영어 논술 자동 평가를 위한 언어 유창성 측정 방법

양민철○, 김민정, 임해창
고려대학교
{mcyang, mjkim, rim}@nlp.korea.ac.kr

Assessment of Writing Fluency For Automated English Essay Scoring

Min-Chul Yang○, Min-Jeong Kim, Hae-Chang Rim
Korea University

요 약

영어 논술 자동 평가 시스템은 수험자가 쓴 에세이에 대하여 전문 평가자가 직접 읽고 평가하는 방식에서 벗어나 웹상에서 자동으로 평가 받을 수 있는 실시간 시스템이다. 하지만 비영어권 수험자에게는 논리력 혹은 작문 능력보다 그것을 영어로 표현하는 유창성에서 더 큰 문제가 있을 수 있는데 기존 연구에서는 이런 측면에 대한 평가가 부족하였다. 본 연구에서는 보다 정확한 비영어권 수험자의 영어 논술 평가를 위해 어휘력, 문장 구조의 다양성, 문장의 혼잡도를 평가하여 언어 유창성에 집중된 기계학습 방법의 추가적인 자질을 제안한다. 실험 결과 전문 평가자의 점수와 1) 상관관계 2) 정확도 측면에서 제안하는 방법은 기존의 방법에 비해 더 나은 성능을 보였다.

주제어: 영어 논술 자동 평가 시스템, 유창성, 작문 평가, 기계학습

1. 서론

글로벌 시대가 되면서 전 세계적으로 영어는 가장 중요한 언어로 자리 잡고 있다. 특히 우리나라는 치열한 교육열 속에서 어려서부터 영어 교육을 시키는 등 영어 학습에 지대한 관심을 가지고 있다. 그래서 영어 실력을 인증 받기 위해 TOEIC, TEPS 등의 영어 인증 시험이 유행하고 있고 이런 인증 점수는 대학 입시, 회사 입사에 많은 영향을 끼치고 있다. 이 중 TOEFL, GRE, GMAT에서는 영어 논술 평가를 실시한다. 이 시험을 통해 수험자의 논리력, 글의 구성, 어휘력 등을 평가하여 최종적으로 영작문 능력을 측정한다. 최근 들어 실생활 영어가 중요해짐에 따라 영어 말하기, 쓰기 영역 또한 중요해지고 있다. 하지만 이런 영역들은 읽기, 듣기 영역에 비해 사람이 평가하기도 어려우며, 컴퓨터가 평가하기 위해서는 고도의 자연어 처리가 필요하다. 이와 관련된 응용 분야로 영어 논술 자동 평가가 있다.

영어 논술 자동 평가 시스템은 사람의 개입 없이 자동으로 시스템이 수험자의 영어 에세이에 대해서 점수를 내어주고 피드백을 주는 실시간 시스템이다. 현재 e-rater, IEA (Intelligent Essay Assessor) 등의 영어 논술 자동 평가 시스템은 웹상으로 특정 요금만 지불하면 이용가능하다[1, 2]. 특히 GMAT에서 e-rater는 전문 채점자의 평가와 더불어 최종 점수를 내주는데 사용되고 있다. 하지만 이런 시스템들은 상업적인 용도로 쓰이기 때문에 구체적인 연구 내용은 보고되지 않는 상태이다.

영어 논술 자동 평가 시스템의 장점은 다음과 같다. 1) 실시간으로 평가를 받을 수 있다. 전문 평가자와 같은 사람의 개입이 없기 때문에 언제든지 평가가 가능하고 처리속도 또한 빠르다. 2) 적은 비용으로도 평가 받을 수 있다. 자동 채점 시스템만 잘 갖춘다면 전문가의 비용이 들지 않으므로 적은 비용의 평가가 가능하다. 3) 사용 접근성이 용이하다. 전문 평가자와의 복잡한 연결 없이 웹상으로 손쉽게 접근이 가능하고 그 자리에서 평가 또는 피드백을 받을 수 있다.

이미 많은 연구에서 기존 영어 논술 자동 평가 시스템은 전문 평가자와 비슷한 결과를 내준다고 보고하였다[3, 4]. 하지만 이는 영어권 수험생 대상으로 실시하고 있고 우리나라 수험생과 같은 비영어권 수험생은 대상이 아니다. 이 둘의 가장 큰 다른 점은 영어권 수험생은 논리력 또는 글을 잘 쓰는 능력만 갖추면 좋은 점수를 받을 수 있지만 비영어권 수험생들은 이와 더불어 좋은 아이디어를 잘 표현/번역할 수 있는 유창성 또한 고려되어야 한다는 점이다. 특히 비영어권 수험생들은 모국어에 따라 오류 유형이 다르고, 자연스럽지 않은 혹은 원어민들이 잘 쓰지 않는 표현을 사용한다. 하지만 기존의 시스템은 이런 언어 유창성이 중요한 고려대상이 아니기 때문에 이에 대해 충분한 평가가 이루어지지 않고 있다. 그러므로 보다 정확한 비영어권 수험생의 영어 논술 평가를 위해 언어 유창성의 심도 있는 평가가 필요하다.

주제 : 좋은 이웃의 조건

잘 쓴 에세이 [5 점]
... to **interrupt** other's privacy ... to **interfere** in almost cases. ... **bothers** me ... not **meddling** in other people's concern ...

못 쓴 에세이 [2 점]
... **I will describe that** what are the **quality of good neighbors** ... **I would think that good neighbors quality** are ... **I always think that** ...

그림 1 영어 에세이 예제

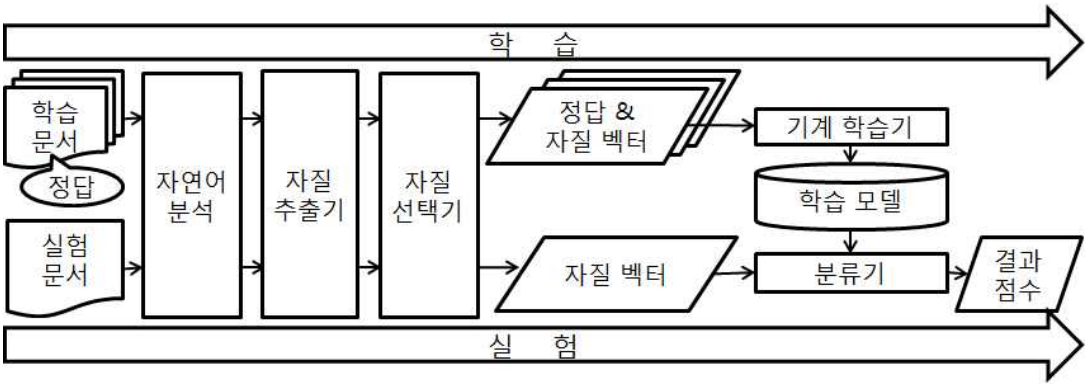

그림 2 시스템 구조

본 연구에서 언어 유창성의 정의는 다음과 같다. 글을 읽었을 때 글의 흐름이 자연스럽고[5], 원어민들이 사용하는 언어에 가까운 표현을 다양하게 사용하는 것이다. 이와 관련된 예제는 그림 1과 같다.

먼저 잘 쓴 에세이를 보면'간섭하다'의 뜻을 가진 자리에 동의어를 이용하여 같은 단어로 되풀이 사용하지 않았다. 하지만 못 쓴 에세이를 보면 같은 문장 구조(주어+조동사+동사+that)를 여러 번 반복 하였고 어휘 열 또한 반복 사용하였다. 이와 같이 같은 어휘 또는 문장 구조의 반복은 읽는 사람에게 글이 부자연스러움을 느낄 수 있고 영작문의 지침으로 보아도 반복 사용은 되도록 지양하는 편이다. 또한'neighbors'와'quality'간의 수일치가 이루어지지 않는 문법적 오류도 보였다. 이런 비영어권 수험생 에세이의 오류를 분석하여 이를 잘 반영한 언어 유창성 측정 방법을 제안하고자 한다.

이후 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련된 기존 연구를 살펴보고 3장에서는 기계학습 기반의 언어 유창성 측정 방법에 대해 설명한다. 4장은 제안하는 방법의 결과와 전문 평가자 점수 간 상관관계와 정확도에 대한 실험 및 평가를 보여준다. 마지막으로 5장에서는 결론과 향후 연구에 대해 논의하도록 한다.

2. 관련연구

본 연구와 관련된 영어 논술 자동 평가에 대한 기존 연구를 살펴보면 다음과 같다. 초기의 연구는 문서 분류 방법을 그대로 적용하였다[6]. 사용한 자질로는 단어 수, 어휘 수, 문장 수, 특정 길이 이상 단어 수 등의 단순한 자질만 사용하여 문서의 표면적인 특성만 고려하였고 언어 유창성에 대한 평가는 하지 않았다.

영어 논술의 작문 스타일을 자동 평가한 연구[7]에서는 이전 연구에서 제안한 자질 이외에 각 단어 빈도수별 비율과 반복어 간 사이의 길이도 측정하였다. 시스템의 평가는 작문 스타일 점수만 정답으로 이용하여 유창성 평가에 중점을 두었다. 하지만 반복어구에만 집중 하였고 문서의 내재적인 언어 유창성에 대한 평가는 이루어지지는 못했다.

다음은 영어 논술 자동 평가 시스템 이외의 연구에서 언어 유창성을 평가하는 기존 연구에 대해 살펴본다.

문장 내 오류를 인식하는 연구[8]에서는 오류 인식의 자질로써 어휘, 품사 특정 N-gram 단위 혼잡도를 측정하였다. 하지만 영어 논술 자동 평가에서는 문서 전체 단위의 혼잡도 측정이 필요하고 비영어권 수험생의 전형적인 오류에 특화된 측정은 부족하였다.

말하기 유창성을 자동 평가한 연구[9]에서는 유사 3-gram의 빈도수 별 비율을 추가적으로 측정하였다. 유사 3-gram 비교 시 삽입, 삭제, 대체 등의 방법을 이용하여 유사도를 구하였지만 표면적으로 드러난 어휘 열에 대한 고려만 하였고 내재된 품사 열을 이용한 문장 구조에 대해서는 고려하지 않았다.

한국 대학생의 영어 작문 실수 유형을 분석한 연구[10]에서는 영어 에세이 등급 별로 자주 틀리는 오류에 대해 상관관계를 이용하여 분석하였다. 대부분 한국 대학생들은 관사, 수일치 등의 오류가 많았다. 이를 통해 한국 수험생들이 자주 범하는 오류가 언어 유창성과 많은 관련이 있으며 이를 파악하는데 중점을 두어 영어 논술 자동 평가시스템에 적용하고자 한다.

3. 언어 유창성 측정에 기반한 영어 논술 자동 평가 방법

본 연구에서 영어 논술 자동 평가는 기계 학습을 이용한 분류 방식으로 접근한다. 따라서 전체 시스템 구조는 그림 2와 같다. 영어 논술문서가 입력으로 주어지면 문장 분리, 형태소 분석, 품사 태깅과 같은 기본적인 자연어 처리 과정을 거친다. 자연어 처리 결과를 이용하여 학습 및 분류에 필요한 자질을 추출하며 이때 자질 선택기에서 유용한 자질을 선별한다. 학습 과정에서는 전문 평가자가 채점한 결과와 각 학습 문서의 자질들을 이용하여 분류 모델을 학습하며 실험 과정에서는 이 학습 모델을 이용하여 새로 입력된 각 실험 문서의 등급을 분류한다.

본 연구에서 언어 유창성을 영어 논술 자동 평가에 반영하기 위해 사용하는 자질은 다음과 같다.

3.1 어휘력 측정 자질

앞서 기술한 것과 같이 본 연구에서의 유창성은 다양한 표현을 사용하는 것을 포함한다. 따라서 본 연구에서는 작성자가 바꿔쓰기 (Parapharsing)를 능숙히 할 수 있

는 어휘력을 가지고 있는지 측정한다. 바꿔쓰기란 비슷한 뜻을 표현할 때 같은 단어를 반복적으로 사용하는 것이 아니라 동의어나 다른 표현을 사용하는 것이다. 이를 위해서는 어휘력이 풍부해야 하며, 실제로 그림 1의 예시와 같이 잘 쓴 에세이에서 많이 볼 수 있다. 본 연구에서는 같은 품사의 어휘들을 쌍으로 묶고 각 쌍의 관계가 유사 관계인지를 워드넷[11]을 이용하여 판별한다. 이때 고려하는 품사는 명사/명사구, 동사, 부사, 형용사의 어휘만으로 국한한다.

3.2 문장 구조의 다양성 측정 자질

그림 1의 예시와 같이 문장 구조가 단조로운 글을 판별하기 위해 문서 내 문장 구조의 다양성을 측정한다. 문장 구조를 파악하기 위해 품사 열을 이용하며, 다양성 측정은 특정 품사 N-gram의 반복 빈도수로 측정한다. 품사는 어휘와 달리 가짓수가 한정 되어 있으므로 짧은 N-gram으로는 모호할 수 있고, 긴 N-gram을 사용하여도 자료 부족 문제가 심각하지 않다. 따라서 어휘 열 2-gram과 품사 열 3-gram을 이용하였다. 이 자질을 통해 읽는 사람이 부자연스럽게 느낄 수 있는 특정 N-gram의 과도한 반복을 파악할 수 있다.

3.3 문서의 혼잡도 측정 자질

작문 실력이 검증된 사람이 쓴 영어와 비슷한 영어로 썼다면 그 영어 에세이는 유창한 영어로 썼다고 볼 수 있다. 이러한 직관을 반영하기 위해 본 연구에서는 대량의 외부 문서 집합으로 학습한 언어 모델을 이용하여 에세이의 혼잡도 (Perplexity)를 측정하여 영어 논술 자동 평가의 자질로 이용한다. 어휘 열과 품사 열을 모두 고려하며 어휘 열은 3-gram, 품사 열은 5-gram을 각각 언어 모델의 단위로 사용한다. 혼잡도는 이전 열과 얼마나 어울리는지 측정하는데 언어 모델의 학습 데이터가 많을수록 좋은 성능을 보인다. 어휘 열 언어 모델은 수동태, 수일치, 전치사 등의 주로 문법적인 오류 인식이 가능하며, 품사 열 언어 모델은 문장 구조 등의 오류 인식이 가능하다.

3.4 자질 집합

본 연구에서 사용한 전체 자질 집합은 다음 표 1과 같다. 이 표는 기존 연구에서 사용한 주요 자질[F_1, F_2, F_3]과 3.1절 ~ 3.3절에서 설명한 제안하는 자질[F_4]을 모두 포함한다.

표 1 자질 설명

분류	설명
표층적 자질 (F_1)	[글자/단어/문장] 수
스타일 자질 (F_2)	어휘 수 [단어/문장] 평균 길이 [어휘/특정 길이 이상 단어] 수 품사 별 [단어/어휘] 수 고급 사전 매칭 [단어/어휘] 수 특정 어구 [단어/어휘] 수
	어휘 밀도
유창성 자질 (F_3)	반복어 간 사이 거리 출현 빈도 별 단어 어휘 수
제안하는 유창성 자질 (F_4)	품사 별 유사 관계인 단어 쌍의 수 출현 빈도 별 [어휘/품사] N-gram 수 [어휘 열/품사 열] 혼잡도

먼저 표층적 자질은 자연어 분석이 필요 없어도 추출할 수 있는 표면적인 자질[6]이다. 스타일 자질은 작성자의 작문 스타일에 관련된 자질로 대부분 사용하는 어휘의 특성에 관한 자질들로 묶여 있다. 여기서 고급 사전과 매칭되는 단어 및 어휘 수는 고등학생 수준의 사전과 GRE 시험 대비용 사전을 이용하여 얼마나 어려운 단어를 사용하는지 측정하는 자질이다. 또한 언어 밀도는 '어휘 수 / 단어 수'로 구하며 대부분 문서 품질을 측정할 때 가장 유용한 자질로 알려져 있다. 유창성 자질은 1장에서 정의한 유창성을 측정하는 자질로 기존 연구[7]에서 제안한 자질[F_3]과 본 연구에서 제안한 자질[F_4]을 따로 분류하였다.

최종적으로 기계학습에 사용되는 자질 값은 각 자질별로 파악하고자 하는 특성에 맞게 추출한다. 예를 들면 혼잡도 자질에서는 문서 내 문장 혼잡도의 평균을 구하여 문서의 대표 값을 구하고, 특정 임계 값 이상의 문장 수를 구하여 언어 모델과 어울리지 않는 문장 수를 구한다. 또한 문서 내 문장 혼잡도의 최대/최소 값과 분산을 구하여 평균만을 통해서 알 수 없는 다른 특성에 대해서도 파악이 가능하다. 또 다른 예로 많이 쓰이는 방식은 문서의 특성에 영향을 미치지 않기 위해 단어 수에 기반한 절대적인 수치가 아닌 정규화를 통한 상대적인 수치를 이용하는 방법이다.

4. 실험 및 평가

3장에서 제시한 언어 유창성이 영어 자동 평가에 얼마나 영향을 미치고 제안한 자질의 유용성을 평가하기 위해 시스템 결과와 전문 평가자 점수 간 상관관계와 정확도를 평가하였다. 본 장에서는 먼저 학습 데이터 구축 등 실험 환경에 대해 살펴보고 실험 결과를 소개한 후 이에 대한 분석을 하도록 한다.

4.1 실험 환경

실험에 사용된 문서들은 모두 한국 수험생들이 작성한 것으로 총 10개의 주제에 대한 영어 에세이이다. 각 주제 별 에세이 수는 차이가 있지만 모두 3,093개로 두 명의 전문 평가자에 의해 최종 점수와 유창성이나 작문 스타일만 고려한 세부 점수까지 채점 되었다. 등급 점수는 TOEFL Writing채점 양식에 맞게 1점 ~ 6점까지의 범위로 채점하였다. 분류 모델의 학습과 시스템 결과의 평가에 사용하는 점수는 두 채점자의 평균의 값의 올림을 사용하였다.

3.3절의 언어 모델을 구성하는데 사용한 자원은 NEWS 데이터로 약 3300만 문장으로 구성되어 있다. 언어 모델

구축과 혼잡도 계산은 SRILM toolkit[12]을 이용하였다.

분류 모델을 위한 기계 학습 방법은 Maximum Entropy[13]를 이용하였다. 이 방법은 경험적으로 얻어진 확률 분포로부터 가능한 정답에 가까운 모델을 찾는데 제한된 자질 안에서 분류하는 방법에 적합하다. 각 실험 환경 별로 최적의 상태를 찾기 위하여 상관관계 기반 자질 선택을 적용하였고, 각 자질 값 간 범위를 맞추어주기 위해 순위 기반 정규화 방법을 사용하였다. 또한 학습 문서와 실험 문서를 분류하기 위해 10-묶음 교차 검증 방법을 이용하였다.

4.2 실험 결과 및 분석

각 분류 방법이 얼마나 전문 평가자와 유사하게 채점하였는지 비교하기 위하여 다음 식 (1), (2), (3)과 같이 상관관계와 정확도를 평가한다.

$$Correlation(X, Y) = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y} \quad (1)$$

E, μ: 기대 값, σ: 표준편차

$$Exact\ Match = \frac{D_0}{N} \quad (2)$$

$$Adjacent\ Match = \frac{D_{-1} + D_0 + D_1}{N} \quad (3)$$

N: 전체 문서 개수
D_n: [(시스템 출력 − 정답) = n]인 문서 개수

상관관계는 시스템이 출력한 결과와 정답 간 유사한 정도의 통계적 차이를 계산한다. 영어 논술 자동 평가에서 식(1)의 상관관계는 시스템과 전문 평가자가 잘 쓴 에세이에는 좋은 점수로 채점했는지, 못 쓴 에세이에는 안 좋은 점수로 채점하는지를 서로 비교하여 얼마나 유사한 경향으로 채점하였는지 검증할 수 있다. 이 평가 방법은 영어 논술 자동 평가 성능 측정에서 가장 많이 사용되는 방법이다. 정확도는 두 가지로 나누어질 수 있는데 정확하게 맞추었는지[식(2)], 아니면 정답과 1점 이내로 맞추었는지[식(3)] 평가한다. 식(3)의 방법은 전체 점수가 6등급으로 나누어졌기 때문에 정확하게 맞추는 것이 어려워서 정답의 범위를 넓힌 것으로 영어 논술 자동 평가의 성능 측정에서 흔히 사용되는 평가 방법이다.

표 2 전체 점수에 대한 성능

사용 자질	Correlation	Exact	Adjacent
F_1+F_2	0.4890	64.89%	98.74%
F_1+F_2+F_3+F_4	0.5054	65.60%	98.80%
향상 폭	+ 3.36%	+ 1.10%	+ 0.07%

표 2는 유창성 자질이 영어 논술 자동 평가 시스템에 얼마나 도움이 되는지 살펴보기 위한 실험 결과이다. 그래서 기존 연구에서 제안한 자질 중 유창성 자질을 제외한 실험[F_1+F_2]과 유창성 자질까지 포함한 실험[F_1+F_2+F_3+F_4] 결과를 비교하였다. 모든 평가 방법에서 소폭의 성능 향상을 보였는데 이것은 영어 논술 평가에서 유창성을 평가하는 부분이 의미가 있음을 알 수 있다. 두 실험 모두 Adjacent Match 결과를 볼 때 1점 이내로는 거의 맞춤을 볼 수 있다.

표 3 세부 점수에 대한 성능

사용 자질	Correlation	Exact	Adjacent
F_3	0.2020	55.42%	98.06%
F_3+F_4	0.2270	56.81%	98.12%
향상 폭	+ 12.39%	+ 2.51%	+ 0.07%

TOEFL Writing의 평가 기준 중 [언어 유창성/어휘 수준/작문 스타일]만 평가하는 세부 점수를 정답으로 둔 비교 실험 결과는 표 3과 같다. 이 실험은 유창성 자질 중에서 본 연구에서 제안한 자질이 얼마나 유용한지 살펴보기 위함이다. 그래서 기존 연구에서 제안한 유창성 자질만 이용한 실험[F_3]과 본 연구에서 제안한 자질까지 포함한 실험[F_3+F_4] 결과를 비교하였다. 표 3을 보면 유창성 자질로만 평가하는데 제한이 있어서 전체적인 성능은 대체로 낮지만 모든 평가 방법에서 성능 향상을 보였고 이를 통해 제안하는 자질이 유창성 측정에 도움이 됨을 알 수 있었다.

5. 결론

본 논문에서는 비영어권 수험생들의 영어 논술을 보다 정확하게 자동 평가하기 위하여 언어 유창성 평가에 집중된 측정 방법을 제안하였다. 제안하는 방법은 크게 작성자의 어휘력, 에세이의 문장 구조 다양성, 혼잡도를 계산하여 기존 연구 방법에 추가하였다.

실제 비영어권인 한국 수험생들을 대상으로 실험해보았고, 이 실험을 통해 언어 유창성 자질들이 영어 논술 자동 평가의 성능 향상시킴을 알 수 있었다. 또한 제안하는 유창성 자질들이 기존 연구에서 제안한 유창성 자질보다 더 성능 향상에 도움이 됨을 알 수 있었다. 이런 결과는 비영어권 수험생들이 자신의 생각을 영어로 표현/번역하는 과정에서도 문제가 있음을 알 수 있었고 영어 논술 자동 평가 시스템에서도 이를 반영해주어야 한다. 또한 제안하는 시스템은 영어 유창성 평가만 할 수 있는 시스템으로 변형이 가능하기 때문에 영어 학습용으로도 활용이 가능하다.

향후에는 언어 모델 구축 시 에세이 성격과 비슷한 대용량의 문서집합을 이용할 것이고, 의미적 파싱 결과를 이용하여 기존보다 자세하게 문장 구조와 단어 간 의미적 관계를 추출하여 더 정확한 영어 논술 자동 평가 방법을 연구할 것이다.

참고문헌

[1] Attali & Burstein, Automated essay scoring with e-rater V.2, The Journal Of Technology, Learning and Assessment, Vol 4-3, 2006

[2] Foltz, Laham & Landauer, The intelligent essay assessor: Applications to educational technology, Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, Vol 1-2, 1999

[3] Wang & Brown, Automated Essay Scoring Versus Human Scoring: A Comparative Study, The Journal Of Technology, Learning and Assessment, Vol 6-2, 2007

[4] Valenti, Neri & Cucchiarelli, An overview of current research on automated essay grading, Journal of Information Technology Education, 2003

[5] Quinlan, Higgins & Wolff, Evaluating the Construct Conerage of the e-rater Scoring Engine, ETS Research Report, No. RR-09-01, 2009

[6] Larkey, Automatic Essay Grading Using Text categorization Techniques, SIGIR-98, pp. 90-95, 1998

[7] Burstein & Wolska, Toward Evaluation of Writing Style: Finding Overly Repetitive Word Use in Student Essays, EACL-03, pp. 35-42, 2003

[8] Sun, Liu, Cong et al., Detecting Erroneous Sentences using Automatically Mined Sequential Patterns, ACL-07, pp. 81-88, 2007

[9] Deshmukh, Kandhawy, Verma & Audhkhasi, Automatic Evaluation of Spoken English Fluency, ICASSP, 2009

[10] 임희석, 박종원 & 남기춘, 한국 대학생이 보이는 영어 작문 실수 유형, 한국 음성 학회, pp. 176-179, 2003

[11] Miller, Wordnet: a lexical database for English, Communications-ACM, vol 38-11, 1995

[12] Stolcke, SRILM - an extensible language modeling toolkit, ICSLP, pp. 901-904, 2002

[13] Adam, Stephen & Vincent, A Maximum Entropy Approach to Natural Language Processing, Computational Linguistics, vol 22-1, 1996