# 기계학습기법을 이용한 영어작문 문장 수준평가 시스템

엄진희*, 곽동민**
*고려대학교 컴퓨터 정보통신 대학원
**(주)아카북 AKA Artificial Intelligence Center
e-mail : arkaisally@gmail.com, gsclub@gmail.com

# A English Composition Level Assessment System Using Machine Learning Techniques

Jin-Hee Eom*, Dong-Min Kwak**
*Graduate School of Computer & Information Technology, Korea University
**AKABOOK Inc. AKA Artificial Intelligence Center

요 약

본 논문은 문장 내에서 나타나는 어휘간의 관계를 통해 표현 수준을 자동으로 평가할 수 있는 시스템을 제안한다. 제안하는 방법은 영어에세이 코퍼스 내의 문장에서 발생하는 철자 및 문법의 오류와 함께 어휘와 문법 패턴에 따른 표현난이도를 평가할 수 있는 자질을 생성하고 다양한 기계학습기법을 사용하여 문장의 수준을 평가하고자 하였다. 또한 기존에 연구되어온 규칙기반의 문장 평가시스템을 구현하고 기계학습기법을 이용한 문장 평가시스템과 비교하였다. 이를 통해 철자 및 문법의 오류율뿐만 아니라 표현난이도를 평가할 수 있는 자질들이 유용함을 확인할 수 있었다.

영어작문 문장의 수준평가를 위해서 국내 학생들의 토플 에세이 코퍼스를 수집하여 2,000문장을 추출하였고, 4명의 전문평가자들을 통해 6단계로 평가하여 학습 및 테스트 세트를 구성하였다. 성능척도로는 정확률과 재현율을 사용하였으며, 제안하는 방법으로 67.3%의 정확률과 67.1%의 재현율을 보였다.

## 1. 서론

현재 TOEFL, TOEIC 등의 공인어학시험은 수험자의 어학능력을 평가하기 위한 방법으로 영작문을 포함하여 시행하고 있다. 또한 국내 중학생 이상의 영어 교과시험에서는 의무적으로 30% 이상의 서술형 문제가 출제되고 있어 영작문에 대한 관심과 열정이 뜨겁게 이어지고 있다.

이러한 추세 속에 많은 수험자들은 인터넷 등을 통하여 영작문에 대한 교정을 전문가로 부터 받고 있다. 그러나 반복적이고 지속적인 훈련과 피드백을 통해서 배양되는 영작 실력은 그에 따른 평가 작업에 많은 시간과 비용이 소요된다는 문제점을 안고 있다. 그리고 이것은 곧 수험자의 부담으로써 전이되고 있다. 이에 2000년대 초반부터 ETS와 자연어처리 기반의 연구소 등에서는 자동평가시스템에 대한 연구가 진행되어 왔으며, 현재는 상용시스템으로까지 발전하여 사용되고 있다. [1]

지금까지 진행되어온 연구와 결과물들은 에세이 등과 같은 텍스트 작문의 경우 구조적이고 의미적인 특성을 통해 전반적인 평가가 진행되어온 반면 문장 단위의 자동평가는 철자 및 문법 오류와 같은 기계적인 측면에서의 평가가 주를 이루어왔다.

본 연구에서는 영어작문 문장의 수준을 평가함에 있어 문법지식, 어휘의 사용, 형식적 규칙-mechanics, 내용수행의 4개 항목을 기준으로 하였다.

이에 철자 및 문법 오류와 같은 기계적인 자질과 함께 어휘와 문법 패턴에 따른 표현난이도를 평가할 수 있는 자질을 새롭게 생성하여 다양한 기계학습알고리즘에 적용하여 실험하였다. 또한 성능비교를 위하여 기존에 연구되어온 규칙기반의 문장수준평가시스템을 구현하여 제안하는 방법의 유효성을 검증하였다.

2장에서는 기존 연구에 대한 조사를 실시하고 3장에서는 규칙기반 문장수준 평가시스템을 소개하고 구현하며, 4장에서는 표현난이도까지 고려한 다양한 자질을 생성하고, 기계학습기법을 적용하여 높은 성능의 상위 2개의 알고리즘을 선택한다. 이어지는 5장에서는 규칙기반시스템과 기계학습기반의 시스템을 비교·분석하고, 마지막 6장에서는 결론과 향후 연구 과제를 제시하도록 한다.

## 2. 관련연구

기존 연구들은 크게 에세이의 전체 텍스트 단위와 개별 문장 단위의 평가로 나누어 볼 수 있다.

에세이 전체 텍스트의 수준을 평가하기 위하여 전반적인 구조와 의미적인 특성을 바탕으로 자연어처리기술과 검색 및 클러스터링 등의 기술들이 활용되어 왔다.

자연어처리기술 기반의 평가방법으로는 “어휘 기반 계량 함수(lexical-based metrics)”를 이용한 PEG(The Project Essay Grade) 시스템(Page and Peterson 1995), 문법과 철자 검색기를 이용하는 방법(Park et al. 1997), 담화 분석을 이용하는 E-Rater 시스템(Miltsakaki and

Kukich, 2000) 등을 살펴볼 수 있다.[2][3]

정보검색이나 패턴인식, 클러스터링 등의 기술을 기반으로한 평가한 방법으로는 LSA(Latent Semantic Analysis)를 이용하는 IEA(The Intelligent Essay Assessor)시스템(Landauer et al., 2003), 베이지안 확률(Bayesian probability)을 이용하는 Rainbow 시스템(McCallum 1996), 하이브릿 방법의 CarmelTC 시스템(Rose, et al. 2003)등을 들 수 있다.[4]

에세이의 개별 문장을 평가하기 위한 방법으로는 철자 및 문법오류와 같은 기계적인 자질을 바탕으로 규칙기반의 시스템이 연구되었다.

Londsale and Strong-Krause(2003)은 Link-Grammar Parser를 이용하여 규칙기반의 문법적 오류 정도를 채점에 감안하는 방법을 제안하였다.[5]

최근 국내에서 김지은, 이공주(2007)는 EFL학습자인 한국의 중학생 영작문을 대상으로 문장 단위의 구문오류 분석을 위주로 하는 규칙기반의 시스템을 소개하고 있다. 이들은 한국어를 모국어로 하는 사람들이 영어 작문에서 주로 어떤 종류의 오류를 범하는지를 파악하여 규칙화하였다.[6]

첫 번째로 살펴본 에세이 전체 텍스트의 수준을 평가하는 시스템은 두 번째로 살펴본 문장단위의 평가시스템의 성능향상을 통해 발전을 도모할 수 있다. 그러나 문장단위의 수준을 평가하는 시스템은 철자 및 문법오류에 대한 기계적 오류를 체크하는 반면 어휘나 문법패턴에 따른 표현난이도를 고려하지 못하고 있어 인간채점자와 평가시스템의 평가 간에는 차이가 존재할 수밖에 없다.

본 연구에서는 인간채점자가 철자나 문법패턴 뿐만이 아니라 표현난이도도 고려하여 채점한다는 사실에 입각하여 기존 시스템이 갖추지 못한 표현난이도에 대한 자질을 추가하여 실험하였다.

## 3. 규칙기반 문장수준 평가시스템

본 연구에서는 비원어민의 영어에세이 분석에서 Link-Grammar에 근거한 규칙기반 평가시스템과 원어민 평가자의 평가 간에 높은 일치도를 보였다는 연구결과를 바탕으로 Londsale and Strong-Krause(2003)이 제안한 Link-Grammar Parser기반의 규칙기반 평가시스템을 문장단위의 평가에 맞게 재구성하여 비교시스템으로 구현하였다.

Lonsdsale and Strong은 Link-Grammar Parser가 에세이 코퍼스를 잘 분석할 수 있도록 미등록어를 사전에 추가하는 작업을 하였다. 이에 본 비교시스템에도 국내 학생들이 주로 사용하는 단어를 통계적 기법을 통해 추출하여 등록하였다.[7]

*예) Acronym과 같이 R.O.K(Republic of Korea), ELC(the English Language Center)*

(1) 에세이 문장을 Link-Grammar Parser에 입력
(2) Parser에 의한 출력결과에서 에러 단어를 추출
(3) 에러단어의 빈도수에 따라 0~5까지의 평가

(그림 1) 규칙기반시스템의 처리순서

## 4. 기계학습기반 문장수준 평가시스템

### 4-1. 수준평가를 위한 자질정보

선행연구[8]에서 밝혀진 낮은 유의미성을 지닌 TTR(Type-Token Ratio), 문장 길이, 어휘 길이 등과 같은 구문의 통계적 정보와 에러율 보다는 문장 내 어휘들 간의 관계를 고려한 문법과 의미적 정보를 다각도로 살피고 활용하는 방법을 수준평가를 위한 자질로써 사용하기로 하였다. 이에 단어, 품사, 의존문법기반의 문장성분 N-Gram과 Link-Grammar Parser를 통해 추출된 에러패턴을 그림 2와 같이 활용하였다. 각각의 추출된 자질은 그림 3과 같이 TF-IDF 기법을 이용하여 가중치를 계산하였다.

품사태거는 Standford Pos-Tagger를 사용하였으며, Dependency-Parser는 ClearNLP그룹의 Parser를 활용하였다.

(1) 단어 자체의 N-Gram
  예) a, apple, absurd ..., a apple, apple absurd, ...
(2) 품사태거에 대한 설명과 단어-품사 N-Gram
  예) a/DT, apple/NN ..., a/DT apple/NN, ...
(3) Dependency-Parser를 통한 문장성분 N-Gram
  예) This/SUBJ, is/HEAD..., This/SUBJ is/HEAD,..
(4) Link-Grammar Parser의 에러패턴
  예) [Thes] is [an] [books]

(그림 2) 어휘간의 관계를 고려한 자질

$$tf \cdot idf_{t,d} = tf_{t,d} \times \log \frac{N}{df_t}$$

$tf_{t,d}$: 문서 내 특정 단어의 빈도 수
$N$ : 코퍼스 내 문서 수
$df_t$ : 코퍼스 내에서 단어의 출현 횟수

(그림 3) tf-idf 가중치 계산식

### 4-2. 기계학습기법 선택

WEKA Workbench라는 기계학습 툴킷(Toolkit)을 사용하여 표 1과 같이 8개 알고리즘의 정확률(Precision)과 재현율(Recall), 그리고 F-measure를 측정하였다.

실험결과 8개의 알고리즘 중에서 NaiveBayesMultinomial과 SVM 알고리즘이 세 가지의 항목 모두에서 가장 높은 성능을 보여주었으며, 이 두 가지 알고리즘을 규칙기반의 평가시스템과 비교하기 위하여 선택하였다.

<표 1> 각 기계학습기법에 따른 학습 후 실험결과

| Classifier | Precision | Recall | F-measure |
|---|---|---|---|
| NaiveBayesM* | **0.673($1^{st}$)** | **0.671** | **0.652** |
| SVM | **0.631($2^{st}$)** | **0.627** | **0.592** |
| NaiveBayes | 0.532 | 0.534 | 0.512 |
| KNN | 0.431 | 0.441 | 0.402 |
| AdaBoost | 0.556 | 0.551 | 0.521 |
| Bagging | 0.432 | 0.421 | 0.412 |
| J48(C4.5) | 0.457 | 0.459 | 0.427 |
| RandoForest | 0.458 | 0.451 | 0.438 |

## 5. 실험 및 결과

### 5-1. 실험환경

본 논문에서는 국내 유명 어학원을 통해 수집된 서로 다른 500개의 영어에세이 코퍼스에서 2,000문장을 추출하였다. 영어에세이 문장의 평가는 기존에 첨삭 및 교정을 해본 경험이 있는 영문학 및 언어학 석/박사 학위를 받은 4명의 전문평가자들이 ETS 스코어링 가이드라인[9]을 비롯한 다양한 채점표를 취합하여 문법지식, 어휘의 사용, 형식적 규칙-mechanics, 내용수행의 4개 항목을 바탕으로 2,000문장을 교차 평가하였다.

표 2와 같이 평가된 2,000개의 문장을 WEKA Toolkit을 이용하여 다양한 알고리즘에 10-fold cross validation 기법을 적용하여 실험하였으며 정확률이 높은 상위 2개의 알고리즘(NaiveBayesMultinomial과 SVM 알고리즘)과 규칙기반 평가시스템과의 비교・분석을 실시하였다.

<표 2> 평가된 문장 데이터 개수

| Level | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Count | 320 | 310 | 350 | 300 | 350 | 370 |

평가척도로는 그림 4와 같이 정확률(Precision)과 재현율(Recall), F-measure를 사용하였다.

$$\mathrm{Precision} = \frac{\text{시스템이 추출한 정답 문장 수}}{\text{시스템이 추출한 문장 수}}$$

$$Recall = \frac{\text{시스템이 추출한 정답 문장 수}}{\text{전체 문장 수}}$$

$$F-measure = \frac{2 \times \mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall}$$

(그림 4) 평가 척도

### 5-2. 실험결과

(1) 규칙기반 평가시스템과 상위 2개의 기계학습알고리즘

표 3과 같이 기계학습기법을 이용한 평가시스템이 규칙기반 평가시스템에 비해 정확률과 재현율에서 높은 성능을 보였다.

<표 3> 각 기계학습기법에 따른 학습 후 실험결과

| Classifier | Precision | Recall | F-measure |
|---|---|---|---|
| NaiveBayesM* | **0.673($1^{st}$)** | **0.671** | **0.652** |
| SVM | **0.631($2^{st}$)** | **0.627** | **0.592** |
| Rule | 0.431 | 0.434 | 0.391 |

(2) 각 특성별 정확률과 재현율

표 4와 같이 각 자질을 제거하고 실험을 하였을 경우에 해당 자질이 얼마나 전체 정확률을 높이는 것에 기여하는지를 테스트 해보았다. N-Gram의 경우 4-Gram까지 추출하였다.

<표 4> 자질 별 기여도

| Removed Feature | Precision | Recall | F-measure |
|---|---|---|---|
| Word N-Gram | 0.677 (+0.004) | 0.674 (+0.003) | 0.655 (+0.003) |
| Word-Pos N-Gram | 0.665 (-0.008) | 0.662 (-0.009) | 0.643 (-0.009) |
| Dependency-Grammar based constituent N-Gram | 0.632 (-0.041) | 0.631 (-0.04) | 0.613 (-0.039) |
| Link-Grammar based Error Pattern | 0.643 (-0.03) | 0.644 (-0.027) | 0.621 (-0.031) |
| **All Features** | **0.673** | **0.671** | **0.652** |

1) 선택된 자질을 제거하고 실험한 결과 단어 N-Gram은 제거된 후에 정확률을 높이는 긍정적인 영향을 미치는 것으로 나타났다.

2) 반면에 단어-품사 N-Gram과 에러패턴, 그리고 문장성분 N-Gram이 제거된 후에 정확률을 떨어뜨리는 현상을 보였다.

### 5-3. 결과분석

(1) 기계학습기반 평가시스템 및 자질들의 유효성

기계학습기반의 평가시스템이 기존 연구되어온 문장단위의 규칙기반 평가시스템보다 우수한 성능을 보였다. 또한 개별 자질들 중에서도 문장성분 N-Gram, 에러패턴, 단어-품사 N-Gram의 자질들이 평가에 중요한 역할을 하는 것으로 나타났다. 이러한 결과를 미루어 볼 때 평가자에 의해 평가되는 기준이 단순히 철자와 문법의 오류뿐만이 아니라 어휘와 문법의 패턴에 따른 표현난이도까지도 판단한다는 것을 알 수 있다. 또한 문장이 길고 어려운 문법을 사용하여 표현을 한 문장일수록 오류가 발생할 확률이 높기 때문에 단순히 규칙기반으로 평가하는 것 보다는 본 연구에서 제시한 기계학습기법과 여러 유의미한 자질들을 사용하는 것이 유용하다는 것을 보여준다.

(2) 각 분류별 영향력 높은 자질분석

표 5와 같이 각 수준별로 참-긍정(true-positive)의 확률을 높일 수 있었던 주요 자질들을 분석해 보았다.

<표 5> 영향력 높은 수준별 자질분석

| Level | Summary |
|---|---|
| 0 | 에러율이 80~100%로 매우 높았으며, 문장의 성분이 하나 이상 빠져 문장을 제대로 구성하지 못하여 의미전달이 되지 않음. |
| 1 | 단순한 문장 패턴을 갖고 있으며 to be, Let's와 같은 동사구가 주로 사용됨. |
| 2 | 현재형, 현재진행형, 단순 부정문, 대등접속사(and, or, but), There is/are, Here is/are가 주로 사용됨 |
| 3 | 과거형, 미래형의 시제와 can, may 조동사, 대등접속사(for, so), to부정사와 단순 의문문이 주로 사용됨. |
| 4 | 불규칙 과거동사, 분사구문, 동명사, 비교급(most, least), 과거형 의문문, 기본적인 종속접속사(because, when, if, before, after, that, as, while, wether, once), in that ~, now that ~, so ~that 이 주로 사용됨 |
| 5 | 에러가 거의 없으며 if 가정문, 과거완료형(has been, will have been), could, should, would, might 등의 조동사/부정문, 수동태, although, whenever, until, since, even, though, even if, so that 등이 주로 사용됨. |

(3) 오분류 분석

오분류된 문장들을 살펴본 결과 문법 및 형식적 규칙-mechanics에는 문제가 없었으나 의미적으로 매우 어색한 문장들이 많았다. 그러한 많은 문장들이 의미적으로 자연스러운 문장이 되기 위해서는 문장 내에서 어색함을 발생시키는 부분이 전혀 새로운 내용으로 대체되거나 삭제가 되어야만 한다. 따라서 평가자들은 그러한 부분을 평가함에 있어 감점을 주게 되는데, 이는 곧 의미적으로 어색한 문장에서 생성된 올바른 문법패턴 자질들이 오분류 영역으로 평가되는 주요한 원인을 제공하는 것으로 분석된다.

*예) Also, we can get power to keep our trip.*

또한 레벨 1~3에서 오분류가 더 높은 것으로 나타나는데 그 원인은 에세이 코퍼스의 특성상 학생들이 에세이에서 표현해야 되는 어휘와 문법의 영역이 유사하기 때문인 것으로 분석된다.

## 6. 결론 및 향후과제

본 논문에서는 유의미한 자질들을 생성하고 기계학습기법을 이용함으로써 기존에 연구되어온 규칙기반 문장평가 시스템에서의 단점을 개선하는데 도움을 줄 수 있다는 것을 밝혔다. 특히 문서분류 등에서 최적의 성능을 내고 있는 NaiveBayes나 SVM이 영어작문 문장 수준의 분류에서도 더 나은 성능을 나타내는 것으로 확인되었다.

또한 Link-Grammar나 Dependency-Grammar Parser를 통하여 문장 내 어휘들 간의 유의미한 정보를 출력함으로써 기계학습을 위한 유용한 자질을 생성할 수 있었다.

그러나 Parser의 정확도가 떨어지는 기술적 한계로 인하여 문장성분이나 에러패턴의 검출이 정교하지 못한 것은 앞으로 해결해 나가야할 과제이다.

향후 연구로는 더 많은 영어에세이 코퍼스를 수집하고 더욱 정교한 구문분석 도구들을 활용하여 분류 모델의 성능을 개선하고 나아가 에세이 전체 텍스트의 전반적인 평가를 해보고자 한다.

## 참고문헌

[1] J. Burstein and D. Higgins, "Advanced Capabilities for Evaluation Student Writing: Detecting Off-Topic Essays Without Topic-Specific Training," Proceedings of the International Conference on Artificial Intelligence in Education, July 2005.

[2] Jill Burstein and Martin Chodorow. 1999. Automated essay scoring for nonnative English speakers. In Computer Mediated Language Assessment and Evaluation in Natural Language Processing, pages 68-75. Association for Computational Linguistics.

[3] Jill Burstein. 2003. The E-rater scoring engine: Automated essay scoring with natural language processing. In Mark D. Shermis and Jill C. Burstein, editors, Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum, Mahwah, NJ.

[4] Martin Chodorow and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In Proceedings of ANLP-NAACL 2000, pages 140-147. Morgan Kaufmann Publishers.

[5] Deryle Lonsdale and Diane Strong-Krause 2003. Automated Rating of ESL Essays. HLT-NAACL-EDUC '03 Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing - Volume 2, Pages 61-67

[6] 김지은, 이공주 2007. 중학생 영작문 실력 향상을 위한 자동 문법 채점 시스템 구축. 한국콘텐츠학회논문지 '07 Vol. 7 No.5

[7] K. Gonjenola. and M. Oronoz. "Corpus-based Syntactic Error Detection Using Syntactic Patterns," Proceedings of the Workshop on Student Research, pp.24-29, 2000.

[8] 최인철, 임해창, 박정 2003. 영작문의 전산언어학적 채점 타당성. Multimedia Assisted Language Learning Vol.6 No.2, 2003.12, 221-241 (21 pages)

[9] Scoring Guides(Rubrics) for Writing Response www.ets.org/Media/Tests/TOEFL/pdf/WritingRubrics.pdf