KIPS Tr. Software and Data Eng.
Vol.5, No.11 pp.527~534 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2016.5.11.527

Automated Scoring System for Korean Short-Answer Questions Using Predictability and Unanimity

Min-Ah Cheon† · Chang-Hyun Kim†† · Jae-Hoon Kim††† · Eun-Hee Noh†††† · Kyung-Hee Sung††††† · Mi-Young Song††††

ABSTRACT

The emergent information society requires the talent for creative thinking based on problem-solving skills and comprehensive thinking rather than simple memorization. Therefore, the Korean curriculum has also changed into the direction of the creative thinking through increasing short-answer questions that can determine the overall thinking of the students. However, their scoring results are a little bit inconsistency because scoring short-answer questions depends on the subjective scoring of human raters. In order to alleviate this point, an automated scoring system using a machine learning has been used as a scoring tool in overseas. Linguistically, Korean and English is totally different in the structure of the sentences. Thus, the automated scoring system used in English cannot be applied to Korean. In this paper, we introduce an automated scoring system for Korean short-answer questions using predictability and unanimity. We also verify the practicality of the automatic scoring system through the correlation coefficient between the results of the automated scoring system and those of human raters. In the experiment of this paper, the proposed system is evaluated for constructed-response items of Korean language, social studies, and science in the National Assessment of Educational Achievement. The analysis was used Pearson correlation coefficients and Kappa coefficient. Results of the experiment had showed a strong positive correlation with all the correlation coefficients at 0.7 or higher. Thus, the scoring results of the proposed scoring system are similar to those of human raters. Therefore, the automated scoring system should be found to be useful as a scoring tool.

Keywords : Machine Learning, Korean Automated-Scoring System, Unanimity, Predictability, Natural Language Processing

기계학습 분류기의 예측확률과 만장일치를 이용한 한국어 서답형 문항 자동채점 시스템

천 민 아† · 김 창 현†† · 김 재 훈††† · 노 은 희†††† · 성 경 희††††† · 송 미 영††††

요 약

최근 정보화 사회에서는 단순 암기보다는 문제 해결 능력과 종합적인 사고력을 바탕으로 창의적인 생각을 할 수 있는 인재를 요구한다. 이에 따라 교육과정도 학생들의 종합적인 사고력을 판단할 수 있는 서답형 문항을 늘리는 방향으로 변하고 있다. 그러나 서답형 문항의 경우 채점자의 주관에 의존하여 채점이 진행되기 때문에, 채점 결과의 일관성을 확보하기 어렵다는 단점이 있다. 이런 점을 해결하기 위해 해외에서는 기계학습을 이용한 자동채점 시스템을 채점 도구로 사용하고 있다. 한국어는 영어와 언어학적으로 다른 분류에 속하므로 영어권에서 사용하는 자동채점 시스템을 한국어에 그대로 적용할 수 없다. 따라서 한국어 체계에 맞는 자동채점 시스템의 개발이 필요하다. 본 논문에서는 기계학습 분류기의 예측확률과 만장일치 방법을 사용한 한국어 서답형 문항 자동채점 시스템을 소개하고, 자동채점 시스템을 이용한 채점 결과와 교과 전문가의 채점 결과를 비교하여 자동채점 시스템의 실용성을 검증한다. 본 논문의 실험을 위해 2014년 국가수준 학업성취도 평가의 국어, 사회, 과학 교과의 서답형 문항을 사용했다. 평가 척도로 피어슨 상관계수와 카파계수를 사용했다. 채점자가 개입했을 때와 개입하지 않았을 때의 상관계수 모두 0.7 이상으로 강한 양의 상관관계를 보였다. 이는 자동채점 시스템이 교과 전문가가 채점한 결과와 유사한 방향으로 답안에 점수를 부여한 것이므로 자동채점 시스템을 채점 보조도구로서 충분히 사용할 수 있을 것이다.

키워드 : 기계학습, 한국어 자동채점 시스템, 만장일치제, 정답 예측확률, 자연어 처리

※ 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.
† 비 회 원 : 한국해양대학교 컴퓨터공학과 박사과정
†† 정 회 원 : 한국전자통신연구원 언어처리연구실 책임연구원
††† 종신회원 : 한국해양대학교 IT공학부 교수
†††† 비 회 원 : 한국교육과정평가원 연구위원
††††† 비 회 원 : 한국교육과정평가원 부연구위원
Manuscript Received : October 4, 2016
Accepted : October 12, 2016
* Corresponding Author : Jae-Hoon Kim(jhoon@kmou.ac.kr)

1. 서 론

21세기 정보화 사회에서는 단순히 많은 것을 아는 것보다 자신이 지닌 지식과 정보를 바탕으로 새로운 것을 창조할 수 있는 사고방식과 능력을 지닌 인재를 요구한다[1]. 이에 정부는 '창의인성교육'을 위해 개정된 교육과정을 도입하였으며, 서답형 문항과 수행평가의 비중을 늘려 학생들의 종합적인 사고능력을 평가하고자 노력하고 있다[2-6]. 서답형 문항은 선택형 문항과 달리 학생이 직접 정답을 구성하는 문항 형태로 학생들의 비판적 사고력, 문제 해결 능력과 창의력을 측정하는 데 적합하다[7-9]. 그러나, 채점비용이 많이 들고, 채점 시간이 많이 소요될 뿐 아니라, 채점자의 주관적 판단에 의존하여 채점하므로 채점 결과에 대한 공정성과 신뢰성을 보증하기 어렵다[8-9]는 문제가 있다. 이런 이유로 해외에서는 기계학습을 이용한 자동채점 시스템[10-13]을 활용하고 있다. 국내에서는 한국교육과정평가원에서 한국어 서답형 문항을 채점하기 위한 자동채점 시스템의 프로토타입(prototype)을 개발하여 실제 시험에 적용하려는 방안을 연구 중이다[14-17].

본 논문에서는 정답 예측확률과 다수의 분류기의 만장일치제를 이용한 한국어 서답형 문항 자동채점 시스템을 제안한다. 준지도학습에서 초기에는 정답을 예측할 확률을 높게 설정하여 학습자료의 문제를 극복하여 학습자료가 충분해지면 서서히 줄여 많은 미채점 답안을 채점할 수 있도록 하였다. 중요한 시험에서 채점오류는 자동채점 시스템의 신뢰성을 떨어뜨릴 뿐 아니라 때로는 중대한 사회적 문제를 야기할 수도 있다. 본 논문에서는 이러한 문제를 최소화하기 위해서 3 종류의 서로 다른 특성을 가진 분류기를 사용하고 이들의 채점 결과가 완전히 일치할 때만 채점된 것으로 간주하는 만장일치제를 사용한다. 제안된 자동채점 시스템의 평가를 위한 실험에 2014년 국가수준 학업성취도 평가의 국어, 사회, 과학 교과의 서답형 문항을 사용했다. 평가 척도로 피어슨 상관계수와 카파계수를 사용했다. 채점자가 개입했을 때와 개입하지 않았을 때의 상관계수 모두 0.7 이상으로 강한 양의 상관관계를 보였다. 이는 자동채점 시스템이 교과 전문가가 채점한 결과와 유사한 방향으로 답안에 점수를 부여한 것이므로 자동채점 시스템을 채점 보조도구로서 충분히 사용할 수 있을 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 서답형 문항 자동채점 시스템 개발을 위해 필요한 관련 연구들을 소개한다. 3장에서는 한국어 문장 수준 서답형 문항 자동채점 시스템에 관해 설명한다. 4장에서는 구현된 자동채점 시스템으로 실제 문항을 채점한 결과와 사람이 채점한 결과를 비교·분석한다. 마지막으로 5장에서는 결론 및 향후 연구에 관해서 논의한다.

2. 관련 연구

2.1 자동채점 시스템

해외의 자동채점 시스템은 TOFEL(Test of English as a Foreign Language), GMAT(Graduate Management Admission Test), GRE(Graduate Record Examinations) 등의 대규모 시험에서 사용되고 있다[10-13]. 현존하는 자동채점 시스템들은 어디까지나 채점자가 채점을 쉽게 할 수 있도록 보완하는 수단으로 사용되고 있으며[9-13], 대부분 기계학습 알고리즘으로 구현되어 있다[10-13]. 해외에서의 자동채점 시스템의 채점 대상은 여러 문단으로 이뤄진 논술형 문항이며, 국내의 자동채점 시스템의 채점 대상은 서답형에 초점을 맞추고 있다[14-22]. 국내에서는 한국교육과정평가원에서 한국어 서답형 문항을 자동채점하기 위한 프로토타입[14-17]을 개발하여 실용성을 검증하고 있으며, 실제 대규모 시험에 적용하려는 방안을 연구 중이다[14-16].

2.2 준지도 학습

기계학습(machine learning)[23, 24]은 인공지능의 한 분야로, 컴퓨터가 새로운 지식을 학습하고, 학습한 정보를 효율적으로 사용할 수 있도록 하는 기술과 알고리즘을 총칭하는 말이다. 기계학습의 학습 방법으로는 지도 학습(supervised learning), 자율 학습(unsupervised learning), 앞의 두 가지 방법을 혼합한 준지도 학습(semi-supervised learning) 방법으로 나눌 수 있다[23, 24].

지도 학습 방법은 정답이 부착된(labeled) 데이터를 이용하여 분류기(classifier)를 학습하고, 학습된 분류기를 통해 새로운 입력 데이터에 가장 적절한 정답을 찾는 방법이다. 지도 학습 방법은 학습 데이터를 만들기까지 많은 시간과 노력이 필요하다. 자율 학습은 정답이 부착되지 않은 데이터로부터 데이터의 구조나 관계를 파악하여 패턴을 통해 자료를 분류하는 방법이다. 이 두 가지 방법의 장점을 취한 것이 준지도 학습 방법으로, 다양한 분야에서 사용되고 있다[24]. 준지도 학습 방법은 정답이 부착된 데이터로 학습하고, 정답 예측 확률이 기준값 이상인 데이터에만 정답을 부착한다. 정답이 부착된 데이터는 학습 데이터로 추가되고, 늘어난 학습 데이터를 이용하여 새로 학습을 진행한다. 이 과정이 반복되면 학습 데이터의 양이 점점 늘어나서 분류기의 정확률 및 신뢰성을 지속해서 개선할 수 있다[24].

2.3 앙상블 학습 알고리즘

앙상블 학습(ensemble learning)은 여러 개의 분류기를 조합하는 것이다[22, 23]. 즉, 여러 개의 분류기 중 성능이 가장 좋은 분류기의 결과를 채택하는 것이 아니라, 다양한 분류기의 결과 중 다수의 의견을 따르는 투표 방식(voting)과 분류기의 결과가 모두 일치할 때만 정답으로 인정하는 만장일치(unanimity) 방법이 있다. 앙상블 학습도 여러 가지 방법으로 구현할 수 있다. 앙상블 방식으로는 균일한 확률 분포에 따라 학습 데이터로부터 반복적으로 샘플링을 하여 여러 분류기의 결과를 조합하는 배깅(bagging), 분류기가 분류하기 어려운 사례에 집중하도록 학습 데이터의 분포를 변경시켜서 학습하는 부스팅(boosting), 다수의 의사결정 트리에 의한 예측을 조합하는 무작위 추출(randomization) 등이 있다[23, 24].

3. 한국어 서답형 문항 자동채점 시스템

3.1 시스템 구조

한국어와 영어는 언어학적으로 다른 분류에 속하기 때문에[25], 영어권에서 사용하는 자동채점 시스템을 한국어에 그대로 적용하는 것은 어려운 점이 있다. 따라서 한국어 서답형 문항을 위한 자동채점 시스템을 새로 설계할 필요가 있다. 한국어 서답형 자동채점 시스템의 구조도는 Fig. 1과 같다. Fig. 1에서 실선은 작업 흐름을 나타내고, 파선은 데이터의 흐름을 나타낸다. 한국어 서답형 문항 자동채점 시스템은 크게 언어 처리 단계와 채점 단계로 나뉜다.

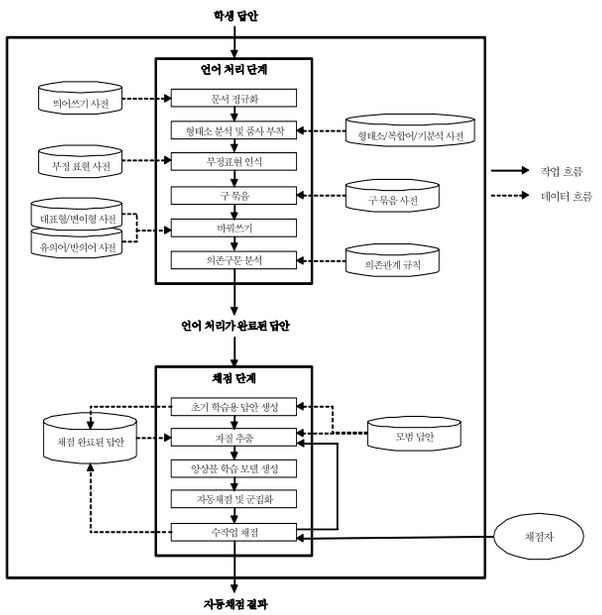

Fig. 1. The Overall Architecture of the Korean Automated Scoring System

언어 처리 단계의 입력은 학생 답안이다. 이 단계에서는 자동채점에 필요한 언어 정보를 분석하기 때문에 많은 언어 자원을 필요로 한다. 언어 처리가 완료된 답안 중, 언어 처리 결과가 완전히 일치하는 답안들은 하나의 유형으로 묶인다. 이 유형은 채점 단계에서 유형에 속한 답안들이 많은 순서대로 채점자에게 보이게 된다. 고빈도로 정렬된 유형들 중 상위 몇 개의 답안에 관해서 채점자가 채점을 진행한다. 이렇게 채점이 완료된 답안들은 출제자가 모범 답안으로 구축해놓은 답안들을 학습 데이터로 사용하게 된다. 학습 데이터로부터 자질을 추출하여 앙상블 학습 모델을 생성한다. 생성한 모델을 바탕으로 채점되지 않은 답안들에 가장 적절하다고 생각되는 점수를 부여한 뒤, 예측 확률과 점수가 일치하는 군집(cluster)들을 생성하여 채점자에게 확인을 받는다. 채점자가 확인을 끝낸 답안들은 채점 완료된 답안으로 다음 채점 단계의 학습 데이터로 사용된다. 각 단계의 자세한 내용은 이하의 절에서 자세히 설명한다.

3.2 분석 단계

이 절에서는 한국어 서답형 문항 자동채점 시스템의 언어 처리 단계에 관해 설명한다. 언어 처리 단계는 언어 정보를 분석하는 단계로서 문서정규화, 형태소 분석 및 품사 부착, 부정표현 인식, 구 묶음, 바꿔쓰기, 의존구문 분석으로 이루어져 있다.

문서정규화는 철자교정, 문장 분리, 띄어쓰기 교정, 문장부호 제거로 이루어져 있다. 철자교정은 대칭 삭제 교정 알고리즘[26]으로 구현했다. 문장 분리 기능은 문장부호 정보와 앞 어절의 어미 정보, 뒤 어절의 어두 정보를 사용한 기계학습을 통해 구현했다. 띄어쓰기 교정은 [27]의 기계학습 방법으로 구현했다. 문장부호 제거는 정규표현식(regular expression)을 통해 구현했다.

형태소 분석은 [28]에 나와 있는 것과 같이 형태소의 경계에서 기계학습을 이용한 띄어쓰기 후, 형태소 사전을 탐색하여 가능한 모든 형태소 후보들을 찾는다. 품사 부착은 찾아낸 형태소 후보, 문맥확률과 어휘확률을 이용하여 가중치 네트워크를 만든 뒤, 가중치 네트워크에서 가장 적절한 경로를 찾는 방법으로 구현했다.

부정표현 인식은 부정부사(‘못’, ‘안’, ‘아니’), 부정 보조용언구(‘~지 못하/않/아니하’, ‘~지 마라’, ‘~지 말자’), 부정 구묶음(‘~ㄹ/을 수 없’), 부정용언(‘아니다’, ‘없다’) 등을 이용하여 부정문과 이중부정 표현(‘~지 않으면 안 된다’)[29]을 인식하기 위한 방법으로 [15]에서 언급한 것과 같이 세종말뭉치에서 부정문 패턴을 분석하여 정규 표현식을 통해 추출했다.

구 묶음은 띄어쓰기에서 사용했던 기계학습 방법과 구 묶음 사전을 이용하여 구현했다. 구 묶음을 수행하기 위해서는 형태소 분석 및 품사 부착 과정이 반드시 선행되어야 한다.

바꿔쓰기는 단어 치환을 사용하여 조사를 대표조사로, 어미를 대표어미로, 단어를 동의어로 치환한다. 대표형/변이형 사전에 대한 해당 정보가 없으면 입력된 단어를 그대로 출력한다.

의존구문 분석은 의존문법을 기반으로 구문분석을 수행한다. 의존문법은 의존관계에 있는 언어요소 중 의미의 중심이 되는 지배소(governor)와 지배소가 갖는 의미를 보완해주는 의존소(dependent)의 관계를 문법으로 표현한 것이다. 이 기능은 스웨덴의 Vaxjo 대학교와 Uppsala 대학교에서 공동으로 개발한 의존구문 분석기인 MaltParser[30]를 이용하여 구현했다.

3.3 채점 단계

이 절에서는 채점 단계에서 수행하는 기능들과 자동채점에서 사용하는 앙상블 학습 알고리즘에 관해 설명한다. 채점 단계는 초기 학습용 답안 생성, 자질추출, 앙상블 학습 모델 생성, 자동채점 및 군집화, 수작업 채점의 순서로 진행된다. 초기 학습용 답안 생성을 제외한 나머지 과정은 계속 반복한다.

초기 학습용 답안은 언어 처리 결과가 완전히 일치하는 학생 답안들을 하나의 유형으로 간주하고, 각 유형에 속한 답안 수의 빈도순으로 정렬한다. 그 뒤, 채점자가 고빈도 답안의 유형에 관해 수동으로 채점을 진행하여 학습 모델 생성에 사용할 학습 데이터를 생성한다. 이 때, 유형을 몇 개나 채점할 것인지는 채점자의 판단에 맡긴다.

자질추출은 해당 시점까지 확장된 학습 데이터에서 단어 자질, 어절 자질, 구문 자질을 추출한다. 단어 자질은 내용어만 추출하고, 어절 자질은 내용어와 정규화된 기능어를 자질로 추출한다. 구문 자질은 의존어와 지배어, 그리고 의존관계를 자질로 추출한다. 자질 가중치는 정보검색에서 널리 사용되는 TF-IDF[31]를 사용하여 구한다.

앙상블 학습 모델 생성과 자동채점 및 군집화는 Fig. 2에 나와 있는 알고리즘으로 구현한다. 자동채점 시스템의 목적은 채점자의 개입을 최소화하면서 채점 결과의 신뢰성과 공정성을 확보하는 것이다[14-18]. 따라서 Fig. 2에 나타난 것과 같이 분류기의 예측확률이 임계값(threshold)이며, 앙상블 학습 중 만장일치제로 결정이 된 미채점 답안에 관해서만 점수를 부여한다.

Fig. 2에서 trainAnswers와 trainLables는 각각 해당 단계까지 채점이 완료된 답안들과 각 답안의 정답들이다. testAnswers는 해당 단계까지 채점되지 않은 답안들이다. 이 세 가지 변수들이 자동채점 및 군집화 알고리즘의 입력이다. 입력이 들어오면 trainAnswers와 testAnswers에서 자질을 추출하여 자질 행렬 trainX와 testX를 생성한다.

생성된 자질 행렬 trainX와 정답 리스트 trainLables를 이용하여 Logistic Regression, Nearest Centroid, 그리고 AdaBoosting 분류기를 학습한다[23]. 시스템에서는 자동채점 결과인 show_testAnswers, show_testLables과 show_probs를 이용하여 결과를 군집화(clustering)하여 보여준다. 임계값은 이전 시스템과 마찬가지로 0.99에서 시작하여 자동채점이 진행될 때마다 0.03씩 감소하도록 설계했다.

```
입력 : trainAnswers, trainLables, testAnswers
출력 : results of clustering

1. //trainAnwers로부터 자질 행렬 trainX를 생성
create a feature matrix trainX with trainAnwers
2. //testAnswers로부터 자질 행렬 testX를 생성
create a feature matrix testX with testAnswers

3. //trainX와 trainLables를 이용해서
//Logistic Regression 분류기를 학습
classifier1 = LogisticRegression(trainX, trainLables)
4. //Nearest Centroid 분류기를 학습
classifier2 = NearestCentroid(trainX, trainLables)
5. //AdaBoosting 분류기를 학습
classifier3 = AdaBoosting(trainX, trainLables)

6. //미채점 답안의 자질 행렬 testX을 통해 분류 확률을 계산
testProb = classifier1.predict_proba(testX)

7. //학습된 분류기들을 이용해서 미채점 답안을 채점
testy1 = classifier1.predict(testX)
testy2 = classifier2.predict(testX)
testy3 = classifier3.predict(testX)

8. //분류 결과를 확인
create list new_testAnswers
create list show_testAnswers
create list show_testLables
create list show_probs

for i in range(len(testy1)):
if testProb[i] > threshold
        and testy1[i] == testy2[i] == testy3[i]:
  add testAnswers[i] to show_testAnswers
  add testLables[i] to show_trainAnswers
  add testProb[i] to show_probs
   else:
     add testAnswers[i] to new_testAnswers

testAnswers = new_testAnswers

9. //show_testAnswers, show_tesLables,
    //show_probs의 정보를 이용하여 군집화
   make results of clustering
    with show_testAnswrs, show_testLables, show_probs

return results of clustering
```

Fig. 2. The Algorithm for Automated Scoring and Clustering

4. 실험 평가 및 분석

자동채점의 결과를 채점자가 확인하고 수정하는 경우를 개입, 채점자가 수정하지 않은 경우를 미개입이라고 정의한다[14-17]. 본 논문에서는 채점자가 개입 여부에 따른 자동채점 시스템의 채점 결과와 인간 채점자의 채점 결과의 상관관계와 정확률을 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[15, 32]. 실험에 사용한 서답형의 문항의 간략한 정보는 표 1과 같다. 실험에 사용한 문항의 평균 학생 답안은 약 7,600개이며, 초기 학습용 답안의 개수는 평균 약 5,600개이다.

Table 1. The Informations of Short-Answer Questions Used in the Experiment

구분 / 문항 번호	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
평균	7,617	5,645

각 문항에 적용한 언어 처리 옵션은 띄어쓰기 교정, 문장 부호 제거, 형태소 분석, 구 묶음이다. 언어 처리 결과가 완전히 일치하는 답안들을 묶어 빈도순으로 정렬한 유형 중, 상위 10개 유형을 채점하여 초기 학습용 답안을 생성했다.

4.2 평가 척도

인간 채점자의 채점 결과는 각 문항의 교과 전문가가 여러 단계를 거쳐 확정한 점수로 채점 상의 오류가 없다고 가정한다[15]. 본 실험에서는 이 인간 채점자의 채점 결과와 자동채점 시스템의 채점 결과가 얼마나 유사한지를 피어슨 상관계수[33]와 카파계수[34, 35]를 통해 판단한다. 피어슨 상관계수는 두 변량 X, Y 사이의 상관 정도와 방향을 나타내는 수치이며, 카파계수는 점수 분류(class) 정보에 대한 평가자 2명의 일치도를 측정하는 통계적인 지표이다. 상관계수 r과 카파계수 κ가 가지는 값의 범위는 $[-1.0, 1.0]$이다. 상관계수 r이 양의 큰 값이라면 X가 커질 때, Y도 커지므로 두 변량은 양의 관계를 가진다고 판단할 수 있으며[33], 카파계수 κ는 두 평가자가의 평가 결과가 일치할수록 높은 값을 가진다[34, 35].

4.3 자동채점 시스템의 채점 결과와 인간 채점자의 채점 결과의 피어슨 상관계수

Fig. 3은 문항별 자동채점 시스템의 채점 결과와 인간 채점자의 피어슨 상관계수 결과이다. 채점자가 개입했을 때와 개입하지 않았을 때의 평균 피어슨 상관계수는 각각 0.945, 0.886으로 나타났다.

문항별 피어슨 상관계수를 비교해보면 채점자가 개입한 경우가 미개입한 경우보다 교과 전문가의 채점 결과와 높은 상관계수를 보인다. 이는 채점자가 자동채점 시스템의 앙상블 모델에서 채점된 결과를 확인하고, 점수가 잘못 부여된 답안의 결과를 수정함으로써 다음 자동채점 단계에서 사용되는 학습 데이터의 오류를 제거했기 때문이다. 피어슨 상관계수 r이 $(0.7, 1.0]$에 해당하면 강한 양의 상관관계를 가진다[33]는 것을 뜻한다. 이는 자동채점 시스템을 이용하여 채점한 결과와 교과 전문가가 채점한 결과가 매우 유사한 양상을 보인다는 것으로 해석할 수 있다.

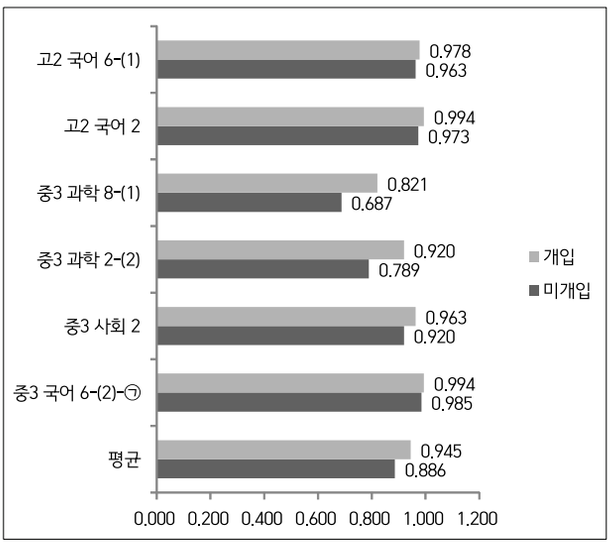

Fig. 3. Pearson Correlation Coefficient Results Per Short-Answer Questions

4.4 자동채점 시스템의 채점 결과와 인간 채점자의 채점 결과의 카파계수

Fig. 4는 문항별 자동채점 시스템의 채점 결과와 인간 채점자의 카파계수 결과이다. 채점자가 개입했을 때와 개입하지 않았을 때의 평균 카파계수는 각각 0.931, 0.862으로 나타났다.

문항별 카파계수를 비교해보면 피어슨 상관계수와 마찬가지로 채점자가 개입한 경우가 미개입한 경우보다 교과 전문가의 채점 결과와 높은 일치도를 보였다. 카파계수 κ의 값의 범위가 $(0.75, 1.0]$에 해당하면 두 분류 간 일치하는 정도가 매우 높다[34, 35]고 해석된다. 따라서 자동채점 시스템을 이용한 채점 결과와 교과 전문가의 채점 결과가 매우 유사하다고 판단할 수 있다.

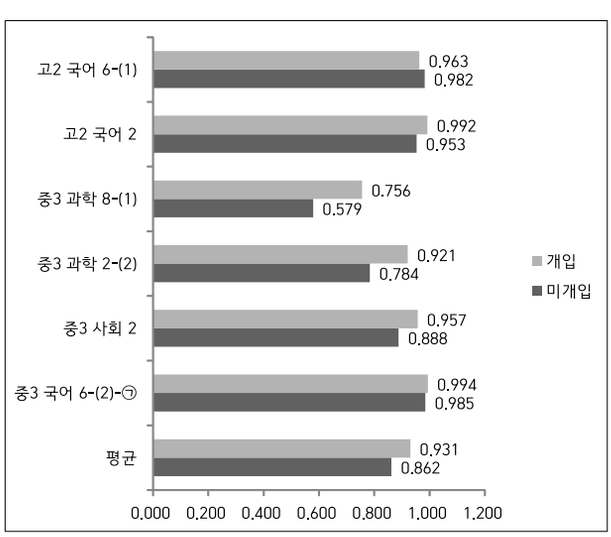

Fig. 4. Kappa Correlation Coefficient Results Per Short-Answer Questions

5. 결 론

본 논문에서는 현재 개발 중인 기계학습 분류기의 예측확률과 만장일치 기법을 사용한 한국어 서답형 자동채점 시스템의 채점 결과와 인간 채점 결과를 비교·분석하여 자동채점 시스템의 실용성에 관해 검증하였다. 개발 및 성능 개선을 진행 중인 자동채점 시스템은 채점이 완료된 답안과 모범 답안을 학습 데이터로 사용하여 앙상블 분류기를 학습한다. 앙상블 분류기는 Logistic Regression 분류기와 Nearest Centroid 분류기, Adaboosting 분류기의 채점 결과를 조합으로 구현했다. 자동채점 시스템의 실용성을 검증하기 위해 채점자가 채점 결과에 개입, 미개입했을 때의 채점 결과와 교과 전문가가 채점한 결과 사이의 상관계수를 비교했다. 비교결과 채점자의 개입 여부와 상관없이 평균 피어슨 상관계수와 평균 카파계수가 모두 0.75 이상의 값이었다. 이는 자동채점 시스템이 교과 전문가가 채점한 결과와 매우 유사하게 답안에 점수를 부여한 것이라고 해석할 수 있다. 자동채점 시스템을 이용한 채점 결과의 일관성과 신뢰성, 시간의 절약성은 [14-18]에 증명되어 있다. [14-18]의 실험 결과와 본 논문의 실험 결과를 조합해보면 자동채점 시스템을 사용하면 채점의 신뢰성을 확보할 수 있고 시간을 절약할 수 있으며 교과 전문가가 채점한 채점 결과와 유사한 방향으로 채점이 진행되므로, 자동채점 시스템을 실제 채점 환경에서 유용한 도구로 사용할 수 있다.

향후에는 문항별로 다양한 기계학습 분류기의 성능을 테스트하여 초기 학습용 답안으로 채점할 유형의 수와 문항별로 어떤 분류기의 조합이 좋은 성능을 내는지, 어떤 자질을 사용해야 하는지, 적절한 자질의 수는 어느 정도인지를 실험을 통해 판단할 예정이다. 또한 실험 결과에 따라 문항에 맞춰 분류기의 조합을 달리하여 보다 효율적인 결과를 낼 수 있도록 시스템을 개선하는 방안을 연구할 예정이다.

References

[1] S-D. Choi, J-Y. Kim, S-J. Ban, K-J. Lee, S-J. Lee, and H-Y. Choi, “Education Strategy to Foster Creative Talent for the 21st Century,” Korean Educational Development Institute Research Report, RR 2011-01, 2011.

[2] Ministry of Education, Science, and Technology, “Introduction to 2009 Revised National Curriculum,” Ministry of Education, Science, and Technology Notification (2009-41), 2009.

[3] Ministry of Education, Science, and Technology, “The Future Korea to Open Using Creative Talent and Advanced Science and Technology,” 2011 Business Report, 2010.

[4] Ministry of Education, Science, and Technology, “The Master Plan for Creativity-Character Education,” Press Release, 2011.

[5] Ministry of Education, Science, and Technology, “The Plan for Improving Education Management for Secondary Schools,” Press Release, 2011.

[6] Ministry of Education, “Introduction to National Curriculum for Elementary and Secondary Schools,” Ministry of Education Notification (2015-74), 2015.

[7] Korean Society for Educational Evaluation, “Dictionary of Educational Evaluation Terms,” Seoul: Hakjisa, 2004.

[8] J-S. Kim, “Guidelines for Short-Answer Questions in Korean Subject,” Secondary Schools Policy Division in Chungcheongnamdo Office of Education, p.7, 2009.

[9] K-A. Jin, “Development of Automated Scoring System for English Writing,” *English Language & Literature Teaching*, Vol.13, No.1, pp.236-237, 2007.

[10] Y. Attali and J. Burstein, “Automated Essay Scoring with E-rator v.2.0,” ETS Research Report RR-04-45, 2005.

[11] M. D. Shermis and J. Burstein, “Automated Essay Scoring: A Cross-Disciplinary Perspective,” Inc., Publishers. Mahawah, New Jersey, 2003.

[12] L. M. Rudner, V. Garcia, and C. Welch, “An Evaluation of the IntelliMetricSM Essay Scoring System,” *The Journal of Technology, Learning, and Assessment*, Vol.4, No.4, 2006.

[13] ETS, ETS Automated Scoring Technologies, ETS Report, 2010.

[14] N-H. Noh, S-H. Lee, E-Y. Lim, K-H. Sung, and S-Y. Park, “The Development and Evaluation for Automatic Scoring Programs in Korean Large-Scale Assessments,” Korea Institute of Curriculum & Evaluation, Research Report RRE 2014-6, 2014.

[15] E-H. Noh, M-Y. Song, K-H. Sung, and S-Y. Park, “Refinements and Application of Automatic Scoring Programs for Korean Large-scale Assessments,” Korea Institute of Curriculum & Evaluation, Research Report RRE 2015-9, 2015.

[16] M.-Y. Song, E.-H. Noh, and K.-H, Sung, “Analysis on the Accuracy of Automated Scoring for Korean Large-scale Assessment,” *The Journal of Curriculum and Evaluation*, Vol.19, No.2, pp.255-274, 2016.

[17] M-A. Cheon, H-W. Seo, J-H. Kim, E-H. Noh, K-H. Sung, and E-Y. Lim, “Semi-Automatic Scoring for Short Korean Free-Text Responses Using Semi-Supervised Learning,” *Korean Journal of Cognitive Science*, Vol.26, No.2, pp.147-165, 2015.

[18] M-A. Cheon, H-W. Seo, J-H. Kim, E-H. Noh, and K-H. Sung, “Effects of Human Raters on Results of an Automatic Scoring System Based on Semi-Supervised Learning,” *Proceedings of Korea Computer Congress 2015*, pp.666-668, 2015.

[19] D. Y. Jung, “Evaluation of Short and Long Essay Questions By Using Vector similarity and Thesaurus,” Master's Thesis in Graduate School of Education Dongguk University, 2001.

[20] H. J. Park and W. S. Kang, “Design and Implementation of a Subjective-type Evaluation System Using Natural Language Processing Technique,” *The Journal of Korean Association of Computer Education*, Vol.6, No.3, pp.207-217, 2003.

[21] W.-S. Kang, “Automatic Grading System for Subjective Questions Through Analyzing Question Type,” *The Journal of the Korea Contents Association*, Vol.11, No.2, pp.13-21, 2011.

[22] W. J. Cho, J. S. Oh, J. Y. Lee, and Y.-S. Kim, “An Intelligent Marking System based on Semantic Kernel and Korean WordNet,” *The KIPS Transactions: Part A.*, Vol.12, No.6, pp.539-546, 2005.

[23] P. Harrigton, “Machine Learning in Action,” Manning Publications, 2012.

[24] A. Sogaard, “Semi-Supervised Learning and Domain Adaptation in Natural Language Processing,” Morgan & Claypool Publishers, 2013.

[25] S.-S. Kang, “Korean Morphological Analysis and Information Retrieval (Korean edition),” Hong Reunggwahakchulpansa, 2002.

[26] Romoku, [Internet] http://blog.faroo.com/2012/06/07/improved-edit-distance-based-spelling-correction/.

[27] K. S. Shim, “Automatic Word Spacing based on Conditional Random Fields,” *Korean Journal of Cognitive Science*, Vol.22, No.2, pp.217-233, 2011.

[28] M.-A. Cheon, “Morphological Analysis and Part-of-Speech Tagging for Applying Korean Automated Scoring of Short-Answer Questions,” Master's Thesis in Graduate School of Korea Maritime and Ocean University, 2016.

[29] The National Institute of The Korean Language, “Korean Grammar for Foreigners 1,” Seoul: Communicationbooks, 2005.

[30] J. Nivre, “Algorithms for Deterministic Incremental Dependency Parsing,” *Computational Linguistics*, Vol.34, No.4, pp.513-553, 2008.

[31] G. Casella, S. Fienberg and I. Olkin, An Introduction to Statistical Learning with Applications in R, Springer.

[32] Korea Institute for Curriculum & Evaluation, “Test Paper and Answers in 2014 National Assessment of Educational Achievement of Korea,” 2014. (http://www.kice.re.kr/boardCnts/list.do?type=default&page=2&searchStr=&m=030302&C06=&boardID=1500208&C05=&C04=&C03=&searchType=S&C02=&C01=&s=kice).

[33] D. M. Corey, W. P. Dunlap, and M. J. Burke, “Averaging Correlations: Expected Values and Bias in Combined Pearsrs and Fisher's z Transformations,” *The Journal of General Psychology*, Vol.125, No.3, pp. 245-261, 1998.

[34] J. Carletta, “Assessing Agreement on Classification Tasks: The Kappa Statistic,” *Computational Linguistics*, Vol.22, No.2, pp.249-254, 1996.

[35] J. L. Fleiss, B. Levin, and M. C. Paik, “Statical methods for rates and propositions 3rd Edition,” John Wiley & Sons, Inc., pp.598-626, 2003.

천 민 아

e-mail : minah0218@kmou.ac.kr

2014년 한국해양대학교 IT공학부(학사)

2016년 한국해양대학교 컴퓨터공학과 (석사)

2016년～현 재 한국해양대학교 컴퓨터공학과 박사과정

관심분야 : Machine Learning, Natural Language Processing, Automated Scoring

김 창 현

e-mail : chkim@etri.re.kr

1993년 한국과학기술원 전산학과(석사)

1996년 한국과학기술원 전산학과 (박사수료)

2001년～현 재 한국전자통신연구원 언어처리연구실 책임연구원

관심분야 : Machine Translation

김 재 훈

e-mail : jhoon@kmou.ac.kr

1986년 계명대학교 전자계산(학사)

1988년 한국과학기술원 전산학과(석사)

1996년 한국과학기술원 전산학과(박사)

1997년～현 재 한국해양대학교 IT공학부 교수

관심분야 : 자연언어처리, 기계학습, 자동채점, 정보추출

노 은 희

e-mail : noro@kice.re.kr

1991년 홍익대학교 국어교육(학사)

1993년 서울대학교 국어교육(석사)

1999년 서울대학교 국어교육(박사)

현 재 한국교육과정평가원 연구위원

관심분야 : Automated Scoring, 국어교육, 학업성취도 평가, 서답형 문항

성 경 희

e-mail : kelly9147@kice.re.kr
2000년 이화여자대학교 사회생활학과(학사)
2006년 서울대학교 사회교육과(석사)
2012년 서울대학교 사회교육과(박사)
2013년～현 재 한국교육과정평가원 부연구위원

관심분야 : 사회과교육, 사회과 평가, 교수매체 및 교실수업 연구, 한국어 자동채점

송 미 영

e-mail : mysong@kice.re.kr
1992년 이화여자대학교 수학교육과(학사)
1994년 이화여자대학교 교육학과(석사)
2001년 이화여자대학교 교육학과(박사)
2005년～현 재 한국교육과정평가원 연구위원

관심분야 : Test Equating, Test Theory, Differential Item Function, Large-Scale Assessment