

말뭉치를 기반으로 한 한국어 철자 교정기의 구현

이 병훈, 윤 준태, 송 만석
연세대학교 전산과학과

Korean Spelling Corrector Based on Corpus Analysis

Byeong-Hun Lee, Jun-Tae Yun, Man-Suk Song
Dept. of Computer Science, Yonsei University

요 약

대량의 말뭉치에서 나타나는 맞춤법 오류의 대부분은 타자수의 입력 실수로 인한 것이다. 맞춤법 오류의 유형은 크게 띄어 쓰기 오류, 철자 오류, 띄어 쓰기와 철자의 복합 오류의 세 가지로 나타난다. 이 중, 철자 오류를 표층 형태만으로 표준어 오류, 조사/어미 오류, 자소 대치 오류로 유형을 분류하였다. 본 논문은 300만 말뭉치에서 형태소 분석이 실패한 맞춤법 오류 어절 중에서 띄어 쓰기와 철자 오류를 분석하여, 각 오류 유형에 따른 교정 방법과 자소 대치 규칙 베이스를 이용한 교정 방법을 구현하였다. 또한 형태소 분석기를 거친 40만 어절 사전을 이용한 분석기로 기존의 형태소 분석기를 대체시켜 교정 어절을 검증하였고, 위의 사전에서 추출한 순위 결정 요소와 Heuristic 정보를 이용하여 각 후보 어절에 대한 가중치를 계산하고 가능성이 높은 교정 어절을 제시하는 시스템을 구현하였다.

I. 서론

산업 사회에서 정보화 사회로 변해 감에 따라 컴퓨터, 정보 산업이 급속히 발달하고 있다. 또한 한글을 컴퓨터로 처리하기 위한 연구도 매우 활발하게 진행되고 있고, 특히 형태소 분석[1]과 이를 이용한 철자 검색/교정에 대한 관심이 증가되고 있다. 철자 교정의 최근 연구에는 접속정보를 이용한 최장일치 해석과 양방향 접근 방법을 사용한 [2]와 기능어 역사 전 탐색을 통한 전처리 정보의 이용과 글자의 자리에 따른 n-gram 정보, 맞춤법 규칙을 사용한 [3]이 있고, 개선된 CYK 알고리즘을 이용한 형태소 해석과 n-gram 사전을 이용한 교정 방법을 제시한 [4] 등이 있다.

언어에서 나타나는 현상은 문법적인 면보다는 실생활에서 쓰이는 예가 현재의 언어 현상을 더 정확히 반영하기 때문에 용례를 바탕으로 자연 언어를 처리하는 것이 바람직하다. 이미 유럽에서는 80년대 초부터 코빌드 계획(COBUILD Project)을 수행하면서 말뭉치 언어학(Corpus Linguistics)라는 새로운 분야가 발생하였다 [5]. 또한 일본에서는 전자사전 계획

(EDR Project)을 통하여 대량의 일본 언어 자료를 컴퓨터를 이용하여 수집, 분석하고 이를 영어와 일본어 간의 기계번역 연구에 이용하고 있다 [6, 7]. 우리 나라에서는 현재 연세 대학교 한국어 사전 편찬실에서 말뭉치를 기반으로 한 새로운 사전의 편찬을 위해 대량의 말뭉치를 다양하게 구성해 놓고 있다 [8]. 본 논문은 이 중 330여 만 어절 말뭉치(이후 300만 말뭉치)를 바탕으로 실제 오류를 분석하고 통계학적인 방법으로 철자 교정 시스템을 개발하였다.

300만 말뭉치에서 나타나는 맞춤법 오류의 대부분은 타자수의 입력 실수로 인한 것이다. 맞춤법 오류의 유형은 크게 띄어 쓰기 오류, 철자 오류, 띄어 쓰기와 철자의 복합 오류의 세 가지로 나타난다. 이 중, 철자 오류를 표준 형태만으로 표준어 오류, 조사/어미 오류, 자소 대치 오류로 유형을 분류하였다.

본 논문은 300만 말뭉치에서 형태소 분석이 실패한 맞춤법 오류 어절 중에서 띄어 쓰기와 철자 오류를 분석하여, 각 오류 유형에 따른 교정 방법과 자소 대치 규칙 베이스를 이용한 교정 방법을 구현하였다. 자소 대치 규칙 베이스는 철자 오류 중에서 어미나 조사가 틀린 것을 제외하고 하나의 자소만 틀린 것을 대상으로 구축하였다. 또한, 300만 말뭉치를 형태소 분석하여 이에 성공한 어절들을 모아서 455,000 어절 사전(이후 40만 어절 사전)을 구축하였다. 교정 어절의 검증은 기존의 형태소 분석기를 사용하지 않고 40만 어절 사전을 이용하였다. 그리고, 위의 사전에서 추출한 순위 결정 요소(어절 빈도, 음절 빈도, 조사/어미 빈도)와 Heuristic 정보를 이용하여 각 후보 어절에 대한 가중치를 계산하고 가능성이 높은 교정 어절을 제시하는 시스템을 구현하였다.

II. 맞춤법 오류의 분석

기존의 오류 어절의 유형을 분류한 것들을 보면, 형태소 분석기에서 나온 정보를 이용하기 때문에 분류 기준이 자세하게 나누어져 있다 [1, 2]. 본 논문에서 사용하는 분류 기준은 단지 어절의 표준 형태로만 교정하기 때문에, 띄어 쓰기 오류, 철자 오류, 띄어쓰기와 철자 오류의 복합오류로 크게 세 가지로 분류한다. 오류의 종류는 본 연구실에서 구현한 형태소 분석기[9]를 이용하여 300만 말뭉치를 분석한 후, 분석에 실패한 어절들 중에서 미등록 어절과 외래어 어절을 제외한 24,000 여 어절을 대상으로 분류하였다 <표1>.

오류 종류	빈도	비율(%)
띄어쓰기오류	19,283	78.1
철자오류	5,103	20.7
복합오류	301	1.2
전체 갯수	24,687	100.0

<표 1> 오류의 종류와 빈도

1. 띄어 쓰기 오류

띄어 쓰기 오류는 다시 불 띄 오류와 띄 불 오류의 두 가지로 나눌 수 있다. [2, 3] 여기서 띄 불 오류는 띄어 써야할 것을 잘못 붙여 쓴 오류를 말하고, 불 띄 오류는 그 반대이다. 불 띄 오류의 경우는 앞, 뒤 어절을 현재 어절과 합치면 띄 불 오류와 같은 형태가 되기 때문에, 이

에 대한 분류는 따로 조사하지 않았다.

띄어 쓰기 오류 중에서, 용언과 용언의 결합, 용언과 체언의 결합이 많은 빈도를 나타냈다 <표 2>. 또한, 용언과 용언의 결합에서는 본용언/보조용언 결합이 가장 많은 오류의 형태를 나타냈고, 용언과 체언, 체언과 체언의 결합에서는 의존 명사와의 결합의 형태가 가장 많은 비율을 나타냈다. 아래 <표 3>에는 말뭉치로부터 구해진 202개의 보조용언에서 두 용언만 결합된 형태 중의 일부[10]와 띄어 쓰기 오류에 나타난 의존 명사 중에서 높은 상위 5개를 보였다.

띄붙 오류의 유형	용례
용언 + 용언	살펴야할, 앓아가버리고, 운용해본, 풀어가며, 피해보겠다는,
용언 + 체언(조사)	늘어복지, 구성하는등, 하는것이, 개최하는등, 이뤄질것이라며, 느낄것이, 나설생각이라고,
체언(조사) + 용언	매듭짓고, 결론내야, 고통알아야, 전제로한, 연말에가서, 원료로하는,
체언(조사) + 체언(조사)	판문점중립국, 호텔등, 유가하락으로, 회의등을, 규제등으로,

<표 2> 띄붙 오류의 유형과 용례

본용언/보조용언	용례	의존 명사	발생 빈도
게 - 되	만들게 되었다	등	947
아 - 가	돌아 가기를	것	408
야 - 할	살펴야 할	수	307
어 - 지	이루어 지지	중	117
지 - 않	쉽지 않았다	및	43

<표 3> 본용언/보조용언과 의존 명사의 예

2. 철자 오류

철자 오류는 오류 원인에 따라 다음과 같이 분류될 수 있다.

① 사용자의 철자법 지식 부족으로 인한 오류

타자수는 문서를 입력할 때, 자신이 알고 있는 맞춤법을 그대로 입력하게 된다. 따라서 이로 인한 오류는 보통 사람도 자주 틀리기 쉬운 어절들이 나타나게 된다 <표 4>.

② 입력 문서의 잘못으로 인한 오류 (표준어 규정에 어긋나는 오류)

말뭉치의 입력 문서에는 맞춤법 개정 이전의 문서가 많이 있고, 또한 소설에서 의도적인 사투리의 사용으로 인한 오류가 있다 <표 4>.

종류	용례
지식부족으로인한 오류	역활, 원지, 아뭇든, 슛컷, 안밖으로
맞춤법에 어긋나는 어절	~읍니다, ~요, 하니바람,
사투리 사용 어절	종재잉, 영낙없당께, 아그덜은, 있었어야재잉,

<표 4> ①과 ②의 유형에 나타난 오류의 예

③ 입력자의 실수로 인한 오류

타자수의 입력 실수로 인한 오류는 철자 오류의 90 % 이상을 차지 하고 있다. 이를 다시 여러 요인 별로 분류해 보면 아래 <표 5>와 같다.

오타 요인	용 예
인접자판을 잘못친 오타	속입수, 염탐꾼이자, 전통을, 분면히,
불필요한 음소의 첨가	아닐할, 앉으으며, 특수 11성에, 화확물질,
음소의 삭제	필어적으로, 트유한, 행동요어로, 최그에,
조사오류	행위를, 세계를, 세포에세, 정으로보터,
어미오류	있느지를, 오른다고, 있었고, 하느데에,
Shift Key 오류	에리함, 필신, 폐지, 의례로서,
복모음오류	필썬, 원시, 으무사항, 학출을,

<표 5> 입력자의 오타로 인한 오류의 요인과 예

위의 분류는 표층 형태로만 판단하면 오류 원인이 달라도 같은 오류 유형을 가진다. 그러므로, 원인에 따른 분류보다 오류 유형에 따른 분류가 교정 시스템을 구현하는데 효과적이기 때문에, 본 논문에서는 철자 오류를 표준어 오류, 조사/어미 오류, 자소 대치 오류의 세가지로 분류하였다.

표준어 오류에는 맞춤법 개정으로 과거에는 맞는 것으로 쓰이던 것이 현재에는 틀린 것으로 분류되는 어절들과 사투리, 입력자의 지식 부족이나 습관적인 오타로 높은 빈도를 나타내는 어절들을 포함한다.

조사/어미 오류는 보통의 어절보다는 그 어절 형태가 특이하고 오류 형태도 일정하다. 조사나 어미는 어절의 뒷부분에 위치하므로 찾기 쉽고, 이 오류를 따로 처리하게 되면 모든 글자에 대하여 자소 대치 교정을 하지 않아도 되므로 시스템의 성능을 향상시킬 수 있다.

자소 대치 오류는 한 어절내의 자소가 다른 자소로 대치되거나 삽입, 삭제된 오류를 말한다. 단, 대상이 되는 오류 어절은 조사나 어미 오류를 제외한 부분에서 한 글자에서 자소 대치 오류가 단 하나만 나타나는 것으로 한정하였다.

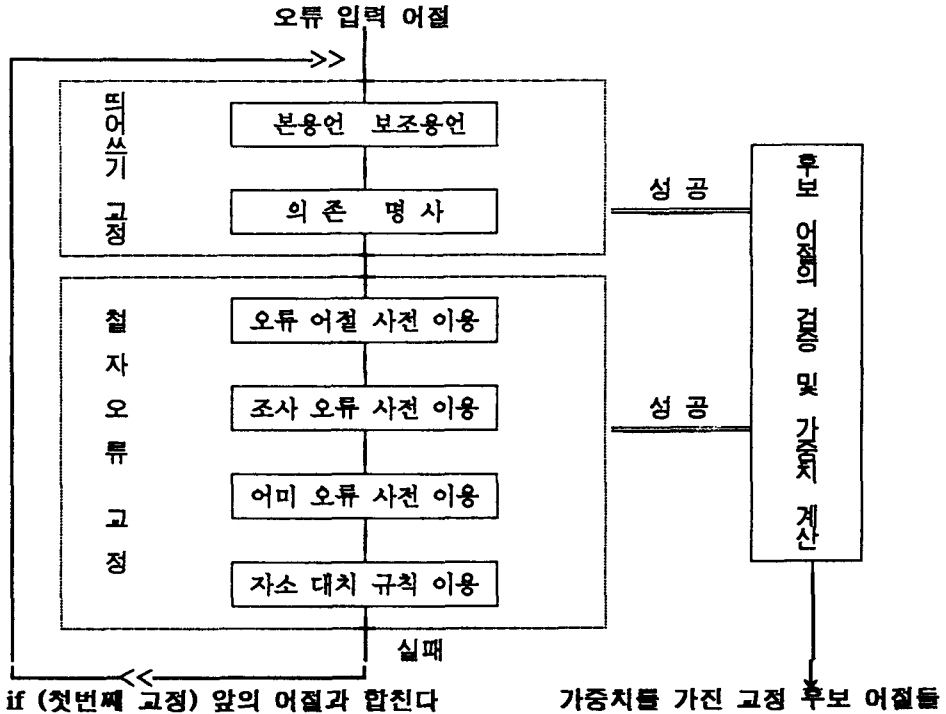
Ⅲ. 시스템의 구성

시스템은 크게 띄어 쓰기 교정과 철자 교정의 두 단계로 구성되어 있다 <그림 1>. 각 세부 단계에서 교정된 후보 어절들은 40만 어절 사전을 통하여 검증되고, 그 가중치가 일정한 수준을 만족하면 다음 어절을 교정한다.

띄어 쓰기 교정은 본용언/보조용언과 의존 명사를 찾아 교정을 하고, 이것이 실패하면 앞에서부터 뒤로 한 글자씩 두 부분으로 나누어 교정한다. 만약 두 부분 중에서 하나만 성공하게 되면, 이를 복합 오류로 간주하고 철자 교정으로 보내게 된다.

철자 오류 교정 과정은 입력 어절을 조사나 어미를 찾아 두 부분으로 나누어 조사나 어미 부분을 제외한 나머지에서 교정을 한다. 만약, 조사나 어미를 찾지 못하면 이를 조사/어미 오류 사전을 통하여 교정을 하고, 이 단계가 실패하면 철자 오류 교정의 마지막 교정 단

계인 자소 대치로 전체 어절에 대하여 교정을 한다.



<그림 1> 시스템의 구성 - 한 어절의 교정

1. 사전의 구성

시스템이 사용하는 사전은 크게 4개로 구성되어 있다. 각 사전마다 이를 관리하는 관리 시스템을 가지고 있어서 사전의 수정과 보강을 책임지고 있다. 특히, 어절 오류 사전은 교정 작업을 수행하면서 오류 어절의 발생 빈도에 따라서 새로운 어절을 계속 보강해야 한다.

1.1. 40만 어절 사전

이 사전은 300만 말뭉치를 형태소 분석기로 처리하여 성공한 어절과 빈도를 가지고 있다. 이 사전을 이용하여 교정된 후보 어절을 형태소 분석기를 대신하여 검증을 하게 된다. 아래 <표 6>은 어절 사전에서 빈도가 1,000 이상인 어절 중의 일부이다.

1.2. 어절 오류 사전

철자 오류의 분류에서 나타난 표준어 오류와 자주 틀리는 어절에 대하여 오류 어절과 교정 어절을 일대일로 저장한 사전이다.

1.3. 조사 오류 사전, 어미 오류 사전

조사나 어미의 형태는 보통의 낱말과는 다른 특성이 있다. 예를 들면, ‘를’이나 ‘으’ 등과 같이 첫 음절에 나오지 않는 글자들과 같은 형태론적 특성이 나타난다. 또한, 조사나 어미는 특성상 어절의 뒷부분에 위치하므로 후방향에서 탐색을 하게 된다. 따라서, 이의 오류는 어절의 뒷부분에서만 발생하고 일정한 패턴을 보이므로 사전을 구성하여 처리하는 것이 바람직 하다 <표 7>.

표제어	빈도	표제어	빈도	표제어	빈도
가장	3,091	대한	7,175	특히	2,529
가지고	1,589	모든	3,593	하고	4,492
같은	6,283	수	20,276	하는	6,711
것은	8,508	어떤	3,274	한	11,589
그	23,217	위해	3,526	한다	7,091
것이다	12,475	있다	20,256	할	6,642
나는	3,812	지난	3,900	함께	2,888
다른	4,027	큰	2,955	했다	4,014

<표 6> 40만 어절 사전 중의 빈도 1,000 이상인 어절

조사의 오류 형태	조사의 교정 후보	어미의 오류 형태	어미의 교정 후보
를, 르	를	두룩	도룩
에거, 에세	에게	음벼	으며
으틏서	으르씨	엳으며	엳으며
으, 으이, 니, 디, 윗	의	엇다	었다
느, 느, 느	는	어도	어도
와, 와, 오아	와	헛다	했다
엳, 엳, 엳	에	는데	는데
으릅터	으로부터	는가	는가

<표 7> 조사 오류 사전과 어미 오류 사전의 일부

1.4. 자소 대치 규칙 테이블

한 글자 내에서 한 자소가 다른 자소로 잘못 입력되거나 삭제, 첨가된 오류를 자소 대치 오류로 보고, 약 2,500 어절로부터 자소 대치 규칙을 얻어 냈다. 여기서, 한 어절 내의 자소 대치 오류는 조사나 어미의 오류를 제외한 나머지 글자에서 한번 나타나는 어절을 대상으로 하였다. 자소 대치 교정은 어절의 앞에서부터 교정을 하게 되기 때문에 어절의 자리 별로 대치 규칙을 구성해야 한다. 그러나, 한 어절에서 조사나 어미가 차지하는 글자를 제외하면 평균 어절길이는 2 ~ 3 글자이므로 4자리 이상에서의 대치는 1.3에서 조사/어미 오류로 처리하는 것이 더 효과적이다.

규칙 테이블은 각 어절에 대하여 초성, 중성, 종성 별로 따로 구성하였고, 그리고 첫 글자와 나머지 글자로 나누어 규칙을 구성하여 전체 규칙의 수는 6개로 이루어 졌다. 각 사전은 모든 자소마다 많게는 8 ~ 9개 적게는 0 ~ 1개의 대치 자소를 가지고 있다. 아직 적은 수의 어절을 분석했기 때문에 빈도가 1, 2인 대치 자소가 있고, 대치 자소가 없는 것도 있다. 이는 앞으로 더 많은 어절을 분석하여 보강해야 할 것이다. 분석 결과 자소 대치 오류는 앞, 뒤 글자의 자소나 현재 글자의 자소에 따라 대치 오류가 발생하고, 자판 배열의 영향을 크게 받지 않았다. 아래 <표 8>은 중성의 첫 글자에 대한 대치 규칙을 보였다.

1.5. 조사/어미 대치 사전

자소 대치 규칙으로 어절을 교정하기 전에 조사나 어미 부분은 사전을 이용하여 미리 앞 부분과 갈라 놓아 조사나 어미에 대하여 자소 대치 규칙을 적용하지 않게 하였다. 이렇게 하면 1.3. 에서 교정하지 않는 오류가 있는데, 이는 뒷부분이 비록 틀린 것이지만 조사나 어

⑤ 철자 오류 교정 방법 : 오류 어절 사전을 이용한 교정이 가장 확실하게 교정되는 것이므로 철자 오류 교정 중 가장 높은 가중치를 가지고, 조사와 어미 오류 어절 사전을 이용한 교정은 그 다음 높은 가중치를 가지게 된다. 마지막으로 자소 대치 규칙 테이블을 이용한 교정인데, 여기에서 초성보다는 중성이나 종성이 높은 가중치를 갖게 되고, 규칙의 순위에 따라 가중치가 다르게 주어져야 한다.

2.2. 가중치 결정 방법

가중치 결정의 기본적인 계산 방법은 두 개 이상의 값에 대하여 가장 작은 값을 취하는 min-max composition 방법이다. 이 방법은 Zadeh에 의해 처음 제안되었고 실수의 곱보다 빠르고 실효성이 높은 것으로 실험적 결과로 입증되었다. 띄어 쓰기 교정에서 한 어절을 둘로 나누었을 때, 각 어절의 가중치를 계산하면 그 중 작은 값을 전체 어절의 가중치로 결정한다.

IV. 실험 및 결과

시스템은 Solbourne Workstation에서 C로 구현하였다. 철자 오류 50개와 띄어 쓰기 오류 45개, 복합 오류 5개로 총 100개의 어절을 실험 대상으로 하였다. 아래에 입력 오류 어절과 그 결과를 일부 보였다.

<u>오류 어절</u>	<u>띄어 쓰기 교정</u>	<u>철자 교정</u>	<u>복합오류 교정</u>
역활을	역 활을 10	역활을 18	
듬듬히		듬듬이 14	
하루에	하루 에 13	하루에 15	하 루에 8, 하 우에 5, 하 구에 5 하 루이 6
해야할	해야 할 15		
잡자기	잡자 기 10	잡자기 14 잡자기 12	잡 자기 7, 남 자기 7, 삼 자기 7 밥 자기 10, 집 자기 4, 접 자기 8 잡 자기 6, 잠 자기 9, 잔 자기 9 자 자기 4
폐품	폐 품 7	제품 18, 폐품 12	
싸람과		사람과 14	싸 암과 7, 싸 남과 8 싸 감과 5, 싸 립과 6
할것이라는	할 것이라는 17		
경쟁등	경쟁 등 15		
어려한	어려 한 10	이러한 15 어떠한 14	어 거한 10
없었을것이다			<u>없었을 것이다</u> 7
점말	점 말 12	전말 12, 정말 17 점만 12, 점막 12	
사회현상을			<u>사회 현상을</u> 5
전사회를		전사회를 13	<u>전 사회를</u> 17

V. 결론

본 논문은 300만 말뭉치에서 실제로 나타난 맞춤법 오류를 분류하고 그 교정 방법을 제시, 구현하였다. 맞춤법 오류는 띄어 쓰기 오류, 철자 오류, 복합 오류로 분류하고, 특히 철자 오류는 표준어 오류, 조사/어미 오류, 자소 대치 오류로 분류하였다. 표준어 오류와 조사/어미 오류에 대하여 각각 오류 어절 사전과 조사/어미 오류 사전을 구성하여 처리하고, 자소 대치 오류는 실제 오류의 용례에서 추출한 규칙을 이용하여 교정하게 된다. 300만 말뭉치를 분석하여 얻어낸 40만 어절 사전을 이용한 검사기로 교정 어절을 검증하여 형태소 분석기를 이용한 교정보다는 빠른 시간 안에 결과를 얻을 수 있도록 하였다.

앞으로의 과제는 시스템의 실제 사용을 위하여 속도를 향상시켜야 한다. 이를 위하여서는 쓸데없는 후보 어절의 생성/검증을 막아야 할 것이다. 더 많은 오류를 분석하여, 어절 내에서 자릿수 별로 자소 대치 규칙을 구축하고 더 많은 정보를 첨가하여 규칙 테이블을 확대, 보강하여야 할 것이다. 또한 일괄 처리로 자동 교정을 하기 위해서는 교정기에서 어느 정도의 문맥 정보나 속어 정보를 이용해야 한다.

VI. 참고 문헌

- [1] 박종만, "철자 검색기에서 틀린 어절의 처리", 우리말 정보화 잔치 논문집, 1991
- [2] 조영환, "한글 맞춤법 교정기의 설계 및 구현", 한국과학기술원 석사학위 논문, 1991
- [3] 이영식 외, "한국어 철자 오류 교정 시스템", 정보과학회 인공지능 연구회 춘계 학술발표회 논문집, pp25 - pp38, 1993
- [4] 이원일 외, "Binary N-gram과 형태소 분석기를 이용한 한국어 철자 교정기", 정보과학회 봄 학술발표논문집, 제 20권, 제 1호, pp813 - pp816, 1993
- [5] 남기심 외, "사전편찬학 연구", 제 2집, pp162 - pp193, 탑출판사, 1988
- [6] EDR, An Overview of the EDR Electronic Dictionary, TR-024, Apl., 1990
- [7] EDR, Proceedings of International Workshop on Electronic Dictionary, TR-031, Nov., 1990
- [8] 정찬섭 외, "사전편찬학 연구", 제 3집, pp70 - pp76, 탑출판사, 1990
- [9] 박영환, "말뭉치에 기반한 형태소 분석기 및 철자 검색기의 구현", 연세대학교 석사학위논문, 1992
- [10] 윤준태, "다양한 통사현상을 통한 한국어 구문 분석기 연구", 연세대학교 석사학위논문, 1993