교육과정평가연구
The Journal of Curriculum and Evaluation
2013, Vol. 16, No. 3, pp. 147~165

쓰기 자동채점 알고리듬의 성능 비교: 최대 엔트로피 기법과 서포트 벡터 회귀 기법

이용상(한국교육과정평가원 부연구위원)* 시기자(한국교육과정평가원 연구위원)
박도영(한국교육과정평가원 부연구위원) 윤경아(SK 텔레콤 매니저)
구슬기(한국교육과정평가원 전문연구원) 임황규(한국교육과정평가원 전문연구원)

≪ 요 약 ≫

한국교육과정평가원에서는 국가영어능력평가시험에 최적화된 쓰기 자동채점 프로그램을 개발 중에 있다. 자동채점 프로그램의 성능은 기계학습에 적용되는 알고리듬에 의해 많은 영향을 받으며, 현재 국가영어능력평가시험 쓰기 자동채점 프로그램에서는 가장 일반적으로 사용되고 있는 알고리듬인 최대 엔트로피 기법과 서포트 벡터 회귀 기법을 적용하고 있다. 본 논문에서는 쓰기 자동채점 프로그램에 적용된 두 가지 알고리듬에 따른 채점 성능을 비교·분석하였다. 이를 위해 n-fold 방식을 이용한 모의실험과 실제 채점자료를 이용한 실증 연구를 통하여 두 가지 알고리듬의 성능을 비교하였다. 분석 결과, 모의실험 및 실증 분석 모두에서 서포트 벡터 회귀 기법이 최대 엔트로피 기법보다 국가영어능력평가시험 쓰기 영역 채점에 더 적합한 알고리듬인 것으로 확인되었다.

주제어: 자동채점 알고리듬, 최대 엔트로피 기법, 서포트 벡터 회귀 기법, n-fold 모의실험

Ⅰ. 서론

국가영어능력평가시험은 학생들의 의사소통 능력을 향상시키고, 듣기, 읽기, 말하기, 쓰기 능력을 균형 있게 평가함으로써, 국외 영어시험에 대한 의존도를 낮추고 영어 공교육에 대한 바람직한 기준을 제시하기 위한 목적으로 개발된 검사이다. 이와 같은 국가영어능력평가시험은 기존

* 제1저자 및 교신저자, yong21c@kice.re.kr

의 영어 시험들과 달리 구성형인 말하기, 쓰기 시험을 포함하고 있으며, 이들 구성형 시험에 대한 학생들의 답안은 컴퓨터에 의하여 자동채점되는 듣기, 읽기 영역의 선다형 시험과는 달리, 별도의 인증과정을 거친 채점위원에 의해 채점이 이루어지고 있다. 따라서 채점에 대한 신뢰성은 국가영어능력평가시험의 공신력 확보를 위한 선결 조건이며 한국교육과정평가원에서는 채점자 신뢰성을 높이기 위해 채점자 직무 연수를 실시하는 등 다양한 노력을 기울이고 있다. 그러나 국가영어능력평가시험이 대규모로 시행될 경우 구성형 답안 채점을 위해 대규모 채점 위원들이 필요하지만, 채점자 훈련에 소요되는 시간과 비용을 고려할 때 대단위 채점자 훈련을 통해 신뢰도 높은 채점자들을 육성하는 데는 일정한 한계가 있다.

이러한 구성형 시험 채점의 현실적인 어려움을 해결하기 위해 ETS나 Pearson사와 같은 국외 주요 평가 기관들에서는 자연어 처리 기술을 기반으로 한 자동채점 솔루션을 연구·개발하고 이를 공인시험 및 모의시험에 적용하고 있다(Dikli, 2006). 이와 같이 자동채점이 국외 시험에 적용되고 있는 주된 이유는 채점의 효율성과 일관성을 확보하기 위해서이다. 인간채점과 비교할 때 자동채점은 채점에 소요되는 시간과 비용을 획기적으로 감소시킬 수 있을 뿐만 아니라 인간채점에서 발생하는 채점 주관성, 피로 및 외부 상황의 영향 등을 통제함으로써 채점의 일관성을 유지할 수 있다. 그동안 국내에서는 자동채점 시스템 개발에 필요한 전문적 기술과 대규모 투자비용 등의 이유로 상용 프로그램 개발이 전무한 실정이었으나, 한국교육과정평가원은 국가영어능력평가시험의 구성형 채점이 갖는 현실적인 어려움을 극복하고 자동채점의 장점들을 적극 활용하기 위해 국내 기술력으로 국가영어능력평가시험에 적합한 자동채점 프로그램을 개발 중이다(시기자 외, 2012; 신동광 외, 2012).

자동채점이 국내에 성공적으로 도입되기 위해서는 자동채점의 정확성이 검증되어야 하며, 이러한 자동채점의 정확성은 본질적으로 자동채점 프로그램이 채택하고 있는 채점 알고리듬에 의해 좌우될 수밖에 없다. 따라서 현재 개발 중인 국가영어능력평가시험 쓰기 자동채점 프로그램에 포함되어 있는 최대 엔트로피(Maximum Entropy, ME)와 서포트 벡터 회귀(Support Vector Regression, SVR) 알고리듬 중에서 자동채점 결과를 산출하는데 실제적으로 사용하게 될 알고리듬을 확정해둘 필요가 있다. ME와 SVR은 비교적 최근에 소개된 기계학습 알고리듬으로서 채점 성능이 규칙 기반 알고리듬과 같은 기존의 알고리듬에 비해 우수한 것으로 보고된 바 있다(Li & Yen, 2010; Sukkarieh, 2010).

이와 같은 분석의 필요성을 바탕으로 본 연구는 한국교육과정평가원에서 개발 중인 국가영어능력평가시험 쓰기 자동채점 시스템의 기계학습기에 적용된 ME와 SVR 알고리듬의 정확성 비교에 그 목적을 두었다. 이를 위해 2011년 국가영어능력평가시험 예비평가 쓰기 2급 2문항에 대한 1,000명의 답안을 대상으로 모의실험과 실증 자료 분석을 실시하여 ME와 SVR 알고리듬에 따른 자동채점의 성능을 비교하였다. 모의실험에서는 10-fold 교차실험을 실시하였으며, 알고리듬에 따른 자동채점의 성능 비교를 위해 인간채점 간, 인간채점과 ME 자동채점 간, 인간채점과 SVR 자동채점 간 상관계수를 살펴보았다.

Ⅱ. 이론적 배경

국가영어능력평가시험 쓰기 자동채점 프로그램은 인간채점자가 채점하는 방식, 즉 채점 훈련(학습) 후 형성된 직관(지식)에 의해 채점하는 방식을 컴퓨터에 적용한 기계학습을 도입하여 크게 채점모델 생성과 자동채점의 단계를 거친다. 채점모델 생성 단계에서는 채점자들의 채점 결과가 있는 응시자의 답안들(국가영어능력평가시험 코퍼스)로부터 점수대별 채점 답안과 답안의 평가적 특성을 채점자질(feature)로 추출하고(Thomas, 2003), 자동채점 단계에서는 앞 단계에서 생성된 채점모델을 활용하여 새롭게 입력된 답안에 대한 채점을 자동으로 수행한다. 위와 같은 작업을 효과적으로 수행하기 위해, 자동채점 프로그램은 언어 처리기, 오류 분석기, 그리고 자동채점기로 구성되어 있다.

언어 처리기와 오류 분석기는 형태소 분석 등의 언어분석 및 영문법적 오류 검사 등을 수행하여 답안의 특성을 추출하기 위한 기본 정보들을 생성하는 역할을 한다. 자동채점기는 언어 처리기와 오류 분석기에서 생성된 기본 정보를 활용하여 기계학습을 통한 채점모델 생성과 자동채점을 수행하는데, 이는 크게 자질 추출/선택기, 기계학습기, 점수 산출기로 구성된다. 자질 추출/선택기는 응시자의 답안에 대한 언어 처리 결과와 오류 정보를 입력 받아 각 답안에서 이미 정의된 채점자질을 기반으로 값을 추출/선택하여 자질 벡터(feature vector), 즉 기계학습기의 입력 데이터를 생성한다. 기계학습기는 국가영어능력평가시험 코퍼스의 채점자 점수들과 자질 추출/선택기에서 생성된 자질 벡터들을 활용하여 ME와 SVR 기계학습 알고리듬을 기반으로 하는 채점모델을 알고리듬별로 생성한다. 점수 산출기는 기계학습 알고리듬에 의해 생성된 채점모델을 기반으로 하여 점수가 정해지지 않은 학생의 새로운 답안을 자동채점하는 기능을 한다. 여기에서 어느 알고리듬에 의한 채점모델을 사용하는가에 따라서 자동채점의 결과도 다르게 산출된다.

1. 기계학습 알고리듬

기계학습기는 채점자가 채점이라는 학습 과정을 통해 생성된 경험과 직관을 이용하여 인간채점자가 새로운 답안의 채점 시에 이전에 채점했던 답안과 유사한 특성을 갖는 답안에 유사한 점수를 주는 방식을 컴퓨터(기계)에 그대로 적용한 것이다. 기계학습기는 점수대별 답안의 특성을 정량화할 수 있을 정도로 다양한 특성을 갖는 충분한 양의 답안들로부터 추출된 자질 벡터 값들과 이에 대응되는 인간채점자 점수와의 정량적인 관계를 확률 및 통계이론에 기반한 학습 과정을 거쳐 채점모델을 생성한다. 자동채점에 주로 사용되는 기계학습 알고리듬은 분류(classification) 분석에 기반한 방법과 회귀(regression) 기반 방법으로 나뉘며, 본 연구에서는 국가영어능력평

가시험 쓰기 채점 데이터에 적합한 방법을 찾기 위해 분류분석에 기반한 방법의 대표적인 기법인 ME 기법과 회귀분석에 기반한 방법의 대표적인 기법인 SVR 기법을 이용하여 채점모델을 형성하여 자동채점을 실시하고 그 결과를 비교・분석하였다.

가. 최대 엔트로피(ME) 기법

최대 엔트로피(ME) 기법은 공간 물리학, 컴퓨터 비전, 자연어 처리 등 여러 분야에서 성공적으로 적용된 다목적 기계학습 기법(Adam et al., 1996; Valenti et al., 2003)으로서, 서로 다른 정보를 분류하는 데 주로 사용되는 분류분석에 기반한 기법 중 하나이다. 이 기법은 특정 해(solution)가 다른 것에 우선한다는 증거가 없으면 모든 해는 같은 가능성을 가져야 한다는 직관을 구현한 것으로, 미리 정의된 제한 조건들은 만족하면서 그 이외의 경우 동일한 확률값을 갖게 하는 확률에 기반한 지수 모델 기법이다. 답안 x가 특정 점수 y가 될 ME 모델을 수식으로 표현하면 다음과 같다.

$$p(x,y) = \frac{\Pi_i \mu_i^{f_i(x,y)}}{Z}$$

여기서, $\mu_i = \exp(\lambda_i)$이고, λ_i는 실수 패러미터이다. 또한 Z는 정규화 상수이고 $f(x,y)$들은 한 답안과 점수의 쌍인 (x,y)에 대한 자질이다.

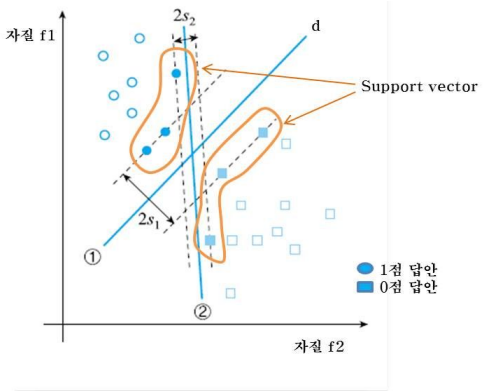
각 자질 $f_i(x,y)$에 대해서, 학습코퍼스 $S = \{(x_1,y_1),...,(x_N,y_N)\}$에 대한 분포 p의 기댓값은 S의 경험적 분포 $\tilde{p}$의 기댓값으로도 설정된다. 따라서 제약은 아래 수식과 같고, 이 제약을 만족시키는 해를 찾는 것이 기계학습 과정이 된다.

$$E_{\tilde{p}}[f_i] = \sum_{j=1}^{N} \tilde{p}(x_j,y_j) f_i(x_j,y_j) = E_p[f_i]$$

이를 본 연구에 적용하여 설명하면 학습 대상이 되는 채점 코퍼스 S의 답안 $(x_1, y_1), \dots, (x_N, y_N)$들의 언어 처리 및 오류 분석 결과들을 기계학습기에 입력하면 이 결과로부터 자질 f_i를 추출하고 ME 알고리듬에 의해 $E_{\tilde{p}}[f_i]$ 를 구한 후 1점부터 5점까지 각 점수대별 확률을 산출하는 5개의 식 $p(x,y)$들을 채점모델로 생성한다. 새로운 답안 x에 대한 점수 y는 이 점수대별 채점모델을 사용하여 가장 높은 확률을 갖는 점수로 결정된다. 예를 들어 답안 x에 대한 점수대별 확률이 $p(x,1) = 0.2, p(x,2) = 0.2, p(x,3) = 0.4, p(x,4) = 0.1, p(x,5) = 0.1$라면 가장 높은 확률 0.4를 갖는 점수 3점이 자동채점 결과인 y가 된다.

나. 서포트 벡터 회귀 기법

서포트 벡터 머신(Support Vector Machine) 기법은 원래는 분류를 위해 개발된 방법으로서 이를 회귀에 활용하도록 변형한 기법을 서포트 벡터 회귀 기법이라 한다. 서포트 벡터 머신은 각 점수대별로 데이터를 분류하기 위해 자질 벡터들을 입력으로 받아 각 자질들을 하나의 차원으로 간주하여 차원별로 점수를 구분하는 선형적 또는 비선형적 최적 분리 경계면(linear or nonlinear optimal separating hyperplane)을 찾는다(Cortes et al., 1995; Vapnik et al., 1997). 최적의 경계면은 분류할 두 집단으로부터 가장 멀리 떨어진 초평면(hyper-plane)으로 정의되는데, 경계면에 가장 가까이 있는 데이터를 서포트 벡터라고 한다. 〔그림 1〕에 2개의 자질(2차원)로 구성된 평면에서 0점과 1점 답안을 구분하는 간단한 서포트 벡터 머신을 제시하였고, 여기서 d가 초평면이 된다.



〔그림 1〕 서포트 벡터의 개념

서포트 벡터 회귀 기법은 데이터(x_i, y_i)가 주어진 서포트 벡터 머신에서 손실함수를 이용하여 모든 데이터에서의 거리를 최소로 하는 초평면 y를 예측하는 최적의 경계면 $f(x)+ w_k + b$을 찾는 기법이 된다. 특히 본 연구에서는 ε-집약적 손실함수(ε-intensive loss function)를 사용하여 경계면에서 각 데이터 사이의 거리가 ε보다 작아지도록 제약조건을 설정하고 w_k를 찾아내어 회귀모델을 생성한다. 이렇게 생성된 모델의 독립변수들은 기계학습을 위해 선정된 자질들이 되고, 종속변수는 예측하려는 점수가 된다. 이 모델을 적용하여 새로운 답안을 자동채점한 결과는 0과 5사이의 실수가 되는데 이 결과를 정수화하기 위해 필요에 따라 소수점에서 반올림하여 최종 점수를 산출한다. 일반적인 회귀 기법은 모델 생성에 사용될 데이터 내에 상관관계가 높은 자질들이 많거나 채점자질들의 개수가 많아지는 경우 모델의 정확도가 감소하는데, 이 기법은 데이터가 희소하거나 채점모델 생성에 사용한 자질들의 개수가 많은 고차원 데이터

(multi-dimensional data)의 경우에 대해 정확도가 높다는 장점이 있다.

2. 알고리듬 비교 선행 연구

상용화된 대표적인 영어 쓰기 자동채점 시스템으로는 PEG(Project Essay Grade; Page, 2003), ETS사의 e-rater(Attali, 2006), Pearson사의 IEA(Intelligent Essay Assessor; Dikli, 2006) 그리고 Vantage Learning사의 IntelliMetric(Vantage Learning, 2005)이 있다. 이 시스템들은 자연언어처리 분야의 다양한 기법을 활용하여 답안으로부터 채점을 위한 자질을 추출하고 이를 바탕으로 기계학습 과정을 통해 채점모델을 생성한다는 측면에서 국가영어능력평가시험 자동채점 시스템과 자동채점 과정이 매우 유사하다. 그러나 대상이 되는 시험(예, 국가영어능력평가시험, TOFEL 등)에 적합한 자질들을 어떻게 정의하고 구현할 것인가와 채점모델 생성을 위해 어떤 기계학습 알고리듬을 적용할 것인가에서 큰 차이가 존재한다. 본 연구에서는 자동채점 알고리듬 비교에 초점을 두고 있으므로 위에서 언급한 자동채점 시스템들이 도입한 기계학습 알고리듬에 대해 다음 〈표 1〉과 함께 소개하고자 한다.

〈표 1〉 대표적인 영어 쓰기 자동채점 시스템과 이들이 도입한 기계학습 알고리듬

자동채점 시스템	PEG	e-rater	IEA	IntelliMetric
기계학습 알고리듬	선형회귀	선형회귀	잠재 의미 분석	다중 모델 채택 (여러 알고리듬 사용)

먼저 PEG와 e-rater는 점수(종속변수)를 예측하기 위해 자질값(독립변수)들과 오차를 선형결합을 통해 모델로 생성하는 선형회귀(linear regression) 기법을 사용하고 있다. PEG는 1973년에 개발된 최초의 성공적인 자동채점 시스템으로 학생들의 에세이로부터 채점에 필요한 간단한 자질들을 추출하고 교사가 채점한 점수 간의 상관관계를 분석하여 선형회귀 기법을 통해 채점모델을 생성한다. 이 시스템은 채점 방법이 단순하고 특히 채점자질들의 추출에 있어 자연언어처리 기법을 활용하지 않고 단순히 통계처리 방법에만 의존하는 단점을 가지고 있다. e-rater(Attali, 2006)는 인간채점자들이 사용하는 것과 동일한 자질을 자동채점에 사용하도록 개발되었다. 따라서 구문의 다양성, 어휘 사용의 복잡성, 주제의 부합여부 등을 자연언어처리 기법을 활용하여 50여개의 자질정보로 추출하고 이들을 선형회귀 기법을 사용하여 채점 모델을 생성한다. 선형회귀 기법은 독립변수들이 비선형(nonlinear)의 관계를 갖는 경우 성능이 낮아질 수 있는 단점을 가지고 있다.

IEA(Dikli, 2006)는 응시자가 작성한 에세이를 잠재 의미 분석(Latent Semantic Analysis)

기법을 활용하여 답안 내 단어들 간, 단어와 답안 간, 답안들 간의 의미적 유사도(semantic similarity)를 구하여 채점하는 방식을 채택하고 있다. 즉, 새로운 답안의 채점을 위해 이 답안에 나타나는 단어들과 의미적으로 유사성이 가장 높은 단어들을 포함하고 있는 답안을 이미 채점이 완료된 답안들로부터 추출하여 그 답안의 점수를 예측점수로 제시하게 된다. 이 방식은 어순이나 문법적 관계 등에 기반한 내용 평가는 어렵기 때문에 자연언어처리 분야에서 주로 주제 분류에 사용되는 수준이고, 철자 및 문법 오류와 같은 요소들에 대한 평가가 어렵다. 한편 IntelliMetric은 응시자 답안으로부터 400여개의 자질을 선별하여 의미, 구문, 담화 단계에서 분석되어 내용과 구성의 범주에서 작문을 평가한다. 자동채점에 사용되는 기계학습 알고리듬을 한 가지로 규정하지 않고 선형 회귀 분석, 잠재 의미 분석, 베이시안(Bayesian) 기법 등을 적용한 여러 개의 채점모델을 생성하고 이들로부터 나온 값들을 활용하여 최종 점수를 생성하는 것으로 추정된다.

본 연구에서 사용할 ME 기법과 SVR 기법은 주로 학계에서 많이 활용되고 있는데, 이는 상용화되어 있는 시스템에서 사용하고 있는 기법들에 비해 상대적으로 최근에 나온 기계학습 알고리듬이기 때문이다. 이 두 가지 기법을 영어 쓰기 자동채점에 사용한 최근 연구로는 Sukkarieh (2010)의 연구와 Li와 Yen(2010)의 연구가 있다. 우선 Sukkarieh(2010)의 연구는 에세이의 내용 평가에 ME를 사용하였으며, 이는 단문의 내용 평가에 사용되는 ETS의 c-rater에서 도입된 기법인, 실험에서는 규칙 기반 기법에 비해 좋은 성능을 보였다. Li와 Yen(2012)의 연구에서는 에세이로부터 표층적 자질 및 문법 오류 및 내용 평가 등과 관련된 자질들을 추출하여 서포트 벡터 회귀 기법을 통해 자동채점 모델을 생성하여 실제 CET(Chinese English Test) 데이터에서 인간채점자의 점수와 86%의 정확도를 보였다. 영어가 모국어가 아닌 응시자들을 대상으로 수행하는 시험에 적용하여 높은 정확도를 보였다는 점에서 매우 의미 있는 연구라 할 수 있다.

Ⅲ. 모의실험 연구

1. 분석 방법

본 연구에서는 우선 알고리듬에 따른 자동채점의 성능을 비교하기 위해 모의실험을 실시하였다. 모의실험에 사용한 자료는 2011년도 예비평가 쓰기 2급의 2개 문항 유형별 채점 점수로, 일상생활 글쓰기 문항은 999개, 자기의견 쓰기 문항은 1000개의 채점 답안들로 구성되어 있다. 자료의 양호도를 나타내는 두 명의 인간채점 점수 간의 상관계수를 채점 영역별로 살펴보면 〈표 2〉와 같다.

〈표 2〉 인간채점 간 상관분석

문항 유형	주제	과제 완성	내용	구성	언어 사용
일상생활 글쓰기	만나고 싶은 인물	0.87	0.83	0.80	0.77
자기 의견 쓰기	학생 아르바이트에 대한 찬반	0.85	0.78	0.75	0.74

본 모의실험에서는 다음 〈표 3〉에 제시되어 있는 자동채점을 위한 옵션들과 성능 실험 방법을 사용하였다.

〈표 3〉 자동채점 옵션과 성능 실험 방법

분류			모의실험에 사용된 기법
자동채점 옵션	기계학습 알고리듬		ME, SVR
	채점모델 생성 방법		두 채점자 점수의 평균값으로 모델 생성
	전문가 선정 채점자질	과제완성	43개 (전체 단어 수, 특정 어구의 출현 회수 등)
		내용	33개 (고득점 답안과의 의미상 거리 등)
		구성	24개 (특정 담화표지 출현 회수 등)
		언어사용	23개 (철자 오류 개수, 문법 오류 개수 등)
	채점자질 선택 방법		전문가 선정 채점자질로 고정
	상관계수 비교 단위		소수점
성능 실험 방법			10-fold 교차분석

모의실험을 위한 조건들은 채점모델 생성 방법, 자질 선택 방법, 상관계수 비교 단위이며, 본 연구에서는 이러한 조건들을 동일하게 설정하고 ME 기법과 SVR 기법 간 비교를 위한 실험을 수행하였다.

채점모델 생성 방법은 두 채점자 점수의 평균값을 기준으로 각 영역별 채점모델을 생성하는 방법을 사용했다. 한 개의 답안에 대해 두 개의 채점자 점수가 존재하므로 원 점수들을 기반으로 2개의 채점모델을 생성한 후 자동채점을 수행할 때 각 모델을 통해 예측된 두 점수의 평균을 최종 점수로 선택하는 방법도 있다. 그러나 이 경우는 두 채점자 점수의 평균값을 기준으로 한 개의 채점모델을 생성하는 방법에 비해 정확도 측면에서는 큰 차이가 없으나 속도 면에서 느리다는 단점이 있어 본 연구에서는 선택하지 않았다.

채점모델의 생성에 사용되는 채점자질과 관련해서 전산학의 자연언어처리 분야의 전문가들이 약 100개의 상위 자질들을 정의하여 구현했고, 채점자들로 구성된 영어평가 전문가 그룹에 의해 이들 중 실제 자동채점에 사용될 영역별 채점자질들이 선정되었다. 과제완성 영역의 경우를 예로 들면, 지문에 제시된 단어 수나 주어진 과제에 대한 내용이 포함되도록 답안을 작성했는지

여부가 채점기준에 포함되어 있으므로 자연언어처리 전문가들은 단어 수를 값으로 갖는 자질이나 지문에 제시된 과제와 의미적인 관련성을 가지는 표현들이 답안에 나왔는지 여부를 확인하는 자질들을 컴퓨터가 이해하고 처리할 수 있도록 수식이나 알고리듬 형태로 정의하여 생성하였다. 이렇게 생성된 자질들은 영어평가 전문가들의 검토를 통해 각 영역별 채점에 영향력이 있을 것으로 판단되는 영역별 자질들로 분류되었다.

상관계수 비교 단위는 채점모델 생성 방법과 예측된 점수를 소수점 또는 정수로 생성하는지의 여부와 관련되어 있다. 본 연구에서는 두 채점자 점수의 평균값을 기준으로 각 영역별 채점모델을 생성하는 방법을 채택하고 있으므로 채점모델의 정확도를 정보의 손실 없이 비교하기 위해서는 기준값이 되는 두 채점자 점수의 평균과 예측 결과 간의 상관계수를 구해야 한다. 따라서 두 채점자 점수의 평균이 0.5 단위의 소수점 형태이고 예측할 점수도 동일한 형태로 생성해야 하므로 상관계수 비교 단위는 소수점 수준이 된다. 향후에 자동채점을 실제 적용하게 될 때에는 정수 형태의 점수를 응시자에게 제공해야 하므로 소수점 형태의 점수를 반올림하거나 답안별로 정수 형태의 채점결과를 갖는 데이터를 생성하여 기계학습을 시킴으로써 정수 형태의 결과물을 산출하는 채점모델을 생성하는 방식을 고려하는 것이 필요하다.

실증 데이터만을 가지고 알고리듬의 성능을 비교할 경우 그 결과의 일반화에 한계가 있으므로 본 연구에서는 이러한 한계점을 극복하기 위해 데이터를 분할하여 다양한 조합의 데이터를 사용하여 성능을 비교하는 n-fold 교차실험을 선택했다. n-fold 교차실험은 채점자 점수가 부여된 채점 데이터만 가지고 자동채점의 정확도를 분석하기 위해 주로 사용되는 실험 방법이다. 먼저 채점 데이터를 n개의 세트로 나누어 n-1개 세트로는 채점모델을 생성하고 나머지 1개 세트는 이 채점모델에 적용하여 자동채점을 수행한다. 이렇게 예측된 자동채점 결과와 채점자 점수 간의 상관계수를 비교하게 되며, n-fold 방법에서는 이러한 실험을 데이터 세트를 바꿔가며 n번 반복한다. 본 연구에서는 약 1000개의 데이터를 10-fold 교차실험을 실시하였으며, 이를 위해 100개씩 10개의 데이터 세트(S1, S2, …, S10)를 생성한 후, 첫 번째 실험에서는 900개 데이터 {S1, S2, S3, …, S9}로 채점 모델을 생성하고 S10에 적용하여 S10에 대한 자동채점 결과를 얻고, 두 번째 실험에서는 {S1, S2, S3, …, S8, S10}로 채점모델을 생성하고 S9에 적용하여 S9에 대한 자동채점 결과를 얻는 방식으로 모의실험을 진행하였다. 이러한 실험을 10번 반복하여 S1, S2, …, S10에 대한 자동채점 결과를 얻게 되고, 이 결과와 인간채점자의 채점 결과 간의 상관계수를 분석하였다.

2. 분석 결과

ME 기법과 SVR 기법을 각각 적용하여 10-fold 교차 실험을 한 결과는 〈표 4〉와 같다.

〈표 4〉 기계학습 방법의 변경에 따른 상관계수의 변화

데이터	구분		채점 영역				평균
			과제 완성	내용	구성	언어 사용	
만나고 싶은 인물	인간채점자 간		0.87	0.83	0.80	0.77	0.82
	인간채점과 자동채점	ME	0.77	0.85	0.82	0.81	0.81
		SVR	0.84	0.88	0.84	0.81	0.84
학생 아르바이트에 대한 찬반	인간채점자 간		0.85	0.78	0.75	0.74	0.78
	인간채점과 자동채점	ME	0.76	0.83	0.78	0.82	0.80
		SVR	0.77	0.85	0.81	0.82	0.81

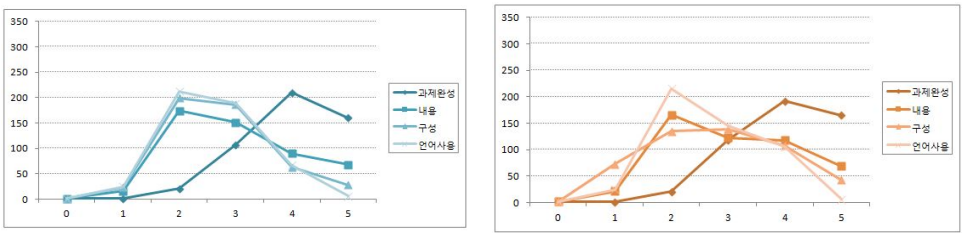
〈표 4〉에서 보여주듯이 "만나고 싶은 인물"에 대한 인간채점자 간 상관계수를 분석한 결과 4개 채점 영역의 평균은 0.82 정도로 나타났다. 이러한 채점 데이터를 이용하여 10-fold 모의실험을 통해 ME 알고리듬의 성능을 분석해 본 결과, 4개 채점 영역의 평균 상관계수는 0.81로 인간채점자 간 상관계수보다 다소 낮게 나타났으나 채점 영역별로 살펴보면 과제 완성 영역을 제외한 모든 채점 영역에서 인간채점자 간 상관계수보다 높다는 것을 알 수 있다. SVR 알고리듬을 적용한 10-fold 모의실험 결과에서도 과제 완성 영역을 제외한 모든 영역에서 인간채점자 간 상관계수보다 높은 상관계수를 보여주었으며, ME 알고리듬을 이용한 모의실험결과와 비교하면 모든 채점 영역에서 SVR을 적용했을 때 자동채점의 성능이 다소 향상되는 것으로 확인되었다.

한편 "학생 아르바이트에 대한 찬반" 문항의 4개 채점 영역에 대한 인간채점자 간 상관계수 평균은 0.78 이었으며, ME 알고리듬을 적용하여 10-fold 교차실험을 수행한 결과, 채점 영역별인간채점과 자동채점 간 상관계수의 평균은 약 0.80으로 나타나 인간채점자 간 상관계수의 평균과 유사한 결과를 얻을 수 있었다. 채점 영역별로 살펴보면 과제 완성 영역을 제외한 내용, 구성, 언어 사용 영역에서 인간채점자간 상관계수보다 높은 상관계수를 보여주었다. 이와 같은 결과는 ME 알고리듬을 적용하여 자동채점을 수행할 경우 과제 완성을 제외한 모든 영역에서 인간채점보다 더 신뢰할 만한 채점 결과를 얻을 수 있음을 보여준다. 한편 SVR 알고리듬을 적용할 경우에는 4개 채점 영역에 대한 상관계수의 평균이 0.81로서 ME 알고리듬을 적용했을 때보다 상관계수가 더 높게 나타났다. 채점 영역별로 살펴보면 ME 알고리듬을 적용한 경우와 마찬가지로 과제 완성 영역을 제외한 모든 채점 영역에서 인간채점자 간 상관계수보다 높게 나타났으며, ME 알고리듬을 적용한 자동채점보다 모든 영역에서 다소 높은 상관계수를 보여주고 있음을 알 수 있다. 이상의 모의실험 결과를 종합하면, ME 알고리듬과 SVR 알고리듬 모두 과제 완성 영역을 제외한 모든 채점 영역에서 인간채점에 비해 채점의 신뢰도를 향상시키는 것으로 확인되었으며, ME와 SVR 중에서는 SVR이 자동채점의 성능을 보다 향상시키는 것으로 나타났다.

Ⅳ. 실증 연구

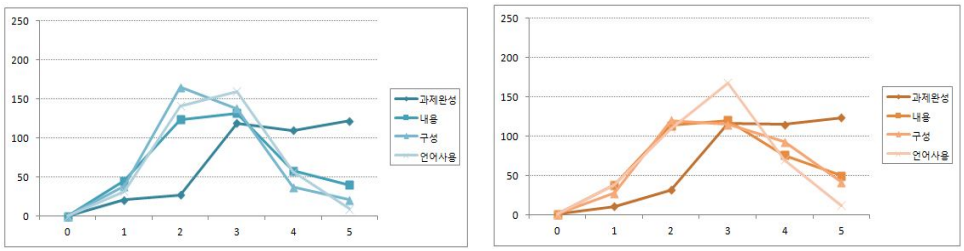
1. 분석 방법

ME와 SVR 알고리듬의 양호도를 분석하기 위하여 두 명의 인간채점자가 채점한 점수의 평균값 자료를 활용하여 기계학습을 실시한 후, ME와 SVR 알고리듬을 적용하여 자동채점한 점수와 개별 인간채점자가 채점한 점수 간의 상관을 살펴보았다. 비교를 위해 두 명의 인간채점자 점수 간의 상관도 함께 제시하였다. 본 연구에서 사용된 2011년 5월 예비평가 1번 문항과 2번 문항에 대해 각각 500개 답안에 대한 채점 점수가 사용되었으며 채점 영역별, 점수 구간별 빈도를 요약하면 〔그림 3〕, 〔그림 4〕과 같다.



인간채점자 1 인간채점자 2

〔그림 3〕 인간채점자 1, 2의 점수 분포: 2011년 예비평가 1번 문항



인간채점자 1 인간채점자 2

〔그림 4〕 인간채점자 1, 2의 점수 분포: 2011년 예비평가 2번 문항

2011년 5월 예비평가 쓰기 2급 1번 문항과 2번 문항에 대한 채점 영역별, 점수 구간별 빈도 분석 결과, 인간채점자 1의 내용, 구성, 언어 사용 영역의 점수대별 분포는 2-3점에 집중적으로 분포한 반면, 과제 완성 영역에서는 3점, 4점, 5점에 대부분 분포하고 있는 것으로 나타났다. 이와 같은 현상은 내용, 구성, 언어 사용 영역보다 과제완성에서 점수를 받기 쉬움을 의미하는

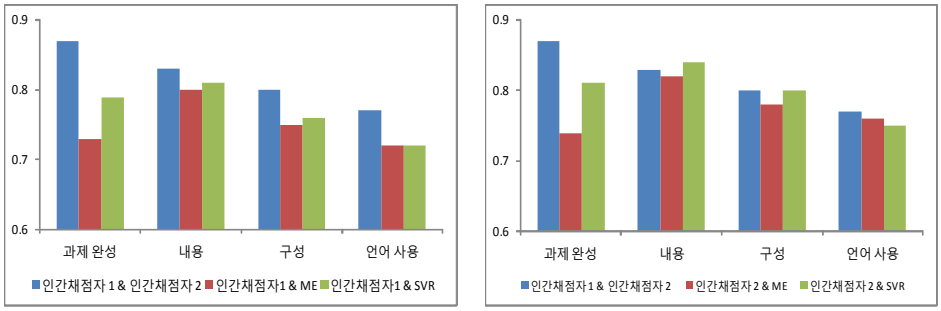
것이다. 과제완성은 문항에서 요구되는 과제에 대한 달성 정도를 평가하는 영역이다. 즉, 과제완성에서는 문항에서 요구하는 최소 단어수를 사용하여 답안을 작성하였는지, 주어진 조건이나 근거를 모두 사용해서 답안을 작성하였는지 평가한다. 이렇듯, 요구되는 과제가 단어 수 충족, 주어진 정보나 조건의 포함 여부 등 매우 단순하기 때문에 과제완성에서는 다른 채점 영역에 비해 수험생들이 쉽게 과제를 달성 수 있다. 따라서 내용의 구체성이나 충실성, 글의 응집성이나 논리성, 사용된 문법이나 표현의 정확성이 다소 부족하더라도, 단어 수에 대한 조건을 만족 시켰거나, 요구되는 정보를 포함하기만 하면 과제가 달성된 것으로 간주되기 때문에, 다른 영역에 비해 상대적으로 높은 점수를 받을 수 있다. 이와 같은 경향은 인간채점자 2의 채점 점수, 채점자 평균에서도 유사하게 나타나고 있다.

2. 분석 결과

ME와 SVR의 성능을 비교하기 위해 두 가지 알고리듬을 적용한 자동채점 결과와 인간채점 결과 간 상관을 비교하였다. 우선 채점 영역별로 두 명의 인간채점자(인간채점자 1, 인간채점자 2) 간 상관계수와 개별 인간채점자와 자동채점 간 상관계수를 비교하였다. 2011년 5월 예비평가 1번 문항과 2번 문항에 대한 상관분석 결과를 요약하면 〈표 5〉, 〈표 6〉과 같으며, 〔그림 5〕와 〔그림 6〕에서는 알고리듬에 따른 상관계수 차이를 시각적으로 비교하였다.

〈표 5〉 인간채점과 자동채점에 따른 상관계수 비교: 2011년 예비평가 1번 문항(만나고 싶은 인물)

채점 영역	인간채점자 1 & 인간채점자 2	인간채점자 1 & ME	인간채점자 1 & SVR	인간채점자 2 & ME	인간채점자 2 & SVR
과제 완성	0.87	0.73	0.79	0.74	0.81
내용	0.83	0.80	0.81	0.82	0.84
구성	0.80	0.75	0.76	0.78	0.80
언어 사용	0.77	0.72	0.72	0.76	0.75

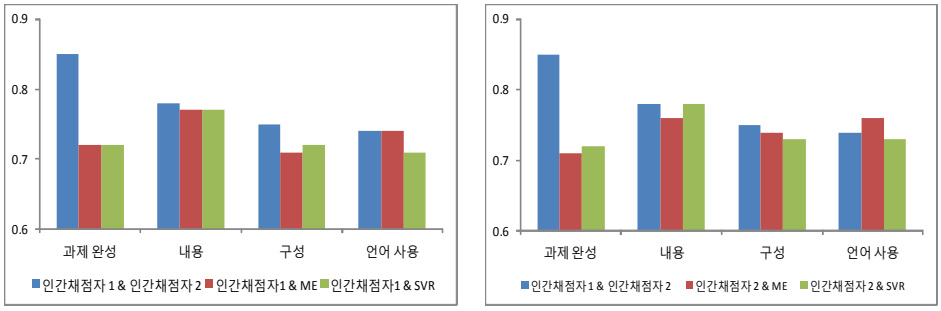


〔그림 5〕 알고리듬에 따른 인간채점과 자동채점 상관계수: 1번 문항

2011년 5월 예비평가 1번 문항에 대한 상관분석 결과 인간채점자 점수 간의 상관은 채점 영역에 따라 0.77~0.87로 나타났으며, 인간채점자 1과 ME에 의한 점수 간의 상관은 0.72~0.80, 인간채점자 2와 ME에 의한 점수 간의 상관은 0.74~0.82로 나타났다. 또한 인간채점자 1과 SVR에 의한 점수 간의 상관은 0.72~0.81, 인간채점자 2와 SVR 에 의한 점수 간의 상관은 0.75~0.84로 나타났다. 채점 영역별 결과를 살펴보면 과제 완성, 내용, 구성 영역에서는 인간채점자와 SVR 에 의한 점수 간의 상관계수가 인간채점자와 ME에 의한 점수 간의 상관계수보다 높았으며, 언어 사용에서는 인간채점자와 ME에 의한 점수 간의 상관계수가 인간채점자와 SVR에 의한 점수 간의 상관계수보다 높게 나타났다. 이와 같은 현상은 2번 문항에 대한 분석에서도 동일하게 나타났으며 그 결과는 〈표 6〉과 〔그림 6〕에서 확인할 수 있다.

〈표 6〉 인간채점과 자동채점에 따른 상관계수 비교: 2011년 예비평가 2번 문항(학생 아르바이트 찬반)

채점 영역	인간채점자 1 & 인간채점자 2	인간채점자 1 & ME	인간채점자 1 & SVR	인간채점자 2 & ME	인간채점자 2 & SVR
과제 완성	0.85	0.72	0.72	0.71	0.72
내용	0.78	0.77	0.77	0.76	0.78
구성	0.75	0.71	0.72	0.74	0.73
언어 사용	0.74	0.74	0.71	0.76	0.73

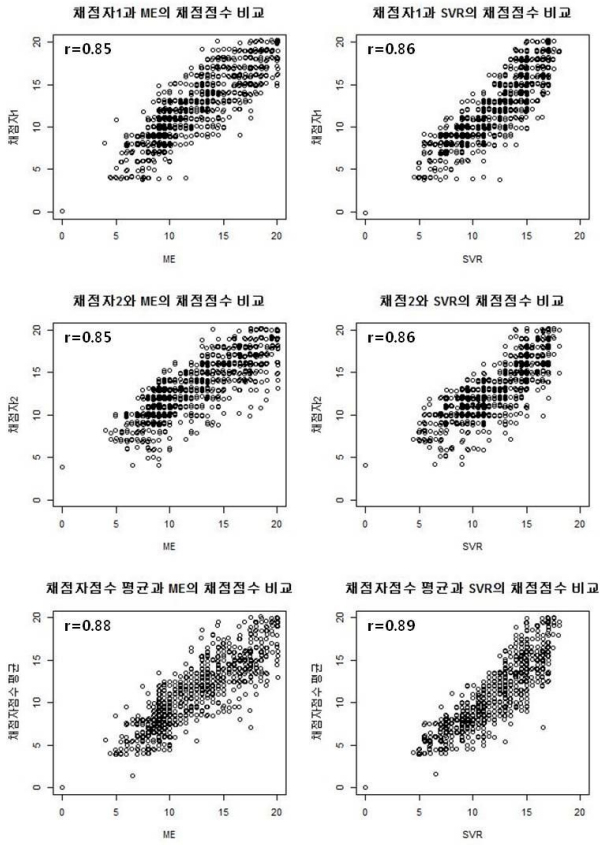


〔그림 6〕 알고리듬에 따른 인간채점과 자동채점 상관계수: 2번 문항

2011년 5월 예비평가 2번 문항에 대한 상관분석 결과 인간채점자 점수 간의 상관은 채점 영역에 따라 0.74~0.85로 나타났으며, 인간채점자 1과 ME에 의한 점수 간 상관은 0.71~0.77이며, 인간채점자 2와 ME에 의한 점수 간의 상관은 0.71~0.76으로 나타났다. 한편 인간채점자 1과 SVR 에 의한 점수 간의 상관은 0.71~0.77, 인간채점자 2와 SVR에 의한 점수 간의 상관은 0.72~0.78로 나타나 알고리듬에 따른 상관계수 차이가 크지 않았다. 그러나 채점 영역별로 살펴본 결과 과제 완성과 내용 영역에서는 인간채점자와 SVR에 의한 점수 간의 상관이 인간채점자와 ME 에 의한 점수 간의 상관과 같거나 높게 나타난 반면, 언어사용 영역에서는

반대로 인간채점자와 ME에 의한 점수 간의 상관이 인간채점자와 SVR에 의한 점수 간의 상관보다 높게 나타났다. 이와 같은 결과는 알고리듬에 따른 자동채점의 성능이 채점 영역에 따라 변화될 수 있음을 보여주는 것으로, 향후 채점의 공신력을 제고하기 위해 자동채점 알고리듬을 채점 영역에 따라 달리 적용하는 방안에 대한 시사점을 얻을 수 있었다. 또한 과제 완성 영역의 인간채점자 간 상관이 가장 높았으나 알고리듬에 관계없이 자동채점과 인간채점자 간 상관이 가장 낮은 것으로 나타나고 있어, 과제 완성 영역에서 자동채점이 타 채점 영역에 비해 가장 성능이 좋지 않은 것으로 드러났다. 이와 같은 결과는 과제 완성 영역에서 채점을 위해 요구되는 조건이 많은 동시에 조건을 충족시켰는지 여부를 판단하는 기준이 인위적인 경향이 있어 자동채점에 이러한 기준을 구현하는 것이 한계가 있기 때문인 것으로 판단된다.

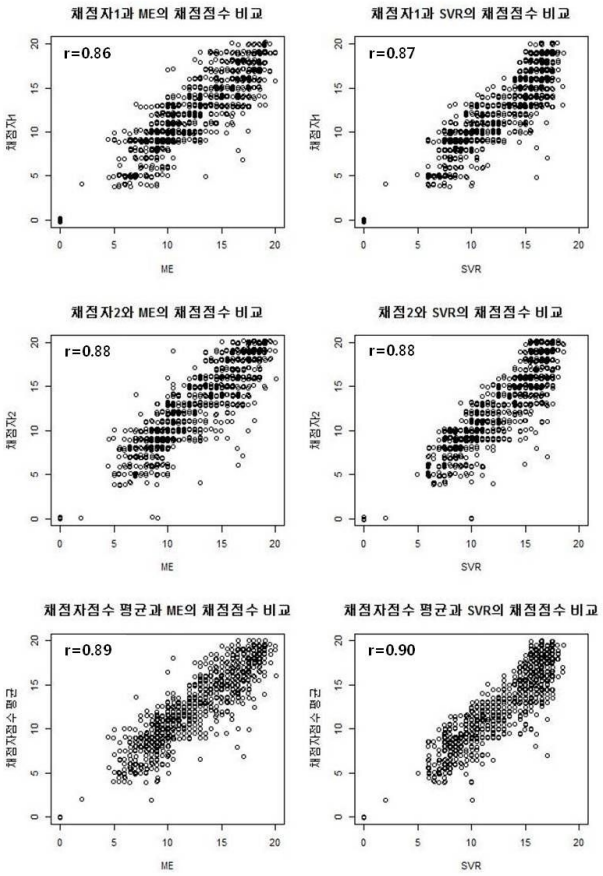
다음으로 두 가지 알고리듬을 적용하여 채점 영역별로 점수를 부여한 결과를 합산한 문항별 점수와 인간채점자의 문항별 점수 간 상관분석을 통해 다시 한 번 알고리듬에 따른 자동채점의 성능 차이를 검증하였다. 〔그림 7〕과 〔그림 8〕은 2011년 5월 예비평가 1번과 2번 문항에 대한



〔그림 7〕 인간채점과 자동채점(정수)에 따른 점수 분포 비교: 2011년 예비평가 1번 문항

인간채점자 점수와 ME와 SVR에 의한 문항별 총점의 상관관계를 보여 주는 산포도이다. 우선 문항 1번에 대한 인간채점자 점수와 ME와 SVR에 의한 점수 간 산포도를 보여주는 〔그림 7〕은 일관되게 SVR에 의해 산출된 점수가 ME에 의해 산출된 점수보다 인간채점자 점수에 더 근접한 점수를 산출하는 것으로 보여주고 있다. 인간채점자 1과 인간채점자 2, 그리고 두 명의 인간채점자 점수의 평균과 ME에 의해 산출된 점수 간의 상관계수는 각각 0.85, 0.85, 0.88인 반면 SVR에 의해 산출된 점수와의 상관계수는 0.86, 0.86, 0.89로 ME에 의한 결과와 비교할 때 일관되게 높은 상관계수를 나타내고 있다.

1번 문항에서 확인된 ME에 대한 SVR의 비교 우위는 2번 문항에서도 확인할 수 있었다. 〔그림 8〕에서 보여주듯이 인간채점자 1, 인간채점자 2, 인간채점자 평균과 ME 간 상관계수는 0.86, 0.88, 0.89로 확인되었으며, SVR과의 상관계수는 0.87, 0.88, 0.90으로 ME와 같거나 높은 것으로 나타났다.



〔그림 8〕 인간채점과 자동채점(정수)에 따른 점수 분포 비교: 2011년 예비평가 2번 문항

Ⅴ. 결론 및 제언

국가영어능력평가시험에서는 채점의 효율성과 공신력을 제고하기 위해 자동채점 프로그램을 개발 중에 있다. 자동채점 프로그램은 학생들의 답안들로부터 추출된 자질들과 답안에 대한 점수들 간의 관계를 확률 및 통계이론을 기반으로 분석하는 학습 과정을 거쳐 채점모델을 생성하며, 자동채점에 주로 사용되는 기계학습 알고리듬은 분류분석에 기반한 방법과 회귀분석에 기반한 방법이 있다. 본 연구에서는 국가영어능력평가시험 쓰기 채점 데이터에 적합한 방법을 찾기 위해 분류분석에 기반한 방법의 대표적인 기법인 ME 기법과 회귀분석에 기반한 방법의 대표적인 기법인 SVR 기법을 이용하여 채점모델을 형성하여 자동채점을 실시하고 그 결과를 비교·분석하였다. ME 기법은 주어진 자질들을 기반으로 해당 답안이 받을 수 있는 점수들에 대해 확률을 계산하고, 가장 확률이 높은 점수를 부여하는 방법이며, SVR 기법은 자질 값과 답안의 점수 간의 관계에 대한 정보 함수를 구현하고, 주어진 자질값들을 기반으로 정보 손실을 최소화할 수 있는 점수를 예측하는 방법이다.

현재 국가영어능력평가시험 쓰기 2급 자동채점 프로그램에서는 두 가지 기법들을 적용하여 프로그램을 개발하고 있으며, 본 연구에서는 자동채점 프로그램에 적용되고 있는 ME와 SVR 알고리듬의 성능을 비교 분석하여, 국가영어능력평가시험 쓰기 자동채점 프로그램에 적합한 알고리듬을 제안하고자 하였다. 이를 위해 n-fold 방식을 이용한 모의실험과 2011년 5월 예비평가 쓰기 2급 시험 채점 자료를 이용하여 두 가지 알고리듬의 성능을 비교하였다. n-fold 방식을 적용한 모의실험 결과 ME보다 SVR 알고리듬을 적용할 경우 자동채점의 성능이 보다 향상되는 것으로 나타났으며, 이와 같은 결과는 2011년 5월 예비평가 자료를 활용한 ME와 SVR 간 비교에서도 동일하게 나타났다. 예컨대, 2011년 5월 예비평가 쓰기 2급 문항 채점 자료를 이용하여 ME와 SVR의 성능을 비교한 결과, SVR이 ME에 비해 인간채점자 간의 상관계수와 더 근접한 결과를 산출하는 것으로 확인되었다. 그러나 채점 영역별로 살펴보면 언어 사용 영역에서는 ME가 SVR에 비해 더 우수한 알고리듬이 될 수 있는 가능성을 보여주었다. 이는 채점 영역의 특성에 따라 알고리듬의 우수성에 차이가 있을 수 있음을 보여주는 것으로, 향후 자동채점 알고리듬 관련 연구에서 이에 대해 심층적으로 검토해 볼 필요가 있다.

이상의 연구 결과에 기초할 때, 국가영어능력평가시험 쓰기 2급에서 인간채점자 1명의 점수를 자동채점으로 대체할 경우 ME보다는 SVR을 적용하는 것이 더 적합한 것으로 판단된다. 그러나 이와 같은 결과를 일반화시키기 위해서는 보다 다양한 유형의 문항에 대한 채점 자료에 적용해볼 필요가 있으며, 채점 영역의 특성이 알고리듬의 성능에 미치는 영향에 대해서도 지속적으로 검증할 필요가 있다. 이를 통해 자료의 특성에 최적화된 채점 알고리듬을 선별적으로 적용할 수 있는 근거 자료를 도출할 수 있을 것이다.

참 고 문 헌

시기자, 박도영, 이용상, 박상욱, 임은영, 구슬기, 임황규, 최연희, 이공주, 김지은, 김성, 이은숙, 김성묵, 윤경아, 이순웅(2012). 국가영어능력평가시험 쓰기 자동채점 프로그램 개발. 한국교육과정평가원 연구보고 RRE 2012-10.

신동광, 민호기, 박상복, 정채관, 주헌우, 김미지, 김연희, 이은숙, 김동남, 김영준(2012). 국가영어능력평가시험 말하기 자동채점 프로그램 도입 방안. 한국교육과정평가원 연구보고 RRE 2012-9.

Adam, L. Berger, Stephen, A., Pietra, D., & Vincent, J. (1996). A maximum entropy approach to natural language processing. *Journal of Computational Linguistics, 22*(1), 39-71.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment, 4*(3), 1-30.

Cortes, C., & Valpnik, V. (1995). Support vector networks. *Machine Learnings, 20,* 273-297.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1). Retrieved from http://www.jtla.org(2012. 5. 21).

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157-1182.

Li, Y., & Yan, Y. (2012). *An effective automated essay scoring system using support vector regression.* International Conference on Intelligent Computation Technology and Automation.

Nigam, K., Lafferty, J., & MacCallum, A. (1999). Using maximum entropy for text classification. *IJCAI Workshop on Machine Learning for Information Filtering*, 61-67.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective,* 43-54. Mahwah, NJ: Lawrence Erlbaum Associates.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of Database Systems,* 532-538.

Smola, A. J., & Scholköpf, B. (2004). A tutorial on support vector regression. *Journal of Statistics and Computing, 14*(3), 199-222

Sukkarieh, J. Z. (2010). *Maximum entropy for the automatic content scoring of free-text responses.* Proceedings of the 30th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering.

Thomas, K. L. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10*(3), 295-308.

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education, 2.*

Vantage Learning (2005). *How IntelliMetric Works* [On-line]. Retrieved from http://www.vantagelearning.com/docs/intellimetric/IM_How_IntelliMetric_Works.pdf (2012. 5. 21).

Vapnik, V., Steven, E., Golowich, & Smola, A., J. (1997). Support vector method for function approximation: Regression estimation and signal processing, *Advances in Neural Information Processing System 9*, MIT, 281-287.

• *논문접수 : 2013-08-27/ 수정본접수 : 2013-09-30/ 게재승인 : 2013-10-16*

ABSTRACT

The comparison of performances of the automated scoring algorithms: the maximum entropy and support vector regression methods

Yong-Sang Lee
(Korea Institute for Curriculum and Evaluation)
Ki-Ja Si
(Korea Institute for Curriculum and Evaluation)
Do-Young Park
(Korea Institute for Curriculum and Evaluation)
Kyuing-A Yoon
(SK Telecom)
Seul-Ki Koo
(Korea Institute for Curriculum and Evaluation)
Hwang-Kyu Lim
(Korea Institute for Curriculum and Evaluation)

Korea Institute for Curriculum and Evaluation is developing an automated scoring system for the writing test that is optimized to the National English Ability Test(NEAT). It is expected that the performance of an automated scoring system is influenced by the algorithms to be applied to machine learning. The current automated scoring system for the NEAT adopts the maximum entropy and support vector regression algorithms, which are most commonly used. In this paper, we analyzed and compared the performances of two algorithms that have been applied to the automated scoring system for the NEAT. For this purpose, we conducted a simulation study using the n-fold method and empirical study using the actual data set. Our study supports that the support vector regression method is performing better than the maximum entropy method in the writing test of the NEAT in both empirical and simulation studies.

Key Words : automated scoring algorithm, maximum entropy, support vector regression, n-fold simulation study