韓國컴퓨터情報學會 論文誌
第16卷 第12號, 2011. 12.

2011-16-12-1-7

문서내 단어간 비교를 통한 철자오류 검출

김 동 주*

Detecting Spelling Errors by Comparison of Words within a Document

Dong-Joo Kim*

요 약

일반 출판물과는 달리 문서 편집기를 사용하여 작성중에 있는 문서에는 사용자의 실수에 의한 오타 오류가 자주 발생한다. 이와 같은 온라인 문서에서 맞춤법 오류의 다수를 차지하는 사용자의 오타 오류는 대부분 자판을 입력할 때 주위 문자를 잘못 입력하는 경우이다. 통상적인 철자 검사기는 이러한 오류들을 형태소 분석기를 이용하여 검출하고 교정하게 된다. 즉, 입력된 어절에 대해 형태소 분석을 시도하고 분석되지 않은 어절을 철자 오류로 간주하게 된다. 그러나 오타 입력된 어절임에도 불구하고 형태소 분석에 성공한 경우에는 이와 같은 방법으로는 검출이 불가능하다. 본 논문에서는 기존 방법들이 검출하지 못했던 철자 오류들을 검출해 낼 수 있는 방법을 제시한다. 이 방법은 문서 작성자의 오타 입력은 반복하여 입력되지 않는 경향이 있으므로 저빈도로 발생한다는 특성에 기반하여 제안되었다. 저빈도의 어절의 자소 대치를 통해 문서의 특정 구간 내의 다른 단어와 비교하여 오타일 확률이 적은 단어인 자주 나오는 단어와 매칭이 된다면 일단 오류 후보로 가정하는 것이다. 여기에는 몇 가지 경험적인 제약이 추가되어야 한다. 이러한 단어간 비교에 의한 추정은 기존에 발견하지 못했던 구문오류뿐만 아니라 일부 의미오류까지 검출할 수 있으며, 교정 후보 선정시 가중치 적용에도 사용될 수 있다.

▸ Keyword: 오타, 철자오류, 형태소분석, 철자검사기

Abstract

Typographical errors by the author's mistyping occur frequently in a document being prepared with word processors contrary to usual publications. Preparing this online document, the most common orthographical errors are spelling errors resulting from incorrectly typing intent keys to near keys on keyboard. Typical spelling checkers detect and correct these errors by using

• 투고일 : 2011. 10. 31, 심사일 : 2011. 11. 25, 게재확정일 : 2011. 12. 14.
* 안양대학교 컴퓨터공학과(Dept. of Computer Engineering, Anyang University)

morphological analyzer. In other words, the morphological analysis module of a speller tries to check well-formedness of input words, and then all words rejected by the analyzer are regarded as misspelled words. However, if morphological analyzer accepts even mistyped words, it treats them as correctly spelled words. In this paper, I propose a simple method capable of detecting and correcting errors that the previous methods can not detect. Proposed method is based on the characteristics that typographical errors are generally not repeated and so tend to have very low frequency. If words generated by operations of deletion, exchange, and transposition for each phoneme of a low frequency word are in the list of high frequency words, some of them are considered as correctly spelled words. Some heuristic rules are also presented to reduce the number of candidates. Proposed method is able to detect not syntactic errors but some semantic errors, and useful to scoring candidates.

▸ Keyword:Mistyping, Typographical Error, Morphological Analysis, Spell Checker

I. 서 론

오늘날 개인 업무, 출판, 신문사 등 사회의 전 분야에 걸쳐 컴퓨터를 활용한 전자출판 및 전자 문서편집기의 사용이 이루어지고 있으며 편집되는 전자 문서의 양도 날로 증가하고 있다. 문서를 작성하는데 있어 맞춤법 규칙에 따라 문서를 작성한다는 것은 쉬운 일이 아닐 뿐만 아니라, 문서 편집기를 사용해 입력할 때 입력자의 실수에 의한 오타 오류는 매우 빈번하게 발생하게 된다. 이러한 철자 오류를 포함한 맞춤법 오류를 검출하고 교정하기 위한 많은 한국어 철자 검사기가 상용화되었다.

현재 상용화되어 있는 한국어 맞춤법 검사기들은 맞춤법 오류와 철자오류를 포함하여 구문오류와 의미오류, 심지어 문체오류까지도 검출하고 있다. 철자오류 특성을 이용한 철자오류의 검사와 교정에서는 충분히 신뢰할만한 결과를 보이고 있는 반면 문법오류와 의미오류의 처리는 제한된 형태에 대해서만 검출 및 교정이 이루어지고 있다. 이로 인해 한국어의 문법 및 의미 체계에 익숙하지 않은 일반인에게 당연히 검출되어야 하는 것이라고 생각되는 오류들이 검출되지 않아 맞춤법 검사기의 신뢰도를 떨어뜨리고 있는 실정이다. 이에 따라 일반인이 쓰기에 적합하고 신뢰도가 높은 맞춤법 검사기를 구현하기 위해서는 검출 형태에 대해 적용범위가 넓은 구문오류 검출기와 의미오류의 검출기, 그리고 각각의 교정기가 요구되고 있다.

철자오류는 크게 입력자의 무지에 의한 오류와 입력 오타에 의한 오류로 나눌 수 있다. 입력자의 무지에 의한 오류는 철자법 지식의 부족으로 잘못 입력하는 경우를 말하며, 유사 발음으로 인한 잘못된 단어를 입력하는 경우가 많다. 이러한 오류들은 반복적으로 발생하고 예측이 가능하다는 특징으로 인하여 어느 정도 검출/교정이 가능하다. 그러나 입력자의 오타에 의한 오류는 반복적인 패턴을 보이지 않아 예측이 불가능해 지식을 통한 검출/교정이 쉽지 않다.

기존의 맞춤법 검사기에서 입력 오타에 의한 철자오류는 형태소분석을 통해 검출을 하고, 자소 대치와 형태소 분석을 통해 올바른 어절로의 교정을 시도하고 있다[1,7,8,9]. 형태소 분석을 통한 오류 어절의 검출은 형태소 분석기가 검사 대상이 되는 어절에 대해 성공적으로 분석 해낸다면 적법한 어절로 판단하고 분석하지 못한다면 철자오류 어절로 간주하게 된다. 그런데 대부분의 오타로 인한 철자오류는 주로 자판 입력시 의도한 자소 주위의 다른 자소를 잘못 입력하는 경우가 대부분이다. 따라서 기존의 철자검사기에서 철자오류의 교정은 어절내의 모든 자소에 대해 오타 자소로 가정하고 각 자소에 대해 키보드상 인접 자소로 대치한 후 형태소 분석을 통한 적법성 검사로 교정 후보를 생성해낸다. 그러나 이러한 기존 방법들은 오타 입력된 어절에 대한 형태소분석이 성공한 경우 철자오류를 검출해내지 못한다.

본 논문에서는 형태소분석기에서 적법하다고 판단한 철자오류 어절을 검출하고 교정할 수 있는 방법을 제안한다. 제안하는 방법은 오타 입력된 어절이 특정 구간내의 문서에서 반복되지 않는 특성을 이용한 방법이다. 반복되지 않는 어절을 자소변환을 통해 교정을 시도하고, 자소변환된 어절을 문단내의 다른 어절과 비교하여 오타일 확률이 적은 어절과 매칭이 된다면 오류 후보로 제시하게 된다. 이러한 단어간 비교에 의한 추정은 오류 패턴에 대한 지식의 구축없이도 구문오류 및 의미오류를 검출/교정할 수 있게 해준다.

본 논문의 구성은 다음과 같다. 제2장에서 오류 유형을 분류하고 기존의 구문 및 의미오류 처리 방법에 대해 논한다.

제3장에서는 본 논문에서 제안하는 단어간 비교에 의한 추정 방법에 대해 서술하고, 제4장에서는 제안하는 방법의 성능을 평가하기 위한 몇 가지 실험결과를 제시하고, 마지막 5장에서 이 논문의 결론을 맺는다.

II. 관련 연구

1. 오류 유형 분류

중고 교과서에서의 오류 유형을 나타내는 표 1[2]에서 보는 바와 같이 일반적인 인쇄물인 경우, 띄어쓰기, 맞춤법 오류 및 어미나 조사의 오용이 전에 오류의 약 60%를 차지한다.

표 1. 교과서의 오류 유형
Table 1. Error types in a textbook

오류 유형	개수	비율(%)
띄어쓰기 오류	808	29.5
어미와 조사의 오용	689	25.1
적절하지 못한 낱말	379	13.8
문맥에 호응하지 않는 문장	147	5.3
맞춤법 오류	132	4.8
기타	589	21.5
합계	2,744	100.0

표 2. 문서편집기로 작성되는 문서에서 오류의 유형
Table 2. Error types in documents written with editor

오류 유형	비율(%)
띄어쓰기 오류	26.1
맞춤법 오류	10.7
철자 오류	47.5
구문 및 기타 오류	9.1
기타	6.1

그러나 문서 편집기의 경우에는 표 1과는 다소 다른 유형을 나타낸다. 문서 편집기를 사용하여 문서를 작성할 때는 문서의 내용을 자판을 통해 직접 입력하게 되므로 자판을 잘못 입력하게 되는 경우가 빈번히 발생한다. 표 2는 몇몇 신문사를 통해 얻은 교정 작업 전의 신문 기사와 인터넷상의 소설 등을 조사한 오류 유형이다[1]. 표 2에서 알 수 있듯이 문서 편집기를 사용하여 문서를 작성할 때, 오타에 의한 철자 오류가 전제 오류의 절반을 정도를 차지할 정도로 매우 중요하다.

문서편집기를 사용할 때 문서에 나타나는 오류는 크게 입력자의 무지에 의한 오류와 오타 입력에 의한 오류로 나눌 수 있다. 사용자의 무지에 의한 오류는 (1)과 같은 경우로 사용자가 맞춤법 규정을 잘 모르거나 잘못 알아서 생기는 오류이다. 표 2의 분류로 (1)의 예는 구문 및 기타 오류에 속하는 것으로, 문법 요소간의 불일치(Disagreement), 문체 오류, 오용어 오류(Misused words)와 문맥상의 의미오류(Wrong meaning in context)에 해당한다.

(1) a. 밥을 먹으러 하다. (먹으려)
b. 검붉은 피가 엉키다. (엉기다)
c. 선생님께서 수학을 가리키신다. (가르치신다)

(2) a. 고성능 → 고선능 (근접자소를 잘못 입력)
b. 등으로 → 등오로 (자소를 잘못 입력)
c. 사슴을 → 가슴을 (잘못된 입력, 형태소분석 성공)
d. 포옹 → 포용 (유사어 오류)

입력자의 오타에 의한 오류는 의도한 자소를 실수로 다른 자소를 입력하는 경우이다. 이 오류는 철자오류의 대부분에 해당할 정도로 매우 많이 발생한다. 입력자의 오타에 의한 오류의 예는 (2)와 같다. 기존의 철자검사기에서는 철자오류 단어를 형태소 분석하여 어절의 적법성 검사를 실시하고, 분석 실패하여 적법하지 않은 어절로 판단하면 인접자소 대치를 통하여 교정을 시도하게 된다. 이와 같은 방법으로 (2-a)와 (2-b)의 오류는 검출이 가능하다. 그러나 (2-c). (2-d)와 같이 '사슴을'이라는 단어를 실수로 '가슴을'이라고 입력했을 경우 형태소 분석에 성공하여 적법한 어절로 판단하기 되므로 기존의 철자검사 방법으로는 검출 불가능하다. 물론 이는 오타입력이 의미오류로 발전한 경우이며 기존 맞춤법 검사기에서는 목적어와 용언 사이의 의미오류 관계에 관한 지식정보를 참조하여 검출 및 교정을 수행하게 된다. 그러나 이러한 지식정보는 입력자가 반복적으로 틀리는 것에 대해서만 구축되기 때문에 (2-c)의 예와 같이 예측불가능한 오류에 대해서는 대안이 없다. (2-d)와 같이 발음이나 철자가 유사해 잘못 입력하기 쉬운 유사어 오류 또한 기존 방법으로는 검출 및 교정이 용이하지 않다.

2. 교정 방법

기존 방법에서는 구문 및 의미오류를 검사하기 위해서 형태소 분석, 구문 분석을 수행한 결과뿐만 아니라, 오류를 찾아내기 위해서 문단 또는 문장에서 나타날 수 있는 문법 오류나 의미 오류에 대한 지식이 필요하다. 이러한 지식들은 규칙과 교정 방법의 형태로 표현되어 지식베이스(Knowledge base)를 구성하게 되는데, 구문 및 의미 오류 검사에서는 이 부분이 검사기의 성능을 결정하는 중요한 부분이다. 축적된

지식베이스를 이용하여 구문 오류를 검출하고 교정하는 방법에는 연어(collocation) 정보를 이용한 방법[3]과 이에 더 나아가 부분 구문 분석(partial sentence analysis)을 이용해 여러 단어에 거쳐 존재하는 오류를 발견하는 방법[4,5] 등이 연구되고 있다.

연어정보를 이용한 방법은 한 단어 단위의 분석을 다수 단어로 확장하여 근접한 구문 오류를 검출하는 것이다. 예를 들어 예문과 같이 '땀'이라는 단어는 '배다'라는 단어와 호응이 되지 않고 '베다'라는 단어와 호응하여 쓰인다는 것이다.

(3) a. 옷에 땀이 배어서 입을 수가 없다. (×)
b. 옷에 땀이 베어서 입을 수가 없다. (○)

이러한 방법을 사용하려면 오류 유형별 처리 규칙과 지식베이스가 필요하다. 이 규칙은 일차적으로 단어들간의 연관관계에 기인한다. 즉, 사람이 혼동하기 쉬운 단어를 '검사 단어'로 정하고 그 단어와 연어 관계(Collocation relation)에 있는 단어, 연어 오류 관계(Anti-collocation relation)에 있는 단어를 '검사 기준'으로 삼아 의미 오류의 발생 여부를 검사한다.

예를 들어 '소화시키다, 마음을 가라앉히다'란 뜻의 '삭이다'와 연어 관계에 있는 단어는 '분노, 음식물, 돈, 화' 등이고 연어 오류 관계에 있는 단어는 '술, 감주' 등이다. 반대로 '발효시키다'란 뜻의 '삭히다'와 연어 관계에 있는 단어는 '술, 감주'이고 연어 오류 관계에 있는 단어는 '분노, 음식물, 돈, 화' 등이다. 즉 '분노를 삭히다'란 문장은 검사 단어 '삭히다' 앞에 연어 오류 관계에 있는 '분노'가 있으므로 의미상 틀린 단어이다. 이때는 '분노'와 연어 관계에 있는 '삭이다'로 교정해야 한다.

이러한 단어간 연어 정보에 기반한 오류 검사를 위해서 고려해야할 몇 가지 사항이 있다. 첫째, 검사 단어와 연어 관계, 연어 오류 관계에 있는 단어를 분류하는 기준의 타당성과 정확성의 문제이다. 이를 위해 지식베이스를 구축하는 과정에서 많은 조사와 전문가의 의견 수렴이 필요하다. 둘째, 검사 단어와의 연어 관계, 연어 오류 관계에 있는 단어를 정의하는 방법의 효율성 문제이다. 앞의 예처럼 연어 관계에 있는 단어를 단지 나열하는 것이 아니라 각 단어를 의미 분류하여 사용해야 효율적이 된다. 따라서 사전에 있는 명사들의 의미와 기능에 따라 최대한 하위 범부로 나누고 그 정보만으로 검사 단어와 연어 관계 여부를 확인할 수 있도록 사전을 구성하게 된다[3].

단어간 연어 정보에 기반한 방법으로 오류를 검출할 때, (4)와 같은 경우에서는 오류 단어가 있는 단어가 인접해 있지 않기 때문에 검사 단어를 기분으로 의미 오류가 있는지를 찾기 힘들다.

(4) 주차 문제로 시비를 이웃간에 몹시 심하게 가린다.

(4)에서 '옳고 그름을 분간하다'의 뜻으로 '가르다'가 바른 표현이다. '가리다'는 '여럿 가운데 일전한 것을 구분해 내다', '낯선 사람들을 대하기 싫어하다' 등의 뜻이다. 따라서 '시비'와 '가리다'는 연어 오류 관계에 있다. 그런데 두 단어의 문장 내 관계는 목적어와 술어 관계이다. 따라서 검사 단어 '가리다'와 연관 오류 관계에 있는 단어, 즉 '시비'를 검색하기 위해 목적어와 술어 사이에 삽입 가능한 모든 문장 성분을 고려해야 한다.

이를 해결하기 위해 부분 구문 분석기법에 의한 교정 방법이 제안되었다[4,5]. 이 방법은 의존 문법에 기반한 부분 구문 분석을 수행하여 여러 단어에 걸쳐 떨어져 있는 오류 단어를 검색하게 된다.

의존 문법은 문장 내에 나타나는 단어들의 어순은 고려하지 않고 언어 요소 상호간의 의존 관계만을 밝히려는 이론으로 어순이 자유로운 한국어 분석에 적합하다. 의존 문법에서는 두 구문 단위 사이의 문법적 관계가 존재하면 의존관계에 있다고 말하며, 이 관계에 있는 구문 요소들을 지배소와 의존소로 크게 구분한다. 지배소란 의존 관계에서 의미의 중심이 되는 요소이며, 의존소란 지배소가 갖는 의미를 보완하는 요소이다. 일반적으로 지배소는 여러 개의 의존소를 가질 수 있으나 의존소는 하나의 지배소만을 갖는다[6]. 이와 같은 특성을 가진 의존 문법은 부분 구문 분석에 쉽게 적용될 수 있다.

의존 문법에 기반한 부분 구문 분석은 검사 단어와 연관 오류 관계에 있는 단어를 찾기 위해 지배소와 의존소 관계에 놓인 어절들을 분석한다. 지배소와 의존소 관계표[5]에 따라 (4)를 의존 문법의 지배소와 결합 가능한 의존소를 고려하면서 연관 관계에 있는 문장 요소(목적어)가 발견될 때까지 분석하게 된다.

검사 단어를 기준으로 오른쪽 우선 부분 구문 분석을 수행한다. 검사 단어 '가리다'는 지배소로서 주어, 목적어, 용언, 수식어 등을 의존소로 가질 수 있다. 앞 단어('심하게')를 받아 형태소 분석한 결과 '심하게'는 형용사 어간에 부사형 어미가 결합한 용언 수식어로 지배소와 의존 관계에 있다. 다음 단계에서 '심하게'를 지배소로 삼으면 용언 수식어가 의존소로 올 수 있다. 용언 수식어는 '몹시'라는 부사가 의존소로 왔다. '몹시'를 지배소로 하여 다음 단어를 받아 형태소 분석한 결과 의존 관계가 성립하지 않는 '명사+부사격조사'로 분석되는 '간에'가 왔다. 그러나 아직 여기서 검사 단어와 목적어 관계에 있는 연어 관계 혹은 연어 오류 관계 단어를 검색하지 않았으므로 부분 구문 분석을 계속 수행한다. 즉, '간에'를 지배

소로 하여 다음 단어와의 의존 관계를 검사한다. 지배소 '이웃'은 의존소로 명사나 명사 수식어를 취할 수 있다. 그러나 형태소 분석 결과 의존관계가 없는 '명사+목적격조사'인 '시비를'이 앞 단어에 있다. 이때 의존 관계는 성립하지 않지만 연어 관계를 검사해야 하는 문장 요소가 나왔으므로 더 이상 부분 구문 분석을 하지 않는다. 즉, 연어 오류 관계 여부를 검사해야 할 문장 요소가 분석되었으므로 부분 구문 분석을 끝내고 의미 오류 처리 규칙에 기반한 검사 및 교정을 수행한다.

이 같은 부분 구문 분석 방법으로 여러 단어에 걸쳐 떨어져 있는 검사 단어와 연어 관계 단어간의 의미 및 문체 오류 발생 여부를 검사할 수 있으나, 문형 비교(Pattern matching)에 의해 문서 오류를 검출하는 접근 방법은 사람들이 자주 틀리는 오류를 정확하게 잡아주며 구문 오류 검출 및 교정(Syntactic critique)이라는 문제에 쉽게 접근할 수 있는 장점이 있다. 반면 관련 규칙을 만드는데 오랜 시간이 걸리며, 예측하지 못하는 오류를 처리하지 못하는 문제점이 있다.

문서의 구문 오류를 모두 검출해내기 위해서는 원칙적으로 문장 전체를 구문 분석한 후 오류가 존재하는 단어를 찾아내야 한다. 그래서 한국어 분석에 관한 연구가 계속되고는 있지만 문법 검사기에 사용할 수준의 한국어 문장 분석 기법은 아직 연구되어 있지 않다. 문장 전체를 구문 분석하여 오류를 검출한다는 것은 현재 기술로는 힘들기 때문에 구분 분석을 수행한 후 교정하는 방법에 대한 연구도 미흡한 상태이다.

가능한 또다른 방법으로는 모든 단어들을 검사한다는 것이 어렵기 때문에 범위를 축소하여 오류의 가능성이 있는 단어나 구나 나타날 때마다 그와 대치되는 필요한 문장 요소를 부분 구문 분석하여 찾는 방법인데, 본 논문에서 제안하는 단어간 비교에 의한 추정도 이러한 오류의 가능성이 높은 단어를 문장내의 단어간 비교를 통해 추정하게 된다.

III. 본 론

앞 장에서 기술한 구문 오류 교정 방법은 구문 오류를 검출하기 위해 전통적인 방법 아래 연구되어온 것들이다. 본 논문에서는 전통적인 방법과는 달리 교정 후보를 선정하는 방법에서 착안하여 형태소 분석만으로도 구문 오류뿐만 아니라 의미오류를 추정 또는 검출할 수 있는 단어간 비교에 의한 추정 방법을 제안한다. 3.1절에서는 단어간 비교에 의한 추정 방법을 위한 가정을 살펴보고 3.2절에서는 기본적 알고리즘에 대해 설명하며 3.3.절에서는 구문 오류를 검출하기 위해 추가된 휴리스틱한 제약과 전반적 알고리즘을 기술한다.

1. 철자 오류의 특징

제안하는 방법을 설명하기 전에 철자 오류의 빈도를 살펴볼 필요가 있다. [7]에 따르면 철자오류의 대부분은 95%가 자소가 대체, 탈락, 추가된 오류이며, 73%가 자소가 대체되어 발생한 경우임을 알 수 있다. 이를 근거로 철자 검사기는 오류 단어를 발견하고 교정하기 위해 오류 단어의 각 자소를 자판 주위의 자소와 바꾸어 형태소 분석을 수행해 분석에 성공하면 이 단어를 교정 후보로 제시하게 된다. 보통 이때 철자 오류는 한 음절, 단 하나의 자소에서 발생한다고 가정한다. 이는 한 단어에서 두 개 이상 발생하는 경우는 극히 드물다는 통계자료에 근거하며, 이 가정을 기반으로 하는 알고리즘은 계산량을 줄일 수 있다.

그러나 입력 단어가 형태소 분석에 실패하는 경우 철자 오류 검사로 교정이 가능하지만, (5)의 예와 같이 사용자가 오타 입력한 단어가 형태소 분석에 성공한다면 오류 단어를 검출해낼 수 없다.

(5) a. 구문 오류 → 고문 오류
b. 사슴을 → 가슴을
c. 가능 → 기능
d. 기반한 → 기발한

(5-a, b, c)는 주위 자판을 잘못 눌러 입력한 경우이다. 그러나 (5-d)의 예는 'ㄴ' 주위의 자판이 아닌 'ㄹ'을 잘못 입력한 경우이다. 기존의 철자 검사기에서는 형태소 분석에 실패한 단어에 대해서만 교정을 시도하기 때문에 (5)의 오류들은 검출할 수 없다. 또한 (5)의 오류는 예측 불가능하게 발생하기 때문에 앞서 2장에서 언급한 문형 비교에 의한 교정 방법으로도 검출이 거의 불가능하다. 이러한 오류는 구문 또는 의미적 교정 단계에서 처리하게 되지만 아직까지 구문 교정에 쓰일 만큼의 구문 분석 기법이 연구되지 않은 것이 현실이다.

본 논문에서는 이러한 구문 오류를 검출할 수 있는 경험적 방법을 제안한다. 이 방법은 구문적인 전통의 방법과는 별도로 형태소 분석만으로도 구문 오류를 검출 또는 추정할 수 있다. 즉, 구문 분석 없이 형태소 분석 결과만을 가지고 문서 내의 단어간 비교에 의해 구문 오류가 있는 단어를 추정하는 것이다.

본 논문에서는 이러한 오류를 검출하고 교정하기 위해 단어간 비교에 의한 추정(Deduction using words comparison)이라는 방법을 제안한다. 이는 단어간 비교에 의한 추정을 통한 구문 오류 검출은 사용자 a) 오타의 대부분은 자판을 입력할 때 주위 문자를 잘못 입력하는 경우이다, b) 입력자는 반복하여 오타 입력하지 않으므로 문서내의 특정

구간에서 저빈도로 발생한다는 경험을 바탕으로 한다. 이에 해당하는 예들은 (6)과 같다.

(6) a. 구문적 분석 ... 구문적 분석 ... 고문적 분석
b. 별 ... 별 ... 벌
c. 계제 기반 기계번역 ... 예제 기반 기계번역, ... 예제기반기계번역

(6-a)는 학술발표 논문집에 실렸던 논문의 오류이며, (6-b)는 소설 '어린 왕자'에서 발견된 오류이다. 그리고 (6-c) 문서 교정을 보기전의 자연언어처리 분야의 석사학위 논문 초록에서 발견된 오류이다. 실제 문서 작성자는 자주 나오는 단어를 틀리게 쓰는 경우가 많으며, 이러한 오류들은 마지막 교정을 거친 후에도 고쳐 써지지 않을 가능성이 매우 높다.

2. 오류의 추정

단어간 비교에 의한 추정은 오타일 확률이 높은 단어를 자소 대치를 통해 특정 범위내의 다른 단어와 비교하여 오타일 확률이 낮은 단어와 매칭이 된다면 오류 후보로 간주하는 것이다. 따라서 오타일 확률이 낮은 단어와 오타일 확률이 높은 단어를 각각 정의할 필요가 있다. 입력자는 자주 발생하는 단어에서 오타를 입력할 확률이 높다는 경험을 바탕으로 문서의 특정 범위 내에서 오타일 확률이 낮은 단어와 오타일 확률이 높은 단어는 표 3과 같이 정의한다. 표의 오른쪽에 나오는 비율은 '국민 교육 헌장'을 분석하여 얻은 결과이다.

표 3. 단어가 오타일 확률
Table 3. Error types in a textbook

오류 가능성	대상 단어	비율(%)
낮음	3회 이상 출현한 단어	3
	2회 출현한 단어	10
	1음절 단어	12
높음	1회 출현한 단어	75

단어간 비교에 의한 추정에서 대상 어절(혹은 단어)은 오타일 확률이 높은 어절을 의미하며, 비교대상 어절(혹은 단어)은 오타일 확률이 낮은 단어를 의미한다. 이에 따라 기본적인 아이디어를 기술하면 다음과 같다.

1번 발생한 어절(혹은 단어)을 검사대상 목록에 추가하고, 특정 구간내에 2번 이상 발생하는 단어는 해당구간의 비교대상 목록에 추가

대상 어절을 자소 대치 후에 비교대상 목록의 어절과 비교하여 일치하는 것이 존재하면 자소대치된 대상 어절을 오류 후보로 간주

이들 규칙만으로도 자주 발생하는 단어를 잘못 입력하는 오류의 대부분을 검출할 수 있다. 이 규칙을 적용한 전체적인 단어간 비교에 의한 추정 알고리즘은 그림 1과 같다.

```
문서의 끝까지 반복 {
    형태소분석기를 이용하여 적법성 검사
    검사 성공한 어절을 검사대상목록에 저장
    만약 어절이 비교대상목록에서 2회 이상의 빈도를 가지면
        검사대상목록에서 제거, 비교대상목록에 추가
    비교대상목록이 충분하다면 {
        입력 어절이 비교대상목록에 존재한다면
            처음으로
        자소 대치된 어절 목록 ← 원형을 자소대치
        자소 대치된 어절 개수만큼 반복 {
            자소대치된 어절이 비교대상목록에 존재한다면
                자소대치된 어절을 교정후보목록에 추가
        }
        자소 대치된 단어 목록 ← 표층형을 자소대치
        자소 대치된 단어 개수만큼 반복 {
            자소대치된 단어가 비교대상목록에 존재하다면 {
                비교대상목록의 단어의 문맥 어절로 변환
                형태소분석기를 이용하여 적법성 검사
                검사 성공하면 {
                    검사대상목록 단어의 문맥어절로 변환
                    변환된 어절을 교정후보목록에 추가
                }
            }
        }
    }
    입력어절 ← 다음어절
}
```

그림 1. 검사 및 교정 알고리즘
Fig. 1. Detection and Correction Algorithm

이와 같이 단어간 비교에 의한 추정 방법은 여러 번 나타나는 단어 중 하나를 오타로 입력하였을 때 오류로 판정할 수 있게 해 준다. 즉 오타를 발견하기 위해 입력 어절에 대해 자소 대치를 수행한 뒤 빈번히 발생하는 어절과 비교하게 된다.

[7]에 따르면 잘못 입력된 자소의 약 85%가 원래 맞는 자모의 자판에 바로 인접하여 위치하며, '등으로'를 '등오로'와 같이 바로 인접하지 않는 자소로 대치된 경우도 15%정도라고 한다. 인접하지 않은 자소로 대치된 경우는 (7)과 같다.

(7) a. 오늘 → 오를
b. 갑으로 → 갑오로

(7-a)는 종성 'ㄹ'의 영향을 받아 초성을 잘못 입력한 경우이고, (7-b)는 뒤 어절의 모음에 영향을 받아 '-으로'를 '-오로'로 잘못 입력한 경우이다. 두 경우 모두 인접하지 않은 자소로 대치된 것이며, 철자 검사기에서는 교정되지 않는 단어

들이다. 본 논문에서는 인접하지 않은 오류도 최대한 검출하고 교정하기 위해 다음과 같은 경험적 규칙을 추가하였다.

규칙 1: 초성과 종성이 같은 자소일 경우 초성 자소를 다른 모든 초성 자소로 대치

규칙 2: 뒤 음절과 현재 음절의 중성 자소가 같을 경우 현재 음절의 자소를 다른 모든 중성 자소로 대치

규칙 3: 1음절 단어는 비교 대상에서 제외

규칙 1과 규칙 2는 인접하지 않은 오류를 검출하기 위해 첨가한 규칙이며, 규칙 3은 보통 1음절 단어에서의 오타 오류는 잘 발생하지 않으며 1음절 단어까지 자소 대치할 경우 과검출 또는 과추정이 증가하므로 1음절 단어를 비교대상에서 제외하였다.

3. 시스템의 구성

제안하는 알고리즘을 포함하는 전체 맞춤법 검사기의 구조는 그림 2와 같다. 일반적인 맞춤법 검사기는 입력 어절에 대해검사 단계에서 형태소 분석을 실시한 후 분석에 실패한 경우 오류 단어로 판단하여 교정 단계로 넘어가게 된다. 교정 단계로 넘어온 단어는 자소 대치 및 여러 알고리즘을 거쳐 교정 후보를 생성하게 되고 교정 후보 단어에 대해 다시 형태소 분석을 시도하게 된다. 교정 후보 단어들 중 형태소 분석에 성공하면 성공한 단어들을 최종적으로 교정 후보로 제시하게 된다. 철자 검사가 끝난 문장들을 대상으로 구문 오류 검사와 의미 및 문체 오류 검사가 수행된다.

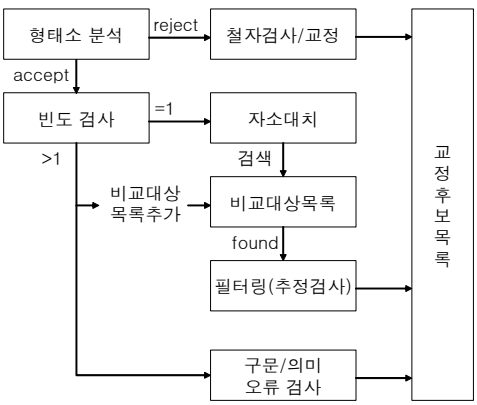

그림 2. 제안하는 맞춤법 검사기 구조
Fig. 2. Architecture of Proposed Critiquing System

그림 2와 같이 철자 검사 단계에서 적법하다고 판단한 단어들을 목록 생성 단계에서 해당 단어가 2회 이상 출현한 비교대상 목록에 포함되어 있는지에 대해 검사한다. 만약 해당 단어가 특정 범위 내에서 1회만 출현하였고 2음절 이상의 단어라면 오류일 가능성이 존재하는 어절이라고 판단한다. 그런 다음 이 어절들은 오류 단어 추정 단계에서 단어간 비교에 의해 구문 오류를 추정하게 된다. 만약 비교대상 목록에 자소 대치된 변형 어절(혹은 단어)이 존재하면 해당 어절을 오류 단어로 추정하며 대치어로는 자소 대치된 변형 어절을 제시한다. 오류로 추정된 어절은 추정 검사 단계에서 추정이 올바른지에 대한 검사를 수행한다.

예를 들어 (8-a)는 대상 단어의 원형인 '분석기'와 비교대상의 표층형인 '분석이'가 서로 일치된 경우이다. 이 경우에는 '분석기'의 자리에 '분석이'를 넣어 형태소 분석을 해 보면 (8-a)와 같은 추정은 잘못된 추정임을 알 수 있으므로 오류 추정을 취소하게 된다. (8-b)는 '가정하기'의 원형인 '가정'과 부사어 '가장'이 일치되는 경우이다. 여기에도 '가정'의 자리에 '가장'을 넣어 형태소 분석하면 추정을 취소할 수 있다. (8-c)와 (8-d)는 동일한 원형을 가진 어절이 표층형이 달라 잘못 추정되는 경우이다. 이 경우에는 각각의 원형이 같은지를 비교한 후 같다면 추정을 취소하게 된다. 이처럼 추정 검사 단계에서는 불필요한 구문 오류 추정을 줄이기 위해 형태소 분석을 통하여 구문 오류 추정에 대한 검증을 실시하게 된다.

(8) a. 분석기/의 → 분석이
b. 가정/하기 → 가장
c. 대하/여 → 대해 (대하 + 어)
d. 나타나/ㄴ → 나타내 (나타나 + 어)

마지막으로 추정 검사단계를 통과한 오류로 추정된 어절은 문서 작성자에게 오류 가능성이 높은 단어가 있음을 알리게 되고 작성자로 하여금 교정을 수행할 것인지 선택하게 된다.

IV. 실험 및 결과

본 장에서는 앞서 언급된 시스템으로 논문에서 제안하는 방법의 적합성을 평가하기 위한 검증을 실시한다. 그런데 검증을 위해 자연스러운 오타가 포함된 문서를 구하기란 쉽지 않은 일이다. 검증을 위한 데이터의 규모에 있어서 비록 제한적이기는 하지만 단어간 비교에 의한 추정에 의해 오류 검출이 가능한지에 대해 실험한 결과를 제시하고, 가장 적합한 방법을 제시한다. 제안하는 방법에 의한 오류 추정의 적합성을 평가하기 위해 오류 검출 여부, 오류 추정비율, 그리고 문서내 단어간 비교의 범위에 대해 조사하였다.

먼저 본 논문의 핵심이라고 할 수 있는 단어간 비교에 의

한 추정을 했을 때 구문 오류를 검출할 수 있는지에 관한 실험을 수행하였다. 이 실험을 위해서 본 논문에서는 테스트 코퍼스로 자연언어처리 분야의 석사논문 초본을 택하였다. 이 논문 초본은 본 논문과는 무관하게 작성된 것이며, 일차 교정을 거친 문서로 띄어쓰기 오류를 제외하고는 철자 오류가 없는 문서다. 전체 어절 수는 3,662개이고, 영문자 및 특수 문자를 포함하는 어절을 제외한 어절의 수는 3,062개이다.

테스트 문서를 단어간 비교에 의한 추정으로 문서를 분석한 결과는 표 4에서 보는 바와 같다. 표 4에서 O 표시는 문서상 구문 오류를 올바로 검출한 경우이고 X표시는 불필요한 과추정의 가리키며 △ 표시는 과추정이지만 추정 검사 단계에서 취소할 수 있는 경우를 가리킨다.

표 4. 오류 검출 여부 실험 결과
Table 4. Experimental Results for Predicting Errors

추정 단어	발견 대치어	검출 여부
따라	나라의	X
이르러서는	이루어진다	△
나누어진다	나무	△
기존의	기본	X
계제를	예제를	O
이식하는	지식	△
전에	전체	X
변역에	번역에	O
기존은	기준	X
이루어진	이후	△

표 5. 자소대치어절의 형태소 분석
Table 5. Morphological Analysis for Generated Eojeol

생성어절	형태소 분석 결과	검색성공유무
자라	자라/n 자/v + 라/e	실패 실패
나라	나라/n 나/v + 라/e	성공 성공
아라	실패	-
가라	가/v + 라/e	실패

표 4에서 검사 대상 어절제약에 맞는 추정 대상 어절 '따라'에 대해 'ㄸ'과 키보드상의 인접 자소인 'ㅈ', 'ㄴ', 'ㅇ', 'ㄱ', 'ㄹ'로 대치해 '자라', '나라', '아라', '가라'라는 어절들을 생성한다. 생성된 어절은 형태소 분석을 거치게 되면 표 5과 같은 분석 결과를 얻을 수 있다. 분석 성공한 어절의 형태소 분석 결과 중 '자라', '자', '나라', '나', '가'의 어간을 오류의 가능성이 적은 비교되어지는 단어 목록에서 검색을 실시한다. 석사학위 논문에서 오류 가능성이 적은 단어 목록에는 '나라'만 유일하게 등록되어 있어 '나라'만이 검색 성공하게 된다. 그런데 '나라'는 사용된 문맥에서는 '나라/n+의/j'로 형태소 분석되었으므로 '나/v+라/e'는 교정 후보에서 제외되고 '나라/n'만 남게 된다.

테스트 코퍼스에서는 10개의 단어를 구문 오류로 추정하였으며 이 중 2개가 실제 오류였다. 이처럼 결과를 낸 것은 논문이라는 테스트 코퍼스 성격상 동일한 단어를 많이 사용하게 되고 일상적인 단어의 사용이 거의 없다는데 기인한 것이다. 이 실험을 통해 단어간 비교에 의한 추정을 통한 검출만으로도 추가적인 구문 및 의미 오류를 검출할 수 있음을 알 수 있었다.

실험에서 고려되어야만 하는 또다른 한 가지는 단어간 비교에 의한 추정에서 맞는 것을 틀리다고 판단하는 과검출 또는 과추정이 많아질 수 있다, 이를 평가하기 위해 두 번째로 문서상에서 추정이 일어나는 정도를 실험하였다. 과추정 정도를 평가하기 위해 사용된 텍스트 문서는 표 6과 같다. 각 텍스트는 철자 오류나 유사어 오류가 없는 문서들이다. 과추정 비율 실험은 문서내의 단어간 비교의 범위를 앞 뒤 150 단어로 정하고 추정 검사 단계를 수행하지 않은 시스템에서 수행되었다.

표 6. 과추정 평가를 위한 텍스트 문서
Table 6. Texts for Evaluation of Over-Predicting

텍스트	단어수	비고
소설 1	1,024	인터넷 소설
소설 2	2,765	인터넷 소설 (대화체)
수필 일부분	1,194	블로그 글
학술발표 논문 1	1,959	형태소분석 관련 논문
학술발표 논문 2	1,309	중의성해소 관련 논문
석사학위논문 초본	3,062	기계번역 관련 논문

표 7. 과추정 평가를 위한 텍스트 문서
Table 7. Texts for Evaluation of Over-Predicting

텍스트	단어수	소계	O	△	X	비율
소설 1 소설 2 수필 일부분	1,024 2,765 1,194	4,983	0	5	16	0.3
학술발표 논문 1 학술발표 논문 2	1,959 1,309	3,268	0	5	5	0.15
석사학위논문초본	3,062	3,062	2	4	4	0.13
총계	11,313		2	14	25	0.21

표 7에서 알 수 있듯이 대부분의 텍스트에서 추정되는 단어의 개수는 1% 미만이었다. 1음절 단어는 처리하지 않고 한 번 출현한 단어에 대해서도 처리하지 않기 때문에 추정되는 단어의 개수는 매우 적다. 또한 추정된 단어 중에서도 추정 검사 단계에서 형태소 분석 및 규칙을 통해 추정을 검증하면 추정을 취소할 수 있는 후보들이었다.

세 번째 실험은 단어간 비교를 수행할 때 어느 정도의 범

위가 적당한지에 관한 실험이다. 인터넷의 글들과, 학술발표 논문, 그리고 두 번째 실험에서 사용하였던 석사학위 논문 초본을 대상으로 실험을 수행하였다. 그림 3은 참조하는 앞뒤 비교 단어의 범위를 변화하면서 실험한 결과이다. 그러나 이 결과만으로는 문서내의 단어간 비교의 최적 범위를 구할 수 없다. 그런데 150 단어 이하에서 오류 단어를 검출하지 못하는 경우가 존재했으며, 그림 3에서 보다시피 범위가 증가할수록 추정 수가 꾸준히 증가하다가 400 단어 이상에서는 증감 변동이 있음을 알 수 있다. 이는 2회 이상 출현하는 단어가 적고 비교 대상이 될 단어가 많은 상태에서 400 단어 이상을 기점으로 비교되어질 단어가 많아지고 대상 단어가 적어지는 상태로 바뀜을 의미한다. 따라서 본 논문에서는 150~300 단어 범위가 적당하다는 결론을 얻었다.

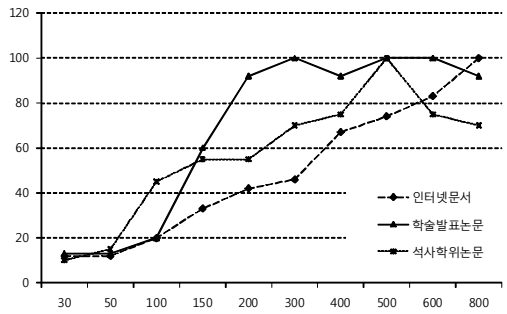

그림 3. 범위에 따른 추정 개수
Fig.3. Number of Estimated Words by Varing the Range

V. 결론

본 논문에서는 형태소 분석만으로도 구문 및 의미 오류의 일부를 검출할 수 있는 방법인 단어간 비교에 의한 추정 방법을 제시하였다. 이 방법은 문서 작성자는 자주 발생하는 단어에 대해 오타를 입력할 가능성이 높다는 가정하에 제안되었으며, 자소 대치의 방법으로 단어간 비교를 수행하여 구문 오류를 추정하였다.

실험을 통해 단어간 비교에 의한 추정은 성공적으로 오류를 검출함을 알 수 있었으며, 또한 우려할 수준의 과추정이 발생하지 않음을 알 수 있었다. 따라서 철자 검사기나 문법 검사기와 이 방법을 통합하여 사용함으로써 추가적인 오류를 검출 할 수 있을 뿐만 아니라 교정 후보 선정시 가중치 적용에도 사용할 수 있다.

올바른 구문 교정을 위해서는 단어간 비교에 의한 추정을 수행한 뒤 오류 후보에 대해 구문분석을 수행해야 할 것이다. 구문 분석을 위해서 오류를 검출해낼 수 있도록 분류된 단어들의 의미 분류를 필요로 한다. 향후 이러한 구문/의미 교정을 위한 단어들의 의미 분류에 관한 연구가 함께 진행해야 할 것이다.

참고문헌

[1] Dong-Joo Kim, "*A Critiquing System with Tight Morphological Constraints*," MS Thesis, Hanyang University, 1997.

[2] Sung-U Mi, "*Sae Machumpop kwa Kyojong ui Sirche*," Omungak, 1994.

[3] Chul-Min Sim and Hyuk-Chul Kwon, "Implemen ntation of Korean Spelling Checker based on Collocation of Words," Journal of Computing Science and Engineering, Vol. 23, No. 7, pp. 776-785, 1996.

[4] Kil-ja So and Hyuck-chul Kwon, "A Korean Grammar Checker using Lexical Disambiguation Rule and Partial Parsing," Journal of Computing Science and Engineering, Vol. 28, No. 3, pp. 305-315, 2001.

[5] Hyun-Jin Kim, Chul-Min Sim and Hyuk-Chul Kwon, "Implementation of a Korean Grammar Checker using Partial Sentence Analysis," Proceedings of the 8th Annual Conference on Human and Cognitive Language Technology, pp. 469-475, Oct. 1996.

[6] Youngkook Hong, Jonghyeok Lee and Geunbae Lee, "A Korean Syntactic Analyzer based on the Dependency Grammar," Journal of Computing Science and Engineering, Vol. 19, No. 5, pp. 191-194, 1990.

[7] Hankyu-kyu Lim, Ung-Mo Kim, "A Spelling Correc tion System Based on Statistical Data of Spelling Errors," Journal of Korea Information Processing Society, Vol. 2, No. 6, pp. 839-846, 1995.

[8] G. E. Heidorn, K. Jensen, L. A. Miller, R. J. Byrd and M. S. Chodorow, "The EPISTLE text-critiquing system," IBM System Journal, Vol. 21, No. 3, pp. 305-326, 1982.

[9] Perterson J. L., “Computer Programs for Detecting and Correcting Spelling Errors,” CACM, Vol. 23, No. 12, pp. 676-687, 1980.

저 자 소 개

김 동 주

1996 : 한양대학교 전자계산학과 공학사.

1998 : 한양대학교 전자계산학과 공학석사.

2007 : 한양대학교 컴퓨터공학과 공학박사

현 재 : 안양대학교 컴퓨터공학과 교수

관심분야 : 맞춤법검사, 기계번역, 한국어정보처리, 의견검색, 감정인식

E-mail: djkim@anyang.ac.kr