Journal of the Korean Association for Science Education
Journal homepage: www.koreascience.org

전문가의 형태소 분류를 활용한 과학 논증 자동 채점

이만형, 유선아*
한국교원대학교

Automated Scoring of Scientific Argumentation Using Expert Morpheme Classification Approaches

Manhyoung Lee, Suna Ryu*
Korea National University of Education

ARTICLE INFO

Article history:
Received 21 May 2020
Received in revised form
23 June 2020
30 June 2020
Accepted 30 June 2020

Keywords:
scientific argumentation, automated scoring, machine learning, scientific language, experts system

ABSTRACT

We explore automated scoring models of scientific argumentation. We consider how a new analytical approach using a machine learning technique may enhance the understanding of spoken argumentation in the classroom. We sampled 2,605 utterances that occurred during a high school student's science class on molecular structure and classified the utterances into five argumentative elements. Next, we performed Text Preprocessing for the classified utterances. As machine learning techniques, we applied support vector machines, decision tree, random forest, and artificial neural network. For enhancing the identification of rebuttal elements, we used a heuristic feature-engineering method that applies experts' classification of morphemes of scientific argumentation.

Ⅰ. 서론

2019년 미래세대를 위한 과학교육표준(Korean Science Education Standards, KSES)에서는 4차 산업 혁명과 빅데이터의 시대에서 과학 교육이 추구하는 목표로 의사소통과 협력을 바탕으로 창의적 사고력과 문제해결력을 지닌 '과학적 소양을 갖추고 더불어 살아가는 창의적인 사람'을 강조한다(Song *et al*., 2019). KSES에서 역량으로 제안한 '의사소통과 협업'이란 협업 상황에서 과학적 지식과 소양을 바탕으로 과학적 설명과 주장을 전달, 교환, 공유하는 능력을 의미한다(KOFAC, 2019). 이러한 역량을 바탕으로 과학 교육은 지식을 어떻게 알게 되었는지, 알고 있는 것이 왜 그러하다고 믿는지를 이해할 수 있는 학습자의 능력을 개발해야하며, 논증(argumentation)은 과학 언어의 구조적 요소로서 이러한 목적을 이루기 위한 필수적인 역할을 수행한다(Jiménez-Aleixandre, Rodriguez, & Duschl, 2000).

과학적 의사소통으로써 논증이란 단순히 과학적 용어를 나열하고 교과 지식의 참과 거짓을 논하는 것이 아니다. 논증은 사회적 상호작용에서 이루어지는 실천적 행위로써(Driver, Newton, & Osborne, 2000), 증거를 얻고 사용하는 과정 속에서 대화를 활용하여 설명과 예측을 개발하고 자신의 합리적 신념체계를 구축해가는 활동이다(Duschl, 2008). 그러므로 논증을 통한 과학수업은 증거에 기반한 과학적 설명을 만들고 평가할 수 있는 능력과 과학적 설명을 정당화할 수 있는 능력을 개발해야 한다(Kang, & Lee, 2013).

이미 많은 과학 교육자들은 탐구 과정과 학생 논증의 유형을 분석하고 평가하는 것이 논증 수업의 목적을 달성하기 위해 꼭 수행되어야 하는 연구임을 인식하고 관련된 연구를 수행해왔다(Jiménez-Aleixandre, Rodriguez, & Duschl, 2000; Zohar, & Nemet, 2002; Erduran, Simon, & Osborne, 2004). 그러나 말하기 논증 연구의 경우, 논증을 평가하기 위해 학생의 담화를 전사하고 녹화된 영상을 비교한 후, 연구자가 논증 요소를 파악하는 과정에 많은 시간과 노력이 요구된다.

인공지능의 기법과 기계 학습을 기반으로 한 논증의 자동 채점과 논증 마이닝 연구는 이러한 문제를 해결하는데 도움을 줄 수 있다(Lee *et al*., 2018; Mao *et al*., 2018; Lee *et al*., 2019). 자동 채점은 많은 시간과 노력 없이 실시간으로 학생의 반응을 평가할 수 있는 기계 학습 영역 중 하나이며, 짧은 시간 내에 많은 양의 텍스트를 평가할 수 있기 때문에 대규모 평가에 사용되어 왔다(Lee *et al*., 2019; Liu *et al*., 2016). 외국의 경우 컴퓨터를 활용한 자동 채점에 대한 연구는 1960년대 Ellis B. Page가 에세이 자동 채점을 최초로 제안하면서 관심을 갖게 되었으며 이후 컴퓨터와 자연어 처리 기술(natural language process)의 발달에 따라 언어의 특징과 수준을 평가할 수 있는 수준까지 발전하게 되었다(Lee, & Park, 2019). 특히 2000년대 이후 기계 학습의 발달로 인하여 학생의 논증과 에세이에 대한 연구가 활발히 진행되어 미국교육평가원(Educational Testing Service, ETS)은 c-rater와 e-rater를 개발하였다. c-rater에 기계 학습 기술이 추가된 c-rater-ML은 현상을 설명하기 위해 증거를 사용해야 하는 8개의 과학 문항을 채점할 수 있도록 개발되었다(Liu *et al*., 2016).

* 교신저자 : 유선아 (sunaryu@knue.ac.kr)
http://dx.doi.org/10.14697/jkase.2020.40.3.321

또한 Nehm, Ha, & Mayfield(2012)는 진화에 대한 학생의 설명을 기계 학습을 활용한 자동 채점 프로그램(SIDE)에 적용하는 연구를 진행하여 기계 학습을 통한 자동 채점의 효과성을 연구하였고, Lee *et al*.(2019)은 학생의 과학적 논증을 c-rater-ML을 활용한 자동 채점 프로그램(HASbot)에 적용하는 연구를 통하여 자동 채점을 통한 피드백 제공의 효과성에 대한 연구를 수행하였다. 국내의 경우 한국교육과정평가원이 영작문 채점을 위한 자동채점 프로그램을 개발하는 것으로부터 본격적인 자동 채점 연구가 시작되었다(KICE, 2006). 이후 국어, 사회, 과학 학업성취도 평가의 문장 수준 응답 대한 자동채점 연구(Song, Noh, & Sung, 2016), 과학 문항에 대한 자동 채점 프로그램의 개발(Ha *et al*., 2019) 등의 연구가 진행되고 있다.

자동 채점의 기본적인 개념은 동일한 점수를 얻은 텍스트 데이터를 통해 공통되는 단어, 의미, 특징을 자동으로 추출하여 모델을 구축한 후, 새롭게 제시된 텍스트에 대한 예측을 통해 채점을 수행하는 것이다(Beggrow *et al*., 2014). 자동 채점은 평가자의 시간과 노력을 감소시켜 평가의 효율성을 향상시킬 뿐만 아니라 평가자의 편견을 제거함으로써 신뢰성을 확립하는 것에 도움이 된다(Lee *et al*., 2019). 또한 학생의 서술형 답안뿐만이 아니라 과학적 논증 수업 중 학생의 응답에 대한 실시간 평가와 피드백이 가능해지기 때문에 학습 발달에 큰 도움이 될 수 있으며 교육 현장에 큰 도움이 될 수 있다. 그러나 기술적 제약이나 데이터 구축 등의 문제로 교육 현장에서는 적극적으로 활용되고 있지 못하고 있다(Ha *et al*., 2019).

국외의 자동 채점 연구는 내용에 기반한 서술형 텍스트에 대한 평가(Nehm, Ha, & Mayfield, 2012; Lee *et al*., 2019; Liu *et al*., 2016)와 전체 작문의 품질이나 구조를 평가에 대한 연구(Bridgeman, Trapani, & Atali, 2012) 등이 있다. Linn *et al*.(2014)은 연구를 통해 과학적 설명의 실시간 자동 채점을 통해 얻어진 자동화된 피드백이 교실에서 중요한 도구로 사용할 수 있음을 보여주었다. 또한 Lee *et al*.(2019)은 교실 활동을 실시간으로 자동 채점하고 피드백을 제공함으로써 과학 논증 글쓰기를 향상시켰다는 결과를 얻었다. 이처럼 자동 채점 기술 발전은 설명, 논증 또는 커뮤니케이션과 같은 언어에 의해 유발된 작업에 대하여 개인화되고 맞춤화된 피드백을 창출하는데 있어 실행 가능한 길이 되고 있다(Martin, & Sherin, 2013).

우리나라의 과학 교육 분야에서 기계 학습을 활용한 학생 평가 연구로써는 국가수준 학업성취도 평가의 서답형 문항의 평가에 대한 한국교육과정평가원의 연구가 진행되었고(Song, Noh, & Sung, 2016), Ha(2016)는 한국어로 작성된 자연선택 개념의 서술형 답안을 영어로 번역한 후 채점 수행하는 알고리즘을 적용한 결과 자동 채점의 유의미한 결과를 얻었다. 또한 Ha *et al*.(2019)는 학생의 학습을 지원하는 도구로써 활용할 수 있는 서술형 평가 문항 자동 채점 프로그램 WA^3I를 개발하는 자동 채점 연구를 수행하였다. 그러나 이러한 연구는 특정한 과학 지식에 대한 학생의 서술형 답안을 활용하여 과학적 지식의 자동 채점과 피드백의 활용에 초점을 맞춘 연구이므로 과학 논증 구조와 모델링과 같은 과학적 실천을 평가하기 위한 추가적인 연구가 필요하다. 과학 논증에 대한 평가와 피드백을 위한 기계 학습의 국내 선행 연구로는 Lee *et al*.(2018)이 있다. Lee *et al*.(2018)은 논증에 대한 선행 문헌에 제시된 발화 및 교실 환경에서 수집된 학생의 논증 발화에서 추출된 데이터를 활용하여 자동 채점 모델을 구축하였고, 이를 통해 논증 자동 채점의 타당성과 활용 가능성에 대한 연구를 시작하였다.

자동 채점 연구에 있어 전문가 시스템(expert system)은 채점 모델의 성능을 개선시키는데 도움이 될 수 있는 인공지능 기술이다. 전문가 시스템이란 인공지능이 연구되는 초기에 제안된 기법으로서 특정 지식(domain knowledge)의 복잡한 문제를 해결하는 전문가의 사고 과정을 분석한 후 휴리스틱(heuristic)으로 프로그래밍에 적용하는 방법이다(Buchanan, & Feigenbaum, 1980). 전문가 시스템의 대표적인 예로 1969년 Stanford대학에서 개발된 DENDRAL이 있다. DENDRAL은 화학자의 질량 스펙트럼 분석 과정을 프로그램에 if-then과 같은 규칙으로 적용시킨 프로그램으로써 미지 화합물의 스펙트럼을 통해 분자 구조를 예측한다(Russell, & Norvig, 2016). 교육 분야의 전문가 시스템의 경우 교수 계획, 학생 상담의 도구로써 활용될 수 있을 뿐만 아니라 개별화된 교수 평가 도구로써 활용될 수 있다(Baek, 1989; Engin *et al*., 2014). 자동 채점의 선행 연구에서는 전문가 시스템과 비슷한 방법으로써 논증 과정에서 논증 요소를 분류 할 수 있는 형태소와 규칙을 정의한 후 자동 채점을 수행하는 연구가 진행되었다(Ong, Litman & Brusilovsky, 2014; Kim, 2019).

본 연구는 2017년~2019년에 수집된 한 고등학교의 1-3학년을 대상으로 한 소집단 논증 전체 발화에 대한 데이터를 기반으로 자동 채점에 대한 연구를 진행하였다. 교실 현장에서 얻어진 많은 양의 논증 발화는 기계 학습 연구에서 가장 중요하다고 할 수 있는 훈련 데이터의 양과 질의 확보로 연결되는 것이다. 특히 교실 현장에서 얻어진 실제적인 논증 발화는 수업 상황 속의 모든 발화를 포함하므로 학생들의 실제적 사고와 논증 패턴을 반영할 뿐만 아니라 일상적 대화 과정에서 발생할 수 있는 데이터의 잡음도 포함한다. 이러한 실제적인 논증 발화에 대한 연구는 교실 현장에 자동 채점을 적용할 수 있는 가능성을 높인다.

이에 본 연구는 다양한 방법의 기계 학습을 활용한 자동 채점을 통해 어떠한 기계 학습이 자동 채점에 적합한지 알아보는 것을 목적으로 한다. 또한 기계 학습을 통해 얻어진 채점 모델의 분석을 통하여 채점의 정확도를 개선하기 위한 방안으로써 형태소를 활용한 전문가 시스템의 적용 가능성을 탐색하고자 한다.

II. 이론적 배경

최근 많은 과학교육 연구자들이 기계 학습 기법을 연구에 활용하는 것에 관심을 갖고 있으나 과학 교육 분야에서 기계 학습에 대한 이해는 아직 걸음마 단계라고 할 수 있다. 이에 이 연구에서는 먼저 논증의 맥락에서 기계 학습과 데이터 마이닝 등, 인공지능 빅데이터 교육 시대에 필요한 기본 개념에 대한 설명을 먼저 제공하여 본 연구에 관한 이해를 높이고 필요성을 강조하고자 한다. 이를 바탕으로, 본 연구는 다양한 기계 학습 기법을 적용하여 논증 요소 분석을 자동화한 연구 결과 및 한계점을 설명하고 미래 연구에 대한 방향을 제시할 것이다.

1. 논증 자동 채점을 위한 논증의 선행 연구 검토

논증 과정에 대한 연구는 논증의 형식, 논증의 상호과정 그리고 논증과 개념 학습에 관한 연구 등 3가지 범주로 나눌 수 있으며(Yang

et al., 2009), 논증의 형식에 대한 연구에서는 Toulmin(1958)이 제시한 논증구조패턴(Toulmin's Argumentation Pattern, TAP)이 대표적으로 사용되고 있다. TAP은 주장(claim), 자료(data), 보장(warrant), 보강(backing), 반박(rebuttal), 한정어(qualifier)를 논증의 구성 요소로 정의한 후 요소 사이의 기능적 관계를 나타내는 모델로써(Driver, Newton, & Osborne, 2000), 증거를 바탕으로 한 추론으로 구성된 결론의 정당화를 논증의 개념으로 본 후 논증의 구조를 분석하였다(Kang, & Lee, 2013). Toulmin은 특정 주장의 정당성을 얻기 위해 자료, 보장 및 뒷받침을 사용하는 과정으로 과학적 주장을 구성하는 과정을 설명하였으며, 논증의 강도는 구성 요소의 조합과 존재에 기초한다(Sampson, & Clack, 2008).

Toulmin의 논증 구조와 요소를 기반으로 한 과학 논증에 대한 연구로써 Kelly, Druker, & Chen(1998)은 학생의 논증을 TAP의 구성 요소로 나눈 후, 학생들의 추론 과정 중 보장과 주장의 정당화에 초점을 맞춤으로써 논증의 정당화에 대한 전략과 방법에 대한 연구를 수행하였다. Erduran, Simon, & Osborne(2004)은 교실에서 일어나는 과학 논증에 TAP을 적용한 후, 논증 요소를 정량화하여 논증에 대한 양적, 질적 평가를 수행하였다. 특히 반박을 논증의 수준을 나타내는 중요한 기준으로 설정하여 논증의 타당도와 강도를 평가하였다. 또한 Simon, Erduran, & Osborne(2006)은 교사의 담론을 TAP을 활용한 분석을 통해 교사들이 사용하는 논증의 빈도, 복잡성을 분석하며 교실 논증을 발달시킬 수 있는 방향성을 제시하였다. 과학 논증을 분석하는 국내 과학 교육 연구에서도 TAP을 기반으로 한 논증 분석, 논증 과정의 분석틀 개발, 논증과정 평가를 위한 루브릭 개발 등 다양한 분야에서 진행되고 있다(e.g., Kang, Kwak, & Nam, 2006; Kwon, & Kim, 2016; Shin, & Kim, 2012; Yang *et al*., 2009).

논증 분석에 TAP이 널리 활용되고 있지만 논증 분석 과정에서 언어가 지닌 모호함으로 인하여 각각의 논증 요소로 구분하는 것에 어려움이 있다(Kelly, Druker, & Chen, 1998; Erduran, Simon, & Osborne, 2004; Simon, Erduran, & Osborne, 2006). 또한 논증 구조를 분석하더라도 논증의 정확성에 대한 판단으로 이어지기 어렵다는 문제점(Driver, Newton, & Osborne, 2000)과 논증 분석 결과가 단순히 요소의 빈도로만 제시된다는 문제점(Shin, & Kim, 2012)등이 드러났다.

기계 학습을 논증 분석을 위해 사용할 경우 대규모 데이터에 분석을 적용할 수 있으며 인간 채점자들에게 발생할 수 있는 실수와 불일치를 방지할 수 있어 기존의 문제점을 개선하는데 도움을 줄 수 있다(Lee *et al*., 2019; Nehm, Ha, & Mayfield, 2012; Lee *et al*., 2018; Kim, 2019). 또한 논증 분석에 많은 시간이 걸려 분석 결과가 교실 수업에 즉시적으로 반영되지 못해 논증 수업의 올바른 피드백으로써 활용되지 못하는 점이 개선될 수 있다.

2. 기계 학습

그렇다면 기계 학습과 데이터 마이닝은 무엇일까? 컴퓨터와 정보통신기술의 발달로 인하여 현대 사회의 모든 분야에서는 구조화 되어 있는 정형 데이터부터 텍스트, 오디오와 같이 구조화되어있지 않은 비정형 데이터까지 다양한 형태의 수많은 자료가 매우 빠른 속도로 생산되고 기록되고 있다. 이러한 대용량 자료에 대하여 데이터 사이의 관계, 패턴, 규칙 등을 탐색하고 모형화함으로써 유용한 지식을 추출해내는 과정을 데이터 마이닝(data mining)이라고 한다(Jun, 2015).

기계 학습이란 인공지능(artificial intelligence)의 한 분야로서 컴퓨터가 데이터로부터 판단을 할 수 있는 알고리즘을 개발하는 학문 분야이다. 특히 자료로부터 복잡한 패턴을 인식하여 예측을 수행하는 모델을 만들어내면서 미래 상황에 대한 의사결정을 이룰 수 있도록 도움을 준다(Park *et al*., 2015). 기계 학습을 수행하는 과정은 데이터 수집, 데이터의 전처리, 모델의 훈련(학습), 모델 평가, 모델 개선의 5단계로 나눌 수 있다. 데이터의 전처리란 기계 학습이 모델을 구성할 수 있도록 수집된 데이터를 병합하거나 정돈하는 과정이다. 일반적인 기계 학습의 모델을 구축하는 과정에서 전처리된 데이터는 모델을 만들기 위한 훈련 데이터(training data)와 만들어진 모델의 성능을 평가하기 위한 테스트 데이터(test data)로 분할된다(Yoo, 2019). 모델의 훈련과정에서는 교차검증(cross validation)을 수행함으로써 모델의 과적합(overfitting)을 방지한다(Figure 1).

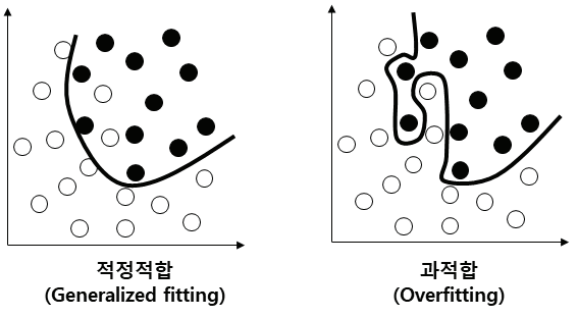

Figure 1. Generalized fitting and Overfitting

기계 학습은 모델이 학습되는 방식에 따라 지도 학습(supervised learning), 비지도 학습(unsupervised learning), 강화 학습(reinforced learning)으로 나눌 수 있다. 지도 학습이란 특정한 입력 변수에 대한 올바른 정답이 있는 데이터에 대하여 정해진 학습 방법에 따라 모델이 구성되는 것을 뜻한다. 즉, 데이터의 입력 변수와 출력 변수 사이의 관계를 찾는 모델을 최적화하는 학습 방식이다. 지도 학습은 출력 변수의 값이 레이블(label)이라 불리는 범주로 구성된 범주형 데이터를 분류(classification)하거나, 연속적 특징을 지닌 데이터의 수치를 예측하기 위한 대한 회기 분석(regression)에 주로 사용된다(Lantz, 2013). 비지도 학습이란 특정한 입력 변수에 대하여 올바른 정답이 없는 데이터에 대하여 정해진 학습 방법에 따라 모델이 구성되는 것을 뜻하며, 주로 데이터의 숨겨진 구조와 특징을 찾아내는 패턴 감지(pattern discovery)나 데이터를 비슷한 그룹으로 묶어주는 군집화(clustering)를 수행하는데 사용된다(Lantz, 2013). 마지막으로 강화 학습이란 모델이 학습하는 과정에서 얻어지는 결과를 통해 보상을 얻으면서 보상이 최대가 되는 방향으로 학습이 일어나는 것을 의미한다.

이 연구에서는 논증 요소에 대한 채점이 이루어진 학생의 논증을 바탕으로 새로운 논증 과정에 대한 채점을 수행하는 것이 목적이므로 지도 학습 중 서포트 벡터 머신(Support Vector Machine, SVM), 의사결정나무(Decision Tree, DT), 랜덤 포레스트(Random Forest, RF), 인공신경망(Artificial Neural Network, ANN)을 논증 채점의 기계 학습 기법으로 선정하였다.

요약하자면, 기계 학습이란 컴퓨터를 활용하여 데이터를 계산함으로써 새로운 데이터에 대한 예측이나 데이터 사이의 관계를 얻는 과정이다. 그러므로 기계 학습은 자료 주도적으로 진행된다고 표현할 수 있는데, 이 과정에서 연구자가 데이터에 대한 내용학적 지식을 가지고 있지 못하다면 모델이 표현하는 의미와 가치를 평가하기 어렵다. 따라서 기계 학습을 사용하고자하는 연구자는 기계 학습에 대한 기본적 이해뿐만 아니라 모델의 결과를 해석하기 위한 내용학적 지식을 갖추는 것이 반드시 필요하다(Yoo, 2019). 즉, 교육 분야에서 기계 학습은 교육 현장에서 산출된 데이터를 기계 학습이라는 도구를 활용하여 모델을 형성한 후, 교육학 지식이라는 전문성을 바탕으로 모델을 해석, 적용하는 과정으로 진행된다(Yoo, 2019).

3. 기계 학습의 알고리즘

앞서 이야기한바와 같이 기계 학습에는 다양한 알고리즘이 사용되며 논증 요소를 분류하기 위해 이 연구에서 사용된 기법들인 서포트 벡터 머신, 의사결정나무, 랜덤 포레스트, 인공신경망을 간단히 소개하고자 한다.

서포트 벡터 머신(Cortes, & Vapnik, 1995)은 Figure 2와 같이 2차원의 공간에 두 가지의 데이터가 주어졌을 때, 하나의 직선(초평면, hyperplane)으로 공간을 분류함으로써 자료를 분류하는 알고리즘이다.

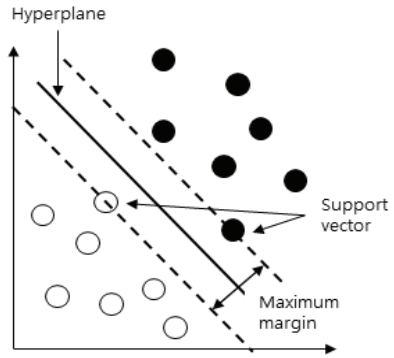

Figure 2. An example of a separable problem in a 2 dimensional space using support vector machine

그러나 교육 현장에서 발생하는 데이터뿐만 아니라 대다수 데이터의 변수들은 2차원으로 구현되지 않으며 비선형적이다. 이러한 문제를 해결하기 위해 Figure 3와 같이 서포트 벡터 머신은 비선형 데이터를 차원을 변화시킨 새로운 공간(특성 공간, feature space)으로 옮기는 방법(매핑, mapping)을 통해 경계면을 찾는다. 이러한 기법을 커널 트릭(kernel trick)이라 하는데 커널이란 변수의 차원을 확장시킬 때 사용하는 계산 기법을 의미한다(Lantz, 2013).

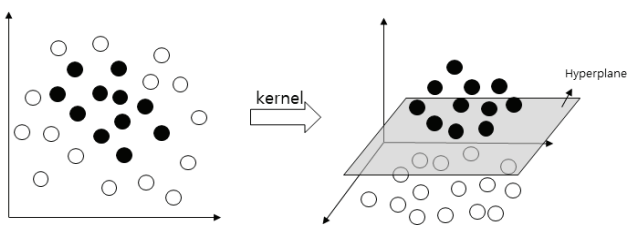

Figure 3. An example of using kernel trick

의사결정나무는 훈련 데이터의 입력 변수를 기준으로 하향식(top-down)으로 데이터를 이진 분할(recursive binary splitting)하는 방법을 사용한다. 변수의 관계를 모델링하기 위해 트리(tree) 구조를 활용하기 때문에 의사결정나무로 불리며(Lantz, 2013) 의사결정나무의 트리는 Figure 4와 같이 변수를 표현하는 마디(node)와 가지(branch)로 구성되어 있다(Park *et al*., 2015).

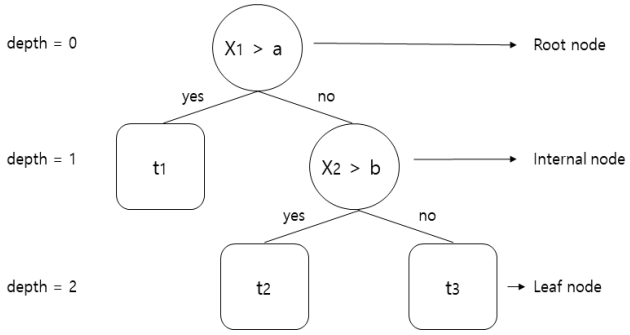

Figure 4. An example of decision tree model

의사결정나무의 모델을 만드는 과정은 크게 성장(growing)과 가지치기(prunning)로 이루어진다. 성장이란 각 마디에서 최적의 분할을 수행하기 위하여 입력 변수를 선택하는 과정이 반복되는 과정으로, 마디에 있는 데이터가 거의 같은 범주로 분류가 되거나 정해진 나무의 깊이(depth)와 같은 정지 규칙(stopping rule)에 도달할 때까지 성장이 진행된다(Lantz, 2013). 그러나 모델의 분류가 너무 세분화되는 경우 훈련 데이터에 모델의 과적합이 발생하게 된다. 의사결정나무에서는 과적합을 해결하기 위하여 불필요한 마디와 가지를 제거하고 병합(merge)하는 가지치기(pruning)를 수행한다(Park *et al*., 2015).

랜덤 포레스트는 Breiman(2001)에 의해 개발된 기계 학습으로 주어진 훈련 데이터로부터 다수의 독립적인 의사결정나무를 만든 후, 개별 트리의 예측 결과에 대한 투표를 통해 하나의 예측 결과를 만들어낸다. 랜덤 포레스트와 같이 약한 성능을 지닌 기계 학습 모델을 결합하여 최종적으로 강한 성능을 지닌 하나의 기계 학습 모델을 만드는 과정을 앙상블(ensemble)이라고 한다(Lantz, 2013). 또한 랜덤 포레스트는 모델을 구성하는 과정에서 최대한으로 무작위성을 높이는 과정을 통해 분산을 줄임으로써 예측력을 높이게 된다(Park *et al*., 2015).

인공신경망은 뇌의 생물학적 신경망 구조인 뉴런(neuron)의 네트워크를 착안하여 입력 변수에 따른 출력 변수의 관계를 모델링하는 기계 학습 기법이다. 인간의 뇌는 여러 개의 뉴런이 상호 연결되어 정보를 처리하는 것과 같이 인공신경망은 노드를 연결선(link)으로 상호 연결하여 예측을 수행하는 것이다. 초기에는 컴퓨터가 복잡한 신경망의 데이터를 효과적으로 처리하지 못하는 문제로 인하여 활발히 연구가 진행되지 못하였으나, 1980년대에 들어서면서 컴퓨터의 발달 및 새로운 알고리즘의 개발에 따라 다시 각광을 받게 되었다(Park *et al*., 2015). 인공신경망 모형은 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)로 이루어져있으며 입력층의 노드의 수는 입력 변수와 같으며 출력층의 노드의 수는 출력 변수의 수와 같다.

Table 1은 위에서 설명한 4가지 기계 학습의 특징을 간략히 정리한 것이다.

4. 전문가 시스템

전문가 시스템이란 간단히 정의하면 인간 전문가의 의사 결정 능력을 모방하는 컴퓨터 시스템이다(Giarratano, & Riley, 1998). 전문가의 의사 결정 과정을 모방하기 위해 휴리스틱 기법이 사용되는데 이는 경험이나 직관, 상식 등을 결론 도출의 기초로 사용하는 것이다. 예를 들어 이 과정은 if-then구조의 생성 규칙(production rule)의 집합으로 표현될 수 있다. 전문가 시스템 중 가장 널리 사용되는 규칙기반 전문가 시스템(rule-based expert system)은 지식을 생성 규칙의 집합으로 표현하여 문제를 해결하는 시스템이다(Negnevitsky, 2005). 불확실성을 지닌 문제에서 휴리스틱 알고리즘의 사용은 정확도와 정밀도의 손실을 초래할 수도 있지만, 전문가의 경험적 지식을 통해 문제를 빠르고 효율적으로 해결할 수 있다는 장점이 있다(Negnevitsky, 2005; Russell, & Norvig, 2016).

규칙기반 전문가 시스템은 기반 지식(knowledge base), 추론 엔진(inference engine), 사용자 인터페이스(user interface) 등의 요소로 구성된다(Figure 5). 기반 지식은 문제 해결에 필요한 특정 분야의 전문 지식을 뜻하며, 이 지식은 규칙(rules)의 집합과 사실(facts)의 집합으로 이루어져있다. 추론 엔진은 기반 지식의 규칙과 사실을 연결하여 결론을 추론하는 시스템의 사고과정이다(Baek, 1989; Negnevitsky, 2005). 기반 지식과 추론 엔진을 설계하는 과정은 기존 자동 채점 연구에서 자질 설계(feature engineering) 과정과 유사하다(Kim, 2019). 즉, 인간 전문가의 지능을 모방한 규칙기반 전문가 시스템은 인간 전문가의 규칙과 사실로부터 기반 지식을 구성한 후, 문제 해결을 위한 생성 규칙을 휴리스틱 알고리즘으로 적용한 인공지능이라고 할 수 있다. 전문가 시스템의 장점은 실시간으로 전문가 수준의 반응을 얻을 수 있고, 전문가의 문제 해결 과정을 쉽게 사용할 수 있으며, 비용이 저렴하다는 것이다. 또한 여러 명의 전문가들에 의해 설계된 하나의 전문가 시스템은 한 명의 인간 전문가보다도 많은 지식과 문제해결능력을 확보할 수 있게 된다(Giarratano, & Riley, 1998).

이러한 전문가 시스템은 기존 컴퓨터 보조 수업(computer assisted instruction)의 문항 설계 과정과 유사하다. 하지만 컴퓨터 보조 수업에서는 전문가가 문항을 개발하는 역할을 수행하지만 전문가 시스템에서는 전문가의 문제 해결 과정을 탐색하고 규칙을 설계하여 인공지능에 적용한다는 점에 큰 차이가 있다(Lippert, 1989). 또한 전문가의 축적된 기반 지식이 빅데이터로 기계 학습에 적용된다면 전문가 시스템은 컴퓨터 보조 수업을 크게 발전시킬 수 있는 도구가 될 것이다.

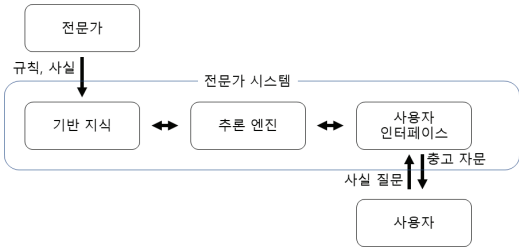

Figure 5. Architecture of an expert system

Ⅲ. 연구 방법 및 내용

이 연구는 기존 R 패키지를 활용하여 논증 자동 채점의 정확도와 논증 요소를 추출하는 방법을 개선하기 위해 논증의 자동 채점을 수행하는 과정의 각 단계마다 도출된 모델을 평가하고 기계 학습의 성능을 개선하는 과정을 반복 수행하였다.

특히, 본 연구는 새로운 자동채점 접근 방법으로 한국어 채점의 정확도를 높이기 위해 불용어 처리를 도입하고 전문가의 논증 채점 과정을 분석하여 기계 학습에 휴리스틱으로 적용하는 전문가 시스템을 적용하였다.

1. 과학적 말하기 논증의 발화 자료 수집

논증의 발화 자료는 2018학년도 1~2학기에 걸쳐 경기도 Y시에 소재의 B고등학교의 1학년 9개 학급 250명(남132명, 여118명)과 2학

Table 1. Features of machine learning method(Kevin, 2012; Lantz, 2013; Fernández-Delgado *et al*., 2014; Park *et al*., 2015; Yoo, 2015)

	서포트 벡터 머신	의사결정나무	랜덤 포레스트	인공신경망
기본 개념	・초평면으로 공간을 분류함으로써 자료를 분류	・데이터를 여러 개의 작은 부분 집합으로 분할하는 과정을 반복하여 부분 집합의 자료가 동질적이도록 분류	・다수의 독립적인 의사결정나무를 만든 후, 개별 트리의 예측 결과에 대한 투표를 통해 하나의 예측 결과를 생성	・뇌의 생물학적 신경망 구조인 뉴런의 네트워크를 착안하여 입력 변수에 따른 출력 변수의 관계를 생성
장점	・예측력이 전반적으로 높음 ・다양한 형태의 데이터에 적용이 잘됨	・생성된 모델을 통해 변수의 중요성과 관계를 쉽게 해석하고 설명할 수 있음	・각각의 의사결정나무가 수행한 예측보다 예측력이 높음 ・연구자가 나무의 성장 변수를 결정하지 않아도 됨	・분류와 예측의 문제에 모두 적용할 수 있음 ・실제 생활에서 발생하는 복잡하고 다양한 자료에 대한 예측이 뛰어남
단점	・모델을 찾는 과정에서 여러 커널 함수를 결정해야 함 ・예측 결과를 변수로 해석하기 어려운 모델이 생성	・설명력에 비하여 예측력이 떨어짐 ・하나의 마디에는 하나의 변수만을 선택하게 되므로 모형의 안정성이 부족함	・의사결정나무에 비하여 생성된 변수간의 관계를 해석하기 어려움	・훈련 데이터에 대한 과적합 문제가 발생할 수 있음 ・생성된 모델에서 변수 간의 관계를 해석하기 어려움
특징	・커널 트릭을 통해 변수의 차원이 확장된 모델을 생성할 수 있음	・나무를 얼마나 성장시킬 것인지에 대한 변수를 결정해야 함 ・출력변수가 범주형인 경우 불순도(impurity)를 감소시켜 마디의 동질성이 증가하도록 모델을 생성함	・의사결정나무의 수, 마디에서 설명 변수의 수를 결정해야 함 ・변수를 해석하는 문제를 해결하는 방안으로서 중요도 지수(variable of importance index)나 부분 의존성 도표(partial dependence plot)을 사용할 수 있음	・은닉층의 수와 은닉 노드의 수를 결정해야 함 ・과적합을 해결하기 위해 은닉 노드 중 중요도가 떨어지는 노드를 제거하는 드롭 아웃(drop out)이나 가중치 감소(weight decay)를 수행할 수 있음

년 자연계열 3개 학급 103명(남 76명, 여 27명)을 대상으로 한 소집단 말하기 논증 연구의 전사본이다. 이 중 2학년 학생들은 2017학년도 1학년 2학기에 7개의 과학관련 사회쟁점(SSI) 주제에 대하여 총 21차시의 과학 논증 수업에 참여한 경험이 있다. 2018학년도에 진행한 과학 논증 수업은 1학년의 경우 6개의 주제에 대하여 총 18차시 진행하였으며, 2학년의 경우에는 8개의 주제에 대하여 총 22차시 진행하였다. 선정된 주제의 예시는 Table 2와 같다.

한 주제 당 논증 문항은 평균 3개로 구성되어있으며, 2015개정 교육과정 1학년 과학과 2009개정 교육과정 2학년 물리Ⅰ, 화학Ⅰ, 지구과학Ⅰ을 고려하여 과학 교육 전문가 1인, 현직 과학교사인 박사과정 1인 및 석사 과정 5인, 그리고 연구 참여 학교의 과학 교사들과의 합의, 수정, 검토과정을 통해 개발되었다. 논증의 기록을 위해 모든 소집단을 대상으로 360도 VR 카메라를 설치하여 논증의 전체 과정을 촬영을 하였다. 360도 VR 카메라는 소집단 중심에서 촬영을 진행하기 때문에 고른 음량으로 언어적 상호작용을 기록할 수 있는 장점을 가진다. 2018년의 자료 수집 결과, 총 1,167차시(1학기 516차시, 2학기 651차시)의 논증이 촬영되었다.

이 중 2017학년도에 과학 논증 수업에 참여한 경험이 있는 2학년 학생들의 말하기 논증 중 화학Ⅰ 분자 구조 주제의 논증 문항 1번과 2번에서 발생한 발화를 연구 대상으로 선정하였고, 논증에 적극적으로 참여한 5개의 소집단을 초점 집단으로 선정하여 총 10차시의 영상을 전사하였다. 그 결과 2,605개의 학생 발화를 분석 자료로 선정하였다.

2. 과학적 말하기 논증에서 채점 모델 설정

과학 논증을 분석하는 연구는 탐구에 따른 사고의 인식론적 변화(Kwon, & Kim, 2016; Lee, & Nam, 2016), 탐구 활동 중 논증의 초점 요소에 따른 분석(Lee, & Nam, 2018), 과학 글쓰기의 논증 분석(Cho, & Nam, 2014)등 다양한 방법으로 진행되고 있는데, 이러한 연구는 과학적 말하기 논증의 전체 과정을 분석하기에는 어려움이 있다. 또한 과학 수업에서 이루어지는 소집단 토론(Park, & Kim, 2018; Kwon, & Kim, 2016; Lee, & Nam, 2016)에 대한 연구가 이루어지고 있으나 각 연구에서 제시하는 연구 방법 및 평가 모델이 서로 다르다는 문제점이 있다. 그러므로 본 연구에서는 기존 연구를 바탕으로 하여 과학적 말하기 논증을 분석하는 과정에 적용할 수 있는 포괄적인 기계 학습 채점 모델을 설정하였다.

Toulmin(1958)은 논증 유형(TAP: Toulmin's Argumentation Pattern)을 분석하는 과정에서 논증의 요소로서 주장(claim), 자료(data), 보증(warrants), 보강(backing), 한정어(qualifier) 그리고 반박(rebuttal)으로 분류하여 논증을 분석하였다. 과학 교육에서는 Driver, Newton, & Osborne(1998)로부터 본격적으로 과학 수업 중 논증 유형에 대한 연구가 진행되었으며, Osborne, Erduran, & Simon (2004)와 Erduran, Simon & Osborne(2004)은 TAP을 바탕으로 한 논증 요소를 개발하여 과학 논증 수업 중 발생한 학생의 발화를 분석하였다. 이 외의 연구에서 각각의 논증 요소를 개발하여 과학 논증을 분석했으며, 각 논문에서 제시된 논증 요소는 Table 3와 같다.

Toulmin(1958)은 논증 요소에 따른 논증 구조와 패턴에 초점을 맞추어 논증을 분석하였으나 분류 기준이 모호하여 하나의 발화에 대하여 여러 범주로 분류할 수 있기 때문에 교실에서 발생한 과학 논증의 요소를 분석하는 과정에서 연구자 사이의 코딩의 신뢰도가 떨어지는 문제점이 있다(Erduran, Simon, & Osborne, 2004; Sampson, & Clark, 2008). Zohar, & Nemet(2002)은 TAP의 자료, 보장, 보강을 하나의 범주로 묶으면서 TAP의 분석 과정에서 발생하는 신뢰성과 타당성 문제를 해결하는 장점(Sampson, & Clark, 2008)을 가졌지만, 명시적 결론(explicit conclusion), 암묵적 결론(implicit conclusion)과 같이 세부적 분류 기준으로 인하여 자동 채점에서 신뢰성을 확보하기 어렵다. McNeill *et al.*(2006)은 논증의 구조로서 주장,

Table 2. Examples of topics in the scientific argumentation(Kim, & Ryu, 2019)

학년	과목	연관 단원	논증 주제	논증 문항
2018 학년도 1학년	통합 과학	Ⅱ-1. 역학적 시스템	놀이기구의 운동과 안전	1. 세계의 유명한 놀이기구 1개를 소개하고, 이에 적용된 운동량과 충격량에 대해 설명해 보자. 2. 1번에서 소개한 놀이기구와 관련된 안전사고(가능성)를 2가지 이상 설명하고, 각각의 해결책을 과학적 근거를 들어 제시해 보자. 3. 본인이 세계적으로 유명한 놀이기구 제작자라고 가정하고, 스릴 있으면서도 안전한 놀이기구를 고안해 보자. 그리고 그렇게 만든 이유는 무엇인지 설명해 보자.
2018 학년도 2학년	화학Ⅰ	Ⅲ-3. 분자의 구조	분자 구조	1. 물과 암모니아는 분자량이 비슷함에 불구하고 물의 끓는점은 100℃인 반면 암모니아는 −33.4℃이다. 그 이유를 분자 구조와 연관 지어 설명해 보자. 2. 물이 굽은형이 아닌 경우 나타날 수 있는 자연 현상 및 생활 속의 변화를 생각해보자. 3. 물이 굽은형인 이유를 설명할 수 있는 모형을 재활용품을 이용하여 제작하고 이 모형이 갖는 장점과 한계에 대해 토론하시오.

Table 3. Argumentation code of reference articles

Reference	Argumentation code					
Toulmin(1958)	Claim	Data	Warrant	Backing	Rebuttal	Qualifier
McNeill *et al.*(2006)	Claim	Evidence	Reasoning		-	-
Erduran, Simon, & Osborne(2004)	Claim	Data	Justification		Rebuttal	-
Osborne, Erduran, & Simon(2004)	Claim	Ground			Rebuttal	-
Zohar, & Nemet(2002)	Argument, Counter-argument, Rebuttal, Conclusion, Opposition, Justification					

증거, 추론을 제시하였으나 반박 요소를 논하지 않았다. 반박이 논증의 질을 나타내는 중요한 지표(Erduran, Simon, & Osborne, 2004)임에도 불구하고 반박 요소가 누락된다면 좋은 논증 평가가 이루어질 수 없다는 문제가 발생한다. 이에 따라 본 연구에서는 논증 자동 채점 과정에서 논증 요소를 적절하게 분류하였으며, 연구 과정에서 TAP을 통한 교육적 전략을 탐색한 Erduran, Simon, & Osborne(2004)의 논증 요소를 바탕으로 주장, 자료, 정당화, 반박, 비논증이라는 5가지의 채점 모델을 개발하였다. 또한 Erduran, Simon, & Osborne(2004)의 논증 요소는 논증 자동 채점에 관한 선행 연구(Lee *et al*., 2018)와 동일한 요소이므로 기계 학습을 적용하는 과정의 채점 모델로서의 타당성을 확보할 수 있다. 각 논증 요소 채점의 세부 기준은 학생의 발화를 분석하는 과정에서 연구자 간의 협의를 통하여 개선하였다.

주장(Claim, C)이란 논증 과정에서 자신의 의견에 대한 일반적인 수용을 위해 공개적으로 제시된 발화이며 자신의 의견에 대한 추측과 예측을 주장의 범주로 포함한다. 그리고 학생의 지식, 관찰 사실 등을 바탕으로 구성된 자신의 의견을 전달하는 행위도 주장으로 분류한다. 자료(Data, D)란 과학적 사실(현상, 단어, 수치 등)을 언급하거나 과학적 사실에 대한 단순한 재확인 및 수정에 대한 발화를 자료로 분류한다. 정당화(Justification, J)란 보장이나 보강과 같이 자료와 주장, 자료와 자료 사이의 논리적인 연결을 하며 결론을 뒷받침하는 발화이다. 또한 학생의 논증 과정에 많이 나타나는 주장을 뒷받침하기 위하여 학생이 과학 지식을 언급하는 경우를 주장과 자료 사이의 논리적 연결의 역할로 해석하여 정당화로 코딩하였다. 반박(Rebuttal, R)이란 다른 학생의 주장의 힘을 약화시키는 발화로서 특별하거나 예외적인 자료와 의견을 전달하는 것을 포함할 뿐만 아니라 다른 학생 주장의 정당화를 추가적으로 요구하여 주장의 힘을 약화시키는 것 또한 포함한다(Erduran, Simon, & Osborne, 2004; Jiménez-Aleixandre, Rodriguez, & Duschl, 2000; Kang, Kwak, & Nam, 2006; Lee *et al*., 2018; McNeill, & Krajcik, 2007). 제시된 4가지의 코드로 분류되지 않는 일상적 대화, 논증 주제와 무관한 수업 활동을 위한 의사진행 발화는 비논증(None, N)으로 분류하였다. Table 4는 각 코드에 분류된 학생 발화의 예시이다.

본 연구에서는 코딩의 타당성을 확보하기 위하여 과학 교육 전문가 1인, 박사과정 1인 및 석사 과정 4인의 협의를 통하여 코딩을 진행하였다. 또한 코딩 과정 중 한 문장의 발화 속에 여러 가지의 코드 요소가 혼재되어있을 때에는 접속 부사나 어미를 기준으로 발화를 분리하는 과정을 진행하였다. 코딩의 결과 각 요소에 따른 발화의 수는 총 2,605개(주장 441개, 자료 608개, 정당화 475개, 반박 141개, 비논증 940개)이다. 코딩 결과 비논증이 가장 많은 비율을 차지하는데, 이는 학생의 발화에서 논증과 논증이 아닌 것이 섞여 있는 것이 자연스러운 보통의 교실 속 상호작용의 모습이 반영된 결과이다

3. 연구 개괄

연구의 수행 과정은 R의 기계 학습 패키지를 활용하여 논증 요소를 예측하는 과정이다. 기계 학습을 활용하는 과정에서 발생하는 문제점을 각 단계별로 분석하며 개선해나가는 과정을 반복하였으며, Figure 6를 통해 본 연구의 진행 과정을 요약하였다. 텍스트 전처리 및 기계

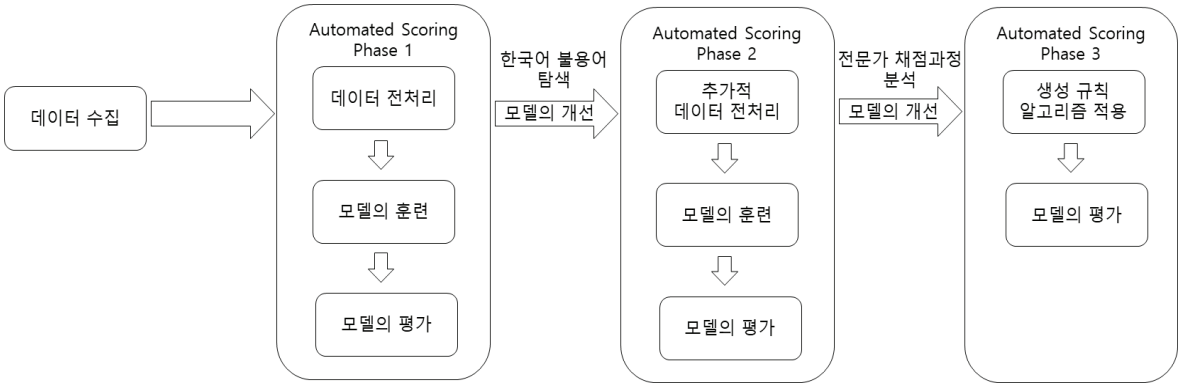

Figure 6. A flow of research process

Table 4. Example of student's argumentation

코드	정의	학생 발화 예시
주장	・의견에 대한 일반적인 수용을 위해 공개적으로 제시된 발화	・결합각이 크다는게 반발력이 크다는거지. ・인력이 높을수록 끓는점이 높아요.
자료	・과학적 사실의 언급 ・과학적 사실에 대한 재확인 및 수정	・물의 구조식을 보면 결합각이 107도인가 그렇데. ・암모니아 극성 맞지?
정당화	・자료와 주장, 자료와 자료 사이의 논리적인 연결을 하는 발화	・무극성은 분산력만 있고, 극성은 분산력과 정전기적 인력이 있어서 그래. ・근데 여기서 기하 벡터 개념이 살짝 들어가면 전자기적 힘이 여기는 -를 띄게 되고, 여기도 -를 띄게 되겠죠?
반박	・다른 학생의 주장의 힘을 약화시키는 발화	・근데 둘 다 수소결합하는거 아니야? ・일단 각이 큰 이유는 비공유 전자쌍하고 공유 전자쌍 사이의 반발력 때문이 아니야?
비논증	・일상적 대화 혹은 논증 주제와 무관한 발화 ・의사진행을 위한 발화	・미안한데 좀만 더 크게 말해줘 ・근데 어차피 분자 간의 결합인데 지금 여기서 말하는 건 분자 구조랑 관련해서 말하래잖아. ・음, 내가 조사한건 다 말했어.

학습은 R 3.6.0, R studio 1.2.1335, Python 3.7.3과 Jupyter Notebook 6.0.0을 통해 수행되었다.

4. 기계 학습을 위한 데이터 전처리

논증 발화에 대한 자연어 처리의 시작점은 비정형 데이터인 언어를 기계 학습에 적용할 수 있도록 정형 데이터인 숫자로 변환하는 것이다. 이를 위하여 긴 문자열을 작은 단위인 토큰(token)으로 변환시키는 토큰화(tokenization)를 진행해야 한다. 영어의 경우는 띄어쓰기를 기준으로 토큰 생성이 가능하지만 한국어는 교착어(agglutinative language)의 특성으로 인해 형태소 분석을 통하여 토큰을 생성해야 한다. 형태소 분석이란 학생 발화와 같은 말뭉치를 말의 가장 작은 단위인 형태소로 분리한 후 각각의 형태소에 품사(part of speech, POS)를 부착(tagging)하는 과정을 의미한다.

자연어 처리를 위한 Python의 대표적인 패키지로서 'KoNLPy'가 있으며, 'Hananum', 'Kkma' 등 총 5가지의 형태소 분석기를 지원한다. 본 연구에서는 형태소 분석기 중 기존 텍스트 마이닝 연구에서 많이 활용되고 우수한 성능을 나타내는 '꼬꼬마(Kkma)' 형태소 분석기(Lee *et al*., 2010; Park, & Cho, 2014)를 활용하여 코딩된 학생의 발화에 대한 토큰화를 수행하였다. 'Kkam' 형태소 분석기는 한글 형태소의 품사를 '체언(N), 용언(V), 관형사와 부사(M), 감탄사(I), 조사(J), 어미(E), 접사와 어근(X), 부호(S), 한글 이외(O)'의 기준으로 자연어의 형태소와 품사를 구분한다. 예를 들어 'Kkma' 형태소 분석기를 통해 '과학은 재미있다'라는 문장을 분석하면 '과학/NNG 은/JX 재미있/VA 다/EFN'와 같이 '형태소/품사'의 형태로 토큰화된 결과를 얻을 수 있다.

코딩된 2,605개의 발화를 Jupyter Notebook과 'Kkma' 형태소 분석기를 통해 토큰화를 진행한 후 생성된 결과를 excel파일로 추출하여 형태소 분석기의 성능을 검토하였다. 그 결과 학생의 발화 중 일부가 구어체의 특성과 형태소 분석기의 성능으로 인하여 올바르게 토큰화가 일어나지 않는 문제점을 확인하였다. 이렇게 전처리 과정에서 문제가 발생하면 비정형 데이터인 자연어가 정형 데이터로 수치화되는 과정에서 왜곡이 발생하게 되므로 기계 학습의 데이터가 오염된다. 예를 들어 '그니까'라는 단어는 학생이 이유를 이야기하는 과정에서 많이 쓰이는 단어로서 '그러니까'의 축약어이다. 두 단어에 대한 형태소 분석을 진행했을 때 두 단어가 같은 의미를 가졌음에도 불구하고 서로 다른 형태소로 분류되는 문제가 발생한다. 이로 인하여 하나의 동일한 의미를 가졌음에도 서로 다른 2개의 정형 데이터로 작용하여 데이터가 흩어지는 문제점이 발생한다. 그러므로 연구자는 정형 데이터가 흩어지지 않도록 적절한 데이터 전처리 과정을 수행해야 한다. 이외의 검토된 문제점들을 해결하기 위해 발화의 전처리 과정에서 구어체와 같은 불확실한 발화에 대해서는 맞춤법 검사 등을 통해서 전사된 문장의 구조와 의미를 유지하며 수정하였다. 또한 형태소 분석기의 미등록 단어를 검토하여 사용자 사전에 추가하였으며 잘못된 토큰화가 이루어진 형태소에는 추가적인 작업을 수행하여 올바르게 수정하였다. Python의 'Kkma' 형태소 분석기를 통해 토큰화된 학생의 발화와 형태소는 excel파일을 통해 R로 입력하였고, 이후 기계 학습 과정은 R에서 진행되었다.

입력된 학생의 발화와 형태소는 R의 'tm' 패키지의 Corpus()와 DocumentTermMatrix()함수를 통해 문서단어행렬(document-term matrix, DTM)로 변환하였다. 문서단어행렬이란 문서에 나타난 각 단어의 빈도를 행렬로 표현한 것으로 일반적으로 한국어의 경우 토큰화된 형태소가 단어의 역할을 수행한다. 각 단어의 빈도는 문서단어행렬의 가중치(weighting)로 표현되며 weightTfIdf() 함수를 통해 TF-IDF(term frequency-inverse document frequency)값으로 계산하였다. 생성된 DTM의 행에는 2,605개의 문장이 배열되었으며, 열에는 토큰화된 형태소 1,968개가 배열되었다.

5. 기계 학습과 자동 채점의 수행

가. Phase 1. 모든 학생 발화와 형태소를 대상으로 한 논증의 자동 채점

Phase 1에서는 2,605개의 모든 학생 발화로 구성된 문서단어행렬을 데이터로 활용하여 자동 채점을 수행하였다. 기계 학습을 수행하기 전에 수집된 데이터를 모델 구성을 위한 훈련 데이터와 모델의 평가를 테스트 데이터를 분리해야 한다. 데이터의 분리는 'carte' 패키지의 createDataPartition()함수를 통해 이루어졌으며, 2,605개의 문장에 대하여 채점된 논증 요소를 기준으로 각각 80%의 훈련 데이터와 20%의 테스트 데이터로 분류하였으며 결과는 Table 5와 같다.

분류가 이루어진 훈련 데이터를 활용하여 채점 모델을 만들기 위해 서포트 벡터 머신, 의사결정나무, 랜덤 포레스트, 인공신경망을 활용한 지도 학습을 수행하였다. 앞서 이야기한 것과 같이 지도 학습의 기본적인 단계는 레이블이 주어진 훈련 데이터와 테스트 데이터를 구성한 후, 훈련 데이터를 기계 학습에 적용하여 학습된 예측 모델을 구축한다. 이후 만들어진 모델을 테스트 데이터에 적용하며 기계 학습 모델의 성능을 평가하게 된다. 이 연구에서는 기계 학습의 R 패키지로써 'e1071', 'rpart', 'randomForest', 'nnet'을 사용하였다.

논증 자동 채점을 위한 기계 학습 모델을 구성하는 과정에서 각 기법의 변수 역할을 수행하는 파라미터(parameter)와 하이퍼 파라미터(hyper parameter)를 결정해야한다. 기계 학습을 활용한 채점 모델의 성능을 높이기 위해서는 하이퍼 파라미터의 최적값을 찾아야하며 이 과정을 하이퍼 파라미터 튜닝(tunning)이라 한다. 하이퍼 파라미터 튜닝은 다양한 하이퍼 파라미터 값을 변화시키면서 기계 학습에 적용하는 과정을 반복 수행하여 모델의 예측 정확도가 높은 최적의 값을 찾는 휴리스틱 방법을 통해 결정된다.

서포트 벡터 머신은 Figure 2와 같이 최적의 분류를 수행할 수 있는 경계면을 찾아서 데이터 사이의 거리인 마진(margin)을 최대로 하는

Table 5. Classified result of coded argumentations

	주장	자료	정당화	반박	비논증	총
훈련 데이터	335	485	380	118	747	2,085
테스트 데이터	86	123	95	23	193	520

것이다. 서포트 벡터 머신을 사용하는 과정에서 연구자가 결정해야하는 하이퍼 파라미터는 cost, gamma와 커널 함수이다. cost는 경계면이 어느 정도의 오차를 허용할 것인지를 결정하는 파라미터이며 gamma는 하나의 훈련 데이터가 어느 정도의 영향력을 미치게 되는지를 결정하는 파라미터이다. 하이퍼 파라미터를 결정하기 위해 cost, gamma, 커널함수를 변화시키면서 구한 예측 결과를 비교하여 가장 좋은 결과를 나타낸 파라미터를 선택하였다.

의사결정나무는 하향식의 트리 구조를 통해 데이터를 예측하는 기계 학습으로써 연구자는 모델의 트리 구조를 통해 분류에 대한 통찰력을 얻을 수 있다. 이 과정에서 연구자는 어느 정도 나무를 성장시킬지, 성장한 나무를 얼마나 가지치기를 할지 판단해야한다(Yoo, 2015). 이 때 트리의 오분류(misclassification)와 복잡도(complexity)를 표현하는 복잡성 매개변수(complexity parameter)는 가지치기와 트리의 최대 크기를 조절하기 위한 변수로 사용할 수 있다(Jun, 2015). 본 연구에서는 의사결정나무 기계 학습을 수행한 후 모델의 복잡성 매개변수의 최적값을 구한 후, 가지치기를 수행하여 얻어진 개선된 모델을 통해 의사결정나무의 모델을 얻었다.

랜덤 포레스트는 같은 데이터에서 다르게 만들어진 여러 개의 의사결정나무의 앙상블을 통해 예측을 수행하는 방법이다. 랜덤 포레스트는 의사결정나무에서 설정한 나무의 성장, 가지치기와 같은 하이퍼 파라미터를 설정하지 않아도 되지만(Yoo, 2015) 하이퍼 파라미터로써 어느 만큼의 의사결정나무를 만들 것인지(ntree)와 각 마디에서 선택되는 특징 개수(mtry)를 판단해야한다. 본 연구에서 mtry는 Breiman(2001)이 범주형 데이터에 대해서 제시한 $\sqrt{p}$ (p는 훈련 데이터의 특징 개수)를 따랐다.

인공신경망은 입력 변수에 따른 출력 값 사이의 관계를 나타내는 인공 뉴런을 모델링하여 변수 사이의 복잡한 비선형적 관계를 예측하는데 좋은 성능을 지닌 기계 학습 기법이다(Lantz, 2013). 인공신경망이 구성되는 과정에는 뇌의 시냅스와 같은 은닉층이 존재하며 연구자는 하이퍼 파라미터로서 은닉층의 수(size)를 결정해야 한다. 본 연구에서는 은닉층의 수를 1~5개로 변화시키며 error값의 평균을 비교하여 파라미터를 결정하였다. 또한 가중치 감소를 통해 과적합을 해결하는 하이퍼 파라미터 decay는 Ripley, & Venables(2020)의 nnet 수행과정을 따랐다.

본 연구에서 사용한 기계 학습 R 패키지와 파라미터 값은 Table 6와 같다. 이외의 파라미터는 패키지의 기본(default) 파라미터로 진행되었다.

Phase 1에서는 Table 6의 하이퍼 파라미터가 적용된 기계 학습 기법에 훈련 데이터(2,085개의 학생 발화)의 모든 형태소(1,968개)를 활용하여 논증 자동 채점 모델을 생성하였다. 이후 생성된 채점 모델을 평가하기 위하여 테스트 데이터(520개의 학생 발화)에 대한 자동 채점을 수행하였다.

나. Phase 2. 불용어 처리된 학생 발화와 형태소를 대상으로 한 논증의 자동 채점

Phase 1 훈련 데이터의 형태소에는 의미를 지닌 체언과 용언뿐만 아니라 문법적 역할을 수행하는 조사(e.g., ~이, ~가) 등이 포함되어 있다. 하지만 문법적 역할을 수행하는 조사는 학생의 논증 과정에서 의미를 지니지 않기 때문에 자동 채점 과정에서 제외되어야 한다. 즉, 논증 자동 채점에서 논증 요소에 영향을 미치지 않는 조사와 같은 문법적 형태소는 불용어(stop word)로써 처리되어야한다(Kil, 2018). 그러나 영어와는 다르게 한국어는 아직 불용어에 대한 연구가 많이 진행되어있지 않아 텍스트 전처리의 어려움이 있었으나 Kil(2018)이 텍스트 마이닝을 위한 한국어 불용어 목록을 제안하였다.

이 연구에서는 Kil(2018)이 제안한 불용어 목록을 참고하여 학생 논증 발화에 대한 불용어 처리를 수행하였다. 이 과정에서 Kil(2018)은 '어서/ecd', '이니까/ecd'등과 같은 어미가 문법적 역할을 수행하기 때문에 불용어로 분류하였으나 본 연구에서 논증을 코딩하는 과정에서 어미가 논증을 표현하는 주요 형태소로 작용하는 점을 확인하였다. 따라서 본 연구에서는 어미에 대한 불용어 처리를 수행하지 않았다. 불용어 처리를 통해 조사, 접두사, 접미사, 어근 그리고 분석이 불가능한 형태소를 추출하여 210개의 형태소를 제거하였다. Phase 2에서는 불용어 처리가 진행된 2,085개의 학생 발화와 1,758개의 형태소를 활용하여 DTM을 새롭게 구성한 뒤 기계 학습을 수행하여 채점 모델을 생성하였다. Phase 1과 동일한 방식으로 Phase 2의 채점 모델을 평가하기 위하여 테스트 데이터에 대한 자동 채점을 수행하였다.

다. Phase 3. 전문가 형태소를 생성 규칙 알고리즘으로 적용한 한 논증의 자동 채점

Phase 2에서 불용어 처리가 추가된 텍스트 전처리 과정을 수행했음에도 불구하고 5가지의 코드 요소 중 반박에 대한 정확도는 여전히 낮게 얻어지는 문제점이 있었다. 반박 채점의 어려움은 선행연구(Lee *et al*., 2018)에서도 지적되었던 문제이다. 어려움의 이유는 교실의 논증 과정에서 반박이 활발하게 일어나지 않기 때문에(Erduran, Simon, & Osborne, 2004) 훈련 데이터가 부족한 것이 그 원인이 될 수 있다. 또한 학생의 발화를 분석하는 과정에서 반박을 서술하는 방식(Ah, 2018)이 다른 논증 요소를 서술하는 방식과 큰 차이가 나타나지 않는다는 점과 반박 과정에서 사용되는 과학 용어가 자료, 정당화에서 사용되는 용어와 차이가 없다는 점은 형태소 단위로 자동 채점을 기계 학습의 모델을 구축하는 과정에 문제점이 될 수 있다.

이 연구에서는 논증 요소 중 반박의 채점 성능을 향상시키는 방법으로 전문가 시스템을 활용하였다. 자동 채점 과정에 전문가 시스템을 추가하기 위해 수행된 절차를 간략히 기술하면 다음과 같다. 먼저

Table 6. Set machine learning parameters

기계 학습 모델	R package	하이퍼 파라미터
서포트 벡터 머신	e1071	gamma = 0.5, cost = 8, kernal = linear
의사결정나무	rpart	cp = 0.01
랜덤 포레스트	randomForest	ntree = 500, mtry = 44
인공신경망	nnet	size = 3, decay = 5e-4

전문가들의 논증 채점 과정을 조사하여 전문가들이 사용한 논증 요소들의 분류 기준들을 분석한다. 이 기준들을 형태소 단위로 추출함으로써 전문가 형태소 목록을 구성하였다. 구성된 전문가 형태소 목록을 자동 채점 과정에 생성 규칙으로 적용함으로써 개선된 채점 결과를 얻고자 하였다. 전문가 형태소 생성 규칙 알고리즘은 Phase 2의 자동 채점을 통해 채점이 이루어진 논증 요소 중 비논증(None)을 제외한 4가지 논증 요소의 학생 발화를 대상으로 하였다. 비논증은 예측 정확도가 다른 논증 요소의 예측보다 높게 나타났기 때문에 생성 규칙 알고리즘으로 인한 오분류를 발생시키지 않기 위해 제외되었다. 정리하면, Phase 3에서는 Phase 2의 모델을 통해 채점이 이루어진 학생 발화를 대상으로 전문가 형태소 포함 여부를 if-then의 규칙으로 재분류를 수행하는 과정을 추가하였다. 마지막으로 이를 통해 얻어진 결과를 통하여 Phase 3 채점 모델을 평가하였다.

Ⅳ. 연구 결과 및 논의

기계 학습 모델의 자동 채점 결과는 R의 'caret' 패키지의 confusionmatrix() 함수를 사용하여 혼돈 메트릭스(confuse matrix), 정확도(accuracy), kappa값을 구하였다. 혼돈 매트릭스란 실제 값과 예측 값에 대한 참, 거짓의 2x2 행렬로서 모델의 성능을 평가하기 위한 지표로서 많이 사용된다. 또한 구해진 혼돈 매트릭스를 활용하여 각 논증 요소의 정밀도(precision), 재현율(recall) 그리고 정밀도와 재현율을 모두 고려한 f1-score를 구하였다. 정밀도란 모델의 참이라고 예측한 것 중에서 실제 참인 비율을 의미하며, 재현율이란 실제 참인 것 중에서 모델이 참이라고 예측한 비율의 값이다. 기계 학습 모델의 성능은 단지 정확도와 kappa만을 활용하는 것보다는 정밀도, 재현율, f1-score를 종합적인 관점으로 판단하는 것이 좋다.

1. Phase 1. 모든 학생 발화와 형태소를 대상으로 한 논증의 자동 채점

Phase 1의 논증 자동 채점 모델은 추가적인 전처리를 진행하지 않은 2,605개의 학생 발화와 1,968개의 형태소로 구성된 훈련 데이터를 통해 생성되었다. Phase 1의 채점 결과는 Table 7~Table 9과 같다. Table 7의 Human과 Prediction은 인간 채점자와 자동 채점 모델의 채점 결과이며 C, D, J, R, N은 주장, 자료, 정당화, 반박, 비논증 요소를 의미한다.

Table 8. Analyze results by machine learning model (Phase 1)

	SVM	DT	RF	ANN
Accuracy	61.15%	50.96%	65.96%	53.08%
kappa	0.476	0.3177	0.5298	0.3573

4가지의 기계 학습 기법 중 랜덤 포레스트(RF)가 65.96%의 가장 높은 정확도를 나타냈으며 의사결정나무(DT)는 4가지 기법 중 가장 낮은 50.96%의 정확도를 나타내었다. 앞서 말한바와 같이 랜덤 포레스트는 여러 개의 의사결정나무를 통해 하나의 예측 결과를 얻어내는 방법으로 높은 예측력을 지니는 특징을 지닌다는 점을 통해 4가지 기계 학습 기법 중 가장 높은 정확도를 나타내는 것을 설명할 수 있다.

이 연구에서 랜덤 포레스트의 결과는 서답형 응답에 대한 자동 채점에 대한 선행 연구(Song, Noh, & Sung, 2016; Ha *et al*., 2019)와 비교했을 경우 그 성능이 약간 떨어진다. 선행 연구들의 경우 비교적 짧은 문장으로 구성된 정답이 있는데 비해, 이 연구의 논증 발화 데이터는 정해진 정답이 없으며, 문장 형태가 정형화되어 있지 않아 잡음

Table 7. Confusion matrix of automated scoring(Phase 1)

		Human.C	Human.D	Human.J	Human.R	Human.N
SVM	Prediction.C	48	12	21	3	9
	Prediction.D	8	75	24	6	10
	Prediction.J	22	18	38	5	12
	Prediction.R	0	0	2	4	9
	Prediction.N	8	18	10	5	153
DT	Prediction.C	7	0	4	0	0
	Prediction.D	26	44	31	7	11
	Prediction.J	39	35	52	5	20
	Prediction.R	0	0	0	0	0
	Prediction.N	14	44	8	11	162
RF	Prediction.C	56	8	15	1	4
	Prediction.D	11	72	36	7	8
	Prediction.J	9	18	37	5	4
	Prediction.R	0	0	0	1	0
	Prediction.N	10	25	7	9	177
ANN	Prediction.C	47	17	38	3	20
	Prediction.D	8	69	23	8	14
	Prediction.J	5	12	6	6	6
	Prediction.R	1	3	5	2	1
	Prediction.N	25	22	23	4	152

Table 9. Analyze results by coded argumentation(Phase 1)

		precision(%)	recall(%)	f1-score
SVM	Prediction.C	51.61	55.81	0.5363
	Prediction.D	60.98	60.97	0.6098
	Prediction.J	40.00	40.00	0.4000
	Prediction.R	26.67	17.39	0.2105
	Prediction.N	78.87	79.27	0.7907
DT	Prediction.C	63.63	8.13	0.1443
	Prediction.D	36.97	35.77	0.3636
	Prediction.J	34.43	54.74	0.4228
	Prediction.R	0	0	0
	Prediction.N	67.79	83.94	0.7500
RF	Prediction.C	66.67	65.12	0.6588
	Prediction.D	53.73	58.54	0.5603
	Prediction.J	50.68	38.95	0.4405
	Prediction.R	100.00	4.35	0.0833
	Prediction.N	77.63	91.71	0.8409
ANN	Prediction.C	37.60	54.65	0.4455
	Prediction.D	56.56	56.09	0.5633
	Prediction.J	17.14	6.32	0.0923
	Prediction.R	16.67	8.70	0.1143
	Prediction.N	67.26	78.76	0.7255

이 많았던 것이 성능에 영향을 끼친 것으로 보인다.

그러나, 국내에서 이루어진 영작문을 대상으로 하는 자동 채점의 초기 연구(KICE, 2006)에서 인간 채점 결과와 자동 채점 결과 간 합치도(exact agreement)가 최저 49.02%, 평균 68.23%인 점을 참고하였을 때, Phase 1의 자동 채점 결과는 자연스러운 교실 상황 속의 논증 발화에 대한 채점 결과로서는 긍정적이라 할 수 있다. 또한 보편적 내용(domain general)의 과학적 말하기 논증을 다룬 선행연구의 정확도가 평균 51.51%인 점을 고려하면 특정한 주제(domain specific)를 대상으로 한 본 연구의 자동 채점 모델이 보다 효과적일 수 있음을 보여준다.

서포트 벡터 머신(SVM)과 인공신경망(ANN)은 61.15%와 53.08%의 정확도와 0.476, 0.3573의 kappa값을 나타내어 적당한(moderate) 일치도와 어느 정도(fair) 일치도(Landis, & Koch, 1977)를 얻었음에도 불구하고, 자동 채점 과정에서 어떠한 요소(형태소)가 코딩의 예측에 영향을 미쳤는지 확인할 수 없다는 단점이 있다. 즉, 서포트 벡터 머신과 인공신경망 기법은 유의미한 채점이 가능한 기법이지만 채점 모델에 대한 유의미한 추론과 결과의 해석이 어렵다는 것이다. 하지만 의사결정나무와 랜덤 포레스트의 경우 트리의 마디와 중요도 지수를 통해 어떠한 형태소가 자동 채점 과정에서 중요한 역할을 수행하는지 확인 할 수 있다. 이는 사람이 수행하는 논증 채점 과정에서 고려하는 형태소를 기계 학습에서 활용한 형태소와 비교함으로써 두 가지 방법에 어떠한 차이가 있는지 확인할 수 있는 방법이다. 기계 학습이 논증 요소를 자동 채점하는 과정에서 어떠한 형태소가 중요한 변수로 작용하는지 확인하기 위해 의사결정나무에서 나타나는 마디와 랜덤 포레스트의 MDG (mean decrease gini)를 활용하여 특성 중요도(feature importance) 상위 30개를 확인하였다. MDG란 모든 의사결정나무에서 특정 예측 변수의 지니 계수 감소를 평균한 값이다. 의사결정나무가 생성되는 과정에서 지니 계수의 감소가 클수록 분류가 잘 일어나게 되는 것이므로, 랜덤 포레스트의 MDG가 크다는 의미는 해당 변수가 분류를 하는데 더 중요하게 작용한다는 것이다. Figure 7는 의사결정나무의 트리를 나타낸 그림이며, Figure 8는 랜덤 포레스트의 MDG 특성 중요도 상위 30개를 나타낸 것이다.

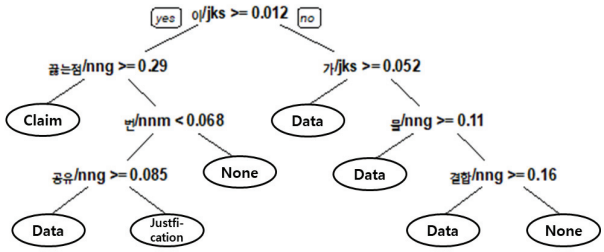

Figure 7. Decision tree model by phase 1

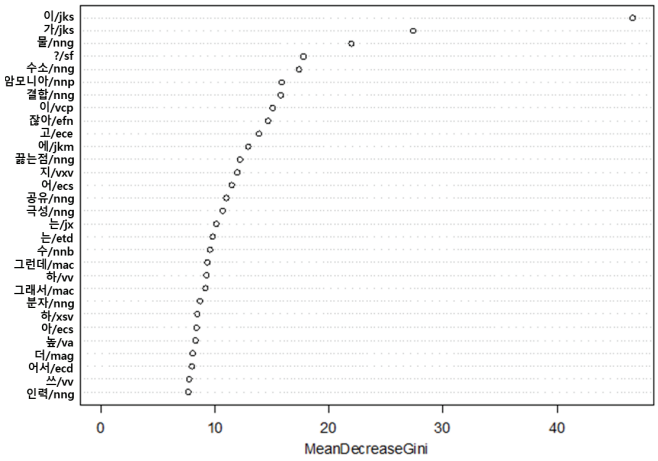

Figure 8. Mean decrease Gini of variables in random forest analysis by phase 1

Figure 7와 Figure 8은 논증 요소를 채점한 전문가들은 일반적으로 채점 과정에서 주요한 형태소로 고려하지 않는 '이/jks', '가/jks', '에/jks', '는/jx'과 같은 조사가 주요 형태소로 나타났음을 보여준다. 특히 '이/jks'의 경우는 의사결정나무에서 뿌리 노드(root node)이며 랜덤 포레스트에서는 가장 중요한 특성 중요도를 갖는 것으로 확인하였다. 조사가 기계 학습에서 주요한 역할을 수행하는 이유는 한국어 문장에서 조사의 활용이 많기 때문에 코퍼스의 많은 비율을 차지하는 주요 변수로서 작용하기 때문이라고 생각된다. 하지만 연구 방법에서 언급한 바와 같이 조사는 문법적 역할을 수행함으로 학생의 논증 과정에서 의미를 부여할 수 없다. 보다 개선된 자동 채점을 위해서는 학생의 발화 데이터의 전처리 과정에서 불용어 처리를 수행해야 한다.

2. Phase 2. 불용어 처리된 학생 발화와 형태소를 대상으로 한 논증의 자동 채점

앞선 연구 방법에서 말한바와 같이 문법적 역할을 수행하는 조사는 자동 채점 과정에서 불용어로써 처리되어야 한다. Phase 2의 훈련 데이터는 Kil(2018)이 제안한 불용어 목록을 참고하여 J(조사), X(접

두사, 접미사, 어근), U(분석 불능) 형태소가 제외된 1,758개의 형태소와 2,085개의 학생 발화로 구성되었다. Phase 2의 채점 모델을 활용한 자동 채점 결과는 다음과 같다(Table 10, Table 11).

Table 10. Analyze results by machine learning model (Phase 2)

	SVM	DT	RF	ANN
Accuracy	60.58%	49.42%	64.81%	52.88%
kappa	0.4669	0.2775	0.515	0.368

Table 11. Analyze results by coded argumentation(Phase 2)

		precision(%)	recall(%)	f1-score
SVM	Prediction.C	57.83	55.81	0.5680
	Prediction.D	57.26	57.72	0.5749
	Prediction.J	40.91	37.89	0.3934
	Prediction.R	27.27	26.09	0.2667
	Prediction.N	75.86	79.79	0.7778
DT	Prediction.C	36.67	25.58	0.3014
	Prediction.D	40.97	47.97	0.4419
	Prediction.J	35.71	10.53	0.1626
	Prediction.R	0	0	0
	Prediction.N	57.64	86.01	0.6902
RF	Prediction.C	59.38	66.28	0.6264
	Prediction.D	56.00	56.91	0.5645
	Prediction.J	50.70	37.89	0.4337
	Prediction.R	0	4.35	0.0833
	Prediction.N	76.21	89.64	0.8238
ANN	Prediction.C	51.92	62.79	0.5684
	Prediction.D	49.70	67.48	0.5724
	Prediction.J	37.84	14.74	0.2121
	Prediction.R	6.67	8.70	0.0755
	Prediction.N	67.03	63.21	0.6507

결과를 살펴보면 문법적 역할을 하는 형태소를 제거하는 과정이 자동 채점의 정확도를 크게 변화시키는 요인이 아니라는 것을 확인할 수 있다. 이는 영어 등 외국어에서 사용되는 불용어 처리 알고리즘을 한국어에 적용하는 처치가 정확도에 큰 영향을 끼치지 않음을 보여주기 때문에 자동 채점을 수행하는 과정에서 불용어 처리 알고리즘의 적용이 큰 무리가 없음을 시사한다.

코드 요소에 따른 f1-score를 세부적으로 비교하였을 때 8개의 항목에 증가, 2개의 항목에서 변화가 없었으며 10개의 항목에서 미미한 감소가 나타난다. 또한 Phase 1과 Phase 2의 f1-score의 총 변화량의 합은 증가하는 결과를 얻었다. 이를 통하여 전체적인 정확도가 소폭 감소하였음에도 논증 요소별 채점 결과는 큰 차이가 나타나지 않는 것을 알 수 있다. 불용어를 처리하기 전과 후의 정확도를 비교하였을 때 불용어 처리 후 정확도가 1.54%~0.20% 감소하는 결과가 나타났는데 이는 새로운 훈련 데이터를 활용하여 새로운 모델을 만들었기 때문이다. 또한 Phase 1의 의사결정나무에서 주요한 노드 역할을 하고 랜덤 포레스트에서 MDG값이 큰 형태소 '이/jks', '가/jks'가 훈련 데이터에서 제외되는 결과를 초래해 정확도의 감소가 나타난 것으로 보인다.

Phase 1의 평가 과정과 동일하게 불용어 처리가 된 학생의 발화에서 어떠한 형태소가 논증 채점에 중요 요소로써 작용하는지 확인하기 위해 의사결정나무와 랜덤 포레스트의 노드와 MDG를 활용하여 특성 중요도 상위 30개를 확인하였다.

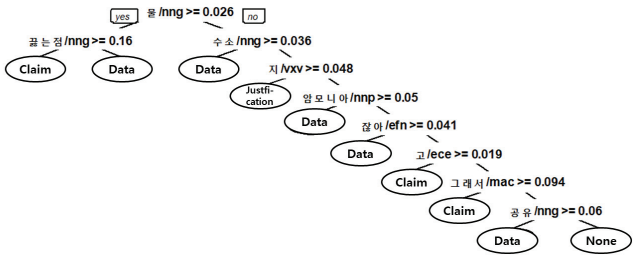

Figure 9. Decision tree model by phase 2

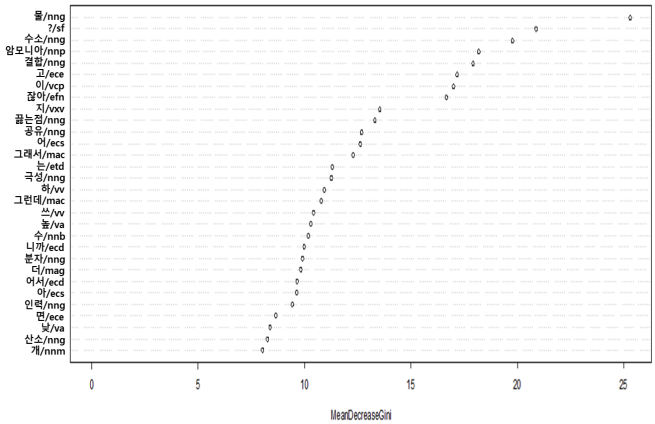

Figure 10. Mean decrease Gini of variables in random forest analysis by phase 2

불용어 처리가 이루어진 학생의 발화에 대한 의사결정나무 모델의 마디를 확인한 결과 Phase 1에서 마디로 도출되지 않았던 부사 '그래서/mac', 어미 '잖아/efn', '고/ece'와 용언 '지/vxv', 명사 '수소/nng', '암모니아/nnp'가 도출되었다. 또한 랜덤 포레스트의 특성 중요도로 어미 '니까/ecd', '면/ece'와 명사 '개/nnm' 등이 새롭게 도출되었으며, 부사 '그래서/mac', '그런데/mac'와 어미 '고/ece', '잖아/efn', 용언 '지/vxv' 등의 중요도 순위가 상승하였음을 확인하였다.

이 결과에서 주목할 점은 불용어 처리를 함으로써 과학적 주제에 대한 논증 과정에서 사용되는 과학 용어가 의사결정나무 트리의 노드로써 도출되는 점과 랜덤 포레스트의 중요도 지수 순위가 함께 상승한다는 것이다. 앞서 이야기한바와 같이 Phase 1의 경우 조사가 채점 모델의 주요한 변수로써 나타나는데 조사는 단순한 문법적 역할을 수행하기 때문에 논증 자동 채점 모델의 질을 평가하기에는 좋은 기준이 아니다. 따라서 과학적 논증을 평가하는 과정에서 유의미한 형태소로 분류할 수 있는 과학 용어의 중요도가 높아진 점은 긍정적인 결과이다. 또한 본 연구에서 사용된 학생의 논증 주제는 수소 결합과 분자 구조로 정해져있기 때문에 한정된 과학 용어가 논증 과정에서 주요한 역할을 수행하고 반복적으로 사용된다. 이는 과학 용어가 모

Table 12. Example of student's argumentation

형태소/품사	코드 요소	학생 발화
그래서/mac	주장	그래서 생물이 나타날 수가 없음. 그래서 물이 더 쎄니까 그게 끓는점이 쎄다고 했거든요.
잖아/efn	자료	아까 이 말했듯이 동물의 그 대부분이 수분이 있는데 몸 안에, 그런데 인체의 70%가 물이잖아. 얘도 극성이긴 극성이잖아, 암모니아.
지/vxv	정당화	물의 극성이 약해지기 때문에 … 비열이 약해지면 그럴 수 있겠다.
그런데/mac	반박	그런데 2개니까 극성이 더 심할 것 아냐? 그런데 용매 무극성이 되는 거잖아. 굽은형이 아닌 직선형이 되면.

델의 평가 과정에서 더욱 중요한 역할을 차지하게 되는 원인이다. 과학 용어 이외의 형태소가 논증 자동 채점 요소로 새로 등장하거나 중요도가 논증 발화 예시는 다음과 같다(Table 12).

예를 들어 Figure 9에서 자료 요소의 마디로 나타난 '잖아/efn'라는 형태소는 학생의 발화에서 자료를 표현할 때 주로 사용되는 형태소이다. 이 형태소는 불용어 처리 전에는 의사결정나무의 마디로 나타나지 않았지만 불용어 처리를 통해 의사결정나무에서 자료를 결정하는 마디로 작용함을 확인 할 수 있다. 또한 '그래서/mac'라는 형태소 또한 학생의 발화에서 주장을 표현할 때 서두에 자주 등장하는 형태소로써 불용어 처리를 통해 의사결정나무에서 주장을 결정하는 마디로 나타나는 것을 확인하였다.

이러한 점들은 단지 문법적 역할을 수행하는 조사가 논증 자동 채점에서 주요한 변인으로 고려된 Phase 1의 자동 채점 모델과는 다른 결과이다. Phase 2의 정확도가 Phase 1과 비슷하게 나타났을지라도, Phase 2에서 생성된 채점 모델에서는 과학 용어의 중요도가 높아졌을 뿐만 아니라 논증 과정에서 담화표지(discourse function)의 역할을 수행하는 어미, 접속부사, 용언(Shen, 2019)이 자동 채점 모델을 구축하는 과정에서 중요한 특성으로 작용하였음을 확인할 수 있다. 즉, 학생 발화에 대한 적절한 텍스트 전처리 과정 및 자동 채점을 위한 형태소의 특성 선택(feature selection)이 잘 이루어질수록 기계학습을 통한 논증 자동 채점은 인간의 채점 과정과 유사한 형태를 지니게 된다는 의미이다.

3. Phase 3. 전문가 형태소를 생성 규칙 알고리즘으로 적용한 한 논증의 자동 채점

전문가의 논증 분석 과정에서 채점 과정에 영향을 미치는 형태소를 알아보기 위하여 논증이 활발하게 일어난 연속된 발화 50개를 대상으로 논증 요소를 분류한 후 평가의 기준이 된 형태소를 기록하는 방식의 논증 채점을 수행하였다. 이 채점은 2017년부터 2019년까지 과학적 논증의 발화를 연구한 박사과정 1인과 석사과정 4인이 수행하였으며 이 5인은 모두 5년 이상의 교직 경험이 있다. 채점은 5인이 각자 독립적으로 수행한 후 논증 요소를 선택하는 과정에 선택의 기준이 된 형태소를 기록하게 하였다. 또한 기록한 형태소를 활용하여 의사결정나무의 트리와 같은 전문가 의사결정 트리를 제작하도록 하여 전문가의 채점 과정을 가시적으로 확보하였다. 5인 전문가의 분석 결과를 종합하여 각 논증 채점에 영향을 미치는 형태소를 분석하였다. 전문가 형태소를 종합하는 과정에서 모든 논증 요소에 공통적으로 사용되는 과학 용어(e.g., 수소, 산소, 결합)와 3인 이상 전문가가 공통적으로 선택하지 않은 형태소는 제외하였다(Table 13). 전문가의 형태소를 분석한 결과 Phase 2에서 언급한바와 같이 전문가가 논증을 채점하는 과정에는 과학 용어만큼이나 어미, 접속부사, 용언이 중요한 역할을 수행하는 것을 확인할 수 있었다.

Table 13. Expert-selected morphemes

코드 요소	논증 요소에 영향을 미치는 형태소
주장	~같아, 그래서
자료	~잖아,~진다,~진대
정당화	~겠지,~인데,~진대, 때문에, 왜냐하면(왜냐면)
반박	?,~다고?(~라고?),~아니야?, 근데(그런데)

Phase 2에서 얻어진 의사결정나무의 트리와 전문가 의사결정 트리를 비교한 결과 '그래서/mac' 형태소는 주장을 코딩하는 주요한 마디로 동일하게 도출되었다(Figure 11). 또한 '~진대'로 지적한 어미는 'Kkam' 형태소 분석기의 '지/vxv' 형태소이며, 이는 정당화로 코딩하는 주요한 노드로서 공통적으로 도출된 하나의 예시이다. 이를 통해

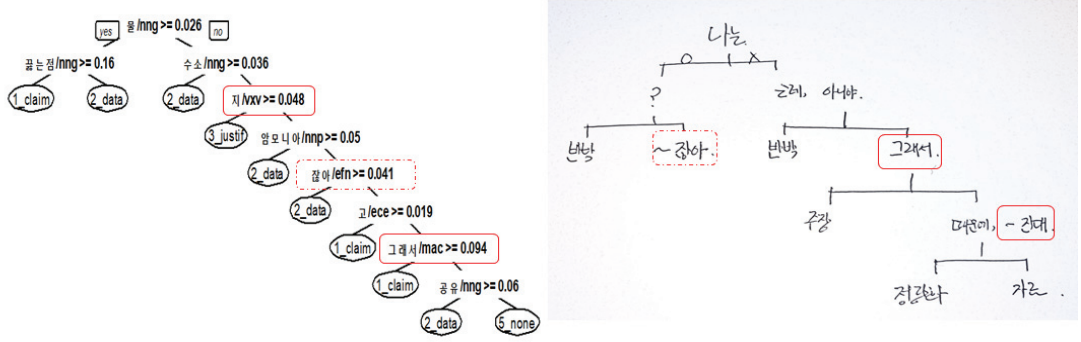

Figure 11. Comparison of Decision Tree and Expert Decision Tree

전문가의 사고 과정을 기계 학습으로 얻어진 모델에 적용한다면 보다 개선된 평가 모델을 구축할 수 있는 가능성을 확인하였다. 다만 전문가 형태소 분석 과정 중 2인의 전문가가 공통 요소로 추출한 일부 형태소의 경우 다른 논증 요소와 중복되어 나타나는 경우가 발생하였다(e.g., 자료, 정당화, 주장:~잖아; 자료, 정당화:~진대). 이러한 문제를 해결하기 위해서는 언어학에서 연구되고 있는 담화표지를 참고하여 자동 채점 과정에 반영하거나, 보다 많은 전문가가 참여하여 논증 채점 과정에서 고려되는 형태소의 분석을 통해 더욱 많은 데이터의 확보가 필요하다.

위 과정을 전문가 시스템 개발의 관점에서 정리하면 다음과 같다. 5인의 인간 전문가는 문제 해결과정으로 말하기 과학적 논증의 채점을 수행하였다. 인간 전문가의 채점 과정은 본 연구자의 분석을 통해 전문가 시스템의 규칙과 사실로 이루어진 기반 지식과 추론 엔진으로 표현되어 규칙 기반 전문가 시스템으로 개발되었다. 개발된 전문가 시스템은 Phase 2을 통해 얻어진 모델의 후속 과정으로써 논증의 자동 채점에 적용되었다. 또한 Phase 3는 기계 학습을 반복 수행하거나 Phase 2에서 얻어진 모델을 수정하는 과정을 거치지 않았다. 그러므로 Phase 2에서 얻어진 의사결정나무와 랜덤 포레스트의 노드와 특성 중요도는 변하지 않는다.

Phase 2의 채점 모델에 전문가 형태소 생성 규칙 알고리즘을 추가한 Phase 3의 자동 채점 결과는 아래와 같다(Table 14, Table 15).

Table 14. Analyze results by machine learning model(Phase 3)

	SVM	DT	RF	ANN
Accuracy	56.92%	48.27%	60.00%	47.08%
kappa	0.428	0.2762	0.4585	0.3106

Table 15. Analyze results by coded argumentation(Phase 3)

		precision(%)	recall(%)	f1-score
SVM	Prediction.C	53.93	55.81	0.5486
	Prediction.D	58.06	43.90	0.5000
	Prediction.J	37.33	29.47	0.3294
	Prediction.R	20.00	52.17	0.2892
	Prediction.N	75.86	79.79	0.7778
DT	Prediction.C	40.30	31.40	0.3529
	Prediction.D	44.05	30.08	0.3575
	Prediction.J	32.00	16.84	0.2207
	Prediction.R	16.13	21.74	0.1852
	Prediction.N	57.64	86.01	0.6902
RF	Prediction.C	55.21	61.63	0.5824
	Prediction.D	58.82	40.65	0.4808
	Prediction.J	41.56	33.68	0.3721
	Prediction.R	11.43	17.39	0.1379
	Prediction.N	76.21	89.64	0.8238
ANN	Prediction.C	44.95	56.98	0.5026
	Prediction.D	47.62	48.78	0.4819
	Prediction.J	27.87	17.89	0.2179
	Prediction.R	4.76	8.70	0.0615
	Prediction.N	67.03	63.21	0.6507

의사결정나무에서 반박에 대한 재현율과 f1-score의 경우 Phase 1과 2에서 모두 0을 나타냈으나 전문가 형태소 생성 규칙 알고리즘이 추가된 후 각각 21.74%와 0.1852로 증가하였다. 또한 서포트 벡터 머신과 랜덤 포레스트의 반박에 대한 f1-score는 각각 0.0225, 0.0546만큼 증가하였다. 이를 통해 전문가 형태소 생성 규칙 알고리즘은 Phase 1, Phase 2에서 나타난 반박의 자동 채점에 대한 문제를 해결할 수 있는 좋은 방안이 될 수 있음을 알 수 있다. 또한, 전문가 시스템을 채점 모델의 후속 단계에 휴리스틱으로 추가하는 것이 아니라, 기계 학습에서 모델이 만들어지는 과정에 전문가의 채점 과정을 적용할 수 있는 알고리즘이 개발된다면 보다 개선된 자동 채점 결과를 얻을 수 있을 것이다.

V. 결론 및 제언

이 연구에서는 기계 학습 기법을 활용한 논증의 자동 채점의 성능을 개선할 수 있는 방향을 모색하였다. 논증 자동 채점에 관한 연구는 자칫 기계 학습의 적용이라는 기술적 측면의 연구로만 이해되기 쉽다. 이에, 결론 및 제언에서는 본 연구의 과정과 결과를 간략히 정리하고, 자동 채점 연구가 과학 논증의 연구 및 교수 학습에 기여할 수 있는 시사점을 고찰하고자 한다.

실제 고등학교의 과학 논증 발화 데이터에 4가지의 기계 학습 기법을 활용한 자동 채점을 수행하였다. 과학 용어뿐만 아니라 문법적 역할을 수행하는 조사가 주요 형태소로 고려되는 문제점을 발견하였다. 이에 불용어를 활용하여 학생의 발화에 추가적인 전처리를 수행하여 과학 용어 이외에 유의미하게 고려되는 형태소(e.g., 그래서, 잖아, 지, 그런데)를 새롭게 도출하였다. 또한, 반박 요소에 대한 채점 정확도를 높이기 위해 전문가의 논증 코딩 과정을 형태소 단위로 탐색한 후, 얻어진 결과를 생성 규칙 알고리즘으로 적용시켜 반박에 대한 채점 정확도를 높였다.

논증 자동 채점 연구는 논증 기술과 사고방식을 체계적으로 밝혀 과학 논증에 대한 이해를 높일 수 있다. 예를 들어, 기술적으로 자동 채점의 정확도를 높이기 위해서는 서포트 벡터 머신 등의 기법이 주로 사용되어왔지만, 의사결정 나무와 같은 기법을 사용하면 논증 텍스트를 구성하기 위해 어떤 과정을 거치는가에 대한 분석이 가능하다. 본 연구에서는 전문가 시스템을 활용하여 자동채점의 성능을 높이는데 이 기법을 사용하였다. 이러한 기법을 전문가와 학생들의 논증 구성 과정을 비교, 분석하기 위해 사용한다면 과학 논증의 교수 학습 방법에도 적용할 수 있을 것이다. 자동 채점을 통해 논증이 자동으로 분석되어 학생 발화에 대한 채점이 이루어진다면 이전 연구에서 시도하였던 다이어그램을 통한 논증의 시각화(Shin, & Kim, 2012), 양적 요소를 통한 논증 수준의 파악(Erduran, Simon, & Osborne, 2004) 등의 분석 결과와 종합되어 학생의 논증에 대한 분석을 효과적으로 수행할 수 있을 것이다.

두 번째로, 논증 자동 채점 연구는 수업과 평가를 연계하며, 실시간 평가와 피드백 제공이 중심이 되는 과정 중심 평가의 연구와 적용에 활용될 수 있다. 예를 들어, 최근 과학 논증 글쓰기에서 주장과 근거의 연결 관계를 자동 채점하고 실시간 피드백을 제공한 연구에서 학생들의 과학적 논증에 대한 이해도가 유의미하게 높아지는 결과를 보였다(Lee *et al*., 2019). 말하기 논증의 경우, 본 연구와 같이 실제 교실

현장에서 지속적으로 구축된 학생의 발화 데이터로 구성된 채점 모델의 분석 결과는 논증 수업의 설계와 피드백 모델의 구성에 도움이 될 수 있다. 또한, 현재는 실시간 피드백을 제공하는 데는 어려움이 있지만 가까운 미래에 음성 인식 기술과 함께 자동 전사의 정확률이 높아지면 실시간 피드백을 제공할 수 있을 것이다.

최근 교육 연구에서도 빅데이터, 인공 지능, 기계 학습의 활용에 많은 관심을 기울이고 있다. 방법이나 기술에 대한 연구와 더불어 지속적인 데이터 베이스 구축 및 연구를 위한 데이터 공용화, 이에 따른 윤리적 문제에 대한 고민과 고찰이 필요하다. 또한, 컴퓨터 과학, 데이터 과학 등 다양한 분야와의 적극적인 협업을 통한 연구가 활발히 이루어져야 할 것이다. 기계 학습을 적용한 자동 채점 및 관련된 연구에 대한 지속적인 관심과 후속 연구를 통해 논증 수업을 효과적으로 교실 현장에 정착시킬 수 있으리라 기대한다.

국문요약

본 연구는 실제 교실에서 이루어진 학생의 과학 논증과정을 기계 학습을 활용한 자동 채점에 적용함으로써, 논증 자동 채점의 가능성 및 개선 방향을 탐색한다. 분자 구조에 대한 고등학생의 과학 논증 수업 중 발생한 2,605개의 모든 발화를 대상으로 연구를 진행하였다. 지도 학습을 위해 5가지의 논증 요소로 발화를 분류하였고, 분류된 발화를 대상으로 텍스트 전처리를 수행하였다. 전처리된 학생 발화를 활용하여 서포트 벡터 머신, 의사결정나무, 랜덤 포레스트, 인공신경망의 기계 학습 방법으로 자동 채점 모델을 구성하였다. 불용어 처리가 되지 않은 학생 발화를 활용한 자동 채점의 결과 랜덤 포레스트의 정확도는 65.96%, kappa는 0.5298의 유미한 결과를 얻었다. 불용어 처리를 수행한 학생 발화를 활용한 새로운 채점 모델의 결과 채점의 정확도가 크게 변화하지 않음에도 논증 발화 중 과학 용어 및 논증 요소의 담화표지가 채점 모델의 분류 기준이 되는 결과를 얻었다. 또한 인간 전문가의 논증 채점 과정을 분석하여 얻어진 전문가 형태소를 자동 채점 모델에 생성 규칙 알고리즘으로 적용하였다. 그 결과 의사결정나무에서 반박에 대한 재현율(recall)이 21.74% 증가하였다. 이에 본 연구 결과는 과학 교육 연구에서 기계 학습 및 논증에 대한 자동 채점의 활용 가능성과 연구 방향성을 제안하였다.

주제어 : 과학적 논증, 자동 채점, 기계 학습, 과학적 용어, 전문가 시스템

References

Ah, H. Y. (2018). A Study of Counter-Arguments as Observed in Debate of High School Students (Unpublished Master's Thesis). Pusan national university, Pusan.

Baek, Y. K. (1989). The Educational Potentials of Expert Systems. Journal of Educational Technology, 5(1), 79-91.

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.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. Applied Measurement in Education, 25(1), 27–40.

Buchanan, B. G., & Feigenbaum, E. A. (1980). The stanford heuristic programming project: Goals and activities. AI Magazine, 1(1), 25-30.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Cho, H. S., & Nam, J. (2014). The impact of the argument-based modeling strategy using scientific writing implemented in middle school science. Journal of the Korean Association for Science Education, 34(6), 583-592.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. Science education, 84(3), 287-312.

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. Review of research in education, 32(1), 268-291.

Engin, G., Aksoyer, B., Avdagic, M., Bozanlı, D., Hanay, U., Maden, D., & Ertek, G. (2014). Rule-based Expert Systems for Supporting University Students. Procedia Computer Science, 31, 22-31.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. Science education, 88(6), 915-933.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. The journal of machine learning research, 15(1), 3133-3181.

Giarratano, J. C., & Riley, G. (1998). Expert systems: Principles and programming. Boston: PWS.

Ha, M. (2016). 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.

Ha, M., Lee, G.-G., Shin, S., Lee, J.-K., Choi, S., Choo, J., Kim, N., Lee, H., Lee, J., Lee, J., Jo, Y., Kang, K., & Park, J. (2019). Assessment as a Learning-Support Tool and Utilization of Artificial Intelligence: WA3I Project Case. School Science Journal, 13(3), 271-282.

Jiménez-Aleixandre, M. P., Bugallo Rodríguez, A., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": Argument in high school genetics. Science Education, 84(6), 757-792.

Jun, C.-H. (2015). Data mining techniques. Seoul: Hannarea Publishing Co.

Kang, N.-H., & Lee, E. K. (2013). Argument and argumentation: A review of literature for clarification of translated words. Journal of The Korean Association For Science Education, 33(6), 1119-1138.

Kang, S. M., Kwak, K. H., & Nam, J. H. (2006). The effects of argumentation-based teaching and learning strategy on cognitive development, science concept understanding, science-related attitude, and argumentation in middle school science. Journal of the Korean Association for Science Education, 26(3), 450-461.

Kelly, G. J., Druker, S., & Chen, C. (1998). Students' reasoning about electricity: Combining performance assessments with argumentation analysis. International journal of science education, 20(7), 849-871.

Kevin P. Murphy. (2012). Machine Learning: A Probabilistic Perspective. Cambridge, MA: The MIT Press.

KICE(Korea Institute Of Curriculum & Evaluation). (2006). A Study on the Development and Introduction of an Automated Scoring Program(RRI 2006-6). Retrieved from http://kice.re.kr/resrchBoard/view.do?seq=19388&s=kice&m=030102

Kil, H.-h. (2018). The Study of Korean Stopwords list for Text mining. The Korean Language and Literature, 78, 1-25.

Kim, M., & Ryu, S. (2019). Development of Scientific Conceptual Understanding through Process-Centered Assessment that Visualizes the Process of Scientific Argumentation. Journal of the Korean Association for Science Education, 39(5), 651-668.

Kim, S. J. (2019). A Study on the Extraction and Validation of Automatic Argumentative Writing Scoring Feature Using Text Mining (Unpublished Master's Thesis). Korea national university of education, Chung-Buk.

KOFAC (Korea Foundation for the Advancement of Science & Creativity) (2019). Scientific Literacy for All Koreans, Korean Science Education Standards for the Next Generation. Seoul: KOFAC.

Kwon, J.-S., & Kim, H.-B. (2016). Exploring small group argumentation shown in designing an experiment: Focusing on students' epistemic goals and epistemic considerations for activities. Journal of the Korean Association for Science Education, 36(1), 45-61.

Lantz, B. (2013). Machine learning with R. Birmingham, UK: Packt Publishing Ltd.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159-174.

Lee, D., Yeon, J., Hwang, I., & Lee, S. (2010). KKMA: A tool for utilizing sejong corpus based on relational database. Journal of KIISE: Computing Practices and Letters, 16(11), 1046-1050.

Lee, G.-G., Ha, H., Hong, H.-G., & Kim, H.-B. (2018). Exploratory Research

on Automating the Analysis of Scientific Argumentation Using Machine Learning. Journal of The Korean Association For Science Education, 38(2), 219-234.
Lee H.-J., & Park Y.-M. (2019). A Study on the Search for Automatic Scoring Variables by Comparison of Text Studies Using Natural Language Processing. The research in writing, 41, 255-287.
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.
Lee, S., & Nam, J. (2016). Impact of Student Assessment Activities on Reflective Thinking in High School Argument-Based Inquiry. Journal of The Korean Association For Science Education, 36(2), 347-360.
Lee, S., & Nam, J. (2018). Impact of Student Assessment Activities on Claim and Evidence Formation in High School Argument-Based Inquiry. Journal of the Korean Chemical Society, 62(3), 203-213.
Linn, M. C., Gerard, L., Ryoo, K., McElhaney, K., Liu, O. L., & Rafferty, A. N. (2014). Computer-guided inquiry to improve science learning. Science, 344(6180), 155-156.
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.
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.
Martin, T., & Sherin, B. (2013). Learning analytics and computational techniques for detecting and evaluating patterns in learning: An introduction to the special issue. Journal of the Learning Sciences, 22(4), 511−520.
McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. The Journal of the Learning Sciences, 15(2), 153-191.
McNeill, K. L., & Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In M. Lovett, & P. Shah (Eds.), Thinking with data: The proceedings of 33rd Carnegie symposium on cognition. Mahwah, NJ: Erlbaum.
Negnevitsky, M. (2005). Artificial Intelligence: A Guide to Intelligent Systems. Harlow: Addison-Wesley.
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.
Ong, N., Litman, D., & Brusilovsky, A. (2014). Ontology-based argument mining and automatic essay scoring. In Proceedings of the First Workshop on Argumentation Mining (pp. 24-28). Baltimore, MD.
Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. Journal of research in science teaching, 41(10), 994-1020.
Park, C., Kim, Y, Kim, J, Song, J., & Choi, H. (2015). R data mining. Seoul: Kyowoo.
Park, E. L., & Cho, S. (2014). KoNLPy: Korean natural language processing in Python. In Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology (pp. 133-136). Chuncheon, Korea.
Park, J., & Kim, H.-B. (2018). Exploring Teachers' Responsive Teaching Practice in Argumentation-Based Science Classroom: Focus on Structural and Dialogical Aspects of Argument. Journal of The Korean Association For Science Education, 38(1), 69-85.
Ripley, B., & Venables, W. (2020). Package 'nnet'. R package version, 7.3-14. Retrieved April 28, 2020, from https://cran.r-project.org/web/packages/nnet
Russell, S., & Norvig, P. (2016). Artificial intelligence: A modern approach. Harlow: Pearson.
Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. Science education, 92(3), 447-472.
Shen, L. (2019). A Study on Functions of Korean Discourse Markers (Unpublished Doctoral Dissertation). Yonsei University, Seoul.
Shin, H. S. & Kim, H.-J. (2012). Development of the Analytic Framework for Dialogic Argumentation Using the TAP and a Diagram in the Context of Learning the Circular Motion. Journal of the Korean Association for Science Education, 32(5), 1007-1026.
Simon, S., Erduran, S., & Osborne, J. (2006). Learning to teach argumentation: Research and development in the science classroom. International journal of science education, 28(2-3), 235-260.
Song, J., Kang, S. J., Kwak, Y., Kim, D., Kim, S., Na, J., ... & Son, Y. A. (2019). Contents and Features of' Korean Science Education Standards (KSES)'for the Next Generation. Journal of The Korean Association For Science Education, 39(3), 465-478.
Song, M.-Y., Noh, E.-H., & Sung, K.-H. (2016). Analysis on the Accuracy of Automated Scoring for Korean Large-scale Assessments. The Journal of Curriculum and Evaluation, 19(1), 255-274.
Toulmin, S. (1958). The uses of argument. Cambridge, UK: Cambridge University Press.
Yang, I. H., Lee, H. J., Lee, H. N., & Cho, H. J. (2009). The development of rubrics to assess scientific argumentation. Journal of The Korean Association For Science Education, 29(2), 203-220.
Yoo, J. E. (2015). Random forests, an alternative data mining technique to decision tree. Journal of Educational Evaluation, 28(2), 427-448.
Yoo. J. E. (2019). Machine Learning for Large-scale/Panel Data and Learning Analytics Data Analysis. Journal of Educational Technology, 35(2), 313-338.
Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 39(1), 35-62.

저자정보

이만형(한국교원대학교 학생)
유선아(한국교원대학교 교수)