면 다음과 같다.

$$\mathrm{Attentio}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$h_i = \mathrm{Attentio}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\mathrm{MultiHea}(Q,K,V) = \mathrm{Concat}(h_1, \cdots, h_h)W^O$$

여기서 Attention은 가중 헤드 어텐션을, Multihead는 다중 헤드 어텐션을, head는 다중 헤드 어텐션에 이용되는 구조를 의미하고, 각 W는 훈련 과정을 통해 학습되는 가중치 행렬을 의미한다. 즉 BERT는 다중 헤드 어텐션을 통해 동일한 문장 간의 가중 곱 어텐션을 여러 차례 진행하여 문장의 깊은 의미를 파악하는 구조체이다.

BERT의 모델구조는 다음과 같다. 입력을 구성할 때 문장의 가장 앞부분에는 [CLS] 토큰을, 문장의 가장 마지막 부분에는 [SEP] 토큰을 추가함으로써 문장의 시작과 끝을 설정해주고, 모델을 통해 출력되는 가장 첫 번째 토큰을 통해 문장의 분류군을 결정한다. 본 연구에서는 분류 성능 향상을 위하여 가중치 행렬로 이루어진 선형 분류기(Linear Classifier) 구조를 추가로 활용한다. Bi-LSTM에서와 마찬가지로, 최종적으로 Softmax함수를 적용함으로써 입력 문장이 각 분류군으로 분류될 확률값을 얻을 수 있다.

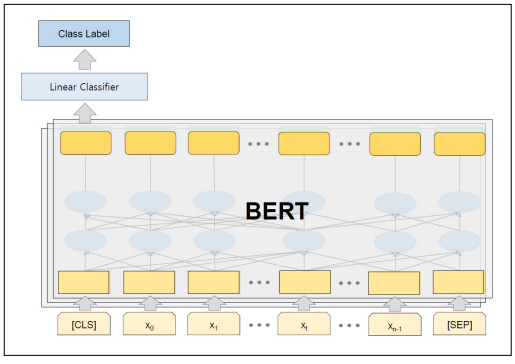

Fig. 1. Proposed Model for Automated Industry and Occupation Coding System

2) https://github.com/SKTBrain/KoBERT

4. 실험

4.1 데이터

본 연구에서는 [4]에서 활용한 데이터와 동일한 산업분류, 직업분류 데이터를 사용한다. 인구주택총조사 데이터는 각 개인이 종사하고 있는 일과 재직중인 사업체, 사업체에서의 부서 및 직위, 해당 사업체의 사업내용으로 이루어져 있고, 알맞은 산업분류코드와 직업분류 코드가 수작업으로 레이블링 되어있다. 사업체 기초 조사 데이터는 사업체의 이름, 사업체의 사업내용과 사업체가 다루는 주요 품목으로 이루어져 있고, 알맞은 산업분류 코드가 수작업으로 레이블링 되어있다. 인구주택총조사의 산업/직업분류 코드가 4자리(세분류)로 되어있고, 사업체 기초 조사는 산업분류 코드가 5자리(세세분류)로 되어있다. 또한, 산업분류는 8차 개정, 그리고 직업분류는 5차 개정 기준 레이블을 사용하였다. 자연어처리 모델의 입력을 구성하기 위하여 인구주택총조사 데이터에서는 사업체명, 사업체의 사업내용, 사업체에서의 부서 및 직위, 하는 일 4가지 항목을 하나의 빈칸(white space)으로 연결하였고, 사업체 기초 조사 데이터에서는 사업체명, 사업체의 사업 내용, 사업체가 다루는 주요 품목 3가지 항목을 빈칸으로 연결하였다. 이를 통해 인구주택총조사 데이터를 통한 직업코드 및 산업코드 자동분류, 사업체 기초 조사 데이터를 통한 산업코드 자동 분류 작업을 위한 데이터셋을 구축하였다.

인구주택총조사 데이터는 총 1,376,657개의 데이터, 그리고 사업체 기초 조사 데이터는 총 3,198,556개의 데이터로 이루어져 있다. 본 연구에서는 원활한 학습을 위하여 중복 항목 제거 등 데이터 정제작업을 진행하였고, 이를 통해 최종적으로 인구주택총조사 데이터 1,144,871개와 사업체데이터 1,280,090개를 구축하였다. 인구주택총조사 데이터를 통해 분류하고자 하는 직업코드 분류군의 개수는 496개, 산업코드 분류군의 개수는 472개이며, 사업체 기초 조사 데이터를 통해 분류하고자 하는 산업코드 분류군의 개수는 1,107개이다.

훈련(train), 검증(dev), 테스트(test) 데이터셋을 8:1:1비율로 임의로 분절하여 데이터셋을 구축하였고, 임의 분절로 인해 발생한, 훈련 데이터셋 안에 포함되지 않은 레이블을 가진 검증 데이터셋, 테스트 데이터셋의 데이터는 삭제하였다 훈련에 이용한 데이터의 개수는 다음 Table 1과 같다.

Table 1. Data Statistics

	Train	Dev	Test
Population and Housing Census - Industry Code Classification	915,964	114,492	114,497
Population and Housing Census - Occupation Code Classification	915,964	114,490	114,499
Census on Basic Characteristics of Establishments	1,024,072	128,008	128,009

4.2 정량적 평가 지표

각 작업에 대한 실험 결과는 완전 일치 정확도(EM)와 상위 10개 포함 정확도 (Top10)를 기준으로 평가한다. 본 연구에서 설계한 Bi-LSTM모델과 BERT모델은 자연어 입력을 처리하여 최종적으로 해당 입력 문장이 각 분류군으로 분류될 확률값을 도출해낸다. 완전 일치 정확도란 기존 데이터셋에서의 label이 모델의 출력 결과에서 가장 확률값이 높은 분류군과 일치하는 비율을 나타내며, 상위 10개 포함 정확도란 기존 데이터셋에서의 label이 출력 결과에서 확률값이 높은 10개의 분류군 안에 포함되는 비율을 나타낸다. 이를 수식으로 나타내면 다음과 같다.

평가대상이 되는 데이터셋을 $D=\{(x_i, y_i)\}_{i=1}^{n}$이라 하고, 총 분류군의 개수를 l라고 했을 때, 데이터 셋 내의 각 y_i는 0과 l사이의 정수로 설정된다. x_i를 입력으로 받아 모델이 출력한 결과물을 $\hat{y_i}$라 할 때, $\hat{y_i} \in R^l$이라 할 수 있다. $\hat{y_i} \in R^l$ 을 입력으로 받아 l개의 확률값들 중 값이 가장 높은 상위 k개의 확률값이 위치한 인덱스(본 실험에서, 이는 분류군을 의미함)를 출력해주는 함수를 $\phi(\hat{y_i}, k)$라고 정의하면 우리는 완전 일치 정확도(EM)와 상위 10개 포함 정확도(Top10)를 다음과 같이 표현할 수 있다.

$$EM(D) = \frac{1}{n}\sum_{i=1}^{n} 1\{y_i \in \phi(\hat{y_i}, 1)\}$$

$$\text{Top10}(D) = \frac{1}{n}\sum_{i=1}^{n} 1\{y_i \in \phi(\hat{y_i}, 10)\}$$

$$1\{x\} = \begin{cases} 1, \text{ if } x \text{ is True} \\ 0, \text{ if } x \text{ is False} \end{cases}$$

본 실험에서는 각 모델의 성능을 검증데이터셋과 테스트데이터셋에 대한 EM과 Top10을 통하여 증명한다.

4.3 실험

본 연구에서는 인구주택총조사 데이터를 통한 직업 분류, 인구주택총조사 데이터를 통한 산업 분류, 그리고 사업체 기초 조사 데이터를 통한 산업 분류 세가지 실험을 수행한다. 본 실험에서 Bi-LSTM은 은닉층의 크기와 임베딩층의 크기를 512로 하여 직접 설계하였고, Bi-LSTM의 입력 데이터를 생성하기 위하여 훈련 데이터를 통해 8,000개의 단어 사전으로 구성된 센텐스피스(sentencepiece) 모델[20]을 생성하였다.

BERT모델은 SKT Brain에서 발표한 KoBERT모델[3]을 사전학습모델로 활용하여 훈련 데이터를 통해 미세조정(fine tuning)을 진행하였다. learning rate은 0.00003으로 설정하였고 배치사이즈는 128로 설정하였다.

4.3.1 인구주택총조사 데이터를 통한 산업 및 직업 분류

Table 2를 통해 확인할 수 있듯, KoBERT기반 산업코드 분류 모델은 테스트 셋에서 EM 66.34 %, Top10 95.65 %를 보이며, Bi-LSTM 분류 모델보다 각각 9 %p, 23.2 %p 높은 성능을 보인다. 딥러닝 기법을 통한 직업 분류는 테스트셋에서 각각 EM 42 %, Top10 51 %이상의 분류 성능을 기록한다. 특히 KoBERT기반의 분류 모델은 EM 51.06 %, Top10 91.45 %로 Bi-LSTM 베이스라인 성능을 압도적으로 뛰어넘는 성능을 보인다. 이를 통해 한국어 문장을 이해하는 데에, BERT의 양방향적 문맥 해석 능력이 매우 큰 역할을 한다는 사실을 확인할 수 있다. 또한 많은 양의 말뭉치로 학습시킨 KoBERT가 데이터의 정보를 잘 파악하고 적당한 코드를 부여하는 데에 Bi-LSTM 보다 더 적절한 모델이라는 점을 시사한다.

3) https://github.com/SKTBrain/KoBERT

Table 2. Population and Housing Census - Industry Code

	Dev		Test	
Model	EM (%)	Top10(%)	EM (%)	Top10(%)
Bi-LSTM	57.20	72.52	56.98	72.45
KoBERT	**66.62**	**95.66**	**66.34**	**95.65**

Table 3. Population and Housing Census - Occupation Code

	Dev		Test	
Model	EM (%)	Top10(%)	EM (%)	Top10(%)
Bi-LSTM	42.35	64.75	42.30	64.91
KoBERT	**51.26**	**91.51**	**51.06**	**91.45**

본 실험 결과는 사람의 개입 없이 딥러닝 모델만으로도 매우 높은 정확도로 서비스를 제공해줄 가능성을 제시하기에 이를 통해 서비스 제공에 걸리는 시간과 비용에 대한 문제를 효과적으로 줄일 수 있을 것으로 기대한다.

4.3.2 사업체기초조사 데이터를 통한 산업 분류

사업체 기초 조사 데이터를 통한 산업 분류 실험 결과는 Table 4와 같다. Bi-LSTM 기반 분류 모델은 EM 71.18 %, Top 10 에서는 76.91 %의 성능을 보이고, KoBERT 기반 분류 모델은 Bi-LSTM보다 훨씬 높은 성능인 EM 84.74 %, Top10 97.66 %를 달성한다. 특히 KoBERT의 Top10에서는 분류 레이블이 1107개의 클래스를 가지고 있음에도 불구하고 다른 분류보다 가장 높은 성능를 보인다. 이는 유의미하게 높은 실험 결과라 결론내릴 수 있고, 딥러닝을 이용한 자동 코드 분류가 매우 현실적이며 실현 가능한 작업임을 시사한다. 또한, Bi-LSTM이 KoBERT보다 성능이 떨어지는 이유는 인구주택총조사와 마찬가지로 문맥을 더 잘 파악하여 적절한 코드를 예측한 것으로 볼 수 있다. 이를 통해 KoBERT가 데이터셋 종류와 무관하게 트랜스포머 구조와 적절한 문맥을 이용하여 올바른 정답 코드를 예측할 수 있다는 점을 시사한다.

Table 4. Census on Basic Characteristics of Establishments - Industry Code

	Dev		Test	
Model	EM (%)	Top10(%)	EM (%)	Top10(%)
Bi-LSTM	71.58	77.23	71.18	76.91
KoBERT	**84.90**	**97.78**	**84.74**	**97.66**

4.4 오류 분석

4.4.1 데이터 오류

데이터 오류는 모델이 제대로 예측하였는데도 불구하고 수동코딩의 결과가 잘못되어서 틀렸다고 예측하는 경우이다. KoBERT모델을 이용하여 사업체 기초 통계조사

Table 5. Data Error - Census on Basic Characteristics of Establishments

	Data	Predicted Code	Predicted Code Description	Answer Code	Answer Code Description
1	○○ 시어터 가전제품도매 오디오 (음향기기)	51430	가전제품 도매업	52511	가전제품 소매업
2	◇◇ 곱창 한식업점 곱창찌개, 삽겹살	55211	한식점업	55221	피자, 햄버거 및 치킨 전문점
3	△△ 약국 약국 약소매, 조제	52311	의약품 및 의료용품 소매업	52214	과실 및 채소 소매업
4	☆☆ 문구, 편시 악세사리, 문구, 선물, 공책, 연필	52622	문구용품 소매업	52611	철물 및 난방용구
5	◎◎ 중고명품 위탁판매 가방,옷	52709	기타 중고품 소매업	52702	중고 가전제품 소매업

Table 6. Model Prediction Error - Census on Basic Characteristics of Establishments

	Data	Predicted Code	Predicted Code Description	Answer Code	Answer Code Description
1	□□ 역학원 역학강의, 역학서적소매 강의도서	52621	서적 및 잡지류 소매업	93992	점술업
2	▽▽ 가로판매점 공구 중고공구	52709	기타 중고품 소매업	52611	철물 및 난방용구 소매업
3	◉◉ 정수설비공사 도매 정수장치	41010	생활용수 공급업	51812	건설 및 광업용 기계장비 도매업
4	◈◈ 건설 주택건설 주택건설및분양	45212	아파트 건설업	45211	단독 및 연립주택 건설업
5	▣▣ 숯불닭바베큐 치킨호프점 바베큐 닭	55221	피자, 햄버거 및 치킨 전문점	55233	간이 주점업

를 하였을 때의 오류들은 하단의 Table 5와 같이 정리할 수 있다. 예를 들어 “○○시어터 가전제품 도매 오디오(음향기기)”라는 사업체 설명이 들어왔을 때 모델이 예측한 코드 51430 은 '가전제품 도매업'이지만 정답 코드 52511은 '가전제품 소매업'이기 때문에 예측된 코드가 더 정확한 것을 볼 수 있다.

4.4.2 모델예측 오류

모델 예측 오류는 여러 가지가 있지만, 대표적으로 특정 단어 때문에 예측을 잘 못 하는 경우가 있다. 예를 들어 “□□역학원 역학강의, 역학서적소매 강의도서” 라는 사업체 설명이 입력되면 모델이 예측한 코드 52621 설명은 '서적 및 잡지류 소매업'이다. 모델이 이 코드를 예측한 이유는 서적 및 도서 라는 단어에 주의 집중 값이 커 예측을 한 것으로 유추되며, 이러한 오류들은 향후 연구에서 개선해야 할 점이 될 것이다. 또한, 기타 예시들은 하단 Table 6 에서 확인할 수 있다.

5. 결론

본 연구에서는 현재 다양한 분야에 적용되고 있는 딥러닝 기술을 이용하여 산업/직업 통계분류 자동코딩을 위한 시스템을 개발하였다. 특히 기존과 다르게 분류 수준에 상관없고 자연어 처리에 특화된 KoBERT모델을 이용하여 높은 자동코딩 정확도를 보여주었다. 또한 모델과 데이터에 대한 오류 분석을 통하여 향후 연구에서 개선할 점들을 분석하여 개선된 모델 개발 가능성을 제시하였다.

REFERENCES

[1] Y. K. Kang. (2001). Automatic coding system for industry and occupation classification. The Korean Association for Survey Research. *Fall Conference 2001,* 33-45.

[2] Population and Housing Census. (2020) Understanding of the Census. https://www.census.go.kr/cui/cuiDefView.do?q_menu=3&q_sub=1

[3] Statistics Korea. (Year Unknown) Statistics Korea Census on Establishments . https://kostat.go.kr/understand/info/info_kost/1/index.action?bmode=read&cd=S010004

[4] H. S. Lim. (2004). An automated Classification System of Standard Industry and Occupation Codes by Using Information Retrieval Techniques. *The Journal of Korean Association of Computer Education 7(4),* 51-60.

[5] C. K. Woo. (2020). *A Study on Automatic Coding of Korean Standard Industrial Classification Based on Deep Learning.* Masters dissertation. Korea University, Seoul.

[6] H. D. Cheol. (2007). *A Research on the Design and Implementation of the Automated Industry and Occupation Coding System.* Masters dissertation. Hannam University, Daejeon

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez & I. Polosukhin. (2017, December). Attention is all you need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000-6010).

[8] M. Thompson, M. E. Kornbau & J. Vesely. (2012). Creating an Automated Industry and Occupation Coding Process for the American Community Survey. Seattle : U.S Census Bureau.

[9] S. Wood, R. Muthyala, Y. Jin, Y. Qin, N. Rukadikar, A. Rai & H. Gao. (2017, December). Automated industry classification with deep learning. *In 2017 IEEE International Conference on Big Data (Big Data)* (pp. 122-129). IEEE. DOI : 10.1109/bigdata.2017.8257920

[10] K. He, X. Zhang, S. Ren & J. Sun. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). DOI : 10.1109/cvpr.2016.90

[11] J. S. Lee, S. P, Jun, & H. S. Yoo. (2018). A Study on Automatic Classification Model of Documents Based on Korean Standard Industrial Classification. *Journal of Intelligence and Information Systems, 24(3),* 221-241 DOI : 10.13088/jiis.2018.24.3.221

[12] S. Hochreiter & J. Schmidhuber. (1997). Long short-term memory. *Neural computation, 9(8),* 1735-1780. DOI : 10.1162/neco.1997.9.8.1735

[13] S. M. Park, C. W. Na, M. S. Choi, D. H, Lee & B. W. On. (2018). KNU Korean Sentiment Lexicon - Bi-LSTM-based Method for Building a Korean Sentiment Lexicon -. *Journal of Intelligence and Information Systems, 24(4),* 219-240. DOI : 10.13088/jiis.2018.24.4.219

[14] M. S. Choi, & B. W. On. (2019). A Comparative Study on the Accuracy of Sentiment Analysis of Bi-LSTM Model by Morpheme Feature. *Proceedings of KIIT Conference, 2019(6),* 307-309.

[15] Y. T. Oh, M. T. Kim & W. J. Kim (2019). Korean Movie-review Sentiment Analysis Using Parallel Stacked Bidirectional LSTM Model. *Journal of KIISE, 46(1),* 45-49 DOI : 10.5626/JOK.2019.46.1.45

[16] J. Devlin, M. W. Chang, K. Lee & K. Toutanova. (2019,

June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186).
DOI : 10.18653/v1/N19-1423

[17] H. J. Park & K. S, Shin. (2020). Aspect-Based Sentiment Analysis Using BERT: Developing Aspect Category Sentiment Classification Models. *Journal of Intelligence and Information Systems, 26(4),* 1-25
DOI : 10.13088/jiis.2020.26.4.001

[18] K. H. Kim, C. E. Park, C. K. Lee, & H. K. Kim. (2020). Korean End-to-end Neural Coreference Resolution with BERT. *Journal of KIISE, 47(10),* 942-947.
DOI : 10.5626/JOK.2020.47.10.942

[19] Y. S. Choi & K. J. Lee. (2020). Performance Analysis of Korean Morphological Analyzer based on Transformer and BERT. *Journal of KIISE, 47(8),* 730-741.
DOI : 10.5626/JOK.2020.47.8.730

[20] T. Kudo & J. Richardson. (2018, November). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 66-71).
DOI : 10.18653/v1/D18-2012

임 정 우(Jungwoo Lim) [학생회원]

· 2019년 8월 : 성균관대학교 문헌정보학과 전공 (문헌정보학사)
· 2019년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : 상식추론, 딥러닝, 자연어처리
· E-Mail : wjddn803@korea.ac.kr

문 현 석(Hyeonseok Moon) [학생회원]

· 2021년 2월 : 고려대학교 수학과(이학사)
· 2021년 2월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : 인공지능, 자연어처리, 기계번역
· E-Mail : glee889@korea.ac.kr

이 찬 희(Chanhee Lee) [학생회원]

· 2013년 8월 : 서강대학교 컴퓨터공학심화(학사)
· 2016년 8월 ~ 현재 : 고려대학교 컴퓨터학과 석박사 통합 과정
· 관심분야 : 인공지능, 자연어처리, 딥러닝
·E-Mail : chanhee0222@korea.ac.kr

우 찬 균(Chankyun Woo) [학생회원]

· 2008년 2월 : 순천향대학교 컴퓨터공학과 (공학사)
· 2020년 8월 : 고려대학교 컴퓨터정보통신대학원 (공학석사)
· 현재 : 통계청 조사시스템관리과 전산주무관
· 관심분야 : 통계분류, 머신러닝
·E -Mail : ckwoo@korea.kr

임 희 석(Heuiseok Lim) [종신회원]

· 1992년 2월 : 고려대학교 컴퓨터학과 (학사)
· 1994년 2월 : 고려대학교 컴퓨터학과 (석사)
· 1997년 2월 : 고려대학교 컴퓨터학과 (박사)
· 2008년 ~ 현재 : 고려대학교 컴퓨터학과 교수
· 관심분야 : 통계분류, 머신러닝
· E-Mail : limhseok@korea.ac.kr