

한국어 서답형 문항 자동채점 시스템의 성능 개선

천민아*^o, 김창현[†], 김재훈*, 노은희[‡], 성경희[‡], 송미영[‡], 박종임[‡], 김유향[‡]

*한국해양대학교, 컴퓨터공학과

[†]한국전자통신연구원

[‡]한국교육과정평가원

minah0218@kmou.ac.kr, chkim@etri.re.kr, jhoon@kmou.ac.kr,
(noro, kelly9147, mysong, pji0310, kyh15}@kice.re.kr

Performance Comparison of Automated Scoring System for Korean Short-Answer Questions

Min-Ah Cheon*^o, Chang-Hyun Kim[†], Jae-Hoon Kim*,
Eun-Hee Noh[‡], Kyung-Hee Sung[‡], Mi-Young Song[‡], Jong-Im Park[‡], Yuhyang Kim[‡]

*Department of Computer Engineering, Korea Maritime and Ocean University

[†]Electronics and Telecommunications Research Institute

[‡]Korea Institute for Curriculum and Evaluation

요 약

최근 교육과정에서 학생들의 능력 평가는 단순 암기보다 학생들의 종합적인 사고력을 판단할 수 있는 서답형 문항을 늘리는 방향으로 변하고 있다. 그러나 서답형 문항의 경우 채점하는 데 시간과 비용이 많이 들고, 채점자의 주관에 따라 채점 결과의 일관성과 신뢰성을 보장하기 어렵다는 문제가 있다. 이런 점을 해결하기 위해 해외의 사례를 참고하여 국내에서도 서답형 문항에 자동채점 시스템을 적용하는 연구를 진행하고 있다. 본 논문에서는 2014년도에 개발된 ‘한국어 문장 수준 서답형 문항 자동채점 시스템’의 성능 분석을 바탕으로 언어 처리 기능과 자동채점 성능을 개선한 2015년도 자동채점 시스템을 간략하게 소개하고, 각 자동채점 시스템의 성능을 비교 분석한다. 성능 분석 대상으로는 2014년도 국가수준 학업성취도 평가의 서답형 문항을 사용했다. 실험 결과, 개선한 시스템의 평균 완전 일치도와 평균 정확률이 기존의 시스템보다 각각 9.4%p, 8.9%p 증가했다. 자동채점 시스템의 목적은 가능한 채점 시간을 단축하면서 채점 기준의 일관성과 신뢰성을 확보하는 데 있으므로, 보완한 2015년 자동채점 시스템의 성능이 향상되었다고 판단할 수 있다.

주제어: 자동채점, 기계학습, 준지도 학습, 자동채점 시스템의 성능 비교

1. 서론

21세기 지식정보화 사회에서는 획일화된 지식과 사고 방식보다 자신이 알고 있는 것을 바탕으로 새로운 것을 창조할 수 있는 사고방식을 중요시한다. 이에 정부는 2009 개정 교육과정을 도입하여 ‘창의인성교육’을 강조하고, 그것에 맞게 학생들의 능력 평가하고자 노력하고 있다[1-4]. 특히, 서답형 문항과 수행평가의 비중을 늘리도록 권고하고 있다[1,5]. 서답형 문항은 학생이 직접 정답을 구성하는 문항 형태로 선택형 문항보다 문제 해결능력이나 비판적 사고력, 창의력 등을 측정하는 데 적합하다[6-7]. 그러나 서답형 문항의 채점은 선택형 문항보다 채점에 걸리는 시간과 비용이 많고, 채점자의 주관적 판단에 의존하므로 채점자마다 채점 기준이 다를 수 있어 채점의 공정성과 채점 결과의 신뢰성을 확보하는데 어렵다는 단점이 있다[7].

해외에서는 서답형 문항을 효율적으로 채점하기 위해 자동채점 시스템[8-10]이 활용되고 있다. 국내에서도 한국교육과정평가원에서 한국어 서답형 문항을 자동채점하기 위한 프로토타입(prototype)을 개발하여 실제 시험에

적용하는 방안을 연구하고 있다[11-15]. 본 논문에서는 2014년에 개발한 ‘한국어 문장 수준 서답형 문항 자동채점 시스템[11-15]’의 결과를 바탕으로 성능을 개선한 시스템의 결과를 비교 분석한 결과를 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 서답형 문항 자동채점 시스템 개발을 위해 필요한 관련 연구들과 기존의 한국어 서답형 자동채점 시스템을 소개한다. 3장에서는 언어 처리 기능과 자동 채점 기능을 보완한 한국어 자동채점 시스템에 대해 설명한다. 4장에서는 기존의 시스템과 개선한 시스템의 성능을 비교 분석한다. 마지막으로 5장에서는 결론 및 향후 연구에 대해서 논의한다.

2. 관련 연구

이 장에서는 한국어 서답형 자동채점 시스템에 이용하기 위한 이론들과 2014 한국어 서답형 자동채점 시스템에 대해 간략하게 소개한다.

2.1 준지도 학습

기계학습(machine learning)은 컴퓨터가 새로운 정보를 학습하고, 습득한 정보를 효율적으로 사용할 수 있도록 하는 알고리즘과 기술을 말한다[16]. 기계학습의 학습 방법은 크게 지도 학습(supervised learning)과 비지도 학습(unsupervised) 방법으로 나눌 수 있다[16].

지도 학습 방법은 정답이 부착된 데이터를 학습 자료로 사용하여 분류기(classifier)를 학습한 후, 새로운 데이터에서 정답을 찾는다. 지도 학습 방법은 학습 자료를 구축하기 위해서 전문가의 도움이 필요하고 이 과정에서 많은 시간과 노력이 소요된다. 비지도 학습 방법은 정답이 부착되지 않은 자료를 사용하여 그 자료의 구조나 관계를 패턴으로 정의하여 자료를 분류한다. 위의 두 가지 학습 방법을 결합한 기계학습 방법을 준지도 학습(semi-supervised learning)이라고 한다. 준지도 학습은 정답이 부착된 자료와 정답이 부착되지 않은 자료를 모두 학습 자료로 사용한다. 학습이 진행될 때마다 정답이 부착되지 않은 자료에 대해 정답이 부착되고 이를 학습 자료로 사용함으로써 학습 자료의 양이 점점 늘어나서 분류기의 정확률 및 신뢰성을 지속적으로 향상시키는 방법이다.

2.2 2014 한국어 문장 수준 서답형 문항 자동채점 시스템

2014년도에 개발된 한국어 문장 수준 서답형 문항 자동채점 시스템[11,13-15]은 채점자가 채점한 답안과 채점하지 않은 답안을 복합적으로 사용하여 분류하는 준지도 학습 방법을 사용하고 있다. 시스템은 크게 언어 처리 단계와 채점 단계로 나뉜다. 언어 처리 단계에서는 형태소 분석과 같은 한국어 정보처리 시스템을 이용하여 학생들의 답안을 분석한다. 이 단계는 문서 정규화, 형태소분석 및 품사부착, 구문분석, 바뀌쓰기, 의존구조분석으로 구성된다. 언어 처리 단계가 끝나면 채점을 진행한다. 채점 단계는 Logistic Regression과 k-Nearest Neighborhood 알고리즘[16]을 이용하여 구현하였으며, 크게 다섯 단계로 구성되어 있다. 1) 초기 채점 단계로서 언어 처리 결과가 완전히 일치하는 답안들을 하나의 유형으로 간주하여 해당 유형에 속한 답안의 빈도순으로 정렬한다. 그 뒤, 인간채점자가 고빈도 학생 답안을 수동으로 채점한다. 2) 현재까지 채점된 결과와 모범답안을 학습말뭉치로 간주하여 자질을 추출하고, 선택한 자질의 가중치를 학습한다. 3)미채점 답안의 군집화로 학습된 결과를 바탕으로 미채점 답안을 군집화하여 분류한다. 4) 검증에 위한 과정으로서 분류된 결과 중에서 신뢰성이 높은 채점 답안에 대해서 인간 채점자가 확인하고 학습말뭉치에 추가한다. 5) 2)~4)의 과정을 미채점 답안이 존재하지 않을 때까지 반복한다. 2013년 학업성취도 평가의 서답형 문항을 사용하였다. 문항별 1,000개의 학생 답안을 추출하여 채점시간과 정확률을 평가한 결과, 채점 정확률은 평균 약 97.95%로 나타났다.

3. 2015 한국어 문장 수준 서답형 문항 자동채점 시스템

이 장에서는 2015 한국어 문장 수준 서답형 문항의 자동채점 시스템에서 개선한 내용을 중심으로 서술한다.

3.1 전체 시스템 구조

그림 1은 2015 한국어 문장 수준 서답형 문항 자동채점 시스템의 구조도다. 2014 자동채점 시스템과 같이 2015 자동채점 시스템도 ‘언어 처리 단계’와 ‘채점 단계’로 나뉜다.

‘언어 처리 단계’는 학생 답안이나 모범 답안(문제 출제자가 작성한 예시 답안)에 대해 자동채점에 필요한 언어 정보를 분석하는 단계다.

‘채점 단계’는 학생 답안과 모범 답안의 언어 처리 결과를 군집화하여 각 군집에 점수를 부여하는 단계다. 초기 학습용 답안은 언어 처리 결과가 완전히 일치하는 문장들을 하나의 유형으로 간주한다. 각 유형에 속한 문장들의 수를 내림차순으로 정렬하여 채점자가 상위 유형에 점수를 부여한다. 이렇게 채점된 답안들과 모범 답안들을 학습 자료로 간주하여 자질을 추출하고 자동채점 분류기를 학습한다. 학습된 시스템은 미채점 학생 답안을 자동으로 채점하고, 채점 결과의 신뢰성이 답안에 대해서 인간 채점자가 확인한다. 이 과정은 미채점 답안이 없거나 채점 확률이 0이 될 때까지 반복한다.

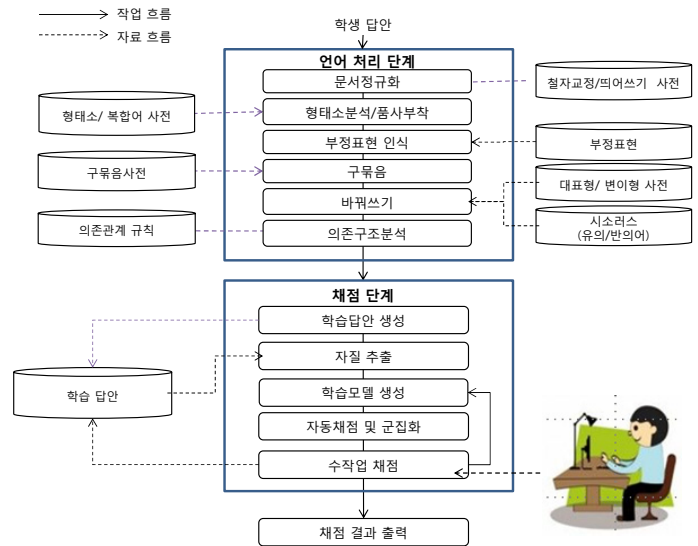


그림 1. 2015 한국어 문장 수준 서답형 문항 자동채점 시스템 구조도

3.2 ‘언어처리 단계’에서 보완된 사항

기존 언어처리 단계의 문서 정규화는 [문장부호 제거-띄어쓰기 교정-철자 교정]로 이루어졌으나 띄어쓰기 성능을 높이기 위해 [철자교정-띄어쓰기 교정-문장부호 제거] 순서로 변경했다. 또한, 2014 자동채점 시스템은 단문장 답안에 초점을 맞춰 설계되었으나 2015 자동채점 시스템은 다문장(多文章) 답안에 대비하기 위해 문장 분리 기능이 추가되었다. 문장 분리 기능은 문장부호 정보

와 앞 어절의 어미 정보, 뒤 어절의 어두 정보를 사용한 기계학습을 통해 구현했다. 그 외에도 부정부사(‘못’, ‘안’, ‘아니’), 부정 보조용언구(‘~지 못하/않/아니하’, ‘~지 마라’, ‘~지 말자’), 부정 구뭉음(‘~르/을 수 없’), 부정용언(‘아니다’, ‘없다’) 등을 이용하여 부정문과 이중부정 표현(‘~지 않으면 안 된다’)[17]을 인식하기 위한 부정표현 인식 기능[18]을 추가했다.

3.3 ‘채점 단계’에서 보완된 사항

2014 자동채점 시스템의 채점 단계에서는 초기 학습용 답안을 토대로 Logistic Regression과 k-Nearest Neighborhood 알고리즘으로 각각 채점 모델을 학습한 뒤, 두 모델의 결과를 결합한 앙상블(ensemble) 기법으로 미채점 답안들을 채점했다[11, 13-15]. 자동채점 시스템의 목적 중 하나는 채점자의 개입을 최소화하면서 신뢰성 있는 채점 결과를 얻는 것이다. 따라서 채점자가 개입하지 않은 자동채점 결과의 정확률을 높일 필요가 있다. 채점 단계의 알고리즘 조합을 실험을 통해 변경했다. k-Nearest Neighborhood 분류기의 경우 문항에 따른 채점 결과의 편차가 심해서 그와 유사하지만 안정적인 성능을 보이는 Nearest Centroid로 변경했다[12]. 또한 자동채점의 신뢰성을 더 높이기 위해 AdaBoosting 분류기를 추가하여 세 가지의 분류기의 결과가 일치하고, 분류 확률이 해당 단계의 임계값(threshold)보다 높은 결과만 자동채점의 결과로 확정하여 채점자에게 보이도록 알고리즘을 보완했다.

4. 실험 평가 및 분석

자동채점의 결과를 채점자가 확인하고 수정하는 경우를 개입, 채점자가 자동채점 결과를 수정하지 않는 경우를 미개입이라고 정의한다. 프로그램의 성능을 비교하기 위해서는 채점자에 의한 점수 확인 및 오류 정정의 영향을 덜 받는 ‘채점자 미개입’의 경우를 비교하는 것이 더 적절하다. 이 장에서는 채점자가 미개입한 2014 자동채점 시스템과 2015 자동채점 시스템의 성능을 비교 분석한다.

4.1 실험에 사용한 서답형 문항 및 정답

평가에 사용된 서답형 문항은 2014년에 시행된 “국가수준 학업성취도 평가”의 국어, 사회, 과학 교과 문항 중 1문장 수준의 문항을 선택했다[19]. 실험에 사용한 서답형 문항의 정보를 간략하게 정리하면 표 1과 같다. 각 문항의 학생 답안은 약 7,600개 이상이며, 동일한 상황에서 프로그램의 성능을 비교하기 위해 언어분석 옵션은 ‘띄어쓰기 교정, 문장부호 제거, 형태소 분석, 구뭉음’을 동일하게 적용했다.

언어분석 결과가 완전히 일치하는 답안들을 묶어 빈도순으로 정렬한 유형 중, 상위 10개 유형을 채점하여 초기 학습용 답안을 생성했다.

표 1. 문장 수준 자동채점 시스템 문항별 채점 정보

문항 번호	구분	전체 답안 수	초기 학습용 답안 수
고2국어6-(1)		7,965	7,317
고2국어2		7,965	6,448
중3과학6-(1)		7,440	4,539
중3과학2-(2)		7,440	5,369
중3사회2		7,442	4,262
중3국어6-(2)-㉠		7,453	5,940

4.2 평가 척도

정확률(accuracy)과 재현율(recall)은 일반적인 분류 시스템의 성능 평가 척도로 많이 사용된다[20-21]. 그러나 이 척도들은 자동채점 시스템의 경우, 기계학습을 위한 ‘초기 학습용 답안’의 생성, 자동채점의 반복 수행 후에 채점 결과가 산출되지 않은 미판단 답안의 ‘수작업 채점’에는 채점자의 판단을 필요로 하기 때문에 자동채점 시스템의 성능평가에 그대로 적용하기에는 적절하지 않다. 따라서 본 논문에서는 표 2와 같이 채점 답안을 분류하여 본래 의미를 최대한 유지하는 방향으로 정확률과 재현율을 계산한다. 표 2의 각 항목의 의미는 아래와 같다[12].

- a : 바르게 자동채점한 답안 수
- b : 틀리게 자동채점한 답안 수
- c : 올바르게 수작업 채점한 답안 수
- d : 틀리게 수작업 채점한 답안 수
- e : 올바르게 채점한 초기 학습용 답안 수
- f : 틀리게 채점한 초기 학습용 답안 수
(초기 학습용 답안이 정확하면 f=0)

표 2. 자동채점 결과의 정확성 평가를 위한 답안 분류

		채점 결과 맞음	채점 결과 틀림
자동 채점 대상 답안	자동 채점된 답안	a	b
	수작업 채점한 답안	c	d
초기 학습용 답안		e	f

인간채점 결과가 오류가 없다는 가정했을 때, 채점 결과가 인간채점 결과와 일치하면 채점이 바르게 된 것으로 간주한다.

정확률은 (식 1)로 계산한다. 이는 ‘자동채점 시스템이 채점한 답안(a+b)’ 중에서 실제로 ‘올바르게 채점된 답안(a)’의 비율로서, 자동채점 시스템의 결과가 얼마나 정확한지를 판단하는 척도로 나타낸다. 재현율은 (식 2)로 계산하며, ‘자동채점 대상 답안(a+b+c+d)’ 중 ‘올바르게 자동채점한 답안(a)’의 비율로, 자동채점 대상답안 중 자동채점 시스템이 정답을 얼마나 찾아냈는지를 나타낸다. (식 3)은 전체 답안 중에서 채점 결과가 올바른 답안의 비율로서, 완전 일치도(accuracy)를 의미한다.

$$\text{정확률} = \frac{a}{a+b} \quad (\text{식 1})$$

$$\text{재현율} = \frac{a}{a+b+c+d} \quad (\text{식 2})$$

$$\text{완전 일치도} = \frac{a+c+e}{a+b+c+d+e+f} \quad (\text{식 3})$$

4.3 2014 자동채점 시스템과 2015 자동채점 시스템의 성능 비교

그림 2는 2014 자동채점 시스템과 2015 자동채점 시스템의 문항별 완전일치도, 정확률과 재현율의 평균이다. 2015 자동채점 시스템의 완전일치도와 정확률이 2014 자동채점 시스템의 결과보다 증가한 것에 비해, 재현율은 감소한 것을 볼 수 있다.

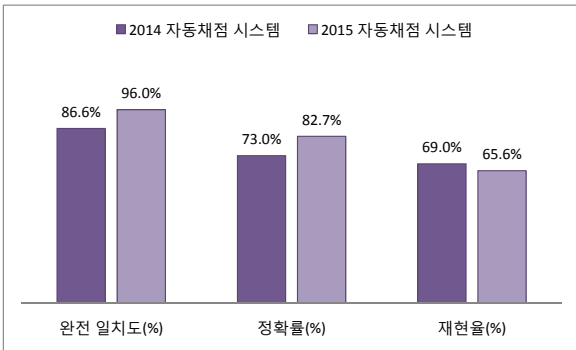


그림 2. 문장 수준 자동채점 시스템의 정확성 평균

완전일치도와 정확률이 높아졌으나 재현율이 낮아진 이유는 2014 자동채점 시스템의 채점 단계에서는 2개의 분류기의 채점 결과를 결합하였으나, 2015 자동채점 시스템에서는 3개의 분류기의 채점 결과를 사용하기 때문이다. 3개의 분류기의 결과가 완전히 일치해야만 채점이 되었다고 인식하기 때문에, 미채점 답안의 수가 증가한 것이 재현율 하락의 원인이다. 완전일치도가 평균 9.4%p, 정확률이 8.9% 증가한 것과 비교했을 때 재현율은 3.4%p 정도 하락했으므로, 2015 자동채점 시스템의 성능이 향상되었다고 판단할 수 있다.

문항별 문장 수준 자동채점 시스템 채점 결과의 완전일치도는 그림 3과 같다. 그림 4는 문항별 자동채점 시스템 채점 결과의 정확률을 비교한 결과이며, 그림 6은 재현율을 비교한 결과이다. 그림 3과 그림 4에서 2015 자동채점 시스템의 결과가 2014 자동채점 시스템보다 좋은 성능을 보였다.

그러나 그림 5에서 볼 수 있듯이 재현율은 하락한 것을 볼 수 있는데 이는 그림 3에서 분석한 것처럼 자동채점의 기준을 엄격하게 적용했기 때문에 수작업 채점 답안의 수가 늘어났기 때문이다.

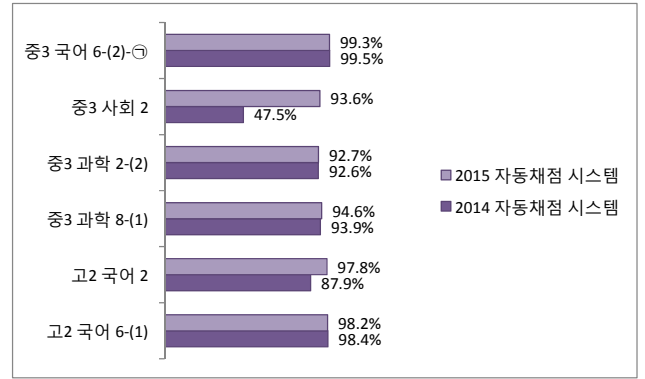


그림 3. 문장 수준 자동채점 시스템의 문항별 완전 일치도

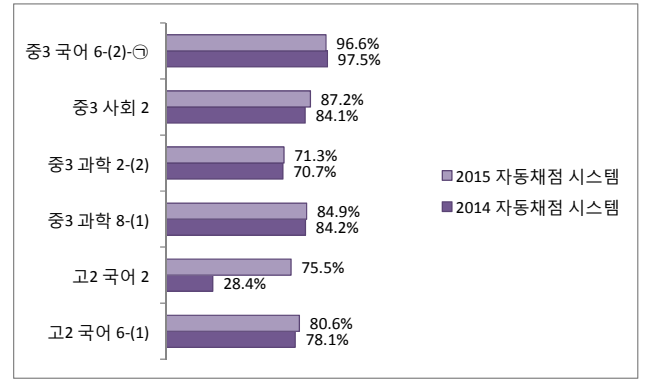


그림 4. 문장 수준 자동채점 시스템의 문항별 자동채점 정확률

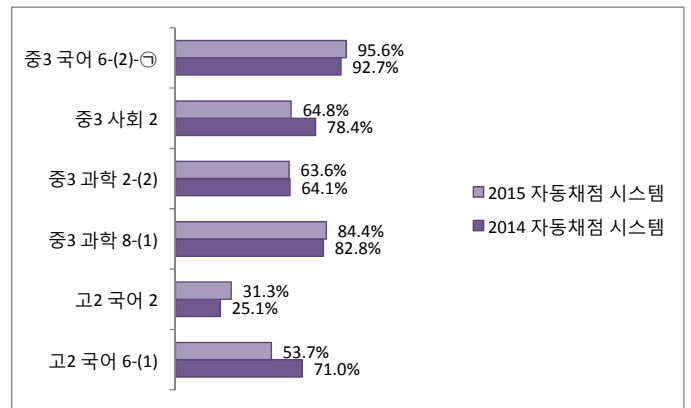


그림 5. 문장 수준 자동채점 시스템의 문항별 자동채점 재현율

5. 결론

본 논문에서는 2014년도에 개발된 프로토타입을 개선한 한국어 서답형 자동채점 시스템을 간략하게 소개하고 2014년도에 개발된 자동채점 시스템의 성능과 보완된 자동채점 시스템의 성능을 비교 분석하였다. 2015 자동채

점 시스템은 ‘언어 처리 단계’에서 여러 문장으로 구성된 답안에서 문장들을 분리하는 기능을 추가하였으며, 부정표현을 인식하는 기능도 추가하였다. 또한, 문서 정규화 단계의 순서를 [철자교정-띄어쓰기 교정-문장부호 제거] 순으로 변경하여 언어 처리 단계의 성능을 높였다. ‘채점 단계’에서는 기존의 Logistic Regression 분류기와 k-Nearest Neighborhood 분류기를 결합한 방식에서 채점 결과의 안정성과 신뢰성을 높이기 위하여 k-Nearest Neighborhood 분류기 대신 안정적인 채점 결과를 내는 Nearest Centroid 분류기를 사용하였으며, AdaBoosting 분류기를 추가하였다. 그 결과 자동채점의 평균 재현율은 기존의 시스템보다 약 3.4%p 감소하여 65.6%의 성능을 보였으나, 평균 완전 일치도가 96.0%로 기존의 시스템의 완전 일치도보다 약 9.4%p 증가하였으며, 자동채점의 평균 정확률은 82.7%로 약 8.9%p 증가했다. 재현율이 다소 감소하였으나 자동채점 시스템의 목적은 가능한 채점 시간을 단축하면서 채점 기준의 일관성과 신뢰성을 확보하는 데 있으므로, 2015년 자동채점 시스템은 2014년도에 개발된 프로토타입보다 자동채점의 성능이 향상되었다고 판단할 수 있다.

향후에는 문항별로 다양한 기계학습 분류기의 성능을 테스트하여 초기 학습용 답안으로 채점할 유형의 수와 문항별로 어떤 분류기의 조합이 좋은 성능을 내는지 실험을 통해 판단하고, 실험 결과에 따라 문항에 맞춰 분류기의 조합을 달리하여 효율적인 결과를 낼 수 있도록 시스템을 개선하는 방안을 연구할 예정이다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업[R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발]과 한국교육과정평가원의 “한국어 문장 수준 서답형 문항 자동채점 프로그램 고도화 개발 및 적용” 사업의 일환으로 수행하였음

참고문헌

[1] 교육과학기술부, 2009 개정 교육과정 총론, 교육과학기술부 고시 제 2009-41호, 2009.

[2] 교육과학기술부, 창의인재와 선진과학기술로 여는 미래 대한민국, 2011년 업무보고, 2010.

[3] 교육과학기술부, 2011 창의·인성 교육 기본 계획, 보도자료, 2011.

[4] 교육과학기술부, 중등학교 학사 관리 선진화 방안, 보도자료, 2011.

[5] 교육부, 초·중등학교 교육과정 총론, 교육부 고시 제 2015-74호, 2015.

[6] 한국교육평가학회, 교육평가용어사전, 서울: 학지사, 2004

[7] 김종성, 국어과 서답형 평가 길라잡이, 충청남도교육청 중등교육정책과, p. 7, 2009.

[8] Attali, Y. and Burstein, J., Automated Essay Scoring with E-rator v.2.0, ETS Research Report,

RR-04-45, 2005.

[9] Shermis, M. D. and Burstein, J., *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum Associates, Inc., Publishers. Mahwah, New Jersey, 2003.

[10] Chapelle, O. and Scholkopf, B., *et al, Semi-Supervised Learning*, Cambridge, England, MIT Press, pp. 1-8, 2006.

[11] 노은희, 이상하, 임은영, 성경희, 박소영, 한국어 서답형 문항 자동채점 프로그램 개발 및 실용성 검증, 한국교육과정평가원, 연구보고서 RRE 2014-6, 2014.

[12] 노은희, 송미영, 성경희, 박소영, 한국어 문장 수준 서답형 문항 자동채점 프로그램 개발 및 적용, 한국교육과정평가원, 연구보고서 RRE 2015-9, 2015.

[13] 천민아, 서형원, 김재훈, 노은희, 성경희, 임은영, “준지도 학습 방법을 이용한 한국어 서답형 문항 자동채점 시스템”, 한국인지과학회, 한국정보과학회, 제26회 한글 및 한국어 정보처리 학술대회 논문집, pp.112-116, 2014.

[14] 천민아, 김재훈식, 서형원, 노은희, 성경희, “준지도 학습기반 자동채점 시스템에서 채점자가 채점 결과에 미치는 영향”, 한국정보과학회, 2015 한국 컴퓨터종합학술대회 논문집, pp.666-668, 2015.

[15] 천민아, 서형원, 김재훈, 노은희, 성경희, 임은영, “준지도 학습 방법을 이용한 한국어 서답형 문항 자동채점”, 한국인지과학논문지, 제26권, 제2호, pp.147-165, 2015.

[16] Harrington, P., *Machine Learning in Action*, Manning Publications, 2012.

[17] 국립국어원, 외국인을 위한 한국어 문법 1, 서울: 커뮤니케이션북스, 2005.

[18] 이정태, 천민아, 김재훈, 세종 말뭉치로부터 용언 언어 추출, 제 27회 한글 및 한국어 정보처리 학술대회 논문집, pp.121-123, 2015.

[19] 한국교육과정평가원, 2014년도 국가수준 학업성취도 평가 기출답안 및 정답, 2014.

[20] 이준석, 정보검색론, 숭실대학교 정보검색연구실 (미간행), 2006.

[21] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern Information Retrieval (the concepts and technology behind search, 2nd edition)*, Addison Wesley, 2015.