KIPS Tr. Software and Data Eng.
Vol.3, No.9 pp.381~388 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2014.3.9.381

Swear Word Detection and Unknown Word Classification for Automatic English Writing Assessment

Gyoung Ho Lee† · Sung Gwon Kim†† · Kong Joo Lee†††

ABSTRACT

In this paper, we deal with implementation issues of an unknown word classifier for middle-school level English writing test. We define the type of unknown words occurred in English text and discuss the detection process for unknown words. Also, we define the type of swear words occurred in students's English writings, and suggest how to handle this type of words. We implement an unknown word classifier with a swear detection module for developing an automatic English writing scoring system. By experiments with actual test data, we evaluate the accuracy of the unknown word classifier as well as the swear detection module.

Keywords : Automatic Writing Assessment, Swear Word Detection, Unknown Word Classifier

영작문 자동평가를 위한 비속어 검출과 미등록어 분류

이 경 호† · 김 성 권†† · 이 공 주†††

요 약

본 논문에서는 중·고등 수준 단문형 영어 작문시험의 자동채점 시스템을 위한 사전 미등록어 분류기 구현에 대해 다룬다. 영어 자동채점 과정에서 발생하는 사전 미등록어의 유형을 정의하고 각 유형에 대한 검출 방법에 대해 논의하였다. 또한 영작문 답안에서 나타날 수 있는 비속어의 유형을 정의하고 검출 방법에 대해 연구하였다. 영작문 자동평가 시스템의 모듈로서 비속어 검출 기능이 포함된 미등록어 분류기를 구현하였다. 미등록어 분류와 비속어 검출 방법에 대한 성능을 실제 시험 데이터에 적용하여 그 성능을 평가하였다.

키워드 : 자동채점, 비속어 검출, 미등록어 분류기

1. 서 론

영어 쓰기 능력평가 시험에서 인간 채점자에 의한 채점은 시간과 비용이 많이 들고 모든 학생들에게 객관적인 점수판별 기준을 적용하기 어려운 문제가 있다. 이와 같은 채점비용 문제, 채점 객관성 문제를 해결할 수 있는 방법으로 기계에 의한 자동채점기의 필요성이 커지고 있다. 그러나 자동채점 과정을 어렵게 하는 문제 중 하나는 학생들이 답안에 입력하는 단어의 오류와 이에 대한 평가이다.

일반적으로 영어시험 자동채점을 위한 자동채점 프로그램에서는 시험의 목적이나 수준에 맞게 정의된 영어사전을 사용한다. 하지만 학생들의 답안에는 철자 오류가 있는 단어, 비속어, 고유명사, 아무 의미 없이 마구잡이로 입력한 단어 등 자동채점기의 사전에 포함되어 있지 않아 그 의미를 알 수 없는 단어(이하 미등록어)들이 포함되는 경우가 있다. 이 때, 그 시험의 수준과 목적을 고려하여 미등록어의 유형에 따라 그 단어가 채점에 미치는 영향을 다르게 반영할 필요가 있다. 본 논문에서는 중·고등학교 학생들을 대상으로 주어진 상황에 맞는 단문(1~2문장) 영작문 능력을 측정하는 시험*의 자동채점 프로그램에서 생길 수 있는 문제를 해결하기 위한 미등록어 분류기를 연구하였다. 이 분류기의 특징은 다음과 같다.

1) 입력된 미등록어를 7가지 유형으로 분류
2) 중·고등학교 학생들의 어휘수준을 고려
3) 비속어에 대한 강력한 검출
4) Aspell을 이용한 시스템 구성

※ 본 논문은 국가영어능력평가시험(NEAT) 쓰기 3급의 자동채점프로그램을 위해 연구되었음.
† 준 회 원 : 충남대학교 정보통신공학과 박사과정
†† 준 회 원 : 안전행정부 공무원 연수과정
††† 정 회 원 : 충남대학교 정보통신공학과 교수
Manuscript Received : May 14, 2014
First Revision : July 17, 2014
Accepted : July 17, 2014
* Corresponding Author : Kong Joo Lee(kjoolee@cnu.ac.kr)

본 논문의 분류기는 입력으로 들어온 미등록어를 7가지 유형(철자오류, 대소문자오류, 공백오류, 고유명사, 비속어, 노이즈단어, 기타)으로 분류한다. 분류된 미등록어와 그 유형은 자동채점 단계에서 채점요소로 각 유형별로 다르게 반영된다.

본 논문에서 연구하는 분류기가 사용되는 시험은 중·고등학교 학생을 대상으로 하는 시험이다. 그렇기 때문에 미등록어의 유형을 추정하고 복원할 때, 중·고등학교 학생의 어휘 수준을 고려하도록 설계하였다. 또한 시험의 채점 규정 중, 답안에 비속어를 포함한 경우 0점으로 처리하는 규정이 있기 때문에 비속어 검출을 강력하게 수행한다. 그렇기 때문에 본 논문의 미등록어 분류기는 전체 자동채점 과정에서 중요한 역할을 한다. 미등록어 분류기의 기본 동작은 영어권 철자교정기에 많이 사용되고 있는 Aspell 철자교정기의 기능을 이용하여 수행하였다.

본 논문의 구성은 다음과 같다. 이어지는 2절에서 미등록어 분류기와 관련된 관련 연구를 소개한다. 3절에서는 본 논문에서 연구한 미등록어 분류기의 시스템 구성에 대해 설명하고 분류기의 성능을 측정하였다. 마지막으로 4절에서는 본 논문의 결론과 향후 연구 방향에 대해 고찰하였다.

2. 관련 연구

사전 미등록어 검출의 경우 검출 프로그램이 실제 적용되는 응용이나 시험의 종류에 따라서 검출 유형과 범위가 다양하게 달라질 수 있다. 그렇기 때문에 기존의 연구들과 직접적인 검출 성능 비교나 검출 방법의 비교는 적당하지 않을 수 있다. 여기에서는 기존의 수행된 사전 미등록어 검출 방법과 본 논문에서 제시하는 검출 방법 간의 차이를 비교하여 본 논문에서 제시하는 방법의 특징에 대하여 설명한다.

[1]의 연구에서는 미등록어의 품사를 추정하는 연구를 수행하였다. 이 논문에서 기계학습 방식을 이용하여 미등록어의 품사를 추론한다. 미등록어의 품사태깅을 위하여 이 논문의 기계학습기는 미등록어의 양쪽에 있는 단어들의 품사, 미등록어의 양쪽에 있는 단어들의 문법형태, 미등록어의 접두사와 접미사, 관사의 사용여부, 대문자 사용여부나 하이픈 사용여부를 기계학습의 자질(feature)로 사용하여 미등록어의 품사를 분류한다. 본 연구는 기계학습을 이용하여 미등록어의 품사를 결정하는 방식이 아닌 각 미등록어가 가지고 있는 특징을 기반으로 순차적으로 미등록어의 유형을 분류한다는 점에서 이 연구와 차이점이 있다.

사전에 등록되어 있지 않은 미등록어가 실제로 사용하는 단어이지만 사전에 포함되어 있지 않은 단어인지 아니면 오류나 노이즈단어인지를 인식하기 위한 연구가 진행된 바 있다. 이 연구에서는 웹문서를 기반으로 출현빈도를 계산하여 사전에 등록되지 않은 미등록어를 인식하는 방식을 연구하였다[2]. 이 연구에서는 미등록어 처리를 위해 단계별 접근

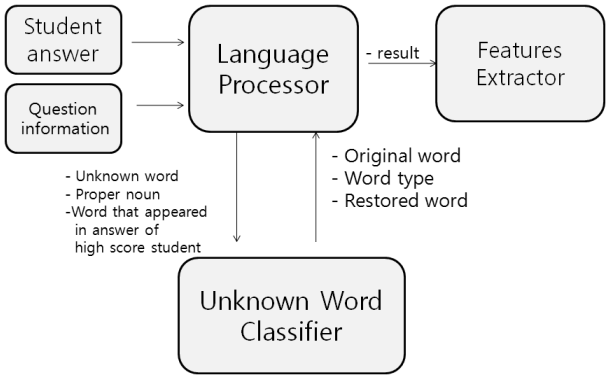


Fig. 1. Operating environment of unknown word classifier

을 취하고 있다. 먼저 전문분석 기반 인식 단계에서는 전문에서 2회 이상 반복되어 나타나는 사전에 없는 단어를 미등록어로 인식한다. 하지만 이렇게 하면 1회만 나타나는 미등록어를 처리할 수 없으며 본 연구에서 주요 연구대상으로 삼는 텍스트는 단문이므로 위의 방법은 단문 시험에서는 적절하지 않다. 그 다음 단계로는 미등록어를 용언(동사)과 명사로 나누어 웹 출현빈도를 구글검색엔진을 이용해 얻고 이것이 일정한 임계값보다 높으면 미등록어로 인식하는 방식을 취하고 있다. 이 논문에서 용언과 명사 검사를 따로 행하는 것은 각각에 붙는 어미(또는 조사)가 보통 다르기 때문이다. 따라서 각각에 어미리스트와 조사리스트를 가지고 어미와 조사를 제거하고 웹 검색을 이용해 출현빈도수를 추출해낸다.

3. 시스템 구성

3.1 미등록어 분류기 동작 환경

본 논문의 분류기는 Fig. 1과 같은 시스템에서 사용된다. 미등록어 분류기는 영어 자동채점 시스템의 일부로 동작한다. 이 자동채점 시스템은 기계학습 알고리즘을 기반으로 하고 있다. 답안 중 일부에 대해 미리 채점을 하고 이를 이용하여 학습 모델을 생성한다. 생성된 학습 모델을 이용하여 다른 답안을 채점하게 된다. 미등록어 분류기는 전체 시스템 중, 언어처리기와 연동되어 사용된다. 언어처리기는 자동채점 시스템에서 학생답안에 대하여 자연언어처리 기술을 이용하여 학생의 영어작문답안을 분석하는 역할을 한다. 그렇기 때문에 언어처리기는 학생들의 답안과 채점하고자 하는 문항에 대한 정보를 담고 있는 문항정보를 입력으로 받는다. 문항정보에는 문항에 대한 기본 정보 외에 문항 출제자에 의해 입력된 답안에서 사용될 수 있는 영어 또는 한국어 로마자표기 고유명사의 목록과 미리 채점된 답안 중 고득점 학생의 답안에서 사용된 단어와 그 단어의 고득점학생 답안에서의 빈도수가 저장되어 있다.

미등록어 분류기는 언어처리기가 답안을 처리하는 과정에서 답안에 나타난 단어가 시스템에서 사용하는 기본사전에 없어 그 의미를 알 수 없을 때 호출되어 사용된다.

이때 언어처리기는 미등록어 분류기에 자신이 알 수 없는 미등록어와 문항정보로부터 입력받은 고유명사목록, 고득점 학생답안의 단어와 그 빈도를 함께 넘겨준다. 분류기는 이 정보들을 이용하여 미등록어의 유형을 결정하고 미등록어, 미등록의 유형, 미등록어의 복원단어를 언어처리기에 돌려준다. 미등록어의 유형이 철자오류(철자오류, 대/소문자오류, 공백오류)로 분류된 미등록어는 단어의 철자오류가 복원된 단어로, 미등록어의 유형이 고유명사인 경우에는 고유명사의 후보군 중 가장 첫 번째에 있는 단어가 복원단어로 결정된다. 그 외의 경우들은 입력된 원본단어가 복원단어이다.

"Kyoung Bok Gung"과 같이 한국어 고유명사의 로마자 표기를 한국어의 음절 단위로 띄어쓰는 경우가 있다. 또한 "facebook, twitter"와 같이 미등록어들의 결합, 또는 단순히 노이즈 단어에 포함된 공백 등 다양한 이유로 연속적으로 미등록어가 나타날 수 있다. 이러한 경우, 연속된 단어에서 공백을 제거한 후(예 : Kyoung Bok Gung → KyoungBokGung) 고유명사 여부를 검사하고, 고유명사 유형이 아닌 경우 공백으로 분리된 각각의 단어에 대해 미등록어분류절차를 수행한다.

3.2 Aspell

본 논문의 분류기는 기존에 유닉스 시스템에서 철자교정기로 많이 사용되고 있는 Aspell[3]을 활용하고 있다. Aspell은 Metaphone 알고리즘[4]을 이용하여 발음을 기반으로 한 철자 교정 작업을 수행한다. Aspell에는 Aspell의 기본 영어사전 외에 사용자가 직접 만든 사전을 Aspell의 기본사전으로 사용할 수 있다. 그렇기 때문에 철자교정을 위해 단어의 수준과 규모가 조정된 영어사전, 영어 비속어 사전, 한국어-로마자 변환 비속어 사전, 등을 Aspell이 참조하는 기본 사전으로 설정하여 Aspell의 기본 기능을 미등록어 분류에 활용할 수 있다. 또한 철자교정후보 단어 제안에 사용되는 기본사전의 단어와 입력단어 사이의 차이(edit distance)의 수준을 설정하여 철자교정의 범위를 조절할 수 있다.

Procedure of Detecting unknown word
1) if (Word is a proper noun) :
return [word, type:proper_noun, representative_proper_noun]
2) else if (Word is a swear_word) :
return [word, type:swear, word]
3) else if (Word has a space error)
return [word type:space_error, correct_word]
4) else if (Word has a spelling error) :
return [word, type:spelling_error, correct_word]
5) ***else if (Word has a case error) :***
return [word, type:case_error, correct_word]
6) else if (Word is similar to a swear_word) :
return [word, type:swear, word]
7) else if (Word contains swear_word) :
return [word, type:swear, word]
8) ***else if (Word is a noise word) :***
return [word, type:noise, word]
9) else :
return [word, type:etc, word]

Fig. 2. Overall process

3.3 미등록어 분류기 동작

언어처리기로부터 미등록어와 기타 정보를 입력받은 미등록어 분류기는 Fig. 2에 나타낸 절차를 통해 미등록어의 유형을 판별한다.

1) 먼저 언어처리기로부터 받은 고유명사 목록에 미등록어가 포함되는지 여부 및 고유명사 목록의 단어와 미등록어의 차이의 크기 등을 고려하여 미등록어가 고유명사인지 여부를 확인한다. 그 결과, 단어가 고유명사로 판명된다면, 미등록어의 유형을 고유명사 유형으로 결정한다. 만약 고유명사가 아니라면, 다음단계로 미등록어의 비속어 여부를 확인한다. 2) 이 단계에서의 비속어 확인 여부는 미리 정의되어 있는 비속어 사전과 비교를 통해 이루어진다. 이 단계에서는 미등록어가 비속어 사전의 단어와 정확하게 일치하는 경우를 비속어로 판별한다. 그렇지 않은 경우 미등록어의 철자오류 여부를 판별하는 단계로 넘어간다. 고등학생 수준의 단어를 사전으로 사용하는 Aspell 철자교정기를 통해 3) 띄어쓰기 오류가 판별되면 공백오류, 4) 철자오류로 판별되면 철자오류, 5) 대소문자 오류로 판별되면 대소문자 오류로 각 미등록어의 유형을 결정하고 결과를 반환한다. 6) 미등록어의 철자오류가 없다고 판별되면 다시 한 번 비속어 여부를 검사한다. 이때는 입력 단어와 비속어 사전을 비교하여 edit distance차이가 N 이하인 단어를 비속어 유형으로 결정한다. 7) 다음으로 입력 단어를 4글자 이상의 substring으로 나누어서 각 substring의 비속어 여부를 판별한다. 이를 통해 미등록어 속에 섞여 있을 수 있는 비속어를 검출 할 수 있다. 본 논문에서 다루는 영작문 자동채점 시스템은 비속어가 쓰인 영작문에 대해 0점 처리라는 강력한 규정이 있기 때문에 2), 6), 7)의 세 단계에 걸쳐 최대한 false positive를 줄일 수 있도록 검출을 하였다. 이러한 비속어 검출 단계에서도 비속어가 아님으로 판단되었다면, 8) 단어의 노이즈 판별단계를 수행한다. 노이즈 단어란 학생 답안에 나타난 아무런 의미가 없는 문자들의 나열을 의미한다. 판별 결과 노이즈가 아니라면, 9) 최종적으로 그 단어의 유형을 기타유형으로 판단한다. 각 단계에서 미등록어 검출 방법은 이어지는 장에서 상술하도록 한다. 각 단계에서 검출되는 미등록어의 유형과 그 예시는 Table 1과 같다.

3.4 고유명사 유형 확인

미등록어 분류기는 언어처리기로부터 문제 출제자에 의해 문제와 관련이 있는 영어, 또는 한국어 단어의 로마자 표기 형태의 고유명사 후보군을 입력받는다. 사전 미등록어 분류기에서는 미리 입력된 고유명사 후보군을 이용하여 미등록어의 고유명사 여부를 판별한다. 입력받은 고유명사목록을 Aspell의 사전으로 사용하여 Aspell의 교정능력 수준에서 학생답안의 미등록어와 고유명사목록의 단어를 비교하여 고유명사 여부를 판별한다.

Table 1. Example of unknown words

Word	Example	Origin / Description
Spelling error	spel	spell
	foregin	foreign
Case error	i	I
	korea	Korea
Space error	waitmy	wait my
	taxidriver	taxi driver
Proper nouns	KyoungBokGung	romanization of proper noun
	Kyoung-Bok-Gung	romanization of proper noun
Swear word	fuckyou	English swear word
	SSibal, ahsibal	romanization of Korean swear word
	rotoRl	romanization of Korean swear word
Noise word	eoifjoiwofwiweiofo	word without meaning
	kkkkk	repetition
etc.	twitter	neologism
	lol, gotta	abbreviations, colloquial
	Namsan	proper noun assumption
	tenosynovitis	jargon

만약 학생 답안에 '경복궁'을 영어로 나타내기 위해 "KyoungBokGung", "GyoungBokGung" 등의 표현이 나올 수 있다. 이러한 표현의 경우 로마자 표기원칙에 따라 표기하는 것이 맞지만, 로마자 표기법을 정확히 알지 못하는 경우 고유명사를 소리 나는 대로 적는 경우가 빈번하다. 이러한 경우 로마자 표기 원칙에는 어긋나지만 전체적인 관점에서 틀린 의미라고 보기 어렵다. 이러한 고유명사가 답안에 나타났을 경우, 이를 미등록어에서 고유명사로 판별하기 위해 문제 출제자가 미리 예상되는 고유명사의 다양한 범주를 입력하고, 이를 이용하여 고유명사 여부를 판별한다. 출제자가 입력한 고유명사 목록 중에 판별해야 하는 미등록어가 포함되는 경우와 입력된 고유명사와 미등록어사이의 edit distance 차이가 N 이하일 경우, 미등록어를 고유명사로 판별한다.

3.5 비속어 검출

본 논문에서는 검출해야 하는 비속어의 유형을 4가지로 분류하였다. 영어 비속어의 경우, 1) 영어 비속어를 그대로 쓰는 경우와 2) 영어 비속어를 한국어 발음나는 대로 키보드의 한글 자판배열에 맞춰 로마자로 작성하는 경우이다. 한국어 비속어의 경우, 3) 한국어 비속어를 소리 나는 대로 로마자로 표기하는 경우와 4) 한글 자판배열에 맞춰 로마자로 작성하는 경우이다. 이러한 비속어의 예는 Table 2와 같다.

Table 2. Type of swear words

Language	Type	Example
English	1. Swear word	fuckyou bitch
	2. Romanizaton of Korean pronunciation (typing in Korean keyboard layout)	Qjrzb qltcnl
Korean	3. Romanization	AhSSibal geseggi
	4. Romanization (typing in Korean keyboard layout)	dkTlqkf rotoRl

비속어 검출을 위하여 영어 비속어 126개와 한국어 비속어 1,990개를 수집하였다. 한국어 비속어의 경우 인터넷 게시판에서 사용되는 금지어 목록을 수집하고 이중 자주 쓰이는 한국어 비속어를 골라내었다. 이렇게 수집된 영어, 한국어 비속어를 각각 2가지 형태로 변형하였다. 한국어 비속어의 경우 한글로 작성된 "한국어 비속어 사전"과, 한글-로마자 변환 규칙에 의해 한국어 비속어를 로마자로 표시한 "한국어 비속어 로마자표기 사전"을 만들었다. 영어 비속어의 경우 영어 비속어를 모은 "영어 비속어 사전"과 영어 비속어에서 중·고등학생 수준에서 많이 사용되는 비속어를 한국어 발음대로 변환한 "영어 비속어 한국어발음표기 사전"을 만들었다. 이와 같은 비속어 분류와 비속어 사전을 이용하여 분류기는 아래와 같이 총 3단계에 걸쳐 비속어 검출을 수행한다.

1) 비속어 사전의 단어와 일치여부 (Fig. 2의 2단계)
2) 비속어 사전의 단어와 edit distance 차이 비교 (Fig. 2의 6단계)
3) 단어 내의 substring으로 포함된 비속어 검출 (Fig. 2의 7단계)

1) 비속어 사전의 단어와 일치여부

첫 번째 비속어 검출단계에서는 미등록어가 비속어 사전의 엔트리와 정확히 일치하는지 여부를 판별한다. Table 2에서 정의한 각 비속어 유형에 대해 본 단계에서의 비속어 판별 방법은 Table 3과 같다. 이와 같은 방법으로 미등록어와 비속어 사전의 단어와의 일치여부를 판별한다.

2) 비속어 사전의 단어와 edit distance 차이 비교

비속어 사전의 엔트리와 정확 일치여부가 아닌 경우 철자검사를 거쳐 다시 비속어 검사 단계를 수행하게 된다. 이때는 비속어 사전을 기본 사전으로 이용하는 Aspell 철자교정기를 사용하여 비속어를 검출한다. 이 단계에서는 영어 비속어(Table 2의 1번 비속어 유형)과 한글 비속어 로마자 표기(Table 2의 3번 비속어 유형)만을 검사한다. 각 유형별 검출방법은 Table 4와 같다. 이 단계를 통해서 변형된 영어 비속어와 한국어 비속어를 로마자로 표기한 경우의 비속어를 검출할 수 있다.

Table 3. Swear word detection process 1

Type	Detection Process
1.	***if*** an unknown word ***A*** in [dictionary of English swear words]
2.	convert an unknown word ***A*** into Korean ***B*** using automata and keyboard layout; ***if*** an unknown word ***B*** in [dictionary of Korean Phonetic representation of English swear words]
3.	***if*** an unknown word ***A*** in [dictionary of romanized Korean swear word]
4.	convert an unknown word ***A*** into Korean ***B*** using automata and keyboard layout; ***if*** an unknown word ***B*** in [dictionary of Korean swear words]
5.	***if*** an unknown word ***A*** begins with slang prefix (A slang prefix is frequently used in vulgar language, but rarely used in common words. eg. “fuck”, “jot”, “gae”)

Table 4. Swear word detection process 2

Type	Detection Process
6.	***if*** Aspell() can generate a correction word for an unknown ***A*** with Aspell’s ultra mode option(within 1-edit distance) and [dictionary of English swear words]
7.	***if*** Asepell() can generate a correction word for an unknown ***A*** with Aspell’s ultra mode option and using [dictionary of romanized Korean swear word]

3) 단어 내의 substring으로 포함된 비속어 검출

미등록어가 서브스트링으로 비속어를 포함하는 경우가 있다. 이런 비속어를 검출하기 위해 미등록어를 4글자 이상(영어 비속어 사전 단어들의 평균 길이와 자주 나오는 비속어의 길이를 고려하여 선정)의 문자개수를 가지는 substring으로 나누고, 각 단어들을 비속어 검출 단계 1) 비속어 사전의 단어와 일치여부 작업을 다시 수행한다. 검출 단계 1만 수행하는 이유는 단순한 노이즈 단어가 substring으로 쪼개지는 과정에서 비속어로 검출되는 위험을 막기 위함이다. 이에 대한 예를 Table 5에 나타내었다.

3.6 철자 오류 검출

기본적인 철자 오류와 띄어쓰기 오류는 Aspell 철자교정기의 기능을 이용하여 처리한다. 채점과정에서 사용되는 사전에서 중・고등학생 수준의 단어 사전과 고득점답안에 나타난 단어 목록을 이용하여 철자 오류와 띄어쓰기 오류를 검출한다. 중・고등학생 수준의 사전을 기본사전으로 사용하는 Aspell에 미등록어를 입력으로 주면, Aspell는 기본 사전 중 미등록어와 일정 edit distance 차이 이하의 단어들을 후보단어로 제시해준다. 이때, 고득점 답안에서 나타난 단어가 후보단어 중에 있다면 이 단어로 미등록어를 복원한다.

Table 5. Swear word detection process 3

Type	Detection Process
8.	***Do*** Swear word Detection Process 1 (Table 3) for each substring of an unknown word ***A*** with 4 letters and more. ex) asdfuckzxcv → 4 letters : asdf, sdfu, **fuck**, uckz ... → 5 letters : asdfu, sdfuc, dfuck ... → ... → word length -1 : asdfuckzxc, sdfuckzxcv

고득점 답안에서 사용된 단어가 다른 단어보다 복원단어로 더 타당할 것이기 때문이다. 만약 후보 단어 중 고득점 답안의 단어들이 2개 이상 있다면 고득점 답안들에서 단어가 나타난 빈도가 높은 단어를 복원단어로 지정한다.

철자 오류는 없으나 대문자로 쓰여야 하는 단어(예 : America, USA)를 소문자로 쓴 경우 Case 오류로 분류한다.

3.7 노이즈 단어 판별

아무런 의미가 없는 문자의 나열을 노이즈 단어로 정의한다. 학생의 영작문에서 노이즈 단어로 판별되는 것은 Table 6와 같은 유형이다.

Table 6. Type of noise words

Type	Description
1.	***if*** an unknown word ***A*** has at least three repetitions of the same substring (ex lollollollollol)
2.	***If A*** has upper letter at the middle or end of a word
3.	***if A’***s length is 1
4.	***if A’***s length is 20 or more
5.	***if A*** has high perplexity

노이즈 단어 유형 5번은 미등록어의 복잡도(Perplexity)[6][7]를 이용하여 노이즈 단어 여부를 판별한다. 20,399개의 단어를 가지고 있는 영어 사전에 나타난 단어들의 알파벳 character 단위의 tri-gram을 학습한 언어 모델을 생성하고, 그 사전에서 나타난 단어의 복잡도 중 가장 높은 복잡도보다 미등록어의 복잡도가 높으면 노이즈 단어로 판별하도록 하였다[5].

3.8 기타 유형 판별

위의 단계를 모두 거쳤을 때 어떠한 오류 타입도 부여되지 않은 미등록어는 기타유형으로 판별하게 된다. 기타 유형을 노이즈 단어와 따로 구분하는 이유는 두 유형이 영작문 자동채점의 관점에서는 서로 다른 가중치를 갖기 때문이다.

4. 실 험

중·고등학생들을 대상으로 한 단문형 영어작문 시험의 데이터를 이용하여 분류기의 성능을 평가하기 위한 실험을 수행하였다. 총 1,209개의 학생 답안에서 발생한 1,212개의 미등록어에 대해, 본 논문의 분류기의 분류 결과와 사람의 분류 결과를 비교하는 실험을 수행했다.

4.1 분류 및 복원 정확도 실험

Table 7, 8에서 전체 개수는 전체 미등록어에 대해 분류기를 통해 분류된 각 유형별 미등록어의 개수를 의미한다. 분류오류개수는 해당 유형으로 판별된 미등록어의 유형분류가 잘못된 단어의 개수이다. 분류 정확도는 {1-(분류오류개수)/(전체개수)}×100를 나타낸다. 복원오류개수는 해당 미등록어의 분류는 맞지만, 단어를 복원했을 때, 복원된 단어가 잘못된 경우를 나타낸다. 복원 정확도는 {1-(분류오류개수+복원오류개수)/(전체개수)}×100로 계산된다. 복원오류개수는 원래 단어를 추론해야 하는 단순철자오류와 띄어쓰기 오류, 대소문자오류에 대해서 복원오류 개수와 복원 정확도를 계산하였다.

실험결과 미등록어로 분류된 1,212개의 단어에 대해 약 92.4%의 분류 정확도와 89.4%의 복원 정확도를 보였다. 고유명사나, 노이즈, 대/소문자 오류 같은 경우에는 분류 정확도가 높았지만, 기타유형 판별에 있어서는 낮은 정확도를 보였다.

Table 7. Accuracy of unknown word classifier

Word	Total number	Number of misclassified words	Accuracy
Proper noun	39	0	100%
Space error	71	6	91.6%
Spelling errors	668	53	92.1%
Case error	307	0	100%
Swear	1	0	100%
Noise word	96	10	89.6%
etc.	30	23	23.3%
Total	1212	92	92.4%

Table 8. Accuracy of spelling correction

Word	Total number	Number of incorrect spelling correction	Accuracy
Spelling errors	668	37	86.5%
Space error	71	0	91.6%
Case error	307	0	100%
Total	1212	37	89.4%

실험 결과를 살펴보면, 철자오류에서 분류오류나 복원오류로 잘못 분류되는 경우들에는 “ah-ha”를 “ahead”로, “Heiioa”를 “Hero”로 복원하는 것 같이 2 edit distance를 갖는 단어를 잘못 복원하는 경우들이 보였다. 그리고 ‘Iam’이나 ‘min’ 같이 특정하게 많이 잘못 쓰이는 표현이나 축약어 문제를 해결하기 위해서는 철자교정 단계의 Aspell에서 사용하는 사전을 보강해야 할 것으로 보인다. 또한 ‘adoctor’를 ‘doctor’로 잘못 복원하는 경우와 같이, 관사 사용 빈도수는 철자오류 검출의 입력으로 사용하지 않아서 해당 단어를 띄어쓰기 오류가 아닌 철자오류로 분류하는 경우가 있었다. 기타유형으로 판별된 단어로는 “Namsan”, “Dondaemoon”과 같이 한국어 고유명사의 로마자 표기이지만 출제자에 의해 입력받은 고유명사 목록에 없는 단어가 포함된다. “pinkscaf”(“pink scarf”의 오류)같이 올바르게 쓰인 단어와 철자오류를 포함한 단어가 공백 오류로 결합된 경우 기타 유형으로 분류되었다.

4.2 분류 및 복원 정확도 실험

본 논문에서 구현한 미등록어 분류기의 주요 특징은 정확한 비속어 검출에 있다. 비속어는 영작문 점수에 큰 영향을 줄 수 있기 때문에 신중하고 정확히 검출되어야 한다. 분류 및 복원정확도 실험 결과에서 비속어가 아닌 단어를 비속어로 분류하는 경우(false positive)는 나타나지 않았다. 비속어 검출 실험에서는 분류기의 한국어 및 영어 비속어 검출 능력을 실험하였다.

실험을 위해 영어권 사용자가 사용할 수 있는 비속어와 한국어 비속어 및 그 비속어들의 변형 형태를 각각 110개, 154개 수집하였다. 한국어 비속어의 경우 수집한 한국어 비속어를 한글-로마자변환규칙에 따라 로마자로 변환하여 비속어 검출을 수행하였다. 실험결과는 Table 9과 같다.

Table 8에서 “정확검출”은 미등록어가 영어, 한국 비속어 사전의 엔트리와 정확하게 일치하여 비속어로 검출된 경우의 개수를 의미한다. 비속어 사전에 “fucks”가 엔트리로 있을 때, 미등록어 입력으로 들어온 “fucks”가 비속어로 검출된 경우 정확검출이다. “변형검출”은 비속어 사전의 단어와 미등록어의 관계가 일정 edit distance차이 이하로 검출되는 비속어의 개수이다. 비속어 사전의 엔트리로 “bullshit”

Table 9. Accuracy of swear word detection

Word	English swear	Korean swear
Total number of words	110	154
Number of detection	85	127
Exact detection	2	9
Deformation detection	83	118
Accuracy	77.2	82.5

과 “shit”이라는 단어가 있을 때, “bulllshits”, “sssshit”과 같이 사전의 단어에서 약간의 변형된 형태를 비속어로 인식하는 경우이다.

실험결과에서 작성자가 의도적으로 비속어 “cocks”이나 “sucks”을 입력하였지만, 판별과정에서 “clocks”과 “ducks”의 철자오류로 비속어를 분류하는 경우가 있었다. 실제 시험 상황에서 이러한 단어가 나왔을 때 작성자의 의도를 파악하는 것은 어려운 일이다. 이러한 문제는 미등록어의 문맥정보 등을 활용하여 해결할 수 있을 것이다.

한국어 비속어 검출 실험에서도 정확검출은 미등록어 “inyeona”가 한국어 비속어 사전에 엔트리로 존재하는 경우이다. 또한 한국어 비속어의 경우에도 정확 검출보단 변형검출의 결과가 더 많았는데, 실험에 사용한 한국어 비속어들이 비속어 사전의 단어와 같은 경우보다 약간씩의 변형을 가지고 있는 경우가 많아서이다. 실제로 학생들이 작성한 비속어를 보면 “sipallom” 같은 단어를 “ssipallom”, “ssiipalom”과 같이 같은 글자를 반복하여 쓰는 식의 비속어 변형이 많았다. 또한 사전에 비슷한 단어가 없지만 비속어에 주로 쓰이는 글자인 “gae”의 포함여부로 인해 “gaejjos”같은 단어를 비속어로 검출할 수 있었다.

비속어의 경우 학생들이 생각하여 답안에 쓸 수 있는 단어들의 종류가 제한적이다. 하지만 답안을 작성할 때 나타날 수 있는 변형은 다양하다. 실험결과를 보면 어느 정도 단어의 변형에 대해 대응이 가능함을 보였다. 이전에 비속어검출에 관한 공개된 실험결과를 논문의 저자가 아는 수준에서는 발견하지 못하였으므로 다른 실험의 방법과 결과에 대한 엄밀한 비교가 어렵다. 하지만 임의로 수집된 비속어에 대한 분류 실험에서 사전의 질에 전적으로 의지하는 정확검출보다 단어의 변화에 대처하는 변형검출의 결과가 더 좋은 것을 보아, 분류기가 적용되는 시점에서 응시생들의 수준을 고려하여 적절한 비속어 사전을 구축할 수 있다면 실제 시험에서의 비속어 검출에 충분히 적용할 수 있을 것이다.

5. 결 론

본 논문에서는 중・고등 수준 단문형 영어 작문시험의 자동채점 시스템에서 사용하는 미등록어 유형 분류기에 대하여 연구하였다. 이 분류기를 통해 자동채점기가 확인할 수 없는 단어의 유형에 대한 정보를 제공할 수 있게 된다. 이를 위해 7가지의 미등록어 유형을 정의하고 각 유형에 대한 검출 방법을 소개하였다. 그 결과 미등록어의 유형에 대한 92.4%의 분류 정확도와 89.4%의 복원 정확도를 보였다.

실험 결과 중 기타유형 대한 분류 정확도가 다른 항목에 비해 낮았는데, 이는 사전미등록어를 구분하는 별도의 방법 없이 앞서 분류되지 않은 것을 기타유형으로 분류하는 데서 기인한다고 볼 수 있다. 기타유형 미등록어의 보다 정확한 판별은 향후 연구 과제로 남아 있다. 또한 각 오류 유형의 검출 단계에 맞는 적절한 유사어 검출 방법을 적용하여 검출 성능을 향상 시킬 수 있을 것으로 생각한다.

본 논문에서 소개한 미등록어 자동 분류기는 향후 학생 영어 시험뿐만 아니라 사전을 재정의 함으로써 성인이나 기타 다양한 목적을 가진 영어 능력평가 자동채점 시스템이나 SNS, 대화 시스템 등의 다양한 응용의 미등록어 자동 분류기로 확장할 수 있다. 이를 위하여 적절한 사전 정의 방식과 데이터 수집, 분류 기법의 지속적인 연구를 필요로 한다. 본 논문의 미등록어 분류기는 이러한 발전을 위한 시스템 구현과 실제 적용 사례로 그 의의가 있다.

References

[1] Tetsuji Nakagawa, Taku Kudoh, and Yuji Matsumoto, “Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines”, in Proc. of the 6th NLPRS, pp.325-331, 2010.

[2] Park So-Young, “Phase-based Model Web Documents for Korean Unknown Word Recognition”, in *Journal of the Korea Institute of Information and Communication Engineering,* pp.1898-1904, 2009.

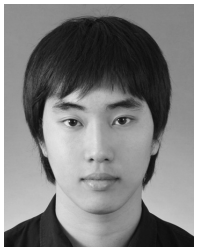
[3] Atkinson, Kevin, “Gnu aspell 0.60. 4”, 2006.

[4] Philips, Lawrence, “Hanging on the metaphone,” Computer Language 7.12, Dec., 1990.

[5] Kukich, Karen, “Techniques for automatically correcting words in text,” ACM Computing Surveys (CSUR) Vol.24, No.4, pp.377-439, 1992.

[6] Bahl, L., Baker, J., Jelinek, E., and Mercer, R., “Perplexity-a measure of the difficulty of speech recognition tasks,” In Program, 94th Meeting of the Acoustical Society of America 62:$63, Suppl. No.1, 1997.

[7] Jia, Zhongye, Peilu Wang, and Hai Zhao, “Grammatical Error Correction as Multiclass Classification with Single Model,” CoNLL-2013 p.74, 2013.



이 경 호

e-mail : lee6boy@empal.com

2011년 충남대학교 정보통신공학과(학사)

2013년 충남대학교 정보통신공학과(석사)

2013년~현 재 충남대학교 정보통신공학과 박사과정

관심분야 : 자연언어처리, 기계학습, 인공지능

김 성 권

e-mail : mooshu@naver.com

2014년 충남대학교 정보통신공학과(학사)

2014년～현 재 안전행정부 공무원 연수과정

관심분야 : 기계학습, 인공지능

이 공 주

e-mail : kjoolee@cnu.ac.kr

1992년 서강대학교 전자계산학과(학사)

1994년 한국과학기술원 전산학과 (공학석사)

1998년 한국과학기술원 전산학과 (공학박사)

1998년～2003년 한국마이크로소프트(유) 연구원

2003년 이화여자대학교 컴퓨터학과 대우전임강사

2004년 경인여자대학 전산정보과 전임강사

2005년～현 재 충남대학교 정보통신공학과 교수

관심분야 : 자연언어처리, 기계번역, 정보검색, 정보추출