http://dx.doi.org/10.17232/KSET.36.2.353
Journal of Educational Technology
2020, Vol 36, No 2, pp.353-382
〔Theories & Issues〕

The Present and Future of AI-based Automated Evaluation: A Literature Review on Descriptive Assessment and Other Side

Gyeong-Geon Lee (Seoul National University Doctoral Student)
Minsu Ha† (Kangwon National University Professor)

This study primarily reviews the present of AI(Artificial Intelligence)-based automated evaluation of descriptive assessment with its technological/ethical issues, to derive short-term strategies for utilization of it. AI-based automated evaluation system of descriptive assessment have been driving force of paradigm change of evaluation, being used for scoring usual essays and even scientific concept problems. However, it also encounters some challenges such as (1) technological limitation of natural language processing, (2) a matter of scoring reliability, (3) user trust and deliberate cheating, and (4) ethical issues with respect to high-stake exams. Thus in short-term range, AI-based automated evaluation system can be used as (1) adaptive supporting tool for students' learning, (2) fast learning analytic tools for students' response, and (3) analysis of evaluation criteria. Yet, once those problems are overcome in technological/ethical aspects, AI-based system could be used for automated evaluation of (1) visual representation of mental model, (2) authentic practice, and (3) real-time feedback and process-based evaluation, in the dimension of image processing over text processing, in subject-specific sense. Though, as nobody predicted deep learning algorithm to be mainstream of AI technology like nowadays, a future possibility of AI-based automated evaluation system would surely be divergent.

Keywords : artificial intelligence(AI), descriptive assessment, automated evaluation, automated scoring, learning analytics

† Correspondence : Minsu Ha, Kangwon National University, msha@kangwon.ac.kr

인공지능 기반 자동평가의 현재와 미래: 서술형 문항에 관한 문헌 고찰과 그 너머*

이 경 건 (서울대학교 박사과정)
하 민 수†(강원대학교 교수)

〈요 약〉

본 연구에서는 인공지능 기반의 서술형 문항 자동평가 시스템의 현재와 기술적 및 윤리적 도전을 살펴보고, 그 단기적인 활용 방안을 살펴보았다. 인공지능 서술형 문항 자동평가 시스템은 이미 광범위한 주제의 에세이 채점뿐만 아니라 정교한 과학적 개념을 묻는 문항에 대한 채점에도 사용될 만한 성능을 보이며 평가의 패러다임을 바꾸어가고 있다. 하지만 인공지능 서술형 자동평가 시스템은 (1) 자연어 처리의 기술적 한계, (2) 채점 신뢰도 문제, (3) 인간 사용자의 신뢰와 인위적 속임, (4) 고부담 시험에서의 윤리적 한계 등의 도전을 안고 있다. 이에 단기적으로는 인공지능 자동평가 시스템이 (1) 학습자에게 적응적인(adaptive) 학습 지원 도구로서, (2) 학습자에게 빠른 피드백을 제공하고, (3) 평가 준거의 빠른 분석을 가능케 하는 역할을 감당할 수 있다. 하지만 이러한 문제들이 기술적/윤리적으로 극복된다면, 인공지능 자동평가 시스템의 미래는 과학 등의 교과-특수적인 영역에서, 텍스트를 넘어서는 이미지 처리를 통해 (1) 시각적 표상에 대한 자동평가, (2) 실제 수행에 대한 자동평가, (3) 실시간 피드백과 과정중심 자동평가에 활용될 수 있을 것이다. 하지만 딥 러닝이 인공지능 기술의 급속한 발달을 주도하게 될 것을 누구도 예측하지 못하였듯이, 인공지능 자동평가 시스템의 미래적 가능성 역시 여전히 열려 있다고 하겠다.

주요어 : 인공지능, 서술형 문항, 자동평가, 자동채점, 학습분석

* 이 논문은 2019년도 교육부의 재원으로 한국과학창의재단 창의교육 거점센터사업의 지원을 받아 수행된 연구임.
† 교신저자 : 하민수, 강원대학교, msha@kangwon.ac.kr

Ⅰ. 서 론

2016년 이세돌 九단과 알파고(AlphaGO)의 대국은 인공지능의 성능이 기존에 비하여 비약적으로 상승하고 있음을 우리에게 알린 신호탄과 같았다. 이후로 인공지능 기술에 의한 공장 자동화를 핵심으로 하는 이른바 '4차 산업 혁명' 담론이 한국 사회 전반에서 퍼져나갔을 뿐 아니라, 교육학계에서도 이와 관련한 논의들이 상당수 이루어지게 되어 인공지능에 의한 교육 실천의 변화가 조만간 도래할 것으로 예상되기도 하였다(e.g. 김홍겸 외, 2018). 하지만 딥 러닝(deep learning)[1]으로 대표되는 오늘날의 인공지능 기술은 그 개발자들까지도 온전히 이해하기 쉽지 않을 만큼 복잡하고 규모가 큰 모델을 생성하는 특징을 지닌다. 이에 따라 교육학 내용 전문가(subject-matter expert, SME)들은 인공지능 기술에 의한 교육 혁신의 필요성 혹은 불가피성을 절감하면서도 이를 실현하기 위한 구체적인 방법에 있어서는 어려움을 겪을 수 있다. 그러므로 오늘날의 교육학자들이 살펴보아야 하는 지점은 이러한 테크놀로지의 발달과 기존의 교육학 이론 체제가 맞닿으며 발전이 이루어지고 있는 부분이라고 할 것이다.

오늘날 교실 현장에서는 학습자의 선개념이나 사전 지식 등과 같은 발달 수준에 맞추어 수업을 계획하고 학습자의 개념 변화를 종단적으로 확인하는 구성주의적 패러다임이 강조되고 있다(Magnusson et al., 1997). 이러한 구성주의적 수업에서는 학생들의 선개념 또는 사전 지식을 수업 전에 확인하고 수업 후에 다시 확인하여야 하므로, 진단평가와 형성평가가 교수학습활동의 중요한 단계로 여겨진다. 교사는 이러한 평가를 통해 자신의 수업이 효과적으로 학생들의 지식을 발달시키거나 선개념을 변화시켰는지를 확인하고, 차후 수업을 계획하기 위한 중요한 정보를 얻는다. 결국 진단평가와 형성평가는 교수학습 상황에서 교사와 학생 모두에게 중요한 과정이다.

교수학습을 지원하기 위한 맥락에서 효율적인 평가란, 학습자의 개념이나 지식 수준에 관하여 더 많은 정보를 획득하는 평가일 것이다. 특히 학습자가 가지고 있는 '진짜 이해'를 확인하는 것이 중요하다. 이처럼 진단의 기능이라는 면에서 볼 때 여러 평가

1) 인공신경망(Artificial Neural Network, ANN)의 한 형태이다. 인공신경망은 마치 인간의 개별 뉴런과도 같은 퍼셉트론(perceptron)을 컴퓨터상에서 논리적으로 구현하여, 그 복잡한 연결을 통해 인공지능을 구현하고자 하는 시도이다. 인공신경망은 일반적으로 입력층, 은닉층, 출력층으로 이루어지며 개별 층마다 퍼셉트론의 개수 및 그 연결 구조는 사용자가 구성할 수 있다. 은닉층의 개수 또한 사용자가 임의로 늘려나갈 수 있는데, 일반적으로 신경망의 층이 3층 이상이 될 때를 깊은 신경망(Deep Neural Network, DNN)이라고 부른다. 딥 러닝은 이러한 깊은 신경망에서 퍼셉트론 간의 연결 강도로 주어지는 모수들을 데이터에 기반하여 학습시키는 과정이다.

의 방법 중 서술형 평가가 선택형 평가보다 더 효율적이라고 할 수 있다. 선택형 평가에서 교사는 학생들이 가지고 있을 것이라 짐작되는 여러 개념을 예측하여 제시하고, 학생은 그 중에서 답안을 선택한다. 여기서 학생들이 가진 개념의 일부가 드러날 수 있겠지만, 선택형 평가의 결과가 학생들의 진정한 이해를 반영한다고 단정할 수는 없다(Beggrow et al., 2014). 반면 서술형 평가에서 학생들은 자신의 생각을 표현할 때 자신이 알고 있는 단어를 조합하여 직접 답안을 생성하므로, 그것이 학생들이 가지고 있는 진짜 이해라고 판단할 수 있다(Opfer et al., 2012).

하지만 교수학습 과정에서 서술형 문항으로 학생들의 개념이나 지식을 평가하려면 물리적으로 많은 시간이 요구된다. 선택형 평가에서와 달리, 서술형 평가에서 학생들이 작성한 응답에는 매우 다양한 단어들이 나타나며 표현 양식도 학생들에 따라서 매우 다를 수 있기 때문이다. 교사가 학생 개개인이 작성한 내용의 의미를 이해하여 즉각적인 피드백을 제공하기 위해서는 빠른 학습분석(learning analytics)이 필수적이겠지만, 이는 교사가 투입하여야 하는 시간과 노력을 감안할 때 현실적으로 쉽지 않은 일이다. 이는 연간 실시 횟수가 적은 편인 총괄평가에서도 마찬가지로서, 서술형 평가는 교사에게 큰 부담으로 작용한다고 알려져 있다. 결국 교수학습 상황에서 매번 교사에게 서술형 평가에 대한 분석을 요구한다면 교사의 업무부담은 매우 높아질 것이다. 이에 대한 대안으로 제시되는 것이 컴퓨터 자동채점이다.

컴퓨터(인공지능)를 활용한 서술형 평가 자동채점 시스템은 교수학습에서 요구되는 진단 및 형성평가를 서술형으로 바꾸고자 하는 교육적 요구에 힘입어 많은 연구진들에 의하여 개발되고 있다. Educational Testing Service (ETS)에서 개발하는 C-Rater-ML (Liu et al., 2016) 뿐만 아니라 자연선택 개념에 관한 자동채점연구(Ha et al., 2011; Moharreri et al., 2014; Nehm, Ha, & Mayfield, 2012), 산・염기 개념에 관한 자동채점연구(Haudek et al., 2012), 통계 개념에 관한 자동채점연구(Kaplan et al., 2014), 광합성 개념에 관한 자동채점 연구(Weston et al., 2015) 등의 다양한 연구가 그 사례에 해당한다. 최근에는 논증에 관한 자동채점연구(Mao et al., 2018; Zhu et al., 2017)도 이루어지며, 컴퓨터 자동채점을 활용하여 학생들에게 학습을 촉진하는 형태의 프로그램의 개발 또한 상당수 진행되고 있다.

위에서와 같이 컴퓨터 또는 인공지능을 활용한 자동채점연구가 널리 이루어지고 있지만, 여전히 자동채점에 대하여 많은 오해와 우려가 있는 것도 사실이다. 그중 대표적인 것이 인공지능에 대한 부정적 인식, 정확도와 오류, 정확하지 않은 예측에 대한 부작용, 문장에 포함된 복잡한 의미를 확인하는 기술적 한계 등이다.

인공지능에 대한 부정적 인식은 인간의 편리함을 위하여 만든 인공지능이 다시 인

간을 평가하는 평가자의 위치에 있음에 대한 불편함, 인공지능이 인간의 역할을 대체하여 직업적 선택을 줄일 수 있을 것이라는 인식 등을 포함한다. 예컨대 인공지능이 향후 교사를 대체하여 학생들을 지도할 수 있을 것이며 그에 따라 교사라는 직업이 사라질 수 있다는 우려이다. 그런가 하면 인공지능의 정확도와 채점의 한계에 관한 우려도 많다. 'Professionals Against Machine Scoring Of Student Essays In High-Stakes Assessment'라는 제목의 온라인 청원이 대표적이다.[2] 여기서는 전문가들이 고부담시험에서 자동채점을 반대하는 의견을 내고 있다. 또한 2018년 호주의 학력평가고사(NAPLAN, National Assessment Program-Literacy and Numeracy)에서도, 학생들의 에세이를 컴퓨터로 자동평가하려는 시도에 대해 창의성 등과 같은 고차원적 사고보다는 내용이 없이 장황한 에세이에 더 많은 점수가 제시될 수 있다는 우려가 있었다. 예컨대 기존의 데이터에 기반하여 훈련된 자동채점모델은 새롭고 독창적인 아이디어가 있는 학생 답안에 더 높은 점수를 주지 못할 수 있으며, 컴퓨터가 어떤 자질(feature)에 더 높은 점수를 주는지에 대해 학생들이 인식하게 되면 시험에 높은 점수를 받을 수 있도록 컴퓨터를 속일 수 있는 기술을 찾아낼 수도 있을 것이다. 또한, 채점에는 오류가 반드시 따르기 마련인데 이에 대한 책임 소재도 중요한 논점이 될 수 있다.

이와 같은 우려는 충분히 이해 가능한 것들이지만, 교수학습을 위한 즉각적인 평가도구, 학습자 맞춤형 수업, 학습자의 자기주도적 학습, 교사의 업무부담 감소 등 다양한 목적을 위하여 평가에서 인공지능의 역할은 늘어날 수밖에 없다는 것도 전문가들의 의견이다(Zhu et al., 2017). 이러한 현재 시점에서는, 자동평가 및 채점과 관련하여 인공지능이 가지고 있는 기술적 수준과 한계를 이해하고, 사용자들의 부정적 인식을 최소화하면서 이를 교실 현장에서 그 목적에 맞게 활용할 수 있는 방안을 탐색하는 일이 필요하다.

본 연구의 목적은 첫째 현재 진행되고 있는 인공지능 자동평가의 현재 기술과 여러 문제점들을 고찰하고, 둘째 단기적으로 우리가 어떻게 인공지능을 학교 현장에서 평가도구로 사용할 수 있는지 그 범위를 탐색하며, 마지막으로 최근에 개발되고 있는 다양한 인공지능 기술들을 활용하여 장기적으로 인공지능 자동평가를 어떻게 활용할 수 있는지 방향을 탐색하는 것이다. 그러므로 본 논문은 일종의 리뷰 논문에 해당하며, 이를 위하여 연구자들은 논리적 문헌 고찰(logical literature review)을 연구 방법으로 택하였다. 살펴보고자 하는 영역이 상당히 전문적인 연구 분야로서 관련 문헌들이 상대적으로 적은 편이므로, 문헌을 체계적으로 표집하기보다는 해당 분야에서 중요하게 여겨지는 사례들을 주로 제시하게 되었다. 이에 먼저 Ⅱ장에서는 인공지능 기반 자동평가

2) http://humanreaders.org/petition (2019.12.31. 확인)

와 관련된 중요 연구 사례들을 통시적으로 파악하면서 그 흐름을 정리하였다. 이러한 통시적 분석은 선구적으로 이루어졌던 해외 사례들에 비추어 국내의 관련 연구들이 어떻게 이루어지고 있는지를 자연스럽게 파악하게 해줄 것이다. 다음 Ⅲ장에서는 테크놀로지 측면 및 교육평가 측면의 관점에서 인공지능 자동평가의 한계와 도전을 어느 정도 심도 있게 살펴본다. Ⅳ장에서는 지금까지 해당 기술이 주로 활용되어 온 서술형 문항을 중심으로 인공지능 자동평가의 단기적 활용 방안을 살펴보며, 마지막으로 Ⅴ장에서는 최근 급속도로 발전하고 있는 인공지능 기술을 고려하여 그 너머의 가능성을 제시한다. 이 때, 자동평가 시스템의 사례로서 학생의 정확한 개념 이해를 중요시할 뿐 아니라 관련 연구가 비교적 많이 이루어진 편인 과학 교과와 관련된 문헌들을 상당수 언급할 것임을 밝혀 둔다.

Ⅱ. 인공지능을 활용한 서술형 문항 자동평가의 현재

인공지능을 활용한 자동평가 연구를 본격적으로 고찰하기에 앞서, 과연 그것이 무엇을 의미하느냐를 먼저 짚어볼 필요가 있을 것이다. 인공지능을 활용한 서술형 자동 평가는 학생의 답안에 대하여 이론상으로 상정되는 참 점수(true score)에 가깝게 채점을 수행하겠다는 거시적 목표를 지니며, 이 자체는 기존에 이루어지던 교육평가 및 그 이론과 본질을 같이한다. 다만, 자연어로 작성된 답안을 읽고 처리하며 채점하는 지능적 행위를 인간이 아닌 인공지능이 수행하도록 한다는 점이 다를 뿐으로 이해할 수 있다. 그러므로 그 적용 또한 후술할 바와 같이 기존에 이루어지던 평가제도 및 체제 안에서 이루어지는 것이 보통이다. 예컨대, 인공지능 기반 자동평가는 TOFLE 등 이미 상용화되어 잘 알려진 시험이나 국내의 전국 단위 학력평가 등에서 적용 및 시험되고 있는 것이다. 다만 인공지능이 마치 인간과 같은 지능적 행위를 할 수 있기를 지향한다는 점에서, 평가라는 고도의 지적 작업을 인공지능이 수행할 수 있는지를 검토하고 그것이 향후 교육에 가져올 가능성을 예견하는 일이 관련 연구들의 주요 논지가 된다.

서술형 응답을 컴퓨터를 활용하여 자동으로 채점하는 방법은 Ellis Page에 의하여 시작되었다(Page, 1966). Page는 영어 교사로서의 경험과 전산 언어학, 인공지능 등의 학문을 융합하여 영어 에세이를 자동채점하는 PEG software의 초기 버전을 완성하였다. PEG는 영어 에세이의 수준을 확인할 수 있는 중요한 자질(feature)을 확인하고, 그것들을 자동으로 추출하는 시스템으로서, 자질들과 전문가 채점으로 생성된 점수와의 관계식을 활용하여 채점이 필요한 에세이의 대략적인 점수를 예측하는 형태이다. PEG

software가 상업적으로 성공하던 1980년~2000년 사이에 PEG software와 유사한 많은 상업용 에세이 자동채점 프로그램들이 개발되어 활용되었으며, Intelligent Essay Assessor, IntelliMetric, e-rater 등이 그 대표적 사례이다.

2000년대 들어서는 과학 문항에 관한 학생들의 서술형 응답을 자동채점하는 엔진들이 개발되기 시작하였다. 대표적으로 Educational Testing Service에서 개발한 C-rater ML이다. C-rater ML을 활용하여 중학생들의 과학 영역 평가를 실시한 결과 인간 채점과 상응하는 수준의 신뢰도가 있었다는 보고가 있다(Liu et al., 2014; Liu et al., 2016). C-rater ML의 경우 상업적 목적으로 개발된 자동채점 모델이지만, 최근에는 상업적 목적이 아닌 연구 집단에 의하여 개발되는 인공지능 자동채점 모델 연구도 활발히 이루어지는 추세이다. 미국과학재단의 지원의 Automated Analysis of Constructed Response 프로젝트는 대표적인 서술형 자동채점 연구의 집단 연구모임이다. 이 연구 집단의 웹페이지 명인 'beyondmultiplechoice.org'[3]에서 확인할 수 있듯이 학습을 위한 평가(진단 및 형성 평가)에서 선택형 평가(multiple choice assessment)를 넘어서서 서술형 평가를 활용해야 함을 강조하고, 그것이 강의실에서 효율적으로 활용될 수 있도록 서술형 평가의 자동화를 연구하고 있다. 실제로 산과 염기에 관한 학생들의 개념(Haudek et al., 2012), 광합성 개념(Weston et al., 2015), 체중 감소에서 나타나는 학생들의 혼합 개념(Sripathi et al., 2019) 등 다양한 과학개념에 관한 자동채점 모델 개발이 이루어져 왔다.

Nehm과 그의 동료들이 수행한 일련의 연구는 학생들의 과학개념에 관한 형성평가 문항 개발로부터 웹기반의 자동채점포털 개발에 이르기까지의 전 과정을 자세히 보여준다. 먼저는 상황기반추론과 혼합개념에 관한 이론적 모델에 기반하여 학생들의 자연선택개념에 대한 이해수준을 확인할 수 있는 검사도구(Assessing COntextual Reasoning about Natural Selection)가 개발되었다(Nehm & Ha, 2011; Nehm et al., 2012; Opfer et al., 2012). 다음으로는 자연선택 개념에 관한 학생들의 응답을 자동채점하는 모델에 대한 개발이 이루어졌다. 그 초기 모델은 SPSS Text Analysis 3.0을 활용하여 훈련한 것이었다(Nehm, & Haertig, 2012). 곧, 학생들의 응답을 수집하고 전문가 채점을 한 뒤 전문가 채점에서 사용되는 학생들의 언어(단어)를 통해 라이브러리와 표현 규칙 등을 입력하여 컴퓨터가 학생들의 점수를 예측하게 하였다. 한편 그 이후에 개발된 채점 모델은 기계학습을 활용하였다(Ha et al., 2011; Nehm, Ha, & Mayfield, 2012). 학생들의 응답과 전문가 채점자료를 입력하면 응답 속에 나타나는 다양한 자질들과 전문가의 채점 사이의 관계를 바탕으로 채점모델을 훈련시키는 것이다. 여기서 기계학습 방법은 라이브러리를 구축하거나 표현의 규칙을 입력하는 등의 추가적인 노력이 요구되지 않기

3) https://beyondmultiplechoice.org/ (2019.12.31. 확인)

때문에 효용이 높았다. 한편 기계학습 방법을 활용하여 개발된 자동채점모델은 100명 이상의 학생들의 면담 자료를 비교하는 연구에서 그 효용성이 드러나기도 하였다(Beggrow et al., 2014). 자연선택개념에 관한 선택형 평가도구, Assessing COntextual Reasoning about Natural Selection(ACORNS)의 전문가 채점과 컴퓨터 채점, 임상 면담을 통해 확인한 학생들의 자연선택 개념 총 4가지를 비교한 결과, 컴퓨터 채점이 임상 면담 결과와 가장 높은 수준에서 일치하였다. 이후 자동채점모델들은 웹에서 자동으로 학생 응답을 분석하고 그 결과를 보고하는 형태의 포털로 구성되어 강의실에서 형성평가의 도구로 활용될 수 있는 형태로 개발되었다(Moharreri et al., 2014).

인공지능을 활용한 자동평가는 학생들의 개념을 빠르게 확인하고 그 결과를 확인할 수 있는 장점이 있기 때문에 강의실 내에서 즉각적인 피드백을 가능하게 해준다. 그러므로 이와 같은 장점을 활용하여 학생들의 과학개념 수준을 이끌어 주는 학습 도구로서 활용이 가능하다. Zhu, Liu, et al.(2020)은 기후변화에 관한 교수 학습에서 인공지능 자동채점을 활용하여, 학생들의 논증을 즉시 분석하는 시스템과 그 분석 결과를 바탕으로 학생들이 논증을 수정할 수 있는 피드백을 제공하는 시스템을 구성하였다. 이 시스템을 통해 학생들은 자신의 생각을 피드백에 따라 지속적으로 변화시킨다. 연구자들은 학생들의 로그파일(log file)을 분석하여 어떻게 논증을 피드백에 근거하여 변화시켰는지 확인하였고, 인공지능 자동채점 시스템이 학생들의 논증 수준의 발달을 이끌 수 있음을 실증적으로 증명하였다. 해당 연구는 논증(argumentation)에 관한 자동채점과 즉각적인 피드백에 근거한 학생들의 논증 수정에 관한 Lee et al.(2019), Mao et al.(2018), Zhu et al.(2017)의 연구 결과들을 바탕으로 이루어진 것이었으며, 인공지능 자동채점이 학생들의 학습발달을 이끄는 도구로써 활용될 수 있음을 실증적으로 확인한 연구는 이전에 Gerard et al.(2016), Tansomboon et al.(2017)에서도 진행된 바 있었다. Gerard et al.(2016)은 학생들이 작성한 에세이, 다이어그램, 그림 등을 자동으로 분석하는 모델을 활용하여 학생들의 발달을 지도하는 시스템을 개발하였다. 이 시스템은 교사의 협력자가 될 수 있었으며, 학습의 도움이 필요한 학생들이 교사의 도움 없이 컴퓨터의 안내로 학습을 발달시킬 수 있음을 확인하였다. 그런가 하면 Tansomboon et al.(2017)은 자동채점에 기반한 자동화된 시스템을 어떻게 디자인하는 것이 더 효율적인지 확인하는 연구를 수행하였다.

우리나라에서도 인공지능 활용 자동평가에 관한 연구가 이루어지고 있다. 상술한 바와 같은 에세이에 대한 인공지능 자동채점 연구의 역사에서 짐작할 수 있듯이, 국내에서 진행되는 자동채점 선행연구의 상당수는 영어 작문채점에 관한 것이다(e.g. 김지은, 이공주, 2007). 영어 에세이의 자동평가에 관한 많은 선행연구와 그 채점 모델 및 시스

템이 누적되어 있기 때문에, 영어 작문 평가는 자동채점 연구의 진입장벽이 상대적으로 낮은 편이라고 할 수 있다. 과학 개념에 관해서는 하민수(2017)가 미국에서 개발된 자동채점 알고리즘을 활용하여 한국 학생들의 자연선택 개념을 평가한 바 있다. 한국어로 된 서술형 응답을 영어로 자동 번역하여 채점 자료로 사용하였고, 일부 과학 개념 평가에서는 효용성이 있음을 입증하였다. 한편 한국교육과정평가원 소속 연구자들은 전국 단위의 대규모 학력 평가에서 한국어 평가 문항 자동채점 연구들을 지속적으로 보고하고 있다. 성태제 외(2010)는 학업성취도 평가에 사용된 서답형 문항 중 일부에 대해 컴퓨터 채점 방안을 탐색하고 그 실현 가능성을 확인한 바 있다. 노은희 외(2014)는 학생들의 응답을 유사한 정도에 따라 분류하여 빠르게 평가하고, 일부의 평가 자료를 활용하여 인공지능을 학습시킨 뒤 다시 채점하는 방법을 활용하여 대규모 평가 자료를 빠르게 채점하는 방안을 연구하였다.

인공지능을 활용한 자동평가 연구들을 살펴보면 빠른 분석을 통해 채점의 효율성을 높이고, 강의실 내에서의 즉각적인 피드백을 제공하기 위한 형성평가용 시스템이 개발되고 있음을 확인할 수 있다. 더욱이 최근 연구 동향에서는 인공지능의 빠른 피드백을 바탕으로 학생들의 학습을 지원하는 도구로서의 기능이 연구되고 있다. 학습도구로서의 인공지능 자동평가 활용은 고부담 평가에서의 인공지능 활용에 대한 거부감 최소화, 교사의 업무 부담 감소, 학생들에 대한 맞춤형 학습발달 등의 많은 장점이 있다. 한편 지금까지 다양한 컴퓨터 튜터 플랫폼이 오랫동안 개발되어 왔으며 디지털 교과서와 같이 교실 환경의 디지털화가 빠르게 진행되어 왔음을 고려한다면, 학습지원도구로서의 인공지능 자동평가 활용이 가질 유용성은 더욱 증대될 것이다.

Ⅲ. 인공지능 서술형 평가의 한계와 도전

지금까지의 장들에서는 인공지능 기반의 서술형 문항 자동평가 시스템이 활용될 수 있는 가능성을 연구 사례를 중심으로 살펴보았다. 본 장에서는 선행 문헌들에서 간접적으로 드러나 있거나 다루지 않은 한계와 도전을 다루되, 특히 자동채점 시스템 내부에서 작동하는 알고리즘 및 평가 이론을 근래의 테크놀로지와 함께 보다 심도 있게 살펴보고자 한다.

1. 자연어 처리의 기술적 한계

인공지능을 활용한 서술형 평가가 갖는 한계는 우선 기계의 자연어 처리(Natural Language Processing, NLP) 혹은 자연어 이해(Natural Language Understanding, NLU)가 완벽하지 않고, 또 어떤 면에서는 완벽할 수 없다는 점에서 기인한다.[4] 이는 통계적 언어 모델이 본질적으로 언어 단위(문자 혹은 단어)의 시계열적 예측 모델이기 때문이다. 특정 시점 t까지의 언어 단위가 주어졌을 때 $t+1$ 시점에서 어떤 언어 단위가 등장할 것인지를 정확히 예측하기란, 규칙 기반(rule-based) 모델에서는 매우 어려운 일이다. 인간이 사용하는 자연어는 그 특성상 구문론적(syntax)으로나 의미론적(semantics)으로 무한한 변용이 가능하기 때문이다. 이처럼 복잡다단한 인간 응답자의 언어적 표현을 온전히 이해하기란 같은 인간 채점자에게도 쉽지 않은 일이며, 인공지능이 이와 유사한 기능을 하도록 하기 위해서는 매우 많은 모수(parameter)를 활용하는 데이터-기반(data-based) 모델을 구축해야 한다. 예컨대, 비영리 인공지능 연구기관 OpenAI에서 근래에 공개한 언어 모델 중 하나인 GPT-2는 최대 48개 층으로 이루어진 신경망에서 무려 15억 개 이상의 모수를 추정하며, 이를 학습시키는 데 필요한 데이터는 웹 크롤링(crawling)[5]을 이용하여 수집하고 정제한 40GB 이상의 텍스트 데이터 해당한다(Radford et al., 2019). 이렇게 학습된 언어 모델은 특정한 벤치마크(benchmark) 데이터[6] 혹은 일반적인 자언어 생성(Natural Language Generation, NLG)에 대하여 뛰어난 성능을 보이면서 딥 러닝 기반 언어 처리 기술의 발달을 가져오고 있으나, 영역-특수적인 개념을 정교하게 판단하여야 하는 서술형 평가에 이를 단순 적용하기에는 무리가 있는 것이 당연하다. 그렇다고 하여 영역-특수적이며 응답자-특수적인 자연어 데이터를 수많은 개별 교과 및 개념에 따라 충분히 모으는 것도 녹록치 않을 것임을 쉽게 예상할 수 있다.

또한 자연어 처리의 경우, 많은 자원이 투자되며 연구가 이루어져 온 영어와, 한국어를 비롯한 기타 언어의 모델들 사이의 기술적 격차가 결코 작지 않다. 특히 한국어의 경우 일종의 교착어에 해당하므로 조사, 부사, 접미사, 접두사 등이 존재하며, 이와 함께 띄어쓰기를 고려한다면 언어 단위를 정확히 토큰화(tokenize)하기가 상당히 어려운 편이다. 예컨대, 한국어 자연어 처리를 위해 빈번히 사용되는 형태소 분석기들조차 “아

4) 여기서 ‘자연어 처리’ 혹은 ‘자연어 이해’란 인간이 사용하는 자연어를 컴퓨터가 이해할 수 있는 데이터의 형태로 처리하는 일과 함께 이에 대한 의미론적 추론을 가능하게 하는 일을 의미한다.

5) 웹에서 원하는 정보를 자동으로 추출 및 수집하는 기법이다.

6) 인공지능 및 학습 알고리즘의 성능을 비교하기 위하여 표준적으로 활용하는 데이터들이다.

버지가방에들어가신다"는 문장을 기계적 규칙이나 확률에 근거하여 정확하게 토큰화하기란 쉽지 않은 것이다. 백영민(2017)은 이러한 어려움을 지적함과 동시에, 한국어의 경우 폭넓게 받아들여지는 불용어[7] 목록이나 어근 동일화 작업에 사용되는 알고리즘 또한 없음을 언급한 바 있다(pp. 138-139). 말하자면, 현재까지 한국어 자연어 처리를 위한 형태소 분석기의 통일된 규준은 거의 존재하지 않는 것과 마찬가지라고 할 수 있다.

단어[8]를 정확하게 토큰화하였다고 가정한 이후, 근래의 텍스트 마이닝에서는 기계가 해당 단어의 의미를 실제로 이해하듯 연산할 수 있도록 단어를 임베딩(embedding)시키는 작업이 요구된다. 단어 임베딩은 본래 Mikolov et al.(2013)에 의해 제안된 Word2Vec에서부터 출발한 것으로, 기본적인 아이디어는 여러 단어들의 출현 빈도와 상호간의 순서(sequence)를 고려하여 각각을 원래 문서-단어 행렬(document-term matrix) 보다 낮은 차원의 벡터 공간에서 표현하자는 것이다. 이러한 임베딩 과정을 거치게 되면 각각의 단어들은 의미 벡터로서 표현되는데, 의미 벡터 간 거리가 가까울수록 해당 단어들은 유사한 의미를 지닌 것으로 여길 수 있다. 예컨대, 단어(문서) 벡터 $\vec{a}$와 $\vec{b}$를 내적(inner product)한 값을 각각의 절댓값으로 나누어주면 두 벡터가 이루는 각 θ에 대하여 $-1 \sim +1$ 범위에서 결정되는 $\cos\theta$ 값을 얻게 되는데, 이는 한 벡터를 다른 벡터에 대하여 투영(projection)하였다는 의미를 지니므로 그 값이 +1에 근접할수록 두 벡터가 유사하다. 적어도 이러한 단어 임베딩이 이루어진 이후에야 인공지능 모델은 텍스트의 의미를 진정 이해하기 시작하였다고 할 수 있는 것이다. 그런데, 이러한 단어 임베딩의 문제는 중의적인, 혹은 전혀 다른 어의를 갖는 단어들이 동일한 의미 벡터로서 표현되는 경우가 발생한다는 것이다. 예컨대, 기존의 단어 임베딩에서는 '은행'이라는 토큰이 금융 업무를 보는 장소를 의미하는지 혹은 식물의 일종을 의미하는지를 명확히 구분하지 못하고 이를 같은 것으로 처리하게 된다. 만약 서술형 평가에서 이와 유사한 현상이 발생한다면, 인공지능은 학생 응답을 올바른 맥락에서 이해하지 못하고 잘못된 채점 결과를 산출할 수 있다.

물론 2018년 11월에 Google에서 공개한 BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2018) 이후의 딥 러닝 기반 모델들은 토큰의 전후 맥락을 고려할 수 있는 양방향성(bidirectionality)을 고려하면서 임베딩 은닉층(embedding layer)을 여럿 설정함으로써 이러한 문제들을 해결하고 있는 것이 사실이다. BERT는 일반적인 언어 모델로 학습된 여러 은닉층들 위에서, 출력층에 가까운 일부 은닉층을 특정한 과제

7) 텍스트 분석에서 활용하지 않는 사소한(trivial) 용어들

8) 보다 정확히는 형태소 등의 언어 단위이지만, 여기서는 편의상 단어로 지칭한다.

를 위한 미세 조정(fine-tuning)을 거쳐 사용할 수 있는 특징을 지닌다. 한국전자통신연구원(ETRI)에서는 한국어 데이터를 활용하여 구현한 ETRI BERT를 공개하기도 하였으므로,[9] 한국어 자연어 처리에 있어서도 기존 단어 임베딩의 약점을 보완할 수 있는 가능성이 있다. 이러한 경우, 결국 특정 교과 및 개념에 대하여 학생이 가진 개념의 정오를 판별하기 위한 자연어 이해 모델을 구축하기 위한 데이터를 수집하고, 이를 일반적으로 사용되는 언어에 대하여 미리 학습된(pre-trained) 모델 위에서 미세 조정하는 일이 관건이 된다. 다만 BERT 등의 기반 모델들은 학습 과정에서 신문 기사, 서적의 일부 등 정형화된 패턴의 문장 데이터를 사용하여 서술형 평가에서의 학생 응답과는 데이터의 특성이 다를 수 있다는 점, 그리고 특정 교과 및 개념에 특수적인 학생 응답 데이터는 개별 연구자들에게 절대적으로 부족하다는 점 등이 여전한 이슈가 될 수 있다.

2. 채점 신뢰도 문제

인공지능 자동채점에서 중요한 이슈 중 하나가 채점 신뢰도이다. 본 장의 다른 절에서 후술할 내용에서와 같이 고부담 시험에서 발생하는 채점 오류는 심각한 문제가 될 수 있으며, 이로 인하여 발생하는 문제에 대하여 누가 책임을 질 것인가에 대한 쟁점도 부각된다. 물론 고부담 환경이 아닌 학습 과정에서 발생하는 채점 오류 역시 문제가 될 수 있다. 만약 교사가 잘못된 평가 정보를 바탕으로 학습자의 수준을 판단하여 수업을 기획한다면, 불필요한 수업을 계획하거나 또는 중요한 정보를 가르치지 않고 지나칠 수 있다. 학습자의 학습을 돕는 보조교사로서의 인공지능의 경우에도 잘못된 피드백은 학습자가 혼동을 느끼거나 틀린 개념을 학습하도록 하게 될 수 있다.

이에, 인공지능 자동평가에서 채점의 오류를 줄이기 위한 다양한 전략들이 연구되고 있다. 가장 먼저 채점의 오류가 전체 점수에 큰 영향력이 없도록 점수 산정의 수준을 높이는 것이다. Ha & Nehm(2016a)은 채점 정확도에 관한 여러 수준을 제시하였다. 예를 들어서 자연선택개념을 측정하는 문항 4개를 교사가 사용하고, 각 문항은 9개의 개념 요소를 근거로 0과 1점으로 채점하며 이를 바탕으로 각 문항의 종합점수를 만드는 평가가 있다고 하자. 인공지능은 학생들의 각 응답에서 9개의 개념 요소의 유무를 판단한다. 인공지능은 1개 문항마다 9번의 판단을 내리며, 4개 문항에서는 총 36개의 판단을 바탕으로 최종 점수가 판정된다. 이 때 각각의 판단에서는 오류가 발생할 수 있지만 문항 당 점수는 9개의 개별 점수의 합산이며, 학생의 점수는 36개의 개별 점수의

9) http://aiopen.etri.re.kr/service_dataset.php (2019.12.31. 확인)

합산이기 때문에 오류의 영향력이 감쇠된다. 또한, 교사가 학생 개개인의 점수가 아닌 수업 전과 후와 같은 학급당 점수의 변화에 관심이 있다고 가정하면 한 학급의 전체 점수에서 오류가 가지는 영향력은 훨씬 줄어들 것이다. 결국 평가 점수를 어느 수준에서 활용하는가에 따라 채점 오류의 영향력은 달라질 수 있으며, 평가의 신뢰도에 따라 그 활용 범위를 달리하는 것으로 잠재적인 문제를 해결할 수 있다.

다음으로 인공지능 채점모델을 훈련할 때 정밀도(precision)와 재현율(recall) 중에서 평가 오류의 부작용이 최소화 되는 방향으로 채점모델을 훈련시키는 것이다. 인공지능 자동채점의 신뢰도를 판단하는 데는 다음과 같이 크게 5가지의 척도가 활용된다. 예를 들어서 전문가가 판단한 점수와 인공지능의 판정에 대한 혼동행렬(confusion matrix)이 있다고 가정하면, 정확도(Accuracy)는 전체 예측(0점과 1점) 중에서 실제 정확하게 예측된 확률을 의미하며, 정밀도(Precision)는 1점(개념이 있다)이라고 판단한 인공지능의 예측 중에서 실제 맞는 비율을, 재현율(Recall)은 1점(개념이 있다)라고 한 전문가 판단 중에서 인공지능이 성공적으로 예측한 정도를 의미한다. F1(F1-measure)은 정밀도와 재현율의 평균값이며, 카파(kappa)는 채점자간 평가가 일치하는 것을 확률로 표현한 값이다. 이 중에서 가장 널리 활용되는 것은 아마도 카파일 것이지만, 정밀도와 재현율 역시 중요한 정보이다. 인공지능 자동채점 모델이 정밀도가 낮지만 재현율이 높은 경우에는 학생의 응답에 특정개념이 나타나지 않았음에도 개념이 있다고 예측한 경우가 많을 수 있다. 반대로 재현율은 낮지만 정밀도가 높은 경우에는 반대로 학생들의 응답에 특정개념이 있다고 판단한 경우가 매우 적을 수 있다. 곧, 재현율이 높을 경우 학생 점수가 과대평가된 경우가 많을 것이며, 반대로 정밀도가 높은 경우 학생 점수가 과소평가된 경우가 많을 것이다. 이 중, 인공지능 자동평가에서 일반적인 학생들이 민감하게 반응하는 것은 후자의 경우이다. 이런 경우에는 학생들에게 적지 않은 혼동을 줄 수 있으며 인공지능 시스템에 대한 신뢰가 떨어질 수 있다. 학생들을 과대평가할 경우에도 마찬가지로 문제가 될 수 있지만, 인공지능 자동평가의 경우 채점의 부담이 낮기 때문에 많은 문항을 학생에게 제공할 수 있으며, 이에 따라 일부 문항에서 나타나는 채점 오류의 영향이 낮아 드물게 나타나는 과대평가는 큰 문제가 되지 않을 수 있다. 따라서 학습자 피드백을 위한 자동평가 모델을 개발 과정에서 인공지능 자동채점 모델을 훈련할 때, 재현율이 상대적으로 높게 나오도록 훈련하면 상기한 문제점을 일부 해결할 수 있다. 물론, 특수한 상황에서는 예외적인 경우도 발생할 것이다. 예컨대 일반적인 학생들이 아닌 영재를 선발하는 시험에서는 전체 학생 평균보다 훨씬 더 높은 성취도를 지닌 응답자들 내에서의 변별이 중요해지므로, 재현성보다는 정밀도가 더 높은 채점 모델이 필요할 수 있다.

마지막은 인공지능이 채점의 정확도를 확인할 수 있는 다른 방법을 고안하고, 채점의 정확도가 낮을 것으로 예측되는 판단이 내려졌을 때 이를 교사에게 알리는 것이다. 이런 경우 교사는 채점의 정확도가 낮은 경우들만 확인하여 채점의 부담을 최소화할 수 있다. 이와 같이 인공지능의 채점이 정확한지 여부를 확인하는 방법에 대하여 Ha & Nehm(2016b)은 두 가지 방법을 제안한 바 있다. 첫 번째 방법은 자동채점의 확률을 채점 정확도를 판단하는데 활용하는 것이다. 인공지능은 학생 응답 내의 정답 유무를 평가할 때 확률적인 방법으로 접근을 하는데, 판단하는 확률이 모호할수록 채점이 정확하지 않다고 판단한다. 인공지능이 학생 응답에 특정 개념이 있을 확률을 0.1, 0.4, 0.6, 0.9로 판단했다고 가정할 때, 0.1과 0.4는 기준으로 설정한 0.5에 비하여 낮기 때문에 특정 개념이 없다고 보고되며, 0.6과 0.9는 0.5에 비하여 높기 때문에 특정 개념이 있다고 보고된다. 이때 0.4와 0.6은 0.1과 0.9에 비하여 0.5에 더 근접하기 때문에 더 모호한 예측이라고 볼 수 있다. 이런 경우 채점의 정확도가 낮다고 판단하는 것이다. 여기서 정확도가 낮은 문항의 경우 교사가 채점을 수정한다는 것을 전제로 수행한 시뮬레이션을 도입할 수 있다. '변이 개념'의 평가의 경우 인공지능에게 완전히 의존할 경우 일치도(kappa)가 0.861인 반면 교사가 학생응답 중에서 정확도가 낮을 것으로 예상되는 9.2%만 수정하여도 일치도가 0.947까지 높아지는 것을 확인할 수 있었다. 또한 '차별적 생존'의 경우 인공지능에게 의존할 경우 일치도가 0.701인 반면 교사가 정확도가 낮을 것으로 예상되는 11.9%만 수정하여도 일치도가 0.808로 거의 완벽한 수준의 일치도로 올라갈 수 있었다. 두 번째 방법은 여러 채점모델을 만들고 채점모델 간 일치도를 확인하는 것이다. 만약 채점모델들이 생성한 결과가 일치하지 않는다면, 아마도 채점이 부정확할 가능성이 있다고 판단하여 교사가 개입하는 것이다. 이 방법을 적용하면 '변이 개념'의 평가의 경우 교사가 채점모델간 일치도가 낮은 12.4%만 수정하여도 일치도가 0.923까지 높아지고, '차별적 생존'의 경우 교사가 채점모델간 일치도가 낮은 14.0%만 수정하여도 일치도가 0.831로 높아졌다. 이와 같은 방법은 인공지능과 교사의 협력 모델을 제시하는 것으로서, 교사가 다시 확인하지 않아도 될 정도로 확실한 응답은 인공지능에게 맡기고 인간의 개입이 필요한 부분만 교사가 확인하여 교사의 업무 부담을 줄이면서 인공지능을 효율적으로 활용할 수 있는 방안이다.

3. 인간 사용자의 신뢰와 인위적 속임

한편 인공지능 기반 서술형 자동채점 시스템을 학교 현장에서 활용할 때, 인공지능의 문항 채점 결과 신뢰도(reliability) 이전에 해당 시스템에 대한 인간 사용자의 신뢰

(trust)가 문제가 될 수 있다. 인공지능 자동채점 시스템이 신뢰할 만한 결과를 산출한다고 하더라도 그것이 학생, 교사, 행정가, 학부모들에게 신뢰받지 못할 경우에는 교실 현장에서 실질적으로 사용되는 데 지장을 불러올 수 있기 때문이다. 이러한 현상이 지속될 경우, 인공지능을 활용한 서술형 문항 자동채점이 가져다 줄 수 있는 유익함이 퇴색되어 어떤 면에서 단순반복적인 채점 업무에 교사가 여전히 많은 자원을 소모하는 결과를 낳을 수 있다.

이러한 낮은 신뢰는 학생들이 문항에 대한 응답을 성실하게 하지 않거나 심지어는 왜곡된 답안으로 시스템을 인위적으로 속이기까지(cheating) 하는 데까지 영향을 미치는 것으로, 자동채점 시스템이 매긴 점수가 학생의 참 점수(true score)에서 벗어나게 된다는 점에서 채점 타당성(validity) 문제의 하나로 받아들일 수 있다. 예컨대 영어 에세이 자동채점 시스템의 고전에 해당하는 PEG의 경우(Page, 1966) 당시 하드웨어의 계산 능력 문제로 에세이에 내포된(intrinsic) 응답자의 글쓰기 역량을 단어의 평균 길이, 에세이의 길이, 쉼표의 수, 전치사의 수, 일반적이지 않은 단어의 사용 빈도 등의 변수로 간접적으로 측정할 수밖에 없었는데, 학생들이 이를 역이용하여 단순히 더 긴 에세이를 쓰는 등의 방식으로 점수를 올리는 일이 가능하다는 비판을 받았다(Kukich, 2000). 이렇듯 자동채점 시스템을 속이는 일이 가능한 것은 PEG보다 이후에 개발되어 자연어 처리 및 훈련 과정에서 보다 복잡한 알고리즘을 사용하는 e-rater에서도 마찬가지였다. ETS에서 실험한 결과, 언어학 등에 숙련된 전문가들은 초기 버전 e-rater의 내부 구동 방식을 숙지한 후에 작성한 에세이가 인간 채점자가 부여한 점수보다 e-rater에서 높은 점수를 받도록 속일 수 있었다(Powers et al., 2002). 이 외에 비근한 예로 답안에서 채점 키워드에 해당할 것으로 생각되는 단어만을 단순 나열하거나 의도적으로 부정어(e.g. ‘아니지만’, ‘아니다’)를 과다 사용하여 시스템의 채점 결과에 혼동을 주는 사례를 얼마든지 생각할 수 있다. 곧, 이러한 일이 오늘날에도 발생할 수 있다는 우려가 과장된 것만은 아니다.

하지만 서술형 문항 자동채점 시스템의 적절성 및 효용성에 대한 심도 있는 평가가 이루어지면서, 해당 시스템들은 잠재 의미 분석(latent semantics analysis) 등의 기법을 통해 응답의 표면적 정보보다 더 깊은 의미를 탐색하게 되어 알고리즘 측면에서의 진전이 일어났으며 점차적으로 그에 대한 대중성과 신뢰성(credibility)이 증가하게 된 것도 사실이라고 하겠다(Yang et al., 2002; cf. Raczynski, & Cohen, 2018). 이와 함께 인위적인 속임수에 대한 데이터가 증가하면서 이를 패턴화하고 그에 대응하는 방안을 도출하는 일까지도 가능하게 되었다. 실제로 IntelliMetric(Rudner, Garcia, & Welch, 2006) 이나 Intelligent Essay Assessor (IEA) (Lochbaum et al., 2013) 등의 에세이 자동채점 시스템이 이

를 성공적으로 검출한 사례들이 이미 보고된 바 있다.

일반적으로 컴퓨터 기반 자동채점 시스템의 타당성을 높이는 일이 경험 데이터에 기반한 통계적 관계와 인간 채점자가 사용하는 것으로 여겨지는 규칙 양자를 조화시키면서 이루어진다고 할 때(Yang et al., 2002), 이러한 인위적 속임수에 대한 대응 또한 이와 유사한 종합적 방법론을 통해 발전해나갈 것이다. 예컨대, 벡터공간에 위치한 응답들 사이에서 거리가 유독 먼 특이한 응답(outlier)을 찾는 통계적 방법과 함께, 키워드 중심의 응답이나 부정어가 과도하게 사용된 응답을 검출하는 알고리즘을 동시에 사용하는 방식을 고려할 수 있다.

여기서, 학생들이 왜 자동채점 시스템에 대하여 낮은 신뢰를 가지고 있는가를 돌아볼 필요가 있다. 이는 달리 표현하자면 학생들이 기계에 의하여 평가받기를 거부하고 오히려 기계를 평가하는 위치에 서고자 하는 현상이라고 표현하겠다. 또한 기계 평가자에 대한 이러한 감정이 인간 사이의 면대면 상호작용을 선호하는 저연령 학습자에게서 더욱 두드러지게 나타나는 것일 가능성 또한 배제할 수 없으므로, 인공지능 기반 서술형 자동채점 시스템은 되도록 고학년 혹은 성인 학습자에게 적용하는 것이 바람직할 수 있다.

4. 고부담시험에서의 윤리적 한계

자동채점을 위한 데이터를 확보하기 위하여는 해당 문항이 고도로 정교화되고 규준화(normalize)되어야 할 필요가 있다. 그런데 실질적으로 그러한 문항과 이에 대한 응답들은 전국 단위 학력평가에 준하는 대규모 평가에서 얻는 것이 불가피하다. 또한 인력 투입을 최소화하는 신뢰도 있는 채점이라는 면에서 그러한 대규모 평가에 자동채점을 적용하는 것이 합리적이라고도 할 수 있다. 실제로 위에서 언급한 e-rater의 경우 꾸준한 성능 개선을 거쳐 2008년부터는 GRE 에세이 채점의 보조수단으로 활용되고 있으며, GMAT(Graduate Management Admission Test) 및 TOEFL(Test of English as a Foreign Language)에서도 자동채점 시스템을 활용하고 있다. 또한 국내에서는 한국교육과정평가원에서 국가수준 학업성취도 평가 데이터를 기반으로 국어, 사회, 과학 문항을 중심으로 서술형 문항 자동채점에 관한 연구를 진행하여 왔다(노은희, 성경희, 2014; 송미영, 노은희, 성경희, 2016). 하지만 이와 같은 대규모 평가는 대체로 고부담시험이라는 점에서, 과연 여기에 자동채점 시스템을 적용하는 것이 바람직하느냐는 윤리적 이슈가 발생할 수 있다.

앞에서 언급하였던 ‘Professionals Against Machine Scoring Of Student Essays In High-

Stakes Assessment'가 그 대표적인 사례이다. 해당 웹사이트는 고부담 시험에서의 에세이 자동채점에 반대하기 위하여 2013년부터 일종의 공개 진정서(petition)를 표방하고 있다. 이들은 적지 않은 연구 결과들에 근거하여 에세이에 대한 기계 자동채점이 피상적이며(trivial), 환원적이고(reductive), 부정확하고(inaccurate), 진단력이 없고(undiagnostic), 불공정하고(unfair), 비밀스럽다(secretive)며 학교 및 기관에서 이를 사용하지 말아야 한다고 주장한다. 결국 에세이에 대한 채점은 인간에 의해 이루어져야 한다는 것이다. 이들은 에세이에 대한 자동채점에 반대하며 서명을 받기 시작하였는데, 해당 진정서에 대하여 2019년 12월 현재까지 4,337명의 서명을 수집하였다. 물론 여기서 해당 단체의 주장을 무비판적으로 받아들일 것만은 아니다. 자동채점 모델의 사용을 옹호하는 증거들도 위에서 언급하였듯 상당히 누적되어 있으며(e.g. Raczynski, & Cohen, 2018), 많은 기관과 연구자들이 구축하여 개선하고 있는 다양한 자동채점 모델들을 너무 단순화하여 평가절하하기엔 무리가 있는 것이 사실이다. 또한, 2013년 초기 이후 해당 진정서에 포함된 서명들이 모두 교육 및 평가 분야의 전문가들인지를 검증하기도 어렵다. 그럼에도 불구하고, 이들의 주장은 개인 단위에서 일어나는 학습(learning)에 비하여 인간과 인간 사이에서 일어나는 교육(education)을 중시하는 모습으로 바라보아 기계로부터 초래되는 '교육의 학습화'(learnification of education)을 경계하고 있는 것으로도 해석할 여지가 있다.

이는 채점 결과에 대한 책임의 문제와도 직결되어 있다. 우선 복잡한 머신 러닝 모델의 경우 그 해석 가능성(interpretability)이 낮아질 수 있고, 학생 응답에 대한 채점 결과가 왜 그러한지를 이해관계자(stakeholder)들이 이해할 수 있는 형태로 제시하지 못할 수 있기 때문이다. 한편 인공지능의 판단 근거를 이해관계자들이 이해할 수 있다 하더라도, 문제가 제기된 채점 결과에 대한 책임 소재는 여전히 불분명할 수 있다. 말하자면, 그러한 모델 알고리즘을 구축하여 제공한 사업체에 책임이 있는 것인지, 인공지능 채점 모델을 하나의 행위자(agent)와 같이 이해하여 그 자체에 책임을 물을 것인지, 혹은 그 사용자에게 책임을 물을 것인지에 대한 법적 문제가 제기될 수 있다. 이는 자율주행자동차의 사례를 살펴볼 때 전혀 무리하지 않은 전망이다. 예컨대 인공지능 자율주행 시스템에 의해 운전되던 자동차에서 사고가 발생하였을 경우 그 책임 소재가 인공지능 공급자에게 있는지, 인공지능 그 자체를 법적 주체로 인정해야 할 것인지, 그렇지 않으면 인간 탑승자가 책임을 져야 하는지가 법리적인 논의의 주제가 될 수 있는 것이다(e.g. 류병운, 2018). 이와 유사한 일이 고부담 시험에 적용된 인공지능 자동평가에서 일어나지 않으리라는 보장은 없다.

이처럼, 인공지능 기반의 자동평가가 대규모-고부담 시험에 적용된다면 경제적인 측

면에서의 장점이 명확한 만큼 윤리적이고 법적인 측면에서의 문제를 초래할 가능성이 상당히 크다. 결과적으로 서술형 문항에 대한 인공지능 기반 자동평가는 고부담 시험에서의 주요 채점 수단으로 활용하는 데는 무리가 있으며, 채점의 보조수단 및 검토수단 등으로 활용하되 최종적인 판단은 인간 채점자가 내리도록 하는 것이 바람직할 것이다(cf. 성경희, 송미영, 노은희, 2016).

Ⅳ. 인공지능 서술형 자동평가의 단기적 활용 방안

이전 장에서 논의한 바와 같이 인공지능을 활용한 서술형 자동평가의 현재는 자연어 처리 기술의 한계와 그에 따른 낮은 수준의 신뢰도, 사용자의 인위적인 속임수 및 고부담 시험에서의 사용의 제한 등으로 인하여 현재로서는 그 다양한 사용에 제한이 따른다. 하지만, 이와 같은 한계에도 불구하고, 현재 수준의 인공지능 서술형 자동평가를 학교 맥락에서 사용 가능한 영역이 있을 것이다. 이 절에서는 현재의 기술력을 활용하여 인공지능 서술형 자동평가의 단기적 활용 방안에 대해서 논의하고자 한다.

첫째로, 이를 학습자의 학습을 위한 적응적(adaptive) 도구로 활용이 가능하다. 이 경우 학습자의 학습보조가 주목적이기 때문에 컴퓨터를 인위적으로 속이는 행위나 고부담시험에서 발생하는 문제는 고려하지 않을 수 있다. 학습자가 태블릿 PC나 스마트폰을 활용할 경우 학교 교육에 비하여 시공간적 제약이 적기 때문에 학습보조형태의 인공지능 활용이 학생들의 학습 관리에 유용할 수 있다. 먼저 학습자들은 자신의 학습을 관리하는 측면에서 인공지능을 활용할 수 있다. 학습자는 인공지능의 안내를 받아서 스스로 자신을 진단하고, 학습 후 자신의 능력을 점검할 수 있다. McMillan & Hearn (2008)는 학생들의 자가 평가 활동이 학습 동기와 성취수준을 향상시킬 수 있음을 확인하였다. 또한 Ibabe & Jauregizar(2010) 역시 피드백이 있는 온라인 자가 평가 활동이 학생들의 동기를 향상시킬 수 있다고 강조하였다. 다만 인공지능이 정확하지 않은 예측을 할 경우 학생들이 필요한 학습을 안내받지 못할 수 있다는 우려가 가능하다. 학습자가 이미 알고 있는 내용에 대하여 추가적인 학습을 제공하는 경우에는 큰 문제가 되지 않지만, 학습자가 모르고 있는 내용임에도 불구하고 알고 있다고 예측이 될 경우 이와 같은 문제점이 발생한다. 이러한 문제는 다수의 동형 문항을 활용하여 해결할 수 있다. 한 문항으로 진단을 할 경우에는 우연히 발생하는 컴퓨터의 부정확한 예측에 영향을 받을 수 있으나, 예컨대 4개 이상의 문항을 통해 진단할 경우 반복적으로 부정확한 예측이 일어날 가능성은 낮다. 복수의 문항을 통해 학생들의 개념을 종합적으로 진

단할 경우 학습자가 특정 개념을 가지고 있는지 명확하게 확인할 수 있기 때문이다.

진단을 통해 학습자에게서 학습이 필요한 부분이 확인된다면 진정한 의미에서의 개별화 교육이 진행될 수 있다. 교과서가 디지털화되면 교과서나 참고서에서 학습자가 부족한 개념을 설명하고 있는 부분을 추천해 줄 수 있을 것이다. 또는 주요 키워드를 제시해 주어서 학생들이 이를 웹에서 검색할 수 있도록 지원할 수 있다. 이에 따라 학습자는 교사에 대한 의존도를 점차 낮추고, 자기 주도적 학습이 강화됨과 동시에 교사의 업무 부담이 줄어들 것이다. 특히 최근에는 온라인에 많은 학습 자료가 있으며 YouTube와 같은 플랫폼 내에는 학습 동영상과 실감형 콘텐츠들도 상당히 많다. 학습자의 부족한 개념을 보충할 수 있는 학습 소재들을 다양하게 검색하여 제공해 줄 수 있는 엔진이 있다면, 이를 인공지능 자동평가와 연계하여 학습을 촉진시킬 수 있을 것이다. 이러한 가능성이 특히 인공지능 서술형 자동평가에서 더욱 두드러지는 이유는, 상술하였듯이 학습자의 개념 수준을 확인하는 진단 기능의 관점에서 선택형 평가에 비하여 서술형 평가는 우수한 기능을 발휘할 수 있으며(Beggrow et al., 2014) 현실적으로 사람이 아닌 컴퓨터가 그 분석과 피드백을 담당할 때 개별화 교육이 실현될 수 있기 때문이다. 이는 전에 없었던 강력한 적응적 학습 도구가 생성될 수 있음을 함의한다.

인공지능을 활용하여 학생들이 스스로 진단 및 형성평가를 수행할 경우 교사가 학생들의 학습 과정을 관리하는 일이 한층 수월해 질 것이다. 서술형 평가에서는 선택형 평가와 달리 학생들의 언어, 문장 표현, 학습 동기 등이 확실히 나타난다. 따라서 교사가 확인해야 되는 정보 역시 많아질 수밖에 없다. 특히 학생들의 자료가 종단적으로 기록될 경우 학생들의 설명 수준의 발달 과정 역시 중요하게 확인해야 되는 요소이다. 인공지능 서술형 자동평가는 학생 평가 자료를 효율적으로 정리하여 교사에게 제공하는 기능을 지원할 수 있다. 예컨대 숙제나 자가 평가 등을 하지 않는 학생, 학습 발달이 되지 않는 학생, 학생들의 설명의 전체적인 분석이나 수준에 대한 정보 등을 요약하여 제공할 수 있다. 학생들의 전체적인 설명 수준은 BiGram 나무나 토픽 모델링과 같은 시각화 도구로 나타내어 교사에게 제공하면 효과적일 것이다.

둘째로 인공지능 서술형 자동평가는 학생 응답에 대한 빠른 학습분석 시스템으로 활용될 수 있다(cf. 조일현, 박연정, 김정현, 2019, pp. 171-172). 이는 온전히 교사들을 위한 기능을 염두에 둔 것으로서, 학습자는 자연어 처리 기술 및 채점모델의 한계로 인한 부정확한 피드백을 경험하거나 시스템에 대한 신뢰를 잃을 우려가 없다. 교사는 학급 단위로 학생들의 학업 발달을 확인함으로서 자신의 수업에 대한 반성적 고찰을 수행할 수 있다. 예를 들어 학습 전후 학생들의 서술형 평가 응답에서 교사가 제시한 개념이나 언어들이 나타나지 않을 경우 학습이 효율적으로 이루어지지 않았다고 판단

할 수 있다. 다만 여기서 학습 개념이 여러 개일 경우 개념 간 발달의 속도가 다를 수 있다는 점에 유의하여야 한다. 일부 개념의 경우 수업 전후 등장 빈도의 수가 많이 차이가 나는 반면 그렇지 않은 개념도 있을 수 있다. 또한 학습자의 배경 변인을 활용한 빠른 분석도 가능할 것이다. 예를 들어서 학생의 성취도 수준을 상중하로 나눈 자료가 있다고 가정하면, 성취도 수준별로 위와 같은 분석을 수행한 결과가 제공될 수 있다. 컴퓨터의 자동화된 분석을 통해 교사는 자신의 수업이 어떤 집단에게 효율적인지에 대한 통찰을 얻을 수 있는 것이다.

세 번째로 교사에게 필요한 평가 준거에 대한 빠른 분석 역시 가능하다. 서술형 평가 문항을 개발할 때에는 평가 준거 역시 함께 개발하게 된다. 학생 응답을 수집한 뒤에 학생들의 응답을 바탕으로 준거에 대한 수정이 필요한 경우가 있기도 하다. 학생들의 응답을 교사가 정확히 예측하여 평가 준거에 반영하였을 경우에는 문제가 되지 않으나, 학생들의 응답의 범위가 다양하고 모호한 응답이 많을 경우에는 그것을 반영하여 평가 준거를 수정한 뒤 다시 채점해야 한다. 이와 같은 상황에서 비지도학습(unsupervised learning)에 기반한 인공지능모델들을 활용한다면, 자동화된 군집분석 등으로 학생들의 응답을 유형화하고 새로운 평가 준거를 마련하는 데 적지 않은 도움을 받을 수 있다. 이 역시 채점모델이 산출하는 오류를 수정하면서 그 한계를 극복하는 방안이 될 수 있는 것이다.

한편 이러한 인공지능 기반 자동채점 시스템을 어떻게 구축할 수 있는가의 문제를 다루어볼 필요가 있을 것이다. 이는 이를 구축하는 데 필요한 시간적이고 재정적인 자원의 문제와 함께 범용성과 확장성을 확보하는 일을 핵심으로 한다. 예컨대, 많은 예산과 노력을 들여 개발된 훌륭한 시스템이라고 하더라도 대다수의 학습자 및 교사들이 사용하기 어려운 플랫폼 위에서만 작동한다면 그 활용 가능성이 극도로 낮아질 수 있기 때문이다. 이러한 문제들에 대하여는 하민수 외(2019)가 보고한 WA^{3}I(Web-based Automated Assessment using Artificial Intelligence, 인공지능을 활용한 웹기반 자동평가) 프로젝트가 가능성을 보여주고 있다. 해당 프로젝트에서는 2015 개정 교육과정과 관련된 과학 및 사회 30여개 문항을 개발하고, 이에 대하여 대략 1,200개의 학습자 응답을 수집하여 머신 러닝을 수행한 결과 Kappa의 값이 대부분 0.75 이상으로 컴퓨터와 전문가 채점 간의 일치도가 상당히 높이 나타나는 채점모델을 구축하였음을 보고하였다. 그리고 해당 문항을 학습자들이 풀어보고 답안의 정답 유무에 관한 피드백을 받을 수 있는 시스템을 별도의 프로그램 설치가 필요하지 않을 뿐 아니라 PC와 모바일 등에서 비교적 자유롭게 접근 가능한 웹상에서 구현하였다.[10] 이는 별도의 플랫폼을 구축할

10) https://www.crezone.net/?page_id=603485 (2020. 06. 09. 확인)

필요가 없으므로 개발을 위한 자원 소요가 줄어들 수 있는 방안이다. 또한 학습자들의 응답과 그에 대한 채점 기록이 서버에 저장될 수 있으므로 교사들이 학습자들의 수준을 파악하는 데에도 도움을 줄 수 있을 뿐 아니라, 향후의 채점모델 수정 및 업데이트에 있어서도 유리하다. 이는 곧 상술하였던 바와 같은 인공지능 자동채점의 단기적 활용의 가능성을 보여주는 충분한 사례가 된다고 하겠다.

Ⅴ. 인공지능 기반 자동평가의 미래

이전 장들에서는 인공지능 서술형 평가의 한계와 도전, 그리고 그 단기적 활용 가능성을 살펴보았다. 하지만, 위에서 논의한 단기적 활용 방안에 그치지 않고 인공지능 기반 자동채점의 외연을 확장하기 위한 연구들 또한 지속적으로 이루어져야 할 것이다. 여기서, 인공지능 기반 자동채점이 직면한 도전들은 단순히 교육평가 이론의 관점에서 해결할 수 있는 문제들이 아니라는 점에 유의하여야 한다. 테크놀로지의 지속적인 발달은 물론이며, 교육철학적인 고찰뿐만 아니라 이를 중앙집권적 교육 체제 안으로 통합하는 일의 윤리적 함의에 대한 사회문화적인 숙의(deliberation) 과정이 뒤따라야 할 것이다. 인공지능 기술 자체가 지닌 간학문적(inter-disciplinary) 특성과 마찬가지로, 이를 교육평가의 영역에 도입하는 일 역시 이를 둘러싼 다양한 학문 분야에서의 협력적인 고찰이 필요하게 될 것이며, 그 전개 양상 또한 상당히 복잡다단할 것이므로 그 미래를 속단하기는 쉽지 않다.

이에 이번 장에서는 기술적/윤리적인 논의의 진전이 위에서 언급된 이슈들을 언젠가 극복해낼 때 인공지능 자동평가가 활용될 수 있는 미래적 전망을 시도하며 전체 논의를 마무리하고자 한다. 본 장에서 살펴보는 주제들은 상호연관성을 지니고 있으나 편의상 논리적으로 구분한 것이다.

먼저, 상술하였듯 인공지능 기반 언어 처리 시스템 자체가 데이터가 많은 영역-일반적인 수준에서 아직 데이터가 적어 구축이 어려운 영역-특수적이고 전문적인 분야로 발전해나가야 하는 과제를 안고 있으므로, 자동평가에 있어서도 보다 교과-특수적인 정확성을 요구하는 시스템이 개발될 것으로 예상할 수 있다. 근래에는 과학 교과와 관련된 연구들이 적지 않게 보고되었던 바(e.g. Moharreri et al., 2014; Weston et al., 2015; Zhu et al., 2020), 엄밀한 언어로 표현된 개념을 중시하는 과학이야말로 이러한 자동채점 시스템의 미래를 고찰할 좋은 대상이 된다고 하겠다.

이와 함께 인공지능 자동채점의 미래는 무엇보다도 텍스트 데이터를 넘어선 이미지

데이터의 처리에 주목해야 할 것이다.[11] 이는 딥 러닝 및 이미지 처리 관련 테크놀로지의 발달로 인하여 가능해진 면이 크다. 딥 러닝 기반 알고리즘들은 2011년 음성인식 분야, 2012년 사물인식 분야(ImageNet challenge), 2014년 얼굴인식 분야 등에서 최고 성능을 내는 알고리즘으로서 크게 주목 받기 시작하였고(장병탁, 2017, p. 15), 이와 함께 컴퓨터 비전(vision) 분야의 지수함수적인 발전이 초래되었다. 또한 하드웨어의 측면에서는 그래픽 처리 장치(GPU, Graphic Processing Unit)의 성능이 향상되고 연구자 및 사용자들이 접근 가능한 데이터의 양이 크게 증가하면서 이미지 분류, 이미지 검색, 사물 인식(object detection) 등의 기술의 활용장벽이 낮아지게 되었다. OpenCV(Open source Computer Vision)[12]는 대표적인 오픈 소스 컴퓨터 비전 및 머신 러닝 라이브러리로서 이를 적절히 활용할 경우 얼굴 인식, 사물 확인, 인간 행동 분류 등의 작업을 개인 PC 수준에서 수행하는 일이 가능하다. YOLO(You Only Look Once)[13] 또한 실시간 사물 인식을 가능하게 하는 딥 러닝 기반 오픈 소스 시스템의 일부로서 연구자들이 무료로 사용할 수 있도록 배포되어 있는 주요 사례이다.

이와 관련하여 인공지능 자동채점이 나아갈 수 있는 방향으로서는 우선 첫째, 문어 및 구어를 넘어선 시각적 표상(visual representation)에 대한 자동평가이다. 예를 들어, 과학 교과에서는 과학적 개념이나 이론에 대한 학생들의 정신 모형(mental model)을 시각화하여 나타내고 이를 평가하는 일이 매우 중요하게 여겨진다. 특히 화학 교과에서는 분자와 같은 입자 개념에 근거하여 화학적 현상을 설명하는 일이 핵심적인데, 이 때 학생들은 분자의 구조 및 분자들의 공간적 분포를 올바로 시각화하는데 어려움을 겪는 일이 많다. 이에 학생들이 직접 그린 입자적 설명 모형을 평가하고 이를 근거로 특정 교수법의 효과를 주장하는 연구들이 적지 않게 이루어져 왔다(e.g. Dori, & Kabermanm, 2012; Tasker, & Dalton, 2008). 지금까지 학생들의 입자 모형 시각화에 대한 채점이 인간 연구자에 의하여 이루어져 왔음을 고려할 때, 딥 러닝 기반의 이미지 처리 기술을 활용하여 이에 대한 자동채점을 시도하는 일이 가능하다.

둘째, 단답-서술형 응답이 아닌 실제 수행에 대한 자동평가이다. 기실 복잡한 수행에 대한 자동채점 아이디어는 상당히 오래 전에 제안된 바 있었으나(e.g. Clauser et al., 1997) 이러한 연구들도 결국 인간 관찰자에 의해 기록된 데이터에 기반한 평가에 해당

11) OCR(Optical Character Recognition, 광학문자인식) 기술의 발달로 인하여 기존에 비해 텍스트 데이터를 얻기 수월해진 면이 있다는 점도 언급할 수 있을 것이다. 다만 본 장에서는 시각적 이미지 데이터의 처리가 가능해졌다는 점에 초점을 두도록 하겠다.

12) https://opencv.org/ (2019.12.31. 확인)

13) https://pjreddie.com/darknet/yolo/ (2019.12.31. 확인)

하며 이미지 혹은 영상 데이터 자체를 직접 처리하는 채점 형식에는 이르지 못한 면이 있다. 여기서 근래에 그 접근성이 우수해진 사물 인식 및 인간 행동 분류 기술을 적용할 여지가 있다. 예컨대 과학 교과에서 중요시되는 실험 실습(hands-on) 활동에서는 안전 문제가 대두되고 있는데, 학생이 장갑이나 고글 등의 안전장비를 착용하지 않거나 유리 기구를 위험하게 다루는 등의 경우는 중등교육 및 고등교육을 막론하고 실험 수행에 대한 평가 요소에 반영되는 경우가 일반적이다. 인공지능이 이러한 요소들을 자동적으로 검출 및 보고하여 교사의 부담을 줄인다면, 교사는 학생들의 참 탐구(authentic inquiry)를 돕는 실험 지도에 더 많은 자원을 투입할 수 있을 것이다. 이와 유사한 맥락에서, 게임 및 시뮬레이션 게임에서 학생들의 복잡한 수행을 자동으로 평가하는 등(Iseli et al., 2010)의 연구가 보고된 바 있으나, 이미지 처리 등을 실시간 진단에 활용한 사례는 아직 드문 것으로 보인다.

셋째, 휘발성 있는(volatile) 정보에 대한 실시간 피드백과 과정 중심 자동평가이다.[14] 김유정 외(2019)는 2015 개정 교육과정 하에서 중학교 및 고등학교 과학 교사 각 1인의 실험 수업 사례를 중심으로 과정 중심 평가를 지원하기 위한 요소를 다음과 같이 정리하였다. 이를테면, 과학 수업에서 과정 중심의 평가가 실질적으로 이루어지기 위해서는 가장 먼저 교사의 시간 부족을 해결해야 하며, 교사가 수업 중 수시로 평가를 진행하고 학생들은 이를 실시간으로 확인할 수 있도록 돕는 매체가 필요하며, 평가 결과의 객관성이 담보되어야 한다(김유정 외, 2019). 이미지 처리 기반의 인공지능 자동채점 시스템이 모바일 디바이스 및 애플리케이션과 연동될 경우에는 이러한 역할을 충분히 감당할 수 있다. 앞서 논의되었듯이 인공지능 자동채점 시스템은 어느 정도 타당하고 신뢰로운 범위 내에서 실시간으로 텍스트 및 이미지 정보를 처리 가능하므로, 궁극적으로는 교사에게 필요한 시간 자원의 확보를 지원하는 효과적인 수단이 될 수 있는 것이다. 이와 유사한 맥락에서, 비교적 최근에 보고된 Zhu et al.(2020), Lee et al.(2019), Mao et al.(2018), Zhu et al.(2017) 등이 과학 수업에서의 학생들의 논증을 실시간으로 평가하고 피드백하는 테크놀로지의 현재를 보여주는 사례가 된다고 하겠다.

본 연구에서 전반적으로 살펴보았듯이, 인공지능 기반 자동채점 시스템의 도입은 먼 미래에나 가능한 일이라고만 하기는 어렵다. 테크놀로지의 측면에서 상당한 발전이 이루어졌을 뿐 아니라, 이를 교육 현장에 투입하였을 때의 성능을 보고하는 연구들 또한 적지 않게 이루어져 왔기 때문이다. 특히 한국교육과정평가원에서 지속적으로 보고하

14) 여기서 '과정 중심 평가'는 2015 개정 교육과정에서 본격적으로 도입된 용어이기도 하지만, 평가에 대한 완전히 새로운 패러다임을 제시하는 것이라기보다 기존부터 제기되었던 '과정'의 중요성을 상기한다는 의미로 사용하였다.

고 있는 자동채점 관련 연구들은 앞으로 한국의 대규모 시험에서도 이를 활용한 평가가 이루어질 수 있음을 시사하며(성태제 외, 2010; 노은희, 성경희, 2014; 송미영, 노은희, 성경희, 2016), 상술하였듯 2015 개정 교육과정 시기에 본격적으로 도입되고 있는 과정중심평가는 이를 가속화할 가능성이 있다(cf. 김유정 외, 2019)

본 연구에서는 인공지능 기반의 서술형 문항 자동채점 시스템의 현재와 기술적 및 윤리적 도전을 살펴보고, 그 단기적인 활용 방안을 살펴보았다. 인공지능 서술형 문항 자동채점 시스템은 이미 광범위한 주제의 에세이 채점뿐만 아니라 정교한 과학적 개념을 묻는 문항에 대한 채점에도 사용될 만한 성능을 보이며 평가의 패러다임을 바꾸어가고 있다. 하지만 인공지능 서술형 자동채점 시스템은 (1) 자연어 처리의 기술적 한계, (2) 채점 신뢰도 문제, (3) 인간 사용자의 신뢰와 인위적 속임, (4) 고부담 시험에서의 윤리적 한계 등의 도전을 안고 있다. 이에 단기적으로는 인공지능 자동채점 시스템이 (1) 학습자에게 적응적인(adaptive) 학습 지원 도구로서, (2) 학습자의 응답에 대한 빠른 학습분석 도구로서, (3) 교사에게 필요한 평가 준거의 빠른 분석을 가능케 하는 역할을 감당할 수 있다. 하지만 이러한 문제들이 기술적/윤리적으로 극복된다면, 인공지능 자동채점 시스템의 미래는 과학 등의 교과-특수적인 영역에서, 텍스트를 넘어서는 이미지 처리를 통해 (1) 정신 모형의 시각적 표상에 대한 자동평가, (2) 실제 수행에 대한 자동평가, (3) 실시간 피드백과 과정중심 자동평가에 활용될 수 있을 것이다. 하지만 딥 러닝이 인공지능 기술의 급속한 발달을 주도하게 될 것을 누구도 예측하지 못하였듯이, 인공지능 자동채점 시스템의 미래적 가능성 역시 여전히 열려 있다고 하겠다.

참고문헌

김유정, 이경건, 장원형 (2019). 교사의 과정 중심 평가 역량에 관한 사례연구: 중·고등학교 과학 교사 사례를 중심으로. 한국과학교육학회지, 39(6), 695-706.

(Translated in English) Kim, Y., Lee, G., & Hong. H. (2019). A case study on Teacher's Process-centered Evaluation Competency(T-PEC): Focused on the case of a middle-school/a high-school science teacher. *Journal of the Korean Association for Science Education, 39*(6), 695-706.

김홍겸, 박창수, 정시훈, 고호경 (2018). 미래교육에서의 인간 교사와 인공지능 교사의 상호보완적 관계에 대한 소고. 교육문화연구, 24(6), 189-207.

(Translated in English) Kim, H. -K., Park, C., Sihun, J., & Ko, H. k. (2018). A view on complementary relation of human teacher and AI teacher in future education. *Society and Theory, 24*(5), 189-207.

김지은, 이공주 (2007). 중학생 영작문 실력 향상을 위한 자동 문법 채점 시스템 구축. 한국콘텐츠학회논문지, 7, 36-46.

(Translated in English) Kim, J. & Lee, K. J. (2007). Implementing automated english error detecting and scoring system for junior high school students. *Journal of Korea Contents Association, 7*, 36-46.

노은희, 성경희 (2014). 한국어 서답형 문항 자동채점 결과 비교 분석 - 국가수준 학업성취도 평가 국어, 사회, 과학 문항을 중심으로 -. 교육과정평가연구, 17(2), 99-122.

(Translated in English) Noh, E., & Sung, K. (2014). A comparative analysis of scoring results in Korean automatic scoring program for short-answer items-focused on the three subjects in NAEA: Korean, social studies and science. *The Journal of Curriculum and Evaluation, 17*(2), 99-122.

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

(Translated in English) Noh. E., Lee, S., Lim, E. Sung, K., & Park. S. (2014). *Development and applicability validation of automated scoring program for short-answer item written in Korean* (Research Report RRE 2014-6). Seoul: Korea Institute for Curriculum and Evaluation.

류병운 (2018). 자율주행자동차 사고의 법적 책임. 홍익법학, 19(1), 31-58.

(Translated in English) Lyou, B. (2018). Legan responsibility of autonomous vehicle accident. *Hongik Law Review, 19*(1), 31-58.

백영민 (2017). R를 이용한 텍스트 마이닝. 서울: 한울아카데미.

(Translated in English) Baek, Y. (2017). *Text Ming Using R*. Seoul: Hanul Academy.

성경희, 송미영, 노은희 (2016). 자동채점 프로그램을 적용한 대규모 평가 사회과 서답형 문항 채점 결과 분석. 시민교육연구, 48(2), 31-46.

(Translated in English) Sung, K., Song, M., & Noh, E. (2016). Automated scoring and analysis on the scoring result in large-scale social studies assessment. *Theory and Research in Citizenship Education, 48*(2), 31-56.

성태제, 양길석, 강태훈, 정은영 (2010). 학업성취도 평가 서답형 문항 컴퓨터 채점화 방안 탐색 (연구보고 RRE 2010-1). 서울: 한국교육과정평가원.

(Translated in English) Sung, T., Yang, G., Kang, T., & Jeong, E. (2010). *Exploration of computerized scoring of short-answer items in academic achievement test* (Research Report RRE 2010-1). Seoul: Korea Institute for Curriculum and Evaluation.

송미영, 노은희, 성경희 (2016). 대규모 평가 서답형 문항 채점을 위한 문장 수준 자동채점 프로그램의 정확성 분석. 교육과정평가연구, 19(1), 255-274.

(Translated in English) Song, M., Noh, E., & Sung, K. (2016). Analysis on the accuracy of automated scoring for Korean large-scale assessment. *The Journal of Curriculum and Evaluation, 19*(1), 255-274.

장병탁 (2017). 장교수의 딥러닝. 서울: 홍릉과학출판사.

(Translated in English) Zhang, B. (2017). *Deep Learning*. Seoul: Hongreung Publishing Company.

조일현, 박연정, 김정현 (2019). 학습분석학의 이해. 서울: 피와이메이트

(Translated in English) Cho, I., Park, Y., & Kim, J. (2019). *Understanding Learning Analytics*. Seoul: Pymate.

하민수 (2017). 영어기반 컴퓨터자동채점모델과 기계번역을 활용한 서술형 한국어 응답 채점 - 자연선택개념평가 사례 -. 한국과학교육학회지, 36(3), 389-397.

(Translated in English) Ha, M. (2017). Scoring Korean written responses using English - Based automated computer scoring models and machine translation: A case of natural selection concept test. *Journal of the Korean Association for Science Education, 36*(3), 389-397.

하민수, 이경건, 신세인, 이준기, 최성철, 주재걸, ... , 박지선 (2019). 학습 지원 도구로서의 서술형 평가 그리고 인공지능의 활용: WA^{3}I 프로젝트 사례. 현장과학교육, 13(3), 272-282.

(Translated in English) Ha, M., Lee, G. -G., Shin, S., Lee, J. -K., Choi, S., Choo, J., ... , Park, J. (2019). Assessment as a learning-support tool and utilization of artificial

intelligence: WA^{3}I project case. *School Science Journal, 13*(3), 272-282.

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2010). Automated, unobtrusive, action-by-action assessment of self-regulation during learning with an intelligent tutoring system. *Educational Psychologist, 45*(4), 224-233.

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance?. *Journal of Science Education and Technology, 23*(1), 160-182.

Clauser, B. E., Margolis, M. J., Clyman, S. G., & Ross, L. P. (1997). Development of automated scoring algorithms for complex performance assessments: A comparison of two approaches. *Journal of Educational Measurement, 34*(2), 141-161.

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.

Dori, Y. J., & Kaberman, Z. (2012). Assessing high school chemistry students' modeling sub-skills in a computerized molecular modeling learning environment. *Instructional Science, 40*(1), 69-91.

Gerard, L. F., Ryoo, K., McElhaney, K. W., Liu, O. L., Rafferty, A. N., & Linn, M. C. (2016). Automated guidance for student inquiry. *Journal of Educational Psychology, 108*(1), 60-81.

Ha, M., & Nehm, R. H. (2016a). The impact of misspelled words on automated computer scoring: A case study of scientific explanations. *Journal of Science Education and Technology, 25*(3), 358-374.

Ha, M., & Nehm, R. (2016b). Predicting the accuracy of computer scoring of text: Probabilistic, multi-model, and semantic similarity approaches. In *Proceedings of the National Association for Research in Science Teaching*, Baltimore, MD, April, 14-17.

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. *CBE-Life Sciences Education, 10*(4), 379-393.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. *CBE-Life Sciences Education, 11*(3), 283-293.

Ibabe, I., & Jauregizar, J. (2010). Online self-assessment with feedback and metacognitive knowledge. *Higher Education, 59*(2), 243-258.

Iseli, M. R., Koenig, A. D., Lee, J. J. and Wainess, R. 2010. *Automatic assessment of complex task performance in games and simulations*. Los Angeles: National Center for Research on Evaluation, Standards, Student Testing, Center for Studies in Education, UCLA. (CRESST Research Report No. 775).

Kaplan, J. J., Haudek, K. C., Ha, M., Rogness, N., & Fisher, D. G. (2014). Using lexical analysis software to assess student writing in statistics. *Technology Innovations in Statistics Education, 8*(1). Retrieved from https://escholarship.org/uc/item/57r90703

Kukich, K. (2000). Beyond automated essay scoring. In M. A. Hearst (Ed.), *The Debate on Automated Essay Grading*. IEEE Intelligent systems, 27-31.

Lee, H. S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education, 103*(3), 590-622.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice, 33*(2), 19-28.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53*(2), 215-233.

Lochbaum, K. E., Rosenstein, M., Foltz, P. W., & Derr, M. A. (2013). Detection of gaming in automated scoring of essays with the IEA. In *National Council on Measurement in Education Conference (NCME)*, San Francisco, CA.

Magnusson, S. J., Templin, M., & Boyle, R. A. (1997). Dynamic science assessment: A new approach for investigating conceptual change. *The Journal of the Learning Sciences, 6*(1), 91-142.

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H. S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment, 23*(2), 121-138.

McMillan, J. H., & Hearn, J. (2008). Student self-assessment: The key to stronger student motivation and higher achievement. *Educational Horizons, 87*(1), 40-49.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Moharreri, K., Ha, M., & Nehm, R. H. (2014). EvoGrader: An online formative assessment tool

for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach,* 7(1), 15.

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching, 48*(3), 237-256.

Nehm, R. H., & Haertig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: Testing the efficacy of text analytic software. *Journal of Science Education and Technology, 21*(1), 56-73.

Nehm, R. H., Beggrow, E. P., Opfer, J. E., & Ha, M. (2012). Reasoning about natural selection: Diagnosing contextual competency using the ACORNS instrument. *The American Biology Teacher, 74*(2), 92-98.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology, 21*(1), 183-196.

Opfer, J. E., Nehm, R. H., & Ha, M. (2012). Cognitive foundations for science assessment design: Knowing what students know about evolution. *Journal of Research in Science Teaching, 49*(6), 744-777.

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

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior, 18*(2), 103-134.

Raczynski, K., & Cohen, A. (2018). Appraising the scoring performance of automated essay scoring systems-Some additional considerations: Which essays? Which human raters? Which scores?. *Applied Measurement in Education, 31*(3), 233-240.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8).

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. *The Journal of Technology, Learning and Assessment, 4*(4). Available from http://www.jtla.org

Sripathi, K. N., Moscarella, R. A., Yoho, R., You, H. Sun, Urban-Lurain, M., Merrill, J., & Haudek, K. (2019). Mixed Student Ideas about Mechanisms of Human Weight Loss. *CBE-Life Sciences Education, 18*(ar37), 1-17.

Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M. C. (2017). Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education, 27*(4), 729-757.

Tasker, R., & Dalton, R. (2008). Visualizing the molecular world-design, evaluation, and use of animations. In *Visualization: Theory and practice in science education* (pp. 103-131). Springer, Dordrecht.

Weston, M., Haudek, K. C., Prevost, L., Urban-Lurain, M., & Merrill, J. (2015). Examining the impact of question surface features on students' answers to constructed-response questions on photosynthesis. *CBE-Life Sciences Education, 14*(ar19), 1-12.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*(4), 391-412.

Zhu, M., Lee, H. S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education, 39*(12), 1648-1668.

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education, 143*, 103668.

◎ 논문접수: 2020. 04. 30 / 1차 수정본 접수: 2020. 06. 12 / 게재승인: 2020. 06. 19