# MicroCCG를 이용한 용언의 불규칙 활용의 처리와 한국어 형태소 분석[1]

이호준[0] 박종철
한국과학기술원 전산학 전공 및 첨단정보기술 연구센터
{hojoon[0], park}@nlp.kaist.ac.kr

# Morphological Analysis of Irregular Conjugation in Korean with Micro Combinatory Categorial Grammar

Ho-Joon Lee[0] Jong C. Park
Computer Science Division & AITrc, KAIST

요 약

본 논문에서는 형태소 수준의 결합범주문법을 이용하여 형태소 분석을 포함한 자연언어처리의 여러 단계를 한 단계의 유도과정으로 처리하고 형태소 분석 단계에서 증가하는 애매성과 복잡도를 상위 분석 단계의 정보를 사용하여 줄이는 방법에 대해서 논한다. 한국어에서 나타나는 복잡한 언어 현상 중에 하나인 용언의 불규칙 활용을 확률 정보 뿐만 아니라 음운정보를 포함한 통사 정보나 의미 정보 등의 상위 정보를 사용하여 처리하여보고 일반적인 형태소 분석기로서의 발전 가능성에 대해서 알아본다.

## 1. 서론

인터넷의 발달과 컴퓨터의 사용으로 인간이 활용할 수 있는 정보의 양이 하루가 다르게 늘어가고 있다. 이러한 정보의 대부분은 전자 문서의 형태로 존재하는데 양적으로나 질적으로 인간의 처리 능력을 넘어서고 있는 것이 현실이다. 이러한 방대한 양의 정보를 효과적으로 활용하기 위해서, 그리고 음성인식과 음성합성 등 인간 생활에 편리를 제공하기 위한 기술들이 지속적으로 개발되어가면서 그 바탕이라 할 수 있는 자연언어처리에 대한 중요성이 점점 더 부각되고 있다.

자연언어처리는 일반적으로 형태소 분석, 통사적 분석, 의미적 분석, 화용적 분석, 담화적 분석 등의 단계를 순차적으로 수행하면서 이루어지는데 이 중 형태소 분석은 자연언어처리 단계에서 가장 기초가 되는 과정이라 할 수 있으며 그렇기 때문에 가장 중요한 분석 과정 중에 하나라고 볼 수 있다. 일반적으로 형태소 분석이 이루어진 결과를 이용하여 통사적 분석, 의미적 분석 등의 상위 분석이 이루어지기 때문에 지금까지 형태소 분석은 주로 통계나 확률적인 방법 등을 사용하여 왔다. 현재까지 형태소 분석에 대한 많은 연구 결과보다 빠르고 정확하게 형태소 분석을 할 수 있는 방법에 대해서 많은 연구가 진행되었지만 통사나 의미 정보를 사용하여야 해결이 가능한 형태소 분석에서의 애매성으로 인한 복잡도의 증가는 순차적인 분석 단계를 거치는 방법으로는 처리하기 어려운 문제로 인식되고 있다. 본 논문에서는 한 단계의 유도과정으로 여러 단계의 분석과정을 처리할 수 있는 결합범주문법을 이용하여 형태소 분석 단계에서 상위 단계의 정보를 사용하여 애매성의 증가로 발생하는 복잡도를 효과적으로 줄이고 형태소 분석기로서의 발전 가능성을 알아보고자 한다.

## 2. 관련연구

### 2.1. 형태소 분석 방법

형태소 분석이란 주어진 입력문에서 의미 있는 최소 단위인 형태소를 추출하는 과정으로 변형이 일어나지 않는 어간부(Head)와 변형이 일어나는 어미부(Tail)로 분리하는 Head-Tail 구분법[1]과 한국어의 음절 특성에 따라 문법 형태소의 경계와 형태론적 변형 음절을 효율적으로 인식하는 음절 단위 분석법[2] 등이 제시되었다. 형태소의 원형을 복원하는 방법으로는 규칙 학습에 의해 형태론적 변형 규칙을 인식하는 방법[3]과 변형이 일어난 어간을 사전에 수록하여 접속 정보를 이용하는 방법[4], two-level 모델을 한국어에 적용 시키는 방법[5] 등이 있고 형태소들의 조합을 구하는 방법으로는 최장, 최단 일치법[6]과 tabular 파싱법[7] 등이 제시되었다. 이 외에도 한국어의 언어적 특성이 고려된 형태소 분석 방법들에 대한 많은 연구가 이루어져 왔다[8].

[1] 본 연구는 첨단정보기술 연구센터를 통하여 한국과학재단의 지원을 받았음

### 2.2. 결합범주문법

결합범주문법은 범주문법에 결합자(combinator)가 추가된 형태로 소수의 축약 규칙에 의하여 구문분석이 이루어지는 어휘 문법이다. 단일화 기반의 결합범주문법은 각 어휘마다 문법, 의미, 담화 정보를 담는 범주가 할당되는데, 특별한 약정 없이 축약규칙만을 통해 문법 정보 외에 의미 정보나 담화 정보까지 한번의 과정으로 유도할 수 있다는 장점을 지닌다[9, 10]. 결합범주문법을 이용하여 공백정보가 없는 문장에서 띄어쓰기를 복원하는 방법에 대한 연구[11]에서는 형태소 수준으로 범주를 할당하여 형태소 분석을 포함한 여러 분석 과정을 한 단계의 유도 과정을 통해 처리할 수 있음을 보였다.

## 3. 한국어 용언의 불규칙 현상과 처리

### 3.1. 한국어 용언의 불규칙 현상

용언이 활용 할 때 어간과 어미의 기본 형태가 유지되거나 달라진다고 하더라도 일정한 규칙으로 설명이 가능한 경우를 규칙 활용이라 하고 활용을 할 때 어간과 어미의 기본 형태가 유지되지 않고 그 현상을 일정한 규칙으로 설명 할 수 없는 경우를 불규칙 활용이라고 한다. 불규칙 활용은 그 활용 부분에 따라 어간의 불규칙, 어미의 불규칙 그리고 어간과 어미의 불규칙으로 나누어 볼 수 있는데 이러한 불규칙 현상에서도 특징적인 규칙을 찾아볼 수 있다. 이러한 규칙으로는 당연히 불규칙 활용을 전부 처리 할 수 없기 때문에 많은 연구에서 불규칙 현상이 일어나는 환경과 그 규칙을 찾아내어 처리하는 동시에 사전에 변형형태를 등록하여 처리하는 방법을 사용하고 있다. 이러한 불규칙 활용이라는 음운적 현상에서 단순한 음운적 결합 현상 이상의 정보를 추출할 수 있는데 이러한 정보를 이용하여 형태소 분석 단계에서 발생하는 복잡도를 줄이는 방안을 제안한다.

(ㄱ) 구운 고기
(ㄴ) 굽은 고기
(ㄷ) 구운 고기를 먹다

(ㄱ)과 (ㄴ)에서 '구'와 '굽'은 둘 다 원형이 '굽'인 용언이라고 생각할 수 있는데 (ㄱ)의 경우 불규칙 현상으로 인한 음운 현상을 처리하는 과정에서 '은'의 품사를 어미로 제한할 수 있고 어미 중에서 관형사형 어미로 생각한 경우라면 뒤의 '고기'를 명사의 형태로 제한할 수 있다. (ㄷ)의 경우에는 '고기'를 명사로 보았을 때 '를'이 조사로 제한될 수 있고 '를'을 목적격 조사로 생각한 경우 '먹다'를 타동사의 형태로 제한할 수 있다. 또한 반대로 (ㄷ)의 경우 '먹다'를 타동사로 보았을 때 그 앞의 '를'을 목적격 조사의 형태로 제한할 수 있다. 이와 같이 용언의 불규칙 활용에 의한 음소 수준의 결합과 각 어휘의 통사적 정보를 바탕으로 형태소 분석 단계에서의 복잡도를 줄일 수 있다.

### 3.2. 결합범주문법을 이용한 처리

용언의 불규칙 현상 중에서 어간이 변화하는 형태인 ㅂ불규칙 현상은 어간의 말음인 'ㅂ'이 '-어' 나 '-어' 로 시작되는 어미, 그리고 매개 모음을 요구하는 어미 앞에서 '오/우'로 변하는 것으로, 단음절 어간을 가진 '돕-', '곱-'일 때만 '오'로 되고 나머지는 '우'로 변하는 음운 현상이다.

ㅂ불규칙 현상을 결합범주문법으로 처리하기 위해서 형태소 수준의 어휘에 KAIST Tag Set이 혼합된 형태의 범주를 할당하는데 이는 기존 언어 자료와의 호환성을 유지하면서 세밀한 형태의 범주를 할당하여 일반적인 범주를 할당하였을 때 나타날 수 있는 애매성의 증가를 줄이기 위해서 이다. 예를 들어 일반적인 방법으로 일반동사(pvg)와 성상형용사(paa)에 동일한 s\np의 범주를 할당하면 일반동사와 성상형용사를 구분하기가 어려워지고 둘 사이에 애매성이 증가할 위험이 있다. 다음은 ㅂ불규칙 형태의 한 예인 '구운'에 대해서 KAIST Tag Set이 혼합된 형태의 범주를 할당한 예이다.

| 구 | 운 |
|---|---|
| pvg(ㅂ불규칙)굽' | np/np\pvg:은' |

'구운'과 동일한 원형을 가지지만 규칙 활용 형태인 '굽은'에 대한 예는 다음과 같다.

| 굽 | 은 |
|---|---|
| paa:굽' | np/np\paa:은' |

일단 '구운'에서 '구'가 '굽'의 ㅂ불규칙 형태라면 '은'이 가질 수 있는 범주는 어간과 결합할 수 있는 어미로 제한되므로 다음과 같이 제한 규칙을 표현해줌으로써 결합범주문법의 높은 표현력(expression power)으로 생길 수 있는 복잡도의 증가를 막을 수 있다.

| 구 | 운 |
|---|---|
| pvg:(ㅂ불규칙)굽',[pe] | np/np\pvg:은',[pe] |

| 굽 | 은 |
|---|---|
| paa:굽',[pe] | np/np\paa:은',[pe] |

또한 KAIST Tag Set을 사용하여 범주를 할당하기 때문에 나타나는 범주의 중복 할당을 피하기 위해서 할당된 범주가 계층구조를 가질 수 있게 다음과 같이 처리하여 중복된 역할을 하는 범주가 할당되는 것을 방지하였다. 아래의 예에서 보는 바와 같이 '은'에 'np/np\p'의 범주를 주고 'p'와 'pvg', 'paa' 등의 사이에서 호환성이 있도록 처리하였다.

| 구 | 운 |
|---|---|
| pvg:(ㅂ불규칙)굽',[pe] | np/np\p:은',[pe] |

| 굽 | 은 |
|---|---|
| paa:굽',[pe] | np/np\p:은',[pe] |

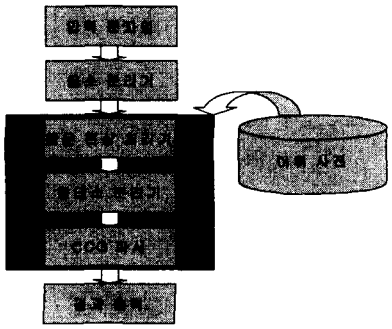
## 4. 구현 및 결과

형태소 단위로 범주를 할당하고 처리하기 위해서 입력 문자열을 음소 수준으로 분리해주는 단계는 다음과 같이 유니코드에서의 한글 코드 표현 알고리즘을 역으로 이용하여 java로 구현하였다.

```
코드값 = 0xAC00 + 초성 * 21 * 28
                + 중성 * 28
                + 종성
```

<그림 1> 한글 코드의 유니코드 값

음소 단위로 분리된 문자열에서 음운 현상을 처리하고 형태소 후보를 찾는 부분은 perl을 이용한 패턴 매칭을 통해 구현하였고 SICStus prolog로 짜여진 CKY 형식의 파서를 이용하여 통사 구조를 검사한다. 전체적인 시스템의 구조는 다음과 같다.



<그림 2> 시스템 구조

대표적인 불규칙 활용 형태라고 볼 수 있는 '굽다'의 ㅂ 불규칙 현상을 처리한 결과는 다음과 같다.

| | Baseline | MicroCCG | 축소율 |
|---|---|---|---|
| 굽은 | 32 | 7 | 78.1% |
| 구운 | 386 | 36 | 90.7% |

<표 1> 실험 결과

실험에서는 하나의 형태소 당 약 7개의 범주가 할당된 사전을 사용하였고 <표 1>에서 보인 바와 같이 음운 현상과 제한 규칙을 이용하여 불필요한 처리 과정이 효과적으로 감소되는 것을 볼 수 있었다.

MicroCCG의 처리 이후에 남아 있는 오분석의 대부분은 복합명사 형태로 오분석 된 경우와 복수개의 어미의 결합 형태로 오분석 된 것이었는데 이와 같은 오분석으로 인한 애매성의 해결을 위해서는 좀 더 세밀한 형태의 제한 규칙 설정과 의미 분석이나 담화 분석 등을 이용한 처리가 필요하다.

## 5. 결론 및 향후 과제

지금까지 음소 수준의 음운 현상과 통사 정보 등을 이용하여 형태소 분석 과정에서 발생하는 애매성과 복잡도를 줄이는 방법에 대해서 논하였다. 본 논문에서는 통사 정보를 바탕으로 용언의 활용에서 나타나는 음운 현상을 이용하여 형태소 분석에서의 애매성을 줄이는 방법을 제안하였는데 '날다'와 '나다'의 동일한 품사를 가지는 형태로 분석이 가능한 '나는'의 경우 애매성을 해결하기 위해서는 의미 분석이 필요하고, 지시 및 인칭 대명사의 경우에는 담화 분석이 필요한데 형태소 분석 단계에서 의미 정보나 담화 정보를 이용한 이러한 처리도 MicroCCG를 이용한 형태소 분석의 방법으로 처리할 수 있을 것으로 예상한다. 확률 정보를 바탕으로 가중치를 부여하여 시스템의 성능을 높이고 미등록어 처리 방안과 통사 정보를 포함하는 상위 정보를 더욱 효율적으로 이용하는 방안을 보완하면 효과적인 형태소 분석기로써 활용이 가능하다고 생각한다.

## 참고 문헌


[1] 최형석, 이주근, " 자연어 어절 처리 알고리즘" , 한국정보과학회 추계 학술발표회 논문집, 11권 2호, 1984.

[2] 강승식, " 음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석" , 서울대학교 컴퓨터 공학과 박사학위 논문, 1993.

[3] 장병탁, 김영택, " 다중언어 형태소 분석 및 합성을 위한 언어 규칙의 기계학습" , 한국정보과학회 논문지, 17권 4호, pp. 463-474, 1990.

[4] H. C. Kwon, Y. S. Chae, and G. O. Jeong, A Dictionary-based Morphological Analysis, *NLPRS*, pp .97-91, 1991.

[5] D. B. Kim, S. J. Lee, K. S. Choi, G. C. Kim, A Two-level Morphological Analysis of Korean, *COLING-94*, Vol. 1, pp. 535-539, 1994.

[6] 김덕봉, 최기선, 강재우, " 한국어 형태소 처리와 사전 -접속 정보를 이용한 한글 철자 및 띄어쓰기 검사기- " , 어학연구, 26권 1호, pp. 87-113, 1990.

[7] 김성용, 최기선, 김길창, " *Tabular Parsing* 방법과 접속 정보를 이용한 한국어 형태소 분석기" , 한국정보과학회 인공지능연구회 춘계 인공지능학술발표회 논문집, pp. 133-147, 1987.

[8] 강승식, " 한국어의 형태론적 특성과 형태소 분석 기법" , 정보과학회논문지, 12권 8호, pp. 47-59, 1994.

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

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

[11] 이호준, 박종철, " 음절단위 결합범주문법을 이용한 한국어 문장의 자동 띄어쓰기" , 제 14회 한글 및 한국어 정보처리 학술대회, pp. 47-54, 2002.