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

영어기반 컴퓨터자동채점모델과 기계번역을 활용한 서술형 한국어 응답 채점 -자연선택개념평가 사례-

하민수*
강원대학교

Scoring Korean Written Responses Using English-Based Automated Computer Scoring Models and Machine Translation: A Case of Natural Selection Concept Test

Minsu Ha*
Kangwon National University

ARTICLE INFO

Article history:
Received 21 March 2016
Received in revised form
4 April 2016
18 April 2016
29 April 2016
Accepted 5 May 2016
Keywords:
automated computer scoring, written response, natural selection, assessment

ABSTRACT

This study aims to test the efficacy of English-based automated computer scoring models and machine translation to score Korean college students' written responses on natural selection concept items. To this end, I collected 128 pre-service biology teachers' written responses on four-item instrument (total 512 written responses). The machine translation software (i.e., Google Translate) translated both original responses and spell-corrected responses. The presence/absence of five scientific ideas and three naïve ideas in both translated responses were judged by the automated computer scoring models (i.e., EvoGrader). The computer-scored results (4096 predictions) were compared with expert-scored results. The results illustrated that no significant differences in both average scores and statistical results using average scores was found between the computer-scored result and experts-scored result. The Pearson correlation coefficients of composite scores for each student between computer scoring and experts scoring were 0.848 for scientific ideas and 0.776 for naïve ideas. The inter-rater reliability indices (Cohen kappa) between computer scoring and experts scoring for linguistically simple concepts (e.g., variation, competition, and limited resources) were over 0.8. These findings reveal that the English-based automated computer scoring models and machine translation can be a promising method in scoring Korean college students' written responses on natural selection concept items.

Ⅰ. 서론

구성주의기반 개념변화 학습을 위해서는 학생들의 선개념을 확인하고, 수업 후 선개념이 과학적 개념으로 변화했는지 확인해야 한다(Magnusson *et al.*, 1997). 수업 후 학생들이 여전히 선개념을 가지고 있거나, 선개념과 과학적 개념을 혼합하여 가지고 있다면 교사는 수업의 문제점을 확인해야 될 것이며, 학생들에게는 추가적인 학습과제를 부여하여 보충할 수 있도록 해야 할 것이다. 구성주의적 개념변화 학습의 시작과 끝 모두 학습자의 개념을 이해하는 것, 즉 평가로 이루어진다. 과학개념평가는 학습자가 이해하고 있는 과학개념의 질적 또는 양적인 자료를 획득하는 일이다(Odom & Barrow, 1995). 지식, 이해, 태도, 감정 등 인간의 정신현상을 정확하게 확인하기 위해서는 효율적인 의사소통이 반드시 필요하다. 평가자는 피평가자들이 이해할 수 있는 수준의 문제를 제시하고, 학생들은 제시된 문항을 이해한 뒤 응답해야 된다. 평가자는 학생들의 응답을 확인하고 그 결과를 질적 또는 양적인 방법으로 나타낸 뒤 그 의미를 이해해야 한다. 이와 같은 과정은 시간을 요구하고, 시간의 제약으로 짧은 시간 동안에 평가는 이루어지기 힘들다. 최근 이런 문제점에 근거하여 컴퓨터를 활용한 평가가 이루어지고 있다(Nehm *et al.*, 2012).

컴퓨터가 평가결과를 수집하고 해석한 뒤 학생들에게 즉각적인 피드백을 줄 경우 개념학습의 경우에는 그 효과가 높다고 알려져 있다. Mathan & Koedinger(2005)는 개념이나 과학적 지식을 학습하는 활동에서는 즉각적인 피드백이 지연된 피드백에 비하여 더 높은 효과를 보인다고 하였다. Shute (2008)의 연구에서도 이 점은 강조되었는데, 인지구조에서 과학적 개념과 선개념이 경쟁하고 있다면 즉각적인 피드백을 할 경우 학습을 더 촉진할 수 있다고 하였다. 특히 최근 대중공개 온라인 강좌(MOOC, massively online open courses)와 같이 교육의 대형화에서 컴퓨터 채점의 장점은 더 강조된다. 평가가 인간의 의하여 이루어질 경우 발생하는 제한된 시간과 비용으로 대중공개 온라인강좌를 수강하는 수만 명의 학생들에게 효율적인 피드백을 할 수 없을 것이다.

구성주의 학습을 위해서는 학습자의 선개념을 보다 명확하게 이해해야 된다. 그와 같은 이유로 인하여 선택형 평가보다 서술형 평가가 더 주목받을 수밖에 없다. 선택형 평가는 제시된 답안들 중에서 가장

* 교신저자 : 하민수 (msha@kangwon.ac.kr)
http://dx.doi.org/10.14697/jkase.2016.36.3.0389

그럴듯한 답안을 선택하는 것으로 그 답안이 정확하게 피평가자의 인지구조에 있는 개념과 일치한다고 할 수 없다(Beggrow *et al.*, 2014). 반면에 서술형 평가의 경우에는 스스로 인지구조에서 단어들을 조합하여 설명을 구성해 내므로 서술형 평가의 응답은 학생들의 정신모형에 가장 근접하다고 할 수 있다(Opfer *et al.*, 2012). 또한 Liu *et al.*(2016)이 강조한 바와 같이 서술형 평가는 학생들의 생각들의 일관성을 측정할 수 있을 뿐만 아니라 생각을 통한 반성적인 경험을 강화하여 교수효과를 확장시킬 수 있다. 선택형 평가의 경우 교실응답시스템(classroom response system)이라 불리는 작은 기계와 소프트웨어로 빠르게 교실현장에서 피드백을 할 수 있다(Crossgrove & Curran 2008; Levesque 2011). 하지만 서술형 평가에 대한 컴퓨터 채점이 교실현장에서 실현가능하기 위해서는 컴퓨터 자동채점 알고리듬을 개발해야 하는 어려움이 있다.

이와 같은 필요성에 근거하여 현재 많은 연구진들이 서술형 평가의 컴퓨터 자동채점 모델을 개발하고 있다. Educational Testing Service (ETS)에서 개발한 개념평가문항을 위한 C-Rater (Leacock & Chodorow, 2003)는 초창기 개발된 과학개념에 관한 서술형 평가 자동채점 도구이다. C-Rater에 기계학습방법을 더하여 ETS 연구진들과 과학교육학자들이 모여서 개발한 C-Rater-ML은 8가지의 서술형 과학탐구 문항의 채점이 가능하다(Liu *et al.*, 2016). 뿐만 아니라 작은 연구그룹에서 개발하고 있는 서술형 평가의 자동채점 모델 역시 많이 개발되고 있다. 이 연구에서 사용하는 자연선택개념평가문항의 자동채점도구뿐만 아니라 산・염기 화학반응에 관한 평가문항(Haudek *et al.*, 2012), 광합성 개념(Weston *et al.*, 2015), 통계 개념(Kaplan *et al.*, 2014) 등 다양하게 개발되고 있다. 특히 전문가 채점자료를 활용하여 자동채점알고리듬을 생성해주는 새로운 기업의 출현(http://turnitin.com)으로 앞으로 다양한 채점 모델이 현장에서 사용될 것이다. 일부 기업체에서 개발된 채점알고리듬의 경우 사용이 제한적이지만 연구단체에 의하여 생성된 채점알고리듬은 교육적 목적에 의하여 무제한적으로 사용되어 질 수 있다. 하지만 사용 언어가 영어이기 때문에 한국어 적용은 어렵다. 이 문제를 해결할 수 있는 방법은 기계번역과 채점알고리듬을 함께 사용하는 것이다. 기계 번역을 통해서 학생들의 응답을 번역하고, 자동채점알고리듬으로 번역된 응답을 자동채점 할 수 있다면 영어기반으로 개발된 채점알고리듬을 한국어에 적용할 수 있을 것이다. 이 연구는 그와 같은 방법의 효용감을 조사하고자 자연선택개념평가문항에 대한 학생들의 응답을 기계 번역하고 그 이후 8가지 세부개념을 자동 채점한 결과와 인간 채점 결과를 비교하였다. 이 연구 결과를 바탕으로 기계 번역과 자동채점의 두 방법이 한국어 응답에 대안이 될 수 있는지 논의한다.

자연선택개념에 관한 컴퓨터 자동채점 모델

글로 이루어진 학생들의 응답을 빠르게 채점할 수 있는 컴퓨터 방법은 다양하다. 첫 번째 제안된 방법은 언어적 유사성을 근거로 학생들의 응답을 비슷한 유형으로 묶어 교사가 채점해야 되는 양을 줄여주는 방법이다. 짧은 단문의 응답을 요구하는 문항일 경우 학생들이 작성한 응답은 대부분 유사할 것이다. 컴퓨터가 언어적 유사성을 근거로 1000개의 응답을 30개의 그룹으로 묶어 준다면 교사는 30번의 채점으로도 전체 1000개의 응답을 채점할 수 있다. 이와 같은 방법으로 개발된 컴퓨터 채점도구가 Microsoft사에서 개발한 Powergrading이다(Basu *et al.*, 2013). 이와 같은 방법의 장점은 특정 문항에 국한하지 않고, 어떤 문항이라도 사용이 가능하다는 점이다. 하지만 응답의 길이가 길면 컴퓨터가 특징적인 그룹으로 응답을 묶을 수 없기 때문에 단문의 서술형 평가에만 유용하며, 교사가 채점을 하는데 요구되는 시간을 줄여줄 수는 있으나 여전히 교사의 채점을 요구하기 때문에 즉각적인 피드백을 위한 도구로는 활용될 수 없다.

두 번째 방법은 사전에 특정문항에 대한 응답을 채점할 수 있는 채점 알고리듬을 학습시켜 두고 학생들의 응답이 수집되면 학습된 채점알고리듬을 사용하여 즉각적으로 채점하는 것이다. 이와 같은 방법에도 학습시키는 방법에 따라 두 가지로 구분된다. 먼저, 채점에 관한 규칙을 컴퓨터에 입력하는 방법이 있다(Haudek *et al.*, 2012). 다른 방법은 전문가 채점 자료에서 나타나는 채점 규칙을 실제 전문가 채점 자료를 바탕으로 기계학습방법을 이용하여 자동으로 추출하는 방법이다(Nehm *et al.*, 2012). 기계학습은 인공지능(artificial intelligence)의 한 영역으로 인간의 의사결정에 관한 자료들을 분석하여 패턴을 찾아낸 뒤, 동일한 방법으로 의사결정을 수행하는 알고리듬을 개발하는 것이다. 기계학습은 다양한 분야(진단 의학, 사진 자동분류, 컴퓨터 바둑게임)에서 활용되는데 컴퓨터 자동채점 알고리듬을 생성하는 방법에도 활용되고 있다. 이 연구에서 사용하고자 하는 자연선택개념에 관한 학생들의 응답을 채점하는 알고리듬 역시 기계학습으로 생성된 알고리듬이다. 인간 채점이 일정하게 채점규칙에 근거하여 이루어졌다면, 인간 채점의 패턴은 수학적으로 표현될 수 있을 것이며 수학적인 방법을 통하여 전문가 채점이 예측될 수 있다. 물론 이와 같은 기계학습은 통계적인 방법을 활용하므로 많은 양의 자료가 필요하다. 이 연구에서 사용한 자연선택개념의 평가도구는 10270개의 인간 채점 자료와 기계학습을 이용하여 개발된 도구이다 (Moharreri *et al.*, 2014).

통계를 활용한 기계 번역

이 연구에서 우리는 한국어 응답을 영어로 번역하기 위하여 기계번역 프로그램 중 무료로 이용이 가능한 구글번역(Google Translate)을 사용하였다. 구글뿐만 아니라 무료로 사용이 가능한 웹번역도구는 많으나(예를 들어서 Bing 번역) 다양한 번역결과를 영어원어민이 살펴본 결과 구글번역이 가장 우수한 것으로 확인되었다. 구글번역은 통계적인 방법을 활용하여 학습시킨 번역 알고리듬을 사용한다. 전문가가 번역한 자료에서 번역의 패턴을 확인한 후 최적화된 단어를 예측하고 조합시켜 번역을 완성한다. 통계적인 방법을 활용하여 기계학습을 시키기 위해서는 자료의 양이 가장 중요하다(Makiko *et al.* 2011). 구글은 번역도구를 만들기 위해 굉장히 많은 번역 자료를 활용하고 있다. 구글번역이 한국어를 영어로 번역한 문장의 질적 수준을 확인할 수 있는 자료나 문헌은 없으나 경험적으로 구글번역의 결과를 신뢰할 수 있다. 예를 들어서 "진화론은 돌연변이와 자연선택으로 구성된다"의 번역결과는 "The theory of evolution composed of mutation and natural selection"로 완벽한 번역이 이루어진다. 물론 이 경우에 오자가 있어서는 안 될 것이다. 예를 들어서 "진화론은 *돌변연이*와 자연선택으로 구성된다" 라는 문장은 '돌연변이'의 오자를 가지고 있다. 이 경우 번역은 "The theory of evolution by natural selection and configuration turns kite"으로 전혀 엉뚱한 문장이 제시된다. 그래서 이 연구에서는 학생들의 응답의 오자를 교정한 것과

교정하지 않은 것을 통해 컴퓨터자동채점의 효용감을 비교할 계획이다.

II. 연구 방법

1. 검사도구 및 참여자

각 참여자는 4문항의 자연선택개념에 관한 개방형 질문에 응답하였다. 사용한 검사도구는 모두 동일한 형태로 구성되어 있으며, 소재는 달팽이, 장미, 펭귄, 느릅나무로 4종류이다. 문항의 형태는 "달팽이(동물)의 한 종은 독성이 있습니다. 이 독성이 있는 달팽이가 어떻게 독성이 없었던 조상 달팽이 종으로부터 진화하게 되었는지 생물학자들은 어떻게 설명할까요?" 구성되어 있다(Nehm *et al.*, 2012; Opfer *et al.*, 2012). 학생들은 자유롭게 자신의 응답을 설문지에 작성하였다. 128명의 생물예비교사(각 학년별 32명)이 작성한 512개의 응답이 이 연구에 사용되었다. 사용한 응답의 평균 길이는 띄어쓰기 기준으로 평균 21.44, 글자 수는 85.99이었다. 모든 참여자는 진화 지식에 관한 검사도구인 Conceptual Inventory of Natural Selection(CINS, Anderson *et al*., 2002)와 진화에 대한 태도 검사도구인 Measure of Acceptance of the Theory of Evolution(MATE, Rutledge & Warden, 1999)를 수행하였다. 이 검사들의 결과는 전문가 채점과 컴퓨터 채점과의 비교를 통하여 인간채점을 대신하여 컴퓨터 채점을 수행할 경우 다른 항목의 점수와의 상관관계가 전문가 채점에 비하여 어느 정도 차이가 나는지 확인하기 위함이다.

2. 인간 채점 및 컴퓨터 자동 채점

참여자들이 작성한 512개의 응답은 전문가들에 의해 채점되었다. 전체 응답은 문항과 채점 루브릭의 개발에 참여한 생물교육전문가가 하였으며, 20%의 문항 (약 100문항)은 채점자간 일치도를 확인하기 위하여 두 명의 생물교육전문가들과 함께 하였다. 채점자간 일치도(kappa)는 0.81 이상이 될 때까지 반복 채점하여 채점의 신뢰도를 확보하였다. 채점은 Nehm *et al*.(2010)이 개발한 채점루브릭을 근거로 이루어졌다. 채점은 5가지의 과학적 개념과 3가지의 비과학적 개념으로 이루어졌다. 과학적 개념은 변이, 변이의 유전성, 경쟁, 제한된 자원, 차별적 생식/생존이며, 비과학적 개념은 목적/필요, 사용/불사용, 적응이다(Nehm *et al.*, 2010).

컴퓨터 자동채점은 미국 학생, 전문가 등의 10270의 응답과 전문가 채점을 기계학습 방법으로 훈련시킨 채점모델을 사용하였다. 이 채점모델은 EvoGrader라는 웹기반자동채점도구에서 탑재되어 있는 모델이다. 8가지 채점모델은 각각 다르며, 다른 방법으로 훈련되었다. 예를 들어서 변이, 경쟁, 제한된 자원은 각각의 단어에서만 정보를 추출한 unigram으로만 구성된 반면 그 외의 알고리듬은 unigram 뿐만 아니라 연속된 두 단어에서 추출된 정보를 같이 포함한 bigram의 정보도 추가되어 있다(Moharreri *et al*., 2014). 그렇기 때문에 각각의 채점모델에 사용된 정보의 총량은 각 모델마다 다른데, 이 정보의 양이 많을수록 채점알고리듬이 복잡하다고 할 수 있다. 예를 들어서 변이개념을 채점하는 모델은 739개의 feature를 사용하였으며, 변이의 유전성은 2951개, 경쟁은 768개, 제한된 자원은 746개, 차별적 생식/생존은 2770개, 목적/필요는 2860개, 사용/불사용은 3003개, 적응은 2944개의 feature를 사용하였다.

컴퓨터 자동채점을 수행하기 전에 한국어로 된 학생들의 응답은 구글번역으로 번역되었다. 학생들의 응답에는 오자가 일부 포함되어 있는데, 번역 전에 오자를 수정하였다. 오자는 단어가 잘못된 것(띄어쓰기 및 오자)을 고친 것으로 문장을 수정한 것이 아니라 응답자의 의도에 맞게(예: 돌언변이를 돌연변이로 수정) 단어를 고친 것이다. 수정한 오자의 수는 411개로 512개의 응답에서 평균 0.80개의 오자가 포함되어 있었다. 오자를 수정한 응답과 수정하지 않은 응답 두 가지를 구글번역으로 번역하였으며, 그 결과를 자동채점알고리듬으로 분석하였다.

3. 분석 방법

컴퓨터 자동채점의 결과에 신뢰가 가는지 확인하는 방법은 매우 다양하다. 이 연구에서는 Ha, Nehm(2016a)가 제안한 방법인 거시적 관점과 미시적 관점 두 가지로 신뢰도를 확인하였다. 컴퓨터 채점의 신뢰도를 확인하는 거시적 방법은 컴퓨터 채점 결과와 인간 채점 결과를 통한 해석의 범위가 어느 정도 차이가 있는지 확인하는 것이다. 예를 들어서, 인간 채점과 컴퓨터 채점으로 생성한 점수의 평균값에 차이가 있는지, 또는 전문가 채점과 컴퓨터 채점을 통해 생성한 점수가 다른 변인들과의 상관관계를 확인할 때 그 통계치가 어느 정도 다른지 등을 통해서 확인할 수 있다. 거시적인 방법으로 정확도를 측정하는데 한계가 있는데, 그 이유는 평균값의 경우 컴퓨터가 감점과 가점의 두 실수를 연달아 할 경우 평균값은 변화하지 않기 때문이다. 그러므로 미시적인 수준에서 보다 엄밀하게 컴퓨터 채점의 정확도를 살펴야 한다. 전체 평균값보다 미시적인 수준에서 확인하는 방법은 두 가지로 구분될 수 있다. 먼저 각 학생별로 제공되는 학생별 점수이다. 평가는 여러 문항으로 구성되어 있을 수 있고, 학생별로 제공되는 점수는 여러 문항들의 합산 점수이다. 학생별 점수가 전문가 채점과 컴퓨터 채점으로 생성되었을 때 그 유사성은 학급단위로 생성된 전체 평균값에 비하여 상당히 미시적이다. 더 미시적인 수준은 각 개념별 점수 예측이다. 컴퓨터는 특정 문항의 응답에서 특정 개념이 존재하는지를 예측한다. 이 연구에서는 각 문항마다 8개의 개념이 존재하는지 예측한 자료를 바탕으로 점수를 생성한다. 각각의 예측이 인간 채점과 일치하는지 살펴보는 것은 가장 미시적인 접근이다. 예를 들어서 이번 연구에서와 같이 128명이 4문항씩 하여 생성된 512개의 응답을 8개의 개념의 유무를 판단하였으므로 총 4096개의 컴퓨터 예측이 있다. 이 전체 예측에서 과연 몇 개가 정확했는지를 따지는 것이 마지막 방법이다.

거시적인 분석을 위해서 사용한 방법은 Chi제곱 검정이다. 문항이나 개념에 따라 인간이 채점하여 생성한 평균 점수와 컴퓨터가 생성한 평균 점수가 일관성이 있다면 (Chi 제곱 검정의 p값이 0.05 이상), 두 점수를 사용할 경우 결과 해석에는 문제가 없을 것이다 (Table 1 참조). 두 번째는 각 방법으로 생성된 점수의 대응표본 t-검정 결과를 확인하였다. 세 번째는 외부 점수와의 상관관계를 통하여 비교할 수 있다. 참여자들은 모두 객관식 자연선택검사도구(Conceptual Inventory of Natural Selection, Anderson *et al.*, 2002)와 진화개념에 대한 수용정도를 수치화하는 검사도구(Measure of Acceptance of the

Table 1. Average score of each item by three different scoring methods

Pseudonym		Variation	Heritability	Competition	Limited resources	Differential survival	Need/Goal	Use/Disuse	Adapt/Acclimation
Experts scoring	Snail	50.0	7.0	3.1	49.2	53.1	25.8	0.8	9.4
	Rose	35.2	4.7	2.3	31.3	35.9	46.9	3.1	9.4
	Penguin	19.5	3.1	0.8	35.2	40.6	48.4	19.5	15.6
	Elm	39.8	7.8	6.3	10.2	64.1	31.3	0.8	6.3
	Total	36.1	5.7	3.1	31.5	48.4	38.1	6.1	10.2
Computer scoring of spelling corrected responses	Snail	49.2	1.6	3.1	44.5	25.8	22.7	0.8	6.3
	Rose	34.4	0.8	1.6	24.2	14.1	33.6	1.6	7.0
	Penguin	17.2	0.0	0.8	31.3	19.5	39.8	10.9	10.9
	Elm	39.8	0.8	4.7	5.5	39.8	12.5	1.6	7.0
	Total	35.2	0.8	2.5	26.4	24.8	27.2	3.7	7.8
Computer scoring of original responses	Snail	50.8	1.6	3.1	43.0	29.7	21.1	0.8	7.0
	Rose	34.4	0.8	1.6	24.2	12.5	33.6	1.6	7.0
	Penguin	18.0	0.0	0.8	30.5	20.3	39.1	10.9	12.5
	Elm	39.8	0.8	4.7	5.5	40.6	11.7	2.3	7.0
	Total	35.7	0.8	2.5	25.8	25.8	26.4	3.9	8.4

Theory of Evolution, Rutledge & Warden, 1999) 수행하였다. 또한 인간 채점과 컴퓨터 자동채점의 결과를 두고 학년별 변화 과정에 관한 ANOVA를 수행하고 각 채점방법별로 통계치(F, 유의도, 효과크기)가 얼마나 다른지를 통하여 전문가 채점과 컴퓨터 채점 결과의 유사성을 확인하였다. 학생별로 생성되는 점수의 유사성을 확인하기 위하여 상관관계분석을 하였다. 사회과학에서 두 변인간 상관관계가 0.5 이상이라고 하면 크다고 하고(Cohen, 1992), 자연과학에서 어떤 두 생물의 정보들의 상관관계가 0.9이상이면 두 생물은 동일한 생물로 이해할 수 있다고 알려져 있다(Sato *et al*., 표 2005; Zhu *et al*., 표 2002). 가장 미시적인 수준에서 컴퓨터 채점의 정확도를 확인하는 방법은 채점자간 일치도 분석인 Cohen의 kappa이다. Landis, Koch(1977)가 제한한 기준에 의하면 0.21-0.40는 그럭저럭(fair), 0.41-0.60는 적당한(moderate), 0.61-0.80는 상당한(substantial), 0.81-1.00은 거의 완벽한(almost perfect)으로 구분된다. 통계분석은 SPSS 22.0 버전을 사용하였다.

Ⅲ. 연구 결과 및 논의

자동채점의 사용자가 문항에 대한 컴퓨터 채점의 평균점수를 사용하고자 할 경우는 컴퓨터가 제시한 평균점수가 전문가 채점의 평균점수와 일관성이 있는지 확인해야 된다. Table 1에는 전문가 채점, 오자를 수정한 컴퓨터 채점, 오자를 수정하지 않은 컴퓨터 채점 세 가지의 각 문항별, 개념별 평균 점수를 보여준다. 예를 들어, 전문가 채점은 변이개념에 관하여 달팽이 문항은 50.0점, 장미 문항은 35.2점, 펭귄 문항은 19.5점, 느릅나무 문항은 39.8점으로 전체 평균은 36.1점이다. 오자를 수정한 컴퓨터 채점은 달팽이 문항이 49.2점, 장미 문항이 34.4점, 펭귄 문항이 17.2점, 느릅나무 문항이 39.8점으로 전체 평균은 35.2점이다. 문항별 점수의 분포가 일치하는지 확인하기 위하여 Chi 제곱 검정을 수행하였다. 인간 채점과 오자를 수정한 컴퓨터 채점의 점수의 Chi 제곱 검정의 p값은 8개 개념별로 각각 0.991, 0.839, 0.988, 0.804, 0.630, 0.254, 0.795, 0.899이었다. 가장 낮은 값이 필요/목적 개념의 0.254이고 그 이외에는 모두 0.5이상의 높은 p값을 보였다. 인간 채점과 오자를 수정하지 않은 컴퓨터 채점의 Chi제곱 검정의 p값은 8개 개념별로 각각 0.995, 0.839, 0.988, 0.843, 0.467, 0.241, 0.601, 0.943이었다.

두 번째 Chi 제곱 검정은 과학적 개념 5가지의 전체 평균점수와 비과학적 개념 3가지의 평균점수들을 토대로 각 채점방법별로 확인하였다. 변이, 유전성 등 과학적 개념의 인간 채점의 평균은 36.1, 5.7, 3.1, 31.5, 48.4이다. 오자를 수정한 컴퓨터 채점방법의 점수와 Chi 제곱 검정을 하면 그 p값이 과학적 개념은 0.181, 비과학적 개념은 0.960이었다. 오자를 수정하지 않은 컴퓨터 채점방법과 비교하면 그 p값이 과학적 개념이 0.204, 비과학적 개념이 0.934이었다. 이와 같은 방법을 통해 확인한 결과 평균값을 사용할 경우에는 여러 채점방법을 사용해도 전문가 채점과 통계적으로 유의미한 일관성을 보이는 값을 생성함을 알 수 있다. 하지만 앞서 논의한 바와 같이 이 결과는 컴퓨터 채점과 전문가 채점의 일치도를 확인하는 것이 아니라 문항과 개념에 따라 생성된 점수의 분포에 일관성이 잇는지 확인하는 것이다. 예를 들어서 Heritability의 전문가 채점은 평균이 5.7%인데, 컴퓨터 채점은 0.8%로 그 차이는 상당하며, 다른 결과들을 통해서 확인할 수 있듯이 Heritability의 컴퓨터 채점 모델은 효용감이 매우 낮다. 그럼에도 불구하고 문항과 개념에 따라 채점결과의 일관성을 확인하는 중요한 이유는 채점결과의 일관성이 있을 경우 다양한 문항을 사용했을 때에도 기존의 일관성을 근거로 컴퓨터 점수를 교정할 수 있기 때문이다.

Table 2에는 전문가 채점과 컴퓨터 채점 간의 대응표본 t-검정 결과를 제시하였다. 연구 방법에서 설명한 바와 같이, 채점 방법간 점수를 비교하는 것은 일치도를 확인하는 것은 아니다. 하지만 컴퓨터가 확인할 수 있는 점수가 인간 채점에 비하여 어느 정도인지를 확인하고, 컴퓨터 채점 결과로부터 인간채점의 점수를 교정하는데 중요한 정보가 될 수 있을 것이다. 대응표본 t-점정 결과 컴퓨터에 비하여 전문가 채점에서 유의미한 수준에서 더 높은 평균점수를 보이고 있다. 이 점은 컴퓨터가 확인하지 못하는 개념들이 더 있기 때문이다. 이와 같은 오류의 비율이 일관성이 있다면 점수의 차이는 일관성을 근거로 수정할 수 있을 것이다. 다시 말하면, 다양한 문항이나 대상을 통해서

Table 2. Paired sample t-test between experts'scores and the scores of two different scoring methods

		Mean	SD	t	p	d
Normative Scientific idea	Experts scoring	4.99	2.90	10.34	0.000	0.53
	Computer scoring of spelling corrected responses	3.59	2.46			
	Experts scoring	4.99	2.90	9.60	0.000	0.51
	Computer scoring of original responses	3.63	2.44			
Non-normative naive idea	Experts scoring	2.17	1.67	6.74	0.000	0.42
	Computer scoring of spelling corrected responses	1.55	1.33			
	Experts scoring	2.17	1.67	6.64	0.000	0.41
	Computer scoring of original responses	1.55	1.36			

수집한 자료들에서도 전문가 채점과 컴퓨터 채점에서 나타나는 값의 차이가 일관성이 있다면, 두 값의 높은 상관관계를 바탕으로 회귀식을 만들 수 있을 것이다. 회귀식을 통해서 컴퓨터채점을 통해 생성한 값을 대입하여 인간채점을 유추할 수 있을 것이다. 이 연구에서 사용한 자료의 수로는 이와 같은 일관성을 확인하기에는 부족하므로 이 연구는 후속 연구로 확인 되어야 할 것이다.

컴퓨터 채점의 정확도를 확인하기 위한 두 번째 거시적인 방법은 Table 3와 4에 제시 되어 있다. Table 3에서는 만약에 어떤 연구자가 서술형 평가 문항(이 연구에서는 ACORNS)의 점수와 선택형 평가 점수 등과 같은 다른 구인들의 점수와의 상관관계에 대해서 알아보고 싶을 경우, 전문가 채점점수와 다른 변인들 간의 상관관계와 구글번역과 기계채점의 점수와 다른 변인들 간의 상관관계에서 차이가 있는지 확인하는 것이다. 큰 차이가 없을 경우에는 기계로 번역을 하고 채점을 해도 그 결과 해석에는 큰 문제가 없을 수 있다. Table 3를 보면 세 가지 다른 방법 채점을 할 경우 과학적 개념과 CINS(선택형 진화개념평가)와의 점수와 상관관계는 각각 0.422, 0.424, 0.426이다. 전문가 채점과 컴퓨터 채점, 전문가 채점과 오자 수정 컴퓨터 채점의 R^2의 차이는 1%가 되지 않는다(-0.002, -0.003). 또한 MATE (진화 수용정도의 평가)와는 0.336, 0.251, 0.255로 같은 방법으로 실시한 R^2의 차이는 약 5% 정도(0.050, 0.048)로 의미 있는 수치는 아닌 것으로 판단된다. 서술형 평가에서의 비과학적 개념의 점수에 대한 세 가지 다른 방식으로 채점한 점수와 CINS 점수와의 상관관계도 -0.415, -0.426, -0.436로 R^2의 차이는 1%~2% 정도 규모이다(-0.009, -0.018). 비과학적 개념의 점수와 MATE점수의 차이 역시 같은 방법으로 수행한 R^2의 차이는 약 1% 미만이다(0.010, 0.000)

Table 3. Pearson correlations between CINS/MATE scores and experts/computer scores(‡ p<0.01)

	Method	CINS	MATE
Normative Scientific idea	Experts scoring	0.422‡	0.336‡
	Computer scoring of spelling corrected responses	0.424‡	0.251‡
	Computer scoring of original responses	0.426‡	0.255‡
Non-normative naive idea	Experts scoring	-0.415‡	-0.257‡
	Computer scoring of spelling corrected responses	-0.426‡	-0.237‡
	Computer scoring of original responses	-0.436‡	-0.257‡

Table 4에서는 과학적 개념과 비과학적 개념의 점수의 학년간 차이를 세 가지 방법으로 확인했을 경우 통계치에서 차이점이 나타나는지 확인하였다. 효과 크기 중심으로 확인하면 과학적 개념의 경우 세 가지 다른 방법으로 수집한 점수의 학년간 차이가 0.057, 0.052, 0.046으로 의미 있는 차이가 나타나지 않았다. 비과학적 개념의 경우에는 세 가지 방법으로 생성한 점수들의 학년간 차이의 효과크기는 0.026, 0.011, 0.011으로 약간의 차이가 확인되었다. 이 결과는 전문가 채점을 사용하지 않고 컴퓨터 채점을 사용할 경우에도 학년간 점수 향상을 비교할 경우 통계적인 해석에서의 차이는 거의 미미하다는 것을 보여주고 있다.

Table 4. ANOVA results of academic years using three different scoring methods

	Method	F	Sig.	Partial Eta Squared
Normative Scientific idea	Experts scoring	2.483	0.064	0.057
	Computer scoring of spelling corrected responses	2.264	0.084	0.052
	Computer scoring of original responses	1.986	0.120	0.046
Non-normative naive idea	Experts scoring	1.088	0.357	0.026
	Computer scoring of spelling corrected responses	0.449	0.718	0.011
	Computer scoring of original responses	0.449	0.719	0.011

앞서 Table 1, 2, 3, 4에서 사용한 방법은 연구방법에서도 논의한 바와 같이 컴퓨터가 Overestimate과 Underestimate의 두 실수가 동시에 발생하면 평균값의 차이는 없어지므로 비록 그 안에는 많은 실수가 있었는지를 확인할 수 없다. 그래서 필요한 분석이 보다 미시적인 수준에서 컴퓨터가 생성한 점수의 정확도를 보는 것이다. Table 5에서는 전문가 채점과 컴퓨터 채점이 생성한 점수의 상관관계(Pearson correlations)를 보여준다. Pearson correlations에 대한 기준은 최근 Educational Testing Service의 연구진들에 의하여 사용되어진 Cohen (1968)의 기준에 따라서 확인하였다. 이 기준에 의하며 Pearson correlations의 계수가 0-0.09이면 상관이 없고, 0.10-0.30이면 작은 상관,0.31-0.50이면 중간 크기, 0.51-1.00이면 크다고 하였다. 전문가 채점과 구글번역의 컴퓨터 채점의 점수간 상관관계는 과학적 개념이 약 0.85 수준, 비과학적 개념이 0.77 수준으로 큰 상관관계의 범주에서도 중간 이상의 수준을 보이고 있다. Liu *et al.* (2016)는 미국의 Educational Testing Service사가 개발한 c-rater ML를 사용하여 8개

Table 5. Pearson correlations between experts'scores and the scores of two different scoring methods(‡ p<0.01)

	Method	(1)	(2)	(3)	(4)	(5)	(6)
Normative Scientific idea	Experts scoring	1.000					
	Computer scoring of spelling corrected responses	0.848‡	1.000				
	Computer scoring of original responses	0.832‡	0.985‡	1.000			
Non-normative naive idea	Experts scoring	-0.582‡	-0.465‡	-0.471‡	1.000		
	Computer scoring of spelling corrected responses	-0.488‡	-0.422‡	-0.433‡	0.776‡	1.000	
	Computer scoring of original responses	-0.487‡	-0.419‡	-0.433‡	0.770‡	0.983‡	1.000

문항의 점수를 컴퓨터 채점을 한 후 인간 점수와의 상관관계를 보고하였다. 이 연구에서 보고한 상관관계 r은 0.66, 0.71, 0.79, 0.79, 0.82, 0.83, 0.87, 0.91으로 본 연구의 0.85와 0.77을 비교했을 때 그 수준은 거의 대등하다고 할 수 있다. 또한 이 수준은 자연과학에서 두 생물종이 일치한다고 주장할 수 있는 수준인 0.9에는 도달하지 못하지만 과학적 개념의 채점 결과의 경우에는 상당히 근접하고 있다(Sato *et al*., 2005; Zhu *et al*., 2002).

가장 미시적인 방법은 각 예측별로 Overestimate, Correct estimate, Underestimate의 비율을 확인하는 것과 채점자간 일치도를 확인하는 Cohen's kappa이다. Table 6에는 각 개념별로 나타난 Overestimate, Correct estimate, Underestimate이 제시되어 있다. 특정 개념이 학생들의 설명에 실제로는 포함되어 있지 않음에도 불구하고 포함되어 있다고 예측하면 하면 Overestimate이고 그 반대는 Underestimate이다. 이 두 비율이 비슷할 경우 전체 평균에는 영향을 미치지 않는다. 두 가지 경우의 비율을 살펴보면 과학적 개념과 비과학적 개념 모두에서 Underestimate의 비율이 더 높은 것을 알 수 있다. 예를 들어서 차별적 생식/생존(Differential survival/reproduction)의 경우 Overestimate는 3.7%인데 비해 Underestimate의 경우에는 27.3%이다. 이와 같은 경우는 기계학습을 사용하는 경우는 흔하게 발생하는데, 그 이유는 기계학습 방식이 특정개념의 존재유무를 이전의 자료에서 나타나는 정보를 활용하기 때문이다. 채점알고리듬이 가지고 있는 정보 이외의 단어나 표현을 사용하여 개념을 표현한 경우는 컴퓨터가 확인할 수 없다. 또한 이 결과가 Overestimate가 일부 개념에서 더 많이 나타나는 Ha(2013)의 결과와 차이가 나타나는 원인은 한국어를 구글번역이 번역하는 과정에서 원래 채점 모델에서 사용되어진 미국의 학생들의 단어들과 상당한 차이가 나타나기 때문인 것으로 이해될 있다.

Table 7에 제시된 자료는 채점자간 일치도에 관한 통계치인 Cohen의 kappa값과 일치도의 비율이다. Kappa값을 확인할 때 가장 널리

Table 6. Overestimate, correct estimate, and underestimate of computer scoring

	Ideas	Estimate	Correcting spelling	Not correcting spelling
Normative Scientific idea	Variation	Overestimate	3.1	3.3
		Correct estimate	92.8	93.0
		Underestimate	4.1	3.7
	Heritability	Overestimate	0.0	0.0
		Correct estimate	95.1	95.1
		Underestimate	4.9	4.9
	Competition	Overestimate	0.0	0.0
		Correct estimate	99.4	99.4
		Underestimate	0.6	0.6
	Limited resources	Overestimate	1.6	1.4
		Correct estimate	91.8	91.6
		Underestimate	6.6	7.0
	Differential survival/reproduction	Overestimate	3.7	4.3
		Correct estimate	68.9	68.8
		Underestimate	27.3	27.0
Non-normative naive idea	Need/Goal	Overestimate	2.9	2.3
		Correct estimate	83.2	83.6
		Underestimate	13.9	14.1
	Use/Disuse	Overestimate	1.6	1.8
		Correct estimate	94.5	94.3
		Underestimate	3.9	3.9
	Adapt/Acclimation	Overestimate	2.1	2.5
		Correct estimate	93.4	93.2
		Underestimate	4.5	4.3

Table 7. Kappa and agreement percentage between experts scoring and computer scoring

	Ideas	Correcting misspelled words	Kappa	agreement
Normative Scientific idea	Variation	O	0.843	0.928
		X	0.847	0.930
	Heritability	O	0.232	0.951
		X	0.232	0.951
	Competition	O	0.894	0.994
		X	0.894	0.994
	Limited resources	O	0.801	0.918
		X	0.795	0.916
	Differential survival/reproduction	O	0.369	0.689
		X	0.365	0.688
Non-normative naive idea	Need/Goal	O	0.623	0.832
		X	0.630	0.836
	Use/Disuse	O	0.413	0.945
		X	0.403	0.943
	Adapt/Acclimation	O	0.595	0.934
		X	0.594	0.932

활용되는 Landis, Koch(1977)의 기준을 보면 0.21-0.40는 그럭저럭(fair), 0.41-0.60는 적당한(moderate), 0.61-0.80는 상당한(substantial), 0.81-1.00은 거의 완벽한(almost perfect)으로 구분된다. 이 기준으로 생각하면 변이개념, 경쟁개념, 제한된 자원 개념은 '거의 완벽한' 수준이며, 목적/필요 오개념, 적응 오개념은 '상당한' 수준, 그 외에 다른 개념들은 그 이하의 수준으로 구분된다. 이 채점모델을 생성한 Ha(2013)의 연구를 살펴보면 변이, 경쟁, 제한된 자원의 세 개념은 Unigram을 사용한 채점 모델이며, 그 외에 모델들은 Unigram과 Bigram을 모두 사용한 채점모델이다. Unigram은 텍스트를 활용한 기계학습에서 자료의 단위가 단일 단어인 것이고 Bigram은 자료의 단위가 연속된 두 단어이다. 예를 들어서 'dogs like cats' 과 'cats like dogs'라는 두 문장을 구분할 때 두 문장의 Unigram(dog, like, cat)은 같으나 Bigram(dog_like, like_cat)은 다르다. Bigram이 필요한 경우는 의미의 구분이 보다 면밀하게 되어야 하는 경우에 추가되어야 한다. 다시 말하면 의미가 복잡한 경우에는 구글번역과 자동채점이 잘 작동되지 않는 것에 비해, 단순한 경우에는 '거의 완벽한'수준으로 채점이 될 수 있음을 보여준다.

두 번째로 앞서 연구 방법에서 설명한 바와 같이, 학생들의 응답에는 오자가 일부 포함되어 있는데, 번역 전에 오자를 수정하였다. 오자는 단어가 잘못된 것(띄어쓰기 및 오자)을 고친 것으로 문장을 수정한 것이 아니라 응답자의 의도에 맞게(예: 돌언변이를 돌연변이로 수정) 단어를 고친 것이다. 수정한 오자의 수는 411개로 512개의 응답에서 평균 0.80개의 오자가 포함되어 있었다. 한 응답에 평균 단어수는 21.44개이므로 100단어에서 약 4개 단어가 오자로 판단된다. Ha, Nehm(2016a)에서도 컴퓨터 채점에서 오자의 영향을 확인하였는데, 오자의 양은 미국학생들의 응답에서도 상당했으나 컴퓨터 채점에는 의미 있는 영향이 없는 것으로 확인되었다. 이 연구에서도 오자의 수정이 전체 일치도에 의미 있는 차이를 보여주지는 않았다. 그럼에도 불구하고 오자는 분명히 컴퓨터 채점의 장애요인이기 때문에 이와 같은 문제점을 해결할 수 있어야 한다. 가장 중요한 것은 학생들이 올바른 단어를 사용하여 응답 할 수 있도록 해야 될 것이다. 학생들의 응답을 수집할 때 사용하는 온라인설문 도구에서 한글오자는 알려주지 않는다. 대안으로는 한글프로그램을 사용하면 오자의 경우 표시를 해 주므로 한글프로그램을 이용해서 답안을 적는 방법이 대안이 될 수 있을 것이다.

이상의 결과를 토대로 전체 논의한다. 먼저 교육평가에서 평가의 정확도는 매우 중요하며, 만약 학습 결과에 대한 평가가 정확하지 않다면 학습자의 학습에 오히려 부정적인 영향을 줄 수 있을 것이다(Ha, Nehm, 2016a). 예를 들어서 과학적 개념을 가지고 있음에도 불구하고 컴퓨터가 잘 못 채점할 경우 학습자는 해당 개념이 비과학적인 것으로 오인할 수 있다. 그러므로 일치도가 100%가 되지 않는 이상 분명히 False Positive와 False Negative의 예측이 있고, 그것은 학습자의 학습에 방해가 될 수 있다. 특히 이 연구에서와 같이 실질적으로 사용되어질 수 있는 채점모델은 8개 중에 3개 밖에 되지 않는다. 그러므로 이와 같은 수준에서 개인별 피드백 보다는 학급단위의 평가로 활용해야 될 것이다. 이 연구에서 사용한 EvoGrader에서도 컴퓨터 채점 결과를 학습자의 개인별 평가로 활용하기 보다는 학급단위의 진단평가와 형성평가의 도구로 활용할 것은 권장하고 있다. 즉, 교사가 수업의 전과 후에 학생들의 개념의 변화가 어느 정도 향상되었는지 학급 전체에서 나타나는 과학적 개념의 수준을 가늠하기 위하여 사용할 수 있다. 또한 학습자의 개인별 피드백의 경우에는 문항별 점수보다는 등급과 같은 방법으로 제시하면 일부 부정확한 요소들의 영향을 줄일 수 있을 것이다.

문항의 각 개념별로 학습자의 개인별 피드백이 가능하기 위해서는 kappa가 0.9이상은 확보되어야 할 것이다. 이와 같은 방법은 최근에 제시었는데, Ha, Nehm(2016b)는 컴퓨터의 각 개념별 예측이 정확한지, 정확하지 않는지 확인할 수 있는 방법을 제안하였다. 이 방법들은 컴퓨터의 예측에서 사용되는 확률, 여러 컴퓨터 채점모델을 개발하여 각 모델별로 일치하는지 여부의 두 가지 방법이다. 이 방법들을 사용하면 컴퓨터의 예측에서 낮은 정확도의 컴퓨터 예측은 교사가 직접 확인할 수 있을 것이다. 예를 들어서, 미국학생들의 응답 채점에서 Adapt/Acclimation개념을 위한 채점모델의 kappa값은 0.683이었으

나, 9.5%의 낮은 정확도를 가지는 문항만 교사가 수정한다면 kappa가 0.940까지 향상될 수 있다. 이와 같은 방법으로 교사의 업무를 약 90% 이상 줄일 수 있다. 컴퓨터에 전적으로 의존하는 방법은 교사의 효용감도 낮출 것이며, 자칫 학습을 방해할 수 있음은 명백하다. 이 연구에서 기계번역과 컴퓨터 채점의 두 방법을 활용했을 때에 비록 일부 개념에서는 일정 수준 이상의 정확도가 나왔으나 여전히 많은 개념에서 오류가 나타나는 것을 확인할 수 있다. 정확도를 신장시키는 방안에 대한 연구와 함께, 정확도가 가장 중요한 교육평가에서 컴퓨터 채점을 어떻게 효율적으로 활용할 수 있는지에 대한 논의도 활발히 진행되어야 할 것이다.

Ⅳ. 결론 및 제언

서술형 평가는 인지구조 속에 형상화 되어 있는 학생들의 개념이나 생각 등을 보다 정확하게 확인할 수 있는 평가방법으로 개념학습을 위한 진단평가와 형성평가로 활용될 필요가 있다. 또한 학생들도 자신의 언어로 생각을 풀어서 설명하거나 또는 논증하는 연습을 수행하고 피드백을 받아야 한다. 효율적인 개념학습 뿐만 아니라 설명과 논증 능력을 향상시키기 위해서 서술형 평가가 많이 활용되어야 하나 채점을 위한 시간과 비용의 현실적인 제약에 의하여 널리 활용되고 있지 않다. 이와 같은 문제점을 해결하고자 컴퓨터 기술을 활용한 자동채점이라는 대안적인 방법이 도입되고 있다. 영어권을 중심으로 교사의 채점과 거의 대등한 수준의 자동채점알고리듬이 많이 개발되고 있다. 이 연구는 해외에서 개발되어 공개된 자동채점모델을 한국어로 적용시키기 위해서 기계 번역(구글번역)과 연동하는 방법을 제안하고 번역과 채점의 두 가지 방법이 기계로 이루어졌을 때 전문가 채점에 비하여 어느 정도의 차이를 보이는지 확인하였다.

결과를 요약하면 첫째, 컴퓨터 자동채점 방법이 생성한 점수는 전문가 채점에 비하여 평균값은 유의미하게 낮았다. 하지만 문항에 따라 자동채점방법으로 생성한 점수의 평균값과 전문가 채점으로 생성한 점수의 평균값의 차이는 일관되었다. 자동채점 평균점수를 통하여 다른 변인간의 상관관계를 확인하거나 학년간 차이를 확인할 때 나타나는 통계치는 전문가 채점으로 생성한 점수를 활용한 결과와 유의미하게 큰 차이는 없었다.

둘째, 자동채점을 통해서 생성되는 학생별 점수와 전문가 채점을 통해 생성한 학생별 점수의 상관관계는 과학적 개념이 약 0.85, 비과학적 개념이 약 0.77로 상당히 높은 수준이었다. 마지막으로 채점자간 일치도의 가장 엄격한 방법을 통하여 확인한 결과 언어적으로 복잡하지 않은 개념들 (변이, 경쟁, 제한된 자원)의 경우에는 거의 완벽한 수준의 일치도를 보였으며, 언어적으로 매우 복잡한 개념의 경우에는 매우 낮은 일치도를 보였다. 즉, 번역과 자동채점의 두 단계를 거치더라도 언어적으로 복잡하지 않은 수준의 개념들은 채점을 위해서 활용 가능하다는 것을 보여준다.

이 연구 결과를 토대로 몇 가지 후속 연구를 제언한다. 이 연구는 이미 개발된 영어기반 자동채점 알고리듬을 사용하고자 한국어 응답을 분석하였다. 그 과정에서 우리나라 학생들이 흔히 사용하는 언어와 그것을 구글이 번역한 언어들이 미국의 대학생들의 언어와 상당한 차이가 있을 수 있다. 미국의 대학생들의 응답을 채점한 자료를 통하여 이 연구에서 사용한 자동채점알고리듬을 생성하였기 때문에 이 과정에서 일치도의 차이가 나타났을 가능성이 있다. 한국어 응답의 구글번역과 자동채점알고리듬을 개발할 때 사용한 미국학생들의 응답의 언어적 유사성을 확인한다면 근본 원인을 이해할 수 있을 것이다.

한국어로 된 기계학습알고리듬을 활용하여 우리나라 학생들의 응답을 통해서도 채점모델을 생성할 수 있을 것이다. 하지만 컴퓨터과학에 관한 비전문가들도 한국어를 통하여 기계학습을 할 수 있는 사용자 편의성이 최적화된 프로그램이 없기 때문에 한국어응답을 채점한 결과로 기계학습을 수행하여 자동채점 모델을 개발하는 것은 한계가 있다. 대신에, 한국어 응답과 채점결과가 상당히 있다면 기계로 번역하고 그 번역결과를 통하여 알파벳 기반의 기계학습알고리듬을 활용하여 채점모델을 생성할 수 있을 것이다. 같은 단어에 대한 기계번역은 당연히 같을 수밖에 없으므로 이 과정을 통해서 생성되는 자동채점모델은 다른 한국어 응답에도 분명히 작동할 것이라 확신한다. 물론 가장 이상적인 환경은 한국어로 된 문자를 통하여 자동채점할 수 있는 알고리듬을 생성하는 기계학습 소프트웨어가 나오는 것이다. 컴퓨터 공학의 발전 속도를 보면 이와 같은 소프트웨어가 개발되는 것도 빠른 시일 내에 가능할 수 있을 것이다. 그렇기 때문에 서술형 평가의 중요성을 인식하는 만큼 이 분야에 대한 관심 역시 가져야 할 것이다.

두 번째는 현재 개발되어 있는 다른 알고리듬을 한국어 응답에 적용해 보는 과정을 수행하였으면 한다. 산・염기 화학반응에 관한 평가문항(Haudek *et al*., 2012), 광합성 개념(Weston *et al*., 2015), 통계 개념(Kaplan *et al.*, 2014) 등은 개발되어 현재 사용되고 있으며, 그 외에도 여러 연구진들에 의하여 다양한 영역에서 자동채점 알고리듬들이 개발되고 있다. 기계 번역을 통한 한국어 적용이 가능하다면 우리나라 과학교육을 위해서 사용되어질 필요가 있을 것이다.

인공지능에 대한 국내 관심이 높아지는 시점에 교육과 관련된 인공지능에 대해서도 관심을 가진다면 자동채점연구에도 많은 발전이 있을 것이다. 자동채점에 관한 기술이 지속적으로 발전하면 디지털 교과서는 학생들의 응답에 즉각적인 피드백을 줄 수 있다. 또한 미래에 상호작용이 가능한 컴퓨터 보조교사 개발도 순조로워진다. 그렇게 될 경우 미래에는 단순한 개념학습에서 컴퓨터의 역할이 높아지고 교사들은 보다 복잡하고 융합적인 영역에 관련된 능력의 교육에 집중할 수 있을 것이다.

국문요약

이 연구는 기계 번역을 활용하여 영어기반서술형 평가의 자동채점 모델을 한국어 응답에 적용하는 방법의 효용감을 조사하기 위하여 이루어졌다. 이 연구를 위하여 예비생물교사 128명이 4문항으로 구성된 자연선택개념평가도구에 응답한 512개의 서술형응답을 활용하였다. 서술형응답은 한글맞춤법을 교정한 것과 교정하지 않은 학생들이 작성한 그대로의 응답 두 가지를 구글번역으로 번역하였다. 8가지 과학적 개념과 비과학적 개념을 채점하는 자동채점모델을 통해 생성한 4096개의 예측자료의 정확도를 독립적으로 수행한 전문가 채점자료와 비교하는 방법으로 확인하였다. 그 결과 컴퓨터로 채점한 점수와 전문가 채점점수의 평균값의 문항별 분포는 유의미한 차이가 없었다. 평균값을 활용하여 생성한 통계치들은 전문가 채점자료를 통하여 생성한 자료들과 유의미한 차이가 없었다. 학생별 점수의

Pearson 상관관계 계수를 확인한 결과 과학적 개념 점수는 0.848, 비과학적 개념 점수는 0.776이었다. 언어적으로 단순한 개념의 경우 채점자간 일치도 (kappa)가 0.8이상이었다. 이 결과는 기계 번역과 영어기반 서술형 평가의 자동채점모델이 우리나라 학생들의 자연선택개념문항을 채점하는데 유용한 방법이 될 수 있음을 보여준다.

주제어 : 컴퓨터자동채점, 서술형응답, 자연선택, 평가

References

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. Journal of Research in Science Teaching, 39(10), 952-978.

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: A clustering approach to amplify human effort for short answer grading. Transactions of the Association for Computational Linguistics, 1, 391-402.

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.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213–220.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159.

Crossgrove, K., & Curran, K. L. (2008). Using clickers in nonmajors-and majors-level biology courses: student opinion, learning, and long-term retention of course material. CBE-Life Sciences Education, 7(1), 146-154.

Ha, M. (2013). Assessing scientific practices using machine learning methods: Development of automated computer scoring models for written evolutionary explanations. Unpublished Doctoral Dissertation. Columbus: The Ohio State University.

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, 358-374.

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

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.

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).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 159-174.

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

Levesque, A. A. (2011). Using clickers to facilitate development of problem-solving skills. CBE-Life Sciences Education, 10(4), 406-417.

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.

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.

Makiko, M., Yuta, T., & Kazuhide, Y. (2011). Phrase-based statistical machine translation via Chinese characters with small parallel corpora. IJIIP: International Journal of Intelligent Information Processing, 2(3), 52-61.

Mathan, S. A., & Koedinger, K. R. (2005). Fostering the intelligent novice: Learning from errors with metacognitive tutoring. Educational Psychologist, 40(4), 257-265.

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), 1-14.

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.

Nehm, R. H., Ha, M., Rector, M., Opfer, J. E., Perrin, L., Ridgway, J. *et al*. (2010). Scoring guide for the open response instrument (ORI) and evolutionary gain and loss test (ACORNS). Technical Report of National Science Foundation REESE Project 0909999.

Odom, A. L., & Barrow, L. H. (1995). Development and application of a two-tier diagnostic test measuring college biology students' understanding of diffusion and osmosis after a course of instruction. Journal of Research in Science Teaching, 32(1), 45-61.

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.

Rutledge, M. L., & Warden, M. A. (1999). The development and validation of the measure of acceptance of the theory of evolution instrument. School Science and Mathematics, 99(1), 13-18.

Sato, T., Yamanishi, Y., Kanehisa, M., & Toh, H. (2005). The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics, 21(17), 3482-3489.

Shute, V. J. (2008). Focus on formative feedback. Review of educational research, 78(1), 153-189.

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(2), ar19.

Zhu, Z., Pilpel, Y., & Church, G. M. (2002). Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. Journal of Molecular Biology, 318(1), 71-81.