인지과학, 제26권 제2호
Korean Journal of Cognitive Science
2015, Vol. 26, No. 2, 147～165.

준지도학습 방법을 이용한 한국어 서답형 문항 반자동 채점

천민아 서형원 김재훈† 노은희 성경희 임은영

한국해양대학교 컴퓨터공학과 한국교육과정평가원

서답형 문항은 학생들의 종합적인 사고력을 평가할 수 있다는 장점이 있으나, 채점 비용이 많이 들고 채점자의 주관이 개입될 수 있다는 단점이 있다. 이런 단점을 개선하기 위해 영어권에서는 자동채점 시스템을 개발하여 사용하고 있으나, 한국어의 경우에는 아직 여전히 연구 단계에 있다. 본 논문에서는 준지도학습 방법을 이용한 한국어 서답형 문항의 채점 시스템을 제안한다. 제안된 시스템은 모범답안을 초기 모델로 학생답안의 일부를 채점하고 그 결과를 이용해서 점진적으로 학생답안의 채점을 늘려가는 준지도학습 방법을 이용한다. 제안된 시스템을 평가하기 위해서 2013학년도 학업성취도 평가의 국어 및 사회 과목의 서답형 문항을 사용했다. 채점 시간과 일관성에 관해서 매우 좋은 결과를 얻었다. 그 결과 채점 시간을 크게 단출할 수 있었으며 다양한 채점 방법을 적용하여 객관성을 확보한다면 현장에서 바로 적용할 수 있을 것으로 기대된다.

주요어 : 자동채점, 기계학습, 준지도 학습

† 교신저자: 김재훈, 한국해양대학교 컴퓨터공학과, 연구분야: 한국어정보처리
E-mail: jhoon@kmou.ac.kr

서 론

선택형(객관식) 문항은 대학수학능력시험과 같이 대규모 시험에서 주로 사용되고 있다. 선택형 문항은 채점 시간이 매우 짧고 채점의 일관성도 매우 쉽게 유지할 수 있으나, 학생에 따라 문항을 잘못 해석할 수 있으며 학생들의 종합적인 사고 능력을 측정하기 매우 어렵다[1-2]. 이런 이유로 최근 많은 시험에서 서답형 문항을 출제하고 있다. 서답형 문항은 채점 비용이 너무 많이 들고, 채점 시간도 많이 소요될 뿐 아니라 채점의 일관성과 신뢰성도 확보하기 어렵다는 단점이 있다[2]. 이러한 문제를 완화하기 위해서 다양한 형태의 자동채점 시스템[2-5]이 활용되고 있다. 영어의 경우에는 ETS[1]를 비롯한 많은 연구 기관에서 매우 활발하게 연구들이 진행되고 실용 단계에 있으나 한국어의 경우에는 KICE[2]에서 기초연구 단계로 진행되고 있다[3-7].

본 논문에서는 한국어 서답형 문항에 대한 반자동채점 시스템을 제안한다. 제안된 시스템은 크게 분석 단계와 채점 단계로 구성되어 있다. 분석 단계는 학생답안을 입력으로 받아 문장에 포함된 다양한 형태의 언어 정보를 분석하며, 문서정규화, 형태소분석, 품사부착, 구묶음, 바꿔쓰기, 의존구조분석으로 구성되어 있다. 채점 단계는 반복적으로 학생답안을 채점하며, 크게 네 가지 과정으로 구성된다. 1) 분석된 학생답안을 빈도순으로 정렬하고 고빈도 학생답안에 대해서 인간채점자(human rater)가 직접 채점하고 이 결과를 모범답안으로 간주한다. 이하에서는 모범답안을 학습말뭉치라고 부르기도 하며 특별한 혼란이 없을 경우에는 서로 섞어서 사용되기도 한다. 2) 학습말뭉치를 이용해서 미채점답안(unscored answer)들을 다시 채점한다. 3) 새롭게 채점된 답안 중에 신뢰도가 높은 답안은 채점자가 확인 후 학습말뭉치에 추가한다. 4) 2)와 3)의 과정을 미채점답안이 존재하지 않을 때까지 계속 반복한다. 한국어 서답형 문항 반자동채점 시스템에서는 위에서 설명한 일련의 작업을 효과적으로 수행할 수 있고 채점자가 채점에만 집중할 수 있도록 사용자 인터페이스를 제공한다. 제안된 시스템은 11개 문항에 대해서 각 문항 당 1000개의 학생답안을 채점했을 때 평균 채점 시간은 약 23분이였고 정답일치율이 95.6%

1) Educational Testing Service, http://www.ets.org/

2) 교육과정평가원(Korea Institute for Curriculum and Evaluation), http://www.kice.re.kr/

로 매우 좋은 결과를 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 서답형 문항 반자동채점 시스템 개발을 위해 필요한 관련 연구들을 소개한다. 3장에서는 구현한 서답형 문항 반자동채점 시스템에 대해 설명한다. 4장에서는 구현된 반자동채점 시스템을 평가하고 결과를 분석한다. 마지막으로 5장에서는 결론 및 향후 연구에 대해서 논의한다.

관련 연구

자동채점 시스템

자동채점 시스템은 GMAT(Graduate Management Admission Test), TOFEL(Test of English as a Foreign Language), GRE(Graduate Record Examinations) 등의 대규모 시험에서 채점 보조 수단으로 널리 사용되고 있다. GMAT는 1999년부터 ETS에서 개발한 E-rater[8]를 이용하여 에세이 자동채점을 실시하였고, 2006년부터는 Vantage Learning 사에서 개발한 에세이 자동채점 프로그램인 IntelliMetric[9]을 사용하고 있다. TOEFL과 GRE 시험에서는 2006년부터 인간채점자를 보완하는 수단으로 ETS에서 개발한 E-rater와 SpeechRater이 사용되고 있다[10]. 이 밖에도 영어 자동채점 프로그램으로 PEG(Project Essay Grade)[3], IEA(Intelligent Essay Assessor)[4], E-rater(electronic essay rater)[5], IntelliMetric[6]과 MY Access![7], BETSY(Bayesian Essay Test Scoring System)[8] 등이 있다. 이들 시스템의 대부분은 논문형 문항을 채점하는 것이다. 한국어의 경우에서도 논술형 문항에 대한 자동채점 시스템은 연구되지 않았고, 수답형 문항에 대해서는 2000년대 초반부터 실험실 수준의 자동채점에 대한 연구[11-14]를 수행하고 있다.

3) http://www.writingplanet.net/news2/

4) http://kt.pearsonassessments.com

5) https://www.ets.org/

6) http://www.mccanntesting.com/products-services/intellimetric/obv

7) https://www.myaccess.com/

8) http://edres.org/betsy/

그러나 반자동 방법으로 진행된 연구는 한국교육과정평가원을 중심으로 한국어 서답형 문항에 대한 자동채점에 대한 연구[15] 외에는 거의 없었다.

준지도학습(semi-supervised learning)

기계학습 방법은 크게 지도학습(supervised learning)와 비지도학습(unsupervised learning)으로 나눌 수 있다[16]. 지도학습은 정답이 부착된 데이터를 학습 말뭉치로 사용하여 분류기(classifier)를 학습하고 정답이 부착되지 않은(학습되지 않은) 데이터를 입력 받아서 정답을 부여하는 기계학습 방법이다. 지도학습 방법은 비교적 좋은 성능을 보이지만 학습 말뭉치를 구축하기 위해 전문가의 도움이 필요하고 이 과정에서 많은 시간과 노력이 소요된다. 비지도학습은 지도학습 방법과 달리 정답이 부착되지 않은 자료를 사용하여 그 자료의 구조나 관계를 파악하여 패턴을 분류하는 방법이다. 최근에는 이 두 가지 학습 방법을 결합한 준지도학습(semi-supervised learning)이 다양한 분야에서 사용된다[17]. 준지도학습은 정답이 부착된 자료와 부착되지 않은 자료를 모두 학습말뭉치로 사용하는 방법이다. 준지도학습은 크게 상호학습(co-training)과 자가학습(self-training)으로 나눌 수 있다. 반자동채점에 적용할 방식은 자가학습이다. 자가학습은 정답이 부착된 학습말뭉치를 이용해서 분류기를 개발하고, 개발된 분류기를 사용하여 정답이 부착되지 않은 자료에 대해 정답을 부착한다. 부착된 정답의 신뢰도가 높을 경우 학습말뭉치에 추가하여 학습말뭉치의 크기를 점진적으로 확장하여 분류기의 성능을 개선하는 방법이다.[18-19].

한국어 서답형 문항 자동채점 시스템

시스템 구조

한국어 서답형 문항 반자동채점 시스템의 구조도는 그림 1과 같다. 한국어 서답형 문항 반자동채점 시스템은 크게 분석 단계와 채점 단계로 나뉜다.

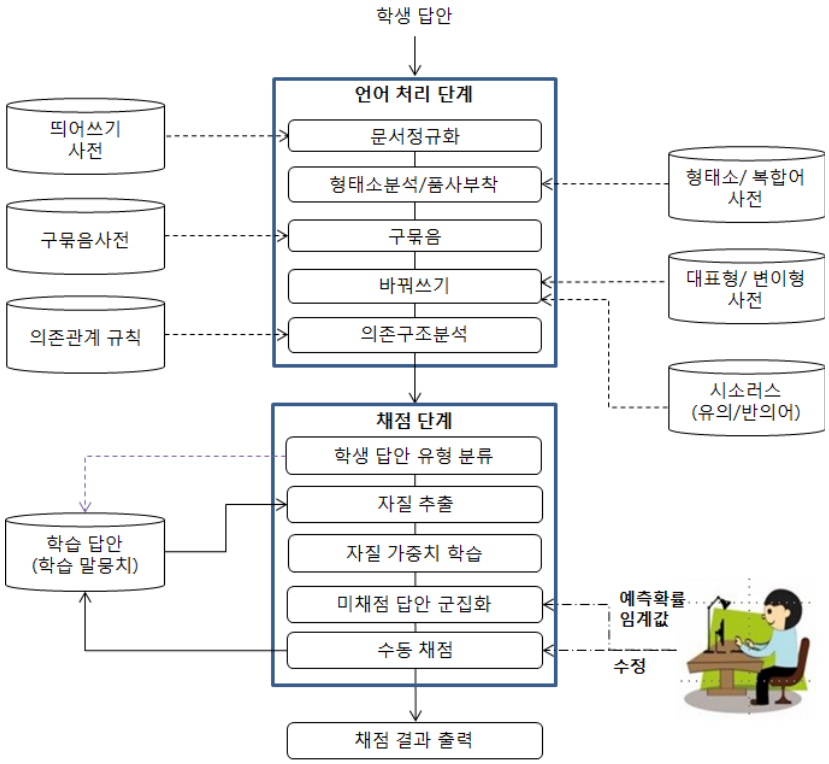

그림 1. 한국어 서답형 반자동채점 시스템의 구조도

분석 단계는 학생답을 입력으로 받아 자동채점에 필요한 언어 정보를 분석하는 단계로서 많은 언어 자원이 필요하다. 분석 단계가 끝난 답안들 중, 분석 결과가 같은 답안들을 묶어 고빈도 순으로 정렬한 뒤 상위 몇 개의 답안에 대해 채점하여 모범답안을 구축한다. 이 모범답안이 초기 학습말뭉치가 된다. 채점 단계는 학생답안과 모범답안의 유사도를 계산하여 점수 단위로 군집화(clustering)하여 각 군집에 점수를 부여하는 단계이며 반복적으로 수행된다. 채점 단계가 끝난 시점에서 채점이 완료된 답안들이 다시 학습 말뭉치가 되고, 채점 되지 않은 답안들이 테스트 답안이 된다. 매 채점 단계마다 해당 시점의 학습말뭉치 전체를 다시 학습한다. 채점 과정이 반복될수록 채점 답안의 신뢰성이 점점 높아지고 미채점된 학생답안 수는 점점 감소하여 모든 학생들의 답안을 채점할 수 있게 된다. 각 단계의 자세한

내용은 이하의 절에서 자세히 설명할 것이다.

분석 단계

이 절에서는 제안된 시스템의 분석 단계에 대해서 설명한다. 분석 단계는 언어정보를 분석하는 단계로서 문서정규화, 형태소분석 및 품사부착, 구묶음, 바꿔쓰기, 의존구조분석으로 구성되어 있다. 분석된 언어정보는 채점 단계의 자질로 사용되기 때문에 가능한 한 정확한 정보를 분석해야 하므로 형태소분석 사전, 의존관계 규칙 등 많은 언어 자원이 요구된다(그림 1 참조).

문서정규화

문서정규화(text normalization)는 문장부호 제거, 띄어쓰기 교정과 철자교정으로 이루어져있다. 문장부호 제거는 정규표현식(regular expression)을 통해 간단하게 구현한다. 띄어쓰기 교정(spacing correction)은 세종말뭉치[20]를 띄어쓰기 사전으로 사용하여 기계학습 방법 중 최대 엔트로피 모델(maximum entropy model)을 통해 구현한다. 철자교정(spelling correction)은 최소편집거리(minimum edit distance) 알고리즘을 이용하여 구현한다.

형태소분석 및 품사부착

형태소분석(morphological analysis)은 CYK 알고리즘[21]을 수정하여 사용한다. 이 알고리즘을 이용하면 어미의 활용이 존재하지 않더라도, 사전 정보만으로 가능한 모든 형태소 분석 결과를 찾아낼 수 있다는 장점이 있다. 그러나 한국어의 경우에는 “아름다운”과 같이 동사의 원형과 어미가 결합할 때(예: “아름답 + ㄴ”) 다양한 형태로 철자의 변형이 발생한다. 이런 문제의 경우에는 CYK 알고리즘으로 처리할 수 없기 때문에 용언(동사와 형용사)에 대해서는 부분적인 기분석 사전을 활용하여 해결한다.

품사부착(part-of-speech tagging)은 형태소분석 결과를 이용해서 격자구조(lattice structure)를 구성하고 구성된 격자 위에 세종말뭉치에서 구한 문맥확률(contextural probability)과 어휘확률(lexical probability)을 적재하여 가중치 네트워크(weighted

network)를 만든다. 가중치 네트워크에서 가장 적절한 경로(longest path)를 찾은 뒤, 그 결과에 품사를 부착한다[22].

구묶음

구묶음(chunking)은 띄어쓰기에서 사용했던 기계학습 방법인 최대 엔트로피 모델과 구묶음 사전을 이용하여 구현한다. 구묶음을 수행하기 위해서는 형태소분석 및 품사부착 과정이 반드시 선행되어야 한다. 구묶음 결과는 '발 없는 말이 천 리 간다'와 같은 속담을 판단하는 데 쓰이거나, 의존구조분석을 위한 자질로 사용될 수 있다.

바꿔쓰기

바꿔쓰기(paraphrasing)는 단어 치환(word replacement)을 사용하여 조사를 대표조사로, 어미를 대표어미로, 단어를 동의어로 치환하도록 구현한다. 대표형/변이형 사전(지식 기반)에 대한 해당 정보가 포함되어 있지 않으면 입력된 단어를 그대로 출력한다. 따라서 대표형/변이형 사전이 완전히 비어 있다면 입력과 출력은 똑같으며 이 사전은 문제의 영역이나 과목 등에 따라 다르므로 문제를 채점할 때 이점을 고려하여 대표형 사전을 보완하거나 수정해야 한다. 바꿔쓰기를 하기 위해서도 형태소분석 및 품사부착 과정이 반드시 선행되어야 한다. 이 기능을 유의어를 확장하여 학생 답안의 채점에 용이하게 사용할 수 있다.

의존구조 분석

의존구조분석(dependency parsing)은 의존관계 규칙(의존문법)을 기반으로 구문분석을 수행한다. 의존관계 규칙은 의존소(dependent)와 지배소(governor)의 관계를 문법으로 표현한 것이다. 지배소는 의존관계에 있는 언어요소들 중 의미의 중심이 되는 요소이며, 의존소는 지배소가 갖는 의미를 보완해주는 요소이다. 이 기능은 스웨덴의 Växjö 대학교와 Uppsala 대학교에서 공동으로 개발한 의존구문 분석기인 MaltParser를 사용하여 구현한다[23].

채점 단계

이 절에서는 채점 단계에서 수행하는 기능들과 반자동채점 알고리즘에 대해 설명한다. 채점 단계는 학생답안 유형 분류, 자질추출, 자질 가중치 학습, 미채점답안 군집화, 수동 채점으로 구성되며 자질추출, 자질 가중치 학습, 미채점답안 군집화, 수동 채점은 반복적으로 수행하면서 학습말뭉치(초기에는 모범답안만 학습말뭉치에 포함됨)를 확장시킨다. 자질추출 및 가중치 학습 계는 해당 시점까지 확장된 학습말뭉치 전체를 이용한다. 이와 같은 과정을 반복해서 모든 학생답안이 학습말뭉치에 포함되면 채점이 완료된다.

학생답안 유형 분류

학생답안 유형 분류는 분석 결과가 완전히 일치하는 학생답안들의 집합을 하나의 군집으로 간주하고 각 군집에 속한 답안 수의 빈도순으로 정렬한 뒤, 인간 채점자가 고빈도 답안의 군집에 대해 수동으로 채점한다. 어느 정도의 답안을 채점할 것인지는 인간채점자가 임의로 결정한다. 이렇게 채점된 답안들을 모범답안(초기 학습말뭉치)로 간주한다.

자질추출 및 가중치 학습

자질추출은해당 시점까지 확장된 학습말뭉치에서 단어 자질, 어절 자질, 구문 자질을 추출한다. 단어 자질은 내용어만 추출하고, 어절 자질은 내용어와 정규화된 기능어(예를 들면, 에게/한테→에게)를 자질로 추출한다. 구문 자질은 의존어와 지배어 그리고 의존관계를 자질로 추출한다. 자질 가중치는 정보검색에서 널리 사용되는 $TF-IDF$를 사용한다[24]. 본 논문에서 하나의 학생답안을 하나의 문서로 간주한다. 내용어는 조사, 어미, 대명사를 제외한 품사 정보를 통해 추출한다.

미채점답안 군집화

그림 2는 미채점답안의 군집화 알고리즘을 기술한다. 본 논문에서 제안한 군집화 알고리즘은 분류 확률과 분류 결과를 이용한다. 즉 분류 확률이 임계값(threshold)보다 큰 각 분류(class)를 하나의 군집으로 간주한다. 또한 분류의 정확률

을 높이기 위해서 그림 2의 일곱 번째 단계와 같이 두 분류기의 결과가 같은 경우에만 정답으로 간주한다.

```
1. 채점이 완료된 답안의 자질을 행렬(trainX)로 변환한다.
2. // trainX를 이용해서 로그 회귀분석 분류기를 학습한다.
   clf1 = LogisticRegression(trainX, trainLabels)
3. // trainX를 이용해서 kNN 분류기를 학습한다.
   clf2 = KNeighborsClassifier(trainX, trainLabels)
4. 미채점답안의 자질을 행렬(testX)로 변환한다.
5. // 미채점답안의 분류 확률을 계산한다.
   testProb = clf1.predict_proba(testX)
6. // 이미 학습된 두 분류기 clf1과 clf2를 이용해서 미채점답안을 채점한다.
   testy1 = clf1.predict(testX)
   testy2 = clf2.predict(testX)
7. // 미채점답안을 군집화한다.
   nCorpus, nLabels = [], []
   for i in range(len(testy1)):
      if testProb[i][int(testy1[i])] > threshold and testy1[i] == testy2[i]:
         nCorpus.append(testCorpus[i])
         nLabels.append(testLabels[i])
8. return nCorpus, nLabels
```

그림 2. 미채점답안 군집화 알고리즘

분류 확률을 계산하기 위해서는 로지스틱 회귀분석(Logistic Regression) 모델[25]을 사용하고, 학생답안의 분류를 위해서는 로지스틱 회귀분석 분류기와 k-NN (k-Nearest neighbors) 분류기[16]를 사용한다. 로지스틱 회귀분석 분류기는 학습 시간을 절약하기 위해 확률 계산과 분류의 목적으로 사용된다.

그림 2의 미채점답안 군집화 알고리즘에서 사용된 변수의 의미는 다음과 같다. trainX는 분류기를 학습하는데 사용하는 데이터로 해당 시점에서 채점이 완료된 답안을 저장하고 있는 변수다. trainLabels는 trainX에 저장되어 있는 각 답안의 점수

데이터이다. clf1 변수는 trainX와 trainLabels를 이용해서 생성한 로그 회귀분석 분류기를 의미하고, clf2 변수는 k-NN 분류기이다. testX는 채점이 되지 않은 답안의 자질을 행렬 형태로 저장하는 변수이다. testProb 변수는 로지스틱 회귀분석 모델을 사용하여 계산한 분류 확률을 저장한다. testy1 변수와 testy2 변수는 각각 분류기와 testX의 정보를 점수 별로 분류한 결과이다. nCorpus 변수와 nLabels 변수는 분류 확률이 임계값보다 크고 두 분류기의 결과가 일치하는 답안과 그 점수를 각각 저장한다.

이렇게 의미적으로 유사한 학생답안들을 묶은 결과가 인터페이스를 통해서 인간채점자에게 제공되므로, 인간채점자가 채점기준의 일관성을 유지하는데 도움이 된다. 인간채점자가 군집 결과의 확인한 결과는 학습말뭉치에 추가되고, 이 학습말뭉치를 이용해서 채점 단계를 다시 수행한다. 이와 같은 과정을 반복하면 학습말뭉치의 양은 증가하고 분류 확률의 신뢰성과 분류 정확률이 증가한다.

시스템 평가 및 분석

평가에 사용된 서답형 문항 및 정답

평가에 사용된 서답형 문항은 2013년에 실시된 “국가 수준 학업성취도 평가”의 국어 과목(중3, 고2)과 사회 과목(중3)에서 선택했다[26]. 각 문항마다 추출한 1,000개의 학생답안[9]과 출제자가 제안한 모범답안으로 구성된 자료를 통해 시스템의 성능을 평가했다. 각 문항의 모범답안 및 배점 분포는 표 1과 같다.

한국어 서답형 반자동채점 시스템 성능 평가

본 논문에서 제안된 반자동채점 시스템을 평가하는데 사용한 점수들로는 인간채점점수, 반자동채점점수, 기준점수가 있다. 인간채점점수는 총 3라운드로 진행되

9) 각 과목 당 1,000개의 학생답안을 표본하여 교육과정평가원에서 직접 입력한 것이다.

표 1. 2013학년도 학업성취도 평가에 포함된 서답형 문항에 대한 모범답안 및 배점분포

문항 번호 \ 구분		모범답안	배점 분포
중3 국어	2-(1)	여학생들이 참여할 만한 종목이 없다.	0, 1, 2
	2-(2)	모두가 즐길 수 있는 체육 대회(학교 행사)가 될 것이다.	0, 1, 2
	4-(2)	사소한 능력이라도 마음만 있으면 할 수 있대.	0, 1, 2, 3
고2 국어	2-(1)	네가 그 과제를 못하겠니?	0, 1
	2-(2)	너는 (그) 과제를 (혼자) 할 수 있다.	0, 1, 2
	4-(1)	B는 예시를 통해 주지 A를 뒷받침한다.	0, 1, 2
	5-(2)	우리 (모두) 천 원의 기적 운동에 참여하자.	0, 1, 2
	6-(1)	동물쇼는 폐지되어야 합니까(?)	0, 1
	6-(2)	동물쇼에 동원된 동물은 야생 상태의 모습이 아니므로, 자연 생태 교육의 효과가 없다.	0, 1, 2, 3
중3 사회	4-(3)	왕권은 신이 부여한 것이다.	0, 1
	8	헌법재판소에 헌법소원(헌법소원심판)을 청구하세요.	0, 1, 2, 3

며, 각 라운드는 두 명의 채점자가 같은 문항의 답안을 채점하여 점수가 일치하는지 확인하고, 해당 문항의 답안이 일치하지 않으면 다음 라운드로 넘겨서 확정한 점수이다. 반자동채점점수는 제안된 서답형 문항 반자동채점 시스템으로 채점한 점수이다. 기준점수는 인간채점점수를 최종적으로 교과전문가가 확인하고 오류가 있을 경우에 이를 수정하여 최종적으로 확정한 점수이며 이 점수를 통해 채점의 정확성을 확인한다. 표 2는 제안된 반자동채점 시스템으로 학생답안을 채점한 결과이다.

분석 단계의 성능 평가

표 2에서 세 번째 열의 답안 유형 수는 입력된 학생답안의 군집 수이며 반자동채점 시스템을 사용하지 않았을 때 인간 채점자가 채점해야 하는 답안의 유형 수이다. 답안의 유형 수가 많을수록 채점이 복잡해지므로, 답안의 유형수를 채점 복

표 2. 국어(중3, 고2) 및 사회(중3) 과목의 서답형 문항에 대한 반자동채점 시스템의 채점 정보 (N=1,000)

과목	문항 번호	답안 유형 수		초기 학습 말뭉치 크기		반복 횟수	소요 시간 (분)	불일치 답안 수	일치율 (%)
		분석 단계 전	분석 단계 후	채점 답안 수	답안 유형 수				
중3 국어	2-(1)	792	580	386	18	10	15	27	97.3
	2-(2)	787	507	503	21	20	25	39	96.1
	4-(2)	865	771	238	17	13	32	80	92.0
고2 국어	2-(1)	65	56	964	18	4	4	0	100.0
	2-(2)	467	292	386	18	20	15	37	96.3
	4-(1)	553	513	491	17	31	35	85	91.5
	5-(2)	635	558	503	26	23	40	77	92.3
	6-(1)	451	338	716	29	14	21	28	97.2
	6-(2)	544	527	543	34	14	35	29	97.1
중3 사회	4-(3)	301	252	750	21	13	11	10	99.0
	8	367	312	669	18	23	16	22	97.8
평균	-	530	428	559	22	17	23	40	95.6

잡도와 동일시 할 수 있다. 예를 들면 문항 “중3 국어 2-(1)”의 채점 복잡도는 79.2%이고, 문항 “고2 국어 2-(1)”의 채점 복잡도는 6.5%이므로 전자가 후자보다 채점이 어렵다는 것을 의미한다. 네 번째 열의 답안의 유형 수는 분석 단계를 거친 후, 학생답안들의 군집 수이며 분석 단계만 수행하더라도 많은 유형 수가 줄어들고 있음을 확인할 수 있다. 평균적으로 102개의 유형이 줄어들었으며 19.2%가 감소되었다. 따라서 분석 단계가 답안의 유형수를 줄여주는데 유용하다고 판단할 수 있다.

초기 학습 말뭉치의 크기

표 2에서 다섯 번째 열은 ‘학생답안 유형 분류’ 과정에서 고빈도 답안에 대해

인간채점자가 수동으로 채점한 결과이며 채점 결과는 초기 학습 말뭉치가 된다. 초기 학습 말뭉치의 크기는 평균 559 답안으로 학생답안의 약 56%가 고빈도 답안으로 채점이 가능하다. 이것은 평균 22개의 답안을 채점함으로써 얻어진 결과이다. 분석 단계를 거친 후 22개의 고빈도 학생답안을 채점하면 학생답안의 약 56%가 채점이 완료된다는 것이다.

반복 횟수 및 소요시간 분석

표 2에서 일곱 번째 열은 채점이 완료되기 까지 채점 단계의 반복 횟수를 나타내고 있다. 평균적으로 17번 반복으로 채점이 완료되었다. 이 반복 횟수는 채점 복잡도와 정비례하지는 않는다. 왜냐하면 '미채점답안 군집화 알고리즘'에서 계산되는 분류 확률로 얼마나 많은 미채점답안이 채점답안으로 제시하느냐에 따라 달라진다. 실험에 사용된 분류 확률의 임계값(threshold)은 0.99이며 이는 횟수를 반복하면서 0.03씩 감소하여 사용하고 있다. 즉 첫 번째 임계값은 0.99이고 두 번째 반복에서 사용된 임계값은 0.96이다. 모든 채점을 완료하는데 걸린 시간[10]은 평균 23분이다. 이 시간은 분석 단계의 수행 시간은 제외된 것이며, 채점 단계의 총 수행 시간과는 다소 거리가 있다. 수동 채점 과정에서 인간 채점자가 채점 답안에 오류가 없는지를 판단하는 시간도 함께 포함되어 있기 때문이다. 라운드 방식에서 하나의 문항을 채점할 때 걸리는 시간이 사흘이라는 점과 비교해봤을 때, 이 방식을 적용하면 시간 비용을 크게 절감할 수 있다[15].

일치도 분석

표 2에서 아홉 번째 열은 기준점수와의 반자동채점 결과[11]가 불일치한 답안 수이다. 평균 40개의 답안이 불일치했으며 역으로 말하면 960개의 답안은 기준점수와 일치했다는 것이다. 즉 기준점수와의 일치율은 평균 95.6%로 높은 수치를 보였으나 여전히 약 4%의 오류가 있다. 이를 개선하기 위해서는 일차적으로는 반자동

10) 걸린 시간 = 고빈도 답안 채점 시작에서부터 채점되지 않은 답안의 수가 0이 될 때까지의 소요 시간

11) 참고로 반자동채점은 한 사람의 인간 채점자가 반자동채점 시스템으로 채점한 결과이다.

채점 시스템의 성능이 개선되어야 할 것이다. 또 인간 채점자들과 같이 여러 명이 같은 문항을 채점하거나 여러 라운드를 통해서 채점함으로써 일치율을 개선할 수 있을 것이다.

각 문항 채점 결과의 일치도 및 불일치율

표 3은 각 문항의 채점 결과의 일치도와 불일치율을 정리해서 나타낸 것이다.

표 3. 각 문항 채점 결과의 일치도 및 불일치율

-		피어슨 상관계수		Kappa계수		불일치율(%)	
과목	문항 번호	인간채점과 기준점수	반자동채점과 기준점수	인간채점과 기준점수	반자동채점과 기준점수	인간채점과 기준점수	반자동채점과 기준점수
중3 국어	2-(1)	0.96	0.82	0.90	0.80	1.4	2.7
	2-(2)	0.97	0.93	0.91	0.87	2.5	3.9
	4-(2)	0.97	0.93	0.93	0.81	3.1	8.0
고2 국어	2-(1)	0.99	1.00	0.99	1.00	0.5	0.0
	2-(2)	0.98	0.87	0.98	0.87	0.5	3.7
	4-(1)	0.99	0.88	0.97	0.83	1.4	8.5
	5-(2)	0.99	0.93	0.99	0.88	0.9	7.7
	6-(1)	0.98	0.94	0.98	0.94	1.1	2.8
	6-(2)	1.00	0.90	0.98	0.84	1.1	7.6
중3 사회	4-(3)	0.86	0.95	0.85	0.95	3.2	1.0
	8	1.00	0.92	0.99	0.93	0.2	2.2
평균 (표준편차)		0.97 (0.04)	0.92 (0.05)	0.95 (0.05)	0.88 (0.06)	1.45 (1.04)	4.37 (3.05)

일치도 분석에는 다양한 분석 방법이 있으나 본 논문에서는 피어슨 상관계수와 Kappa 계수를 분석하였다. 반자동채점과 기준점수의 일치도 중 피어슨 상관계수 평균은 0.92로 강한 양의 선형관계를 나타냈다. 따라서 반자동채점 점수와 기준점수가 의비하는 바가 유사하다고 판단할 수 있다. Kappa 계수 역시 평균 0.88로 반자동채점과 기준점수가 의미하는 바가 거의 일치한다고 판단할 수 있다.

인간채점과 기준점수의 비교를 A, 반자동채점과 기준점수의 비교를 B라고 한다면 B가 A에 비해 일치도는 상대적으로 낮으며 불일치율이 높은 것을 확인할 수 있다. 그 원인을 분석하기 위해서 불일치율이 평균 불일치율보다 높은 문항들에 대해서 살펴보면 원인은 크게 2가지로 생각할 수 있다. 첫 번째로 배점 분포가 많은 문항의 경우 불일치율이 높게 나왔다. 각 점수에 대한 학습 말뭉치의 양이 부족 상대적으로 부족하게 되므로 불일치율이 높게 나온 것으로 추측할 수 있다. 두 번째로 채점 기준이 다른 문항들에 비해 복잡한 문항들이 불일치율이 높았다. 예를 들어 고2 국어 과목의 4-(1) 문항의 경우에 모범답안으로 “B는 예시를 통해 주지 A를 뒷받침한다.”외에도 2점으로 인정될 수 있는 유형의 수가 44가지 보다 많았다. 평균 불일치율보다 불일치율이 높은 다른 문항들의 경우에도 기준 답안 유형수의 평균값보다 많아 상대적으로 복잡한 문항이다.

그 외에도 인간채점점수의 경우 3라운드에 거쳐 확정한 답안이므로 반자동채점 시스템도 라운드 방식을 적용하여 3라운드까지 채점한 결과로 다시 분석할 경우 현재 성능보다 훨씬 좋은 성능을 낼 것으로 예상된다.

결론 및 향후 연구

본 논문에서는 학생들의 종합적인 이해능력 및 사고능력을 판단하는데 적합한 서답형 문항을 채점할 때 발생하는 문제점을 개선하기 위한 반자동채점 시스템을 제안하였다. 먼저 언어처리 기법을 이용하여 학생들의 답안을 분석 및 처리한 뒤, 이 결과가 일치하는 답안들을 묶어 고빈도 순으로 정렬하여 채점자가 채점할 수 있게 했다. 이렇게 채점된 결과에서 자질을 추출하여 각 자질들의 가중치를 학습한 뒤, 미채점답안들에 대해 로지스틱 회귀분석과 k-NN 알고리즘을 이용하여 점수별로 답안을 분류하는 기법을 사용하여 반자동채점 시스템을 구현했다. 2013년에 실시된 “국가 수준 학업성취도 평가”의 국어, 사회 과목의 서답형 문항을 선택하여 샘플로 추출된 1000개의 답안에 대해 시스템의 성능을 분석했다. 그 결과, 반자동채점 결과와 기준채점 결과 평균 95.6%의 높은 일치율을 보였다. 피어슨 상관계수와 Kappa 상관계수에서도 각각 0.92와 0.88이라는 유효한 값을 얻을 수

있었다.

향후에는 본 논문에서 제안한 반자동채점 시스템의 성능을 높이기 위해 초기 학습말뭉치의 크기 선정과 각 배점별로 적절한 수의 학습말뭉치의 확보 등을 연구하여 적용할 계획이다. 채점자의 편의성을 위한 개선방향으로 채점할 때 사용자가 직접 유의어 사전을 등록하고 편집할 수 있도록 하는 인터페이스의 추가와 채점에 필요한 키워드를 제공하는 방식을 적용할 예정이다.

감사의 글

본 연구는 한국교육과정평가원의 “한국어 서답형 문항 자동채점 프로그램 개발 및 실용성 검증” 사업과 미래창조과학부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)[10041807, 지식학습 기반의 다국어 확장이 용이한 관광/국제행사 통역률 90%급 자동 통번역 소프트웨어 원천 기술 개발]의 일환으로 수행하였음.

참고문헌

[1] 이양락 외 (2010). 2014학년도 대학수학능력시험체제 개발을 위한 기초 연구, 한국교육과정평가원 연구보고서 대수능 CAT 2010-3.

[2] 진경애 (2007). “영작문 자동 채점 시스템 개발 연구”, **영어어문교육**, 13(1): 236-237.

[3] S. Dikli (2006). “An Overview of Automated Scoring of Essays”, *The Journal of Technology, Learning, and Assessment,* 5(1): 5-35.

[4] E. Jang, S. Kang, E. Noh, M. Kim, K. Sung, and T. Seong, “KASS: Korean Automatic Scoring System for Short-answer Questions”, Proceedings of the 6th International Conference on Computer Supported Education, pp.226-230.

[5] Y. Chen, C. Liu, C. Lee, and T. Chang (2010). “An Unsupervised Automated Essay-Scoring System”, IEEE Intelligent Systems, September/October, 61-67.

[6] C. Leacock and M. CHodorow (2003). "C-rater: Automated Scoring of Short-Answer Questions", *Computers and the Humanities 37*: 389-405.

[7] 박일남, 강승식, 노은희, 김명화, 성태제 (2013). "정답 템플릿 작성 방식에 의한 한국어 서답형 문항 자동채점 시스템", **정보과학회논문지: 컴퓨팅의 실제 및 레터,** 19(12): 630-636.

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

[9] L. M. Rudner, V. Garcia, and C. Welch (2006). "An Evaluation of the IntelliMetric[SM] Essay Scoring System", *The Journal of Technology, Learning, and Assessment, 4*(4).

[10] ETS (2010). ETS Automated Scoring Technologies. ETS Report.

[11] 최동경 (2001). 벡터 유사도와 시소러스를 이용한 주관식 답안의 채점 방법. 동국대학교 교육대학원 석사학위논문.

[12] 박희정, 강원석 (2003). "유의어 사전을 이용한 주관식 문제 채점 시스템 설계 및 구현", **한국컴퓨터교육학회논문지** 6(3): 207-216.

[13] 강원석 (2011). "질의문 유형 분석을 통한 서답형 자동채점 시스템", **한국콘텐츠학회논문지** 11(2): 13-21.

[14] 조우진, 오정석, 이재영, 김유섭 (2005). "의미 커널과 한글 워드넷에 기반한 지능형 채점 시스템", **정보처리학회논문지 A,** 12(6): 539-546.

[15] 노은희 외 (2014). **한국어 서답형 문항 자동채점 프로그램 개발 및 실용성검증,** 한국교육과정평가원, 연구보고서 RRE 2014-6.

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

[17] O. Chapelle, B. Schölkopf, and A. Zien (2006). Semi-supervised learning. The MIT Press, pp.1-8.

[18] A. Sogaard (2013). Semi-Supervised Learning and Domain Adaptation in Natural Language Processing, Morgan & Claypool Publishers.

[19] S. Bergsma (2010). Large-Scale Semi-Supervised Learning for Natural Language Processing, PhD Dissertation, Department of Computing Science, University of Alberta.

[20] 국립국어원 (2011). 21세기 세종계획, 문화체육관광부, http://korean.go.kr/sejong

[21] 김성용 (1987). Tabular parsing 방법과 접속정보를 이용한 한국어 형태소 분석

기, 한국과학기술원 석사 학위논문, pp.21-37.

[22] 김재훈 (1998). "가중치망 모델을 이용한 한국어 품사 태깅", **한국정보과학회 논문지** *25*(6): 951-959.

[23] J. Nivre (2008). "Algorithms for Deterministic Incremental Dependency Parsing", *Computational Linguistics 34*(4): 513-553.

[24] G. Salton and M. J. McGill (1983). Introduction to Modern Information Retrieval, McGraw-Hill, pp.118-120.

[25] G. Casella, S. Fienberg and I. Olkin (2013). An Introduction to Statistical Learning with Applications in R, Springer.

[26] 한국교육과정평가원 (2013). 2013년도 국가 수준 학업 성취도 평가 기출답안 및 정답, http://kice.re.kr/board.do?boardConfigNo=112&menuNo=372

1차원고접수 : 2015. 03. 12
1차심사완료 : 2015. 05. 08
게재확정일 : 2015. 05. 29

(*Abstract*)

Semi-Automatic Scoring for Short Korean Free-Text Responses Using Semi-Supervised Learning

Min-Ah Cheon Hyeong-Won Seo Jae-Hoon Kim
Korea Maritime and Ocean University
Eun-Hee Noh Kyung-Hee Sung EunYoung Lim
Korea Institute for Curriculum and Evaluation

Through short-answer questions, we can reflect the depth of students' understanding and higher-order thinking skills. Scoring for short-answer questions may take long time and may be an issue on consistency of grading. To alleviate such the suffering, automated scoring systems are widely used in Europe and America, but are in the initial stage in research in Korea. In this paper, we propose a semi-automatic scoring system for short Korean free-text responses using semi-supervised learning. First of all, based on the similarity score between students' answers and model answers, the proposed system grades students' answers and the scored answers with high reliability have been included in the model answers through the thorough test. This process repeats until all answers are scored. The proposed system is used experimentally in Korean and social studies in Nationwide Scholastic Achievement Test. We have confirmed that the processing time and the consistency of grades are promisingly improved. Using the system, various assessment methods have got to be developed and comparative studies need to be performed before applying to school fields.

Key words : Automated scoring, Machine learning, Semi-Supervised learning