문장부호를 고려한 특수어절 분석 알고리즘

김현주, 이영민, 이영상, 천승태
(주)데이터스트림즈 기술연구소
{hjookim, ymynlee, yslee, stchun}@datastreams.co.kr

Special Word Analysis Algorithm Considering Punctuations

Hyun-Joo Kim, Young-Myn Lee, Young-Sang Lee, Seung-Tae Chun
Research and Development, DataStreams Inc.

요 약

자연언어 분석에 있어서 형태소 분석은 핵심적인 기술로 요구되고 있다. 한글 형태소 분석기는 한글을 분석하기 위한 알고리즘을 활용하여 형태소 단위로 분석한다. 하지만 한글과 문장부호가 혼용된 특수어절은 한글을 분석하는 알고리즘을 통하여 정확한 결과를 도출할 수가 없으므로 별도의 알고리즘이 필요하다. 본 논문에서는 이러한 문제점을 특수어절에 공백을 삽입하여 다시 어절로 분리해 내는 알고리즘을 적용하여 해결하고자 한다.

1. 서론

문장부호는 글에서 문장의 구조를 드러내거나 글쓴이의 의도를 전달하기 위하여 사용하는 부호이다. 문장부호 뒤의 띄어쓰기는 그 부호 뒤에 오는 말에 따라 정해지는 것이 원칙이나 글 쓰는 사람에 따라 다양한 형태로 나타난다[4]. 또한 글을 쓸 때 다양한 문장부호를 사용하지만 그 용법에 따라 바르게 쓰기는 쉽지 않다.

최근 글쓰기 환경이 컴퓨터와 인터넷 중심으로 급격히 변화함에 따라 문장부호는 실제 언어생활에 쓰이고 있는 부호와 용법을 고려한 규정으로 변화되어 적용되고 있다. 하지만 각기 그 쓰임이 다르고 다양하여 그만큼 혼동되기 쉽다[1]. 이러한 경우 실생활에서의 정보 전달 등에서는 문제가 되지 않지만 정보 검색 시스템이나 기계 번역 시스템에는 심각한 문제를 야기한다[3]. 그러므로 형태소 분석기에서 문장부호의 적절한 처리가 이루어져야 한다.

(1) 이달 구호는 '친절 봉사' 입니다.
(2) "죽어도 해낼거야." 라고 그가 말했다.

예문 (1)의 '입니다'는 서술격조사인 '이다'의 활용형이고 예문 (2)의 '라고'는 조사이다. 조사는 윗말에 붙여 쓰므로 모두 문장부호 뒤에 붙여 적어야 한다. 또한 따옴표와 문장부호를 함께 쓸 때, 문장부호와 따옴표의 위치가 혼동될 수 있으며 한글 맞춤법에 맞게 쓰지 않는 경우도 있다[2][5]. 따라서 본 논문에서는 이러한 경우에도 적절한 형태소 분석이 될 수 있도록 문장부호를 고려한 특수어절 분석 알고리즘을 제안한다.

본 연구는 2015년 ETRI 연구과제(RO-1261-5106-7000-4003)의 지원을 받아 수행되었습니다.

본 논문은 구성은 다음과 같다: 2장에서는 논의를 위한 전반적인 사항을 살펴보고 3장에서는 문장부호를 고려한 특수어절 분석 알고리즘을 제안한다. 4장에서는 제안한 알고리즘을 적용, 그 결과를 분석하며 마지막으로 5장에서는 본 논문의 결론을 도출한다.

2. 관련 연구

2.1 형태소 분석 유형

한글 형태소 분석은 기본적으로 '명사+조사', '동사+어미'의 유형(pattern)이 주류를 이루지만 그 유형이 상당히 다양하게 나타난다.

(3) 학생들이 찍은 사진만이 점수를 받았다.
(4) 그는 어찌 공부함이 잘하는지 아는 학생이다.

(3)의 어절들은 '명사+조사(학생들+이, 점수+를)', '동사+어미(찍+은)'의 유형으로 분석되지만, '명사+조사복합체(사진+만이)', '동사+선어말어미+어미(받+았+다)'의 유형으로 분석되는 경우도 있다. 그리고 (4)의 예문을 보면 '동사+어미+조사(공부하+ㅁ+이)', '명사+서술격조사+어미(학생+이+다)'와 같이 그 유형이 다양하게 나타나므로 이들을 분석하기 위해서는 별도의 유형을 정해두어야 한다[1][6][7]. 또한 (1), (2)의 예에서 보았듯이 분리된 어절에서 한글과 문장부호가 혼용되는 경우(특수어절), 기존의 유형으로는 분석이 불가능하며 이를 처리하기 위한 알고리즘이 필요하다. 하지만 형태소 분석기 내에 특수어절 처리를 위한 알고리즘이 없을 경우, 모든 특수문자를 공백으로 치환하여 어절 단위로 분리하는 방법을 사용한다[2].

2.2 어절 분리의 문제점

문장을 분리할 때 공백을 기준으로, 즉 어절 단위로 분리하는데 어절 내에 문장부호를 포함하여 기타 기호가 혼용된 경우에는 문제가 발생한다[1].

(5) 3+2=5를, a>b에서는+

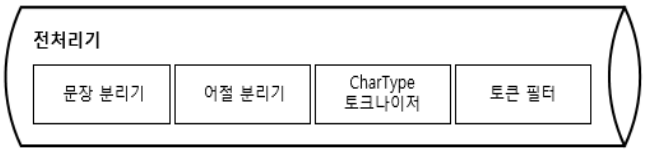
(5)와 같은 경우는 한글 이외의 글자 유무에 따라 토큰을 생성하고 토큰 필터를 활용하여 형태소 분석에 필요 없는 토큰을 제거하면 문제를 해결할 수 있다. 문제는 괄호나 따옴표 등의 문장부호가 어절 내에서 한글과 혼용되어 사용되는 경우이다.

(6) 홍길동(1933)에서는
(7) <누구를 위하여 종은 울리나?>를 읽으면
"얼마나 멀리 가느냐"가 문제이다.

(6), (7)에는 한 어절에 한글과 문장부호 등이 혼용되어 있다. (6)은 조사에 선행하는 부호를 분리하고 선행 명사를 처리할 수 있으므로 '명사+조사'의 유형으로 처리가 가능하다. 반면에 (7)은 '동사+어미+부호+조사(울리+나+?>+를, 가+느냐+"+가)'의 구성으로 부호를 분리하더라도 선행 요소가 명사가 아니므로 별도의 알고리즘으로 처리해 주어야 한다.

2.3 형태소 분석 전처리기

형태소 분석의 전처리 단계에서는 텍스트를 문장으로 분리하고 문장을 다시 어절 단위로 분리하여 한글 이외의 글자 유무에 따라 토큰을 생성한다. 생성된 토큰은 토큰필터에서 형태소 분석에 필요 없는 토큰을 제거하여 최종 분석할 토큰을 결정한다. 최종 토큰에 한글과 문장부호가 혼용되어 있을 경우 본 논문에서 제안하는 알고리즘을 적용한다. (그림 1)은 형태소 분석기의 전처리 단계이다.



(그림 1) 형태소 분석 전처리기

3. 문장부호 분리 알고리즘

3.1 문장부호 분리 방법에 대한 제안

국립국어원에서 지정한 표준 문장부호[1]는 24가지이다. 문장부호는 글에서 문장의 구조를 드러내기 위해 사용되는데 어절 내에서 한글과 혼용되어 있을 경우 형태소 분석에 어려움이 있다. 따라서 24가지의 문장부호를 형태소 분석 방법에 맞게 재정의 할 필요가 있다. <표 1>은 분석 방법에 따라 재정의 한 문장부호의 분류표이다.

<표 1> 분석 방법에 따른 문장부호의 분류

문장부호	분석 방법
마침표, 물음표, 느낌표, 쉼표, 가운뎃점, 쌍점, 빗금, 큰따옴표, 작은따옴표, 소괄호, 중괄호, 대괄호, 겹낫표, 겹화살괄호, 홑낫표, 홑화살괄호, 줄표, 붙임표, 물결표, 줄임표	문장부호 분리 알고리즘
드러냄표, 밑줄, 숨김표, 빠짐표	제외

<표 1>에서 정의한 문장부호가 한글과 혼용되어 사용될 경우 본 논문에서 제안하는 알고리즘으로 처리된다. 문장부호 중 드러냄표, 밑줄, 숨김표, 빠짐표는 형태소 분석에 필요한 기호가 아니므로 문장부호에서 제외한다.

3.2 문장부호를 고려한 특수어절 분석 알고리즘

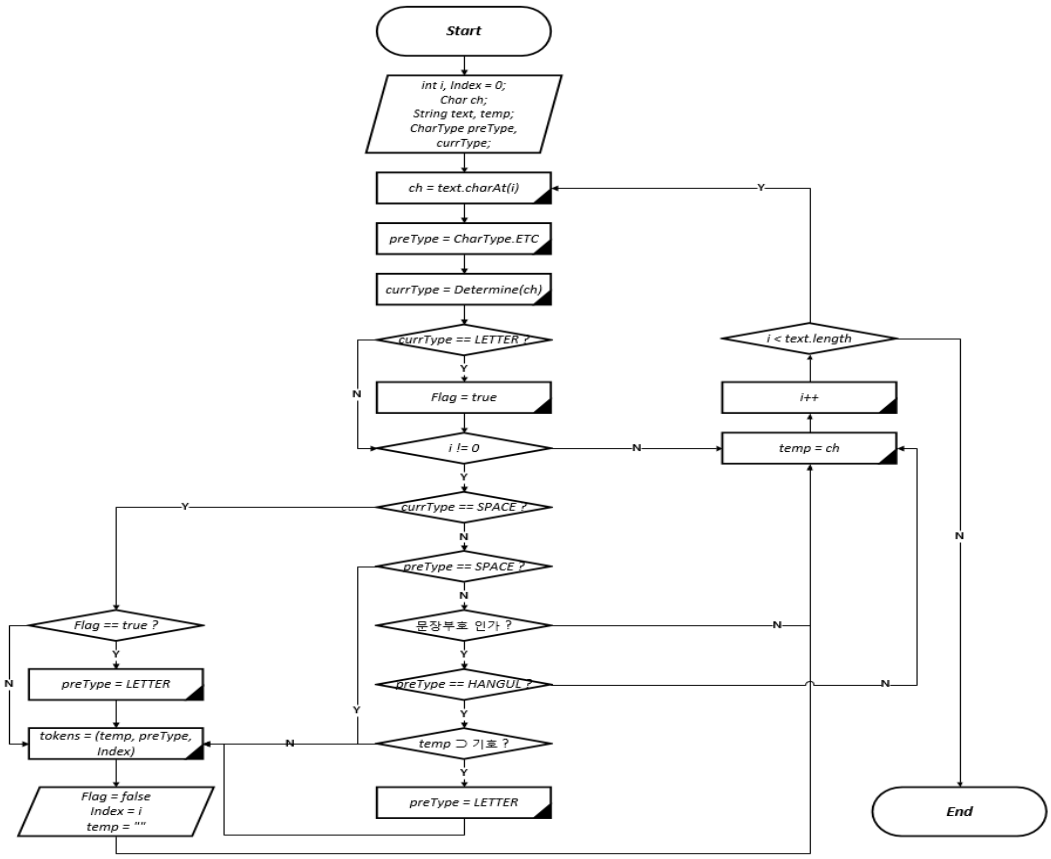
본 논문에서 제안하는 문장부호를 고려한 특수어절 분석 알고리즘은 (그림 2)와 같다. 문장부호를 분리하는 과정은 입력 받은 문장을 공백 기준으로 분리한 후 어절 내의 문장부호의 유무를 확인하여 다시 어절로 분리하는 과정이다.

입력 받은 문자열은 text 변수에 저장되고 문자열의 증감값을 위해 변수 i를 사용한다. CharType은 입력 받은 문자열의 타입으로 HANGUL, LETTER, SPACE, ETC로 나뉘며 첫 음절은 ETC로 지정한다. HANGUL은 한글로 이루어진 문자열, LETTER는 한글 이외의 문자 혹은, 한글과 문자가 혼용된 문자열, SPACE는 공백, ETC는 초기값을 의미한다. Determine() 함수에서 각 음절의 타입이 결정되며 각 음절은 ch 변수에 저장된 값을 이용한다.

첫 음절을 저장한 후 초기값을 preType 변수에 복사한다. 첫 음절에 대한 CharType을 확인하여 한글이 아닌 경우, Flag 값을 이용해 문자로 인식한다. Flag의 초기값은 false이고 문자가 인식된 경우 true로 변경된다. 첫 음절에 대한 CharType을 문자열 임시 저장 공간인 temp 변수에 저장한 후 다음 음절로 이동하여 위의 순서를 반복한다.

해당 음절이 공백일 경우, Flag 값을 확인하여 이전 값에 대한 한글 여부를 확인하고 공백을 토큰에 담아 다음 음절로 값을 변경한다. 공백이 아닌 경우, 이전 CharType에 대한 공백 여부를 확인하여 공백일 경우 토큰에 담고 아닐 경우 문장부호 여부를 확인한다. 문장부호가 인식된 경우 바로 이전 값에 대한 한글 여부를 확인해야 한다. 한글일 경우 현재까지 저장된 temp 값에 대해 한글 이외의 글자가 포함되었는지 확인한 후 토큰에 담는다. 문장부호가 인식되었지만 이전 글자가 한글이 아닌 경우, 다음 음절로 이동한다. 분석된 어절은 해당 문자열, 그 문자열의 타입과 위치값을 tokens에 담아 최종 어절을 생성한다.

1) http://www.korean.go.kr



(그림 2) 문장부호를 고려한 특수어절 분석 알고리즘

위의 동작을 문장이 끝날 때까지 반복하여 특수어절에 대한 새로운 토큰을 생성한다. 따라서 형태소 분석기에 본 논문에서 제안한 알고리즘을 적용하면 한글과 문장부호가 혼용된 어절도 정확한 형태소 분석이 가능해진다.

4. 실험 및 분석

4.1 실험자료

본 논문에서 제안한 문장부호를 고려한 특수어절 분석 알고리즘의 성능 평가를 위하여 인문/사회 논문 3336 어절, 자연/예체능 논문 3126 어절, 웹 게시판 및 SNS의 자유 게시판 글 3147 어절, 인터넷 기사 7191 어절을 선정하였다. 이 중 한글과 문장부호가 혼용된 어절은 총 16800 어절 중 3109 어절로 전체 데이터의 18.51%이다. 실험에 사용된 문장부호의 구성 비율은 <표 2>와 같다.

<표 2> 실험에 사용된 문장부호의 구성 비율

종류	총 어절 수 (어절)	문장부호 어절 수	구성비율 (%)
인문/사회 논문	3336	609	18.26
자연/예체능 논문	3126	537	17.18
웹 게시판 글	3147	724	23.01
인터넷 기사	7191	1239	17.23
합 계	16800	3109	•

4.2 실험결과

총 16800 어절에 대해 본 논문에서 제안한 알고리즘을 적용하여 형태소 분석을 한 결과 평균 97.24%의 분석 정확도를 얻었으며 결과는 <표 3>과 같다. 성능비교를 위해 같은 실험 데이터를 이용하여 알고리즘을 적용하지 않고 테스트한 결과 평균 81.49%의 분석 정확도를 얻었다.

정확도는 각 분야별 총 어절 수에 대해 정확하게 분석된 어절에 대한 비율을 의미한다.

$$\text{정확도}(\%) = \frac{\text{분석된 어절 수}}{\text{총 어절 수}} \times 100$$

본 논문에서 제안한 알고리즘을 적용하지 않고 형태소 분석을 할 경우, 한글과 문장부호가 혼용되어 있는 어절에 대한 형태소 분석이 정확하게 이루어지지 않았다. 이는 특수어절의 경우 한글 형태소 분석의 유형으로 처리되지 않았기 때문이다. 결과를 살펴보면, 인문/사회 논문 81.74%, 자연/예체능 논문 82.82%, 웹 게시판 글 76.99%, 인터넷 기사 82.77%로 나타났다.

제안한 알고리즘을 적용하여 테스트한 결과는 인문/사회 논문 96.88%, 자연/예체능 논문 96.80%, 웹 게시판 글 96.92%, 인터넷 기사 97.73%의 정확도를 확인할 수 있었다. 분석되지 않은 부분은 다수의 경우 형태소 사전에 등록되어 있지 않은 단어와 문장부호가 혼용되어 나온 결과이고 그 외, 오타와 문법오류로 인해 발생하였다.

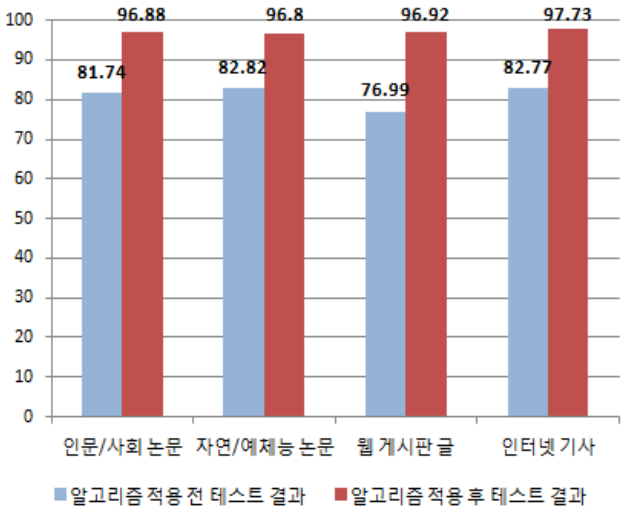
<표 3> 문장부호를 고려한 형태소 분석 정확도

종류	분야	어절 수	정확도(%)	
			적용 전	적용 후
인문/사회 논문	인류학	45	87.29	97.18
	문학	87	77.52	97.42
	철학	54	80.22	96.34
	역사	57	84.43	97.27
	심리	75	74.23	96.56
	교육	27	92.24	97.13
	사회	84	78.79	97.47
	회계/경영	33	90.18	97.02
	경제	33	87.50	96.21
	정치	114	64.49	95.64
평 균		•	**81.74**	**96.88**
자연/ 예체능 논문	과학	78	75.93	96.91
	수학	57	83.62	97.13
	화학	93	75.00	97.31
	물리	21	92.55	96.45
	생물	99	72.50	97.22
	식품	48	81.40	96.12
	공학	27	89.89	96.25
	IT	27	90.43	96.45
	미술	45	86.49	97.00
	체육	42	86.00	96.67
평 균		•	**82.82**	**96.80**
게시판 글	자유게시판	724	76.99	96.92
평 균		•	**76.99**	**96.92**
인터넷 기사	정치	162	79.39	97.96
	경제	141	80.58	96.41
	군사	57	92.49	97.50
	외교	135	81.78	98.65
	사회	165	78.17	97.88
	교육	141	80.42	97.09
	문화	117	83.88	98.62
	과학	99	86.08	98.45
	스포츠	90	86.05	97.67
	국제	132	78.74	96.78
평 균		•	**82.77**	**97.73**

4.3 결과분석

(그림 3)은 본 논문에서 제안한 문장부호를 고려한 특수어절 분석 알고리즘의 적용 전과 적용 후에 대한 형태소 분석 결과를 도식한 것이다.

(그림 3)에서 보는 바와 같이, 알고리즘 적용 후 분석 정확도가 높게 나타남을 알 수 있다. 총 16800 어절에 대한 평균 정확도를 살펴보면 81.49%에서 97.24%로 본 논문에서 제안한 알고리즘이 적용될 경우 15.75%가 높게 나타났음을 알 수 있다. 평균 정확도는 테스트에 사용된 전체 데이터의 총 어절 수에 대해 정확하게 분석된 어절의 비율로 계산되었다. 특히, 웹 게시판 글의 형태소 분석 결과에 대한 정확도가 상당히 높아졌으며 이는 글쓴이가 한글과 문장부호가 혼용된 어절을 많이 사용했기 때문이다. 하지만 본 논문에서 제안한 알고리즘을 적용하면 글쓴이에 상관없이 비교적 고른 정확도를 보였다.



(그림 3) 형태소 분석 결과 비교

5. 결론

정보 검색 분야와 기계 번역 분야 등의 자연어 처리 시스템에서 형태소 분석은 필수적으로 요구된다. 글을 쓸 때 다양한 문장부호를 사용하는데 한글과 혼용되어 사용될 경우 한글과 붙여 적는 것이 원칙이지만 글쓴이에 따라 다양한 형태로 나타날 수 있다.

본 논문에서는 한글과 문장부호가 혼용된 특수어절을 분석할 수 있는 알고리즘을 제안하고 실험하였다. 실험 데이터는 무작위로 선정한 논문과 웹 게시판 글, 인터넷 기사에서 추출한 16800 어절을 사용하였다. 특수어절에 공백을 삽입하여 다시 어절로 분리해 내는 방식을 적용하여 한글과 문장부호를 별도로 분석할 수 있는 알고리즘을 제안하였으며 실험결과 평균 97.24%의 정확도를 보였다.

본 논문에서 제안한 알고리즘을 적용한 형태소 분석기는 향후 한국어 정보 처리 기술 서비스 창출의 기반이 될 것으로 기대한다.

참고문헌

[1] 이호준, “결합 범주 문법에 기반한 한국어 통합 형태소 분석”, 한국과학기술원(석사논문), 2003.

[2] 최호철, “특수분야 및 띄어쓰기 오류문서 분석을 개선한 형태소분석기의 구현”, 중앙대(석사논문), 2003.

[3] 강승식, “한국어 형태소 분석과 정보 검색”, 홍릉과학출판사, 2003, 08.

[4] 고영근, 남기심, “표준국어문법론”, 박이정, 2014.

[5] 이희승, 안병희, 한재영, “개정 한글맞춤법 강의”, 신구문화사, 2015.

[6] Seung Jae Lee, “Grammaticality of Morpheme-1: Passives and Causatives”, 2010.

[7] László Kovács, “Classification Method for Learning Morpheme Analysis”, Journal of Information Technology Research archive Volume 5 Issue 4, October 2012.