효율적인 문서처리를 위한 띄어쓰기 교정 기법 개선*

강미영° 권혁철
부산대학교 전자계산학과 인공지능연구실
{kmyoung°, hckwon}@pusan.ac.kr

Improving Word Spacing Correction Methods for Efficient Text Processing

Miyoung Kang° Hyukchul Kwon
AI Lab. Dept. of Computer Science, Pusan University

요 약

한국어 문서에서 가장 많이 나타나는 띄어쓰기 오류는 의미적이고 통사적인 중의성이나 오류를 야기한다. 이 논문은 부산대 인공지능 연구실에서 개발한 부분 문장 분석을 기반으로 하는 한국어 철자 및 문법 검사기(2.2)에 구현되어 있는 어절 내 한 번 띄어쓰기 오류 교정 기법 및 어절 간 띄어쓰기 오류 교정 기법을 확장하고 개선하며 어절 내 여러 번 띄어쓰기 기법을 개발함을 목표로 한다.

1. 서 론

한국어 문서에서 띄어쓰기 오류가 있는 어절[1]은 형태소 분석이 안 되며 의미적이고 통사적인 오류나 중의성이 생긴다. 이러한 띄어쓰기 오류는 한국어 문서에서 가장 많이 나타난다. 따라서 띄어쓰기 오류를 정확하고 효율적으로 교정하는 기법을 개발하는 것은 시스템 전체 성능 향상을 가져온다. 효율적인 띄어쓰기 교정 기법 개발은 많은 데이터를 분석하고 다양한 오류를 언어학적이고 화용론적으로 분석을 함으로써만 가능하다. 2장에서 선행 연구들에 대해 알아 볼 것이며, 3장에서는 본 연구실에서 개발된 다양한 띄어쓰기 기법을 살펴 보고 개선할 것이다. 4장에서는 여러 번 띄어쓰기 교정을 하기 위해 한국어의 특성에 기반한 단서(clue)를 구축하고 여러 번 띄어쓰기 기법을 제안한다. 5장에서 실험 및 결과를 분석하고 결론을 내린다.

2. 관련 연구

한국어의 띄어쓰기 오류는 띄어쓰기의 복잡성으로 인하여 가장 많이 나타난다. 이러한 띄어쓰기 교정을 개선하기 위해 개발된 많은 연구들 중 심광섭은 말뭉치에서 추출한 음절 bigram 빈도를 이용하여 음절 간 띄어쓰기 확률을 계산하는 방법을 제안하고, 이 상호 정보를 이용한 방법을 자동 띄어쓰기에 활용하여 94.6%의 정확도를 얻었다.[3] 강승식은 어절 블록 양방향 알고리즘을 개발하여 띄어쓰기 오류를 자동 처리하는데 97.3%의 정확도를 보인다. 이 알고리즘은 조사 및 어미의 음절 특성을 이용하여 띄어 쓸 확률이 매우 높다고 판단되는 어절 블록을 설정한 후에 어절 블록 내에서 형태소 분석기를 이용하여 어절을 인식하는 방법이다[1]. 부산대학교 인공지능 연구실에서는 확률과 규칙에 기반한 한국어 문법 검사기가 설계되어, 규칙이 먼저 적용되고 만약 동일한 값을 가진 두 가지 요소가 있다면 말뭉치 연구를 통해 미리 얻어진 통계적인 가치에 따라 통계적인 가중치를 부여함으로써 선택하는 방법을 적용하는 띄어쓰기 시스템이 구현되었으며[5], 최장 일치 기법, 형태소 분석 결과를 이용한 가중치 적용 기법, 휴리스틱을 이용한 기법들을 어절 내 한 번 띄어쓰기 교정 시스템에 구현하였다.[4], [7] 끝으로, 문맥을 고려해야만 분석이 가능한 한국어 띄어쓰기 오류들은 문법검사기에 구현된 부분 문장 분석 기법을 개선하고 보완하여 띄어쓰기에 응용했다.[4], [6]

3. 띄어쓰기 오류 유형 및 띄어쓰기 기법

이 논문은 크게 어절 내 띄어쓰기 오류와 어절 간 띄어쓰기로 나눈다. 어절 내 띄어쓰기 오류는 띄어쓰기를 하지 않으면 형태소 분석이 되지 않는 오류이다.[표 1]

[표 1] 어절 내 띄어쓰기 오류

띄어쓰기 오류 어절	대치어
① 몇달	몇 달
② 갈예정	갈 예정
③ 공부할때가많다	공부할 때가 많다
④ 교사및학생들의모임	교사 및 학생들의 모임

[표 2] 어절 간 띄어쓰기 오류

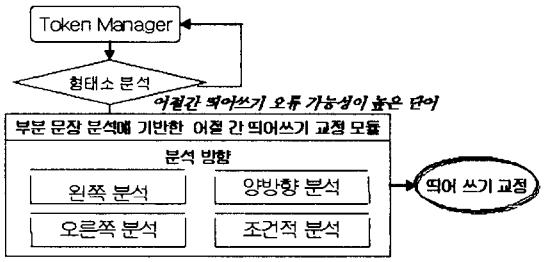
띄어쓰기 오류 어절	대치어
⑤ 나는**보다** *나은*	나는 보다 나은
⑥ *발령*이 **난후**	발령이 난 후
⑦ *폭탄* **수발**이 *터지고*	폭탄 수 발이 터지고
⑧ *자손* **만 대**	자손 만대

* 이 논문은 국가지정연구실사업(과제명: 언어 중심의 지능적 정보처리를 위한 단계적 우리말 분석기술의 개발 (M10203000028-02J0000-01510))의 지원을 받아 이루어진 것임.

1) 이 논문에서 '어절'은 '단어'와 구분되는 개념으로 사용되었다. 띄어쓰기의 한 낱덩이를 지칭한다.

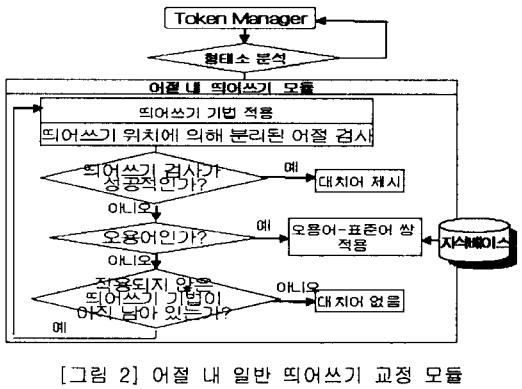
어절 간 띄어쓰기 오류는 문장 분석을 통해서만 띄어쓰기 오류 여부를 정의할 수 있는 오류를 말하며,[표 2] 띄어 써야 하지만 붙여 쓴 오류(⑤, ⑥, ⑦)와 붙여 써야 하지만 띄어 쓴 오류(⑧)로, 다시, 구분될 수 있다.

현 시스템은 어절 단위로 검사를 하여 형태소 분석이 안 되는 어절이 발견되면 띄어쓰기 일반 처리 모듈로 넘기거나, 여러 어절 간 오류 가능성 어절이 발견된 경우 형태소 분석 정보 및 어절 정보를 저장하여 어절 간 오류 처리 모듈로 넘긴다. 어절 간 오류 처리 모듈로 넘겨진 경우 부분 문장 분석으로 띄어쓰기 오류 가능 어절(표 2의 굵은 글씨)과 그 오류 가능 어절과 연어관계(collocation)나 비연어관계(anti-collocation)에 있을 수 있는 피지배소(표 2의 이탤릭체 글씨)가 존재하는지 검사한다. 부분 문장 분석에는 크게 4가지 분석 방향이 가능하다: 오른쪽 분석 ⑤, 왼쪽 분석 ⑥, 양방향 분석 ⑧, 조건적 분석 ⑦로 띄어쓰기나 붙여쓰기에 의해 대치어가 제시된다.



[그림 1] 어절 간 띄어쓰기 교정 모듈

한편, 띄어쓰기 일반 처리 모듈에서는 각 띄어쓰기 기법을 적용하여 띄어쓰기 위치를 정한 다음 분리된 각 어절이 형태소 분석이 되는지 검사한다. 검사가 성공적이면 시스템은 대치어를 제시하고 성공적이지 못하면 검사 중인 어절을 오용어-표준어 사전을 토대로 검사하여 표준어로 대치한다. 남아 있는 띄어쓰기 기법이 없다면 대치어를 제시하지 못한 채 띄어쓰기 루틴을 빠져 나온다.

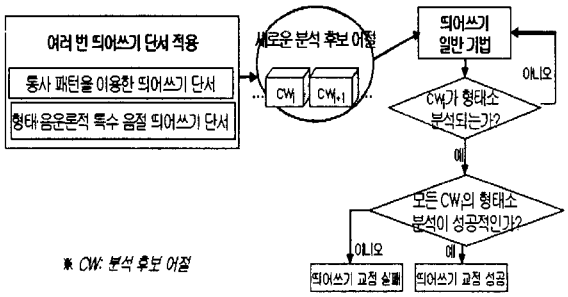


[그림 2] 어절 내 일반 띄어쓰기 교정 모듈

어절 내 띄어쓰기를 위한 기법들을 적용 가중치가 높은 순서로 나열하면 (1) 휴리스틱을 사용한 기법, (2) viable prefix를 이용한 최장 일치 기법, (3) '접두명사' 및 접두사(prefix)와 그 이웃하는 명사 사이의 관계를 살펴보는 기법, (4) 형태소 분석정보를 바탕으로 가중치별로 띄어쓸 위치를 정하고 주로 분석해야 할 형태소 범위를 정하는 기법이 있다. 어절 내 한 번 띄어쓰기 오류 어절 교정은 이러한 다양한 기법들을 사용하여 높은 교정 정확도(97.81%, 표 3 참조)를 얻을 수 있다. 기존의 시스템에서는 띄어쓰기 위치를 중심으로 분리된 각 어절이 모두 형태소 분석되어야만 대치어가 제시되므로 띄어쓰기 기법이 모두 사용된 후에도 형태소 분석이 되지 않는다면 대치어 없이 아래의 띄어쓰기 모듈을 빠져 나간다. [표 1]의 ③, ④와 같이 두 군데 이상 띄어 써야 할 어절이 발견되면 대치어를 제시하지 못한 채 띄어쓰기 처리 루틴이 종료된다.

4. 어절 내 여러 번 띄어쓰기 기법

위에서 발견된 문제점을 개선하기 위하여 현 시스템은 단서를 이용한 여러 번 띄어쓰기 기법을 개발한다. 특정 단서를 이용해서 띄어 쓸 위치를 정한다. 이 절차에 따라 생성된 어절 후보들에 대한 형태소 분석을 한다. 이 때 형태소 분석에 실패한 후보들은 띄어쓰기 일반 모듈로 다시 넘겨지고 일반 기법들의 적용을 거친 후 다시 형태소 분석이 되는지 검사를 거치게 된다.



[그림 3] 어절 내 여러 번 띄어쓰기 교정 모듈

예를 들어 '④ 교사및학생들의모임'과 같은 어절이 발견되면 띄어쓰기 일반 모듈은 어절 중간에 위치하는 형태 '및'을 중심으로 '교사 및 학생들의모임'과 같이 띄어쓰는데 이때 '학생들의모임'은 여전히 형태소 분석이 되지 않는다. 후자의 경우는 띄어쓰기 일반 모듈을 다시 거치면서 1음절 휴리스틱 '및'을 중심으로 새로운 띄어쓰기 위치를 얻게 되고 결과적으로 형태소 분석이 가능한 모든 어절을 얻음으로써 옳바른 띄어쓰기를 할 수 있게 된다. 현 시스템에는 370여 개의 휴리스틱이 여러 번 띄어쓰기에 고유한 단서로 사용되고 있다. 이들은 다양한 문서들로부터 추출한 많은 오류들을 분석하여 통사적 패턴을 이용한 것이 많으며 한 번 띄어쓰기에 사용되는 2음절 이상 띄어쓰기 단서나 1음절 단서도 이용한다. 그러나 1음절 띄어쓰기 단서는 오교정률을 높일 위험이 크므로 항상 그 양쪽을 띄어 써야 하는 1음절 단서나 '없', '않', '밭' 등과 같은 일부 1음절을 제외하고는 사용하지 않는다. 다음은 이 단서들의 대표적 예들이다.

⑨ 오류 어절 내에 단서가 나타나는 위치를 고려하지 않음
- 굴절된 어절의 마지막 특수 용절 뒤: 를#, 까봐#

⑩ 오류 어절 내에 단서가 나타나는 위치 고려함
▷ 오류 어절의 앞부분에 위치
- 굴절하지 않는 특수 단어 뒤: 그#, 즉#

▷ 오류 어절의 뒷부분에 위치
- 특수 용절 앞: #않, #깨끗
- 특수 음절 뒤: 의#

▷ 오류 어절의 중간에 위치:
- 특수 음절의 앞과 뒤: #및#, #또는#
- 관형형 어미 + 의존 명사: …을(ㄹ)#때…
- 조사 + 불완전 동사 : …로#인한#

위의 단서들 중 뒤에서 떼는 규칙들에는 제약 조건이 붙는 경우가 있다. 예를 들어 ⑨의 '까봐#'와 같은 단서에는 '까봐서'와 같이 확장한 형태는 예외로 하는 제약 규칙이 첨가된다. 이러한 현상은 한국어가, 문법 형태소들이 실질 형태소 뒤에 연이어 붙을 수 있는 교착어적 경향이 강한 언어라는 사실과 밀접한 관련이 있는 것이다.

5. 실험 및 결론

이 논문에서 어절 내 한 번 띄어쓰기 기법, 어절 내 여러 번 띄어쓰기 기법, 어절 간 띄어쓰기 교정 기법을 확장하고 개선했다. 이러한 각각의 기법들에 대한 교정 정확도를 실험하고 띄어쓰기 교정 성능 개선에 따른 전체 문법 검사기의 성능을 실험하기 위하여 약 1,800만 어절을 포함하는 실험 데이터를 현 문법 검사기로 돌렸다. 총 1,567,342어절이 시스템에 의해 오류 어절로 인식되었는데, 그중 시스템이 띄어쓰기 오류 어절로 인식한 어절 가운데 18,462어절을 표본으로 추출하여 수동으로 분석한 결과는 아래 표와 같다.

[표 3] 띄어쓰기 오류 교정 실험 결과

시스템이 오류로 인식한 어절 / 띄어쓰기 교정	오류 띄어쓰기 한 번 어절	오류 띄어쓰기 여러 번 어절	오류 띄어쓰기 어절 간 어절	계
바르게 고친 것	15,947	197	1,950	18,094
틀리게 고친 것	356	12	0	368
계	16,303	209	1,950	18,462
빈도(%)	88.31	1.13	10.56	100.00

정확한 시스템 성능을 알아보기 위해서는 재현율(recall ratio)과 정확도(precision ratio)를 평가해야 하지만, 이 논문에서는 실험 데이터에 존재하지만 시스템이 인식하지 못했을 띄어쓰기 오류 어절에 대한 조사 결과를 아직 얻을 수 없으므로 재현율을 측정할 수 없다. 실험 데이터 중 미등록어에 대한 교정 실패는 문법 검사기의 교정 영역을 넘어선 오류로 판단되므로 성능 측정에서 제외한다. 또한, 한국어에서는 띄어쓰기와 붙여쓰기를 모두 허용하는 수의적인 경우가 있는데 올바른 교정으로 판단하였다. 한편, 문맥을 봐야만 교정이 가능한 어절 간 오류 어절들 중 잘못 붙여 쓴 어절을 문맥에 따라 띄어쓰기해야 하는 교정에 대해서는, 해당 오류 어절들의 출현 문맥을 파악할 수 없으므로, 정확한 평가를 못 했음을 밝힌다. 따라서 이 논문은 어절 간 오류 중 입력 어절이 띄어 쓴 어절로서 시스템이 붙여쓰기를 한 것에 대한 평가만 실시한다. 1.13% 빈도로 나타나는 여러 번 띄어쓰기 오류 어절은 이 논문에서 제안한 단서를 이용한 띄어쓰기 기법으로 94.26%의 정확도로 교정된다. 또한, 어절 간 띄어쓰기 오류 중 붙여 써야 하지만 띄어쓴 오류 어절은 전체 띄어쓰기 오류의 3.14%를 차지하는데 어절 간 띄어쓰기 교정 기법에 의해 100%의 교정 정확도를 얻을 수 있었으며 다음과 같은 전체 띄어쓰기 교정 정확도를 얻을 수 있었다.

$$\text{띄어쓰기 교정 정확도} = \frac{\text{시스템이 오류로 인식한 어절 중 올바르게 교정된 띄어쓰기 오류 어절 수}}{\text{시스템이 오류로 인식한 어절 중 띄어쓰기 오류 어절 수}} = \frac{18094}{18462} \times 100 = 98.01\ \%$$

이 논문에서는 기존 문법 및 철자 검사기의 교정 정확도 및 교정 속도를 개선하기 위해 일반 띄어쓰기 오류 교정 모듈과 부분 문장 분석 모듈을 확장하고 개선하였다. 또한, 기존 연구에서 불가능하던 여러 번 띄어쓰기 교정이 가능하게 하기 위해 띄어쓰기 오류가 생겨난 언어학적 요인에 대한 이해를 바탕으로 여러 번 띄어쓰기에 적합한 단서들을 지식베이스에 구축하고 단서를 이용한 여러 번 띄어쓰기 기법을 개발하였다. 이 논문에서 제안한 단서를 이용한 어절 내 띄어쓰기 기법과 잠재적 지배관계 개념과 부분 문장 분석 기법을 이용한 어절 간 띄어쓰기 기법을 통해 현 시스템이 높은 성능을 얻을 수 있었다. 그러나 띄어쓰기 오류를 틀리게 교정하는 문제(전체 띄어쓰기 오류 어절의 1.99%)도 일괄적으로 처리할 수 있는 방법을 마련하여야 한다.

참고문헌

[1] 강승식, "한국어 형태소 분석과 정보 검색", 홍릉과학 출판사, 2002.
[2] 김영택 외, "자연언어 처리", 생능출판사, 2001.
[3] 심광섭, "음절 간 상호정보를 이용한 한국어 자동띄어쓰기", 정보과학회논문지(B), 23-9, 991-1000, 1996.
[4] 심철민, "어절 간 연관 관계와 오류 유형 추정 규칙에 기반한 한국어 철자 교정기", 부산대 전산과 석사 학위 청구 논문, 1995.
[5] 이희승 외, "한글 맞춤법 강의", 신구문화사, 2001.
[6] Kang, M.Y., Park S.H., Yoon A.S., Kwon H.C. "Potential Governing Relationship and a Korean Grammar Checker Using Partial Parsing", Lecture Note in Computer Sience. IEA/AIE, 692-702, 2002.
[7] Kim, S.N., Nam, H.S., Kwon H.C, "Correction Methods of Spacing Words for Improving the Korean Spelling and Grammar Checkers", Proc. 5th NLP Pacific Rim Symposium. 415-419, 1999.