음절단위 결합범주문법을 이용한 한국어 문장의 자동 띄어쓰기*

이호준○ 박종철

한국과학기술원 전산학 전공 및 첨단정보기술 연구센터

{hojoon, park}@nlp.kaist.ac.kr

Word Segmentation for Korean with Syllable-Level Combinatory Categorial Grammar

Ho-Joon Lee○ Jong C. Park

Computer Science Division & AITrc, KAIST

요 약

한국어의 띄어쓰기 현상은 단어별로 정형화된 띄어쓰기를 하는 영어나 띄어쓰기가 발달하지 않은 중국어, 일본어와는 다르게 독특한 형태로 발전되어 왔다. 기존에는 부분적인 띄어쓰기 오류를 바로잡아주는 형태의 연구가 많이 진행되었지만 이제는 문자인식이나 음성인식 등의 연구와 결합하여 띄어쓰기가 완전히 무시된 문장의 띄어쓰기를 자동으로 처리하는 방법에 대한 연구가 활발히 진행 중이다. 본 논문에서는 한국어의 띄어쓰기 현상과 띄어쓰기 복원 방법에 대한 기존의 연구에 대해서 살펴보고 기존의 방법으로는 처리하기 힘들었던 형태를 음절단위 결합범주문법으로 설명한다.

1. 서론

급속한 컴퓨터 사용의 발달로 일상생활의 많은 부분이 컴퓨터와 함께 이루어지고 있다. 이러한 컴퓨터의 일반적인 사용으로 컴퓨터 기기가 소형화 되었고 컴퓨터가 소형화 되어감에 따라 입출력 장치도 기존의 타이핑 방식에서 벗어나 문자인식을 이용한 직접 입력방식이나 음성인식을 이용한 음성 입력방식 등의 형태로 변화되어 가고 있다. 기존의 키보드를 통한 입력에서는 입력된 결과를 기계가 잘못 이해하는 경우가 거의 발생하지 않지만 문자인식이나 음성인식을 통한 방법에서는 인식률이 완벽하지 않기 때문에 입력된 결과가 기계에 의해 잘못 이해되는 경우가 빈번히 발생한다. 특히 단어인식이 아닌 공백 정보가 들어 있는 문장인식의 경우 문자인식이나 음성인식의 특성상 공백 정보가 제대로 나타나지 않기 때문에 더욱 인식률이 떨어질 수 밖에 없다. 또 공백 정보가 인식되었다고 하더라도 문자인식이나 음성인식에서 인식된 공백 정보가 그대로 사용될 수 있는 올바른 정보라는 보장이 없기 때문에 인식된 공백 정보를 그대로 사용하기에는 위험이 따른다. 그래서 기존의 부분적인 띄어쓰기 오류를 교정해주는 방법에 대

* 본 연구는 첨단정보기술 연구센터를 통하여 한국과학재단의 지원을 받았음

한 연구에서 이제는 띄어쓰기가 무시된 채 들어오는 문장에 대해서도 띄어쓰기를 복원해주는 연구가 활발히 진행되고 있다.

본 연구에서는 공백 정보가 무시된 한국어 문장의 띄어쓰기 현상 처리 방법에 대한 기존의 연구에 대해서 알아보고 기존의 방법으로는 다루기 힘들었던 형태의 한국어 문장 띄어쓰기를 처리할 수 있는 방안을 제시하고자 한다. 또한 이러한 형태로 구현된 시스템이 문자인식이나 음성인식에서 후처리기로 사용될 수 있는 가능성과 문법검사기와 형태소 분석기로서의 발전 가능성을 보이고자 한다.

2. 관련연구

한국어는 중국어나 일본어에 비해서[1][2] 띄어쓰기 현상이 발달되어 왔기 때문에 띄어쓰기가 완전히 무시된 문장의 교정에 대한 연구보다는 실질적으로 자주 발생하는 2~3 어절에 걸친 부분적인 띄어쓰기 오류를 교정하는 연구들이 많이 진행되어 왔다.

그러나 음성인식과 문자인식이 단어인식 수준에서 벗어나 문장인식 수준으로 확장됨에 따라 글쓰기나 말하기의 특성상 문장에 띄어쓰기 정보가 생략되고 제대로 인식되지 못하는 경우나 잘못된 말하기나 글쓰기 습관, 잡음 등에 의해 띄어쓰기가 잘못 인식되는 경우에 대한 처리가 필요하게 되었다. 이 때 인식된 모든 공백 정보를 올바른 것이라고 볼 수 없기 때문에 이러한 불완전한 정보를 버리고 공백 정보가 없는 문장에서 공백을 올바르게 재현하는 연구가 활발히 진행되어 왔다[3][4][5][6][1]. 이들 연구는 대부분 통계 정보나 휴리스틱에 바탕을 두고 커다란 하나의 문장을 가능성이 가장 큰 부분부터 나누어가는 형태의 접근 방법을 사용하고 있는데 각각의 경우에 대해 좀 더 자세히 살펴보면 다음과 같다.

2.1. 통계 정보 기반 공백 복원

띄어쓰기가 완전히 무시된 문장에 대해서 통계적인 접근 방법을 시도한 연구[3][6]에서는 대량의 말뭉치로부터 인접한 두 음절 앞에 공백이 나타날 확률, 두 음절 뒤에 공백이 나타날 확률 그리고 두 음절 사이에 공백이 나타날 확률 값을 각각 구해서 문장의 자동 띄어쓰기를 실현하였다. 이들 연구에서는 쉽고 빠르게 얻을 수 있는 통계 정보를 가지고 띄어쓰기가 무시된 문장의 복구에 빠르게 적용할 수 있다는 것을 보여주었지만 띄어쓰기 결과가 학습데이터에 의존적이고 통계적인 오류가 발생할 경우 오류를 수정하기가 힘들다는 단점이 있다.

2.2. 어휘 지식과 휴리스틱 기반 공백 복원

공백 정보가 없는 문장에 대해서 어휘 지식과 휴리스틱을 바탕으로 접근을 한 연구[4][5]에서는 한국어에서 띄어쓰기와 밀접하게 관련이 있는 조사와 어미를 띄어쓰기 구분자로 설정하여 형태소 분석 방법을 통하여 자동 띄어쓰기를 실현하였다. 한국어 문장에서 조사와 어미의 출현 빈도수가 높고 조사와 어미로 사용되는 음절수가 일반적으로 자주 사용되는 음절 수에 비해 적기 때문에 조사나 어미가 띄어쓰기에서 좋은 구분자로 사용될 수 있다. 그래서 [4]에서는 조사와 어미를 어절 블록 구분자로 사용하고, 어절 블록 내에서는 최장 일치법을 사용하여 어절을 인식하였다. 이 연구에서는 어휘 지식에 바탕을 둔 휴리스틱을 띄어쓰기 문제에 적용하는 것이 형태론적인 측면에서 바람직한 방법이 될 수 있음을 보여주었으나 조사나 어미로 사용될 수 있는 음절이 2개 이상 나오는 경우의 오류 현상이나 최장 일치법에서의 전파오류(triggered errors) 등이 발생할 수 있다는 문제점이 있다. [5]에서는 이러한 단점을 보완하기 위해 재결합 단계를 두어서 각각의 단계에서 발생한 오류를 줄이고 있다. 그러나 어휘 지식에 바탕을 둔 휴리스틱의 사용으로 인한 오류가 나타날 수 있다는 단점이 남아있다.

[1] 문자인식이나 음성인식에서 어느 정도의 휴지 공간과 시간을 띄어쓰기로 보아야 하는지 결정하는데 어려움이 따른다.

3. 결합범주문법을 이용한 띄어쓰기 현상의 처리

3.1. 음절 단위 결합범주문법

결합범주문법은 범주문법에 결합자(combinator)가 추가된 것으로 소수의 축약 규칙에 의하여 구문분석이 이루어지는 어휘문법이다. 결합범주문법에서는 기존의 많은 문법체계에서는 하나로 합쳐질 수 없었던 문장 성분들이 하나로 결합 될 수 있기 때문에 병렬 구조 등의 복잡한 문형들을 별도의 제약 조건 없이 처리할 수 있다는 장점이 있다. 그리고 어휘 범주에 통사 정보 이외의 의미 정보, 담화 정보 등을 추가함으로써 여러 단계의 분석과정을 한 단계의 유도과정을 통해 해결할 수 있다는 특징을 가지고 있다 [7]. 한국어를 위해 제안된 축약 규칙을 살펴보면 <표 1>과 같다 [8].

축약 규칙	규칙 이름 (기호)
$X/Y \quad Y \rightarrow X$	Forward Application (>)
$Y \quad X\backslash Y \rightarrow X$	Backward Application (<)
$X/Y \quad Y/Z \rightarrow X/Z$	Forward Composition (>B)
$Y\backslash Z \quad X\backslash Y \rightarrow X\backslash Z$	Backward Composition (<B)
$X \rightarrow T/(T\backslash X)$	Forward Type Raising (>T)
$X \rightarrow T\backslash(T/X)$	Backward Type Raising (<T)
$X \quad conj \quad X \rightarrow X$	Coordination ($<\phi^n>$)
$X/Y \quad Y\backslash Z \rightarrow X\backslash Z$	Forward Crossed Composition ($>B_x$)

<표 1> 한국어를 위해 제안된 축약 규칙

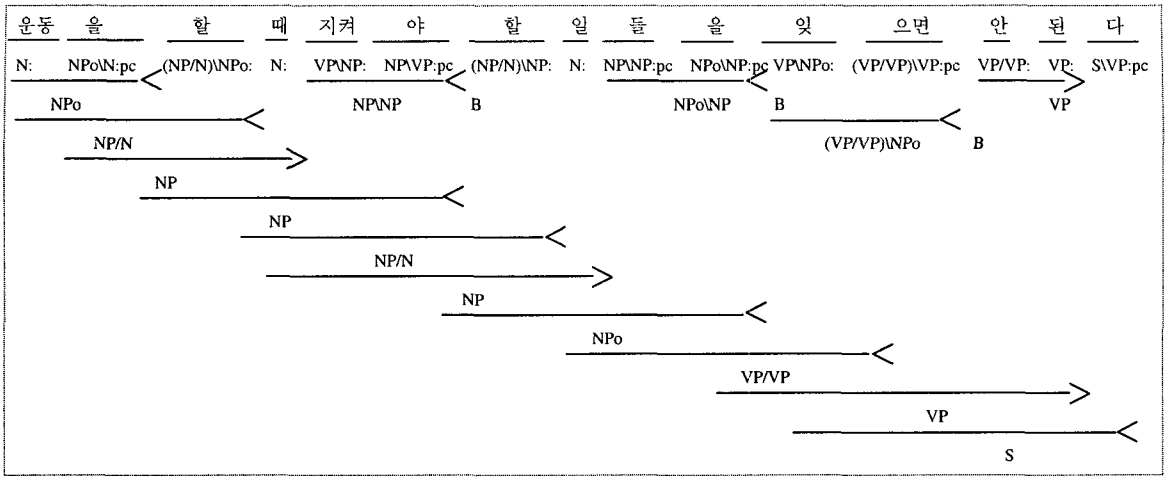
결합범주문법에서 범주를 어디에 할당하느냐에 따라 음소 수준, 형태소 수준, 음절 수준 등으로 세분화된 결합범주문법을 생각해 볼 수 있다. 복잡한 언어 현상의 통사적, 의미적, 화용적, 담화적인 분석을 한 단계의 유도 과정을 통해 처리하기 위해서는 형태소 수준의 결합범주문법을 통한 분석이 가장 좋은 방법이 될 것이다. 그러나 띄어쓰기 현상이 음절 단위에서 발생하는 언어 현상이고 용언의 어간과 어미에서 나타나는 음소 수준의 결합 과정을 제외하고는 음절 수준에서 말하고자 하는 바와 형태소 수준에서 말하고자 하는 바가 동일하기 때문에 본 연구에서는 음절 수준의 결합범주문법을 제안한다. 일부 불규칙 변형 형태의 어간이나 음소 단위로 결합을 하는 어간과 어미 등을 처리하기 위해서는 모든 변형된 형태의 어휘를 어휘 사전에 추가하거나 형태소 수준으로 어휘 사전을 구현하여 변형 형태를 규칙으로 처리해야 하는데 여기서는 우선 변형된 형태를 어휘 사전에 추가하는 방법을 사용하고자 한다. 형태소 수준의 어휘 사전을 구축하고 변형 형태를 규칙으로 처리하면 규칙에 의해서 어휘가 과생성 될 위험이 있는데, 띄어쓰기를 처리하는데 있어서는 음소 수준의 결합이 띄어쓰기에 영향을 미치지 않으므로 본 논문에서는 음절단위로 변형 형태의 어휘를 어휘 사전에 추가해주는 방법을 사용하고 형태소 수준의 어휘 사전과 규칙을 이용한 처리는 향후 계획으로 형태소 분석기에서 다루고자 한다.

한국어 띄어쓰기를 위한 음절단위 결합범주문법에서는 결합범주문법의 특성상 가능한 모든 결합 과정을 결과로 줄 수 있기 때문에 복합명사나 전문용어처럼 여러 가지로 띄어쓰기가 가능한 형태나 여러 가지 의미로 해석이 되면 안 되는 법률 문서 등에 잘 적용될 수 있다. 또한 기존의 방법으로는 처리하기 힘들었던, 구문분석을 필요로 한 <표 2>와 같은 문장에 대해서 복잡한 단계의 분석 과정 없이 한 번의 유도 과정을 통해 올바른 띄어쓰기 방법을 제시할 수 있다는 장점이 있다.

1) 의사는 만성 골수성 백혈병 이라고 말했다.
2) 군은 중거리 탄도 유도탄을 개발했다고 발표했다.
3) 운동을 할 때 지켜야 할 일들을 잊으면 안 된다.
4) 고양이가 누나 가방에 들어간다.
5) 누나가 방에 들어간다.

<표 2> 기존의 방법으로 처리가 어려운 문장

기존의 방법으로는 통계 정보나 휴리스틱을 바탕으로 띄어쓰기를 하고 띄어진 결과를 형태소 분석을 통해 검증하는 수준으로 분석을 하였기 때문에 위의 예문을 제대로 처리하기 위해서는 구문 분석이나 의미 분석 등의 다른 단계의 분석 과정을 두어서 다시 처리를 해야 하는 복잡한 과정이 필요하였다. 음절단위 결합범주문법을 사용하면 다음 <그림 1>과 같이 한 단계의 유도과정으



<그림 1> 음절 단위 결합범주 문법을 통한 띄어쓰기 현상 처리

로 띄어쓰기를 복원할 수 있다. 여기에서는 통사정보를 바탕으로 구문분석을 하였지만 의미정보나 화용정보, 담화정보를 범주에 추가함으로써 쉽게 분석 수준을 확장할 수 있다. 본 연구에서는 <표 1>에서 밝힌 축약 규칙 중에서 forward type raising과 backward type raising을 제외한 6개의 규칙을 사용하였다.

3.2. 한글 맞춤법에 바탕을 둔 띄어쓰기 처리

문교부 고시 한글 맞춤법[9]에서는 띄어쓰기 현상에 대해서 10개의 항목을 들어 설명하고 있는데 이 중 몇 가지를 살펴보면 다음과 같다.

제41항 조사는 그 앞말에 붙여 쓴다.

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.

기존의 연구에서는 위의 정보를 휴리스틱으로 사용하거나 위의 정보가 통계 정보를 구하기 위한 말뭉치에 많이 나타날 것이라는 가정으로 무시하였었다. 그렇기 때문에 통계 정보만을 바탕으로 한 몇몇의 띄어쓰기 교정 시스템에서는 다음과 같은 형태의 교정 오류가 나타나기도 한다.

양심에의호소가적중했다 → 양심에 의 호소가 적중했다
고양이가누나가방에들어갔다 → 고양이가 누나가 방에 들어갔다

본 연구에서는 한글 맞춤법에서 논의하는 한국어의 띄어쓰기 현상이 다음의 세 가지로 표현 된다고 분석하고 세가지 형태의 자질(feature)을 범주에 할당하여 띄어쓰기 현상을 처리하고자 하였다.

1. 앞 음절과 결합하는 형태
2. 뒤 음절과 결합하는 형태
3. 앞, 뒤 음절과 결합하지 않는 형태

기본적으로 모든 음절은 앞, 뒤 음절과 띄어쓰기가 되어 있다고 가정하고 음절단위 결합범주문법을 통해 나온 결과에서 각 음절이 띄어 써야 하는지 혹은 붙여 써야 하는지에 대한 정보를 얻게 된다. 즉 음절단위 결합범주문법을 통해 어떤 어휘가 조사나 어미로 사용되었다는 것을 알게 되면 앞의 체언이나 용언의 어간과 붙여 쓰고, 접두사로 쓰였다는 것을 알게 되면 뒤에 나오는 어휘와 붙여 쓰는 것이다.

음절단위의 결합범주문법에서 위의 세 가지 형태에

해당하는 예제 어휘와 문장 성분, 할당된 범주 그리고 각 범주가 가지고 있는 자질을 살펴보면 <표 3>과 같다.

1. 앞 음절과 결합하는 형태	
예	의사 N: 는 **NPs\N:pc**
문장성분	주격 조사
범주	NPs\N:pc
자질	:pc (앞 음절과 결합)

2. 뒤 음절과 결합하는 형태	
예	**중 N/N:nc** 거리 N:
문장성분	접두사
범주	N/N:nc
자질	:nc (뒤 음절과 결합)

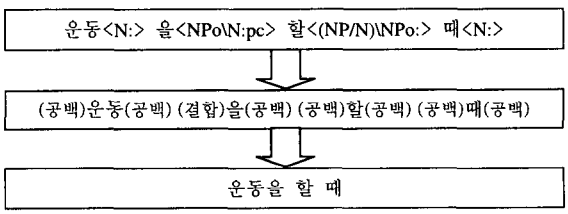
3. 앞, 뒤 음절과 결합하지 않는 형태	
예	**말했 VP\NPs:** 다 S\VP:pc
문장성분	용언 어간
범주	VP\NPs:
자질	: (결합 없음)

<표 3> 띄어쓰기 구조의 간단한 예

위 <표 3>에서 어휘에 할당된 범주는 문장성분과 밀접한 관계가 있고 할당된 자질도 문장성분과 밀접한 관계가 있다는 것을 알 수 있다.[2] 즉 문장성분과 범주, 자질에는 언어학적으로 동일한 정보가 다른 표현 방식으로 나타나 있다. 문장성분이나 범주, 자질은 언어학적인 분석을 통해 주어지기 때문에 통계정보나 휴리스틱에 비해서 더욱 직관적이고 논리적으로 공백 정보를 기술하고 있다. 위와 같은 정보를 바탕으로 <그림 1>의 일부에 대하여 띄어쓰기를 복원하면 <그림 2>와 같다.[3]

어휘에 범주를 할당하는 방법은 <표 3>에서처럼 일반적인 사전의 내용을 바탕으로 어휘의 문장 성분과 예문 등을 통해 많은 경우 자동으로 할 수 있다. 그러나 '굽어', '구워'와 같은 경우나 '잡혔다 (잡히+었다)'와 같은 형태소 단위가 아닌 음절 단위로 어휘 사전을 구축하기 위해서는 수동으로 처리를 해야 하는 경우도 발생하는데 이러한 반자동 형태의 어휘 사전 구축은 형태소 분석 연구를 통해 좀 더 자동화 될 수 있으리라 생각된다.

본 절에서는 한국어의 띄어쓰기 현상을 음절 단위 결합범주문법을 통하여 어떻게 설명할 수 있는지에 대해서 살펴보았다. 다음 절에서는 통계 정보나 휴리스틱만으로는 처리하기 힘들고 여러 단계의 분석 과정이 있어야만 해결 가능한 띄어쓰기 현상을 결합범주문법을 통해 어떻게 한 단계의 유도 과정만으로 처리할 수 있는지 살펴보도록 한다.

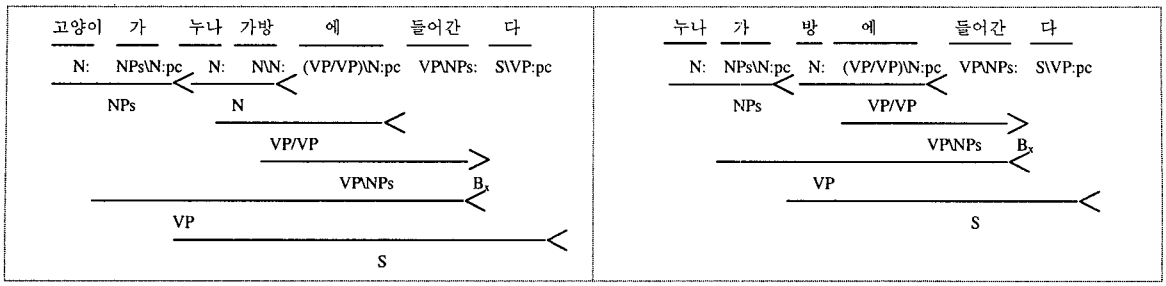


<그림 2> 공백 정보 복원

3.3. 복잡한 형태의 띄어쓰기 처리

<표 2>의 4번과 5번 예문은 4번 예문에 '고양이가'가 추가된 것 이외에는 모든 음절이 동일하지만 두 문장의 구조는 동일하지 않다. 4번 예문 '고양이가 누나 가방에 들어간다'에서는 '고양이'가 주어이고 '누나 가방에'는 장소를 나타내는 부사어로 쓰였지만 '고양이가'가 사라진 5번 예문에서는 '누나'가 주어로 쓰였고 '방에'는 장소를 나타내는 부사어로 쓰였다. 4번 예문을 *'고양이가 누나가 방에 들어간다'*로 본다면 '들어간다'의 주어가 두 개가 되므로 비문이 되고 5번 예문을 *'누나 가방에 들어간다'*로 본다면 '들어간다'의 주어가 없기 때문에 생략 현상을 제외한 경우 비문으로 처리 되어야 한다.

[2] '밥을 먹다가 딴 짓을 했다'에서 연결 어미 '가'로 인해 '밥을 먹다'가 문장 안에 들어가는데 이는 '가'가 '밥을 먹다'에 결합한다고 볼 수 있다.
[3] 복합명사에서 명사간에 붙여 쓸 때 (결합)운동(공백)의 형태로 나타낼 수도 있다.



<그림 3> 복잡한 형태의 띄어쓰기 현상 처리

기존의 통계 정보나 휴리스틱을 이용한 방법에서 이러한 문장을 올바르게 처리하기 위해서는 별도의 구문분석 단계가 필요했다. 구문분석은 띄어쓰기 복원이 이루어진 후에 적용되기 때문에 복잡한 문장의 경우 띄어쓰기 복원과 구문분석을 여러 번 거쳐야 하는 문제가 발생한다. 또한 구문분석을 어떻게 하느냐에 따라서 별도의 말뭉치를 비롯한 데이터를 필요로 하게 되고 통사적 분석 뿐만 아니라 의미적 분석까지 하려면 더 많은 사전 지식을 필요로 하게 되어 전체적인 처리 단계와 복잡도가 증가하게 된다. <표 2>의 예문 4와 5를 음절단위 결합범주문법을 사용하여 처리하는 과정을 보이면 <그림 3>과 같다.

음절단위 결합범주문법을 사용하여 띄어쓰기를 처리하면 기존의 방법과는 다르게 가능한 모든 처리 방법을 결과로 주게 되는데 이러한 방법은 대화형 띄어쓰기 검사 시스템이나 의미 변화에 민감한 법률 문서 등의 띄어쓰기 오류 검사 시스템 등에 효과적으로 사용될 수 있다. 하지만 일반적인 띄어쓰기 검사 시스템에서처럼 음절단위 결합범주문법을 사용하여 하나의 가장 올바른 띄어쓰기 형태를 결과로 찾기 위해서는 확률 등을 이용한 별도의 처리단계가 필요한 것으로 보인다.

그런데 이러한 확률적 처리과정은 가능한 띄어쓰기 형태 중에서 가장 알맞은 것을 선택한다는 점에서 기존의 확률적 처리과정과 의미가 다르다. 기존의 통계 정보나 휴리스틱을 이용한 방법에서 형태소 분석기를 사용해서 결과를 낼 때 첫번째로 나온 결과는 가장 올바른 띄어쓰기 결과이기보다는 가장 띄어쓰기 가능성이 높은 조합이 형태소 분석기를 통과한 결과이기 때문에 확률이 높다고 해서 가장 올바른 형태라고 보기는 어렵고 올바른 결과를 얻기 위해서는 또 다른 구문분석 단계의 이중 처리가 필요하다.

이번 절에서는 음절단위의 결합범주문법을 사용하여 복잡한 여러 단계의 처리 과정을 거치지 않고 한 단계의 유도과정을 통해 올바른 형태로 띄어쓰기를 복원할 수 있음을 보였다. 다음 절에서는 실제 구현된 시스템에 대해서 알아보도록 한다.

4. 시스템 구현 및 실험

음절단위 결합범주문법을 이용한 한국어 띄어쓰기 복원 시스템의 구조는 다음 <그림 4>와 같다.

입력으로 공백 정보가 완전히 무시된 문장이 들어오면 띄어쓰기 후보 검색 모듈에서 간단한 방법으로 띄어쓰기 후보를 찾아준다. 후보 검색이 이루어지면 문장을 파싱하고 그 결과를 바탕으로 후처리 모듈에서 입력 문장의 공백 정보를 복원하게 된다.

본 시스템은 Sparc SunOS 5.8 환경에서 Perl 5.0으로 구현되었다.

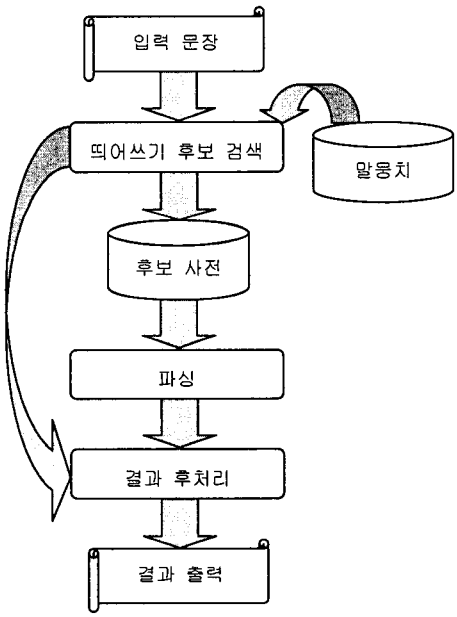
4.1 후보 검색 모듈

결합범주문법을 이용하여 시스템을 구현하면 어휘의 모든 가능한 범주에 대해서 파싱을 해야 하기 때문에 시간 복잡도가 증가할 수 있다. 특히 공백 정보가 무시된 형태를 파싱해야 하는 본 시스템에서는 파싱을 하다가 어휘사전에 어휘가 존재하지 않아서 도중에 파싱에 실패하는 경우가 많이 발생하기 때문에 최소한 모든 어

휘를 어휘사전에서 찾을 수 있을 때까지 파싱을 미루도록 하였다.

어절 단위의 띄어쓰기에서 정확하게 띄어진 어절은 다음 어절의 띄어쓰기에 영향을 미치지 않으므로 입력으로 들어온 띄어쓰기가 무시된 문장의 앞에서부터 사전에 존재하는 후보를 찾아나가는 아주 단순화된 형태의 검색 방법을 사용하였다.

일반적인 통계나 휴리스틱에서 사용하는 방법을 이 단계에 적용하여 확률이 높은 후보를 먼저 검색할 수 있지만 전처리 과정을 단순하게 하기 위해서 순차적인 검색을 하는 것으로 대신하였다.



<그림 4> 한국어 띄어쓰기 복원 시스템

4.2 파싱 모듈

띄어쓰기 후보 검색에서 찾아진 후보들을 CKY를 기반으로 한 CCG 파서에 제공하여 의미 있는 문장 후보들이 선택된 것인가를 확인한다. 제대로 의미 있는 문장 후보들이 선택된 경우라면 파싱 결과로 의미에 맞는 띄어쓰기 복원 정보가 나오게 되고 파싱이 실패로 끝난다면 다시 후보 검색 모듈에서 새로운 후보를 찾게 된다.

4.3 후처리 모듈

파싱의 결과를 가지고 후보로 찾아진 입력 문자열의 띄어쓰기 정보를 복원하는 모듈이다. 띄어쓰기 후보로 선택된 것이 앞 음절과 결합 하는지, 뒤 음절과 결합하는지 아니면 아무런 결합을 하지 않는지에 대한 정보를 가지고 실제 후보 어절에 띄어쓰기 처리를 한다. <그림 2>에서 두 번째 단계에서 세 번째 단계로 넘어가는 과정이다.

4.4 실험

지금까지 음절단위 결합범주문법을 사용하여 한국어의 띄어쓰기 현상을 처리하는 방법에 대해서 알아보았다. 기존의 많은 연구들이 통계정보나 휴리스틱을 사용하여 띄어쓰기 현상을 처리하였는데 음절단위 결합범주문법으로 일반적인 형태 뿐만 아니라 복잡한 형태의 띄어쓰기 현상을 설명하고 공백 정보를 복원할 수 있다는 것을 확인하였다. 본 연구에서 구현한 시스템으로 본 논문의 요약 부분을 처리한 예를 <그림 5>에서 보였고 <표 2>의 문장을 처리한 결과를 <그림 6>에 보였다.

실험 결과 10에서 20 음절의 문장을 처리할 때에는 1초 미만의 시간이 걸렸지만 201 음절의 문장에 대해서 처리를 할 때에는 38초의 시간이 걸려서 처리해야 할 음절이 길어질수록 처리 시간이 급속하게 증가하는 문제점을 보였다. 하지만 형태소 분석 단계와 구문 분석 단계를 거친 공백 정보 복원이기 때문에 공백 정보 복원 결과를 언어학적으로 설명할 수 있고 다른 단계의 분석 과정을 거치지 않고도 분석 결과를 문자 인식이나 음성 인식기의 후처리기로 사용 할 수 있다는 장점이 있다.

5. 결론 및 향후 과제

기존의 연구에서 결과의 정확성을 높이기 위한 후처리로 구문분석이나 의미분석을 하는 경우에 분석 단계에 따라 반복적으로 복잡도가 증가하는 방법이 아닌 한 단계의 유도 과정으로 필요한 분석이 가능하고 다른 처

리 없이 문법적으로 가장 올바른 순서로 결과를 찾을 수 있는 본 연구를 적용하면 좋은 결과를 낼 수 있으리라 생각된다. 또한 본 연구에 통계정보와 휴리스틱을 추가하여 더 나은 성능을 기대할 수 있다.

본 연구에서는 음절단위 결합범주문법을 사용하여 공백 정보가 무시된 문장에서 공백 정보를 복원하는 방법에 대해서 논의하였는데 공백 정보를 복원하면서 형태소 분석과 구문 분석이 이루어지고 있으므로 범주의 할당 수준을 형태소로 낮추고 형태소 수준의 처리에 필요한 언어학적 자질을 추가함으로써 형태소 분석기로 발전 할 수 있는 가능성이 있음을 확인할 수 있었다. 이러한 형태소 분석기는 이번 연구에서 나타난 것과 같이 복잡한 형태의 문장이라도 동일한 수준에서 처리될 수 있고 여러 단계의 분석 과정이 한 단계의 유도 과정으로 처리 될 수 있는 장점을 가지고 있어서 공기정보 뿐만 아니라 구문정보나 의미정보를 동시에 사용하는 형태소 분석기로서의 가능성을 제시해주고 있다. 또한 앞에서 논의한 것과 같이 본 처리기는 문자인식과 음성인식의 후처리기나 문법검사기의 형태로 활용이 가능하다고 생각한다.

<입력문장>
한국어의띄어쓰기현상은단어별로정형화된띄어쓰기를하는영어나띄어쓰기가발달하지않은중국어,,일본어와는다르게독특한형태로발전되어왔다
기존에는부분적인띄어쓰기오류를바로잡아주는형태의연구가많이진행되었지만이제는문자인식이나음성인식등의연구와결합하여띄어쓰기가완전히무시된문장의띄어쓰기를자동으로처리하는방법에대한연구가활발히진행중이다
본논문에서는한국어의띄어쓰기현상과띄어쓰기복원방법에대한기존의연구에대해서살펴보고기존의방법으로는처리하기힘들었던형태를음절단위결합범주문법으로설명한다

<띄어쓰기 결과>
한국어의 띄어쓰기 현상은 단어별로 정형화된 띄어쓰기를 하는 영어나 띄어쓰기가 발달하지 않은 중국어,, 일본어와는 다르게 독특한 형태로 발전되어 왔다

기존에는 부분적인 띄어쓰기 오류를 바로잡아주는 형태의 연구가 많이 진행되었지만 이제는 문자 인식이나 음성 인식 등의 연구와 결합하여 띄어쓰기가 완전히 무시된 문장의 띄어쓰기를 자동으로 처리하는 방법에 대한 연구가 활발히 진행 중이다

본 논문에서는 한국어의 띄어쓰기 현상과 띄어쓰기 복원 방법에 대한 기존의 연구에 대해서 살펴보고 기존의 방법으로는 처리하기 힘들었던 형태를 음절 단위 결합 범주 문법으로 설명한다

<그림 5> 본 논문의 요약 처리 결과

<입력문장>
의사는만성골수성백혈병이라고말했다
군은중거리탄도유도탄을개발했다고발표했다
운동을할때지켜야할일들을잊으면안된다
고양이가누나가방에들어간다
누나가방에들어간다

<띄어쓰기 결과>
의사는 만성 골수성 백혈병이라고 말했다

군은 중거리 탄도 유도탄을 개발했다고 발표했다

운동을 할 때 지켜야 할 일들을 잊으면 안된다

고양이가 누나 가방에 들어간다

누나가 방에 들어간다

<그림 6> <표 2>의 문장 처리 결과

참고 문헌

[1] Chen K. J. and Liu S. H., Word Identification for Mandarin Chinese Sentences, *Proceedings of the 14th International Conference on Computational Linguistics*, pp.101-107, 1992.

[2] Nobesawa S., et al., Segmenting a Sentence into Morphemes using Statistic Information between Words, *Proceedings of the 15th International Conference on Computational Linguistics*, pp.227-233, 1994.

[3] 심광섭, "음절간 상호 정보를 이용한 한국어 자동 띄어쓰기", 정보과학회논문지, Vol.23 No.9, pp.991-1000, 1996.

[4] 강승식, "한글 문장의 자동 띄어쓰기", 한글 및 한국어 정보처리 학술대회, pp.137-142, 1998.

[5] 김계성, 이현주, 이상조, "연속 음절 문장에 대한 3단계 한국어 띄어쓰기 시스템", 정보과학회논문지, Vol.25 No.12, pp.1838-1843, 1998.

[6] 강승식, "음절 bigram 특성을 이용한 띄어쓰기 오류의 인식", 제12회 한글 및 한국어 정보처리 학술대회, pp.85-88, 2000.

[7] Steedman, Mark, 2000, The Syntactic Process. The MIT Press.

[8] 조형준, 박종철, "한국어 병렬문의 통사, 의미, 문맥 분석을 위한 결합범주문법", 정보과학회논문지, pp.448-462, 2000.

[9] 한글 맞춤법, 문교부 고시, 1988.