한국어 철자 검사기의 교정기법 개선

김광영[U] 남현숙 박수호 박진희 권혁철
부 산 대 학 교 전 자 계 산 학 과
{kykim, hsname, suhopark, jinhee}@borame.cs.pusan.ac.kr

Improving of the Correction Methods for a Korean Spell/ Grammar Checker

Kwang-Young kim, Hyeon-Sook Nam, Su-Ho Park, Jin-Hee Park
Dept. of Computer Science, Pusan National University

요 약

본 논문은 부산대 철자 검사/교정기의 기존 성능을 보완하고 기능을 추가하는데 중점을 두었다. 웹 문서, 신문 등을 통해서 사용자들이 자주 틀리는 오류 단어에 대해서 오류 유형을 분류했다 이 결과를 철자 검사 및 교정 시스템에 적용하여 교정기법 개선을 통하여 띄어쓰기 교정 기능을 향상 시켰다. 이렇게 새로 구현한 시스템과 이전 시스템의 성능을 실험을 통해 비교 분석하였다.

본 연구를 진행하면서 발견한 문제점과 한계를 이후 더 발전 해야할 과제로 고찰하고 결론을 맺는다.

1. 서론

최근 컴퓨터 사용이 증가하면서 다양한 형태의 문서가 작성되고 있다. 신문, 서류, 사설, 교과서 등 공식적인 문서 뿐만 아니라 편지, 게시판에 있는 문서, 개인 홈페이지 등 문서의 형태나 쓰임이 다양해졌다. 이렇게 문서의 형식이 여러 가지 형태를 띠면서 오류의 패턴도 각양각색이다. 즉, 문서 작성자가 기본적인 문법 지식을 가지고 쓰는 신문이나 공식적인 문서와 초등학교 학생이 쓴 작문에서 발생하는 오류는 다르다. 이러한 다양한 문서에서 자주 발생하는 오류의 빈도를 측정하여 그 결과를 철자 검사 및 교정 시스템에 적용하여 그 성능을 향상시키고자 한다.

본 논문에서는 첫째 오용어에 대해서 이전에는 오용어가 있으면 띄어쓴 어절이 틀린 어절로 인식되어 대치어 생성에 실패하였다. 현재에는 띄어쓰기를 해본 후 띄어쓴 어절이 오용어일 때는 틀린 어절로 인식하지 않는다. 오용어에 대한 표준어를 대치시켜 띄어쓰기 대치이를 생성 처리하였다.

둘째 이전 시스템에서는 음소대치를 바르게 처리한 때에도 띄어써야 맞는 어절이 대해서 틀린 어절로 인식되어 대치어 생성에 실패했다. 현재는 음소대치를 해본 후 띄어쓰기 검사를 처리함으로써 띄어쓰기 대치어를 생성 처리하였다. 세 째는 일반적 규칙에 의한 띄어쓰기 교정의 기능을 추가하였다

2. 최근 문서에서 발생하는 오류의 빈도

[표 1]은 부산대학교 철자 검사 및 교정 시스템을 사용하여 오류를 분석한 결과이다. 웹 철자 검사기는 부산대학교 웹 철자 검사 및 교정 시스템을 사용자들이 직접 입력한 6,555건의 자료를 분석했다. 오류 유형은 시스템에서 오류로 판단한 어절에 대해 제시해 주는 오류 표시

를 기준으로 하였다.

오류 유형 중 '어미/조사'는 어미나 조사를 잘못 쓴 오류를 가리킨다.

[표1] 유형별 오류의 빈도

오류 유형	웹철자 검사기	신문	중학생 작문	전체
맞춤법	264	25	17	306
표준어	270	2	14	286
오용어	249	4	9	262
어미/조사	347	10	26	383
순화 용어	155	2	5	162
띄어쓰기	822	111	129	1062
의미/문체	318	32	17	367
외래어 표기	43	2	3	48
조어법	2	1	1	4
발음 유사	6	0	2	8
기타				251

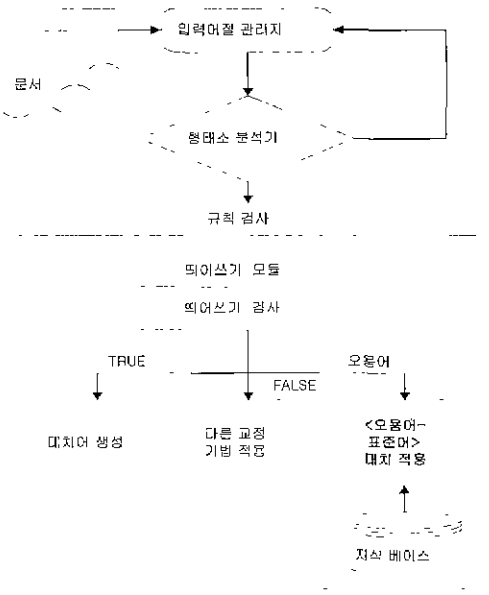
(단위:개)

빈도 계산에 사용한 전체 어절 수는 170,156개이며 전체 오류는 3,139개이다. 시스템에서 정의한 오류 유형으로 검사된 어절이 2,888였고 나머지 오류 어절 251개는 일반 오류인 철자법 오류로 처리되었다. 위의 [표 1]에서 대치어 생성 함수의 기능 별로 오류를 분석했을 때 맞춤법, 오용어, 띄어쓰기 오류 등 띄어쓰기 대치함수에서 대치어를 생성하는 어절이 약 75%, 맞춤법, 표준어, 순화 용어, 외래어 표기 등 음절 혹은 음소 대치어절이 15.6% 였다. 이 오류 빈도의 분석 결과에서 교정기의 성능은 띄어쓰기 대치 함수와 음절 혹은 음소 대치 함수의 성능에 따라 결정함을 알 수 있다.

3. 띄어쓰기 교정 루틴

[그림 1]은 부산대 철자 검사기의 띄어 쓰기 교정 루틴을 나타낸다. 현재의 시스템에 추가된 부분은 오용어 표준 대치어 부분이다.

띄어쓰기 검사에 따라 3가지 방법으로 처리가 된다. 띄어쓰기 검사를 수행하는 중에 오용어일 때는 지식 베이스에서 오용어에 대한 표준어를 대치하고 대치어를 생성한다.



[그림 1] 띄어쓰기 흐름도

· 입력어절 관리자: 문장을 어절 단위로 Token 처리
· 형태소 분석기: 단어의 형태소 분석 처리
· 일반 띄어쓰기 모듈: 여러 가지 띄어쓰기 교정 방법을 사용하여 각 단어의 띄어쓰기 검사
· 검사결과에 따라 3가지 처리
  a. TRUE: 대치어를 생성
  b. FALSE: 음소대치, 붙여 보기 등의 다른 교정 기법 적용
  c. 오용어: 오용어를 지식 베이스 리스트에서 표준어를 대치

4. 이전 시스템과 비교

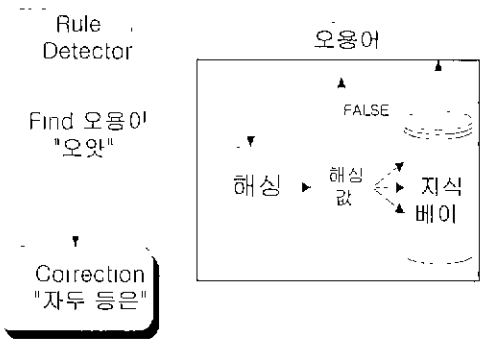
이전 시스템은 띄어쓰기 대치어를 생성할 때 오용어가 있으면 띄어쓴 어절이 틀린 어절로 인식되어 대치어 생성에 실패했다. 현재 시스템은 띄어쓰기를 해본 후 띄어쓴 어절이 오용어일 때는 해 당하는 표준어로 대체하고 난 뒤에 띄어쓰기 대치어를 생성해 낼 수 있다.

오용어에 대해서 형태소 분석기가 형태소 분석에 실패를 하며 지식 베이스의 규칙 검사기에서 오용어를 교정한다

[표 2] 띄어쓰기 대치어 생성할 때 오용어의 처리의 예

틀린 어절	이전 시스템	현재 시스템
오얏등은	대치어 생성 실패	오얏 등은
학교용팩스밀리	대치어 생성 실패	학교용 팩시밀리
친구딸래미가	대치어 생성 실패	친구 딸이

이 지식 베이스는 오용어에 대해서 현재 44,622개의 오용어 쌍을 가지고 있다. 현재의 시스템의 흐름은 다음과 같다



[그림 2] 오용어 교정 처리 예

1. 틀린단어 "오얏등은"에 대해서 교정 처리
2. Find "오얏"을 오용어
3. 오용어 교정
   a. Perfect hashing으로 Knowledge Base에서 오용어 찾기
   b. 오용어 "오얏"을 "자두"로 대치
   c. "자두 등은"으로 대치어를 제시

5. 일반 규칙에 기반한 띄어쓰기 교정 성능 향상

5.1 음소대치 후 띄어쓰기

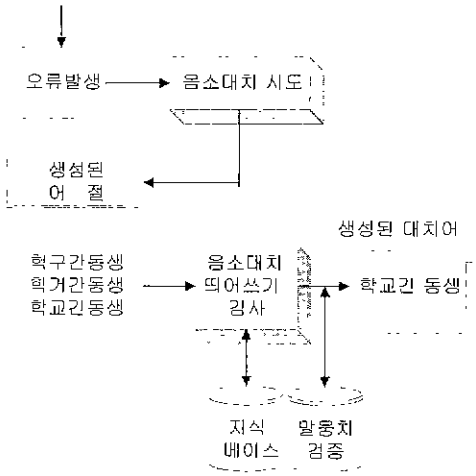
이전 시스템에서는 음소대치를 바르게 처리한다해도 띄어써야 바른 어절에 대해서는 틀린 어절로 인식되어 대치어 생성에 실패했다. 현재 시스템은 음소대치를 해본 후 띄어쓰기 검사를 처리함으로써 띄어쓰기 대치어를 생성해 낼 수 있다. 또한 대치어를 말뭉치로 검증을 함으로써 부적절한 대치어를 걸러내었다.

[표 2] 음소대치 후 띄어쓰기 처리의 예

틀린 어절	이전 시스템	현재 시스템
런구소직원	대치어 생성 실패	연구소 직원
학고간동생	대치어 생성 실패	학교간 동생
팡리소직원	대치어 생성 실패	판리소 직원

현재의 시스템의 흐름은 [그림 3]와 같다.

1 틀린 단어 "학고간동생"에 대해서 음소대치 시도
2 생성된 어절 중에서 말뭉치 사전의 빈도에 따라 부적절한 대치어를 걸러낸다.[그림 3]
3 음소대치 띄어쓰기 검사 루틴은 지식 베이스에서 띄어쓰기 및 검증을 통하여 "학교간 동생"으로 처리하고, 말뭉치를 통해 다시 검증 처리하여 대치어를 제시한다
예) **학고간동생**

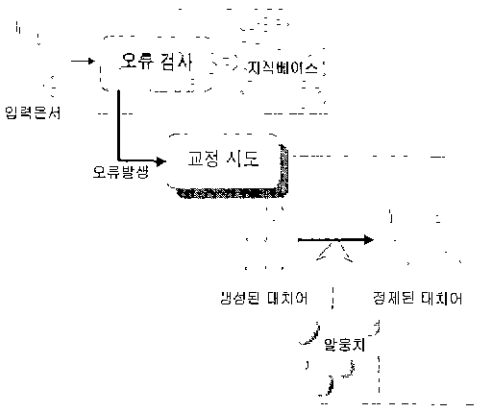


[그림 3] 음소대치 후 띄어쓰기

음소의 대치를 통한 대치어 생성 시도는 초성, 중성, 종성에 대해 각각 대치해보기 때문에 대치어가 많이 생성될 수 있다.

생성된 대치어는 실제 문맥에 맞지 않는 부적절한 대치

어가 있을 수 있다. 이러한 부적절한 대치어는 철자 검사 및 교정 시스템의 신뢰도를 떨어뜨리는 요인이 된다. 따라서 생성된 대치어를 사용자에게 그대로 제시하지 않고 말뭉치 사전의 빈도에 따라 부적절한 대치어를 걸러낸다[7].



[그림 3] 말뭉치를 이용한 대치어 검증작업

[그림3]은 말뭉치를 이용해서 부적절한 대치어를 제거하는 과정이다.

5.2 한 음절 명사 띄어쓰기 성능 향상

앞이나 뒤에서 띄어써야 하는 한 음절 명사에서 띄어쓴 후 나머지 띄어쓴 어절이 또 띄어씨야 하는 어절일 때 이전 시스템에서는 대치어 생성에 실패를 하였다. 현재의 시스템은 다시 한번 띄어써서 대치어를 생성했다.

[표3]한 음절 명사 띄어쓰기 성능 향상의 예

틀린 어절	이전 시스템	현재 시스템
그두사람	대치어 생성 실패	그 두 사람
말한마리	대치어 생성 실패	말 한 마리
권유안한	대치어 생성 실패	권유 안 한

5.3 교정 기법과 적용 가중치

[표 4]는 띄어 쓰기에 사용한 교정 기법에 따른 가중치를 나타낸다.

[표 4] 교정 기법에 따른 가중치

교정 기법		가중치
heuristics 에 의한 띄어쓰기	교정 규칙에 기반한 교정	1
	다수 음절 띄우기	2
	한 음절 띄우기	3
	띄어쓰기 후보 위치에 의한 띄어쓰기	4
음절 정보에 기반한 띄어 보기		5
최장 일치법		6
형태소 분석 정보에 기반한 띄어 쓰기		7

6. 실험

본 논문에서는 3가지 실험을 하여, 한국어 철자 검사기 시스템의 성능을 테스트하였다

첫 번째는 실험에서는 올바르게 교정을 하는지를 검사했다. 두 번째는 첫 번째 실험에서 교정하지 못하는 단어에 대해서 오류 유형을 분류하였다 세 번째 실험에서는 이전 시스템과 현재의 시스템의 성능을 비교하였다.

자료는 연합뉴스, 메일신문사, 중앙일보 기사의 원본을 사용하였다

(1) 철자 검사

[표 5] 철자검사 및 교정 후의 결과

전체 단어	170,150	
올바른 단어	163,451	
틀린 단어	Correction	6,350
	Non-correction	349

오류 빈도가 높은 이유는 기사의 원본을 사용했기 때문이다. [표 5]를 보면 교정률은 약 94.7%로 계산되었다.

(2) 오류 유형 분류

틀린 단어에 오류 유형을 분류하였다.
[표6]는 분류된 오류 유형을 보여 주고 있다.

[표 6] 틀린 단어의 띄어쓰기 오류 유형

오류유형	전체 단어	대치어 생성	대치어 생성 실패
띄어쓰기오류	3101	3101	10
붙여쓰기오류	2	1	1
음소대치후 띄어쓰기	11	11	0
수사 띄어쓰기	942	940	2

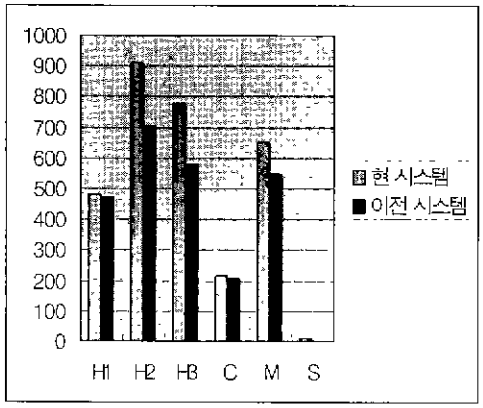
[표 6]에서 나타난 것과 같이 띄어쓰기 오류가 약 60.5%가 된다.

(3) 교정기 성능

본 논문에서는 띄어쓰기 오류 단어 3,112개을 사용하여 이전 시스템과 비교하였다

☆ H1 - Heuristics을 사용한 다수 음절 띄어쓰기
☆ H2 - Heuristics을 사용한 한 음절 띄어쓰기
☆ H3 - Heuristics을 사용한 후보 위치에 따른 띄어쓰기
☆ C - 음절 정보에 기반한 띄어쓰기
☆ S - 음소대치 후 띄어쓰기
☆ M - 형태소 분석 정보에 기반한 띄어 쓰기

Heuristics에 의한 띄어쓰기가 약 74%가 되었다. 다수 음절, 한 음절, 후보 위치에 따른 띄어쓰기의 Heuristics의 확장을 통한 교정 규칙 베이스를 강화하였다.
이전 시스템의 띄어쓰기 성능은 80.6%이고, 현재의 확장된 시스템의 띄어쓰기 성능은 92.9%이다. 그 성능이 12.3%가 향상되었다.



[그림 5] 현 시스템과 이전 시스템 비교

7. 결론

본 논문에서는 철자 검사/교정기의 기존 성능을 보완하고 기능을 추가하는 데에 중점을 두었다. 철자 검사/교정기의 기능을 개선하고 향상시키는 것은 구문 분석에 기반한 한국어 처리의 기반이 되는 기초 기술이다. 따라서 정확한 구문 분석을 위해 기본단위인 어절의 올바른 형태소 분석과 교정이 필요하다.

본 논문의 목적은 이전 시스템이 띄어쓰기 대치어를 생성할 때, 오용어가 있으면 띄어쓴 어절이 틀린 어절로 인식되어 대치어 생성에 실패했기 때문에 이를 보완하려는 것이다. 현재 시스템은 띄어쓰기를 해본 후 띄어쓴 어절이 오용어일 때는 틀린 어절로 인식하지 않고 오용어에 대한 표준어를 대치시키고 나서 띄어쓰기 대치어를 생성해 낼 수 있다.
또한 이전 시스템에서는 음소대치를 바르게 처리한 후에도 띄어써야 맞는 어절에 대해서는 틀린 어절로 인식되어 대치어 생성에 실패했다. 현재 시스템은 음소대치를 해본 후 띄어쓰기 검사를 처리함으로써 띄어쓰기 대치어를 생성해 낼 수 있었다.

향후 연구과제로는 철자 검사 및 분석 작업 과정에서 오류 분석 정보를 제공하는 기능이 더 보강되어야 한다. 이러한 정확한 오류 분석 정보에 기반하여 띄어쓰기 교정 기법을 강화시켜야 한다.

참고 문헌

[1] 심철민 "어절 간 연관 관계와 오류 유형 추정 규칙에 기반한 한국어 철자 교정기",부산대학교 전자계산학과 석사학위 논문,1995

[2] 최성필 "오류분석정보와 복합명사의 의미처리규칙 및 말뭉치를 이용한 철자 교정기의 성능 개선", 부산대학교 전자계산학과 석사학위 논문, 1998

[3] 김수남 "운지 거리와 빈도를 이용한 음소대치의 성능 향상 및 속도 개선",부산대학교 전자계산학과 석사학위 논문, 2000

[4] 채영숙,"연어 규칙에 기반한 한국어 문서 교정 시스템의 구현", 1998

[5] 이영식, 채영숙. 권혁철,"철자 검사기에 있어서의 교정기", 92가을정보과학회, 1992
pp.1001-1004

[6] 심철민, 권혁철, "연어 정보에 기반한 한국어 철자 검사와 교정기의 구현", 정보과학회 논문지, 1996, Vol123 No7. pp.776-785

[7] 이영식, "사전 근사탐색과 heuristics를 이용한 한국어 철자 오류 교정 시스템 구현",1994, 석사학위논문