음절 및 형태소 정보를 이용한 띄어쓰기 일관성 검사

Word Spacing Consistency Check using Syllable and Morpheme Information

이재성
충북대학교 사범대학 컴퓨터교육과

Jae Sung Lee(jasonlee@cbu.ac.kr)

요약

한국어 띄어쓰기 규칙은 경우에 따라 예외 조항이 있어 띄어 쓰거나 붙여 쓰는 것을 모두 허용하는 경우가 있다. 이러한 이중적 규칙에도 불구하고 같은 문서 내의 같은 어절이나 어구들은 일관성 있게 띄어쓰거나 붙여 쓰는 것이 문서 교정상 올바르다. 본 논문에서는 음절 정보 및 형태소 정보를 이용하여 비일관적으로 쓰인 띄어쓰기를 효과적으로 검사하는 방법을 제안하고 실험하여 평가하였다.

■ 중심어 : | 문서 교정 | 띄어쓰기 일관성 | 음절 정보 | 형태소 정보 |

Abstract

Korean word spacing rules have exceptional cases which permit both spacing and no-spacing between words. The exceptional cases, however, do not mean that inconsistent spacing between words or word-phrases is legitimate in a document proof reading. This paper proposes a word spacing consistency check method using syllable and morpheme information, and evaluated it through experiment.

■ keyword : | Proof Reading | Korean Word Spacing Consistency | Syllable Information | Morpheme Information |

I. 서 론

한국어 문장에서 띄어쓰기는 가독성을 높여주며 의미 전달을 명확히 할 수 있도록 한다. 올바른 띄어쓰기를 위해 국립국어원에서는 한글 맞춤법을 정하고 이를 모든 저자들이 지키도록 하고 있다.

그러나 띄어쓰기 규칙의 일부에서는 복수 규칙을 허용하고 있고, 이 때문에 같은 어절이나 단어에 대해 띄어쓰기의 일관성이 없는 경우가 있다. 다음은 한글 맞춤법의 띄어쓰기 규칙 중 복수 규칙을 허용하는 경우이다[1][2].

제47항: 보조용언은 띄어 씀을 원칙으로 하되, 경우에 따라 붙여 씀도 허용한다.

제48항: 성과 이름, 성과 호 등은 붙여 쓰고, 이에 덧붙는 호칭어, 관직명 등은 띄어 쓴다. 다만, 성과 이름, 성과 호를 분명히 구분할 필요가 있을 경우에는 띄어 쓸 수 있다.

제49항 성명 이외의 고유 명사는 단어별로 띄어 씀을 원칙으로 하되, 단위별로 띄어 쓸 수 있다.

제50항 전문 용어는 단어별로 띄어 씀을 원칙으로 하되, 붙여 쓸 수 있다.

* 이 논문은 2008년도 충북대학교 학술연구지원사업의 연구비지원에 의하여 연구되었음.

접수번호 : #100126-001
접수일자 : 2010년 01월 26일
심사완료일 : 2010년 05월 10일
교신저자 : 이재성, e-mail : jasonlee@cbu.ac.kr

이러한 복수 규칙은 맞춤법상 두 가지를 모두 맞는 것으로 처리하지만, 한 문서에서 같은 의미의 어절이나 단어에 대해 서로 다른 띄어쓰기를 하는 것은 문서 교정 원칙상 잘못된 것이기 때문에 문서 출판시 일관성이 유지되도록 수정한다. 예를 들어 한 문서에서 같은 단어에 대해 '문서교정시스템', '문서 교정 시스템', '문서교정 시스템' 등과 같이 다양한 띄어쓰기를 하는 경우, 이를 수정하여 하나의 통일된 띄어쓰기 형태로 표시한다. 이러한 비일관적인 띄어쓰기는 실제 많이 일어나며, 이를 위한 교정 작업은 저술이나 텍스트 콘텐츠 작성시 매우 번거로운 작업 중의 하나이다.

또한, 자동 번역 시스템이나 정보검색 시스템 등에서 대부분 띄어쓰기 단위로 번역 단위나 색인어를 선택하므로 띄어쓰기 일관성은 자연언어 처리 프로그램의 성능에도 영향을 줄 수 있다. 예를 들어 두 단어로 구성된 복합어일 경우, 일관성 있게 붙여 쓰거나 띄어 쓰면 문서 내에서 일정하게 1개의 번역어 혹은 색인어로 사용되지만, 비일관적으로 사용되면 붙여 쓴 경우와 띄어 쓴 경우가 서로 다르게 번역되거나 색인되어 검색될 수 있다. 특히 통계적 번역 시스템에서의 자동 정렬[3][4]이나 정보검색 시스템에서의 단어 유사도 계산시[5][6]에 잘못된 통계 정보를 만들어 낼 수 있다. 따라서, 입력 문서의 정규화를 위해 띄어쓰기 일관성이 활용될 수도 있다.

현재 시장에서 유통되는 한국어 워드프로세서들은 대개 띄어쓰기 자동 교정을 지원하고 있지만, 띄어쓰기 일관성에 대한 처리는 하고 있지 않다. 또한, 학계에서도 자동 띄어쓰기에 대한 연구는 많이 진행되었지만[7-11], 띄어쓰기 일관성 관점에서 이루어진 연구는 현재까지 찾아 볼 수 없다.

본 논문에서는 음절 정보와 형태소 정보를 이용하여 띄어쓰기 일관성을 쉽게 검사할 수 있는 방법을 제안하고 실험한다. 이어 2장에서는 기존의 띄어쓰기 교정 시스템 및 자동 띄어쓰기 시스템을 살펴보고, 3장에서는 띄어쓰기 일관성 검사를 하는 정도에 따라 3가지 수준을 제안하고 설명하며, 4장에서는 어휘 수준의 띄어쓰기 일관성 검사 방법을 설명하고 그에 따른 3가지 처리 모델을 제안한다. 이어 5장에서는 그 실험 결과를 제시하고 평가하며 6장에서 결론을 맺는다.

II. 자동 띄어쓰기 교정 시스템

띄어쓰기 시스템은 철자 교정의 하나로 띄어쓰기 교정을 하기 위해 만들어진 교정 시스템과, 띄어쓰기가 거의 이루어지지 않은 문서를 전체적으로 붙여 쓴 다음, 다시 규칙에 따라 띄어 쓰는 자동 띄어쓰기 시스템이 있다. 교정 시스템은 주로 두 어절 사이의 띄어쓰기를 교정하기 위한 방법을 중심으로 개발되었고 자동 띄어쓰기 시스템은 세 어절 이상의 여러 어절 혹은 문장 전체의 띄어쓰기를 자동으로 하기 위한 방법으로 개발되었다. 두 가지 시스템 모두 규칙에 의한 방법과 통계를 이용한 방법, 그리고 이 둘을 합하여 복합적으로 이용한 방법이 있다[7][10].

규칙에 의한 방법은 한 어절 혹은 두 어절에 대해 형태소 분석을 한 후, 미리 규칙으로 만들어 놓은 붙띄오류(붙여 쓸 것을 띄어 쓴 오류)와 띄붙오류(띄어 쓸 것을 붙여 쓴 오류) 유형을 분석하여 이를 근거로 올바른 띄어쓰기를 제시한다. 만약 형태소 분석에 실패할 경우, 그 어절을 분리하여 복합어로 분리하거나 부분 문자열에 대한 형태소 분석을 시도하여 오류 유형을 분석한다[9]. 이런 방법의 문제점은 형태소 분석기의 한계가 그대로 이 방법에 영향을 미친다는 점이다. 즉, 형태소 분석기 자체의 오류와 미등록어에 대한 처리 한계 등이 성능 향상의 한계가 된다.

통계에 의한 방법은 과도한 형태소 분석을 피하기 위해 통계 정보를 이용하여 대략 어절을 분리하는 방법이다. 이를 위해 단어 내와 단어 경계에서 음절 사이의 분리 횟수나 어절 사이의 분리 횟수를 통계적으로 계산하여 분리에 이용하거나, 조사, 어미 등의 음절 특성을 이용하여 어절을 분리한다. 대부분의 통계적 방법은 통계 자료의 부족 등으로 정확성의 한계가 있기 때문에 복합적 방법의 전단계로 활용한다. 복합적 방법에서는 통계적 방법으로 분리된 어절이 올바른지를 형태소 분석을 통해 검증하여 정확도를 높인다[7][8][10][11].

띄어쓰기 교정 시스템은 잘못 띄어 쓴 부분을 교정하

는 것이기 때문에 복수 규칙을 허용하며, 따라서 일관적인 띄어쓰기에 대해서 처리를 하지 않는다. 또, 대부분의 자동 띄어쓰기 시스템은 모든 단어를 붙여 쓴 후, 시스템이 자동으로 띄어 써 주기 때문에 일관성 있는 띄어쓰기를 하는데 사용할 수도 있다고 생각할 수 있다. 그러나 현재 이런 시스템들의 성능은 대개 어절 정확도가 93% 정도*[10][11]로 아직 완벽하지 않아 많은 오류가 그대로 포함될 수 있다. 더구나 저자가 의도적으로 띄어 쓰거나 붙여 쓴 것을 무시하고 시스템이 규칙이나 통계 정보로 잘못 판단하여 교정할 수 있다. 따라서, 자동 띄어쓰기 처리 후 사람의 개입이 필요하지만, 자동 띄어쓰기 시스템은 비일관적으로 쓴 단어들을 전체적으로 점검할 수 있는 방법은 제공하고 있지는 않다.

본 논문에서 제시한 방법은 비일관성 띄어쓰기의 가능성이 있는 부분을 보기 쉽게 저자에게 제공하고 저자가 판단하여 수정할 수 있도록 한다. 또한, 필요하면 어휘 및 형태소 정보를 이용하여 가능성이 높은 부분만을 선별하여 검사할 수 있도록 한다.

III. 띄어쓰기 일관성 검사 수준

띄어쓰기 일관성 검사는 여러 수준으로 이루어 질 수 있다. 본 논문에서는 크게 어휘 수준, 유형 수준, 의미 수준의 3단계 처리 수준을 이론적으로 제안한다.

1. 어휘 수준의 일관성 검사

어휘 수준에서의 일관성 검사는 기본 어휘가 같은 어절들이 문서 내에서 한 가지 띄어쓰기 형태로 통일하여 사용되는지를 검사하는 것을 말한다. 여기에서 기본 어휘가 같다는 것은 어미 변화나 조사 등은 제외하고 명사나 용언의 어휘가 같다는 것을 뜻한다.

예를 들어, (1)과 (2)는 첫 단어가 '컴퓨터'로 같으며, 뒤의 단어 '과학'도 같다. 따라서 '과학' 단어와 그 앞의 단어를 일관성 있게 띄어 써야 한다. 하지만, (2)와 (3)의 경우, 첫 어절이 서로 다르므로 '과학' 단어와 그 앞의 단어들이 반드시 일관성 있게 띄어 쓸 필요는 없다. 또, (4)와 (5)의 경우는 '꺼져 가다'에서 어미 변화가 된 것으로 기본 어휘가 같고, 본용언과 보조용언 사이에 띄어쓰기 일관성이 없으므로 이를 통일해야 한다. (단, 여기에서는 띄어쓰기의 옳고 그름은 판별하지 않고, 띄어쓰기의 일관성만을 검사한다.)

컴퓨터 과학은	(1)
컴퓨터과학에서	(2)
정보 과학은	(3)
꺼져 가는	(4)
꺼져간	(5)

2. 유형 수준의 일관성 검사

유형 수준의 띄어쓰기 일관성 검사는 같은 유형의 어휘들로 만들어진 어절들이 문서 내에서 통일된 띄어쓰기 형태로 사용되는지를 검사하는 것이다.

예를 들어 (2)와 (3)에서 사용된 '컴퓨터'와 '정보'는 똑같이 과학의 한 종류를 나타내는 어휘들이 '과학'과 함께 쓰인 복합어이다. 이 경우, 유형 수준에서 일관성이 있으려면 띄어쓰기 여부를 일치시켜야 한다. 하지만, 이런 복합어의 경우는 띄어쓰기 규칙상 단어가 합성어일 경우 붙여 쓰고, 합성어가 아닐 경우 띄어 쓴다. 즉, '컴퓨터과학'이 합성어라면 붙여쓰고, '정보 과학'이 합성어가 아니라면 띄어써야 한다.

또한, (4)와 (6)에서처럼 '가다'는 같은 보조 용언의 경우도 두 가지가 서로 다르게 띄어 쓰고 있다. 따라서, 같은 유형의 띄어쓰기 일관성에 문제가 있지만, 각 어휘별 합성어 여부에 따라 다르게 처리해야 한다.

살아가고	(6)

복합어의 띄어쓰기는 각각의 경우마다 합성어 여부가 다르기 때문에 대개 그 기준을 표준국어대사전에 합성어로 등재되었는가의 여부에 따라 결정한다. 하지만, 현재 표준국어대사전에 등재된 합성어 목록도 계속 바뀌고 있으며, 경우에 따라 국어학 혹은 언어학 관점에

* 이 정확도도 띄어쓰기 일관성을 고려하지 않고 복합어에 대해 붙여 쓴 경우와 띄어쓴 경우를 모두 맞는 것으로 계산한 것임

서도 합성어인지 아닌지에 대한 의견이 분분한 경우가 있어 이를 유형 수준의 띄어쓰기 일관성 검사에 적용하기에는 아직 어려운 점이 있다[12].

3. 의미 수준의 일관성 검사

의미 수준의 띄어쓰기 일관성 검사는 문맥상 의미를 파악하여 띄어쓰기 검사를 하는 것이다. 즉, 저자가 의도적으로 그 의미를 다르게 하기 위해 띄어 쓰거나 붙여 쓰는 경우, 이를 파악하여 처리한다. 예를 들어 합성어의 경우, 두 단어가 모여 새로운 의미의 단어를 형성하므로 붙여 쓰고, 이를 띄어 쓴 경우와 구분한다. 예를 들어 '큰아버지' 처럼 붙여 쓴 경우는 아버지의 형님이란 뜻이지만, '큰 아버지'라고 띄어 쓴 경우 키가 크거나 몸집이 큰 아버지를 의미한다. 따라서, 의도적으로 그 의미를 명확히 하기 위해 띄어 쓴 경우와 그렇지 않은 경우를 파악하기 위해서도 같은 의미의 어절들에 대해서는 띄어쓰기 일관성을 유지해야 한다. 이러한 경우는 앞뒤 문맥을 파악하여 처리해야 한다. 하지만 이 경우도 현재의 자연언어 처리 기술 수준이 전체 문맥 파악을 제대로 할 수준이 아니므로 처리하기 어려운 점이 있다.

띄어쓰기 일관성 검사는 간단한 어휘 수준 검사로부터 점점 어려운 단계인 유형 수준 검사, 의미 수준 검사로 진행해 갈 수 있을 것이다. 그러나, 본 논문에서는 현재의 여건과 관련 기술 수준 한계로 인해, 실용적으로 사용가능한 단계인 어휘 수준의 띄어쓰기 일관성 검사에 대해서 구현하고 평가한다.

IV. 어휘 수준 띄어쓰기 일관성 검사 모델

1. 탐색키를 이용한 후보군 추출

(1)과 (2)의 경우처럼 복합어가 비일관적으로 띄어쓴 경우를 찾기 위해서는 우선 명사 부분을 분리하고 이 명사들이 붙여 쓴 경우와 띄어 쓴 경우를 찾아야 한다. 이를 위해서는 모든 어절을 형태소 분석하고 이 형태소들을 기준으로 각 경우를 찾을 수 있다. 하지만, 이 경우, 비일관적인 띄어쓰기의 가능성이 없는 어절들도 모두 포함하여 형태소 분석을 하게 되어 비효율적이다.

본 논문에서는 효율적인 탐색을 위해 비일관적 띄어쓰기 후보들을 간단한 문자열 비교 방법으로 먼저 찾아내고 이 후보들만을 대상으로 형태소 분석을 하는 효율적인 방법을 제안한다. 이 방법은 기본적으로 빈칸을 중심으로 좌우 어절의 일부를 추출하여 탐색키로 사용하는 방법이다.

탐색키로 찾아낸 후보군은 비일관적으로 띄어 쓴 모든 어절을 포함해야 한다. 이를 위해 탐색키는 여러 가지로 정의하여 사용할 수 있지만, 본 논문에서는 탐색키를 빈칸 앞쪽의 어절과 뒤쪽 어절의 첫 글자로 정하였다. 즉, (1)에서 (6)의 예에서 보듯이 띄어쓰기의 비일관성이 나타나는 복합어나 보조용언을 사용한 경우, 대개 앞 어절은 변화가 없는 반면 뒷 어절의 한 글자는 대개 같고 그 이후는 어미나 조사 등이 붙어 변화할 수 있기 때문이다. 물론 뒷 어절의 한 글자도 (5)의 예처럼 변화할 수 있기 때문에 이를 고려하여 3가지 모델을 다음 절에 제안한다.

탐색하는 문자열을 목표 문자열이라 하고, 띄어 쓴 문자열을 A형 목표 문자열, 붙여 쓴 문자열을 B형 목표 문자열이라고 할 경우, (4), (5)에 대한 탐색키는 [그림 1]처럼 정의된다.

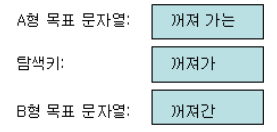

그림 1. 목표 문자열 및 탐색키 예

탐색키는 기본적으로 띄어 쓴 어절에서 추출하므로 문서를 읽는 첫 번째 패스때 띄어 쓴 어절(A형 목표 문자열)로부터 가능한 모든 종류의 탐색키를 구축한다. 이 탐색키를 이용하여 두 번째 패스때 부분 문자열 검색으로 붙여쓴 어절(B형 목표 문자열)들을 추출한다. 만약 붙여쓴 어절이 발견되지 않으면 비일관적 띄어쓰기의 대상에서 그 탐색키로 찾은 어절들을 제거한다. 다음 [그림 2]는 이 방법을 포스트 파일 형태로 구현한 예이다.

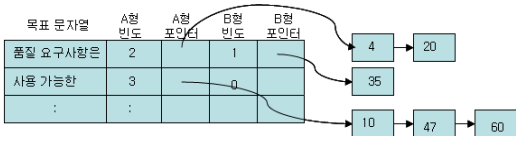

그림 2. 포스트 파일 예

2. 제안 모델

모델은 탐색키를 이용하여 검색하는 방법과 후처리 방법에 따라 3가지로 나누었다. 첫 번째는 '어휘 일치 모델'이며 T1으로 약칭한다. 이 모델은 탐색키와 완전 일치를 하는 부분 문자열만을 찾는 단순한 모델이며, 다른 모델과의 비교를 위해 사용한다.

두 번째는 '동일어간 음절후보 일치 모델'로 Tc로 약칭하며, T1모델을 보완하기 위해 만든 모델이다. 명사의 경우와 다르게 보조용언 어간의 경우, 어절의 앞 음절이 어미 변화에 따라 바뀔 수 있다. 이러한 경우는 정확한 음절 일치만을 허용하는 T1모델로는 찾을 수 없다. (4)와 (5)에서처럼 보조용언의 활용형태를 고려하여 검색문자열과 비교하여야 한다. 즉, 보조용언 어간 '간'이 '가'의 활용형태임을 파악하여 '꺼져가' 가 '꺼져간'과도 일치되는 것으로 처리해야 한다.

보조용언의 종류는 [13]에 의하면 [표 1]과 같이 33종류가 있다. 이 보조용언의 어간은 어미 변화에 의해 바뀌기는 하지만 변화되는 음절의 갯수가 한정적이다. 본 논문에서는 이들 중 어미 변화에 의해 보조용언의 첫 음절이 바뀌는 경우를 조사하여 [표 2]와 같이 정리하였다. 이 표에서 나타난 바와 같이 같은 어간이지만 활용형으로 나타날 수 있는 음절들을 '동일어간 음절후보'라고 정의한다. 또, 이를 이용하여 탐색키의 마지막 음절과 대상 문자열에서 비교되는 음절이 같은 동일어간 음절후보에 속할 경우 문자열이 일치한 것으로 처리한다.

표 1. 보조용언의 종류

가다	들다	싶어지다
가지다	듯싶다	아니하다
갖다	듯하다	않다
계시다	만하다	오다
나가다	말다	있다
나다	먹다	주다
내다	못하다	죽다
놓다	버리다	척하다
달다	보다	체하다
두다	뻔하다	치우다
드리다	싶다	하다

표 2. 첫 음절이 변하는 보조용언 및 동일어간 음절후보

보조용언	동일어간 음절후보
가다, 가지다, 갖다	가, 간, 갈, 감, 갑, 갔, 갖
나다	나, 난, 날, 남, 납, 났
내다	내, 낸, 낼, 냄, 냅, 냈
달다	달, 다, 단, 담, 답
두다	두, 둔, 둘, 둠, 둡, 뒀
들다	들, 든, 드, 듬, 듭
말다	말, 만, 마, 맙
보다	보, 본, 볼, 봄, 봅, 봤
오다	오, 온, 올, 옴, 옵, 왔
주다	주, 준, 줄, 줌, 줍, 줬
하다	하, 한, 할, 함, 합, 했

세 번째 모델은 '형태소 일치 모델'이며 Tm으로 약칭한다. 이 모델은 Tc모델에서 후보로 찾은 어절들에 대해서 형태소 분석을 하고 이를 근거로 비일관적 띄어쓰기 어절을 찾아낸다.

예를 들어, T1과 Tc모델은 아래의 (7)과 (8)의 경우를 모두 비일관성 후보로 추출한다. 따라서, '부품은'을 올바르게 분석하고 명사 부분인 '부품'을 분리하여 (8)의 조사 '부터'와 다른 것으로 처리해야 한다. 이를 위해서 형태소 분석이 필요하다.

메모리용 부품은 (7)
메모리용부터 (8)

또한, 활용형태를 고려하여 검색하는 Tc 모델의 경우, 아래 (9), (10)에 나타난 '간'과 '가'를 동일어간 음절후보로 처리하여 잘못된 결과를 출력할 수 있다. 이를 해결하기 위해서도 앞 어절과 뒤 어절에 대해 형태소 분석을 하여 보조용언과 다른 품사를 구분해 내야 한다.

레지스터 간 (9)
레지스터가 (10)

비교 대상이 되는 두 어절에 대한 형태소 분석 결과는 품사와 어휘수준에서 동시에 비교한다. 즉, 형태소가 명사나 동사와 같은 품사일 경우에는 어휘까지도 같아야 하며, 그 이외의 형태소들은 품사가 같으면 두 어절이 같은 것으로 판별한다. 또, 형태소 분석의 결과가 여러 가지 후보로 나올 경우, 그 중 하나의 후보와만 일치해도 의미소가 같은 어절로 판별한다.

예를 들어, (11)과 (12)는 '버스위의' 와 '버스 위로'를 형태소 분석하여 띄어쓰기 일관성을 검사한 것이다. 두 경우를 모두 붙여 써서 형태소 분석을 하여 두개의 문법 요소로 분리된 것을 보여준다. 첫 번째 요소는 둘 다 '버스위/N' 라는 명사(N)로 분석되고 어휘도 같으며, 두 번째 요소는 어휘는 다르지만 모두 조사(j)이므로 이 어절은 같은 종류의 어절로 판별한다. 따라서 이 어절을 띄어쓰기 일관성이 없는 어절로 표시한다. (13)과 (14)의 경우도, 보조용언 '보다' 가 '본다' 와 '보기로' 로 어미가 변화되어 사용되었지만, 본용언과 보조용언의 동사 어간(V)이 일치하므로 어미(e)의 어휘 변화에 관계없이 같은 종류의 어절로 판별하여 띄어쓰기 일관성이 없음을 지적한다. (15)와 (16)은 '메모리용이라는' 과 '메모리 용량이라도' 의 두 어절에 대한 띄어쓰기 일관성 검사이다. 이 경우, 명사 부분이 어휘가 다르므로 다른 종류의 어절로 판별하여 띄어쓰기 일관성의 비교 대상에서 제외한다.

버스위/N 의/j (11)
버스위/N 로/j (12)

생각해보/V ㄴ다/e (13)
생각해보/V 기로/e (14)

메모리용/N 이라는/j (15)
메모리용량/N 이라도/j (16)

앞에서 설명한 3가지 모델을 다시 정리하면 다음과 같다.

- 어휘 일치 모델(T1): 탐색키에 완전 일치하는 목표 문자열을 찾는 것
- 동일어간 음절후보 일치 모델(Tc): 탐색키의 마지막 음절에 대해 동일어간 음절후보도 일치하는 것으로 처리하여 목표 문자열을 찾는 것
- 형태소 일치 모델(Tm): 모델 Tc로 찾은 결과를 다시 형태소 분석하여 기본 형태소가 같은 목표 문자열을 찾는 것

V. 실험 및 평가

1. 실험

실험을 위해 국립국어원에서 구축한 말뭉치 중 일부를 분야별로 뽑고[14], 또 학회 논문지에 실린 최근의 논문 중 임의로 5편을 뽑아 텍스트 파일로 변환하여 [표 3]과 같이 테스트 문서 집합을 만들었다. 실험의 편의상 책이나 소설 및 논문의 크기를 약 5,000어절 전후의 단편집 혹은 단편 소설로 선정했다. 또, 가능하면 다양한 분야(기록, 교육, 인문, 과학, 예술 등)와 다양한 저자의 책이나 소설로 선정했다. 작문은 대학생들의 비교적 짧은 작문들이며, 뉴스기사도 신문에 난 짧은 기사를 모아 총 어절수가 25,000개 정도가 되도록 했다.

표 3. 테스트 문서 집합

	어절수	문서수	문서당 평균 어절수
책	26804	5	5361
작문	24802	39	636
뉴스기사	24836	100	248
소설	25915	5	5183
논문	25888	5	5178
합	128245	154	833(평균)

실험은 앞 장에서 설명한 3가지 모델을 구현하고 이에 대해 평가했다. 이 중 Tm모델은 공개적으로 사용가능한 형태소 분석기[15]를 사용하였다. [그림 3]은 Tm 모델을 수행하고 그 결과를 출력한 예이다. 이 예에서

보듯이 비일관적으로 나타난 띄어쓰기 부분을 중심으로 좌우 문맥 정보와 파일 이름을 표시하여 사용자가 최종 판단을 할 수 있도록 했다. 각각의 한 줄은 '비일관적 띄어쓰기의 사례' (약칭하여 '사례')로 정의하며, 같은 띄어쓰기가 요구되는 사례들은 점선(중앙에 별표 포함)으로 구분하였고, 이렇게 구분된 집합을 '비일관적 띄어쓰기 그룹'(약칭하여 '그룹')으로 정의한다. (점선의 중앙에 있는 별표는 띄어쓰기가 나타난 어절의 위치를 표시하기 위한 것이다.) [그림 3]은 '기업 간'과 '기업간', '품질 요구사항'과 '품질요구사항' 등이 일관성 없이 띄어 쓴 사례를 찾아 낸 것을 보여준다. 이 프로그램이 더 유용하게 사용될 수 있도록 하려면 각 사례에 실제 나타난 텍스트 위치로 하이퍼 링크를 연결하여 편집이 쉽도록 개선할 수도 있을 것이다.

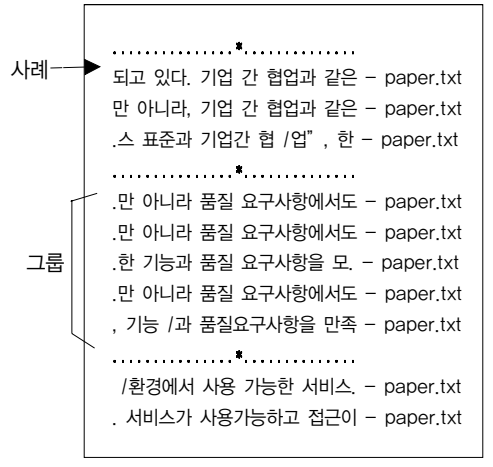

그림 3. 프로그램 출력 예 일부 ('/' 는 줄바꿈 표시)

2. 평가

평가를 위해 테스트 문서 집합에 대해 정답 사례 집합을 수작업으로 작성하였다*. 정답 사례 집합 작성시 Tc모델을 이용하였다. 즉, 이론적으로 Tc는 모든 가능한 비일관적인 띄어쓰기 사례를 추출할 수 있으므로 Tc에서 찾아 낸 비일관적인 띄어쓰기 중 올바른 것을 수작업으로 선택하고 그룹별로 정리하여 정답 사례 집합을 작성했다.

[표 4]는 실험 결과 찾아낸 비일관적 띄어쓰기 오류수를 분야별로 정리한 것이다. 테스트 문서 전체 분야의 띄어쓰기 오류 사례는 평균 74 어절당 1개이었다. 일반적으로 긴 문서에서 비일관적인 띄어쓰기가 나타날 가능성이 높다고 볼 수 있다. [표 4]에서 보듯이 실제 문서 길이가 긴 책, 소설, 논문류에서 비일관적 띄어쓰기가 많이 나왔다. 논문류에서 29어절당 1개씩 오류가 나타나 테스트 분야 중 가장 높은 오류율을 보였다. 예외적으로 뉴스기사류의 경우, 짧은 글임에도 비일관적인 띄어쓰기가 비교적 많았다.

표 4. 테스트 문서에 나타난 비일관적 띄어쓰기 수

	문서당 평균 어절수	오류사례수	문서당 오류사례수	오류사례당 어절수	오류그룹수
책	5361	497	99	54	109
작문	636	140	4	177	44
뉴스기사	248	350	4	71	125
소설	5183	655	131	40	181
논문	5178	881	176	29	107
평균	833†	505	83	74	113

모델 평가는 크게 사례와 그룹에 대한 결과로 나누어 수행했다. 앞에서 설명했듯이 사례는 비일관적으로 띄어 쓴 어절을 포함한 각각의 부분 문자열을 나타내며, 그룹은 서로 상충되는 띄어쓰기를 하고 있는 각 사례들의 집합을 나타낸다. 이 둘의 평가는 다시 각각에 대한 정확률과 재현율 및 F값으로 나누어 계산하였다. 사례에 대한 정확률과 재현율은 (식1), (식2)에, 그룹에 대한 정확률과 재현율은 (식3), (식4)에 나타냈다. F값은 정확률과 재현율의 평가 비중에 따라 다르게 계산할 수 있는데, 이 평가에서는 두 비중을 같게 하여 (식5)와 같이 계산하였다[16]. 이 식(5)을 이용하여 사례 F값과 그룹 F값을 각각 계산할 수 있다.

$$\text{사례정확률} = \frac{\text{정답에 있는 출력사례 갯수}}{\text{출력 사례 갯수}} \quad \text{(식1)}$$

$$\text{사례재현율} = \frac{\text{정답에 있는 출력사례 갯수}}{\text{정답 사례 갯수}} \quad \text{(식2)}$$

* 테스트 문서에 입력 오류가 없다는 가정하에 정답 사례 집합을 작성하였다.

† 문서당 평균어절수를 다시 평균으로 구한 것이 아니고, 테스트 문서 집합에 나타난 총어절수를 총 문서수로 나눈 것이다.

$$\text{그룹정확률} = \frac{\text{정답과 일치하는 출력 그룹 갯수}}{\text{출력 그룹 갯수}} \quad \text{(식3)}$$

$$\text{그룹재현율} = \frac{\text{정답과 일치하는 출력 그룹 갯수}}{\text{정답 그룹 갯수}} \quad \text{(식4)}$$

$$F = \frac{2 \times \text{정확률} \times \text{재현율}}{\text{정확률} + \text{재현율}} \quad \text{(식5)}$$

그룹 평가 식 (식3)과 (식4) 중 '출력 중 정답과 일치하는 그룹 개수'는 정답 그룹에 나타난 모든 사례가 출력 그룹에 포함되어야 개수에 포함한다. 예를 들어, 아래처럼 사례 a, b, c, d, e가 원소인 그룹 A1, A2, B, C, D가 있다고 하자. 집합 A1, A2가 정답 그룹일 경우, B는 정답과 일치하는 그룹이고, C와 D는 일치하지 않는 그룹이다. B의 경우, 일치하지 않는 사례 'f'가 포함되어 있기는 하지만, 검토를 하여 비일관적인 띄어쓰기를 모두 수정할 수 있다. 하지만, C는 사례 'a'가 누락되어 일관적인 띄어쓰기 교정을 못할 가능성이 있다(이를 미그룹 오류라고 정의함). 또, D는 A1과 A2의 두 그룹으로 나뉘어져야 하는데 나뉘어지지 않아 사례수가 많아지거나 혼합된 그룹수가 많을 경우, 수작업으로도 검토가 어려워 질 수 있어 이를 오류로 처리한다(이 경우를 과그룹 오류라고 정의함).

A1 = {a, b, c}
A2 = {d, e}
B = {a, b, c, f}
C = {b, c}
D = {a, b, c, d, e}

띄어쓰기 일관성은 한 문서 내에서의 일관성만을 검사하였으며, 각 실험모델의 평가 결과는 [표 5]와 같다. 사례에 대한 평가 결과를 보면 T1 모델은 Tc보다 재현율은 낮지만, 정확률이 약간 더 높아 F값이 Tc보다 약간 높았다. Tm 모델은 재현율은 93.2%로 Tc에 비해 상대적으로 낮지만, 정확률이 93.2%로 비교적 높아 F값은 93.2%이며, 이는 81.2%와 80.0%를 보인 다른 모델에 비해 현저하게 우수한 것이다. 즉, 단순한 음절 정보만을 사용한 T1, Tc모델보다, 형태소 정보를 사용한 Tm모델이 훨씬 우수함을 알 수 있다. 그룹에 대한 평가 결과도 대체적으로 사례 평가 결과와 유사했지만, 사례 평가와 다르게 Tc의 F값이 T1의 F값보다 높았다. 이는 T1이 같은 그룹에 속한 '비일관적 띄어쓰기 사례'들을 다른 그룹으로 나누어 넣은 경우가 많아 그룹 재현율이 상대적으로 낮게 나왔기 때문이다. 또, 여기에서 Tc의 그룹 재현율이 100%가 아닌 이유는 과그룹 오류가 포함되었기 때문이다. 이는 사실상 수작업을 통해 모든 그룹을 찾을 수는 있음을 나타낸다.

각 모델은 그 필요에 따라 선택해서 사용할 수 있다. 즉, Tc모델은 재현율이 높으므로, 모든 경우의 띄어쓰기 일관성 오류를 검사하고자 할 때 사용자의 시간이 걸리더라도 사용할 수 있다. 반면에 Tm모델은 정확률이 상대적으로 높아, 적은 노력으로 비일관적 띄어쓰기를 수정하고자 할 때 사용할 수 있을 것이다.

또 Tc모델은 Tm모델의 전단계로 형태소 분석 대상 수를 줄여주는 모델이다. 실제 실험 데이터의 총어절수는 128,245개이었고, Tc모델이 비일관성 후보로 제시한 사례수는 3,792개에 불과해 형태소 분석 대상 어절수를 원어절수의 3%로 줄여 주어 Tm 모델이 보다 효율적으로 처리할 수 있었다.

표 5. 각 모델의 사례 및 그룹에 대한 평가 결과(%)

	사례			그룹		
	재현율	정확률	F값	재현율	정확률	F값
T1	95.8	70.9	81.2	92.5	68.5	78.5
T_c	100.0	67.4	80.0	99.6	67.5	80.0
T_m	93.2	93.2	93.2	86.0	86.0	86.0

3. 오류 분석

실험에 사용한 데이터 중 각 분야별로 일부를 뽑아 대략 Tm모델의 오류 유형을 분석해 본 결과, 찾아 내야 할 것을 못 찾아 낸 오류(누락 오류)가 80%, 틀리게 찾아 낸 오류(거짓추출 오류)가 20%정도이었다.

누락 오류는 소설류 등에서 사용한 비표준어인 구어체 등을 형태소 분석기가 분석하지 못해서 추출하는 경우가 누락 오류의 50% 정도이고 그 외는 형태소 분석기 자체의 오류에 의한 것 등이다.

거짓추출 오류로는 붙여 쓰면 다른 의미로 바뀌는 경

우에 주로 나타났다. 현재 Tm모델에서는 두 어절이 의미소가 같은지를 검사하기 위해 띄어 쓴 어절을 붙여 쓴 다음 형태소 분석을 하고 비교한다. 예를 들면 '나 이 머리 좀 봐'에서 '나'와 '이'는 각각 대명사와 관형사로 분석될 수 있으나, 둘을 붙이면 명사 '나이'로 바뀐다. 따라서 '나이가'라는 명사가 포함된 어절과 일치하여 잘못 추출하였다.

일반적으로 비일관적인 띄어쓰기는 복수 띄어쓰기 규칙을 허용하는 복합어 및 보조용언에서 나타난다. 하지만, 실제 텍스트에 나타난 비일관적 오류를 보면 띄어쓰기 규칙이 틀린 오류나 명사 뒤에 '하다' 나 '되다' 등이 붙어 동사로 품사전성이 된 경우에 비일관적으로 띄어쓴 경우, '그때', '이날'과 같이 준합성어로 붙여써도 되고, 분리하여 띄어 써도 되는 경우 등이 나타나고 있다.

VI. 결론

한 문서 내에서 같은 단어에 대해 일관성 있게 띄어쓰기를 하는 것은 보다 정확한 의미 전달을 위해 필요하다. 그러나, 실제 출판된 책이나 기사, 소설, 논문 등을 대상으로 실험해 본 결과, 비일관적인 띄어쓰기가 많이 발견되었다.

본 논문에서는 띄어쓰기 일관성 검사를 어휘 수준, 유형 수준, 의미 수준의 3수준으로 제시하였고, 그 첫 수준인 어휘 수준으로 띄어쓰기 일관성 검사를 하기 위한 방법을 제안하고 구현하여 평가하였다. 제안한 모델 중 '동일어간 음절후보 일치' 모델은 재현율을 중시할 경우 사용할 수 있고, '형태소 일치' 모델은 재현율 및 정확률을 모두 고려할 경우 사용할 수 있을 것이다. 특히, '형태소 일치 모델'은 음절 정보 및 형태소 정보를 이용하여, 비용이 많이 드는 형태소 분석 대상 어절수를 원래 어절수의 3%로 대폭 줄였으며, 실험 결과 사례 재현율과 정확률이 모두 93.2%로 우수한 성능을 보였다. 이를 문서 교정 시스템에 이용하면 효율적으로 띄어쓰기 일관성을 검사하고 수정할 수 있으므로 보다 품질 높은 문서를 작성할 수 있을 것이다. 또한 색인어나 번역 단위를 통일하기 위해 정보검색 시스템이나 자동 번역기 등에서 전처리기로도 활용할 수 있을 것이다.

참 고 문 헌

[1] 국립국어원, *한국 어문 규정집*, (주)계문사, 2007.

[2] _, *문교부고시 88-1 한글 맞춤법 해설*, 국어연구소 간행, 1988.

[3] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," Computational Linguistics, Vol.19, No.2, pp.263-311, 1993.

[4] 신중호, *한국어/영어 병렬 코퍼스에 대한 단어단위 및 구단위 정렬 모델*, 한국과학기술원 석사학위 논문, 1996.

[5] G. Salton, *Automatic text processing*, Addison-Wesley, 1988.

[6] 박영찬, *정보검색을 위한 단어지식의 통계적 구축*, 한국과학기술원 박사학위 논문, 1997.

[7] 심광섭, "음절간 상호정보를 이용한 한국어 자동 띄어쓰기", 정보과학회 논문지(B), 제23권, 제9호, pp.991-1000, 1996.

[8] 신중호, 박혁로, "음절단위 bigram정보를 이용한 한국어 단어인식모델", 한글 및 한국어 정보처리 학술발표 논문집, pp.255-260, 1997.

[9] 최재혁, "양방향 최장일치법을 이용한 한국어 띄어쓰기 자동 교정 시스템", 한글 및 한국어 정보처리 학술발표 논문집, pp.145-151, 1997.

[10] 강승식, "한글 문장의 자동 띄어쓰기를 위한 어절 블록 양방향 알고리즘", 정보과학회 논문지, 소프트웨어 및 응용, 제7권, 제4호, pp.441-447, 2000.

[11] 이도길, 이상주, 임희석, 임해창, "한글 문장의 자동 띄어쓰기를 위한 두 가지 통계적 모델", 정보과학회 논문지, 소프트웨어 및 응용, 제30권, 제4호, pp.358-371, 2003.

[12] 조진현, 김일환, 이현희, 이영제, 강범모, “형태 분석 말뭉치 구축을 위한 합성어의 처리 방법 - 띄어쓰기를 고려하여 -”, 한글 및 한국어 정보처리 학술발표 논문집, pp.9-13, 2002.

[13] 국립국어원, “한국어 학습 자료,” 국립국어원 홈페이지 공개자료실 http://www.korean.go.kr, 2003.

[14] 국립국어원, *세종계획 연구교육용 균형말뭉치*, 2004.

[15] http://nlp.kookmin.ac.kr/.

[16] C. Manning and H. Schutze, “Foundations of Statistical Natural Language Processing,” pp.268-269, The MIT Press, 1999.

저 자 소 개

이 재 성(Jae Sung Lee) 정회원

- 1983년 2월 : 서울대 컴퓨터공학과(학사)
- 1985년 2월 : KAIST 전산학과(석사)
- 1999년 2월 : KAIST 전산학과(박사)
- 1985년 ~ 1988년 : 큐닉스컴퓨터(주) 과장
- 1988년 ~ 1993년 : 마이크로소프트 차장
- 1999년 ~ 2000년 : 전자통신연구원 팀장
- 2005년 ~ 2006년 : 미국 아리조나 대학 방문교수
- 2000년 9월 ~ 현재 : 충북대 컴퓨터교육과 부교수

<관심분야> : 정보검색, 자연언어 처리, 컴퓨터교육