교육과정평가연구
The Journal of Curriculum and Evaluation
2021, Vol. 24, No. 4, pp. 223~238
DOI: https://doi.org/10.29221/jce.2021.24.4.223

순환신경망 장단기 기억(LSTM)을 이용한 자동 채점의 가능성 탐색[1)]

박강윤 (한국지능정보사회진흥원)*
이용상 (인하대학교)**
신동광 (광주교육대학교)

요약

본 연구에서는 순환신경망의 일종인 장단기 기억(Long-Short Term Memory: LSTM)를 이용하여 영어 에세이 자동 채점 가능성을 탐색하였다. LSTM은 기존의 순환신경망(Recurrent Neural Network; RNN)이 갖는 장기의존성의 문제를 극복하기 위해 제안된 학습 모델로, 본 연구에서는 이러한 LSTM 모델을 활용하여 영어 에세이 데이터를 학습시킨 후, 별도의 평가 데이터를 통해 LSTM의 성능을 평가하였다. 이분(二分) 자료의 형태를 갖는 선다형 채점 데이터와 달리 에세이 채점 데이터는 다분(多分) 자료의 형태를 가지므로 본 연구에서는 다항 분류가 가능하도록 학습 모델을 구축하여 점수를 예측하였으며, 이러한 LSTM 학습 모델을 여섯 가지 지표(정확성, 정밀도, 재현율, F1, 카파, 상관계수)로 평가하였다. 그 결과를 살펴보면, 본 연구에서 구축한 LSTM 학습 모델이 학생들의 에세이 점수를 양호한 수준에서 예측할 수 있음을 확인하였다. 또한 학습 모델의 성능을 결정하는 주요 요인 중 하나가 데이터의 질과 양임을 감안할 때 향후 충분한 양질의 데이터를 구축하여 학습할 경우 자동 채점의 정확성을 보다 향상시킬 수 있을 것으로 기대된다. 후속 연구로는 최적의 에세이 자동 채점 알고리듬을 도출하기 위해 향후 다양한 순환신경망 모델을 비교 검토하는 실증 연구들이 수행될 필요가 있다.

주제어 : 순환신경망, 자동 채점, 장단기 기억

1) 이 논문은 2021년도 인하대학교의 지원에 의하여 연구되었음.
* 제1저자, datapark@nia.or.kr
** 교신저자, yong21c@gmail.com

I. 서 론

정부는 “인공지능시대 교육정책방향과 핵심과제”를 통해 인공지능 시대에 능동적으로 대처하기 위한 정책 방안을 제시하였다(관계부처합동, 2020.11.20.). 정부 발표에 따르면 앞으로 공교육에 인공지능을 비롯한 에듀테크(EduTech)의 도입이 확대되고, 이를 바탕으로 인공지능과 인간과의 협업을 통해 양질의 맞춤형 수업을 제공할 수 있는 ‘초개인화 학습 환경’을 조성하겠다고 공언하였다. 이와 같은 정부의 발표는 앞으로 교육현장에서 교수 학습 활동에 인공지능을 적극적으로 활용할 것임을 시사하며, 따라서 교수 학습 방법과 평가의 일관성을 고려한 인공지능을 활용한 평가 방안에 대한 다양한 접근이 이루어질 필요가 있다. 특히 최근 학생평가와 관련하여 4차산업혁명시대에 학생들의 창의성과 고차원적 사고력을 측정하기 위해서는 기존의 선다형 중심의 평가로는 한계가 있다는 지적과 함께 학생들의 고차원적 사고력을 측정할 수 있는 논술형 평가로의 전환이 시급하다는 주장이 제기되고 있다(김태준 외, 2020; 박혜영 외, 2019).

이러한 사회적 요구에 따라 정부는 최근 “2022 개정 교육과정 추진계획”을 통해 미래핵심역량 함양을 위해 학교 현장에서 서·논술 평가를 확대하겠다고 발표한 바 있으며(교육부, 2021.4.20.), 이에 힘입어 논술형 평가 확대에 대한 다양한 논의가 이루어지고 있다. 그러나 이러한 논술형 평가 확대에 대한 기대에도 불구하고 논술형 평가에서의 채점 부담과 채점의 공정성 문제 등으로 인해 실제 논술형 평가를 시행하는 것은 쉽지 않은 상황이다. 예컨대, 논술형 평가를 위해서는 대규모의 전문 채점 인력이 필요하며, 더욱이 인간 채점자의 채점 엄격성 차이나 채점 피로도 등에 따라 학생의 점수가 달라질 경우 논술 채점 결과의 신뢰성을 담보할 수 없다. 따라서 논술형 평가로의 전환이 이루어지기 위해서는 논술 채점이 가지고 있는 현실적인 문제점을 해결할 수 있는 대안 마련이 선행될 필요가 있으며, 이를 위한 기초 연구 수행이 시급한 실정이다.

해외에서는 대규모 논술 채점을 위한 대안으로서 자동 채점에 대한 연구들이 오래전부터 진행되어 왔으며, 다양한 자동 채점 프로그램들이 이미 상용화되기도 하였다. 예를 들어, 1960년대 Ellis Page는 전산 언어학, 인공지능 등의 학문을 융합하여 컴퓨터를 활용하여 자동 평가를 위한 PEG(Project Essay Grade) 소프트웨어를 개발하여 상용화한 바 있으며, ETS에서는 서술형 평가를 위해 개념 기반으로 한 자동 채점 시스템인 C-rater와 기계학습 기반의 e-rater를 개발하여 활용하고 있다(박세진, 하민수, 2020; Leacock & Chodorow, 2003). 국내에서도 영어 에세이 자동 평가를 위한 일련의 자동 채점 시스템 개발 연구가 한국교육과정평가원을 중심으로 수행되었지만(시기자 외 2012; 진경애 외 2006, 2007, 2008, 2011), 정부 정책의 변화에 따라 자동 채점에 대한 지속적인 연구가 수행되지 못하였다. 그러나 최근 인공지능의 발달과 더불어 해외에서 서·논술 평가에 인공지능을 활용하고자 하는 꾸준한 시도들이 있으며, 이러한 시도들(Attali & Burstein, 2006; Kumar & Boulanger, 2020)의 주요 관심사 중의 하나는 인공지능을 활용한 자동 채점이다. 그러나 아직까지 국내에서는 인공지능 알고리듬을 자동 채점에 접목시키고자 하는 시도가 매우 희소한 상황이며, 따라서 이 분야에 대한 지속적인 연구를 통해 향후 서·논술 평가 시대에 채점의 문제를 해결할 수 있는 대안으로 자동 채점을 준비할 필요가 있다.

자동 채점 시스템을 성공적으로 구현하기 위해서는 기계학습을 위한 양질의 데이터 구축과 함께 이들 데이터를 학습하기 위한 자동 채점 알고리듬에 대한 체계적인 연구가 필요하다. 그러나 아직까지 국내에서 자동 채점 알고리듬에 대한 연구들은 매우 제한적으로 이루어졌다(예, 박세진, 하민수, 2020; 시기자 외, 2012; 이용상 외, 2013; 하민수 외, 2019). 이들 연구 중 시기자 외(2012)의 연구와 이용상 외(2013)의 연구는 통계처리 기반의 기계학습 알고리듬을 적용한 연구였으며, 최근 박세진과 하민수(2020)의 연구는 순환신경망(RNN) 기반의 딥러닝을 적용한 연구이다. 이중 특히 박세진과 하민수(2020)의 연구는 한국어 서답형 문항에 RNN을 적용하여 인간 채점과 자동 채점 간 일치율을 0.9 이상 도출하는 고무적인 결과를 보여주고 있다. 그러나 이들의 연구는 과학 교과에 한정하여, 서답형 문항의 답안을 정답과 오답으로만 분류하는 방식의 자동 채점을 구현하였다는 한계점을 가지고 있다. 다시 말해, 박세진과 하민수가 제언한 자동 채점 알고리듬으로는 실제 인간 채점자들의 방식과 동일하게 서답형 답안의 채점 시 부분 점수를 부여할 수 없었다. 따라서 이러한 문제를 극복하기 위해서는 부분 점수에 대한 자동 채점이 가능한 학습 알고리듬 연구가 수행될 필요가 있다.

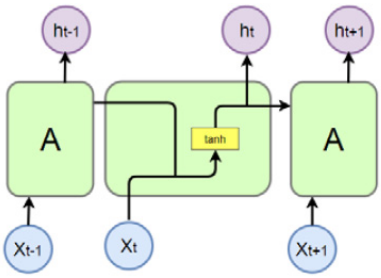
일반적으로 RNN은 자연어와 같은 순차적인 정보를 처리하는 데 적합한 방법으로 알려져 있다(박세진, 하민수, 2020; Mikolov et al, 2010). 이 알고리듬은 직전 데이터와 현재 데이터를 고려하여 다음의 데이터를 예측하는 방법을 사용하므로 RNN의 예측치는 이전 결과가 주된 영향을 준다. 그러나 이러한 RNN은 글이 길어질수록 앞의 정보가 뒤로 충분히 전달되지 못해 학습 능력이 저하되는 장기의존성의 문제(the problem of long term dependencies)가 발생하는 것으로 보고되고 있으며(방성혁 외, 2018; Shewalkar, Nyavanandi, Ludwig, 2019), 이러한 문제점을 극복하기 위한 방안으로 장단기 기억(Long-Short Term Memory: LSTM)가 제안되었다(Hochreiter & Schmidhuber, 1997). RNN에 비해 LSTM의 성능이 우수하다는 점을 실증적으로 입증한 선행 연구(Shewalkar, Nyavanandi, Ludwig, 2019)가 있으며, 본 연구에서는 이러한 선행 연구 결과를 토대로 LSTM을 적용한 자동 채점의 가능성을 탐색하였다. 국내 자동 채점 연구에서는 아직까지 RNN까지만 적용하였고, RNN과 같은 딥러닝 기반의 알고리듬을 활용하여 다분 자료(부분점수를 허용하는 채점 자료)에 대한 자동 채점 연구는 수행된 바 없다. 따라서 본 연구에서는 실제 에세이 채점에서 부분점수를 허용하는 현실을 감안하여 부분점수를 허용하는 서답형 답안 자동 채점에서 LSTM 인공지능 학습모델의 성능을 검증해 보았다.

II. 이론적 배경

1. 순환신경망(RNN)

McCulloch와 Pitts (1943)가 뇌의 생물학적 뉴런의 네트워크에 착안하여 인공신경망(artificial

neural network)을 제안한 이후로 인공신경망 계열의 다양한 학습 모델이 개발되었으며, RNN은 이러한 인공신경망 모델의 한 종류이다. RNN은 순차적으로 입력되는 데이터 패턴을 분석하여 특정 시점의 결과를 예측하기 위해 사용되는 인공신경망 모델로서 내부에 정보가 지속되도록 순환구조를 가지고 있으며 은닉계층에 과거의 데이터를 기억하여 학습을 진행한다.



[그림 1] 순환신경망
* 출처: 주일택, 최승호(2018: p.205)

[그림 1]에서 보여 주듯이 RNN은 현재 시점의 값(x_t)과 이전 시점의 값(x_{t-1})을 함께 입력받아 처리하는 과정을 거치며 이 과정에서 활성화 함수로 하이퍼볼릭(hyperbolic, 쌍곡선) 탄젠트 함수(tanh)를 사용한다. RNN에서는 이러한 활성화 함수를 통해 출력값을 내보내는 역할을 하는 노드(node)를 셀(cell)이라 하며, 이와 같은 셀은 이전 시점의 입력값을 기억하는 메모리 역할을 하기 때문에 메모리 셀이라고도 한다. RNN은 이전 시점의 값과 현재 값을 이용하여 다음 시점의 값을 예측하는 과정을 반복하면서 이전 시점의 입력값이 다음 시점의 출력값에 어떤 영향을 미치는지를 학습할 수 있다는 장점이 있다(주일택, 최승호, 2018). 따라서 이와 같은 RNN은 순차적인 정보를 처리하는 데 적합하며, 순서 길이에 상관없이 입력과 출력을 받아들일 수 있는 네트워크 구조이기 때문에 상황에 맞게 다양하고 유연한 구조를 만들 수 있는 장점을 가진다(박세진, 하민수, 2020).

2. 장단기 기억(LSTM)

RNN은 직전 데이터와 현재 데이터를 고려하여 다음의 데이터를 예측하는 방법을 사용하므로 RNN의 예측치는 이전 결과가 큰 영향을 줄 수밖에 없다. 따라서 RNN은 시점이 길어질수록 앞의 정보가 뒤로 충분히 전달되지 못해 학습 능력이 저하되는 장기의존성의 문제를 가지고 있다. 이와 같은 장기의존성의 문제를 해결하기 위해 Hochreiter와 Schmidhuber(1997)은 LSTM을 제안하였으며, LSTM은 RNN의 장기 의존성을 해결하기 위해 제안된 내부 연산 구조로서 먼 시점의 데이터를 은닉층의 연산 결과와 함께 다음 시점의 은닉층으로 전달하는 특징을 가지고 있다. 이와 같은 LSTM을 수식으로 표현하면 다음과 같다(Shewalkar, Nyavanandi, & Ludwig, 2019).

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_t \bullet [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\widetilde{C_t} = \tanh(W_c \bullet [h_{t-1}, x_t] + b_C) \quad (3)$$

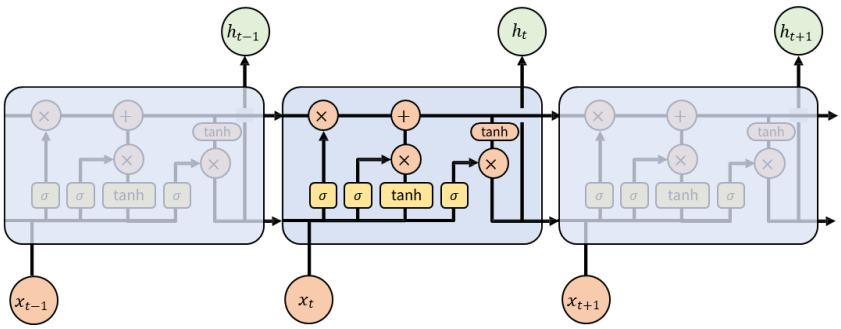
$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \quad (4)$$

$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

σ: 시그모이드 함수, tanh: 하이퍼볼릭 탄젠트 함수, i: 입력 게이트, f: 망각 게이트
o: 출력 게이트, C: 메모리 셀 정보, $\widetilde{C}$: 새 메모리 셀 정보, W: 가중치 행렬, h: 은닉층, b: 편향성

LSTM에서는 입력(i), 망각(f), 출력(o) 게이트를 통해서 불필요한 정보는 제거하고 필요한 정보만 저장하게 된다. 위 수식에서 입력 게이트(i_t)와 망각 게이트(f_t)는 현재 시점에서 입력값(x_t)과 이에 대한 가중치(W_t)와의 곱과 이전 시점 은닉층(h_{t-1})과 이에 대한 가중치(W_t)의 곱을 시그모이드(Sigmoid) 함수로 산출한 값이며, 새 메모리의 셀 정보는($\widetilde{C}$)는 이전 시점의 은닉층과 현재 시점의 입력값에 대한 가중치(W_c)의 곱을 하이퍼볼릭 탄젠트 함수로 산출한 값이다. 망각 게이트에서는 시그모이드 함수를 통해 결과는 0과 1사이의 값을 산출하게 되며 1에 가까울수록 많은 양의 정보를 보전한 것으로 해석한다. 장기기억이라 할 수 있는 C_t는 이점 시점의 셀 정보와 망각 게이트 간의 원소별 곱을 통해서 이전 시점이 셀에서 보전할 정보를 선택하게 되고, 이를 현재 시점의 새로운 셀 정보와 입력게이트 간의 원소별 곱을 통해 선택된 정보와 더하여 산출된다. 이렇게 산출된 현재 시점의 셀 정보(C_t)는 하이퍼볼릭 탄젠트 함수 값으로 산출되어 출력 게이트와의 원소별 곱을 통해 단기 기억이라 할 수 있는 현재 시점의 은닉층(h_t)을 산출한다. 이와 같은 LSTM의 구조를 시각화하면 [그림 2]와 같다.



[그림 2] 장단기 기억(LSTM) 구조

3. 자동 채점 선행 연구

해외에서는 일찍이 영어 에세이에 대한 자동 채점 연구가 시작되었으며, 이에 따라 다양한 자동 채점 프로그램들이 개발 되었다. 기존의 해외 자동 채점 연구는 크게 회귀분석에 기반한 자동 채점, 자연어 처리에 기반한 자동 채점, 인공지능에 기반한 자동 채점으로 나누어 볼 수 있으며(시기자 외, 2012), 최초의 자동 채점 프로그램이라 할 수 있는 PEG(Page, 1966)는 이중 회귀분석에 기반한 프로그램이라 할 수 있다. PEG는 에세이 답안에서 추출된 자질들(features)을 추출하여 인간 채점자가 채점한 점수를 종속변수로 하고 이들 자질을 설명변수로 하는 회귀분석을 통해 점수의 예측력을 높이는 자질을 선별한 후, 이들 자질을 바탕으로 에세이 답안을 채점하는 알고리듬을 가지고 있다. 다음으로 자연어 처리에 기반한 자동 채점은 ETS의 e-rater(Attali & Burstein, 2006)가 대표적이라 할 수 있다. e-rater는 에세이 답안에서 추출된 특질을 예측 변수로 활용하여 점수를 산출하는 회귀분석 모형을 사용한다는 점에서 PEG과 같이 회귀분석에 기반한 자동 채점이라고 할 수 있으나, PEG와는 달리 코퍼스 기반의 회귀모형을 이용한 통계처리 기법과 자연어처리 기법을 모두 사용한다(진경애 외, 2012). 다음으로 인공지능에 기반한 자동 채점은 Vantage Learning의 IntelliMetric이나 MyAccess 등이 있다. 이들 자동 채점 프로그램들은 에세이 답안을 학습하여 채점모델을 구성하고 이 학습 모델을 에세이 답안 및 채점 데이터와 비교하여 타당성을 검증한 후 자동 채점을 시행하는 방식으로 구성된다. 이와 같이 이미 해외에서는 영어 에세이 자동 채점을 위한 여러 연구들을 통해 다양한 자동 채점 프로그램들이 개발되어 상용화되었다. 최근 인공지능에 기반한 자동 채점에서는 RNN과 같은 딥러닝 계열의 알고리듬들이 널리 활용되고 있으나, 이들 알고리듬들은 자동 채점을 통해 산출된 점수를 설명하는 데 어려움이 있다는 한계점을 가지고 있으며, 이에 따라 최근의 자동 채점 연구(Kumar & Boulanger, 2020)는 이러한 한계점을 극복하는 데 초점을 두어 활발히 진행되고 있다.

한편, 국내 자동 채점 연구는 초창기 한국교육과정평가원을 중심으로 수행되었다. 한국교육과정평가원은 국가영어능력평가시험 개발의 일환으로 영어 말하기와 쓰기 자동 채점 연구를 수행한 바 있으며, 이 중 대표적인 연구는 진경애 외(2011), 시기자 외(2012) 등이다. 이들 연구에서는 컴퓨터 기반 서답형 문항으로 시행되는 국가영어능력평가시험의 채점 부담을 해결하기 위해 영어 말하기와 쓰기 자동 채점 시스템을 개발 및 시범 적용하였다. 그러나 이들 연구는 교육 정책의 변화와 함께 국가영어능력평가시험 개발 사업이 중단됨에 따라 더 이상 진척되지 못하는 아쉬움을 남겼다. 그럼에도 불구하고 이들 연구에서 개발한 자동 채점 시스템을 이용하여 채점한 결과와 인간채점 간 상관계수는 0.8 이상, 유사일치도(±1점 차이까지 포함)는 0.9 이상으로 나타나는 등 자동 채점의 가능성을 확인하였다는 점에서 의미가 있다(시기자 외, 2012; 신동광 외, 2015). 이후 자동 채점 연구는 국어과 서답형 문항 자동 채점 시스템 개발을 목적으로 역시 한국교육과정평가원에서 그 명맥을 유지하였지만(노은희 외, 2012, 2013, 2014, 2015), 문장단위의 자동 채점에 그쳐 그 활용도가 매우 제한적이었다. 이상의 한국교육과정평가원을 중심으로 수행된 자동 채점 연구들은 알고리듬 측면에서 초기의 통계처리 기반의 기계학습 알고리듬들(예, 최대엔트로피)을 적용하였으나, 최근에는 RNN 기반의 딥러닝 방식을 적용한 자동 채점 연구로 전환되기 시작되고 있다.

국내에서 RNN을 적용한 자동 채점 연구는 박세진과 하민수(2020)의 연구가 아직까지는 유일하며, 이 연구에서는 RNN에 기반한 초등학생용 과학교과 서답형 문항의 자동 채점 시스템을 개발하였다. 이 연구에서는 서울 경기 지역 초등학생 462명의 학생들로부터 자료를 수집하여 RNN을 적용한 자동 채점 프로그램을 개발하였으며, 이 프로그램은 문항에 따라 정확도 0.997~0.954, 정밀도 0.997~0.963, 재현율 0.997~0.954, F1 0.997~0.953, 카파 0.992~.888의 값을 나타내는 등 매우 고무적인 성능을 보여주었다. 이 연구는 과학교과에서 자동 채점의 가능성을 확인하는 동시에 자동 채점 프로그램과 연계하여 학생들에게 채점 결과에 따른 피드백을 제공할 수 있는 시스템을 개발하여 자동 채점 프로그램을 교수 학습 활동에 활용할 수 있는 가능성을 탐색하였다는 점에서 의미가 있다. 그러나 서답형 문항임에도 불구하고 학생들의 답안을 정답과 오답으로만 구분하여 예측하였다는 점에서 한계가 있다. 이상의 국내에서 수행된 자동 채점 관련 주요 선행 연구를 요약하면 〈표 1〉과 같다.

〈표 1〉 자동 채점 국내 주요 선행 연구

연구	주요 내용
• 진경애 외(2011). • KICE-Pearson 영어 말하기, 쓰기 자동 채점 프로그램 타당성 연구	• 해외 말하기·쓰기 자동 채점 프로그램을 국가영어능력평가시험 말하기·쓰기 문항 유형에 최적화 • 인간채점자와의 상관 및 일치도 분석을 통해 자동 채점 프로그램의 타당성 검증
• 시기자 외(2012). • 국가영어능력평가시험 쓰기 자동 채점 프로그램 개발	• 단문형/에세이형 영어 쓰기 자동 채점 프로그램 개발 및 고도화 • 영어 쓰기 프로그램 타당성 검증 • 사용자 인터페이스 개선
• 노은희 외(2012). 대규모 평가를 위한 서답형 문항 자동 채점 방안연구	• 한국어 단어·구 수준 자동 채점 프로그램 개발 및 채점 신뢰도 분석 • 한국어 자동 채점 프로그램 개발 로드맵 제시
• 노은희 외(2013). • 대규모 평가를 위한 서답형 문항 자동 채점 프로그램 정교화 및 시범 적용	• 한국어 단어·구 자동 채점 프로그램 정교화 및 채점 신뢰도 분석 • 학업성취도 평가 서답형 문항 시범 적용
• 노은희 외(2014). • 한국어 서답형 문항 자동 채점 프로그램 개발 및 실용성 검증	• 한국어 단어·구 자동 채점 프로그램 실용화 개발 및 현장 적용 • 문장 자동 채점 프로그램 프로토타입 설계 및 채점 신뢰도 분석
• 노은희 외(2015). • 한국어 문장 수준 서답형 문항 자동 채점 프로그램 개발 및 적용	• 한국어 문장 자동 채점 프로그램 개발 및 채점 결과 분석 • 학업성취도 평가 서답형 문항 채점 적용
• 하민수 외(2019). • 학습 지원 도구로서의 서술형 평가 그리고 인공지능의 활용: WA3I프로젝트 사례	• 랜덤포레스트 기법을 이용한 한국어 서술형 문항 자동 채점 프로그램 개발 및 채점 결과 분석 • 과학 자식기반 서술형 문장 자동 채점 및 피드백 제공 서비스 적용
• 박세진, 하민수(2020). • 순환신경망을 적용한 초등학교 5학년 과학 서술형 평가 자동 채점시스템 개발 및 활용 방안 모색	• 순환신경망을 이용한 과학과 자동 채점 시스템 개발 및 성능 검증

III. 연구방법

1. 연구 자료

LSTM을 활용한 서답형 문항 자동 채점의 가능성을 탐색하기 위해 본 연구에서는 Kaggle(www.kaggle.com)의 데이터를 사용하였다. 본 데이터는 수기 채점의 고비용, 저효율의 문제를 개선을 위한 연구의 일환으로 Hewlett Foundation에서 총 10개의 에세이 문항을 개발하여 미국 고등학교 1학년(10학년)을 대상으로 수집한 데이터이다. 문항별로 1,704 ~ 2,400명의 학생 응답 데이터가 있으며, 본 연구에서는 이 중 8번 문항(그림 3 참조)을 응시한 2,398명의 에세이 데이터를 활용하였다. 본 연구에서 활용한 에세이 문항에서는 Leonard의 배경에 대한 정보가 Paul에게 어떤 영향을 미쳤는지에 대해서 답하도록 하였으며, [그림 3]의 채점 기준을 활용하여 응답자의 답안을 각각 0점, 1점, 2점으로 채점하였다.

During the story, the reader gets background information about Mr. Leonard. Explain the effect that background information has on Paul. Support your response with details from the story.

Paul finds out that Mr. Leonard was a track star but he could not read. 'No school wanted a runner who couldn't read.' Paul listened to Mr. Leonard about his past and realized that is was similiar to his present. Paul decided that because Mr. Leonard had helped him with track that he needed to help Mr. Leonard out with his reading. 'C'mon, Mr. Leonard, it's time to start your training.'

점수	채점 기준
2점	The response fulfills all the requirements of the task. The information given is text-based and relevant to the task.
1점	The response fulfills some of the requirements of the task, but some of the information may be too general, too simplistic, or not supported by the text.
0점	The response does not fulfill the requirements of the task because it contains information that is inaccurate, incomplete, and/or missing altogether.

[그림 3] 8번 문항 프롬프트 및 채점 기준

본 연구에서 2,398명의 에세이 데이터를 8:2의 비율로 훈련용과 평가용 데이터로 나누어(훈련용 데이터 1,918개, 평가용 데이터 480개) 사용하였으며, 훈련용 데이터는 LSTM 알고리듬이 학습하는 데 사용하고, 평가용 데이터를 활용하여 이렇게 학습한 LSTM의 알고리듬을 적용하여 학생들의 점수를 예측하는 방법으로 LSTM의 성능을 검증하였다. 본 연구에서 사용한 학생들의 에세이는 0점, 1점, 2점으로 채점되었으며, 각 점수의 빈도를 요약하면 〈표 2〉와 같다.

〈표 2〉 응답 자료의 빈도수

구분	0점	1점	2점	계
빈도	725	622	1,051	2,398

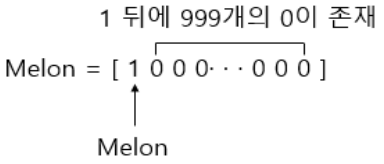
2. 분석 방법

본 연구에서는 Python 3.8.6을 이용하여 LSTM 모델을 구축하였으며 모델 적용을 위해 Tensorflow 2.4.0과 Keras 2.4.0 을 활용하였다. 영어 텍스트 데이터를 전처리 및 분석을 위해 nltk 3.5버전과 사용하였으며 수치화된 데이터를 처리하기 위하여 numpy 1.20.1, pandas 1.2.3을 사용하였고, 시각화를 위하여 matplotlib 3.3.2을 사용하였다. 모델의 학습과 평가를 위하여 영어 텍스트 데이터를 훈련용 데이터(80%)와 평가용 데이터(20%)로 나누어 사용하였다. 훈련용 데이터와 평가용 데어터의 분할 비율에 대한 절대적인 기준은 아직까지 없으며, 일반적으로 8:2의 비율 또는 7:3의 비율로 훈련용 데이터와 평가용 데이터를 나누고 있다. 데이터가 많을 경우 훈련용 데이터의 비율을 더 적게 할 수도 있으며, 본 연구에서는 선행 연구(박세진, 하민수, 2020)에서 적용한 훈련용 데이터와 평가용 데이터 분할 기준을 적용하여 훈련용 데이터와 평가용 데이터로 나누었다.

IV. LSTM 모델을 이용한 기계학습 및 점수 예측

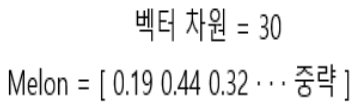
LSTM 모델 학습을 위해 우선 인간 채점자에 의하여 부분 점수가 매겨진 데이터를 자연어 처리 패키지인 NLTK(Natural Language Toolkit)를 통하여 특수문자, 이모티콘 등 분석에 필요가 없는 부분을 삭제하는 전처리 과정을 수행하였다. 다음으로 기계학습을 위하여 토큰화를 실시하였으며, 토큰화 데이터는 Word2Vec 모델을 활용하여 벡터화 작업을 진행 후 Keras의 임베딩 층(embedding layer)에 추가하였다.

본 연구에서는 분석의 효율성을 위하여 최대 길이의 응답을 기준으로 모든 응답 자료의 길이를 통일시킨 후 최대 응답 길이보다 짧은 응답 자료는 0으로 패딩처리하였다. 모델 개발에 사용될 임베딩 층을 구성하기 위하여 텍스트 데이터 수치화 작업을 진행하였다. 텍스트 데이터를 수치화 벡터로 표현하는 방법은 희소 표현(Sparse Representation)과 밀집 표현(Dense Representation)이 있다. 이 중 희소 표현은 단어의 개수가 늘어나면 벡터의 차원이 매우 커진다. 예를 들어, 코퍼스에 단어가 1,000개가 있다면 벡터의 차원은 1,000개여야 한다. 그 중, 표현하고자 하는 단어의 인덱스(index)에 해당하는 부분만 1로 표시가 되고 나머지는 0의 값을 가진다. 이에 대한 예로 Melon이라는 단어의 인덱스가 1이면 희소 표현은 아래의 [그림 4]와 같다.



[그림 4] 희소 표현 예시

이렇게 차원이 높아진 벡터는 공간적 낭비를 불러일으키며 단어의 의미를 담지 못한다. 이러한 희소 표현의 단점을 보완한 표현 방법이 밀집 표현이다. 밀집 표현은 단어 집합의 총 크기로 표현되는 것이 아니라 연구자가 설정한 값으로 표현되며 벡터의 표현이 0과 1이 아닌 실수로 이루어진다. 만약, 연구자가 밀집 표현 차원을 30으로 설정한다면, 모든 단어의 벡터 표현은 차원이 30으로 변경되며 모든 값이 실수로 구성된다. 이에 대한 예는 아래의 [그림 5]와 같다.



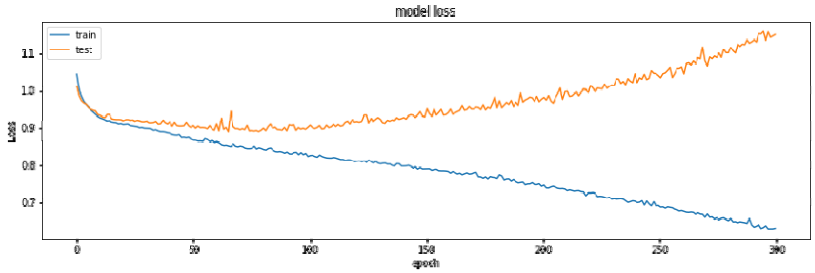
[그림 5] 밀집 표현 예시

본 연구에서는 텍스트 데이터를 수치화 벡터로 표현하는 방법은 밀집 표현(Dense Representation)을 사용하였으며, 다음으로 훈련용 데이터를 LSTM 모델에 적용하여 기계학습을 실시하였다. LSTM 모델의 기계학습 절차를 요약하면 다음과 같다.

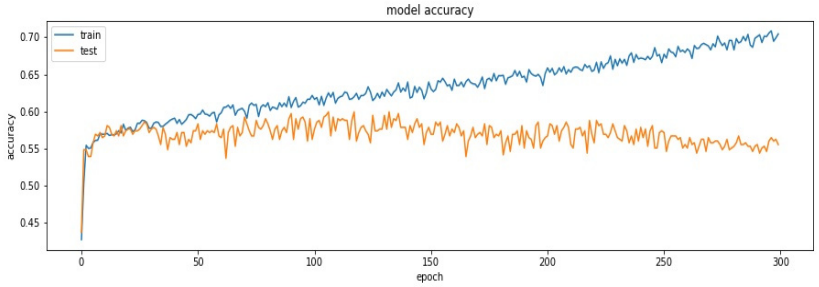
① 먼저 입력층에서 훈련용 데이터가 입력된다. 모형에 입력된 데이터는 임베딩 층에서 사전에 구축한 임베딩 값으로 교체된다.
② 이후 교체된 값은 은닉층인 LSTM 층을 거치게 된다.
③ 마지막인 출력층에서는 3개의 노드와 소프트맥스(softmax) 함수를 활성화 함수로 사용하여 부분 점수를 예측하였으며, LSTM 모델에서 사용한 손실함수는 크로스엔트로피(sparse_categorical_crossentropy)이다.

본 연구에서는 모형의 훈련과정과 손실함수의 최소화 과정 및 과대적합(Overfitting), 과소적합(Underfitting)을 파악하기 위하여 matplotlib.pyplot 모듈을 활용하여 [그림 6] 및 [그림 7]과 같이 시각화하였다. [그림 6]에서 보여주듯이 에포크(epoch, 학습 횟수) 값이 증가됨에 따라 훈련용 데이터의 손실함수 값은 줄어들지만, 평가용 데이터의 손실함수 값은 증가함을 알 수 있으며, 따라서 모형의 과대적합 문제가 발생할 수 있음을 알 수 있다. 더불어 [그림 7]에서 보여주듯이 모형의 정확성에서도 에포크 값이 증대됨에 따라 평가용 데이터에서 정확성은 큰

폭으로 높아지지만 어느 수준부터는 큰 변화가 없음을 알 수 있다. 따라서 모형의 과대적합 및 과소적합의 문제와 정확성을 고려하여 에포크 값을 설정할 필요가 있다.

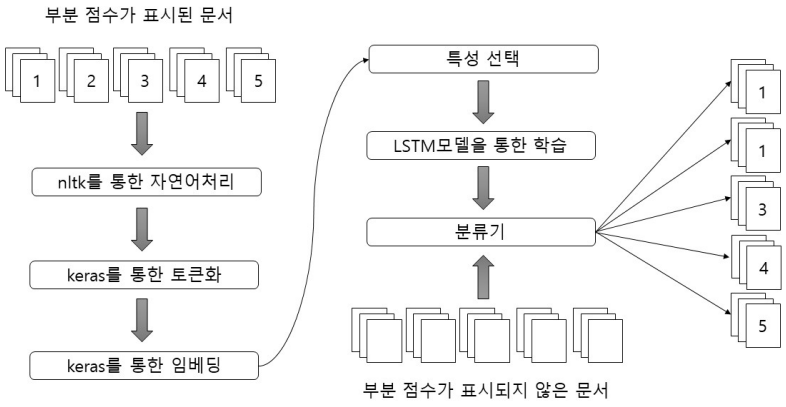


[그림 6] 모형 훈련과정과 손실함수의 최소화 과정



[그림 7] 모형 훈련과정과 모형 정확성 최대화 과정

마지막으로 점수가 표시되지 않은 평가 데이터를 기계학습이 완료된 모형에 적용하여 각 부분 점수를 예측하였으며 이상의 절차를 시각적으로 나타내면 [그림 8]과 같다.



[그림 8] 분석 절차

본 연구에서는 학습된 LSTM 모델의 성능을 평가하기 위해 인간 채점자가 실제로 부여한 점수와 모델이 예측한 점수 간 혼동행렬(confusion matrix)을 만든 후 정확도(accuracy), 정밀도(precision), 재현율(recall), F1(F1-measure), 카파(kappa) 및 상관계수를 산출하였다. 정확도는 전체 예측 중 실제와 같게 예측한 비율이며, 정밀도는 모형이 예측한 값 중 실제로 맞힌 비율, 재현율은 인간 채점자가 정답이라고 판단한 응답 중에서 모델이 실제로 예측한 비율, F1은 정밀성과 재현율의 가중치를 동일한 상태로 고정시킨 후 조화 평균을 의미한다. 카파는 인간채점자의 점수와 LSTM이 예측한 점수 간 일치도를 확인하기 위한 지수이다.

본 연구에서는 LSTM 모델의 성능을 평가하기 위해 1,918개의 데이터로 학습한 후 480개의 평가 데이터의 점수를 예측하여 그 결과를 정확도, 정밀도, 재현율, F1 카파 지수, 상관계수를 기준으로 살펴보았으며, 그 결과는 〈표 3〉과 같다.

〈표 3〉 서·논술형 답안 자동 채점기 모형의 성능 평가

구분	정확성	정밀도	재현율	F1	카파	상관계수
값	0.61	0.60	0.61	0.60	0.49	0.50

〈표 3〉에서 보여주듯이 모델이 예측한 점수와 인간 채점자가 채점한 결과를 비교하였을 때 정확도는 0.61, 정밀도는 0.6, 재현율은 0.61, F1은 0.60, 카파는 0.49, 상관계수는 0.50으로 나타났다. Landis와 Koch(1977)의 기준에 따르면 카파 계수가 0.4 ~ 0.6의 범위를 가질 경우 적정한(moderate)수준을 판단할 수 있으며 본 연구에서 학습한 LSTM의 성능 평가 결과 카파 계수가 0.49로 양호한 성능을 보이는 것으로 판단된다. 한편, 이상의 성능 평가 결과가 RNN 모델을 이용한 자동 채점을 구현한 박세진과 하민수(2020)의 결과와 비교하였을 때 다소 실망스러운 결과로 보일 수 있으나, 박세진과 하민수(2020)가 과학 교과의 구조화된 서답형 답안에 대해 정답과 오답의 이항 분류로만 예측하여 자동 채점을 구현한 반면 본 연구에서는 완전한 에세이 답안에 대해 0점, 1점, 2점의 다항 분류로 예측하였다는 점에서 두 연구 간의 성능 차이는 어느 정도 예측 가능한 결과라 할 수 있다. 또한 영어 에세이에 대한 자동채점 연구인 시지자 외(2012)에서 제시하는 평가지표와도(상관계수가 0.8이상, 유사일치도 0.9이상) 상당한 차이를 보이고 있으나, 시기자 외(2012)에서의 평가지표는 자동채점기의 채점 결과와 인간채점자의 채점 결과 간의 상관계수 및 유사일치도를 제시한 것으로 본 연구에서의 평가 지표와 직접적인 비교는 불가능하다. 다만, 이 연구에서의 쓰기 시험은 자동채점을 고려하여 검사 설계 및 문항 개발이 이루어졌다는 점과 국가영어능력평가시험을 위한 오류 분석 및 어휘 분석 프로그램을 자체적으로 개발하여 적용하였다는 점에서 성능에 차이를 이해할 수 있을 것이다.

V. 결론 및 논의

본 연구에서는 영어 에세이 데이터를 이용하여 순환신경망의 일종인 LSTM을 기반으로 자동 채점 가능성을 탐색해 보았다. 자동 채점 연구는 한국교육과정평가원을 중심으로 일련의 연구들(진경애 외, 2011 〉 시기자 외 2012〉 이용상 외 2013)이 수행된 바 있으며, 최근에는 창의재단을 통해서도 일부 자동 채점 연구가 수행되고 있다. 초기 자동 채점 연구는 영어 쓰기 자동 채점 시스템을 구축하기 위한 목적으로 수행되었으나 정책의 변화에 따라 지속적인 연구가 수행되지 못하였기 때문에 인공신경망과 같은 비교적 최신의 학습모델을 적용한 연구가 수행되지 못했다. 더불어 한국어 서답형 문항 자동 채점 연구들(노은희, 2012, 2013, 2014)도 수행되었으나, 기계학습을 통한 인공지능 기반의 자동 채점이 이루어지지 못하였고, 에세이가 아닌 문장 단위의 자동 채점만을 구현했다는 한계점을 가진다. 한편, 최근 일부 연구자들을 중심으로 인공신경망 중 하나인 RNN을 기반한 자동 채점 연구가 수행되고 있으나, LSTM이나 GRU(Gated Recurrent Unit)와 같은 다양한 인공신경망을 적용한 자동 채점 연구는 아직 찾아볼 수 없었으며, 기존의 RNN을 적용한 자동 채점 연구(박세진, 하민수, 2020)도 정답과 오답으로만 구분하는 이항 분류 방식으로만 수행되었다는 점에서 한계가 있다.

이에 본 연구에서는 기존의 RNN이 가지는 장기의존성의 문제를 극복하는 LSTM 모델을 이용한 기계학습과 예측을 통해 LSTM 기반 자동 채점의 가능성을 탐색해보았다. 일반적으로 에세이 채점은 채점 기준에 의거하여 부분점수를 허용하는 방식(예, 0점, 1점, 2점)으로 채점이 되므로 에세이 채점 데이터는 다항분포를 가지는 특성을 가진다. 따라서 RNN이나 LSTM과 같은 인공신경망을 기반으로 하는 자동 채점도 이러한 채점 데이터의 특성을 반영하여 다항 분류의 방식으로 점수 예측이 이루어질 필요가 있다. 본 연구에서는 이러한 필요성을 감안하여 LSTM을 기반한 다항 분류의 방식으로 점수를 예측하였다. LSTM 학습 모델의 성능을 평가하기 위해 평가 데이터를 이용하여 점수를 예측하고 이를 인간 채점자의 점수와 비교하여 정확도(0.61), 정밀도(0.60), 재현율(0.61), F1(0.60), 카파(0.49), 상관계수(0.50) 등을 살펴보았으며, 그 결과 본 연구에서 사용한 데이터에서 LSTM 모델이 양호한 수준(moderate)의 성능을 보임을 확인하였다.

본 연구에서는 총 2,398명의 데이터를 훈련용 데이터와 평가용 데이터로 나누어 학습 및 검증을 실시하였으며, LSTM 모델의 학습을 위해 사용한 데이터는 전체 데이터의 80%인 1,918개이다. 에세이에 대한 기계학습을 위한 충분한 데이터의 양에 대한 절대적인 규칙은 아직까지 없으며, 에세이의 종류와 답안의 복잡성, 학습 모델 등에 따라 충분한 기계학습을 위해 필요한 데이터의 양이 달라질 것이다. 그러나 본 연구에서 기계학습을 위해 사용된 1,918개의 데이터는 분명 기계학습을 위해 충분한 데이터의 양이라 할 수는 없으며, 특히 점수대(0점, 1점, 2점)에 따른 데이터의 분포가 고르지 않아 점수대별 예측에도 제한점이 있었다. 따라서 후속 연구에서는 데이터의 양이 충분하면서 점수대별 데이터의 분포가 고른 데이터를 확보하여 LSTM 모델을 이용한 자동 채점 가능성에 대한 심층연구를 수행할 필요가 있다. 더불어, 국내 서답형 검사에 적용될 수 있는 자동채점 시스템 구축을 위한 준비단계로서 한국어 서답형 답안을 활용한 실증 연구가 수행될 필요가 있으며, 이러한 실증 연구를 통해서 영어 답안에서 확인된 RNN과 LSTM의 성능 차이를 다시 한번 검증해 볼 필요가 있다.

참고문헌

김태준, 김정아, 임종헌, 이정우, 윤현희, 장근영. (2020). 혁신적 포용국가 실현 방안: 교육분야를 중심으로. 진천: 한국교육개발원.

교육부. (2021.4.20.). 국민과 함께하는 미래 교육과정 논의 본격 착수-「2022 개정 교육과정 추진계획」 발표. https://moe.go.kr/boardCnts/viewRenew.do?boardID=294&boardSeq=84176&lev=0&searchType=null&statusYN=W&page=1&s=moe&m=020402&opType=N (검색일: 2021. 05.01.)

관계부처합동. (2020). 인공지능시대 교육정책방향과 핵심과제. https://moe.go.kr/boardCnts/viewRenew.do?boardID=294&boardSeq=82674&lev=0&searchType=null&statusYN=W&page=1&s=moe&m=020402&opType=N(검색일: 2021. 05.01.)

노은희, 심재호, 김명화, 김재훈. (2012). **대규모 평가를 위한 서답형 문항 자동채점 방안 연구**(RRE 2012-6). 서울: 한국교육과정평가원.

노은희, 김명화, 성경희, 김학수. (2013). **대규모 평가를 위한 서답형 문항 자동채점 프로그램 정교화 및 시범 적용**(RRE 2013-5). 서울: 한국교육과정평가원.

노은희, 이상하, 임은영, 성경희, 박소영. (2014). **한국어 서답형 문항 자동채점 프로그램 개발 및 실용성 검증(RRE 2014-6)**. 서울: 한국교육과정평가원.

노은희, 송미영, 성경희, 박소영. (2015). **한국어 문장 수준 서답형 문항 자동채점 프로그램 개발 및 적용**(RRE 2015-9). 서울: 한국교육과정평가원.

박세진, 하민수. (2020). 순환신경망을 적용한 초등학교 5학년 과학 서술형 평가 자동 채점시스템 개발 및 활용 방안 모색. **교육평가연구, 33**(2), 297-321.

박혜영, 김성숙, 김경희, 이명진, 김광규, 김지영. (2019). **수업-평가 연계 강화를 통한 서·논술형 평가 내실화 방안안**(RRE 2019-6). 진천: 한국교육과정평가원.

방성혁, 배석현, 박현규, 전명중, 김제민, 박영택. (2018). 순환신경망 기반의 사용자 의도 예측 모델. **정보과학회논문지, 45**(4), 360-369.

시기자,박도영,이용상,박상욱,임은영,구슬기,임황규,최연희,이공주,김지은,김성,이은숙,김성묵,윤경아,이순웅. (2012). **국가영어능력평가시험 쓰기 자동 채점 프로그램 개발**(RRE2012-10). 서울: 한국교육과정평가원.

신동광, 박용효, 박태준, 임수연. (2015). 영어 말하기 자동채점의 현재와 미래. **멀티미디어언어교육, 18**(1), 93-114.

이용상, 시기자, 박도영, 윤경아, 구슬기, 임황규. (2013). 쓰기 자동 채점 알고리듬의 성능 비교. **교육과정평가연구, 16**(3), 147-165.

주일택, 최승호. (2018). 양방향 LSTM 순환신경망 기반 주가예측모델. **한국정보전자통신기술학회 논문지, 11**(2), 204-208.

진경애, 남명호, 김명화, 오상철, 김민정, 주형미, 신효필, 반재천, 김수경. (2006). **서답형 문항 자동채점 프로그램 도입 방안 연구(Ⅰ)**(RRE 2006-6). 서울: 한국교육과정평가원.

진경애, 이병천, 주형미, 신동광, 박정, 김지은, 이공주, 이은성. (2007). **서답형 문항 자동채점 프로그램 도입 방안 연구(Ⅱ) : 영작문 채점을 중심으로.**(RRE 2007-4). 서울: 한국교육과정평가원.

진경애, 이병천, 신동광, 박태준, 주헌우. (2008). **서답형 문항 자동채점 프로그램 도입 방안 연구(Ⅲ)**(RRE 2008-6). 서울: 한국교육과정평가원.

진경애, 시기자, 신동광, 송민영, 김인숙, 이용상, 김연희. (2011). **KICE-Pearson 영어 말하기, 쓰기 자동채점 프로그램 타당성 연구.** 서울: 한국교육과정평가원.

하민수, 이경건, 신세인, 이준기, 최성철, 주재걸, 김남형, 이현주, 이종호, 이주림, 조용장, 강경필, 박지선. (2019). 학습 지원 도구로서의 서술형 평가 그리고 인공지능의 활용: WA3I 프로젝트 사례. **현장과학교육, 13**(3), 271-282.

Attali, Y.,& Burstein, J. (2006). Automated essay scoring with e-rater V.2.0. Journal of Technology, *Learning and Assessment, 4*(3), 1-21.

Hochreiter, S., & Schmidhuber, J. (1997) Long short-term memory. *Neural computation, 9*(8), 1735-1780.

Kumar, V.,& Boulanger, D. (2020). Explainable automated essay scoring: Deep Learning really has pedagogical value. *Frontiers in Education, 5*, 1-22.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389-405. 1-22.

McCulloch, W. S., Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics, 5*(4), 115-133.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). *Recurrent neural network based language model.* In Eleventh annual conference of the international speech communication association.

Page, E. B. (1966): The imminence of grading essays by computer, *Phi Delta Kappan, 47*(5), 238-243.

Shewalkar, A., Nyavanandi, D., & Ludwig, S. (2019). Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research. 9*. 235-245.

· 논문접수 : 2021.10.05. / 수정본접수 : 2021.10.29. / 게재승인 : 2021.11.10.

ABSTRACT

Exploring the Feasibility of an Automated Essay Scoring Model Based on LSTM

Kangyun Park
Researcher, National Information Society Agency
Yongsang Lee
Assistant Professor, Inha University
Dongkwang Shin
Associate Professor, Gwangju National University of Education

In the present study, the feasibility of an automated essay scoring of English was explored using Long-Short Term Memory (LSTM), a type of Recurrent Neural Network (RNN). LSTM is a deep learning model proposed to overcome the problem of long-term dependence of the existing RNN. In this study, an automated essay scoring model based on LSTM was adopted to score English essay data extracted from the open huge repository of data 'kaggle,' and the performance of the model was validated. Unlike multiple-choice scoring data which consisted of binary (true/false) data, essay scoring data had multiple facets, thus the data used for the deep learning model was constructed within a multinomial classification to order to predict scores of those essay data. For its validation, the six indices of 'accuracy,' 'precision,' 'recall', 'F1-measure,' 'kappa,' and 'correlation coefficient' were used. As a result, it turned out that the LSTM model could predict students' essay scores at an appropriate level. The performance of the deep learning model is closely related to the quality and quantity of data, thus it is expected that the accuracy of the automated essay scoring could be improved if sufficient quality data is composed and used for the deep learning process. To derive a more valid and reliable algorithm, it is necessary to conduct further empirical studies by testing various RNN models.

Key Words: *RNN, Automated essay scoring model, LSTM*