KIPS Tr. Software and Data Eng.
Vol.4, No.11 pp.509~520 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2015.4.11.509

Prediction of Correct Answer Rate and Identification of Significant Factors for CSAT English Test Based on Data Mining Techniques

Park Hee Jin[†] · Jang Kyoung Ye[†] · Lee Youn Ho[††] · Kim Woo Je[†††] · Kang Pil Sung[††††]

ABSTRACT

College Scholastic Ability Test(CSAT) is a primary test to evaluate the study achievement of high-school students and used by most universities for admission decision in South Korea. Because its level of difficulty is a significant issue to both students and universities, the government makes a huge effort to have a consistent difficulty level every year. However, the actual levels of difficulty have significantly fluctuated, which causes many problems with university admission. In this paper, we build two types of data-driven prediction models to predict correct answer rate and to identify significant factors for CSAT English test through accumulated test data of CSAT, unlike traditional methods depending on experts' judgments. Initially, we derive candidate question-specific factors that can influence the correct answer rate, such as the position, EBS-relation, readability, from the annual CSAT practices and CSAT for 10 years. In addition, we drive context-specific factors by employing topic modeling which identify the underlying topics over the text. Then, the correct answer rate is predicted by multiple linear regression and level of difficulty is predicted by classification tree. The experimental results show that 90% of accuracy can be achieved by the level of difficulty (difficult/easy) classification model, whereas the error rate for correct answer rate is below 16%. Points and problem category are found to be critical to predict the correct answer rate. In addition, the correct answer rate is also influenced by some of the topics discovered by topic modeling. Based on our study, it will be possible to predict the range of expected correct answer rate for both question-level and entire test-level, which will help CSAT examiners to control the level of difficulties.

Keywords : College Ability Scholastic Test(CSAT) Difficulties, English Test, Topic Modeling, Multiple Linear Regression, Decision Tree

데이터마이닝 기법을 활용한 대학수학능력시험 영어영역 정답률 예측 및 주요 요인 분석

박 희 진[†] · 장 경 애[†] · 이 윤 호[††] · 김 우 제[†††] · 강 필 성[††††]

요 약

대학수학능력시험(수능)은 고등학교 3년간의 학업 성취도를 측정하는 대표적인 평가 도구로서 대한민국 대학 입시에 있어 매우 중요한 역할을 하는 시험이다. 응시생들의 학업 성취도를 효과적으로 평가하기 위해서는 수능의 난이도가 적절하게 조절되어야 하나 지금까지는 수능 난이도의 편차가 매우 크게 나타나 매 입시연도마다 여러 가지 문제점을 야기해왔다. 본 연구에서는 전문가의 판단에 의존한 기존 방식에서 벗어나 지금까지 시행된 모의고사 및 실제 시험을 통해 축적된 자료를 바탕으로 데이터마이닝 기법을 적용하여 영어영역 문제의 난이도를 예측하는 모델을 구축하고 난이도 예측에 영향을 미치는 요소를 판별하고자 한다. 이를 위해 각 문항의 특성을 판별할 수 있는 여러 지표와 함께 지문, 문제, 답안 등에 나타난 단어들의 특징을 토픽 모델링(topic modeling) 기법을 이용하여 정량화하고 이를 바탕으로 선형회귀분석 및 의사결정나무 기법을 이용하여 각 문항의 난이도를 예측하는 모델을 구축하였다. 구축된 예측 모델을 실제 문제에 적용한 결과 난이도의 상/하 구분에 대한 예측 정확도는 90% 수준으로 나타났으며, 실제 정답률 대비 오차 비율은 약 16% 이내인 것으로 나타났다. 또한 배점 및 문제 유형이 문제의 난이도에 큰 영향을 미치며 지문이 특정 주제에 관련된 경우에도 난이도에 영향을 미치는 것을 확인하였다. 본 연구에서 제시된 방법론을 이용하여 영어영역 각 문제들에 대한 기대 정답률의 범위를 추정할 수 있으며 이를 종합하여 영어영역 전체 문제에 대한 정답률 예측을 통해 적절한 난이도의 문제를 출제하는 데 기여할 수 있을 것으로 기대한다.

키워드 : 대학수학능력 난이도, 외국어영역, 토픽 모델링, 다중선형회귀분석, 의사결정나무

※ 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2014R1A1A1004648).
† 비 회 원 : 서울과학기술대학교 IT정책전문대학원 산업정보시스템전공 박사과정
†† 비 회 원 : 서울과학기술대학교 글로벌융합산업공학과 부교수
††† 비 회 원 : 서울과학기술대학교 글로벌융합산업공학과 교수
†††† 종신회원 : 고려대학교 산업경영공학부 조교수

Manuscript Received : July 3, 2015
First Revision : August 28, 2015
Accepted : August 30, 2015
* Corresponding Author : Kang Pil Sung(pilsung_kang@korea.ac.kr)

1. 서 론

1994년 첫 시행 이래 우리나라에서 대학수학능력시험(이하 수능)은 수험생들의 대학 진학에 가장 중요한 역할을 하는 국가적 차원의 중요한 시험으로 자리매김해왔다. 지난 2014년 11월 13일 실시된 2015학년도 수능은 역대 최고의 만점자가 배출될 만큼 쉽게 출제가 되어 난이도 조절에 실패했다는 평가를 받았다[1]. 한국교육과정평가원의 채점결과 분석[2-3]에 의하면, 영어영역은 만점자가 전체 수험생의 3.36%인 19,564명으로 교육당국이 목표로 하고 있는 만점자 1%와 매우 큰 차이를 나타내었으며 2012학년도 만점자 비중인 2.67%보다도 높아 '물수능'이라는 오명을 피할 수 없게 되었다.

정부는 변환표준점수 제도, 특차전형 폐지, 9등급제 도입, EBS연계출제발표, 만점자 1% 정책 등 다양한 정책 변화를 통해 수능이 대학입학능력을 객관적으로 검정할 수 있는 타당성 있는 제도로 자리 잡을 수 있게 노력을 하고 있으나[4], 1994년 첫 시행부터 수능의 난이도 조절에 대한 문제점은 지속적으로 제기되고 있는 실정이다. 이러한 난이도 조절 관련 문제점을 해결하기 위해서 학계에서는 다양한 시각에서 분석해왔으며, 영어영역에 대한 대표적인 교육학적 접근 방법론들로는 수능 문항유형의 타당성 분석[5], 지문친숙도 분석[6], 어휘 분석[7], 읽기 난이도 분석[8], 언어영역점수, 지적능력이 수능에 미치는 요인[9] 등이 있다. 그러나 기존의 연구들은 (1) 수능에 영향을 미치는 독립변수를 기존 연구에서 도출하여 설문조사를 실시하는 방식, (2) 수능채점 데이터를 수집하여 집계하는 방식, 또는 (3) 연구실험자를 모의로 추출하여 시험을 실시한 후 결과를 분석하는 방식 등 대부분 표본조사를 통한 분석을 실시함으로써 결과의 신뢰성 및 대표성에 대한 한계점을 드러내고 있다.

따라서 본 연구에서는 수십 년간 누적된 수능 데이터를 수집하여 영어영역 난이도에 대한 정량적인 분석 방법론을 제시하고 높은 정확도를 나타내는 문항별 영어영역 정답률 예측 모델을 개발하고 정답률에 영향을 미치는 주요 요인을 판별하고자 한다. 이를 위해 첫째, 연차적으로 시행되는 수능 모의평가 2회와 수능시험 1회에 해당하는 대량의 실제 수능 문제 전체를 추출하고 텍스트마이닝 기법의 하나인 토픽 모델링(topic modeling) 분석을 실시하여 문항들의 기저에 존재하는 패턴을 찾아내는 연구를 수행하였다. 둘째, 수능 영어영역 문항 및 지문의 특성과 해당 문제의 정답률과의 상관관계를 분석하고 이를 바탕으로 문항별 영어영역 정답률 예측 모델을 구축하였다. 정답률 예측 모델로는 다중선형회귀분석(multiple linear regression)을 이용한 정답률 예측 모델과 분류 및 회귀 나무(Classification and regression trees; CART)를 이용한 이범주 난이도 분류 모델 두 가지가 구축되었으며, 사후 분석을 통해 두 예측 모델의 결과를 비교 및 분석하였다. 본 연구를 통해 영어영역 문항의 지문 및 문제의 특성을 이용할 경우 높은 정확도의 정답률 예측 모델 구축이 가능하다는 것이 밝혀졌으며, 이를 이용하여 기존에 전문가의 지식에만 의존했던 난이도 예측의 어려움이 데이터 기반의 정량적 방법론을 통해 보완될 수 있을 것으로 기대한다.

본 연구의 구성은 다음과 같다. 2절에서 연구의 이론적 배경과 대표적 선행연구 내용들을 소개한다. 3절에서는 연구의 설계방법 및 정답률 예측 모델 구축에 관련된 내용들을 설명하며, 4절에서는 정답률 예측 모델 수행 결과를 분석하고 토의한다. 마지막 5절에서 연구의 내용을 요약하고 결론 및 향후 연구에 대해서 논의한다.

2. 선행연구 고찰 및 이론적 배경

2.1 선행연구 분석

Table 1. Methodologies and Results from Related Works

No	Related Works	Reference	Methodologies and Results
1	Feasibility Analysis of the CSAT question types	[5]	Classify the type of feasibility evaluation using 50 experts
2	Problem count, sentence length, sentence familiarity analysis	[6]	By classifying the real problems that the student experimenter tests conducted after analysis
3	Lexical analysis	[7]	The frequency measurement using corpus analysis tools (NLPTools)
4	Reading Difficulty Analysis	[8]	Extraction survey conducted by the independent variables analyzed previous studies (66 items), regression analysis
5	Language parts scores, intellectual ability Impact Analysis	[9]	CSAT score results collected, regression analysis
6	Qualified foreign language test scores and association studies	[10]	Students participate in physical examination and classification problems (men and women high school third grade 30 students)

우리나라 대학 입시에서 수능이 차지하는 위상과 함께 매년 제기되는 수능 난이도 조절 실패 문제점을 타개하고자 연구자들은 언어학적 측면, 교육적 측면 등 다양한 관점에서 난이도를 분석하는 연구를 수행하였다. [5]의 연구에서는 50명의 출제 전문가를 대상으로 외국어 영역의 문항 유형별 타당성 평가를 수행하여 수준별 영어시험 및 검사지 구성에 반영하고자 하였다. 보다 직접적인 난이도 관련 분석 연구로써 이경숙은 [6]의 연구에서 먼저 학생 실험자를 분류한 뒤 실제 해당 문제를 통한 시험을 실시함으로써 문제 수, 지문 길이, 지문 친숙도 등의 요인과 정답률 간의 관계를 규명하고자 하였다. 실험 결과, 시험방식은 시험수행에 영향을 미치지 않았고, 지문 친숙도와 고숙달도가 영향을 미치는 것으로 나타났다. 또한 텍스트의 지시어, 의문문과 부정어 빈도, 전개

방식 등의 분석 결과, 단서가 되는 텍스트가 앞에 나올 경우 정답률이 높다는 연구를 바탕으로 김남복은 [7]에서 코퍼스 분석 툴을 사용하여 어휘의 빈도 수를 측정한 뒤 이를 통해 어휘 반복률은 연도별로 큰 차이가 없고, 총 어휘 수는 연도별로 증가하는 등 어휘와 수능 간의 상관관계가 높다고 밝혔다. [8]의 연구에서는 선행연구 분석을 바탕으로 추출된 독립변수에 대한 설문조사를 실시하고 그 결과를 회귀분석을 통해 모형화함으로써 도출된 결과는 예측 모형의 설명력은 28.9%, 실제 모의고사를 예측한 결과 실제 대비 1.12점 낮게 예상되었다. 이에 더하여 수능 영어영역이 공인외국어 시험과 연관관계가 높을 것이라는 가정하에, [10]에서는 공인외국어 시험을 분석하여 유의미한 요인을 도출하였다.

앞서 언급된 기존 연구들의 대표적인 한계점은 대량의 수능문제 지문에 대해 전수 데이터를 추출하여 정량적으로 평가한 방법이 없다는 점이다. 선행 적용된 전문가 혹은 실험자에 의한 샘플링 평가는 표본 자제가 가질 수 있는 편향성의 문제로 전체 수능 데이터의 특성을 포괄하기에는 부족하다고 판단된다. 또한 코퍼스툴을 사용하여 정량적 분석데이터를 도출한 [7]의 연구의 경우, 연구의 결론이 사용된 어휘의 단순 빈도 추출에 한정되어 보다 심도 있는 연구가 필요하다고 할 수 있다. 따라서 연도별로 상이한 난이도를 전반적으로 포괄하는 선행연구가 많지 않은 실정에서 수능 영어영역에 대한 정량적인 영향요인 분석 및 정답률 예측 모델 구축에 대한 연구가 필요한 실정이다.

2.2 분석 방법론

본 연구에서는 과거 수능 영어영역 출제 문제들로부터 도출할 수 있는 정량적인 속성들에 더하여 데이터마이닝 기법 중 텍스트를 분석하는 데 주로 활용되는 토픽 모델링 기법을 사용하여 문제 및 지문의 특징을 도출하고 이를 예측 모델의 입력 변수로 사용하고자 한다. 이렇게 도출된 입력 변수를 바탕으로 각 문제의 정답률을 예측하기 위해 통계 및 데이터마이닝 분야에서 널리 사용되는 다중선형회귀분석 및 의사결정나무 기법을 활용하여 정답률 자체를 직접 예측하는 회귀모형과 문제의 난이도(상/하)를 예측하는 분류모형을 구축한다. 본 연구에 사용된 각 기법들에 대한 개괄적인 개념은 다음과 같다.

다중선형회귀분석은 여러 개의 독립변수(X_1, X_2, ..., X_n)와 종속변수(Y) 사이에는 선형 관계식(linear relationship)이 있음을 가정하고 주어진 학습 데이터를 바탕으로 각 독립변수의 영향력인 회귀 계수(β)를 다음과 같은 식을 이용하여 추정하는 모형이다.

$$Y = \beta_0 + \beta_1 \chi_1 + \beta_2 \chi_2 \cdots \beta_n \chi_n + \varepsilon \qquad (1)$$

다중선형회귀분석에서 많은 입력 변수는 노이즈 데이터에 민감하게 반응하거나 불필요한 정보를 포함하여 회귀분석의 예측 성능을 저하시킬 위험이 있다. 따라서 후보 입력 변수 중에서 종속변수를 예측하는 데 있어 통계적으로 유의미한 변수만을 선택하기 위해 교사적 변수 선택(supervised variable selection) 기법인 전진선택법(forward selection), 후방소거법(backward elimination) 및 단계적 선택법(stepwise selection) 등을 적용하여 최적의 변수 집합을 판별한 뒤 이를 이용하여 회귀모형을 구축하는 것이 일반적이다.

텍스트마이닝은 대량의 텍스트 데이터에서 자연어처리 기술과 문서처리 기술, 데이터마이닝 알고리즘을 적용하여 의미 있는 패턴을 발견하는 분석 방법으로, 대량의 텍스트데이터를 효율적으로 분석할 수 있는 연구 분야이다. 텍스트마이닝은 업무처리 산출물, pdf파일, e-mail, 뉴스, 웹페이지 데이터 등 비정형, 반정형 텍스트 데이터에서 효율적인 방식으로 텍스트의 패턴을 추출하고 이에 대한 의미를 발견할 수 있다. 텍스트마이닝은 자연어처리 기술과 문서처리 기술을 적용하여 문서요약, 특성추출 등의 연구에 활발히 이용되고 있다[13-15]. 텍스트마이닝은 데이터수집, 전처리 및 데이터변환, 변수 선택 및 추출, 알고리즘 학습 및 평가의 단계를 거쳐 수행된다. 먼저 수집된 데이터는 텍스트의 파싱, 불용어(Stopword) 제거, 단어표준화, 대표형 변환 등의 전처리 과정을 통해서 불필요한 정보들을 제거하게 된다. 전처리가 수행된 텍스트는 일반적으로 Bag-of-words 표현 방식을 사용하여 문서-단어 행렬(term-document matrix) 형태로 변환되며, 이렇게 변환된 텍스트 데이터를 이용해서 향후 분석 과정에서 유의미한 주요 독립변수를 선택 또는 추출한다. 마지막으로 차원이 축소된 텍스트 데이터를 바탕으로 적합한 데이터마이닝 알고리즘을 적용하여 목적에 맞는 예측 모델을 학습한 뒤 새로운 검증데이터에 적용하여 그 성능을 평가한다.

본 연구에서는 수능 영어영역 지문 텍스트 데이터를 활용하여 텍스트마이닝의 대표적인 분석 기법인 토픽 모델링을 통해 문항별 주제 분포를 추정하여 이를 정답률 예측 모델의 입력 변수로 사용하고자 한다. 토픽모델링은 문서의 주제를 알기 위해 원본 텍스트 내의 단어를 분석하는 통계적인 방법이다. 토픽모델링은 한 문서 내에서 나타나는 빈도가 높은 단어들의 집합을 토픽으로 표현하기에, 문서에 대해 주석 혹은 범주 부여 등의 사전작업을 필요로 하지 않는 장점을 가지고 있어, 대량의 문서 코퍼스(corpus)를 기계적으로 빠르고 정확하게 처리할 수 있다는 장점이 있다.

본 연구에는 토픽모델링용 알고리즘으로 Latent Dirichlet Allocation(LDA)을 사용한다. LDA는 Fig. 1에서 나타난 바와 같이 문서를 생성하기 위한 내재적인 절차를 가정하고 주어진 문서 집합(corpus)이 가장 잘 표현될 수 있는 분포의 모수들을 추정하는 방법론이다. 따라서 LDA는 하나의 문서를 여러 토픽으로 이루어진 집합체로 간주하고, 각 토픽은 디리클레 분포(Dirichlet Distribution)를 따른다고 가정한다. 또한 각 토픽은 모든 단어들의 집합체이며 개별 토픽에서 각 단어들 역시 디리클레 분포를 따른다고 가정한다. 디리클레 분포는 K차원의 실수벡터 중 벡터의 요소가 양수이며 모든 요소를 더한 값이 1인 경우에 대해 확률값이 정의되는 연속확률 분포이다. 이러한 가정을 바탕으로 Fig. 1에서 표현된 문서 생성 과정에서의 결합 확률분포는 다음과 같이 표현된다.

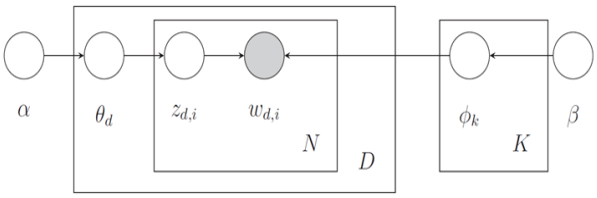

Fig. 1. Document Generation Process of LDA

$$p(w,z,\theta,\phi|\alpha,\beta)=p(\theta|\alpha)p(z|\theta)p(\phi|\beta)p(w|z,\phi) \quad (1)$$

여기서 w,z,θ,ϕ는 각각 단어, 해당 단어가 선택된 토픽, 문서에서의 토픽 분포, 그리고 토픽에서 단어의 발생 확률을 의미하며, α와 β는 각각 문서에서의 토픽 분포에 대한 사전 디리클레 모수(parameter) 및 토픽에서의 단어 분포에 대한 사전 디리클레 모수를 의미한다. LDA는 실제 분포를 알 수 없으므로 Variational Inference 또는 Gibbs Sampling 등의 기법을 통해 전체 문서 집합에서 실제 단어들의 발생 확률이 최대화되도록 w,z,θ,ϕ를 추정한다[16].

3. 연구 설계

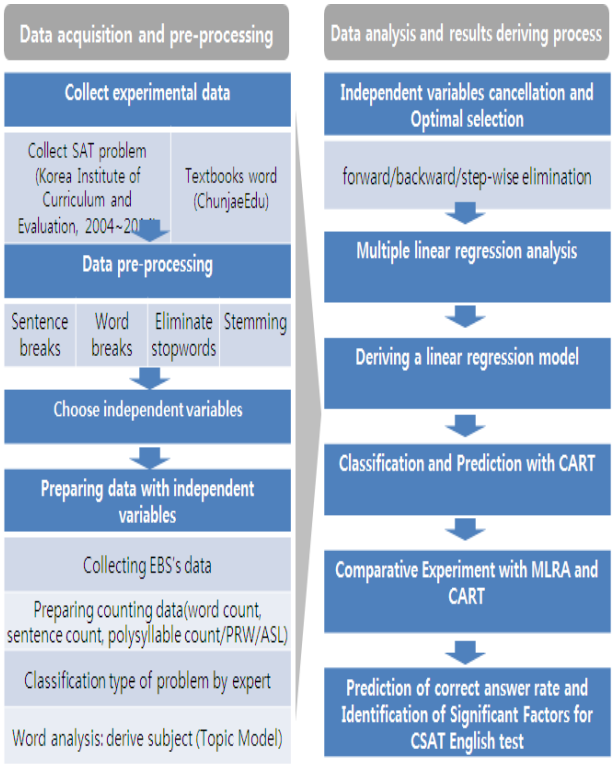

Fig. 2. Research Framework

본 연구의 수행 절차는 Fig. 2에 나타난 바와 같다. 먼저 과거 수능 영어영역에 출제된 문항에 대한 기초 데이터(지문, 문제 유형, 배점, 정답률 등)와 함께 사용된 어휘들에 대한 추가 정보(교과서 빈출도 등)를 수집하였다. 이를 토대로 불필요한 정보를 제거하고 대표형 변환 등을 통해 단어들에 대한 표준화 작업을 수행하였다. 이후 문항 난이도와 관련이 있을 것으로 예상되는 설명변수들을 기존 선행연구 결과를 토대로 설정하였으며, 각 문항 지문의 특성을 추출하기 위하여 토픽 모델링(topic modeling) 기법을 활용하여 지문의 주제 분포를 추정하였다. 이를 바탕으로 변수선택을 시행한 다중선형회귀분석을 통해 정답률 예측 모형 및 예측에 유의미한 영향을 미치는 분석을 실시하였고, 의사결정나무 알고리즘을 활용하여 상/하로 구분된 난이도 분류 모델을 구축하고 분류에 중요한 영향을 미치는 요인을 판별하였다. 최종적으로는 두 예측 모델(회귀모델 및 분류모델)의 비교를 통해 정답률 예측에 큰 영향을 미치는 요인을 분석하였다.

3.1 데이터 수집 및 전처리 설계

본 연구에서는 수능 영어영역 정답률 예측 모델을 구축하기 위하여 2004년부터 2014년까지 출제된 수능 영어영역 및 모의고사 문항을 수집하였고, 수능이 시작된 1994년부터 2003년까지의 영어영역 시험 데이터는 2004년부터 시작된 7차 교육과정과 유형에 차이가 있고 정답률 데이터가 없어 제외하였다. 수능 영어영역 및 모의고사(6월, 9월, 11월) 문항 데이터는 한국교육과정평가원을 통해 수집하였으며, 분석에 사용된 총 문항 수는 듣기평가를 제외한 1,113문항이다. 정답률 예측 모델의 후보 독립 변수를 생성하기 위하여 한국교육과정평가원 및 사설교육기관에서 각 문항에 대한 정답률 및 배점 데이터를 수집하였으며, EBS 웹사이트를 통해 각 문항별 EBS 연계 여부 또한 조사하였다. 중학교 및 고등학교 교과서에 출현한 단어와 수능에 출제된 단어를 비교하기 위하여 중·고교 영어 교과서 중 가장 많이 사용되는 출판사(천재교육)를 선정하여 단어 및 빈출 정보 등을 추가적으로 수집하였다. 결과적으로 2004년부터 2014년까지 독해 영어지문 1,113문항과 함께 중 1 단어 459개, 중 2 단어 365개, 중 3 단어 360개, 고교 공통단어 344개, 고 1 단어 245개, 고 2 단어 29개를 수집하여 분석하였다. 사용된 모든 단어는 불용어 및 특수기호를 제거하고 표준단어로 변환(stemming)하는 작업을 거친 뒤 분석에 사용하였다.

본 연구에서는 수능 영어영역의 문항 난이도에 영향을 미치는 요인들에 대하여 기존 연구 결과를 바탕으로 다음과 같이 네 가지의 가설을 수립하고 이를 검증하기 위한 잠재적 독립 변수를 정의하였다.

- 가설 1: 지문의 문장 구조 및 사용된 단어의 난이도와 정답률은 음의 상관관계를 가질 것이다.
- 가설 2: 문제 유형에 따라 정답률이 달라질 것이다.
- 가설 3: 지문의 내용(context) 난이도(주제)에 따라 정답률이 달라질 것이다.
- 가설 4: 문항 자체의 속성(배점, 문항 위치, EBS 연계 여부 등)에 따라 정답률이 달라질 것이다.

Table 2. Input Variable Candidates for Correct Rate Answer Prediction

No	Type	Variable	Count	Definition
1	Word Difficulty (Hypothesis 1)	ave.word.frequency	1	average of word frequency point in the sentence
2		word.difficulty.80	1	-difficulty 80% or more.(Cumulative frequency ratio is greater than 20%) - The ratio of how much a rare word in the sentence
3		word.difficulty.90.ebs	1	- average of word.difficulty.90 and EBS
4		difficulty.80	1	Of the words appear in sentences, frequently ranking appears in plain English document sub-word ratio 80-100%
5		difficulty.90	1	Of the words appear in sentences, frequently ranking appears in plain English document sub-word ratio 90-100%
6		ex.middle	1	Rate the words of the paragraph does not belong to the middle school textbooks
7		ex.high.common	1	Rate the words of the paragraph does not belong to the high school's common textbooks
8		ex.high	1	Rate the words of the paragraph does not belong to the high school textbooks
9	Paragraph Length (Hypothesis 1)	count.character	1	count of characters in the paragraph
10		count.syllable	1	count of poly syllable(more than 3 syllables)
11		count.word	1	count of words in the paragraph
12		count.setence	1	count of sentences in the paragraph
13	Paragraph Reading Difficulty (Hypothesis 1)	ASL	1	- Average Sentence Length -count.word / count.sentence - The greater ASL, cloudy with long sentences composed of many words in paragraph
14		ASW	1	-Average Sentence Word -count.syllable / count.sentence -The greater ASW, cloudy with sentences composed of poly syllable words in paragraph
15		Flesch.RES	1	- 206.835 - (1.015 x ASL) - (84.6 x ASW), - The greater Flesch.RES, paragraph is easy to read
16		Flesch.kincaid	1	- Flesch Kincaid Grade Level = (ASL x 0.39) + (ASW x 11.8) - 15.9 - The smaller Flesch.kincaid, paragraph is easy to read
17		Fog	1	- Gunning's Fog Index Reading Grade Level = 0.4(ASL + %PSW) - The smaller Fog, paragraph is easy to read
18	Problem Type (Hypothesis 2)	problem.type	8	Classified problem types into 8 categories (1) subject, title (Subject) (2) content correspondence (Correspondence) (3) grammar, vocabulary (Grammar) (4) inference of blank word (BlankWord) (5) sequence of sentences, context, insert sentence (Sequence) (6) paragraph summary (Summary) (7) mood of paragraph(Mood) (8) reading for long setence (LongSentence)
19	Paragraph Subject (Hypothesis 3)	topic	30	- topic modeling for the words in the paragraph - 30 topics
20	Problem Attribute (Hypothesis 4)	diff.point	1	- score point for each problem (2~4 point) - diff.point = each point - average point of the total problems
21		prob.position	1	- percentile point according to the problem position - near to zero if the problem locates prior, near to 100 if locates the end position in the whole problems
22		EBS	1	- whether the problem related with EBS contents - 0: non correspondence 1: correspondence

위와 같은 가설을 바탕으로 Table 2와 같이 총 여섯 개의 범주(단어 난이도, 지문 길이, 이독성, 문제 유형, 지문 주제, 문항 속성)에서 22개의 잠재적인 독립 변수를 사용하였다. 다만, 두 개의 변수(문제 유형, 지문 주제)는 범주형(categorical) 변수로서 향후 선형회귀분석을 수행할 때 각 범주별로 하나씩 더미 변수(dummy variable)를 생성하여 사용하므로 실제 사용 변수의 수는 58개이다.

가설 1에 제시된 지문의 문장 구조 및 단어 난이도와 정답률과의 관계를 파악하기 위해여 총 세 가지의 범주(단어 난이도, 지문 길이, 이독성)에서 총 17개의 변수를 생성하였

다. 첫 번째 범주인 단어 난이도에 대해서는 총 8개의 변수가 생성되었다. 일반적으로 기출문제 지문에 출현 빈도가 높을수록 익숙한 단어이며 빈도가 낮을수록 어려운 단어로 분류할 수 있다. 따라서 본 연구에서는 “단어 빈출 평균”, “단어 난이도 80% 비율” 등을 사용하였다. 또한 중등 및 고등 교과서에 사용된 단어, 사용되지 않은 단어에 비해 난이도가 낮다는 가정을 바탕으로 난이도 점수를 부여하여 세 개의 변수를 생성하였다. 또한 어려운 단어일지라도 EBS 연계된 지문은 난이도가 감소되므로 그 영향을 고려한 “단어난이도 80%+EBS 연계” 변수를 정의하였다. 두 번째 범주인 지문 길이와 관련한 변수로는 지문의 “문자 수”, “음절 수”, “단어 수”, “문장 수” 네 가지를 정의하였고, 세 번째 변수인 이독성에 관련된 변수로는 영어 문장의 이독성에 대한 대표적인 지표인 Flesch Reading Ease Formula, Flesch Kincaid Grade Level 등을 사용하여 다섯 가지의 파생 변수를 생성하였다. Flesch Reading Ease Formula는 단어당 평균 음절 수, 평균 단어 수를 난이도에 반영한 표준이고, Flesch Kincaid Grade Level은 0-100 사이의 level로 부여되는 값이다. 또한 Robert Gunning에 의해서 1952년에 개발된 Gunning's Fog Index Reading Grade Level 역시 이독성 관련 변수로 선정하였다[17].

가설 2에서 제시된 문제 유형과 정답률과의 관계 규명을 위해서 각 문항별로 분류된 “주제/제목/지칭”, “내용일치/불일치”, “어법/어휘”, “빈칸추론”, “글의 흐름/순서/삽입”, “요약”, “분위기”, “장문독해” 8가지를 설명변수로 사용하였다. 가설 3에서 제시된 지문의 내용 난이도와 정답률 간의 관계 분석을 위해서는 지문의 내용에 대한 분류가 필요하다. 이미 정형화되어 존재하는 문제 유형과는 달리 지문의 내용은 사전적으로 분류된 체계가 존재하지 않으므로, 본 연구에서는 토픽 모델링(topic modeling)[16]을 적용하여 지문에 속한 단어를 바탕으로 총 30개의 주제를 도출한 뒤, 각 지문의 대표 주제를 설명변수로 사용하였다. 마지막으로 가설 4에서 제시된 문항 속성에 대해서는 배점 차이, 위치 및 EBS 연계 문항 여부를 설명변수로 사용하였다.

3.2 정답률 예측 모델 및 평가 지표

본 연구에서는 두 가지 방식의 정답률 예측 모델을 구축하고 그 결과를 분석하였다. 첫 번째 방식은 2.2절에서 설명된 다중선형회귀분석(multiple linear regression)을 이용하여 각 문항의 정답률 자체를 예측하는 회귀 모형을 구축하는 것이다. 다중선형회귀분석을 통해 구축된 수능 영어영역 정답률 예측 모델의 예측 정확도를 평가하기 위하여 Table 3에 나타난 바와 같이 네 가지의 성능 평가 지표를 사용하였다. 평균제곱오차(Mean Squared Error; 이하 MSE), 평균제곱오차제곱근(Root Mean Squared Error; 이하 RMSE) 및 평균절대오차(Mean Absolute Error; MAE)는 실제 정답률과 예측된 정답률 사이의 오차를 평가하는 데 중점을 두는 반면, 평균절대비율오차(Mean Absolute Percentage Error; 이하 MAPE)는 실제 정답률 대비 예측된 정답률의 오차 비율에 중점을 두는 지표이다.

Table 3. Performance Evaluation Metrics for Multiple Linear Regression

Evaluation	Definition
Mean Squared Error; MSE	$MSE=\frac{1}{n}\sum_{i=1}^{n}(y-y')^2$
Root Mean Squared Error; RMSE	$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y-y')^2}$
Mean Absolute Error; MAE	$MAE=\frac{1}{n}\sum_{i=1}^{n}\lvert y-y'\rvert$
Mean Absolute Percentage Error; MAPE	$MAPE=100\%\times\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y-y'\rvert}{\lvert y\rvert}$

두 번째 예측 모델은 정답률 자체를 예측하는 회귀모형이 아닌 정답률을 어려움(difficult)과 쉬움(easy) 두 가지 범주로 구분하고 이를 예측하는 분류 모형을 이용한 것이다. 이를 위해 먼저 정답률 70%를 기준으로 어려움과 쉬움 범주를 구분하였으며, 의사결정나무(decision tree)의 일종인 분류 및 회귀 나무(classification and regression tree; 이하 CART)를 이용하여 분류 모델을 구축하였다.

CART를 이용하여 분류 모형을 구축할 경우, 과적합(overfitting)을 방지하기 위하여 분기의 통계적 유의수준, 노드의 분기를 위한 최소 개체 수(minimum split), 최대 나무 깊이(maximum depth) 등 여러 종류의 알고리즘 파라미터(algorithm parameters)가 존재한다. 본 연구에서는 학습 데이터를 이용하여 최적의 파라미터 집합 탐색 및 모형 구축을 한 뒤, 검증 데이터를 이용하여 분류 성능을 평가하였다. 난이도 예측과 같은 이범주 분류 문제(binary classification problem)의 경우 실제 범주와 예측 모델에 의해 예측된 범주를 바탕으로 Table 4와 같은 혼동 행렬(confusion)이 생성되며, 이를 바탕으로 여러 종류의 정확도 지표를 사용할 수 있는데, 본 연구에서는 다음과 같이 각 범주에 대한 정확도(True Difficult Rate; TDR, True Easy Rate; TER), 정확도(Accuracy) 및 균형정확도(Balanced Correction Rate; BCR)를 사용하여 분류 성능을 평가하였다.

Table 4. Confusion Matrix Generated According to the Actual Difficulty Level and the Predicted Difficulty Level

Confusion Matrix		Prediction	
		Difficult (D)	Easy (E)
Real	Difficult (D)	nDD	nDE
	Easy (E)	nED	nEE

$$TDR=\frac{n_{DD}}{n_{DD}+n_{DE}},\quad TER=\frac{n_{EE}}{n_{ED}+n_{EE}} \tag{2}$$

$$Accuracy = \frac{n_{DD}+n_{EE}}{n_{DD}+n_{DE}+n_{ED}+n_{EE}},$$
$$BCR = \sqrt{TDR \times TER} \quad (3)$$

4. 연구 결과 및 토의

4.1 토픽 모델링

수집된 전체 지문에 대해 토픽 모델링을 적용하여 30개의 토픽을 추정하였다. 토픽 모델링은 R의 "topicmodels" 패키지에 구현되어있는 Latent Dirichlet Allocation(LDA) 기법을 이용하였으며 LDA의 모수를 추정하기 위해서는 Variational Expectation-Maximization(VEM)기법을 이용하였다. 각 토픽에서 높은 빈도로 출현하는 대표 단어들은 Table 5에 나타난 바와 같다. 본 연구에서는 토픽 모델링 적용 전에 표준형 변환(stemming)을 수행하였기 때문에 일부 단어들은 어미가 제거된 형태로 나타나는 것을 볼 수 있다. 토픽 모델링 결과, 특정 토픽은 명확하게 하나의 주제를 가지고 있는 경우가 있는 반면(Topic 4: 스토리텔링, Topic 13: 건강, Topic 19: 과학 등), 두 가지 이상의 세부 주제가 결합되어 있는 토픽 또한 존재한다. 본 연구에서는 토픽 모델링의 결과를 바탕으로 각 지문에 대한 토픽의 출현 빈도를 정답률 및 난이도 예측 모델의 입력 변수로 사용한다.

Table 5. Topic Modeling

No.	Representative words of the Topic
Topic 1	averag, group, volunt, hour, zach, annual, rate, toni, book, year
Topic 2	problem, music, suggest, restor, pollut, engin, social, compani, atmospher, drive
Topic 3	parent, peopl, organ, moral, principl, complex, make, person, child, children
Topic 4	speci, idea, stori, make, peopl, good, creativ, communic, import, develop
Topic 5	risk, stoneheng, whale, ticket, crosswalk, pedestrian, digit, interact, student
Topic 6	websit, prefer, product, type, audit, compani, coloni, england, form, green
Topic 7	children, feel, time, lion, traffic, critic, parent, ride, stori, room
Topic 8	time, peopl, experi, task, african, american, feel, posit, mathemat, comment
Topic 9	peopl, read, brain, word, expect, time, technolog, chang, sens, structur
Topic 10	tast, citi, canton, lemon, experi, imag, receptor, solut, sour, compani
Topic 11	fish, call, emot, school, femal, individu, sound, live, squirrel, complet
Topic 12	group, member, societi, myth, number, cultur, season, particip, rule, peopl
Topic 13	muscl, time, walk, peopl, chang, meal, start, power, minut, mind
Topic 14	posit, mother, read, hous, year, didn, grandpa, work, small, home
Topic 15	time, work, materi, peopl, sens, wood, camper, food, read, high
Topic 16	world, blind, hazard, billup, book, dream, music, school, watch, began
Topic 17	student, depart, spend, read, spent, actual, time, alloc, show, advertis
Topic 18	dream, school, individu, tara, church, event, lydia, time, TRUE, love
Topic 19	question, knowledg, scienc, answer, time, direct, mathemat, mediat, role
Topic 20	peopl, interact, person, demand, expect, find, futur, plane, anxieti, qualiti
Topic 21	live, popul, music, number, peopl, reflect, scale, trout, fact, hour
Topic 22	music, time, sens, anim, memori, disadvantag, creat, human, peopl, amnesia
Topic 23	write, chair, good, idea, object, make, understand, concept, reward, point
Topic 24	anim, seed, fish, natur, predat, bear, school, seal, larg, small
Topic 25	percentag, women, employ, nonagricultur, wage, number, year, increas, africa
Topic 26	school, high, earn, lifetim, median, particip, footbal, profession, subject, univers
Topic 27	chris, park, janet, famili, grey, home, piec, work, offic, fenc
Topic 28	scienc, scientif, polici, inform, failur, social, busi, peopl, evid, explain
Topic 29	peopl, success, lake, great, time, live, bottl, messag, salt, thing
Topic 30	consum, camera, electr, generat, imag, earli, film, radio, sourc, world

4.2 다중선형회귀분석을 이용한 정답률 예측

본 연구에서는 사전 실험을 통해 EBS 연계 이전과 이후 데이터의 특징이 확연히 다르게 나타남을 확인하였다. 최근 공교육 정상화의 일환으로 EBS 연계가 중요하게 다루어지고 있는 여건하에서는 EBS 연계 정책 실시 이후의 데이터에 대한 검증이 보다 의미 있을 것으로 판단하여 이후 정답률 및 난이도 예측 모델링에서는 EBS 연계 정책 이후 데이터만을 이용하여 예측 모델을 구축한 뒤, 유의미한 변수 파악 및 예측 성능 평가를 수행하였다. 결론적으로 2011년부터 2013년까지의 데이터를 학습 데이터로 사용하여 주요 변수 선택 및 학습을 수행하였으며 2014년 데이터를 이용하여 모델의 성능을 평가하였다. 먼저 총 58개의 독립변수 중 전진선택, 후방소거, 단계적 선택법을 이용하여 유의미한 변수를 선택하였다. 앞서 언급된 세 가지 변수 선택 기법은 학습 데이터의 변화에 따라 선택되는 변수가 달라질 수 있기

때문에 학습데이터를 다시 모델 구축용과 검증용으로 각각 70% 및 30%로 무작위 분할한 뒤, 변수선택 기법별로 유의확률별 가중치(0.001보다 작을 경우 3, 0.01보다 작을 경우 2, 0.05보다 작을 경우 1)를 계산하는 절차를 30회 반복수행하여 최종적인 독립변수 집합을 구성하였다.

Table 6. Selected Variables by Multiple Linear Regression (T12, T14, and T19 are the Main Topics of the Text Discovered by LDA)

No.	Variable Meaning	Variable	Related Hypothesis
1	diff.point = each point - average point of the total problems	Diff.Point	4
2	inference of blank word	BlankWord	2
3	grammar, vocabulary	Grammar	2
4	content correspondence	Correspondence	2
5	subject, title	Subject	2
6	Topic number 14	T14	3
7	mood of paragraph	Mood	2
8	count of words in the paragraph	Count.Word	1
9	- whether the problem related with EBS contents - 0: non correspondence 1: correspondence	EBS	4
10	Topic number 19	T19	3
11	Topic number 12	T12	3
12	Problem position (0~100: prior~end)	Prob.Position	4
13	count of poly syllable	Count.Syllable	1

각 변수 선택 기법의 30회 반복 수행을 통해 총 13개의 유의미한 변수들이 선정되었으며 각 변수명 및 관련 가설은 Table 6에 나타난 바와 같다. 변수 선택 결과 첫 번째 가설인 지문의 문장 구조 및 사용된 단어의 난이도와 관련해서는 “단어 수” 및 “음절 수” 두 변수가 유의미한 것으로 판별되었으며, 두 번째 가설인 문제 유형과 관련해서는 “빈칸추론 문제”, “어법/어휘 문제”, “내용 일치/불일치 문제”, “주제/제목/지칭 문제” 및 “분위기 문제”의 다섯 변수가 유의미한 것으로 판별되었다. 세 번째 가설인 지문의 내용 난이도(주제)와 관련해서는 12번(사회 관련), 14번(가족 관련), 19번(과학 관련) 주제가 정답률에 유의미한 영향을 미치는 것으로 나타났다. 마지막으로 네 번째 가설인 문항 속성과 관련해서는 배점 차이, 문제 위치 및 EBS 연계 여부 세 변수 모두 정답률과 유의미한 관계가 있는 것으로 파악되었다.

선택된 입력 변수들의 집합을 바탕으로 전체 학습데이터에 적용한 최종 다중선형회귀분석 모델의 계수 및 유의확률은 Table 7에 나타난 바와 같으며, 실제 정답률과 예측된 정답률 사이의 산포도는 Fig. 2에 나타난 바와 같다. 모델의

Table 7. Estimated Variables of Correct Answer Rate Prediction Model

Variables	Estimate	Std. Error	t value	Pr(>|t|)	Signif. codes
coefficient	0.6322227	0.0354650	17.827	< 2e-16	***
Diff.Point	-0.1224359	0.0191699	-6.387	5.92e-10	***
Blank.Word	-0.1281933	0.0241374	-5.311	2.04e-07	***
Grammer	-0.1209579	0.0277888	-4.353	1.81e-05	***
Correspondence	0.1597225	0.0236810	6.745	7.12e-11	***
Subject	0.0856779	0.0217560	3.938	0.000101	***
T14	0.0846276	0.0350574	2.414	0.016337	*
Mood	0.1283670	0.0393323	3.264	0.001218	**
Count.Word	0.0011753	0.0003354	3.504	0.000524	***
EBS	0.0413038	0.0136917	3.017	0.009881	**
T19	-0.0885829	0.0341309	-2.595	0.009881	**
T12	0.1119184	0.0406167	2.755	0.006194	**
Prob.Position	-0.0007376	0.0003301	-2.235	0.026126	*
Count.Syllable	-0.0004661	0.0002483	-1.877	0.061442	.

설명력인 수정 R 제곱합은 0.6285 수준으로 매우 높다고 할 수는 없으나 어느 정도 유의미한 관계를 찾아내고 있다고 판단할 수 있으며, 선택된 모든 변수들이 유의수준 0.1에서 통계적으로 유의미한 결과를 나타내고 있음을 확인할 수 있다. 각 설명변수와 정답률 사이의 관계를 살펴보면 정답률을 낮추는 요인으로는 “배점 차이”, “빈칸추론 문제”, “어법/어휘 문제”, “19번 주제(과학 관련)”, “문제 위치”, “음절 수” 등이 판별되었다. 일반적으로 배점의 경우 출제자가 난이도를 고려해서 부과하게 되는데 배점이 높을수록 정답률이 낮게 나타나는 결과는 출제자들이 부과하는 난이도가 어느 정도 적정하게 부과되었다고 볼 수 있다. 빈칸추론 문제와 어법/어휘 문제는 수험생들에게 어려운 문제로 알려져 있는데 실험결과 역시 정답률을 낮추는 요인으로 나타났다. 토픽 모델링을 통해 도출된 19번 주제의 대표 단어는 ‘question’, ‘knowledge’, ‘science’ 등이 있으며 이는 주로 과학 상식에 관련된 문제들이며, 이 주제에 대해서는 정답률이 낮게 나타나는 것을 알 수 있다. 문제 위치의 경우 뒷 번호에 배치된 문제일수록 정답률이 낮게 나타났는데, 이는 앞부분부터 문제를 풀다가 시간이 부족하거나 심리적인 요인으로 정답률이 낮은 것으로 추정된다. 또한 음절 수가 많을수록 역시 정답률이 낮게 나타남을 확인하였다.

반면, 정답률을 높이는 요인들로는 “내용 일치/불일치 문제”, “주제/제목/지칭 문제”, “분위기 문제”, “단어 수”, “EBS 연계 여부”, “14번 주제(가족 관련)”, “12번 주제(사회 관련)”로 나타났다. 문제 유형 중 내용 일치/불일치 문제, 주제/제목/지칭 문제, 분위기 문제는 상대적으로 쉬운 문제로 받아들여 정답률이 높게 나타났다. 흥미로운 점은 지문에 사용된 단어의 수가 많을수록 정답률이 높게 나타났는데, 이는 단어 수가 많아서 지문이 길어지면 정답을 유추할 수 있는 힌트 내용이 더 많이 들어가 정답률이 높을 수 있기

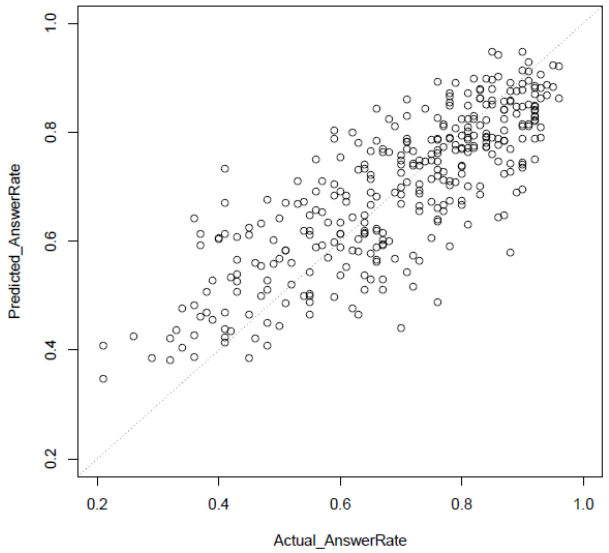

Fig. 3. Actual (X-axis) and Predicted (Y-axis) Correct Answer Rates

Table 8. Prediction Performance of Multiple Linear Regression

Evaluation Indicators	MSE	RMSE	MAE	MAPE
Values	0.0224	0.1495	0.1281	15.97%

때문이다. 또한 EBS와 연계된 문제일 경우 정답률이 높게 나타났으며 지문의 주제 측면에서는 가족 관련 내용과 사회 관련 내용에 대한 문제들의 정답률이 높게 나타나는 것으로 판별되었다.

다중선형회귀분석을 이용한 정답률 예측 정확도는 Table 8에 나타난 바와 같다. 네 가지 평가 지표 중 직관적으로 이해할 수 있는 지표는 MAE와 MAPE이며, MAE가 0.1281이라는 것을 평균적으로 실제 정답률과 예측된 정답률 사이의 오차가 12.81%로서 상당히 정확한 수준으로 예측을 하고 있음을 알 수 있다. 또한 MAPE가 15.97%로 나타나는데, 이는 실제 정답률 대비 오차 비율이 15.97%라는 의미이며 이는 실제 정답률이 80%인 문제에 대해서 69~93%의 범위로 정답률을 예측할 수 있음을 의미한다.

4.3 CART를 이용한 난이도 예측

두 번째로 구축한 예측 모델은 CART를 이용하여 종속변수를 정답률 자체가 아닌 난이도의 상/하로 변환한 분류 모델이다. 다중선형회귀 분석을 설명변수와 종속변수 사이의 관계를 선형으로 가정하는 제약이 있기 때문에 비선형 관계를 추정하고자 하는 목적과 예측 결과에 대한 설명을 제공하고자 하는 목적을 동시에 달성하고자 여러 분류 알고리즘 중에서 CART를 선택하였다. 각 문항에 대해 정답률 70% 미만일 경우 난이도를 어려움(Difficult)으로 정의하고 70% 이상인 문항에 대해서는 난이도를 쉬움(Easy)으로 정의하였다.

Table 9. Selected Variables by CART (T7 and T23 are the Main Topics of the Text Discovered by LDA)

No.	Variable	Related Hypothesis
1	Diff.Point	4
2	BlankWord	2
3	Grammar	2
4	Summary	2
5	Sequence	2
6	T7	3
7	T23	3

CART를 이용하여 난이도 분류 모델을 구축한 결과, 총 7개의 설명 변수가 선택되었으며 해당 변수들은 Table 9에 나타난 바와 같다. 선형회귀분석과는 다르게 첫 번째 가설인 문장 또는 지문 난이도와 관련된 변수들은 선택되지 않았으며 문제 유형과 관련된 변수가 전체의 50% 이상을 차지하는 것을 볼 수 있다. Table 10은 검증 데이터에 대한 CART의 분류 성능을 나타낸 것이다. 검증 데이터에 분류 모델을 적용한 결과, CART는 쉬운 문제에 대한 정확도(TER)는 0.8750을 나타내며 어려운 문제에 정확도(TDR)는 0.8529로서 난이도의 어려움과 쉬움에 대한 예측 정확도의 편차가 작게 나타남을 확인할 수 있다. 이는 Accuracy와 BCR에서도 확인할 수 있는 사항으로, 두 지표가 각각 0.8571, 0.8638로써 큰 차이가 나지 않는다. 만일 한 범주에 대한 정확도가 다른 범주에 대한 정확도와 크게 차이가 날 경우 BCR의 값은 매우 낮게 나타나는 경향이 있는데 본 연구에서 CART를 이용하여 구축한 분류 모델은 난이도에 관계없이 일정 수준 이상의 분류 정확도를 나타내는 것으로 확인되었다.

Table 10. Classification Accuracies of CART

Evaluation Indicators	TDR	TER	Accuracy	BCR
Values	0.8529	0.8750	0.8571	0.8638

CART에 의해 구축된 난이도 분류 의사결정 나무는 Fig. 4에 나타난 바와 같다. 이를 통해 문제 난이도의 어려움/쉬움에 대한 근거를 도출할 수 있다. 예를 들어, 그림에서 가장 왼쪽의 말단 노드의 경우, 총 176문항이 해당되는데 문제에 배당된 점수(Diff.Point)가 낮고, 문제 유형은 빈칸 유형 문제(BlankWord), 어법/어휘 문제(Grammar), 요약 문제(Summary), 또는 글의 흐름/순서/삽입(Sequence)이 아닐 경우, 약 85% 확률로 쉬운(Easy) 문제로 판별된다. 반대로 그림의 가장 오른쪽 말단 노드의 경우에는 총 일곱 문항이 해당되고 문제 배당 점수가 높으며 지문의 주제가 7번 주제(T7)에 해당할 경우 75%의 확률로 어려운(Difficult) 문제로 판별되는 것을 알 수 있다.

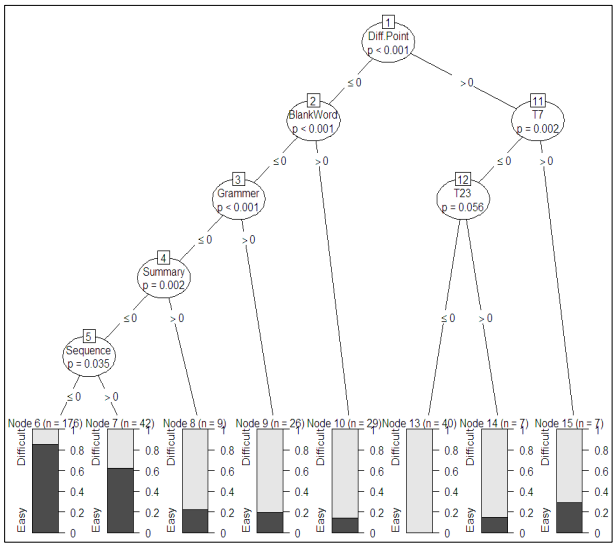

Fig. 4. Generated Tree by CART

4.4 예측 모델 비교 및 토의

Table 11은 다중선형회귀분석과 CART에서 선택된 설명변수들을 비교한 표이다. 배점 차이, 빈칸추론 문제 및 이법/어휘 문제의 세 변수는 두 예측 모델에서 공통으로 도출된 것으로 볼 때, 수능 영어 정답률에 중요한 영향을 미치는 독립변수이므로 이를 활용하여 문제의 난이도 조절 및 정답률 예측을 할 수 있을 것으로 판단된다. 또한, 다중선형회귀분석과 CART의 공통변수로 도출된 것은 아니나 토픽 모델링에 의해서 분류된 변수들이 다중선형회귀분석과 CART에서 유의미한 변수로 도출된 것은, 지문의 내용을 전문가의 구분이 아닌 데이터분석 모델에 의해 자동 분류할 수 있고, 이러한 분류를 통해 정답률을 예측하는 모델에 사용할 수도 있다는 시사점을 주는 의미 있는 결과라고 판단된다.

Table 11. Comparison Between the Selected Variables by Multiple Linear Regression and CART

No.	Type	Multiple Linear Regression	CART	Related Hypothesis
1	common	Diff.Point		4
2		Blank.Word		2
3		Grammer		2
4	difference	Correspondence	Summary	2
5		Subject	Sequence	2
6		T14	T7	3
7		Mood	T23	2
8		Count.Word		1
9		EBS		4
10		T19		3
11		T12		3
12		Prob.Position		4
13		Count.Syllable		1

5. 결론 및 향후 계획

본 연구에서는 수능 영어영역의 정답률에 영향을 미치는 요인을 도출하고 영어문제의 지문만으로 정답률을 예측하기 위해 2004년부터 2014년 수능 및 모의고사 문제와 교과서 단어를 수집하고 영향을 미칠 것으로 예상되는 58개의 독립변수를 선정하여 다중선형회귀분석과 CART에 의한 정답률 예측 및 난이도 분류 모델을 구축하고 그 효과를 검증하였다.

실험 결과 다중선형회귀분석에서는 두 가지 예측 모델에서 공통적으로 선택된 유의미한 설명변수로는 배점 차이, 빈칸추론 문제, 어법/어휘 문제의 세 개가 판별되었으며, 이는 문항 자체의 특성과 문제 유형이 정답률 및 난이도에 영향을 미칠 것이라는 가설을 뒷받침해주는 결과라고 볼 수 있다. 이외에도 다중선형회귀분석에서는 문제 유형과 관련하여 세 가지 설명변수(내용 일치/불일치, 주제/제목/지칭, 분위기)가 추가적으로 유의미한 것으로 판별되었으며, 문장 및 단어의 난이도와 관련된 변수인 단어 수 및 음절 수 또한 중요한 것으로 판별되었다. 또한 토픽 모델링에 의해 분류된 토픽 중 세 개의 토픽이 정답률 예측에 유의미한 것으로 판별되었고, 문항의 특성인 EBS 연계 여부 및 문제 위치 역시 정답률과 높은 연관 관계가 있는 것으로 파악되었다. 이렇게 선택된 총 13개의 설명변수로 정답률 예측 모델을 구축한 결과 MAPE 기준 15% 내외의 예측 오차를 나타내는 것을 확인할 수 있었으며, 이는 상당히 낮은 수준의 예측 오차라고 할 수 있다. 정답률이 아닌 난이도의 쉬움/어려움을 예측하는 분류 문제의 경우, CART에 의해 총 7개의 변수만을 이용하여 85% 이상의 분류 정확도를 얻을 수 있다는 것이 확인되었다. 문제 유형과 관련된 두 변수(요약, 흐름/순서/삽입)와 지문 주제 관련 두 변수(7번 주제, 23번 주제)는 선형회귀분석과 공통적으로 선택된 변수 이외에 CART에서만 선택된 설명변수로 확인되었다.

이러한 실험 결과를 통해 기존 연구 결과인 장경숙(2004)의 연구 결과와 같이 빈칸추론 문제가 정답률을 낮추는 요인이라는 것을 확인할 수 있었으며 문장구조의 복잡도는 정답률에 큰 영향을 미치지 않는 것 또한 확인하였다. 또한 토픽 모델링을 통한 지문 주제 분류를 통해 일상적인 사회나 가정에 대한 주제는 정답률과 양의 상관관계를 보였고 전문지식은 음의 상관관계를 보인 것으로 지문의 주제가 정답률에 영향을 미칠 수 있다는 것을 확인할 수 있었다. 이 결과는 이경숙(1994)의 지문 친숙도 연구와 유사한 결과이나 알고리즘을 통해 주제를 세분화하여 한층 심도 있는 결과를 도출했다는 것에 의의가 있다고 할 수 있다. 또한 기존 설문지 기반 선형회귀분석을 수행한 장경숙(2004)의 연구에서 나타난 설명력 24%보다 우수한 63%의 설명력을 얻을 수 있었다.

본 연구에서 구축된 정답률 및 난이도 예측 모델은 전문가의 지식 없이 영어영역 문제의 특성만을 이용하여 일정 수준 이상의 예측력을 갖는 정답률 및 난이도 예측 모델을 구축했다는 것에 의의를 둘 수 있다. 구축된 예측 모델을

통해 출제자들이 사전에 모의실험을 수행하여 문항별 난이도 예측 및 전체 시험의 난이도를 예측하고 검토함으로써 전반적인 난이도 조절 및 출제의 방향을 보완하는 데 보조적인 도구로 사용될 수 있다. 그러나 본 연구는 수능 영어 지문에 한정하여 텍스트마이닝과 다중선형회귀분석 및 CART를 통한 예측 모델을 구축하였으므로 향후 문제와 문항을 추가한 연구가 필요하다. 또한 지문을 내용에 따라 분류하는 방법으로 본 연구에서는 토픽 모델링을 사용했으나 이에 더하여 텍스트마이닝 분야의 다양한 방법론을 차용하여 정확도를 향상시킬 수 있을 것이다.

References

[1] 2015 school year the CSAT questions headquarters, "2015 school year, the CSAT Press," in *Proceedings 2015 school year the CSAT questions headquarter*, 2014.

[2] Korea Institute for Curriculum and Evaluation, "2015 school year CSAT score results press release," in *Proceedings Korea Institute for Curriculum and Evaluation*, 2014.

[3] Korea Institute for Curriculum and Evaluation, "2015 school year CSAT plan," in *Proceedings Korea Institute for Curriculum and Evaluation*, 2014.

[4] T. C. Kang, "CSAT Improvement Study," *Ministry of Education*, pp.57-77, 2013.

[5] M. K. Kang and Y. M. Kim, "The internal analysis of the validation on item-types of Foreign (English) Language Domain of the current 2005 CSAT for designing the level-differentiated English tests of the 2014 CSAT," *Journal of the Korea English Education Society*, Vol.12, No2, pp.1-35, 2013.

[6] K. S. Lee, "The effects of th number of questions per passage, the length of passage, and the topic familiarity on multiple-choice English listening and reading comprehension tests," *English Teaching*, Vol.54, No.4, pp.327-351, 1999.

[7] N. B. Kim, "A corpus-based lexical analysis of the foreign language(English) test for the college scholastic ability test (CSAT)," *English Language & Literature Teaching*, Vol.14, No.4, pp.201-221, 2008.

[8] K. S. Chang, "A model of predicting item difficulty of the reading test of College Scholastic Ability Test," *Foreign Languages Education*, Vol.11, No.1, pp.111-130, 2004.

[9] Y. M. Sung, "Factor Analysis of English Test Scores in the College Scholastic Ability Test and Implications," Ph.D. dissertation, Inha University Graduate School, 2003.

[10] H. W. Lee and S. Y. Lee, "A study on the relationship between the scores of TOEFIC, TOEIC and TEPS, and college academic performance," *English Language & Literature Teaching*, Vol.9, No.1, pp.153-171, 2003.

[11] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and Regression Trees," Wadsworth, 1984.

[12] D. Hand, H. Mannila, and P. Smyth, "Principles of Data Mining," A Bradford Book The MIT Press, 2001.

[13] F. Sebstiai, "Machine learning in automated text categorization," *ACM Computing Surverys*, Vol.34, No.1, 2002.

[14] J. H. Bae, J. E. Son, and M. Song, "Analysis of twitter for 2012 South Korea presidential election by text mining techniques," *Journal of Intelligent Information Systems*, Vol.19, No.3, pp.141-156, 2013.

[15] H. J. Lee and J. C. Park, "Probabilistic filtering for a biological knowledge discovery system with text mining and automatic inference," *Journal of the Korea Society of Computer and Information*, Vol.17, No.2, pp.139-147, 2012.

[16] D. Blei, "Probabilistic topic models," *Communications of the ACM*, Vol.55, No.4, pp.77-84, 2012.

[17] S. R. Kang, "A Study on the Readability of English Textbooks: Middle School English 1 and 2 Based on the Revised 7th English National Curriculum," Master Dissertation, Inha University Graduate School, 2010.

박 희 진

e-mail : rhaos@seoultech.ac.kr
1993년 인하대학교 전자재료공학과(학사)
2002년 연세대학교 컴퓨터공학과(석사)
2013년 KAIST EMBA(석사)
2014년~현 재 서울과학기술대학교 IT정책전문대학원 산업정보시스템전공 박사과정
관심분야 : Data mining and Security

장 경 애

e-mail : jkalove@hanmail.net
1996년 대구대학교 문헌정보학과(학사)
2014년 연세대학교 컴퓨터공학과(석사)
2014년~현 재 서울과학기술대학교 IT정책전문대학원 산업정보시스템전공 박사과정
관심분야 : 데이터 품질/분석, 최적화 등

이 윤 호

e-mail : younholee@seoultech.ac.kr
2000년 KAIST 전산학과(학사)
2002년 KAIST 전산학과(석사)
2006년 KAIST 전산학과(박사)
2009년~2013년 영남대학교 정보통신공학과 조교수
현 재 서울과학기술대학교 글로벌융합산업공학과 부교수
관심분야 : 응용암호, 데이터보안, 시스템보안

김 우 제

e-mail : wjkim@seoultech.ac.kr
1986년 서울대학교 산업공학과(학사)
1988년 서울대학교 산업공학과(석사)
1994년 서울대학교 산업공학과(박사)
1988년~1991년 동양경제연구소 연구원
1999년~2001년 University of Michigan Visiting Scholar
2003년~현 재 서울과학기술대학교 글로벌융합산업공학과 교수
관심분야 : IT서비스, 소프트웨어공학, 최적화, 스마트그리드 등

강 필 성

e-mail : pilsung_kang@korea.ac.kr
2003년 서울대학교 산업공학과(학사)
2010년 서울대학교 산업공학과(박사)
2011년 현대카드 CVM기획팀 과장
2012년~2014년 서울과학기술대학교 글로벌융합산업공학과 조교수
현 재 고려대학교 산업경영공학부 조교수
관심분야 : 데이터마이닝, 기계학습, 인공지능, 텍스트마이닝