비정형적 아동 학습지 이미지의 OCR 후처리를 위한 신경망 언어 모델 기반 텍스트 편집 도구†

안의진*1, 양희찬*1, 지은경2, 홍신1

1한동대학교 전산전자공학부, 2한국과학기술원 전산학부

21700416@handong.ac.kr, 21800436@handong.ac.kr, ekjee@se.kaist.ac.kr, hongshin@handong.edu

A Neural Language Model-based Text Editing Tool for OCR Result of Unstructured Childhood Education Materials

Eujin Ahn*1, Heechan Yang*1, Eunkyoung Jee2, Shin Hong1

1Handong Global University, 2KAIST

요약

아동 학습 자료는 비정형적 배치와 구성을 갖는 특성으로 인해, 범용 OCR 소프트웨어를 통해 온전한 문장의 형태로 추출이 어려워 OCR 결과 편집에 많은 수작업이 필요한 상황이다. 본 논문에서는 신경망 언어 모델(neural language model)을 활용하여 아동 학습자료 OCR 결과 후처리를 지원하는 도구를 설계하고 구현한 사례를 소개한다. 제안하는 도구는 한국어 언어 모델을 활용하여 여러 줄의 텍스트로부터 동일한 흐름에 따른 텍스트를 자동으로 식별하는 기능과, 하나의 흐름을 구성하는 연속적인 텍스트에서 문장 경계를 자동으로 탐지한 후 문장별로 분리한 텍스트를 출력하는 기능을 제공한다. 실제 동화책을 바탕으로 구상한 40개의 편집 테스크를 바탕으로 한 사례 연구에서 기계학습을 통해 도출한 언어 모델은 총 35개 테스크를 성공적으로 수행하였다.

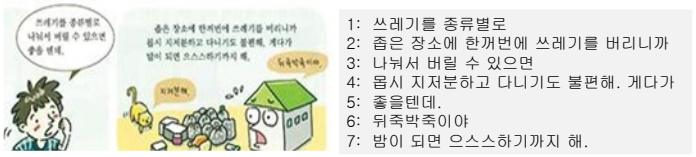
1. 서론

시각장애인 학습자의 통합 교육을 위해서는 비장애인 학습자를 대상으로 제작된 학습 자료를 시각장애인이 쉽게 사용할 수 있는 정제된 텍스트로 편집하여 제공하는 작업이 필수적으로 요구된다. 이러한 학습 자료 편집 작업은 스마트폰 애플리케이션 등의 형태로 제공되는 범용 OCR (Optical Character Recognition) 소프트웨어[1]를 통해 문서 내 글자를 인식하여 텍스트 데이터로 추출한 후, 수작업을 통해 텍스트 데이터로부터 음성합성, 점자 인터페이스에서 사용하기 편리한 온전한 형태의 문장으로 가공하는 방식이 일반적이다[2]. 이러한 텍스트 가공 작업에는 많은 노동이 소요되므로, 자원봉사자들이 온라인 협업함으로써 시각장애인 학습자를 위한 학습 자료를 제작하는 작업을 분담하기도 한다.

균일한 글꼴을 사용하여 단순한 구조로 정렬된 일반적인 문서 이미지와 달리, 비정형적 배치와 구성을 갖는 아동 학습 자료(예: 동화책, 학습 활동지)의 경우, OCR 소프트웨어로 추출한 텍스트가 원본 문서에서 의도한 흐름을 올바르게 재현하지 못하여, 많은 양의 후처리(post-processing) 비용이 소요되는 경우가 많다. 특별히, 범용 OCR 소프트웨어가 올바르게 처리하기 힘든 아동 학습 자료의 특성으로 다음 두 가지가 있다:

(1) 글 상자, 말풍선, 다단 편집으로 인해 서로 다른 흐름의 문장들이 하나의 이미지에 동시에 나타나는 경우가 많다.

(2) 비정형적인 글꼴, 문어체 표현으로 인해 문장 기호가 완비되지 않은 문장이 많으므로 구두점 등 지표를 활용해 문장 경계를 인식하기 어려운 경우가 많다.



(a) 텍스트가 포함된 학습자료 이미지 (b) OCR 텍스트 추출 결과의 예

그림 1. 비정형적인 학습 자료의 예

그림 1은 실제 초등학교 교과서의 일부분에 해당하는 이미지와 이를 일반적인 OCR 소프트웨어를 통해 문자 추출을 수행했을 때 나올 수 있는 결과의 예시이다. 사용자는 이런 OCR 결과로부터 각각의 흐름에 따른 텍스트를 구분한 후, 내용을 파악해 문장 경계를 찾아 분절하는 편집을 수작업으로 진행해야 한다. 초등학교 교과 과정에서 요구하는 많은 양의 비정형적 학습 자료를 고려할 때, 이러한 문서 편집의 비용은 시각장애인 학습자의 학업을 효과적으로 지원하는데 걸림돌이 된다.

본 논문에서는 비정형적 문서에 대한 OCR 결과를 온전한 문장으로 편집하는 작업의 자동화를 위해 신경망 언어 모델(neural language model)을 활용한 OCR 결과 후처리 도구를 설계하고 구현한 사례를 소개한다. 사용자의 수작업 편집 노력을 줄이기 위해, 제시하는 도구는 순환 신경망(recurrent neural network) 형태의 한글 언어 모델을 활용하여 다음 두 가지 기능을 제공한다:

- **흐름 분류**: 여러 줄의 텍스트로부터 동일한 흐름에 따른 텍스트를 자동으로 식별함으로써 흐름별로 텍스트를 분리하여 출력한다.
- **문장 경계 식별**: 하나의 흐름을 구성하는 연속적인 텍스트에서 문장 경계를 자동으로 탐지한 후 문장별로 분리한 텍스트를 출력한다.

이때, 구두점 등 문장 기호를 배제하고 주어진 한글 글자

---

* 두 저자의 본 연구 기여도는 동일함(equally contributed)

† 본 논문의 연구는 정부의 재원으로 소프트웨어중심대학 지원사업(2017-0-00130)과 한국연구재단의 기초연구사업 (NRF-2022R1I1A1A01072004)의 지원을 받음

데이터를 바탕으로 문장 경계 식별을 수행하는 모델을 학습함으로써 아동 학습자료의 특성을 고려하도록 하였다. 이 도구를 활용할 경우, 범용 OCR 소프트웨어를 통해 추출한 텍스트를 편리하게 편집할 수 있도록 지원함으로써 시각장애인 학습자를 위한 학습자료 작성의 생산성을 향상할 수 있다.

2. 여러 흐름이 혼재된 텍스트 조각의 분류

2.1. 흐름 분류를 위한 신경망 언어 모델 구성

본 논문에서는 편집 작업에 들어가는 자원을 줄이기 위해 올바른 문장 흐름을 분류하는 순환 신경망 언어 모델을 구축하는 방법을 제시한다. 그림 1은 이에 대한 예로 볼 수 있다. 그림 1(a)에서 하나의 말풍선과 또 다른 글 상자가 있고, 각 위치에 적혀 있는 문장은 서로 다른 의도의 흐름을 가지고 있다. 그림 1(b)에서는 OCR 소프트웨어를 통해 해당 문서에 적힌 글자를 추출한 결과이다. 좌측 그림에서는 각 문장들의 의도한 흐름을 쉽게 인지할 수 있는 반면, OCR를 통해 추출한 글자들은 다른 흐름을 갖은 문장들로 섞여 있기 때문에 올바른 흐름을 분석하기가 어렵다. 이에 시각장애 아동을 위한 문서를 제작하기 위해 OCR 소프트웨어로 추출한 문장을 편집하여 원본 문서의 흐름대로 올바르게 재배치해 주는 작업이 요구된다.

문자들의 흐름 분류에 대한 실험 결과를 도출하기 위해 두 개의 문장을 받아 서로 같은 흐름의 문장인지 혹은 다른 흐름의 문장인지 분류하는 이진 분류(binary classification) 언어 모델을 구성했다. 해당 모델에서는 임베딩 계층(embedding layer), 두 개의 양방향 순환 신경망(bidirectional recurrent neural network)에 역할을 수행하는 LSTM (long short-term memory) 계층[3], 두 개의 완전 연결 계층(fully connected layer)으로 구성되어 있으며 손실 함수(loss function)로는 log softmax 함수를 사용한다.

해당 모델의 입력 값으로는 두 개의 문장이 있다. 먼저 오는 prefix 문장이 있고 그 뒤따라오는 postfix 문장이 있다. 한글 문장 분류에서는 각 글자를 자모 단위로 분절하여 모델에 학습했을 때 높은 성능을 갖는다[4]. 그러므로 본 신경망 언어 모델은 최대 21개의 자모 단위로 분절된 한글 문장을 입력 값으로 받는다. 이때 postfix 문장을 입력으로 받는 두 번째의 LSTM을 고려했을 때, postfix의 문장이 자모 단위로 분리되기 전 먼저 각 문자 순서를 역순으로 변화하여 분절한다. 문장에서의 자모 개수가 21개 이하 일 때 prefix 문장의 앞부분을 padding으로 채워주며 postfix 문장은 문장의 뒷부분을 padding으로 채운다. 입력 값은 제일 먼저 임베딩 계층을 거치게 되며 각 자모는 300차원의 벡터 값으로 나오게 된다. 이어 두 개의 LSTM은 각 prefix encoder와 postfix encoder로 부르며, prefix 문장으로부터 나온 벡터 값 그리고 postfix 문장으로부터 나온 벡터 값은 각각의 encoder로 입력된다. Encoder의 은닉 층 크기는 800으로 주어지며 양방향 LSTM이기에 마지막 은닉 층 값은 1600의 크기를 갖는다. 각 prefix와 postfix encoder에서 나오는 마지막 은닉 상태(hidden state) 값을 서로 이어 붙여주며(concatenate) 첫 번째 완전 연결 계층으로 입력한다. 활성화 함수로는 ReLu 함수를 사용했으며 해당 완전 연결 계층의 값은 300의 크기로 두 번째의 완전 연결 계층에 입력되며, 이는 첫 번째 완전 연결 계층과 동일하게 ReLu 함수를 거쳐 2개의 값으로 나온다. 해당 2개의 값은 손실 함수에 입력이 되어 그 결과는 각 분류의 확률이 된다.

표 1. 흐름 분류 모델의 검증 성능 결과

	precision	recall
다른 흐름 (False)	80.76%	76.27%
같은 흐름 (True)	77.54%	81.85%

2.2. 분류기 학습 결과

제안한 언어 모델 기반 분류기를 Korean Contemporary Corpus of Written Sentence에서 제공되는 '`KCC150 Korean sentence EUCKR.txt`'와 '`KCCq Korean sentence EUCKR.txt`'를 바탕으로 생성한 데이터를 바탕으로 학습하였다. 해당 한글 코퍼스에서 총 100,000개의 문장 데이터에 구두점을 모두 제거하여 1 대 1 비율로 같은 흐름을 지닌 두 문장 그리고 다른 흐름을 지닌 두 문장으로 나뉜다. 그러므로 해당 100,000개의 데이터는 본 실험에 사용한다. 본 흐름 분류 모델에서는 100,000개의 데이터를 81,000개 훈련 데이터, 9,000개 검증 데이터 그리고 10,000개의 테스트 데이터로 나누어 사용한다. 표 1을 통해 모델 학습 결과 검증 데이터의 성능을 확인할 수 있다. 본 실험에서 테스트 데이터의 경우는 총 정확도는 78.75%를 달성하는 모델을 얻을 수 있었다.

3. 문장기호가 없는 문장의 경계 탐지

3.1. 문장 경계 탐지를 위한 언어 모델 구축

시각장애 아동 학습자를 위해 원본 문서에서 추출한 텍스트 데이터를 온전한 형태의 음성합성, 점자 인터페이스로 사용하기 위해서는 추출한 텍스트를 문장 단위로 경계를 구분하는 과정이 필요하다. 기존에 문장 경계 탐지를 위한 신경망 언어 모델을 구축하는 경우 구두점을 포함하고 있는 학습 데이터를 통해 모델을 훈련시키는 경우가 대부분이다. 하지만 구두점을 지표로 활용하여 모델을 구축하게 되면 문장 기호와 구두점이 없는 텍스트의 사례에서 경계 탐지에 있어서 낮은 정확도를 가지게 되는 한계점이 있다.

본 논문에서는 비정형적인 글꼴, 문어체 표현으로 문장 경계를 인식하기 어려운 아동 학습 자료의 특성을 고려하여 구두점을 포함하고 있지 않는 학습 데이터를 구축하고 순환 신경망 형태의 한글 모델을 사용하여 문장 기호가 없는 문장의 경계 탐지 방법을 제시한다. CRFs 형식의 기계학습을 사용하여 구두점을 가지고 있지 않은 문장의 경계를 탐지하는 기존 연구[5]와 달리, 본 연구에서는 신경망 모델을 이용하였다.

문장 경계 탐지를 이진 분류(binary classification) 모델은, 흐름 분류를 위한 신경망 언어 모델과 같은 구조로, 임베딩 계층(100차원), 두 양방향 LSTM 인코더(은닉층 200차원)로 구성하였고, 두 LSTM 출력 상태를 합친 값을 이층 구조의 ReLu를 사용한 완전 연결 계층(은닉층 300차원)을 거쳐 이진 출력값을 결정하도록 했다. 손실 함수는 log softmax를 사용했다. 값을 서로 이어 붙여주며(concatenate) 첫 번째

완전 연결 계층으로 입력한다.

표 2. 문장 경계 탐지 모델의 학습 결과

	precision	recall
잘못된 문장 경계	99.4%	100.0%
올바른 문장 경계	100.0%	99.4%

3.2. 문장 경계 탐지 모델 학습 결과

학습 데이터는 흐름 분류 모델(2장)의 경우와 동일하게 Korean Contemporary Corpus of Written Sentence에서 제공된 'KCC150 Korean sentence EUCKR.txt'와 'KCCq Korean sentence EUCKR.txt'를 사용하였으며, 전처리 과정을 통해 문장 기호를 모두 제거하고 자모 단위의 순열로 풀어서 표현하였다. 코퍼스로부터 얻은 총 60,000개의 문장을 활용하여 prefix와 postfix가 문장의 종료와 문장의 시작에 해당하는 15,000개 데이터(문장 경계가 맞는 경우)를 생성하였고, prefix와 postfix가 한 문장을 나누는 경우(문장 경계가 아닌 경우)에 해당하는 15,000개 데이터를 생성하였다. 이렇게 생성한 총 30,000 데이터를 훈련 데이터로 24,300개, 검증 데이터로 2,700개, 그리고 테스트 데이터로 3000개로 나누어 모델 학습을 수행했다.

표 2는 문장 경계 탐지 모델의 학습 결과를 나타낸다. 학습을 통해 도출한 모델은 구두점 정보를 사용하지 않고도, 99.7%의 정확도로 문장 경계를 올바르게 탐지하였다.

본 연구에서는 구두점이 존재하는 데이터로 학습한 문장 경계 모델이 구두점이 없는 경우에 어떠한 성능을 보이는지 알아보는 실험을 수행했다. 이 실험에서는 문장 기호를 제거하는 전처리 작업을 수행하지 않고 24,300개의 훈련 데이터와 2,700개의 검증 데이터를 생성한 뒤, 문장 기호를 제거하여 생성한 3000개의 테스트 데이터를 생성하여 문장 경계 탐지 성능을 평가했다. 실험 결과, 문장 기호가 있는 데이터로 학습한 모델은 구두점이 없는 테스트 데이터에 대하여 정확률 49.77%의 낮은 정확도를 보였다.

4. 동화책 문장 사례를 활용한 사례 연구

모델 학습을 통해 구현한 흐름 분류기와 문장 경계 탐지기가 아동 학습자료에 나오는 문장에 효과적으로 작동하는지 알아보기 위해 실제 아동을 위한 동화책을 활용해 실제적인 문서 편집 작업 테스크를 만들고, 이에 대한 사례 연구를 수행하였다.

4.1. 흐름 분류 작업의 사례 연구

실제 흐름 분류 작업을 효과적으로 지원하는지 평가하기 위해, 실제 아동 동화책 본문으로 20개의 흐름 분류 테스크를 구상하여 사례 연구를 수행했다. 각 흐름 분류 테스크는 하나의 prefix 텍스트와 하나의 올바른 post fix 텍스트 후보와 하나의 올바르지 않은 postfix 후보로 구성되어 있다. 예를 들어, prefix 텍스트로는 "새 왕비는 요술거울"이 주어지며, 올바른 postfix 텍스트로는 "을 가지고 있었어요", 올바르지 않은 postfix 텍스트로는 "그 집의 주인이 돌아왔어요"가 주어진다. 이때 학습한 분류기를 통해 두 개의 postfix 텍스트 중 prefix 텍스트에 보다 호응하는 것을 고르도록 하였다.

사례 연구 결과, 총 20개 테스크 중에 15개에서 올바른 분류가 가능하였다. 이를 통해, 학습한 모델은 흐름 분류 작업에 효과적으로 쓰일 수 있음을 확인하였다.

4.2. 문장 경계 탐지 테스크에 대한 사례 연구

도출한 모델이 실제 문장에 대한 경계 탐지 테스크를 잘 수행하는지 평가하기 위해, 실제 아동 동화책 본문으로 20개의 경계 탐지 테스크를 구상하여 사례 연구를 수행했다. 각 테스크는 한 쌍의 prefix 텍스트와 postfix 텍스트를 입력 받아 두 텍스트 사이에 문장 경계가 존재하는지 여부를 판별하도록 했다. 예를 들어, prefix 텍스트로 "나만의 나눔 계획을 세워 봅시다"가 주어지며 postfix 텍스트로는 "꼭 기업가가 되고"를 준 경우는 실제로 문장 경계가 존재하는 사례이며, prefix 텍스트로 "책을 먹는 여우가 새로운"과 postfix 텍스트로 "책을 찾아서 도서관을 찾아왔어요"가 주어진 경우는 문장 경계가 존재하지 않는 사례이다.

총 20개 테스크에 대한 사례 연구에서 문장 경계 탐지 모델은 모든 경우에서 문장 경계를 올바르게 탐지하였다. 이 결과를 통해 학습한 모델을 문장에 대한 경계 탐지 작업에 효과적으로 쓰일 수 있음을 확인했다.

5. 결론 및 향후 연구

본 논문에서는 신경망 언어 모델을 활용한 시각장애 아동의 학습자료 편집을 위한 비정형적 문서의 OCR 결과 후처리 도구를 설계하였다. 제안된 시스템은 현재 여러 단계의 수작업으로 처리되고 있는 문서 편집 과정을 양방향 LSTM 모델의 흐름 분류와 문장 경계 식별을 통해 자동화할 수 있는 방법을 제시하였다.

향후 연구에서는 설계한 모델을 통해 판별된 문장의 흐름을 자동으로 연결해 주는 알고리즘 및 서비스를 구축하고 이를 통해 제시한 두 개의 모델을 하나의 시스템으로 통합하여 완전한 문서 편집 후처리 자동화 서비스를 개발할 계획이다. 실제 시각장애 아동을 위한 학습자료 데이터를 더 많이 확보하고 아동 학습지에서 나타나는 다양한 형태를 고려하여 수작업을 줄이고 서비스를 제공하여 실제 사용자들을 대상으로 편의성 및 실용성 연구를 수행하겠다.

참조문헌

[1] 정유진, "ICT 비전공자, 월드 IT 쇼에 다녀오다" (2022 년 5 월 1 일), 오마이뉴스, 2022

[2] 김지웅, 이강, 김경미, 시각 장애인용 신문 구독 프로그램을 위한 이미지에서 표 구조 인식, 멀티미디어학회 논문지, 19 (11), 2016

[3] H. Sak, A. Senior, F. Beaufays, Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition, arXiv:1402.1128, 2014

[4] 김진성, 김경민, 손준영, 박정배, 임희석, 한국어 단어 및 문장 분류 태스크를 위한 분절 전략의 효과성 연구, 한국융합학회논문지, 12 (12), 2021

[5] 염하람, 김재훈, 한국어 SNS 문서에 적합한 문장 경계 인식, 한글 및 한국어 정보처리 학술대회 논문집, 2021