Journal of The Korea Society of Computer and Information
Vol. 18, No. 8, August 2013
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2013.18.8.017

투표 기반 서술형 주관식 답안 자동 채점 모델의 설계 및 구현

허 정 만*, 박 소 영**

Design and Implementation of an Automatic Scoring Model Using a Voting Method for Descriptive Answers

Jeongman Heo*, So-Young Park**

요 약

본 논문에서는 투표기법을 이용하여 서술형 주관식 문제에 대한 학습자 답안을 자동으로 채점하는 모델을 제안한다. 제안하는 방법은 모델 구축 비용을 줄이기 위해서, 문제 유형별로 세분화하여 서술형 주관식 답안 자동 채점 모델을 따로 구축하지 않는다. 제안하는 방법은 서술형 주관식 답안 자동 채점에 유용한 자질을 추출하기 위해서, 모범 답안과 학습자 답안을 비교한 결과를 바탕으로 다양한 자질을 추출한다. 제안하는 방법은 답안 채점 결과의 신뢰성을 높이기 위해서, 각 학습자 답안을 여러 기계학습 기반 분류기를 이용하여 채점하고, 각 채점 결과를 투표하여 만장일치로 선택한 채점 결과를 최종 채점 결과로 결정한다. 실험결과 기계학습 기반 분류기 C4.5만 사용한 채점 결과는 정확률이 83.00%인데 반해, 기계학습 기반 분류기 C4.5, ME, SVM에서 만장일치로 선택한 채점 결과는 정확률이 90.57%까지 개선되었다.

▸ Keywords : 자동 채점 모델, 기계학습 기반 분류기, 투표 기법

Abstract

TIn this paper, we propose a model automatically scoring a student's answer for a descriptive problem by using a voting method. Considering the model construction cost, the proposed model does not separately construct the automatic scoring model per problem type. In order to utilize features useful for automatically scoring the descriptive answers, the proposed model extracts feature values from the results, generated by comparing the student's answer with the answer sheet. For the purpose of improving the precision of the scoring result, the proposed model collects the scoring results classified by a few machine learning based classifiers, and unanimously selects

•제1저자 : 허정만 •교신저자 : 박소영
•투고일 : 2013. 6. 3, 심사일 : 2013. 7. 26, 게재확정일 : 2013. 8. 2.
* 상명대학교 디지털미디어학부(Dept. of Digital Media, SangMyung University)
** 상명대학교 게임학과(Dept. of Game Design & Development, SangMyung University)
※ 본 연구는 2013년 상명대학교 교내연구비 지원에 의해 수행되었음.

the scoring result as the final result. Experimental results show that the single machine learning based classifier C4.5 takes 83.00% on precision while the proposed model improve the precision up to 90.57% by using three machine learning based classifiers C4.5, ME, and SVM.

▸ Keywords : Automatic Scoring Model, Machine Learning based Classifier, Voting Method

I. 서 론

교육수준이 점점 높아지면서 그에 따라 교육평가의 경향도 변하고 있다[1,2]. 그동안 채점의 효율성을 고려하여 선택형 객관식 문항을 많이 사용했지만, 최근에는 교육평가의 질적 수준이 높은 서술형 주관식 문항이 주목을 받고 있다[3,4]. 서술형 주관식 문항은 자신의 생각을 직접 서술하기 때문에 선택형 객관식 문항보다 분석력과 종합력의 높은 인지 능력을 측정할 수 있다[5-7]. 국가수준의 학업성취도 평가에서 서술형 주관식 문항이 선택형 객관식 문항에 비해 내용 타당도가 높게 나타났으며, 객관식 평가문항이 학습자의 사고능력을 제한한다는 비판은 서술형 주관식 문항의 중요성을 더욱 부각시킨다[8-10].

그러나, 서술형 주관식 문항은 채점 기준이 모호하여 일관성 있게 답안을 채점하기 어렵고, 채점 결과에 대한 신뢰성 확보가 힘들며, 채점에 시간과 노력이 많이 소요된다는 한계가 있다[7,11,12]. 예를 들어, [표1]과 같이 서술형 주관식 답안이 긴 경우 채점 기준을 명확하게 제시하기 쉽지 않고, 채점 시간이 오래 걸릴 수 있다. 이러한 문제를 해결하기 위해 서술형 주관식 답안 자동 채점 모델의 필요성이 대두되었다.

이를 고려하여, 학습자 답안을 벡터화하여 모범 답안과 유사도를 측정하여 자동으로 채점하는 연구[13-16]나 문제유형을 고려하여 문제유형별로 특화하여 자동으로 채점하는 연구[17] 등 다양한 연구가 진행되었다. 자동으로 채점하는 경우, [표1]의 문제 B에서 모범 답안 "~지 않아야 한다."나 정답예제 "~되지 않으며"와 같이 의미를 반대로 만드는 부정어 처리에 유의해야 한다. 또한, 항목별로 나열하여 설명하도록 요구하는 문제 A와 두 가지를 비교하여 설명하도록 요구하는 문제 B는 문제 유형이 다르지만, 모델 구축 비용을 고려하면 두 가지 문제유형을 일반화할 필요가 있다.

따라서, 본 논문에서는 부정어를 고려하고 문제유형을 일반화하여 서술형 주관식 답안을 자동으로 채점하는 모델을 제안한다. 앞으로 2장에서는 서술형 주관식 답안 자동 채점 모델에 관련된 기존 연구에 대해서 살펴본다. 그리고 3장에서는 제안하는 서술형 주관식 답안 자동 채점 모델에 대해 설명하고, 4장에서는 제안하는 모델을 실험 및 평가한다. 그리고 마지막 5장에서는 결론을 맺는다.

표 1. 문제 및 답안 예제
Table 1. Examples of Problems and their Answers

문제 A	GUI를 구성하는 요소 3가지를 설명하고 구체적인 예를 들어 설명하시오.
모범 답안	GUI를 구성하는 세 가지 요소에는 컴포넌트, 이벤트, 이벤트 리스너가 있다. 컴포넌트는 사용자가 프로그램과의 상호작용을 도와주는 화면요소이고, 메뉴, 스크롤, 트리가 있다. 이벤트는 프로그램에 발생하는 모든 이벤트를 정의하는 객체이며, 마우스이벤트, 키 이벤트, 액션 이벤트가 있다. 이벤트 리스너는 대기하다가 이벤트 발생시 응답하는 객체를 말하며, 마우스 리스너, 키 리스너, 액션 리스너가 있다.
정답 예제	컴포넌트는 사용자가 프로그램과 상호 작용할 수 있도록 도와주는 화면요소를 정의하는 객체로 메뉴, 도구, 스크롤, 트리가 있다. 이벤트는 입력장치로 발생하는 모든 이벤트를 포함하는 객체로 마우스, 키, 액션 이벤트가 있다. 이벤트리스너는 이벤트가 발생했을 때 반응하는 객체로 버튼 리스너, 액션 리스너, 마우스 리스너, 키 리스너가 있다.
오답 예제	컴포넌트 - 유저와 데이터의 상호작용 이벤트 - 사건들을 나열 리스너 - 나열한 사건들을 정리
문제 B	오버플로우와 언더플로우를 비교하시오
모범 답안	변수는 자료형에 따라 저장할 수 있는 값의 범위가 제한되어 있고, 그 범위를 벗어나지 않아야 한다. 오버플로우는 자료형에서 허용하는 최대값보다 큰 값을 변수에 저장하여 오류가 발생한 경우이고, 언더플로우는 최소값보다 작은 값을 변수에 저장하여 오류가 발생한 경우이다.
정답 예제	자료형은 표현가능한 수의 범위가 한정되어 있다. 범위의 최대치보다 큰 값을 저장하면 제대로 표현되지 않으며 '오버플로우'라고 하며, 최소치보다 작은 숫자를 저장하면 '언더플로우'라 한다.
오답 예제	오버플로우는 최소값보다 작은 값을 변수에 저장하여 오류가 발생한 경우이고, 언더플로우는 최대값보다 큰 값을 변수에 저장했을 때 용량이 넘쳐 오류가 발생하는 것이다.

II. 관련 연구

서술형 주관식 답안을 자동으로 채점하기 위해서, 다양한 접근방법이 제안되었다. 이들은 크게 벡터 유사도를 고려한 접근방법, n-gram 유사도를 고려한 접근방법, 문제유형을 고려한 접근방법으로 나눌 수 있다. 첫째, 벡터 유사도를 고려한 서술형 주관식 답안 자동 채점 접근방법은 모범 답안과 학습자 답안의 각 단어의 출현 빈도를 각각 벡터화하고 두 단어 벡터의 유사도를 계산하여 정답 여부를 판단한다[13-15]. 따라서, [표1]의 문제 A의 모범 답안과 정답 예제는 동일한 단어를 많이 포함하여, 이 접근방법에서는 두 답안의 유사도가 높게 계산되어 학습자 답안의 정답예제를 정답으로 분류할 수 있다.

그러나, 이 접근방법은 답안의 내용에 포함된 부정어를 올바르게 처리하지 못한다. 예를 들어, 문장 "그 범위를 벗어나지 않아야 한다"와 "그 범위를 벗어날 수 있다"는 반대의 의미이지만 동일한 단어를 많이 포함하고 있으므로, 유사한 문장으로 분류할 수 있다.

둘째, n-gram 유사도를 고려한 서술형 주관식 답안 자동 채점 접근방법은 모범 답안과 학습자 답안을 n개의 어절로 나누고, 이를 벡터화하여 모범 답안과의 유사도를 계산하여 채점한다[16]. 이 접근방법은 여러 어절 열을 고려하므로, 벡터 유사도를 고려한 접근방법과 달리 부정어를 처리할 수 있다. 예를 들어, 세 개 어절을 비교하면, 문장 "벗어나지 않아야 한다"와 "벗어날 수 있다"를 구분할 수 있다. 이 접근방법의 자동 채점 결과는 형태소 분석과 같은 자연어 분석 처리 없이도 수기 채점 결과와 상관관계가 높게 나타났다[16].

그러나, 이 접근방법은 간단하고 짧은 유형의 답안을 잘 분류하는데 반하여, 답안이 길어질수록 성능이 떨어지는 경향이 있다. 예를 들어, [표1]의 문제 B의 모범 답안과 학습자 답안의 정답예제는 동일한 어절열이 많지 않다. 반면에, 문제 B의 모범 답안과 학습자 답안의 오답예제는 "최대값보다 큰 값을 변수에", "값을 변수에 저장하여 오류가 발생한 경우이고", "최소값보다 작은 값을 변수에 저장하여 오류가 발생한"와 같은 동일한 어절열을 포함하고 있다. 따라서, n-gram 유사도를 고려한 서술형 주관식 답안 자동 채점 접근방법은 이러한 경우를 오답예제를 정답으로 잘 못 분류할 수 있다.

셋째, 문제 유형을 고려하는 서술형 주관식 답안 자동 채점 접근방법은 서술형 주관식 문제를 유형별로 구분하고 유형에 따라 채점방식과 기준을 달리하여 채점한다[8,17]. 단어·구 수준의 서답형 자동 채점 접근방법[8]은 문제에서 3단어 이하로 답안을 작성하도록 제한하며, 단순 문자열 일치, 부분 문자열 일치, 구문구조 일치 등으로 답안의 채점기준을 달리한다. 질의문 유형을 고려한 주관식 답안 자동채점 접근방법[17]은 주관식 문제를 단독 과제형, 공통 과제형, 순서 제시형, 설명 제시형, 장단점 제시형, 부분 제시형으로 구분하여 채점한다. 이 접근방법에서는 문제 유형의 특징을 고려할 수 있으므로, 좀 더 신뢰성 있는 채점이 가능하다.

그러나, 이 접근방법은 문제유형에 따라 서술형 주관식 답안 자동 채점 모델을 따로 생성하므로, 앞에서 제시된 유형에 해당하지 않는 문제가 제시될 경우 새로운 적용 기준이 필요하다. 즉, [표1]의 나열 설명 문제 A를 위해 구축한 서술형 주관식 답안 자동 채점 모델을 비교 설명 문제 B에 그대로 적용할 수 없다. 또한, 형태소 분석 정도만 이용하고 있기 때문에 부정문 등으로 인해 의미가 완전히 다른 답안을 정답으로 채점하려면 부가적인 처리가 필요하다.

본 논문에서는 투표기법을 이용한 서술형 주관식 답안 자동 채점 모델을 제안한다. 기존 접근방법이 문제 유형마다 답안 채점 모델을 따로 구축하여 모델 구축부담이 증가한다는 점을 보완하기 위해서, 제안하는 방법은 여러 서술형 주관식 문제를 일반화하고 하나의 답안 자동 채점 모델을 구축한다. 그리고, 부정어가 답안의 의미를 반대로 분석할 수 있다는 점을 고려하여, 제안하는 접근방법은 부정어 관련 자질을 활용한다. 또한, 서술형 주관식 답안 자동 채점 결과의 신뢰성을 높이기 위해서, 제안하는 접근방법은 여러 기계학습 기반 분류기를 사용하여 답안을 채점하고, 다수결 또는 만장일치의 투표방법[18,19]을 바탕으로 최종 채점 결과를 결정한다.

III. 투표기반 서술형 주관식 답안 자동 채점

본 논문에서는 [그림1]과 같이 자질 추출 단계, 답안 분류 단계, 투표 단계로 구성된 투표기반 서술형 주관식 답안 자동 채점 모델을 제안한다. 제안하는 모델은 모범 답안과 학습자 답안을 형태소 분석하고, 그 결과를 비교하여 서술형 주관식 답안 채점에 유용한 자질들을 추출한다. 추출한 자질을 여러 기계학습 기반 분류기에 적용하고, 기계학습 기반 분류기의 결과를 투표하여 최종 결과를 선택한다.

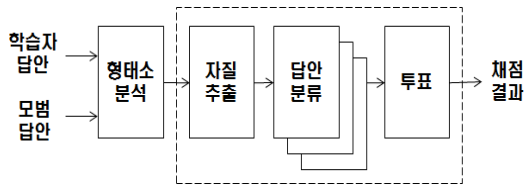

그림 1. 제안하는 모델
Fig. 1. Proposed Model

서술형 주관식 답안 자동 채점에 활용하는 자질은 크게 텍스트 자질, 키워드 자질, 조합 자질로 나눌 수 있다. 첫째, 텍스트 자질은 모범 답안 대비 학습자 답안의 구성비를 의미한다. 이를 위해, 먼저 문제 A의 모범 답안과 학습자 답안에서 문장 개수, 어휘 개수, 음절 개수, 명사 개수, 동사 개수, 접속사 개수, 부사 개수, 조사 개수, 숫자 개수, 영어 단어 개수, 수식 개수, 부정어 개수를 센다. 그리고 모범 답안에 나타난 개수 대비 학습자 답안에 나타난 개수의 비율을 각각 자질로 활용한다. 답안의 길이나 품사별 분포를 나타내는 정보는 정답인지 아닌지 여부를 판단하기에는 미흡하지만, 이러한 정보가 답안 자체의 문서품질을 평가하는 데는 유용할 수 있다[20,21]. 그리고, 부정어는 답안의 의미를 정반대로 바꾸기 때문에 부정어 자질이 필요하다.

예를 들어, [표1]의 문제 B에서 모범 답안은 2개의 문장, 226개의 음절, 18개의 명사 등으로 구성되어 있으며, 중복을 제거하면 어휘의 종류는 총 24개이다. 학습자의 정답 예제 답안은 2개의 문장, 159개의 음절, 10개의 명사, 23개의 어휘 등으로 구성되어 있다. 따라서, 이를 텍스트 자질로 표현하면 문장자질은 1.0(=2/2), 음절자질은 0.7(=159/226), 명사자질은 0.56(=10/18), 어휘자질은 0.96(=23/24) 등으로 표현된다. 반면에, 학습자의 오답 예제 답안은 1개의 문장, 141개의 음절, 14개의 명사, 19 개의 어휘 등으로 구성되어 있다. 이를 텍스트 자질로 표현하면 문장자질은 0.5(=1/2), 음절자질은 0.62 (=141/226), 명사자질은 0.78(=14/18), 어휘자질은 0.79 (=19/24) 등으로 표현된다.

둘째, 키워드 자질은 모범 답안에 포함된 키워드가 학습자 답안에서 어떤 형태로 나타났는지를 표현한다. 이 때, 키워드는 모범 답안에서 중심이 되는 단어들로 구성한다. 키워드 자질은 [표2]와 같이 키워드가 학습자 답안에 포함되어있는지 여부(키워드 유무), 학습자 답안에서 키워드가 차지하는 비율(키워드 비율), 키워드가 모범 답안에 출현한 빈도 대비 학습자 답안에 출현한 빈도의 비율(키워드 출현빈도), 여러 단어로 구성된 키워드의 경우 모범 답안에서 각 키워드가 출현한 빈도의 평균 대비 학습자 답안에서 각 키워드가 출현한 빈도의 평균(키워드 평균), 부정어를 포함한 키워드가 있을 경우 부정키워드가 모범 답안에 출현한 빈도 대비 학습자 답안에 출현한 빈도의 비율(부정키워드), 키워드의 순서가 학습자 답안에서 유지되었는지 바뀌었는지 여부(키워드순서), 여러 키워드가 있는 경우 키워드 사이에 있었던 단어의 수(키워드거리) 등을 포함한다.

표 2. 자질 유형
Table 2. Feature Types

구분	자질
텍스트 자질	문장, 어휘, 음절, 명사, 동사, 접속사, 조사, 부사, 부정어, 숫자, 영어, 수식
키워드 자질	키워드 유무, 답안에서 키워드 비율, 키워드 출현 빈도, 부정 키워드, 키워드 평균, 키워드 순서, 키워드 거리, "(키워드)에서", "(키워드)의", "(키워드)하는데", "(키워드)를", "~에서 (키워드)", "~의 (키워드)", "~하는데 (키워드)", "~를 (키워드)", "(키워드1)에서 (키워드2)", "(키워드1)의 (키워드2)", "(키워드1)하는데 (키워드2)", "(키워드1)을 (키워드2)", 순서 중요도, 부가 순서 중요도
조합 자질	부사 + 숫자, 부정어 + 숫자 + 영어 + 수식, 숫자 + 영어, 숫자 + 키워드 순서, 키워드 거리 + 부가 순서 중요도, 키워드 출현 빈도 + 키워드 평균, 숫자 + "(키워드1)을 (키워드2)", "(키워드1)을 (키워드2)"+"(키워드)의", "(키워드)하는데"+"~에서 (키워드)"

또한, 키워드 자질은 키워드가 답안에서 나타난 패턴을 자질로 포함한다. 즉, 모범 답안과 학습자 답안에서 "(키워드)에서", "(키워드)의 ~", "(키워드)하는데", "(키워드)를", "~에서 (키워드)", "~의 (키워드)", "~하는데 (키워드)", "~를 (키워드)", "(키워드1)에서 (키워드2)", "(키워드1)의 (키워드2)", "(키워드1)하는데 (키워드2)", "(키워드1)를 (키워드2)"의 형태로 나타난 패턴의 출현빈도를 각각 세고, 모범 답안 출현빈도 대비 학습자 답안의 출현빈도의 비율을 각각 자질로 활용한다.

예를 들어, [표1]의 문제 B는 "오버플로우", "언더플로우", "자료형", "범위", "최대", "최소", "값" 등을 키워드로 포함하고 있다. 모범 답안은 패턴 "(키워드1)의 (키워드2)"에 해당하는 "값의 범위"를 포함하고 있고, 학습자 정답 예제는 "범위의 최대"를 포함하고 있다. 따라서, 문제 B의 학습자 정답 예제에서 키워드 패턴 "(키워드1)의 (키워드2)"의 자질 값은 1(=1/1)이 된다. 반면에, 문제 B의 오답 예제에서 키워드 패턴 "(키워드1)의 (키워드2)"의 자질 값은 0(=0/1)이 된다.

셋째, 조합 자질은 텍스트 자질과 키워드 자질을 조합하여 만든 자질이다. 조합 가능한 전체 자질쌍 중에서 서술형 주관식 답안 자동 채점의 성능 개선에 도움이 된 일부자질을 [표2]와 같이 선별하였다. 예를 들어, 모범 답안에 1개의 숫자와 2개의 영어 단어가 포함되어 있고, 학습자 답안에 1개의 숫자와 1개의 영어 단어가 포함되어 있다면, 텍스트 자질에서 숫자 개수 비율 자질은 1(=1/1)이고 영어 단어 개수 비율 자질은 0.5(=1/2)가 된다. 이러한 숫자 개수 비율 자질과 영어 단어 개수 비율 자질을 곱하여 새로운 조합자질 영숫자 비율자질을 0.5(=1×0.5)를 추가한다.

제안하는 서술형 주관식 답안 자동 채점 모델은 이렇게 구성한 자질을 모범 답안과 학습자 답안을 비교하여 추출하고, 추출한 자질을 수식 (2), (3), (4)와 같이 C4.5[22], ME(Maximum Entropy)[23], SVM(Support Vector Machines)[24]의 기계학습 기반 분류기에 적용한다. 기계학습 기반 분류기의 특징이 서로 다르므로, 분류결과가 서로 다를 수 있다. 이러한 점을 고려하여 제안하는 모델은 좀 더 신뢰성 있게 답안을 채점하기 위해서, 수식(1)과 같이 각 기계학습 기반 분류기에서 나온 결과 cC4.5, cME , cSVM이 투표하여 만장일치로 선택한 결과를 최종결과 c로 채택한다.

$$c = \begin{cases} \text{정답} & \text{if } c_{c4.5} = c_{ME} = c_{SVM} = \text{정답} \\ \text{오답} & else\,\text{if } c_{c4.5} = c_{ME} = c_{SVM} = \text{오답} \\ \text{보류} & otherwise \end{cases} \quad (1)$$

$$c_{c4.5} = DecisionTree(x_1, x_2, \ldots x_n) \wedge c_{c4.5} \in \{\text{정답}, \text{오답}\} \quad (2)$$

$$\underset{c_{ME} \in \{\text{정답},\text{오답}\}}{\arg max} \frac{1}{Z(x)} exp\left(\sum_{i=1}^{n} \lambda_i f_i(x,c)\right) \quad (3)$$

$$c_{SVM} = h(\vec{x}) = \begin{cases} \text{정답} & \text{if } \vec{w} \bullet \vec{x} + b > 0 \\ \text{오답} & otherwise \end{cases} \quad (4)$$

수식 (2)는 기계학습 기반 분류기 C4.5에서 생성한 의사결정트리에 n개의 자질 값 x1, x2, … xn을 적용한 결과가 cC4.5, 라는 것을 나타낸다[22]. 수식 (3)은 기계학습 기반 분류기 ME가 각 자질별로 가중치 λi와 자질함수 fi를 적용하여 나온 확률값을 바탕으로 분류결과 cME를 추론한다는 것을 나타낸다[23]. ME는 학습시 제약 조건을 모두 만족하는 확률분포 가운데 엔트로피가 최대가 되는 확률 분포를 찾는다[23]. 수식 (4)는 기계학습 기반 분류기 SVM이 자질 벡터 $\vec{x}$에 가중치 벡터 $\vec{w}$를 곱한 결과를 바탕으로 분류결과 cSVM를 추론한다는 것을 나타낸다[24]. SVM은 학습시 구조적 리스크 최소화를 통해 벡터 공간에서의 최적의 결정경계 영역을 나타내는 가중치 벡터 $\vec{w}$와 상수 b를 찾는다[25][26].

IV. 실험 및 평가

제안하는 모델의 성능을 평가하기 위해서, 전산학 관련 9개 수업에서 출제된 68개의 서술형 주관식 문제에 대한 학습자들의 답안 1,569개를 수집하여 평가말뭉치를 구축하였다. 각 문항의 채점기준이 되는 모범 답안은 전문서적을 참고하여 수기 채점자 두 명이 의논하면서 작성하였다. 이렇게 구축된 평가말뭉치의 각 답안에 대해서 제안하는 모델은 형태소 분석[25]을 수행하고, [표2]과 같은 자질을 추출한다. 그리고 추출된 자질을 기계학습기반 분류기 C4.5[22], ME[23], SVM[24]에 적용하고 투표하여 답안을 분류한다.

제안하는 모델이 서술형 주관식 답안을 얼마나 정확하게 자동으로 채점하는지를 정량적으로 평가하기 위해서, 수식 (5)과 같이 제안하는 모델의 정확률을 측정한다[26]. 그리고, 제안하는 모델이 서술형 주관식 답안을 올바르게 채점한 경우가 얼마나 많은지를 평가하기 위해, 수식 (6)과 같이 재현율을 측정하였다. 수식 (7)는 정확률과 재현율의 조화평균인 f-값을 나타낸다. 신뢰성 있는 실험결과를 얻기 위해서, 평가집합을 90%의 학습집합과 10%의 실험집합으로 나누고, 제안하는 서술형 주관식 답안 자동 채점 모델에 적용하여, 정확률, 재현율, f-값을 측정하는 과정을 10회 반복하였고 평균값을 사용하였다.

$$\text{정확률} = \frac{\text{올바르게 자동채점한 답안수}}{\text{자동채점한 답안수}} \quad (5)$$

$$\text{재현율} = \frac{\text{올바르게 자동채점한 답안수}}{\text{실험집합에 포함된 전체 답안수}} \quad (6)$$

$$f-\text{값} = \frac{2 \times \text{정확률} \times \text{재현율}}{\text{정확률} + \text{재현율}} \quad (7)$$

분류자질의 특징에 따라서 답안 자동 채점의 정확률과 재현율이 어떻게 변하는지 살펴보기 위해서, 실험에 사용된 42개의 자질을 [표2]과 같이 텍스트 자질, 키워드 자질, 조합자질로 나누고, [표3]와 같이 제안하는 모델에 3가지 자질 유형을 활용하여 평가하였다. 기계학습기법을 C4.5, ME, SVM 중 하나만 사용하는 경우, 실험집합에 포함된 전체 답안에 대해서 보류하는 답안 없이 모두 자동 분류하므로, 정확률, 재현율, f-값이 동일하게 나타난다. 따라서, 제안하는 모델의 재현율은 [표3]의 정확률과 동일하다.

표 3. 자질 유형 조합에 따른 정확률
Table 3. Precision on Feature Type Combination

구분	C4.5	ME	SVM
텍스트 자질	74.65	75.11	73.70
키워드 자질	74.48	72.64	72.45
조합 자질	70.28	70.43	70.32
텍스트+키워드	79.48	76.80	76.10
텍스트+조합	75.06	74.83	74.89
키워드+조합	76.32	73.72	73.64
전체 자질	79.81	77.47	79.24

표 4. 각 자질 제외한 모델의 정확률
Table 4. Precision of the Model without Each Feature

구분		C4.5	ME	SVM
전체자질		79.81	77.47	79.24
텍스트	전체-문장 자질	79.35	76.56	79.22
	전체-어휘 자질	77.47	74.87	75.65
	전체-음절 자질	79.87	77.73	78.51
	전체-명사 자질	80.13	77.21	78.64
	전체-동사 자질	80.52	77.4	78.25
	전체-접속사 자질	79.81	76.56	79.03
	전체-부사 자질	77.14	75.32	75.84
	전체-조사 자질	78.25	76.3	77.92
	전체-숫자 자질	78.05	**74.29**	76.69
	전체-영어 자질	80.39	76.88	78.57
	전체-수식 자질	78.96	77.21	77.08
	전체-부정어 자질	80.06	76.62	78.64
키워드	전체-키워드 유무	77.6	77.14	77.21
	전체-답안에서 키워드 비율	80.32	76.56	77.92
	전체-키워드 출현 빈도	80.19	75.39	79.22
	전체-부정 키워드	79.81	76.04	78.7
	전체-키워드 평균	77.53	75.39	77.34
	전체-키워드 순서	79.55	75.97	78.77
	전체-키워드 간의 거리	79.22	76.36	77.21
	전체-순서 중요도	78.9	76.43	77.73
	전체-부가 순서 중요도	79.87	77.92	77.73
	전체-"(키워드)에서"	80.84	**78.77**	78.25
	전체-"(키워드)의"	79.55	77.21	78.25
	전체-"(키워드)하는데"	**81.36**	77.73	79.35
	전체-"(키워드)를"	75.84	76.17	**75.19**
	전체-"~에서 (키워드)"	77.4	75	73.64
	전체-"~의(키워드)"	79.48	77.4	79.48
	전체-"~하는데 (키워드)"	80	77.6	78.7
	전체-"~를 (키워드)"	**75.52**	74.68	76.04
	전체-"(키워드1)에서 (키워드2)"	78.12	75.39	75.58
	전체-"(키워드1)의 (키워드2)"	79.55	76.3	77.86
	전체-"(키워드1)하는데 (키워드2)"	78.96	76.04	76.56
	전체-"(키워드1)를 (키워드2)"	80.71	78.44	**79.87**
조합	부사+숫자	80.52	76.62	78.12
	숫자+영어	78.7	76.69	75.91
	숫자+키워드 순서	78.25	76.95	77.92
	키워드 거리+부가 순서 중요도	77.01	76.82	77.14
	키워드 출현 빈도+키워드 평균	78.96	77.27	78.44
	부정어+숫자+영어+수식	80.65	76.82	78.64
	숫자+"(키워드1)을 (키워드2)	79.61	78.12	79.68
	"(키워드1)을 (키워드2)"+"(키워드)의"	80.06	77.47	79.48
	"(키워드)하는데"+"~에서 (키워드)"	79.61	76.69	78.83

〔표3〕에서 정확률은 텍스트 자질 유형, 키워드 자질 유형, 조합 자질 유형의 순서로 높다. 특히, 텍스트 자질을 사용한 ME의 정확률이 75.11%인데 반해서, 키워드 자질과 조합자질을 모두 사용한 ME의 정확률이 73.72%로 오히려 낮다. 이는 텍스트의 특징을 표현하는 자질들이 키워드 자질들에 비해서 문제의 종류에 영향을 적게 받고 적용 범위가 넓다는 것을 보여준다. 반면에, 학습자 답안이 모범 답안의 키워드를 얼마나 포함하는지를 나타내는 키워드 자질은 키워드와 완전히 일치하지 않은 유사어를 사용한 경우에 취약하고, 문제유형에 다소 민감하게 영향을 받는 경향이 있었다.

또한, 〔표3〕의 C4.5, ME, SVM의 세 가지 분류기 모두에서 텍스트 자질에 다른 자질을 추가하여 활용하는 경우 대체적으로 정확률이 높아지는 경향을 보이며, 전체 자질을 사용하는 경우가 가장 높은 정확률을 보인다. 이는 텍스트 자질, 키워드 자질, 조합 자질의 특징이 서로 다르기 때문에, 함께 사용하면 자질들간에 서로 상호보완할 수 있다는 것을 보여준다.

본 논문에서 제안한 42개 자질의 영향력을 분석하기 위해서, 〔표4〕와 같이 전체 자질에서 각 자질을 제외했을 때 정확률이 얼마나 떨어지는지를 살펴보았다〔19〕. 서술형 주관식 답안 자동 채점에 긍정적으로 도움이 되는 자질이라면, 전체 자질에서 해당 자질을 제외할 경우 정확률이 많이 떨어진다. 반면에 부정적으로 영향을 주는 자질은 전체 자질에서 해당 자질을 제외할 경우 오히려 정확률이 상승된다.

C4.5 기반 답안 채점 모델에서 전체 자질을 사용한 경우 정확률이 79.81%인데, 키워드 패턴 자질 '~를 (키워드)하다'를 제외하면 정확률이 75.52%로 가장 많이 하락하였다. 즉, 키워드 패턴 자질 '~를 (키워드)하다'가 C4.5기반 모델에서 4.29%의 영향력을 가진 중요한 자질임을 나타낸다. 한편, 키워드 패턴 자질 '(키워드)하는데'는 정확률이 81.36%로 오히려 상승되어 답안 채점을 방해되는 자질임을 나타낸다.

ME 기반 답안 채점 모델에서는 숫자 자질이 3.18%로 가장 영향력이 높았고, 키워드 패턴 자질 '(키워드)에서'가 오히려 부정적인 영향력을 나타냈다. SVM기반 답안 채점 모델에서는 키워드 패턴 자질 '(키워드)를 ~하다'가 가장 높은 4.05%의 정확률 감소폭을 나타냈다. 이러한 결과는 기계학습 분류 기법에 따라서 선호하는 자질이 서로 다를 수 있다는 것을 나타낸다.

한편, C4.5 기반 답안 채점 모델은 자질의 포함여부에 따라 정확률의 증감폭이 큰 반면에, 다른 모델은 상대적으로 정확률의 증감폭이 크지 않다. 이는 C4.5기반 모델은 결정 트리 생성시 자질에 영향을 많이 받는 반면에 다른 모델은 자질별로 가중치나 확률을 달리 줄 수 있으므로 상대적으로 자질

에 영향을 덜 받는 편이라는 것을 나타낸다.

표 5. 부정적인 자질 제외한 모델의 정확률
Table 5. Precision of the Model without Negative Features

구분	C4.5	ME	SVM
전체자질	79.81	77.47	79.24
전체자질-부정적인자질	83.00	78.77	79.35

답안 자동 채점의 정확률에 부정적으로 영향을 준 자질을 모두 제외할 경우 정확률이 얼마나 향상되는지를 살펴보기 위해서, 〔표4〕에서 회색바탕으로 표현한 자질을 제외하여 〔표5〕과 같이 실험하였다. C4.5 기반 자동 채점 모델은 17개의 자질을 제외하였고, ME 기반 자동 채점 모델은 8개의 자질을 제외하였고, SVM 기반 자동 채점 모델은 4개의 자질을 제외하였다. C4.5는 자질의 구성에 영향을 많이 받아서 정확률이 3.19%까지 향상된 반면, 다른 자동 채점 모델은 자질의 구성에 영향을 덜 받으므로 상승폭이 미미하였다.

C4.5, ME, SVM 기반 답안 자동 채점 모델의 각 분류 결과를 투표하여 다수가 선택한 결과를 최종 채택한 경우에 성능이 어떻게 변화하는지를 살펴보기 위해서, 〔표6〕과 같이 실험하였다. 각 모델은 주어진 답안에 대해 보류하지 않고 모두 채점 분류하므로 〔표3〕, 〔표4〕, 〔표5〕에서는 정확률, 재현율, f-값이 동일하였지만, 〔표6〕에서는 만장일치하지 않은 결과는 채택하지 않고 보류하여 정확률과 재현률이 일치하지 않는다.

표 6. 투표결과에 따른 성능
Table 6. Performance According to Voting Results

도구	정확률	재현률	F-값
C4.5	83.00	**83.00**	83.00
ME	76.96	76.96	76.96
SVM	78.22	78.22	78.22
C4.5+ME : 만장일치	89.93	79.23	84.24
C4.5+SVM : 만장일치	**90.08**	80.91	**85.25**
ME+SVM : 만장일치	83.01	74.05	78.27
C4.5+ME+SVM : 다수결	80.38	80.38	80.38
C4.5+ME+SVM : 만장일치	**90.57**	63.94	74.96

C4.5 기반 답안 채점 모델은 SVM 기반 답안 채점 모델과 함께 사용하면 정확률이 90.08%까지 상승하며, 이는 C4.5 기반 답안 채점 모델 하나만 사용했을 때에 비해서 7.08% 개선된 결과이다. 이는 규칙기반 분류기법 C4.5와 가중치벡터 기반 분류기법 SVM의 특징이 다르므로 서로 상호보완하면서 긍정적인 영향을 끼쳤기 때문이다. 만장일치 결과만 채택하는 경우 보다 신뢰성있게 채점하여 정확률은 90.57%까지 증가하지만, 보류하는 답안이 증가하면서 재현률은 63.94%로 크게 감소하였다. 한편, 세 가지 모델의 결과에 대해 다수결로 최종결과를 채택하는 경우 C4.5 단독으로 채택하는 경우보다 성능이 오히려 감소하였다. 이는 C4.5가 답안을 올바르게 채점하였음에도 불구하고 ME와 SVM의 틀린 채점 결과가 최종적으로 채택된 경우가 2.32%정도 있었기 때문이다. 다수결로 채택하면 보류하는 경우가 없으므로, 정확률과 재현율이 동일하게 나타난다.

V. 결 론

본 논문에서는 자질 추출 단계, 답안 분류 단계, 투표 단계로 구성된 서술형 주관식 답안 자동 채점 모델을 제안한다. 제안하는 모델은 주어진 모범 답안과 학습자 답안을 비교하여 자질을 추출하고, 이를 바탕으로 기계학습 기반 분류기 C4.5, ME, SVM를 이용하여 답안을 분류하고, 분류 결과를 투표하여 최종 결과로 선택한다. 제안하는 모델의 특징은 다음과 같다.

첫째, 제안하는 서술형 주관식 답안 자동 채점 모델은 투표기법을 이용하여 보다 신뢰성있게 서술형 주관식 답안을 채점할 수 있다. 즉, 답안을 표현하는 자질을 C4.5, ME, SVM의 기계학습 기반 분류기에 적용하여 채점하고, 다수결 또는 만장일치의 투표방법을 바탕으로 최종 채점 결과를 결정한다. 실험결과 기계학습 기반 분류기 C4.5의 정확률은 83.00%인데 반해, C4.5, ME, SVM의 기계학습 기반 분류기의 결과가 일치한 경우만 허용한 경우 정확률이 90.57%까지 개선되었다.

둘째, 제안하는 서술형 주관식 답안 자동 채점 모델은 모범 답안과 학습자 답안을 42개의 다양한 자질을 바탕으로 답안을 표현한다. 텍스트 자질은 모범 답안 대비 학습자 답안의 문장길이 비율, 품사별 분포도 등의 답안 자체의 텍스트 특징을 나타낸다. 키워드 자질은 모범 답안과 학습자 답안에서 키워드가 나타난 패턴이 동일한 경우가 있는지를 나타낸다. 조합자질은 텍스트자질과 키워드 자질을 조합하였을 때 시너지 효과를 내는 자질들을 조합하여 표현한다. 부정어 관련 자질도 텍스트 내에서의 부정어 존재여부와 키워드와 관련된 부정어 존재여부를 분리하여 활용한다.

셋째, 제안하는 서술형 주관식 답안 자동 채점 모델은 모델 구축 비용을 줄일 수 있다. 즉, 문제유형별로 서술형 주관식 답안 자동 채점 모델을 따로 구축하는 기존방법에 비해, 제안하는 모델은 문제유형을 한가지로 일반화하여 채점하므로 모델 구축 비용을 줄일 수 있다.

참고문헌

〔1〕 C. G. Jung, R. I. Choi, "A Developmental Plan on Public Education in Digital Times, 21c," Journal of The Korea Society of Computer and Information, Vol. 8, No. 1, pp.120-128, Mar. 2003.

〔2〕 H. J. Kim, J. H. Choi, "Development of a Teaching and Learning Model for Educational Usage of Web 2.0 and Its Effect Analysis," Journal of The Korea Society of Computer and Information, Vol. 16, No. 10, pp.45-52, Oct. 2011.

〔3〕 K. A. Jin, "Development of automated scoring system for English writing." English Language & Literature Teaching, Vol. 13, No. 1, pp.235-259, Spring. 2007.

〔4〕 K. B. Kim, J. H. Cho, "Performance Assessment System using Fuzzy Reasoning Rule," Journal of The Korea Society of Computer and Information, Vol. 10, No. 1, pp.209-216, Mar. 2005.

〔5〕 O. Y. Kwon, "The Design and Implementation of Web-based Subjective questions Grading Algorithm," Department of Educational(Coumputer Scientce) The Graduate School of Education. Hanseo University, pp.1-3(52), Feb. 2004.

〔6〕 W. S. Kang, "Design and Implementation of a Subjective-type Evaluation System Using Syntactic and Case-Role Information," THE JOURNAL OF KOREAN ASSOCIATION OF COMPUTER EDUCATION, Vol. 10, No. 5, pp.61-69, Sept. 2007.

〔7〕 H. J. Park, 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.207-216, Jul. 2003.

〔8〕 E. H. No, "Developing an Automatic Content Scoring Program for Short Answer Korean Items in Large-Scalse Assessments," Korea Institute for Curriculum and Evaluation, pp.3-48(230), 2012.

〔9〕 Y. R. Kim, B. Y. Cho, "(A)review on NAEA Korean test based on CURRV framework," Korea Institute for Curriculum and Evaluation, pp.1-12p, 2011.

〔10〕 C. Park, "Validating an Automated Scoring System of Constructed-response items for applications," Journal of Educational Evaluation. Vol. 22, No. 3, pp615-631, Sept. 2009.

〔11〕 E. M. Jung, "Design and Implementation of Automatic Marking System for a Subjectivity Problem of the Program," Journal of Korea Multimedia Society, Vol. 12, No. 5, pp767-776, May. 2009.

〔12〕 J. M. Cho, K. H. Kim, "A Study on design of The Internet-based scoring system for constructed responses," THE JOURNAL OF KOREAN ASSOCIATION OF COMPUTER EDUCATION, Vol. 10, No. 2, pp.89-100, Mar. 2007.

〔13〕 W. J. Cho, "An Intelligent Marking System based on Semantic Kernel and Korean WordNet," Graduate School. Hallym University, pp. 3-33, Aug. 2006.

〔14〕 J. S. Oh, W. J. Cho, Y. S. Kim, J. Y. Lee, "A Descriptive Question Marking System based on Semantic Kernels," Journal of Korean Institute of Information Technology, Vol. 3, No. 4, pp.95-104, Sept. 2005.

〔15〕 D. K. Chung, "Subjective Questions Scoring System Using Vector Similarity and Thesaurus," The Graduate School of Education. Dongguk University, pp.1-45, 2002.

〔16〕 J. H. Cho, H. K. Jung, C. Y. Park, Y. S. Kim, "An Autonomous Assessment of a Short Essay Answer by Using the BLEU," HCI2009, pp.606-605, Gangwon-do Phoenixpark Convention center, Korea, Feb. 2009.

〔17〕 S. H. Jang, "Planning and realizing the Automatic Grading System for Subjective

Questions Through an Analysis of Question Types," The Graduate School of Education. Andong National University, 53p, 2008.
〔18〕 Van Halteren, H., Daelemans, W. and Zavrel, J., "Improving accuracy in word class tagging through the combination of machine learning systems," Computational linguistics, Vol. 27, No. 2, pp.199-229, Feb. 2001.
〔19〕 Van Halteren, H., Zavrel, J. and Daelemans, W., "Improving data driven wordclass tagging by system combination," Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Vol. 1, pp.491-497, 1998.
〔20〕 J. T. Lee, Y. I. Song, S. Y. Park, H. C. Rim, "Text-Confidence Feature Based Quality Evaluation Model for Knowledge Q&A Documents," Journal of KIISE : Software and Applications, Vol. 35, No. 10, pp.608-615, Nov. 2008.
〔21〕 J. T. Lee, Y. I. Song, H. C. Rim, "Predicting the quality of answers using surface linguistic features," In Proceedings of the 6th International Conference on Advanced Language Processing and Web Information Technology, pp.111-116, 2007.
〔22〕 Quinlan, J. Ross, "C4.5:Programs for Machine Learning," Morgan Kaufmann Publishers, 1993
〔23〕 H. J. Jung, "A Maximum Entropy Approach to Korean Part-of-Speech Tagging," Information & Communications of Engineering, Myongji University, pp.11-18, 2005.
〔24〕 Vapnik, "The Nature of Statistical Learning Theory," Springer-Verlag, 1999
〔25〕 S. Y. Kim, "Incremental Supervised Learning based on SVM with Unlabeled Documents," The Graduate School of Computer Science. Yonsei University, pp.12-13, 2002.
〔26〕 T. Joachims, "Estimating the Generalization Performance of an SVM Efficiently," Proceedings of the 17th International Conference on Machine Learning (ICML 2000), pp. 431-438, 2000.
〔27〕 S. Z. Lee, "New Statistical Models for Automatic Part-of-Speech Tagging," Ph.D. Thesis, Korea University, 1999.
〔28〕 S. Y. Kim, "Object of Interest Extraction Using Gabor Filters," THE JOURNAL OF KOREAN ASSOCIATION OF COMPUTER EDUCATION, Vol. 13, No. 2, pp.87-94, 2008.

저 자 소 개

허 정 만

현 재: 상명대학교
디지털미디어학부 재학중
관심분야: 컴퓨터공학
Email : vngofgof@naver.com

박 소 영

1997: 상명대학교
전자계산학과 이학사.
1999: 고려대학교
컴퓨터공학과 이학석사.
2005: 고려대학교
컴퓨터공학과 이학박사
현 재: 상명대학교
게임모바일콘텐츠학과 조교수
관심분야: 컴퓨터공학
Email : ssoya@smu.ac.kr