한글 편집거리 알고리즘을 이용한 한국어 철자오류 교정방법

(A Method for Spelling Error Correction in Korean Using a Hangul Edit Distance Algorithm)

박승현*, 이은지**, 김판구***

(Seung Hyeon Bak, Eun Ji Lee, Pan Koo Kim)

요 약

컴퓨터가 상용화되면서 일반인들은 문서를 작성하기 위해 컴퓨터를 이용하는 방법을 자주 사용하게 되었다. 컴퓨터를 이용하여 문서를 작성하는 방법은 작성 속도가 빠르고 손의 피로가 적지만 철자오류가 발생할 확률이 매우 높다. 보통 철자오류는 발견하기 쉽기 때문에 곧바로 수정이 가능하지만, 사용자의 지식 부족 혹은 눈에 잘 띄지 않는 철자오류도 존재하기 때문에 철자오류가 존재하지 않는 문서를 작성하기 어렵다. 온라인상에서는 문서 작성에 대한 규칙 및 예절이 미비하기 때문에 철자오류에 의한 문제가 적지만 중요문서에서 발생하는 철자오류는 신뢰도 하락과 같은 큰 문제를 일으킨다. 철자오류 교정은 전문가 또한 완벽하게 수행하기 힘들기 때문에 비전문가인 일반인들을 위한 교정방법연구가 필요하다. 본 논문에서는 한글 편집거리 알고리즘을 이용해 철자오류를 교정하는 연구를 진행한다. 이전 연구를 통해 검출한 철자오류를 수집한 말뭉치 사전에서 등장하는 단어 중 철자오류 단어와 가장 유사한 단어를 발견하여 주위 단어와의 동시등장빈도를 계산하는 것으로 철자오류 교정을 수행하게 된다.

■ 중심어 : 철자오류; 한글 편집거리; 철자오류 교정; 자연어 처리;

Abstract

Long time has passed since computers which used to be a means of research were commercialized and available for the general public. People used writing instruments to write before computer was commercialized. However, today a growing number of them are using computers to write instead. Computerized word processing helps write faster and reduces fatigue of hands than writing instruments, making it better fit to making long texts. However, word processing programs are more likely to cause spelling errors by the mistake of users. Spelling errors distort the shape of words, making it easy for the writer to find and correct directly, but those caused due to users' lack of knowledge or those hard to find may make it almost impossible to produce a document free of spelling errors. However, spelling errors in important documents such as theses or business proposals may lead to falling reliability. Consequently, it is necessary to conduct research on high-level spelling error correction programs for the general public. This study was designed to produce a system to correct sentence-level spelling errors to normal words with Korean alphabet similarity algorithm. On the basis of findings reported in related literatures that corrected words are significantly similar to misspelled words in form, spelling errors were extracted from a corpus. Extracted corrected words were replaced with misspelled ones to correct spelling errors with spelling error detection algorithm.

■ keywords : Spelling Error; Hangul edit distance; Spelling error correction; Natural Language Processing;

Ⅰ. 서 론

본래 연구의 목적으로 사용되었던 컴퓨터가 일반인들도 사용할 수 있도록 상용화된 이후, 사람들은 문서 작성을 위하여 필기도구를 사용하는 방법뿐만이 아니라 컴퓨터를 이용하는 방법 또한 자주 사용하게 되었다. 컴퓨터를 이용하여 문서를 작성하는 방식은 작성 속도가 빠르며 손의 피로가 적지만 필기도구를 이용하는 방식에 비해 철자오류가 발생할 확률이 높다. 보통 철자오류가 발생할 경우 문서를 작성하는 사용자에게 쉽게 발견

* 학생회원, 조선대학교 소프트웨어융합공학과 ** 학생회원, 조선대학교 컴퓨터공학과 *** 정회원, 조선대학교 컴퓨터공학과

이 논문은 2014년 교육부와 한국연구재단의 지역혁신창의인력양성사업의 지원(NRF-2014H1C1A1073115)과 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원(No. NRF-2016R1A2B4012638)을 받아 수행된 연구임.

접수일자 : 2017년 03월 09일 게재확정일 : 2017년 03월 30일

수정일자 : 2017년 03월 22일

교신저자 : 김판구 e-mail : pkkim@chosun.ac.kr

되기 때문에 곧바로 수정이 가능하다. 하지만 사용자의 지식 부족으로 인해 발생한 철자오류 혹은 눈에 잘 띄지 않는 부류의 철자오류도 존재하기 때문에 문서를 작성하는 사용자의 입장에서는 철자오류가 존재하지 않는 문서를 작성하기가 매우 어렵다. 인터넷 커뮤니티 시스템에 의해 문서 작성 비율이 높은 온라인상에서는 문서 작성에 대한 규칙 및 예절이 미약하기 때문에 철자오류에 의해 심각한 문제가 발생하지 않는다. 하지만 논문과 같은 중요문서에서 발생하는 철자오류는 논문에 대한 신뢰도 하락과 같은 문제를 불러일으킨다. 그로 인해 작성이 완료된 중요문서의 경우 반복적으로 철자오류 교정 작업을 수행하지만, 교정 지식이 풍부한 교열 전문가 또한 문서 내에 존재하는 철자오류를 완벽하게 교정하기 힘들기 때문에 적게나마 철자오류가 잔존하게 된다. 철자오류의 높은 교정 난이도 때문에 맞춤법 비전문가인 일반인들이 철자오류 교정에 대해 큰 어려움을 겪고 있기 때문에 일반인들을 위한 교정방법 시스템에 대한 연구가 필요하다. 본 논문에서는 문장에서 발견한 철자오류를 한글 편집거리 알고리즘을 이용하여 교정하는 연구를 수행하게 된다. 철자오류가 발생한 단어는 본래 입력할 단어의 형태와 상당히 유사하다는 연구를 바탕으로 하여 한글 편집거리 알고리즘을 이용해 철자오류를 교정한다. 본 논문은 2장에서 철자오류 복구와 한글 편집거리 알고리즘에 대한 관련 연구를 소개하고, 3장에서는 철자오류 복구에 대한 연구 방법을 기술한다. 4장에서 연구 결과에 대해 소개하며, 5장에서는 결론 및 향후 연구에 대해 서술하며 본 논문을 마무리한다.

Ⅱ. 관 련 연 구

철자오류는 단어의 철자가 잘못된 철자로 입력되거나 삭제 혹은 철자가 추가적으로 삽입되어 의미 없는 단어가 되거나 문맥에 맞지 않는 단어로 변한 것을 의미한다[1]. 철자오류는 단순 철자오류와 문맥의존 철자오류로 구분할 수 있다. 단순 철자오류는 어휘사전에 존재하지 않는 유형의 철자오류를 의미하며 단순히 형태소 분석을 통해 간단히 오류의 검출과 교정이 가능한 철자오류이다. 문맥의존 철자오류는 오류가 발생한 단어가 어휘사전에 존재하는 단어이지만 문맥에 맞지 않는 유형의 철자오류를 의미하며 오류 검출 및 교정의 난이도가 높아 문맥의존 철자오류에 대한 연구가 다수 진행되고 있다[2]. 문맥의존 철자오류를 교정하는 방법은 규칙 기반 방식과 통계 기반 방식으로 나뉜다. 규칙 기반 방식은 주로 사람이 제작한 규칙을 통해 철자오류를 교정하게 된다. 규칙을 이용한 교정 방식은 추가된 규칙이 많을수록 성능은 향상되지만 고도의 지식을 갖춘 전문가가 필요하며 규칙의 유지 및 보수에 대해 막대한 비용이 필요하다[3]. 통계 기반 방식은 통계 모형을 이용하여 철자오류를 교정하게 된다. 통계적 방식을 이용하는 대표적인 방법으로는 n-gram 언어 모델을 이용한 방법, 교정 어휘 쌍을 이용하는 방법 등이 있다. n-gram 언어 모델을 사용하는 방법은 대용량 말뭉치에서 어절 3-gram을 구하고, 이를 바탕으로 각 문장 또는 부분 문장 확률을 계산하여 철자오류 교정이 이용하는 방법이다. 한국에서는 조사 및 어미의 문제로 인하여 형태소 n-gram을 주로 사용하였으나 한 연구에서는 단어와 조사 간 결합 정보를 그대로 활용하고자 형태소 n-gram이 아닌 어절 n-gram 모델을 사용하여 철자오류 교정을 수행하였다[4]. 영어권 언어에서는 'Web 1T 3-Grams'라는 대용량 말뭉치 사전이 발표되면서 3-gram을 이용한 철자오류 교정 방법들이 연구되어졌다[5, 6]. 교정 어휘 쌍을 이용한 방법은 어의 중의성 해결과 같은 방법론을 이용하게 된다. 교정 어휘 쌍을 이용한 방법은 교정 어휘 쌍에 해당하는 단어가 중의적이라 보며, 통계적 방법으로 중의성을 해결한 이후 결과가 원래 단어와 같으면 철자가 바르다 판단하고 다르면 철자오류로 판단하여 철자교정을 수행하게 된다[7]. 노이지 채널 모델에 기반을 둔 철자오류 교정 방법은 기계 번역, 광학 문자 인식, 음성 인식등과 같은 입력 데이터가 노이지 채널에 의해 데이터의 변형이 일어나 출력 데이터에 오랴가 발생하였다고 가정하여 출력으로부터 입력을 이끌어 내는 디코딩 문제로 간주하여 해결한다[8]. 대부분의 단어와 단어의 편집거리를 구하는 알고리즘은 영어를 기준으로 하여 연구되었지만 한글을 기준으로 하여 편집거리를 구하는 일부 연구가 존재한다[9]. [9]에서는 한글의 편집거리를 음절 단위 기준으로 추출하는 SylED알고리즘과 편집거리를 음소 단위 기준으로 추출하는 PhoED알고리즘을 제안하였다. 그리고 두 알고리즘을 복합적으로 사용하여 두 단어의 음소 단위가 다르다면 α를, 음절 단위가 다르다면 β를 더하는 형태로 문자열의 길이가 다름에도 불구하고 편집거리를 구할 수 있는 KorED알고리즘 또한 제안하였다. KorED알고리즘을 제안 이후 보다 시간이 지나면서 미비한 점이나 성능 향상의 목적으로 추가적인 연구가 진행되었다. 그 결과로 단순히 생각하면 'ㄱ'과 'ㄲ'의 편집거리는 'ㄱ'과 'ㅂ'의 편집거기보다 작다고 판단이 되기 때문에 음소의 종류에 따라 서로 연관성이 있는 음소들의 경우 KorED알고리즘에서 제시한 음소의 거리 단위보다 작은 거리를 가지고 있다고 판단하여 새로운 GrpSIM알고리즘을 제안한 논문을 작성하였다[10].

Ⅲ. 본 론

1. 시스템의 구성도

본 논문에서는 문서에서 철자오류 단어를 검출하고, 검출된 철자오류 단어를 올바른 단어로 교정하는 연구를 수행한다. 그림 1은 제안하는 철자오류 교정 알고리즘의 전체적인 흐름도를

나타내고 있다. 통계적인 방식을 통해 철자오류를 교정하기 위해서는 철자오류가 포함되어 있지 않은 문서를 수집하여 말뭉치 사전을 구축할 필요가 존재한다. 그렇기 때문에 본격적인 연구를 수행하기에 앞서 말뭉치 사전을 구축하게 된다. 말뭉치 사전 구축 이후 철자오류를 교정하기 위해 일차적으로 문장 내에서 철자오류 존재 유무를 검사하며, 철자오류 단어가 존재할 시 철자오류 단어를 추출하게 된다. 본 연구에서는 철자오류 교정을 위해 철자오류 단어와 편집거리가 작은 단어들을 말뭉치 사전에서 추출하여 교정 단어 리스트를 제작하여 교정을 수행하게 된다. 구축한 교정 단어 리스트의 단어들을 입력 받은 문장의 철자오류 단어와 치환한 다음 철자오류 교정 알고리즘을 수행함으로서 치환한 교정 단어의 알맞음 여부를 검사하여, 가장 수치가 높은 단어를 문장에 삽입하는 것으로 철자오류 교정을 끝마치게 된다.

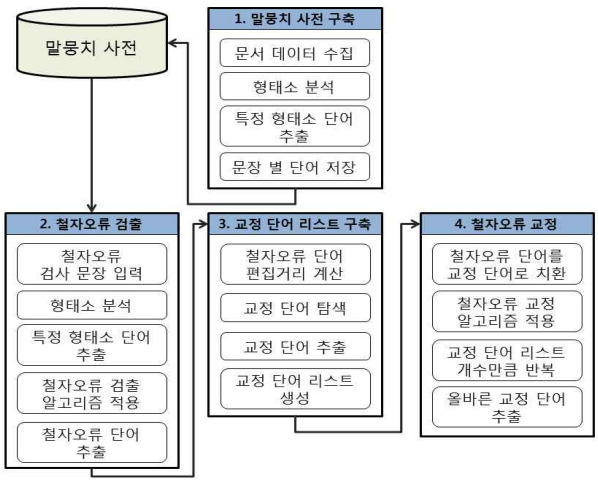

그림 1. 제안된 시스템의 구성도

2. 말뭉치 사전 구축

본 논문에서 제안하는 철자오류 교정 알고리즘은 통계적 기반의 철자오류 교정 방식이다. 통계적 기반의 철자오류 교정 방식은 단어들의 동시 등장 확률을 구하기 위해 말뭉치 사전을 구축할 필요가 있다. 통계적인 방식의 철자오류 교정 모델은 철자오류 교정을 받고 있는 단어가 좌우 문맥에서 등장하는 단어들과의 동시 등장 확률을 계산함으로서 수행되는 형태의 교정 모델을 의미한다[24]. 그렇기 때문에 통계적인 방식의 철자오류 교정 모델을 사용하기 위해서는 단어의 동시 등장 확률을 구하기 위해 말뭉치 사전의 구축이 필수불가결한 문제이다. 말뭉치 사전 구축하기 위해 수집한 문서에서 철자오류가 포함되어 있다면, 특정 단어의 철자오류 검사 시 철자오류 단어와 동시에 등장한다고 계산이 될 수 있기 때문에 말뭉치 사전을 구축할 때에는 되도록 철자오류가 포함되어 있지 않는 문서들 위주로 수집해야 한다. 본 논문은 뉴스 기사들 중 네이버 뉴스에서 2016년 4월 총선의 기사들을 수집하여 말뭉치 사전을 구축하게 되었다. 이 때 수집한 뉴스 기사의 분량은 약 7000여개의 문장에 달하며, 또한 철자오류 검출과 교정 시 문장별로 동시 등장 빈도를 구하여 알고리즘에 적용하기 때문에 문장별로 분리하여 말뭉치 사전을 구축하였다.

3. 철자오류 단어 검출

본 절에서는 코사인 유사도를 이용해 철자오류 검출을 수행하는 과정에 대해 소개한다. 철자오류를 검출하기 위해서는 우선 문장에서 등장하는 모든 단어들을 추출하기 위해 형태소 분석을 할 필요가 있다. 철자오류 검출에 수행되는 품사는 동사, 형용사, 명사 같은 품사 단어들로, 이 단어들은 문장의 의미를 결정짓는데 높은 역할을 맡고 있으므로 중점적으로 철자오류 검사를 수행하게 된다. 부사, 접속사와 같은 품사 단어들은 주로 통계적인 방식이 아닌 규칙 기반 방식을 통해 철자오류 교정이 수행되며, 모든 품사의 단어들을 통계적인 방식을 통해 교정을 수행하게 되면 시스템의 작동 시간이 기하급수적으로 늘어나는 단점이 존재하기 때문에 문장에 등장하는 모든 단어를 통계적 방식으로 철자오류를 검출하는 것은 현실성이 떨어진다. 우선 코사인 유사도를 이용하여 철자오류를 검출하기 위해서는 문장에 등장하는 단어들의 동시 등장 빈도를 구할 필요가 있다. 표 1에서 볼 수 있는 예문은 문장에서 등장하는 단어의 동시 등장 빈도에 대해 설명하기 위해 작성한 문장이다.

표 1. 동시 등장 빈도 예시 문장

친구는 게임을 좋아한다.

표 1의 예문에 등장하는 단어의 동시등장 빈도는 다음과 같이 계산된다. '친구'라는 단어 t_1은 '게임'이라는 단어 t_2와 '좋다'라는 단어 t_3와 한 문장 내에 동시에 등장하고 있다. 그렇기 때문에 t_1과 t_2가 동시에 발생하는 빈도 f_{12}와 단어 t_1과 단어 t_3가 동시에 등장하는 빈도 f_{13}은 각각 '1'이 된다는 것을 알 수 있다. 단어의 동시등장 빈도를 나타내는 f_{ij}는 절대빈도라고 하며, 이전 연구들에서는 절대빈도를 이용하여 철자오류를 검출하였다. 그렇기 때문에 이전 연구들에서는 철자오류 검출 방법은 등장 경향이 유사하더라도 절대빈도의 영향을 많이 받기 때문에 다른 결과가 출력되었지만, 코사인 유사도를 이용하여 철자오류를 검출할 경우, 절대빈도의 영향이 받지 않기 때문에 코사인 유사도를 이용하여 철자오류를 검출하게 된다. 수식 (1)은 두 단어의 동시 출현 확률을 구하기 위해 사용한 코사인 유사도

$$SIM(t_i,t_j) = \frac{\sum_{k=1(k\neq i,j)}^{n} f_{ik} \times f_{jk}}{\sqrt{\sum_{k=1(k\neq i,j)}^{n} f_{ik}^2} \times \sqrt{\sum_{k=1(k\neq i,j)}^{n} f_{jk}^2}} \quad (1)$$

수식을 보여주고 있다.

식 (1)은 문장에서 등장하는 i번째 단어 t_i와 j번째 단어 t_j를 쌍으로 묶어 두 단어의 동시 출현 확률을 구하는 수식을 보여주고 있다. f_{ik}는 단어 t_i와 단어 t_k의 동시 출현 빈도를 나타내며, k는 i나 j가 되지 않기 때문에 두 단어 t_i와 t_j의 동시 등장 빈도, 즉 절대빈도를 이용하지 않는다. 코사인 유사도는 0부터 1까지의 값을 가질 수 있으며, 유사도가 0에 가까울수록 두 객체는 서로 독립적 이라는 의미가 되고, 1에 가까울수록 유사하다고 할 수 있다. 그렇기 때문에 수식 (1)을 통해서 나온 두 단어의 동시 등장 확률이 0에 가까울수록 수식 (1)에 적용한 두 단어 중 한 단어가 철자오류가 발생한 단어라고 할 수 있다. 하지만 단순히 수식 (1)만 가지고는 수식에 적용한 두 단어 중 어느 단어가 철자오류가 발생한 단어인지 알 수 없다. 그렇기 때문에 특정 문장에서 등장하는 단어 t_1이 철자오류 단어인지 아니면 정상 단어인지 여부를 가리기 위해 단어 t_1을 제외한 문장에 등장하는 다른 단어들인 t_2부터 문장에 등장하는 마지막 단어 t_n까지의 코사인 유사도를 계산한 다음 평균값을 추출하여 그 값이 가장 낮으며 통시에 임계값 이하인 단어를 철자오류로 검출하게 된다.

4. 교정단어 리스트 구축 및 철자오류 교정

철자오류 검출이 성공적으로 수행이 완료 되었다면, 검출된 철자오류 단어를 본래 입력 단어로 교정해야할 필요가 존재한다. 본래 철자오류를 교정하는 기존 연구에서는 철자오류가 발생한 단어의 자소를 다른 자소로 치환한 이후 다시 좌우 문맥에 등장하는 단어들과의 동시 빈도를 구함으로서 철자오류 교정을 수행하게 된다. 하지만 이전 연구에서 수행된 교정 단어 탐색 방식은 알고리즘 수행 시간을 크게 증가된다는 단점이 존재한다. 6개의 자소로 이루어진 철자오류 단어를 교정하기 위해서는 원래 입력되어있는 자소를 대치가 가능한 다른 자소로 치환한 다음 좌우 문맥에 등장하는 단어들과의 동시 등장 확률을 계산한다고 생각할 경우 처리 시간이 길다는 것은 금방 알 수 있게 한다. 하지만 단순히 생각하면 사전에 존재하는 단어들 중 철자오류가 발생한 단어와 편집거리가 작은 단어들만 추출하여 좌우문맥 단어들과 동시 등장 확률을 구하게 되면 철자오류 교정이 걸리는 처리시간이 줄어들겠다는 것을 알 수 있다. 그렇기 때문에 본 논문에서는 한글 편집거리 알고리즘을 사용하여 철자오류 단어와 편집거리가 작은 단어들을 말뭉치 사전에서 추출함으로서 철자오류 교정을 수행하게 된다. 철자오류를 검출할 때 식 (1)을 사용하여 가장 낮은 평균값을 기록한 단어를 철자오류 단어라 판단하고 검출하였다면 철자오류 교정은 리스트를 구축한 교정 단어들 중 식 (1)을 사용하여 가장 높은 값을 기록한 단어를 올바른 교정 단어라 판단하고 교정을 수행하게 된다.

Ⅳ. 실험 및 결과

이전 장에서 소개했다시피 철자오류를 교정하기 위해서는 일단 문장에 존재하는 철자오류를 검출할 필요가 있다.

표 2. 철자오류 검출 예시 문장 및 추출단어

예시문장	안철수 국민의당 사임 공동대표가 31일 0시 413 총선 첫 유세에 나서 승리를 다짐했다.
추출단어	안철수 국민의당 사임 공동대표 총선 유세 나서다 승리 다짐

표 2는 철자오류 검출을 위해 수집한 문서에서 등장하는 하나의 문장과 그 문장에서 추출한 품사들을 보여주고 있다. 추출한 단어들 중 '사임'이라는 단어는 본래 '상임'이라는 단어를 입력하다가 첫 번째 음절의 받침 'ㅇ'을 입력하지 못함으로써 철자오류가 발생하였다. 예시문장에 등장하는 단어들의 코사인 유사도 평균을 구하게 되면 최소 '0.66'이 나오게 되지만, '사임'이라는 단어는 다른 단어들과 달리 '0.34'라는 낮은 코사인 유사도 평균값을 기록한다. 단순히 문장 내에 등장하는 단어들 중 낮은 코사인 유사도 평균값을 기록한 단어를 철자오류라고 판단하여 검출할 수 있지만 그리할 경우 단순히 문장 내에 등장하는 단어들 중 가장 낮은 평균값을 기록할 뿐이지 값이 그리 낮지 않은 정상 단어를 철자오류 단어라고 검출할 수 있다. 그렇기 때문에 철자오류를 검출한 코사인 유사도 평균값에 임계값을 두어 평균값이 가장 낮다고 하더라도 임계값 이상인 단어들은 정상단어라고 판단할 필요가 있기 때문에 Precision, Recall, 그리고 F-Measure을 통해 가장 철자오류 검출율이 높은 임계값을 탐색하게 되었다.

표 3. 코사인 유사도 상세 통계

	·····	0.43	**0.44**	0.45	·····
Precision	·····	94.94	**94.05**	91.34	·····
Recall	·····	94	**95**	95	·····
F-Measure	·····	94.47	**94.52**	93.13	·····

임계값을 '0.1'단위로 검색했을 때 임계값을 '0.4'로 두었을 경우 Precision, Recall, F-Measure값이 각각 '93.81', '91', '92.38'로 가장 높게 나왔으며, '0.5'는 '86.6', '97', '91.5'로 차등을 하였다. 그로인해 가장 성능이 높은 임계값은 '0.4'와 '0.5'사이에 존재할거라 예상하였고, 다시 '0.4'와 '0.5'사이의 값을 '0.01'로 검색하였을 때 '0.44'가 가장 높은 성능을 보여주고 있었다. 그렇기 때문에 본 논문에서는 철자오류 검출의 임계값을 '0.44'로 선택하였다. 철자오류 검출 이후 한글 편집거리 알고리즘을 이용해 철자오류 단어와 유사한 단어들을 추출하여 교정 단어 리스트를 구축할 때 올바른 정상단어와 철자오류 단어의 편집거리가 얼마가 되는지 알 수 없다. 그렇기 때문에 수집한 데이터를 통하여 실제 정상단어와 철자오류 단어의 음절 길이에 따른 편집거리에 대한 통계를 추출하여 교정 단어 리스트를 구축하는데 사용하였다.

표 4. 음절 길이에 따른 편집거리

음절길이 / 편집거리	2	3	4
1	129	107	92
2	2	22	34
합	131	129	126

표 4는 단어의 음절 길이에 따른 정상단어와 철자오류 단어의 편집거리를 보여주고 있는데, 음절길이 길어질수록 편집거리가 '1'인 단어의 개수가 줄어들고 있으며 '2'인 단어는 개수가 증가한다는 것을 알 수 있다. 음절 길이가 어느 이상이 되면 편집거리가 '3'인 단어도 등장할 수 있으나 본 논문에서 수집한 데이터에서 음절 길이가 '4'보다 큰 단어는 존재하지 않기 때문에 무시하게 된다. 표 4의 결과를 통해 표 5와 같은 교정 단어 리스트 구축을 위한 규칙을 제정할 수 있었다.

표 5. 교정 단어 리스트 구축을 위한 규칙

1.	최소 하나의 음소에서 철자오류가 발생하므로 음소 단위를 기준으로 하여 교정 단어들을 추출한다.
2.	한 음절에 편집거리가 2회 이상 발생한 단어는 추출할 필요가 없다.
3.	음절 길이가 2인 철자오류 단어는 편집거리가 2인 교정 단어를 추출할 필요가 없다.
4.	음절 길이가 3 이상인 철자오류 단어는 편집거리가 2인 교정 단어를 추출할 필요가 있다.

표 5에서 제정한 교정 단어 리스트 구축이 완료되었다면, 철자오류 단어와 교정 단어 리스트에 존재하는 단어들을 하나씩 치환하면서 코사인 유사도 평균을 구하게 된다. 표 2의 예시문장에서 존재하는 철자오류는 '사임'으로 교정 단어들은 '상임'과 '사이', '짜임'이 등장하였다. 이 중 '짜임'은 문장 내 다른 단어들과의 동시 등장 빈도가 존재하지 않아 제외되었으며, '상임'과 '사이'를 '사임'과 치환한 후 코사인 유사도 평균을 구하게 되면 각각 '0.84', '0.70'이 나오게 된다. 철자오류 교정은 철자오류 검출과 다르게 단순히 가장 높은 코사인 유사도 평균이 나온 단어를 선택하여 치환함으로서 완료한다. 본 논문에서 제안한 철자오류 교정 방법을 통해 Precision, Recall, F-Measure을 구하게 되면 표 6과 같은 결과가 나오게 된다.

표 6. 철자오류 교정 알고리즘의 성능

	Value
Precision	97.80
Recall	89
F-Measure	93.19

V. 결론

본 논문에서는 기존의 코사인 유사도 알고리즘을 이용하여 철자오류를 교정하는 알고리즘과 철자오류 교정 단어를 추출하기 위하여 한글 편집거리 알고리즘을 이용해 철자오류 단어와 편집거리가 작은 단어들을 추출하는 알고리즘을 복합한 새로운 철자오류 교정 알고리즘에 대해서 제안하였다. 그 결과 상당히 높은 수준의 정확도를 보여주며 성공적인 연구가 진행되었다고 할 수 있다. 일반적인 철자오류 교정에 대한 알고리즘은 문장에서 등장하는 각 단어들의 최대 등장 빈도를 구하기 위해 나이브 베이지안의 방식을 많이 사용하고 있으나 코사인 유사도 알고리즘을 사용해서 철자오류를 교정한 본 연구와 같이 각양각생의 방법을 이용하여 철자오류 교정 연구를 수행할 수 있다. 하나의 방법으로만 철자오류 교정을 수행하게 된다면, 해당 연구는 하나의 알고리즘으로만 틀에 박힌 단순한 연구로만 남게 될 수 있다. 차후 철자오류 교정을 수행할 경우 하나의 알고리즘에 틀 박힌 연구가 아닌 다종다양한 알고리즘을 이용하여 철자오류 교정 연구를 수행할 필요가 존재한다.

References

[1] 최철, 박세진, 김철중, 권규식, "쿼르타이 키보드에 기초한 인간공학키보드 설계를 위한 오타율 분석," 대한인간공학회 학술대회논문집, 제 2000-1권, 제-호, 142-145쪽, 2000년
[2] 최현수, 권혁철, 윤애선, "동적 윈도우를 갖는 조건부확률 모델을 이용한 한국어 문맥의존 철자오류 규정 규칙의 재현율 향상," 정보과학회논문지, 제4권, 제5호, 629-636쪽, 2015년
[3] 김경식, 최성기, 권혁철, "극한 언어사용 환경에 적응적인 문맥의존 철자오류 교정 기법," 한국

정보과학회 학술발표논문집, 제2015권, 제6호, 654-656쪽, 2015년
[4] 김민호, 권혁철, 최성기, “어절 N-gram을 이용한 문맥의존 철자오류 교정,” 정보과학회논문지, 제414권, 제12호, 1081-1089쪽, 2014년
[5] Aminul Islam, Diana Inkpen, “Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity”, ACM Transaction on Knowledge Discovery from Data(TKDD), Vol.2, No.2, pp.1241-1249, 2008.
[6] Aminul Islam, Diana Inkpen, “Real-Word Spelling Correction Using Google Web 1T 3-Grams”, Proceedings of The 2009 Conference on Empirical Methods in Natural Language Processing, Vol.3, No.3, pp.1241-1249, 2009.
[7] 김민호, 권경식, 권혁철, “교정 어휘 쌍을 이용한 통계적 문맥 철자오류 교정,” 한국정보과학회 학술발표논문집, 제2013권, 제6호, 607-609쪽, 2013년
[8] Mark D. Kernighan, Kenneth W. Church, William A. Gale, “A Spelling Correction Program Based on a Noisy Channel Model”, Proceedings of The 13th Conference on Computational Linguistics, Vol.2, No.1, 1990.
[9] 노강호, 김진욱, 김은상, 박근수, 조환규, “한글에 대한 편집 거리 문제,” 정보과학회논문지 : 시스템 및 이론, 제37권, 제2호, 103-109쪽, 2010년
[10] 노강호, 박근수, 조환규, 장소원, “음소의 분류 체계를 이용한 한글 편집거리 알고리즘,” 정보과학회논문지 : 시스템 및 이론, 제37권, 제6호, 323-329쪽, 2010년

저 자 소 개

박승현(학생회원)

2012년 조선대학교 컴퓨터공학과 학사 졸업(공학사).
2017년 현재 조선대학교 소프트웨어융합공학과 석사 과정.
<주관심분야 : 자연어처리, 철자오류 교정>

이은지(학생회원)

2012년 조선대학교 컴퓨터공학과 학사 졸업(공학사).
2015년 조선대학교 박사수료
2017년 현재 조선대학교 석박사 연계과정
<주관심분야 : 정보처리, 소셜네트워크, 시맥틴 웹, 온톨로지>

김판구(정회원)

1988년 조선대학교 컴퓨터공학과 학사 졸업(공학사).
1990년 서울대학교 컴퓨터공학과 석사 졸업(공학석사).
1994년 서울대학교 컴퓨터공학과 박사 졸업(공학박사).
1994년 ~ 현재 조선대학교 교수
<주관심분야 : 정보검색, 시맨틱웹, 자연어처리, 빅데이터>