Brain, Digital, & Learning

pISSN : 2384-2474
eISSN : 2586-7490

RESEARCH ARTICLE

A Study on the Development of Automated Korean Essay Scoring Model Using Random Forest Algorithm

Jong-Im Park[1*] · Sook-ki Choi[2]
[1]Korea Institute for Curriculum and Evaluation
[2]Korea National University of Education

랜덤포레스트 알고리듬을 활용한 한국어 에세이 자동채점 모델 개발 연구

박종임[1*] · 최숙기[2]
[1]한국교육과정평가원 · [2]한국교원대학교

*Corresponding Author: pji0310@kice.ac.kr

ABSTRACT

In this study, Random Forest algorithm was applied using data of 402 Korean argumentative essays to develop an automatic scoring model for Korean essays. Recent studies on automatic scoring of essays are developing into studies using deep learning-based algorithms. However when automatic scoring is used in the field of education, deep learning-based algorithms have limitations because the interpretation and explanation of the scoring results must be possible. Therefore, in this study, we tried to develop a model that can interpret the scoring results by applying a machine learning-based algorithm that utilizes scoring features. In the essay, various parts of morpheme-based scoring qualities were derived, and the number of sentences was used. As a result, it was possible to derive a significant level of model performance. However, this study has a limitation in that it utilized basic quantitative feature such as parts of morpheme frequency and number of sentences. In future research, more in-depth grading feature will be developed to explore ways to provide more meaningful educational feedback through automated essay scoring.

Key words: Automated essay scoring, scoring feature, Korean essay scoring, random forest algorithm, machine learning

Brain, Digital, & Learning
2023, Vol. 13, No. 2, 131-146.
https://doi.org/10.31216/BDL.20230008

Received: May 03, 2023
Revised: June 14, 2023
Accepted: June 14, 2023

Introduction

미래의 국어교육은 단순히 교과의 지식 및 개념을 암기하고 회상하는 것을 넘어서 다양한 자료를 비교, 분석, 활용하여 문제를 해결하고 이를 자신의 언어로 직접 구성하여 증명할 수 있는 에세이형 평가로 개선될 전망이다. 선택형 및 단답형 중심의 현행 평가 체제에 익숙한 학생들이 에세이를 작성하는 능력을 갖추고 에세이형 평가에 적응하기 위해서는 학생들의 학습 과정에서 다양한 연습과 피드백이 이루어져야 한다. 그러나 에세이형 평가는 채점 과정에서 많은 시간과 노력이 투입되어야 하므로 즉각적인 진단과 피드백이 어렵다. 또한 다수 응답을 채점하는 과정에서 채점자의 일관성을 확보하는 것이 쉽지 않다. 이와 더불어서 최근 교육부는 국정과제 '82.모두를 인재로 양성하는 학습혁명'에 따라 '학생 개별 맞춤형 교육'을 제안, 핵심정책으로 '디지털기반 교육혁신'을 제시하였다. 이 정책은 '25년부터 인공지능 기반 코스웨어를 운영, 학생 개별 맞춤형 학습을 지원하고 인공지능 기술을 활용하여 수업과 평가를 혁신하고자 하는 정책이다. 이처럼 AI 기반의 맞춤형 학습 구현 시, 선택형, 단답형 문항 위주의 평가를 개선하여 서술형, 에세이형의 평가 도입을 위해서는 한국어에 대해 AI 기반의 자동채점 및 피드백 기술이 필수적으로 요구될 전망이다.

우리나라의 교육 환경을 고려할 때 향후에는 에세이형 평가와 한국어 자동채점 기술이 매우 중요하게 요구될 것으로 예상되지만 한국어 에세이 자동채점 연구 성과는 아직 많이 부족한 상태이다. 한국교육과정평가원에서는 지난 2012년부터 최근까지 한국어 자동채점 연구를 수행하고 있다. 한국어 자동채점 시스템을 개발하기 위해서 2010년대에는 단답형 자동채점을 중심으로 연구가 수행되었으나(Noh et al., 2012, 2013, 2014, 2015, 2016), 2010년대에 이루어진 연구는 주로 단답형 답안에 나타나는 키워드를 활용하여 이진 분류 방식으로 채점을 하였기 때문에 에세이와 같은 다문장을 채점하거나 정답 키워드가 없이 채점 기준에 근거하여 채점하는 것에 한계가 있었다. 이에 최근에는 에세이 수준의 자동채점 연구를 진행하고 있다(Park, et al., 2022).

최근에는 인공신경망 기반의 딥러닝 알고리즘이 발달하면서 자동채점에도 이를 적용하는 연구도 수행되고 있다. 대표적으로 Park & Lee (2022)에서는 딥러닝 기반 학습 모델 비교를 통하여 한국어 에세이 문항 자동채점을 위한 최적의 알고리듬을 탐색하였다. 딥러닝 계열의 알고리듬인 순환신경망(RNN, Recurrent Neural Network), 장단기 기억(LSTM, Long- Short-Term-Memory), 게이트 순환 유닛(GRU, Gated- Recurrent-Unit) 알고리듬을 적용하여, 한국어 에세이 답안 채점을 위한 채점 모델을 구축하고 성능을 비교하였다. 이러한 연구들이 등장하면서 한국어 자동채점 연구는 새로운 도약기를 맞이하고 있다. 그러나 최근에 이루어지고 있는 딥러닝 알고리듬 기반의 자동채점은 대량의 학습 데이터가 있다면 머신러닝 방식보다 정교한 자동채점 모델을 개발할 수 있으나, 채점 과정에서 어떠한 변인이 점수에 영향을 주었는지를 파악하는 것이 불가능하다는 제한점을 가지고 있다. 이와 달리, 머신러닝 기반의 알고리즘은 딥러닝에 비해서는 상대적으로 필요로 하는 학습 데이터의 양이 적기 때문에 자동채점 연구 초기 적용에 효율적이고, 딥러닝 기반의 알고리듬에 비해서 채점 결과의 해석 가능성이 존재하지만 머신러닝 기반의 알고리즘을 적용하기 위해서는 변인에 해당하는 채점자질(Scoring feature)을 사전(事前)에 정의하는 단계가 필요하다. 채점자질을 개발하고 선정하는 과정은 매우 많은 시간과 노력을 필요로 한다는 어려움이 있다. 이에 한국어 자동채점 기술 개발 시 채점자질을 탐색하는 기초 연구가 지속적으로 축적될 필요가 있다. 이러한 기초 연구가 축적되어야 자동채점의 상황이나 목적에 따라서 '성능'과 '해석 가능성' 사이에서 다양한 선택이 가능하기 때문이다.

이에 본 연구에서는 에세이에서 점수에 영향을 미치는 변인에 해당하는 채점자질을 추출하여 머신러닝 알고리즘 기반의 자동채점 모델을 개발하고자 한다. 이를 위해 중등학생들이 작성한 약 400편의 글에 대해 자연어처리를 통해 채점자질을 추출하고, 이를 랜덤포레스트(Random Forest) 알고리듬에 적용하여 자동채점

모델을 개발하고자 한다. 본 연구에서 제시하는 자동채점 모델은 아직 시뮬레이션 단계에 불과하고 사용한 채점자질 또한 단순한 빈도 기반의 채점자질을 사용하고 있어서 제한점이 있다. 그러나 본 연구를 통해서 한국어 자동채점 기술 개발의 다양한 방안을 시도하고, 자동채점이 인공지능이나 컴퓨터 공학 관련 전문가의 영역 외에도 에세이 채점자질 선정과 같이 국어교육 전문가들의 연구가 필요함을 제안하는 데에 목적이 있다.

Background

Scoring Feature

채점자질은 에세이에 대한 정보를 포함하는 측정 가능한 속성으로 머신러닝에 투입되는 변수를 의미한다. 문자로 이루어진 글은 비정형 데이터이기 때문에 기계는 문자 그대로의 데이터를 인간처럼 읽거나 이해하지 못한다. 이에 문자로 이루어진 한 편의 에세이를 다양한 방식으로 분석하여 에세이가 가진 특성을 기계가 인식 가능한 형태의 정형 데이터로 변형해야 한다. 컴퓨터는 특정 점수나 수준의 에세이 데이터에서 나타나는 특성 패턴을 찾아서 이 패턴과 점수와의 관련성을 학습하면서 채점 모델을 구축하고, 이 채점 모델을 활용하여 새로운 에세이에 점수를 예측한다(Park et al., 2022). 채점자질은 기계학습 모델의 성능을 향상시키는 데 매우 중요한 역할을 한다.

자동채점 연구 초기부터도 이러한 채점자질이 유용하게 활용되었는데 대표적으로 Page (1966, 1994, 2003)가 개발한 PEG (Project Essay Grade) 또한 채점자질을 활용하여 자동채점을 수행하고 있다. Page (1966)가 개발한 초기 에세이 자동채점 시스템은 사람이 직접 도출한 계량적 채점자질을 회귀 모델에 적용하는 방식이다. 이때 사용한 계량적 채점자질은 주로 단어 수, 문장 수, 평균 단어 길이, 형용사 수 등의 빈도와 길이 자질이다. 이후에 이러한 채점자질이 텍스트의 의미론적 분석 없이 에세이의 표면적인 특징만을 포착했다는 점에서 비판을 받기도 하였다(Chung & O'Neil, 1997). 그럼에도 불구하고 이러한 계량적 채점자질은 간단히 추출되면서도 에세이 자동채점 시스템의 성능을 높일 수 있기 때문에 오늘날까지도 유용하게 사용되고 있다 (Dascalu et al., 2017; Nguyen & Litman, 2018).

PEG 외에도 영어권 자동채점 연구에서 계량적인 채점자질을 활용하는 사례는 빈번하다. Intelligent Essay Assessor (IEA), e-rater, BETSY (Bayesian Essay Test Score System), IntelliMetric 등 현재 사용되는 대부분의 에세이 자동채점 시스템은 기본적으로 자연어처리를 통해 사람이 사전 정의한 채점자질을 사용한다(Park, et al., 2022). 이들 시스템은 채점자질을 활용하여 회귀 모델 또는 분류 모델을 생성한다. 이들 시스템에서 사용된 채점자질로는 단어 길이, 문장 길이 등의 비교적 간단한 채점자질에서부터 이독성, 문법 오류 등의 좀 더 복잡한 채점자질에 이르기까지 다양하다(Park, et al., 2022).

이들 채점자질 외에도 영어권에서는 미국 멤피스대학교에서 개발한 웹기반 프로그램인 Coh-Metrix를 활용하여 채점자질을 추출할 수 있다. Coh-Metrix는 이독성 지수, 응집성, 어휘 다양성 등의 11개 범주에 대해서 총 106개의 텍스트 언어자질을 산출해 준다. 따라서 영어로 작성한 에세이에 대해서는 이러한 프로그램을 활용하여 채점자질을 손쉽게 도출할 수 있다. 그러나 한국어의 경우 아직은 자동적인 텍스트 분석 도구가 없다. Coh-Metrix처럼 전체적인 요소에 대해 텍스트 자질을 분석하는 도구는 없지만 유사한 도구로 최근에 한국교육과정평가원에서 개발 중인 텍스트 이독성 지수 자동측정 프로그램을 들 수 있다. 이독성 지수 자동측정 프로그램은 국어과 텍스트, 사회과 텍스트, 과학과 텍스트, 교과 통합 텍스트에 대해서 이독성에 영향을 미치는 양적 요인으로 글자 수, 문장 평균 길이, 쉬운 어휘 비율 등의 요인을 제시하였고, 연속되는 올해 연구

에서는 이러한 양적 요인을 활용하여 이독성 지수를 자동으로 측정하는 프로그램을 개발하고 있다. 이러한 이독성 지수 또한 글쓰기 자동채점에 활용할 수 있는 유용한 채점자질이 될 수 있다.

이 외에도 최근에 Nam & Won (2022)에서는 한국어 쓰기 자동채점을 위한 언어자질을 탐색하였는데 이 연구에서는 어휘 다양도, 어휘 밀도, 어휘 세련도와 같은 채점자질을 측정하는 공식을 산출하고, 글의 길이(총 어휘 수, 총 어절 수)와 문장 수, 문장 당 어절 수, 어휘 다양도, 어휘 밀도, 어휘 세련도 자질을 활용하여 상관분석과 회귀분석을 통해 글의 점수에 영향을 미치는 요인을 탐색하였다. 그 결과 '총 어휘 수', '어절 수', '문장 수'의 순서로 글의 점수에 영향을 미치는 것으로 분석되었다.

Machine Learning Based RF algorithm

머신러닝은(Machine Learning)은 데이터를 기반으로 규칙과 패턴을 발견하여 모델이나 알고리듬을 생성하고 새로운 데이터에 대한 예측을 하는 것을 의미한다(Lantz, 2013: Murphy, 2012). 이 단계는 정형화한 에세이 데이터에서 채점자질 정보와 점수(또는 분류 결과) 정보를 활용하여 기계가 이 둘 간의 관계를 학습하면서 예측 모델을 도출하는 단계이다. 전처리한 에세이 데이터를 학습용(train) 데이터와 검증용(test) 데이터로 분할하고, 학습용 데이터를 활용하여 머신러닝을 진행한 다음 검증용 데이터로 모델의 성능을 분석하는 것이 일반적인 방법이다.

인간이 사전 정의한 채점자질을 활용하는 연구는 딥러닝 기반의 알고리듬이 등장하면서 다소 과거의 연구 방식으로 인식되기도 하였다. 그러나 딥러닝 기반의 알고리듬은 점수 예측 결과를 해석하기 어렵기 때문에 교육 분야에서 활용하는 것에 한계가 있다. 의료와 같이 '결과에 대한 해석'이 중요한 영역에서는 여전히 머신러닝 기반의 알고리듬이 주요하게 활용되고 있다. 이에 Park et al.(2022)에서는 교육 목적의 한국어 자동채점 기술 개발을 위해 채점자질을 활용하는 머신러닝 기반의 접근이 우선되어야 한다고 제안한 바 있다.

최근에는 딥러닝 기반의 알고리듬을 사용하더라도 LIME, SHAP와 같은 설명가능한 인공지능 XAI(explainable artificial intelligence)(Adadi & Berrada, 2018; Murdoch et al., 2019)의 다양한 알고리듬이 제안되고 있다(Ribeiro et al., 2016; Lundberg & Lee, 2017). 이에 최근에는 딥러닝 기반의 알고리듬에서도 채점자질을 추출해서 활용하는 채점자질 기반 다층 퍼셉트론(feature-based multi-layer perceptron, MLP) 심층 신경망을 활용하는 연구도 제안되고 있다(Kumar & Boulanger, 2020).

머신러닝 계열의 자동채점 알고리듬에는 나이브 베이즈 분류(Naive Bayesian Classification), 서포트 벡터 머신(Support Vector Machine), 랜덤포레스트(Random Forest) 등 다양한 알고리듬이 존재한다. 본 연구에서는 이 중 랜덤포레스트 알고리듬을 적용하고자 한다. 랜덤포레스트는 의사결정나무(Decision Tree)를 개별 모형으로 사용하여 분류나 회귀 분석 등에 사용되는 앙상블(ensemble) 학습 방법의 일종이다. 앙상블 학습 방법이란 다수의 예측기를 동시에 훈련시켜서 하나의 예측 모델을 사용했을 때보다 좋은 예측 성능을 얻고자 하는 기법이다. 랜덤포레스트 또한 다수의 의사결정나무를 훈련하여 예측 모델을 얻는 방식이다.

랜덤포레스트는 여러 개의 의사결정나무를 만들고, 각각의 나무에서 도출된 결과를 모아서 최종적인 예측 결과를 만들어 낸다. 이때 무작위로 선택된 일부 자질을 사용하여 노드를 분할하는데, 이러한 과정을 랜덤화(randomization)라고 한다. 랜덤한 변수를 사용하기 때문에 특정 변수에 지나치게 의존하지 않고 다양한 변수를 활용하여 예측할 수 있고, 여러 개의 의사결정트리 간의 상관성을 제거할 수 있다. 이러한 특성들로 인해 랜덤포레스트는 매우 안정적이면서도 강력한 분류 및 회귀 모델로 사용된다.

랜덤포레스트 알고리듬의 또 다른 특징 중 하나는 분류 혹은 회귀 문제에서 각 변수, 즉 채점자질들의 중요성에 상대적 순위를 매길 수 있다는 점이다. 분류 모델에서는 지니계수(Gini Index) 또는 엔트로피(cross-entropy)를 이용하고, 회귀 모델에서는 분산을 이용하여 각 노드를 분할하는 변수를 선택한다. 예를 들어 분류 모델에서는 지니

계수를 최소화하는 방향으로 노드를 분할하는데 각 분기점에서 지니계수가 감소하는 폭이 클수록 해당 변수의 중요도가 높다고 판단하고 이를 활용하여 채점자질의 중요도 지수(Importance Index)를 산출할 수 있다.

최근에는 국내 연구에서도 한국어나 영어를 대상으로 랜덤포레스트 알고리듬을 적용하는 연구들이 이루어지고 있다. 랜덤포레스트를 한국어 자동채점 영역에 활용한 연구로는 Ha et al. (2019)가 있다. 이 연구에서는 과학과 서술형 평가 자동채점을 위해서 랜덤포레스트 알고리듬을 활용하여 WA3I라는 서술형 자동채점 프로그램을 개발하였다. 국내 자동채점 연구의 초기에 해당하는 연구로서 후속 연구에 많은 영향을 미친 연구이지만 정답과 오답을 분류하는 이진분류 방식이라서 주로 정답이 있는 서술형 답안에 활용할 수 있다는 제한점이 있다. 최근에는 에세이 수준의 응답에 대한 자동채점 연구들도 이루어지고 있는데 영어 자동채점에 랜덤포레스트 알고리듬을 적용한 연구로는 Lee et al. (2022)가 있다. 이 연구에서는 세종학당재단의 한국어능력시험으로 사용되고 있는 SKA 모의 평가에서 수집한 343편의 에세이 답안에 대해서 20개의 채점자질을 도출하여 채점 모델을 개발하였다. 또한 Park & Lee (2022)에서도 인공지능 학습 데이터가 충분하지 않은 상황에서는 KoBERT를 기반으로 하는 딥러닝 채점 모델보다는 채점 자질을 활용하는 랜덤포레스트 알고리듬이 보다 높은 성능을 보인다고 보고하였고, 채점자질을 활용하는 모델이 형성 평가의 기능까지 확장 가능성이 있다고 제안하였다. 이어서 Shin (2022) 또한 채점자질을 활용하는 랜덤포레스트 기반의 알고리듬을 적용하는 것이 딥러닝 계열의 모델보다 높은 성능을 보이는 것으로 보고하였다.

Method

Essay Database

본 연구에서는 Choi (2018)에서 수집한 약 400여편의 글을 자동채점 모델 개발 및 검증을 위한 데이터로 활용하였다. Choi (2018)에서 수집한 글쓰기 과제는 학생들에게 '소년법에서 규정하고 있는 촉법 소년의 연령 기준을 만 14세 미만보다 낮추어야 한다'는 논제에 대하여 찬성과 반대의 입장 중 하나를 선정하고, 주어진 2편의 자료 글과 자신의 생각을 바탕으로 하여 논증적 글을 작성하도록 하는 과제이다. 학생들에게 제공된 자료 글은 ①소년법과 촉법 소년의 개념 정의를 포함한 기사문, ②촉법 소년의 범죄 현황 자료와 촉법 소년의 연령 조정 필요성을 제안한 논문, ③촉법 소년의 연령 조정이 불필요함을 제안한 논문이다.

Table 1. Writing Task

Task Prompt
[a]~[c] Using the data, write an article arguing for the thesis that 'the age at which criminal punishment is avoided among minor offenders should be lower than 14 years of age'. Please consider the following <Conditions> when writing.
<Condition> ① Choose one of the positions in favor of or against the topic and write an article. ② Write using the contents of [A]~[C] and correctly quote it. ③ When quoting, when writing the same content from [A]-[C], mark it with double quotation marks (" "), and bring and write the content, but do not distort the meaning of the original text when replacing it with your own words and sentences.
※ example Main article: Dr. Jeong Geun-mo said that there is no problem with the safety of Korean nuclear power plants. · When used the same way: Dr. Jeong Geun-mo's argument is that "there is no problem with the safety of Korean-style nuclear power plants." · When used interchangeably: According to Dr. Jeong Geun-mo, Korea's nuclear power plants are stable.
④ When citing, be sure to indicate the source.

데이터 수집 규모는 Table 2와 같다. 중1은 88명, 중2는 101명, 중3은 95명, 고1은 118명으로 학년별 분포를 유사한 수준에서 유지하고자 하였고, 남학생 198명, 여학생 204명으로 성별 비율 역시 유사한 수준을 유지하였다.

Table 2. Essay writer information

Grade	Male	Female	Sum
7	45	43	88
8	49	52	101
9	47	48	95
10	57	61	118
Sum	198	204	402

수집한 글은 아래 Table 3과 같은 평가 기준표에 의하여 채점이 이루어졌다. 각 평가 영역별로 하위 평가 요소를 설정하였고, 각 평가 요소별로 구체적인 평가 기준이 제시되고 있다. 22개의 평가 기준에 대해서 1점과 0점으로 채점이 되어 평가 요소별로 3점~4점 만점이 되도록 구성한 기준표이다.

Table 3. Essay scoring rubric

Scoring category	criteria
Analysis and use of argumentation material	1. The presented data is not distorted and used based on accurate understanding.
	2. Selected the data suitable for the author's argument and evidence.
	3. For rebuttal, even contradictory data were used effectively.
	4. The presented data is cited, but the discussion is developed by interpreting and expanding it.
The validity of the construction of the argument	5. Logical reasons are provided to support the claim.
	6. Various evidence is provided to support the claim.
	7. Specific evidence is provided to support the claim.
	8. Strengthening argument by using the refutation of the conflicting argument.
Coherence of writing	9. Argument is clearly presented.
	10. The author's arguments and details are presented in a coherent manner.
	11. The writing has unity of content without disparate content.
Cohesion of writing	12. Using the structure of 'Introduction-Body-Conclusion', it helps the expected reader's understanding.
	13. Paragraphs are divided appropriately to help the expected reader understand.
	14. The writings are organically connected with a development structure rather than a simple arrangement.
	15. Each sentence is naturally connected using conjunctions or directives.
Expressive fluency and diversity	16. The tone or style that ussed are effective in persuading the prospective reader.
	17. Using accurate vocabulary to express writer’s thoughts.
	18. Expressed concisely without redundant expressions.
	19. The length of the sentence is not long, and it is expressed in a simple and easy-to-understand manner.
Grammar and citation accuracy	20. Spelling and spacing were used correctly.
	21. Correspondence between subject and predicate or closing expression is expressed in an accurate sentence.
	22. The source of the material used is indicated and accurately quoted.

자동채점 모델 생성 시, 에세이 데이터의 특성은 모델의 성능에 중요한 영향을 미친다. 머신러닝을 할 때, 특정 점수대의 데이터가 충분하지 않은 경우, 머신러닝과 검증이 충분히 이루어지기 어렵기 때문이다. 본 연구에 활용한 에세이 데이터는 22점 총점을 기준으로 다음 Table 4와 같은 점수 분포를 보이고 있다.

Table 4. Score distribution

Grade	N	%
.00	8	2.0
1.00	1	.2
3.00	3	.7
4.00	7	1.7
5.00	3	.7
6.00	9	2.2
7.00	4	1.0
8.00	9	2.2
9.00	8	2.0
10.00	17	4.2
11.00	15	3.7
12.00	13	3.2
13.00	20	5.0
14.00	29	7.2
15.00	36	9.0
16.00	48	11.9
17.00	44	10.9
18.00	50	12.4
19.00	19	4.7
20.00	15	3.7
21.00	44	10.9
Sum	402	100.0

Table 4에서 정리하였듯이 22점 만점에서 만점에 해당하는 에세이는 없었고, 각 점수대별로 빈도가 매우 적은 경우가 있었다. 물론 실제 학생들이 작성한 에세이를 수집한 상황에서는 모든 점수대의 에세이가 고르게 분포하기는 어렵다. 그러나 기계가 특정 점수대의 에세이에 대해서는 머신러닝을 진행하기 어렵기 때문에 실제 채점 상황에서도 이러한 점수대에 대해서는 예측 성능이 떨어질 수 있다. 그러므로 머신러닝을 위해 데이터베이스를 구축할 때에는 실제 학생들로부터 데이터를 수집하는 것 외에도 특정 점수대의 에세이를 수집하기 위해서 별도의 계획을 수립할 필요가 있다. 예를 들어, 낮은 점수대나 등급의 에세이를 수집하기 위해서 교사나 대학생이 일부러 낮은 수준으로 작성한 글을 추가로 수집할 수 있다. 이렇듯 추후 연구에서는 머신러닝용 데이터로서의 필요 조건을 분석하여 데이터를 수정하거나 보완하여 재구축하는 연구가 필요하다. 또한 0점에서 21점까지를 1점 단위로 분류하기에는 한계가 있다고 판단하여 이들 점수대를 통합할 필요가 있다고 판단하였다. 추가로, 1점에 해당하는 데이터 빈도가 1건이기 때문에 이 데이터는 검증용 데이터가 존재할 수 없다. 이 경우 머신러닝 알고리듬을 적용할 때 오류가 발생하기 때문에 1점에 해당하는 데이터를 삭제하고 총 401개 데이터에 대해서 분석을 진행하였다.

Park & Lee (2022)에서는 500편의 한국어 에세이에 대해 자동채점 모델을 개발하는 과정에서 본 연구와 마찬가지로 특정 클래스의 빈도가 적은 현상을 해결하기 위해 1점과 5점에 해당하는 데이터를 제외하고, 2, 3, 4점에 해당하는 데이터만을 학습 및 평가용 데이터로 활용하였다. 이처럼 양극단의 데이터가 부족한 경우 다양한 방법으로 데이터 불균형 문제를 해결할 수 있는데 본 연구에서는 데이터 빈도가 적은 클래스를 삭제하기 보다는 데이터를 최대한 활용하는 방안을 선택하여 각 클래스를 통합하는 접근을 취하였다. 이에 본고에서는 분류 클래스를 줄여서 각 클래스에 해당하는 데이터를 늘려주기 위해서 0~10점을 '1점'으로, 11~17점을 '2점'으로, 18~22점을 '3점'으로 재라벨링을 하였다. 이렇게 재라벨링을 한 결과 1점에 해당하는 데이터는 17%, 2점에 해당하는 데이터는 51%, 3점에 해당하는 데이터는 32%로 분포하였다.

The Procedure of RF Algorithm Application

본 연구에서는 파이썬(Python) 기반 구글 코랩(Google Colaboratory) 환경에서 랜덤포레스트 알고리듬을 적용하였다. 먼저 형태소 분석기를 활용하여 에세이를 문장, 단어 단위로 분할하고, 에세이 데이터에서 채점자질을 선택하여 데이터를 전처리하여 머신러닝에 용이한 형태로 가공하였다. 다음으로는 전처리된 에세이 데이터를 학습용(train) 데이터와 검증용(test) 데이터로 분할하였는데 본 연구에서는 검증용 데이터는 20%로 설정하였다. 이때, 데이터를 무작위로 섞은 후 지정된 비율에 따라 학습용 데이터와 검증용 데이터로 분할하였다.

다음으로 랜덤포레스트 알고리듬을 활용하여 머신러닝을 수행하였다. 머신러닝 결과 도출된 모델에 대해서 20%로 설정한 검증용 데이터를 이용하여 모델을 평가하였는데 이때, 모델이 예측한 결과값과 실제 결과값을 비교하여 정확도, 정밀도, 재현율, F1-score 성능 지표를 계산하였다. 구체적으로는 confusion_matrix 함수를 사용하여 혼동 행렬(Confusion Matrix)을 계산하고, classification_report 함수를 사용하여 정밀도(precision), 재현율(recall), F1 점수(F1-score)를 출력하였다. 끝으로 투입한 각 채점자질의 중요도를 확인하였다.

Results and Discussion

Extract Scoring Feature

채점자질은 에세이 평가 기준이나 선행연구, 전문가의 판단 등을 고려하여 다양하게 선정할 수 있으나 본 연구에서는 형태소별 빈도 자질, 문장 수 자질과 같은 가장 기본적인 채점자질을 추출하여 채점 모델을 생성하였다. 문장 수, 형태소별 사용 빈도와 같은 채점자질은 심층적인 채점자질을 도출하기 위한 가장 기본적인 채점자질에 해당한다.

글의 수준을 판단할 수 있는 채점자질로 문장 수, 형태소별 사용 빈도를 선정한 것은 문장의 복잡성이나 통사적 성숙도 등의 쓰기 유창성이 글의 질에 영향을 미친다고 판단하였기 때문이다. 쓰기 유창성은 '학생이 글을 쓰는 데 있어서 단어와 문장 구사가 점점 능숙해지고 그 길이가 증가하는 정도'라고 볼 수 있다(Issacson, 1985). 쓰기 능력이 발달하면서 단어나 문장 등에서 양적 증가가 유창성을 판단하는 하나의 요인인데 가장 대표적으로는 단어 수나 글자 수를 세는 방식이다. 다음으로 유창성이 증가하는 것은 양적 증가 외에 구문 성숙도(syntax maturity) 즉, 보다 복잡한 문장을 사용하는 정도로 판단할 수 있다(Issacson, 1985). 학년이 올라갈수록 문장을 복잡하게 생성하면서 자신의 생각을 보다 심층적으로 표현하는 능력이 발달하고(Hunt, 1965; Owens, 2012 등), 이렇게 통사적 숙달도가 높다는 것은 문장의 길이, 절의 길이, 종속절의 비율이 증가하며(Hunt, 1965), 명사를 수식하는 양상이나 시제 표현 등 문법적 요소의 사용이 증가한다는 것을 의미한다(Owens, 2012).

이에 본 연구에서는 성숙한 문장을 구사한다는 것을 형태소의 사용 빈도를 통해서 측정하고자 하였다. 형태소는 뜻을 가진 가장 작은 말의 단위이다. 형태소는 자립성의 여부에 따라 자립 형태소와 의존 형태소로 나누고, 의미와 기능에 따라 실질 형태소와 형식 형태소(문법 형태소)로 나눈다. 형태소 중에서도 문법적인 기능을 하는 형식 형태소의 사용 빈도는 높은 수준의 문장을 구사하고 있다는 것을 나타내는 표지이면서 동시에 여러 문장으로 구성되는 한 편의 글의 질을 판단할 수 있는 특징이기도 하다. 보통 머신러닝을 할 때 실제적인 의미를 가지지 않는 형식 형태소를 불용어로 처리하는 경우가 많다. 그러나 본 연구에서는 문법적인 의미를 가지고 있는 형식 형태소를 문장을 정확하고 유창하게 구사하기 위해서 필수적인 요소라고 판단하여 중요한 채점자질로 설정하였다.

형태소 기반의 채점자질을 도출하기 위해서는 먼저 형태소 분석기를 활용하여 글을 형태소 단위로 분석하여 품사를 부착하고, 품사 부착된 결과를 활용하여 형태소별 빈도를 도출해야 한다. 본 연구에서는 대표적인 형태소 분석기인 꼬꼬마(Kkma) 분석기를 활용하여 Table 5와 같은 형태소 자질을 분석하였다[1].

Table 5. Scoring feature

Category	Feature	Kkma tag
Noun	botong myeongsa	NNG
	goyu myeongsa	NNP
	ilban uijon myeongsa	NNB
	dan-wi uijon myeongsa	NNM
	susa	NR
	daemyeongsa	NP
Verb	dongsa	VV
	hyeong-yongsa	VA
	bojo dongsa	VXV
	bojo hyeong-yongsa	VXA
	seosulgyeog josa ‘ida’	VCP
	hyeong-yongsa ‘anida’	VCN
Determiner	ilban gwanhyeongsa	MDT
	su gwanhyeongsa	MDN
Adverb	ilban busa	MAG
	jeobsog busa	MAC
Interjection	gamtansa	IC
Particle	jugyeog josa	JKS
	bogyeog josa	JKC
	gwanhyeong-gyeog josa	JKG
	mogjeoggyeog josa	JKO
	busagyeog josa	JKM
	hogyeog josa	JKI
	in-yong-gyeog josa	JKO
	bojosa	JX
	jeobsog josa	JC
Pre-final ending	jonching seon-eomal eomi	EPH
	sije seon-eomal eomi	EPT
	gongson seon-eomal eomi	EPP
Final ending	pyeongseohyeong jong-gyeol eomi	EFN
	uimunhyeong jong-gyeol eomi	EFQ
	myeonglyeonghyeong jong-gyeol eomi	EFO
	cheong-yuhyeong jong-gyeol eomi	EFA
	gamtanhyeong jong-gyeol eomi	EFI
	jonchinghyeong jong-gyeol eomi	EFR
	daedeung yeongyeol eomi	ECE
	uijonjeog yeongyeol eomi	ECD
	bojojeog yeongyeol eomi	ECS
	myeongsahyeong jeonseong eomi	ETN
	gwanhyeonghyeong jeonseong eomi	ETD
Prefix	cheeon jeobdusa	XPN
	yong-eon jeobdusa	XPV
Suffix	myeongsa pasaeng jeobmisa	XSN
	dongsa pasaeng jeobmisa	XSV
	hyeong-yongsa pasaeng jeobmisa	XSA

다음으로 문장 수 자질을 도출하였다. 문장 수를 추출하기 위해서는 먼저 에세이 원본에 대해 문장 분리를 진행해야 한다. 문장 분리는 Kss (Korean Sentence Splitter) 분석기를 활용하였다. Kss는 문장의 종결을 나타내는 문장 경계를 찾아낸다. 예를 들어, 온점을 포함한 문자열에서 온점과 그 다음 문자 사이에 공백이나 줄바꿈이 있는 경우, 해당 온점을 문장 경계로 인식한다. Kss 분석기는 특히 문장 분리를 빠르고 정확하게 처리하는 데에 높은 성능을 보인다.

최근에는 한국어 형태소 분석기의 성능을 분석하고 개선하고자 하는 연구들도 진행되고 있다. 한국어 형태소 분석은 주로 음절을 단위로 이루어지고 있는데 이 방법은 시퀀스 레이블링(Sequence Labeling)과 시퀀스 생성(Sequence Generation)으로 나뉜다(Choi & Lee, 2020). 기존 자연어처리에서는 시퀀스를 다루기 위해서 LSTM (Long Short-Term Memory)을 주로 사용하였는데 이 신경망 모델은 문장이 길어질수록 성능이 떨어지고 순차적 처리를 해야하기 때문에 병렬 처리가 불가능하다(Choi & Lee, 2020). 그러나 Transformer 모델이 제안되면서(Vaswani et al. ,2017) 병렬 처리가 가능하여 기존 LSTM의 문제를 해결하고 있다. 또한 최근에는 BERT와 같은 사전 훈련된 언어 모델이 제안되면서(Devlin et al. ,2018) 양방향의 문맥을 고려할 수 있다. 이 외에도 잘 알려져 있는 ELMo (Peter et al., 2018)와 GPT (Radford et al., 2020) 또한 사전 훈련된 언어 모델이지만 이들 모델은 BERT와 달리 단방향으로 학습을 진행하는 모델이다. 또한 BERT는 단어를 토큰화할 때, WordPiece (Wu, 2016) 방법을 사용하는데 이는 기존의 토큰 활용 방식이 아니라 단어의 subword를 사용하기 때문에 코퍼스의 크기에 비례하여 토큰 수도 많아지는 제한점을 해결할 수 있다. 이에 앞으로의 연구에서는 형태소 분석기의 성능을 보다 개선하기 위해 다양한 모델을 적용해 보고, 에세이 자동채점에 가장 적합한 분석 방법을 찾는 것이 중요하다. 특히 본 연구에서처럼 형태소 분석을 통해 기본적인 채점자질을 선정하는 경우, 형태소 분석기의 정확도가 매우 중요하다.

Result of Application RF Algorithm

앞서 추출한 문장 수, 형태소별 사용 빈도 자질 정보와 에세이 점수 정보를 사용하여 랜덤포레스트 알고리듬으로 머신러닝을 진행하고 그 성능을 검증하였다.

먼저 전체 성능을 살펴보면 Table 6에서 정확도(Accuracy)는 0.70으로 81개의 샘플 중에서 70%에 해당하는 점수를 올바르게 예측하였다. 머신러닝 모델의 성능을 검증할 때, 많은 경우 '정확도' 값만 보고하는데 이는 데이터가 클래스별 불균형이 클 경우에는 해석에 주의가 필요하다. 예를 들어 100개의 샘플 중에서 80개가 1점인 데이터의 경우, 100개의 샘플 모두에 대해서 1점으로 예측하더라도 정확도는 80%라고 보고되기 때문이다. 따라서 데이터 불균형이 있을 경우 정확도만으로는 모델의 성능을 평가하기에 한계가 있어서 정밀도, 재현율, F1 점수 등을 다양하게 살펴보는 것이 중요하다.

Table 6. Random forest model performance(overall)

Test data	81		
Accuracy	0.70		
	precision	recall	f1-score
Macro avg	0.74	0.66	0.67
Weighted avg	0.72	0.70	0.70

이에 본고에서도 다양한 성능 지표를 통합적으로 살펴보고자 하였다. 다음으로 macro avg는 Table 7에 제시한 1, 2, 3점의 각 분류 클래스별로 정밀도(precision)[2], 재현율(recall)[3], F1-score[4]를 계산하고, 이들의 평균을 구한 값이다. macro avg에서 정밀도는 0.74, 재현율은 0.66, F1-score는 0.67로 나타났다. weighted avg는 정밀도, 재현율, F1-score의 가중평균값이다. 클래스별 샘플 수에 따라 가중치를 부여하기 때문에, 크기가 큰 클래스에 더 많은 영향을 받는다. macro avg는 정밀도, 재현율, F1-score가 0.72, 0.70, 0.70으로 도출되었다.

Table 7. Random forest model performance(by score)

score	precision	recall	f1-score	test data
1	0.86	0.43	0.57	14
2	0.72	0.71	0.72	41
3	0.65	0.85	0.73	26

다음으로는 Table 7에서 제시한 각 클래스별 성능을 살펴보겠다. 1점 클래스의 경우 정밀도가 0.86인데 이는 기계가 1점이라고 예측한 글 중 실제로 1점인 글의 비율이 86%라는 것이다. 재현율은 실제 양성 샘플 중 모델이 예측한 양성 샘플의 비율을 나타내므로 1점 클래스의 경우, 실제 1점인 샘플 중 43%만이 모델에 의해 양성으로 예측되었다는 것이다. 1점 클래스의 빈도가 적어서인지 정밀도와 재현율에 차이가 크게 나타났다. 그러나 2점과 3점 클래스의 경우 정밀도와 재현율에 큰 차이가 보이지 않았다.

본고에서는 1점 클래스의 경우 2점과 3점에 비하여 성능이 떨어졌지만 전반적으로 70% 정도의 예측을 하는 성능이라고 해석이 가능하다. 이러한 성능 값을 절대적인 기준으로 판단하기는 어렵지만 선행 연구들과 비교는 가능하다. Park & Lee (2022)에서도 본 연구와 유사하게 500개의 한국어 에세이에 대해서 품사 기반의 채점자질을 활용하여 랜덤포레스트 알고리듬을 적용하고 2, 3, 4점의 3개 클래스로 예측하는 채점 모델을 개발하였다. Park & Lee (2022)에서는 랜덤포레스트를 이용한 머신러닝을 진행하면서 과적합 방지 및 최고 모델 성능 도출을 위하여 다양한 하이퍼 파라미터 미세조정을 진행한 후 F1값이 40% 정도의 성능을 보고하였다. 따라서 선행연구와 비교할 때에는 어느 정도 개선된 모델이라고 해석할 수 있다. 또한 Park & Lee (2022)에서는 머신러닝 계열의 랜덤포레스트 외에도 딥러닝 계열의 사전 학습 언어모델인 KoBERT를 적용하였는데 그 결과 성능은 F1 값이 약 20%인 것으로 보고되었다. 이처럼 학습용 데이터 수가 500개 수준으로 많지 않은 경우에는 사전 정의된 채점자질을 활용하는 모델이 보다 효과적임을 알 수 있다.

다음으로는 랜덤포레스트 모델에서 각 등급을 분류할 때 불순도 개선 값(meandecreaseGINI)을 이용하여 채점자질의 상대적 중요도를 분석하였다. 랜덤포레스트 모델에서 불순도 개선값은 결정 트리의 분할 기준을 선택하는 데에 사용된다. 불순도를 감소시킬 수 있는 방향으로 노드 분할을 선택하게 되는데 GINI 계수나 엔트로피는 분류 문제에서 사용되고 회귀 문제에서는 분산이나 평균제곱오차를 사용하여 불순도를 계산한다. 따라서 노드 분할 전의 불순도에서 분할 후의 불순도를 빼서 개선값을 계산하게 되는데 불순도 개선값이 클수록 보다 효과적인 분할이라고 판단할 수 있다. 랜덤포레스트 모델에서는 이러한 불순도 개선값을 높이는 데에 기여한 채점자질의 상대적 중요도를 분석할 수 있다. 그 결과는 Table 8과 같다.

Table 8에서는 상대적 중요도가 높은 상위 15개의 채점자질을 보여주고 있다. 랜덤포레스트 모델에서 각 등급을 분류할 때 가장 기여도가 높은 채점자질은 '보통명사'이고, 이어서 관형격 조사, 동사, 관형형 전성 어미 등이 순서대로 중요도가 높게 나타났다. 앞서 기술하였듯이 문장 수와 같이 글의 길이를 보여주는 채점자질은 쓰기 유창성을 측정하는 주요 지표이다. 그러나 '문장 수'의 경우 상대적 중요도가 낮게 나타났는데 이후 분석에서는 '문장 수' 대신에 문장 복잡성을 설명할 수 있는 채점자질을 재설정하여 채점 모델을 수정하고자 한다.

Table 8. Feature importance

Feature	Feature importance
NNG	0.089959
JKG	0.064654
VV	0.047990
ETD	0.045409
JKS	0.039567
JKM	0.038767
XSV	0.037304
JKO	0.037025
ECD	0.035690
JX	0.035301
ECE	0.033877
VXV	0.032836
MAG	0.032168
Number of sentence	0.032117
XSN	0.031338

Park & Lee (2022)에서는 '표현' 영역의 점수 분류에 있어서 어휘 수가 약 0.07의 중요도로 가장 높게 나타났고 명사의 경우 0.05에 다소 못 미치는 중요도로 7번째 순위로 나타났다. 어떠한 채점자질이 분류에 높게 기여하고 있는지는 다양한 연구가 반복되면서 보다 유의미한 비교가 가능할 것으로 생각된다. 또한 본 연구에서는 형태소 빈도와 문장 수와 같은 기본적인 채점자질만을 사용하였기 때문에 각 채점자질을 글의 특성에 대한 해석 및 피드백으로 연결하는 데에 한계가 있다. 그러나 이후에는 교육적으로 보다 유의미한 채점자질을 선정하면서 모델의 성능을 비교해보고자 한다.

본 연구에서 사용한 채점자질들은 형태소 분석을 통해 얻은 가장 단순한 형태의 자질이다. Rahimi et al. (2017)에서는 채점자질이 채점기준을 얼마나 잘 설명하고 있느냐에 따라서 자동채점의 정확도에 영향을 미친다고 제안하였다. 우리나라에서 채점자질을 사용하는 연구는 이제 시작 단계이다. 향후 연구에서는 채점기준을 토대로 다양한 채점자질을 설계하면서 글의 내용이나 조직, 표현 등을 보다 구체적으로 설명할 수 있는 채점자질을 도출하고자 한다. 최근 연구에서는 그간 채점자질은 주로 높은 수준의 글에서 드러나는 특성을 중심으로 설계되어 왔으나, 낮은 수준의 글을 충분하게 확보하여 낮은 수준 글을 대표하는 자질을 활용하는 것이 보다 효과적이라는 연구들도 제안되고 있다. 예를 들어 Revision Assistant (West-Smith et al., 2018; Woods et al., 2017)는 채점기준과 관련이 있는 사항들에 대해서 문장 수준의 피드백을 제공하는데 낮은 수준의 글에 드러나는 특성을 채점자질로 활용하여 이를 피드백으로도 연결하고 있다. 이처럼 자동채점의 정확도 외에도 피드백을 고려한 채점자질의 설계 또한 중요한 연구 주제가 될 것이다. 끝으로 영어권에서는 주로 ASAP 데이터베이스를 활용하여 딥러닝 기반의 모델을 적용함으로써, 채점자질을 설계하거나 학습용 데이터를 구축하는 과정 없이, 자동채점의 성능만을 높이려는 시도들이 주로 이루어졌다. 그러나 최근에는 학습용 데이터의 불균형에 따른 시스템의 제한점을 해소하기 위해 자동채점의 연구 영역이 단순히 기계학습 전문가들의 영역이 아니라 질 높은 데이터베이스 구축, 채점자 전문성, 채점자질 설계 등과 같은 작업을 각계의 전문가들이 협력하여 진행해야 한다는 관점이 중요해지고 있다(Kuma & Boulanger, 2020; Madnani et al., 2017; Madnani & Cahill, 2018). 본 연구에서 도출한 형태소 기반의 채점자질 외에 보다 심층적인 채점자질을 도출하기 위해서는 자연어처리 전문가, 기계학습 전문가들과 국어교육 전문가들 간의 긴밀한 협력 연구가 필요할 것이다.

Conclusions

본 연구에서는 중등학생이 작성한 401편의 한국어 에세이 데이터를 활용하여 랜덤포레스트 기반의 알고리듬을 적용하여 자동채점 모델을 생성하고 그 성능을 분석하였다. 그 결과 0.70 정도의 정확도를 확인할 수 있었다. 이 결과를 통해서 다음과 같은 결론을 도출하였다.

첫째, 본 연구에서 도출한 성능은 한국어 에세이를 대상으로 하는 선행 연구와 비교할 때 상대적으로는 높다고 할 수 있다. 상대적으로 성능이 높게 나타난 이유는 여러 가지로 해석할 수 있지만 머신러닝에 사용된 학습용 데이터의 질이 영향을 미쳤다고 볼 수 있다. 사용한 데이터의 수가 유사하고 채점자질 유형, 랜덤포레스트 알고리듬과 같은 방법 또한 유사한 상황에서 결과값이 다른 것은 사용한 데이터의 특성이라고 볼 수 있다. 선행 연구에서 사용한 에세이와 본 연구에서 사용한 에세이 데이터의 채점 신뢰도 등의 특징을 비교분석하기는 어렵겠지만 추후에는 데이터의 특성이 채점 모델 생성에 어떠한 영향을 미치는지를 보다 구체적으로 탐색할 필요가 있다. 본 연구를 통해서 학습용 에세이 데이터가 갖추어야 하는 조건에 대해서 다음과 같은 사항들을 고려할 수 있다. 먼저, 에세이 데이터는 정확하고 신뢰할 수 있는 채점 결과를 포함해야 하고, 각 점수나 등급의 경계에 해당하는 에세이의 경우 특히 정확한 채점이 이루어져야 한다. 다음으로, 에세이 데이터베이스는 각 점수 범위에 대한 적절한 분포를 갖추어야 한다. 즉, 높은 점수와 낮은 점수를 가진 에세이가 적절한 비율로 포함되어야 모든 점수 범위에 대해 충분한 머신 러닝이 수행될 수 있고, 모델의 성능을 높일 수 있다.

둘째, 본 연구에서는 형태소 활용 빈도 및 문장 수 채점자질을 도출하여 채점 모델을 생성하였다. 매우 단순한 계량적 채점자질들이지만 이들 채점자질만으로도 0.70 정도의 성능을 도출할 수 있음을 확인하였다. 추후 연구에서는 에세이의 채점 기준과 관련하여 보다 심층적인 측면을 정의할 수 있는 채점자질을 설정하고, 채점자질의 측정 공식을 개발하고자 한다. 이들 채점자질을 추가하면서 채점 모델의 성능을 지속적으로 비교하는 연구를 축적할 것이다. 특히 에세이 채점자질을 탐색하는 연구는 국어를 전공하는 전공자나 학계에서 적극적으로 관심을 가져야할 연구 영역이다. 실제로 영어권에서 사용하는 다양한 자동채점 시스템들은 수십개의 채점자질을 사용하고 있지만 극소수만 공개하고 있다. 따라서 한국어 에세이에 대해서도 다양한 채점자질을 개발하여 연구 결과를 축적할 필요가 있다.

셋째, 본 연구에서 사용한 랜덤포레스트 알고리듬을 개선하고 나아가 딥러닝 기반의 알고리듬과도 비교, 분석할 필요가 있다. 추후 연구에서는 랜덤포레스트 알고리듬을 적용하는 과정에서 하이퍼파리미터 미세조정을 진행하면서 모델의 성능을 향상시킬 필요가 있다. 물론 본 연구에서 사용한 데이터의 수가 적기 때문에 미세조정을 수행하더라도 과적합 등의 문제가 여전히 발생할 수 있다. 그러나 다양한 조정을 통해서 모델의 예측 성능을 높일 수 있는 후속 연구가 수행될 필요가 있다. 또한 딥러닝 기반의 알고리듬을 적용해 보고, 머신러닝 기반의 알고리듬과 성능을 비교하면서 데이터 컨디션에 따라 어떠한 방법이 보다 적합한지 탐색해 볼 필요가 있다. 이러한 연구를 진행하기 위해서는 다양한 연구 영역의 전문가들과의 협업이 중요할 것이다.

추후 연구에서는 이러한 시사점을 바탕으로 보다 심층적인 채점자질을 설정하고, 보다 정교한 채점 모델을 도출해보고자 한다. 한국어 에세이 자동채점에 대해서도 다양한 연구가 축적되면 당장 고부담 시험 상황에 자동채점을 도입하기는 어렵더라도 인간 채점을 지원하는 정보로 사용하거나, 학생들의 에세이 작성 연습 상황에서 자동채점 기술이 활용될 수 있을 것으로 기대해 본다.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160.

Choi, S. K. (2018). A study on the patterns of source text borrowing performance in the integrated writing of the secondary students focusing on citations and paraphrasing. Research on Writing, 38, 201-245.

Choi, Y. S., & Lee, K. J. (2020). Performance Analysis of Korean Morphological Analyzer based on Transformer and BERT. Journal of KIISE, Vol. 47, No. 8, pp. 730-741,

Chung, G. K., & O'Neil, H. F. (1997). Methodological approaches to online scoring of essays. Center for the Study of Evaluation Technical Report 461, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Dascalu, M., Allen, K. A., McNamara, D. S., Trausan-Matu, S., & Crossley, S. A. (2017). Modeling comprehension processes via automated analyses of dialogism. In: 39th Annual Meeting of the Cognitive Science Society (CogSci 2017). Cognitive Science Society, London.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ha, M, S., Lee, G. G., Shin, S. I., Lee, J. K., Choi, S. C., Choo, J. G., ... Park, J. S. (2019). Assessment as a learning-support tool and utilization of artificial intelligence: WA3I project case. School Science Journal, 13, 271-282.

Hunt, K. W. (1965). Grammatical Structures Written at Three Grade Levels. Champaign, IL: National Council of Teachers of English, Research Report No.3.

Issacson, S. (1985). Assessing written language skills. In C.S. Simon(Ed), CommUnication Skills and Classroom Success: Assessment Methodologoies for Language-Learning Disabled Students. Sandiego: College Hill Press.

Kumar, V., & Boulanger, D. (2020, October). Explainable Automated Essay Scoring: Deep Learning Really has Pedagogical Value. In Frontiers in education (Vol. 5, p. 572367). Frontiers Media SA.

Lantz, B. (2013). Machine Learning with R. Packt Publishing Ltd., Birmingham.

Lee, Y. S., Shin, D. K. & Kim, H. J. (2022). Exploring the feasibility in applying an automated essay scoring to a writing test of Korean language, Bilingual Reseach, 86, 171-191.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al.(Eds), Advances in Neural Information Processing Systems, (pp. 4765–4774). NY: Curran Associates, Red Hook.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee & L. Zettlemoyer. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Madnani, N., & Cahill, A. (2018). “Automated scoring: beyond natural language processing,” In Proceedings of the 27th International Conference on Computational Linguistics, (Santa Fe: Association for Computational Linguistics), 1099–1109.

Madnani, N., Loukina, A., von Davier, A., Burstein, J., & Cahill, A. (2017). “Building better open-source tools to support fairness in automated scoring,” In Proceedings of the First (ACL) Workshop on Ethics in Natural Language Processing, (Valencia: Association for Computational Linguistics), 41–52.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071-22080.

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. Massachusetts: MIT press.

Nam, M. J. & Won, M. J. (2022). A study on exploring language qualifications for automatic scoring of Korean writing. Language Facts and Perspectives, 53, 9-32.

Nguyen, H., & Litman, J. (2018). Argument mining for improving the automated scoring of persuasive essays. In Proceedings of the AAAI Conference on Artificial Intelligence. pp.5892–5899.

Noh, E. H., Kim, M. H., Sung, K. H., & Kim, H. S. (2013). Refinement and trial application of auto-scoring program for short-answer questions for large-scale evaluation. Korea Institute of Curriculum and Evaluation Research Report RRE 2013-5.

Noh, E. H., Lee, S. H., Lim, E. Y., Sung, K. H., & Park, S. Y. (2014). Development oF Korean Self-Answer Questions Automatic Scoring Program and Verification of Practicality(Research Report RRE 2014-6). Korea Institute of Curriculum and Evaluation.

Noh, E. H., Sim, J. H., Kim, M. H., & Kim, J. H. (2012). A Study on Automatic Scoring of Short-Answer Questions for Large-Scale Evaluation(Research Report RRE 2012-6). Korea Institute of Curriculum and Evaluation.

Noh, E. H., Song, M. Y., Park, J. I., Kim, Y. H., & Lee, D. G. (2016). Advancement and Application of Automatic Scoring Program for Korean Sentence-Level Written Answer Questions(Report RRE 2016-11). Korea Institute of Curriculum and Evaluation Research.

Noh, E. H., Song, M. Y., Sung, K. H., & Park, S. Y. (2015). Development and Application of Automatic Scoring Program for Korean Sentence-level Self-answer Questions(Research Report RRE 2015-9). Korea Institute of Curriculum and Evaluation.

Owens, R. (2012). Language Development: An Introduction (8th ed.). Pearson.

Page, E. B. (1966). The imminence of grading essays by computer. Phi Delta Kappan, 48, 238-243.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. Journal of Experimental Education, 62, 127–142.

Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), Automated Essay Scoring: A Cross-disciplinary Perspective (pp. 43-54). Lawrence Erlbaum Associates, Inc.

Park, J. I., Lee, M. B., Lee, S. H., Song, M. H., Lee, M. J., & Choi, S. K. (2022). Design of Automatic Scoring Method for Computer-Based Written and Essay Evaluation (I)(Research Report RRE 2022-6). Korea Institute of Curriculum and Evaluation.

Park, K. Y. & Lee, Y. S. (2022). Deep learning algorithm exploration for automated Korean essay scoring. Journal of Educational Evaluation, 35, 465-488.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. arXiv. arXiv preprint arXiv:2012.11747.

Rahimi, Z., Litman, D., Correnti, R., Wang, E., & Matsumura, L. C. (2017). Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. International Journal of Artificial Intelligence in Education, 27, 694–728.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?": Explaining the predictions of any classifier. CoRR, abs/1602.0. arXiv [Preprint]. Available online at: http://arxiv.org/abs/1602.04938 (accessed September 22, 2020).

Shin, D. K. (2022). Effects of scoring features on the accuracy of the automated scoring model of English. Korean Journal of Teacher Education, 38, 73-91

Taghipour, K. & Ng, H. T. (2016). A neural approach to automated essay scoring. Proceedings of the 2016 conference on empirical methods in natural language processing (pp. 1882–1891). Austin, Texas: Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

West-Smith, P., Butler, S., & Mayfield, E. (2018). “Trustworthy automated essay scoring without explicit construct validity,” in Proceedings of the 2018 AAAI Spring Symposium Series, (New York, NY: ACM).
Woods, B., Adamson, D., Miel, S., & Mayfield, E. (2017). “Formative essay feedback using predictive scoring models,” in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (New York, NY: ACM), 2071–2080.
Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.
Zhao, S., Zhang, Y., Xiong, X., Botelho, A., & Heffernan, N. (2017). A memory-augmented neural model for automated grading. In Proceedings of the Fourth ACM Conference on Learning, 189-192.

Authors Information

Park, Jong-Im: Korea Institute for Curriculum and Evaluation, Researcher, First Author and Corresponding Author
Choi, Sook-Ki: Korea National University of Education, Professor, Co-author

Endnotes

1. 한국어 형태소 분석 범주(Category)는 유사한 영어 형태소 명칭으로 번역하였으나, 세부 항목에 해당하는 채점자질(feature)의 경우 한국어 채점자질에 해당하는 영어가 없고, 국내 문법 학계에서 통일된 번역어가 없기 때문에 소리나는대로 영어 표기를 하였음.
2. 정밀도(Precision)는 모델이 예측한 양성(positive) 샘플 중에서 실제 양성인 샘플의 비율. 다시 말해, 양성으로 예측한 것 중에서 얼마나 실제로 양성인지를 측정함.
3. 재현율(Recall)은 실제 양성인 샘플 중에서 모델이 양성으로 예측한 샘플의 비율. 다시 말해, 실제로 양성인 것 중에서 얼마나 많이 양성으로 예측했는지를 측정함.
4. 정밀도와 재현율의 조화 평균값. 불균형 데이터로 머신러닝을 수행한 경우에는 주로 F1 값으로 성능을 측정함.