# 한국어 학습자 작문 자동 평가를 위한 평가 항목 선정

곽용진○
㈜이르테크
silhuett@iirtech.co.kr

# Evaluation Category Selection For Automated Essay Evaluation of Korean Learner

Yong-Jin Kwak○
IIRTECH Inc

**요 약**

본 연구는 한국어 학습자 작문의 자동 평가 시스템 개발의 일환으로, 자동 평가 결과에 대한 설명과 근거가 될 수 있는 평기 기준 범주를 선정하기 위한 데이터 구축과 선정 방법을 제시한다. 작문의 평가 기준의 영역과 항목은 평가체계에 대한 이론적 연구에 따라 다양하다. 이러한 평가 기준은 자동 평가에서는 식별되기 어려운 경우도 있고, 각각의 평가 기준이 적용되는 작문 오류의 범위도 다양하다. 그러므로 본 연구에서는 자동 평가 기준 선정의 문제는 다양한 평가 기준에 중 하나를 선정하는 분류의 문제로 보고, 학습데이터를 구축, 기계학습을 통해 자동 작문 평가에 효과적인 평가 기준을 선정 가능성을 제시한다.

주제어: 한국어교육, 작문 자동 평가,

## 1. 서론

자연언어처리 기술을 이용해 사람의 작문 결과를 평가하고자 하는 시도는 종종 진행되어 왔다. 그러나 자연언어처리 기술의 정확성이 사람의 언어 능력에 비해 큰 차이가 있어 널리 적용되지는 못했다. 영어권에서는 미국의 ETS(Educational Test Service)는 오랜 기간 작문 자동 평가(Automatic Essay Evaluation) 기술을 개발해 인간 평가 전문가 평가결과와의 유사도를 70% 수준까지 달성하였다.[1] 이러한 성과는 ETS가 보유한 방대한 언어 자원과 평가 인프라 뿐만 아니라, 다양한 평가 요소에 대한 자동 평가 가능성과 방법에 대한 시도로부터 얻어진 결과이다. 그 목적은 평가 주체가 누구(인간 또는 컴퓨터)인가의 문제가 아니라, 일관되고 투명한 평가 체계를 공개함으로써 신뢰성을 제고하는 데 있다.

국내에서도 [2],[3]에서 한국어 모국어 사용자의 작문 평가를 위한 시스템이 개발된 바 있다. 그러나, 이는 제한된 문항과 형식을 전제로 한 대규모 자료처리를 목적으로 한다. 반면에 한국어 교육 분야에서 작문의 평가는 외국어로서의 한국어 능력을 평가하는 것으로 한국어 교육의 확대와 함께 평가 항목과 기준에 있어서도 다양한 방법이 제시되고 있다. 최근 20 여년 동안 한국어에 대한 세계인의 관심이 높아짐에 따라 TOPIK 등의 평가 체계와 객관성, 신뢰성에 제고에 대한 논의도 확대되고 있다.[4]

본 연구에서는 한국어 학습자 작문의 오류와 한국어 교육에서 널리 통용되는 평가범주와 오류 유형을 한국어 교사에게 제공하고, 각각의 오류에 대해 적합한 범주와 유형을 선택하도록 하여 그 자료 분포를 분석한다. 이러한 평가범주 및 오류 유형 레이블 데이터는 기계학습을 이용한 자동 평가 시스템 개발 뿐만 아니라, 한국어 학습자의 오류에 대한 적절한 평가 기준을 설정하는 데 기여할 수 있다.

## 2. 관련 연구

ETS는 Grammar, Usage, Mechanics, Style의 4개 영역, 34개 세부항목에 대해 평가를 수행한다.[1] 세부항목은 Subject-Verb Agreement 등과 같은 영어 특유의 문법 기준 뿐만 아니라, 마침표, 쉼표, 물음표 누락과 같은 기초 정서법을 포함하고 있다.

반면에 [3]에서는 학습 데이터 구축을 배제하고, 고득점자 답안으로부터 기능어를 제외한 어휘집합의 군집화를 통해 개념답안을 생성하여 채점자질(평가기준)으로 활용하였다. 이러한 접근은 전문가에 의한 학습 데이터 구축이 최소화되는 데 반해, 평가 결과에서 대한 설명적 근거 제시가 어렵다는 단점이 있다.

## 3. 연구 방법

한국어 학습자의 작문을 자동 평가하는 과정은 크게 2개의 기능으로 구분할 수 있다. 하나는 학습자의 오류를 식별하는 과정이고, 다른 하나는 식별된 오류의 범주가 무엇인지 판정하는 하는 과정이다. 전자는 입력 어절이 오류인지 아닌지를 추정하는 방법이고, 후자는 현재 어절이 오류일 때 그 범주가 어디에 속하는지를 결정하는 분류의 방법이다.

본 연구는 학습자 작문의 오류를 다수의 한국어 교사에게 제공하고, 그 범주를 결정하도록 한 데이터를 수집, 분석한다. 두 가지 과정을 분리함으로써 학습자의 평가자의 한국어 능력에 내재된 평가기준에 의한 편향을

성을 줄이고, 제시된 오류에 대한 평가(판정)의 기준(명목)이 무엇인지에 집중하도록 한다. 이를 위해 다음과 같은 도구를 제공함으로써 수집과 분석의 효율을 높인다.

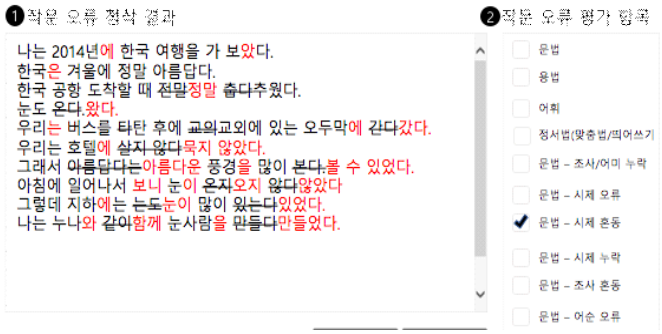


그림 1 작문 오류에 대한 평가 범주 데이터 수집도구

그림 1의 도구는 ①의 화면에 이미 수집된 학습자 작문의 첨삭 결과 자료를 출력한다. 한국어 교사는 ①의 출력 결과중 붉은 색으로 표현된 항목을 클릭하고, ②에서 클릭된 항목에 대한 오류의 평가 기준을 선택한다. ②의 목록에 평가 기준이 없는 경우, 기타를 선택하고 평가 기준명을 작성하여 등록할 수 있다. ①의 화면에 미리 수집된 작문 첨삭 결과를 제공함으로써 한국어 교사는 작문의 오류 여부가 아니라 해당 오류의 평가 기준에 어떤 범주가 적용되어야 하는지만 판단한다.

## 4. 수집 및 결과

데이터 구축은 그림 1의 도구를 이용하여 한국어 교사 10명에게 각각 100개의 작문 첨삭 결과를 제공하여 구축하였다. 구축된 데이터는 다음과 같다.

표 1 작문 오류 평가 기준 학습데이터 구축 결과

| 어절 번호 | 오류어절 | 교정어절 | 평가 범주1 (철자법) | 평가 범주2 (정서법) | 평가 범주... | 평가 범주48 (문법-시제오류) |
|---|---|---|---|---|---|---|
| 1 | 간다 | 갔습니다 | 0 | 8 | ... | 792 |
| 2 | 가리킸니다 | 가르칩니다 | 38 | 51 | ... | 0 |
| 3 | 갓수를 | 가수를 | 91 | 157 | ... | 0 |
| 4 | 계절이 | 계절이 | 108 | 22 | ... | 0 |
| ... | ... | ... | ... | ... | ... | ... |

표 1은 구축 결과에 대해 오류어절과 교정어절이 동일한 항목별로 한국어 교사가 선정한 평가기준을 집계하여 빈도를 산출하였다. 제시된 평가 기준 범주는 철자, 용법, 표현, 문법의 4개 영역, 30개 세부항목이었으나, 최종 수집된 평가 기준 범주는 총 48개 항목으로 증가하였다. 구축된 결과는 Word2Vec 기법을 이용하여 각 평가 기준 범주에 대해 Softmax로 평가하도록 하였다. 오류 어절은 형태분석을 적용하기 어려우나, 교정어절을 보편적인 한국어 형태 분석이 가능하므로 교정어절의 형태분석 결과를 문맥정보로 제공하였다. skip-gram의 윈도우 크기는 교정어절 앞뒤 어절의 1까지도 동적할당 되도록 한 경우의 성능(학습효율)이 가장 좋았다. 또한, 학습된 자동평가 모델에서 평가 범주의 변별력은 표 2와 같다.

표 3 평가범주별 오류 범주 분류 변별력

| 평가범주 | Precision | Recall | F1 |
|---|---|---|---|
| 철자법 | 0.67 | 0.82 | 0.737 |
| 담화-문/구어 구분 | 0.631 | 0.76 | 0.689 |
| 담화-담화표지 | 0.643 | 0.712 | 0.675 |
| 문법-조사 용법 | 0.6 | 0.52 | 0.557 |
| 문법-존대 | 0.42 | 0.63 | 0.504 |
| …. | … | … | … |
| 용법-문맥의미 | 0.16 | 0.48 | 0.24 |
| 문법-어미활용 | 0.14 | 0.27 | 0.184 |
| 어휘 철자법 | 0.13 | 0.25 | 0.171 |
| 내용-주제 완성도 | 0.14 | 0.12 | 0.129 |

철자법, 문법-시제, 문법-존대, 조사 용법, 조사 누락 등 범주는 자동 평가의 가능성이 높았다. 이러한 범주들은 구축결과에서 평가범주가 3개 이하로 나타난 것으로 총 21개 범주가 해당된다.

## 5. 결론 및 향후 과제

본 연구 결과 첨삭 결과 자료를 이용한 자동 평가의 구현이 어느 정도 가능함을 확인하였다. 그러나, 자동 평가 평가 범주의 정교화, 최적 평가 모델과 특성 정보의 조정 등 보다 많은 데이터 구축과 실험이 필요하다.

그러나, 학습자 작문의 오류 식별과 교정, 평가를 구분함으로써 각 모듈의 자동처리 성능대비 한국어 교사에 의한 사후 검토/교정 양이 감소량이 효과적으로 감소한다는 점에서 본 연구 결과의 의의가 있다.

## 참고문헌

[1] Charles A. MacArthur, Steve Graham, and Jill Fitzgerald, Handbook of Writing Research Second Edition, The Guilford Press, 2016.

[2] 노은희, 성경희, 임은영, "한국어 문장 수준 서답형 문항 자동채점 적용 가능성 탐색", 교육평가연구, 제28권, 제2호, pp. 523-551, 2015.

[3] 이경호, 이공주, "기계학습을 이용한 중등 수준의 단문형 영어 작문 자동 채점 시스템 구현", 정보과학회논문지 제41권 11호, pp.911-920, 2014.

[4] 이인혜, “한국어 쓰기 평가의 채점 방식에 따른 채점자 신뢰도 연구 : 종합적 채점 및 분석적 채점을 중심으로”, 고려대학교 대학원, 2012.