# 언어지식을 이용한 형태소 해석의 모호성 축소*

김재훈†○ 김길창‡

† 한국전자통신연구소 음성언어연구실

‡ 한국과학기술원, 전산학과

# Morphological Ambiguity Reduction Using Linguistic Knowledge

Jae-Hoon Kim†○ Gil Chang Kim‡

†Spoken Language Processing Section, ETRI,

‡Department of Computer Science, KAIST

**요 약**

가능한 모든 형태소 해석을 찾아내는 한국어 형태소 해석기들은 필요 이상으로 많은 수의 형태소 해석 결과를 생성하기 때문에, 자연언어 처리 시스템의 상위 과정, 즉 구문해석, 의미해석 등에 큰 도움이 되지 못하고 있는 실정이다. 이러한 문제점을 해결하기 위해서, 본 논문에서는 어휘화된 배열규칙과 형태적 포섭관계와 같은 언어지식을 이용해서, 형태소 해석의 모호성 축소 방법을 제안하고자 한다.

실험 및 평가를 위해서 KAIST 말뭉치를 이용하였으며, 평가의 기준을 설정하기 위해서 품사 쌍의 접속정보를 배열규칙으로 하는 한국어 형태소 해석기를 사용하였다. 어휘화된 배열규칙과 형태적 포섭관계를 이용했을 경우, 각각 54%와 40.4%의 형태소 해석의 모호성 감소율을 보였으며, 이들 두 방법을 동시에 적용했을 경우, 67.5%의 형태소 해석의 모호성 감소율을 보였다.

## 1. 서론

한국어 형태소 해석의 목적은 주어진 한국어 어절에 대한 가능한 모든 형태소 해석을 찾아내는 데 있다. 이와 같은 목적으로 말미암아 '소나무라고?'라는 어절의 형태소 해석 수는 무려 51개나 된다. 이 결과는 자연언어 처리의 상위 과정(품사 태깅, 구문 해석 등)에 아무런 도움을 주지 못할 것이다. 더구나, 이와 같은 목적은 품사 태깅을 비롯하여 대부분의 자연언어 처리 시스템의 목적과는 상충되는 것이다. 따라서, 자연언어 처리의 첫번째 단계인 형태소 해석에서 이와 같은 모호성을 축소 혹은 해소시키지 않을 경우에는 상위 과정에서의 해석이 거의 불가능하게 될 것이다. 본 논문에서는 한국어 형태소 해석의 모호성을 줄이는 방법으로 어휘화된 형태소 배열규칙과 단어의 형성 원리를 모형화한 형태적 포섭관계를 이용한다.

## 2. 어휘화된 배열규칙을 통한 형태적 모호성 축소

많은 한국어 형태소 해석기는 (형태소) 배열규칙(morphotatics)으로 (형태소) 접속정보(connectivity information of morphemes)를 사용한다[1, 5, 7]. 형태소 과잉 해석(morphological over-analysis)의 원인 중 하나는 비교적 단순한 형태소 배열규칙을 사용하기 때문에 발생된다. 일반적으로 명사 접미사(xn)는 거의 모든 명사류와 결합이 가능하다[1]. 따라서, 접속정보에는 <n*, xn>[2]과 같은 관계를 가지고 있다. 그러나, 명사 접미사 '-씨'는 대부분의 명사와 결합 가능한 것이 아니라, 주로 고유명사와 결합된다. 이와 같이 접미어나 기능어의 개별적인 단어의 성질에 따라서, 결합 가능한 품사나 단어가 제한된다. 품사만으로 구성된 접속정보는 이와 같은 단어의 개별적인 성질을 표현할 수 없게 된다. 이와 같은 문제를 해결하기 위해서 일반적으로 세분된 품사를 사용할 수 있다[1]. 그러나, 어느 정도로 많은 품사를 사용해야 하는지는 알 수 없다. 따라서, 본 논문에서는 그와 같은 성질을 가지고 있는 형태소(어휘)를 찾고, 그 형태소를 배열규칙에 포함시키는 방법을 선택하였다.

### 2.1 어휘화된 형태소 배열규칙

일반적인 형태소 접속정보는 표 1과 같다[5]. 이와 같은 접속

*이 연구 결과의 대부분은 첫번째 저자의 학위 과정(한국과학기술원 전산학과 박사과정) 중에 얻어진 것이다.

[1] 이하에서 사용되는 품사태그는 [4].

[2] 여기서, n*의 의미는 n으로 시작하는 모든 품사 태그를 의미한다.

표 1: 어휘화 되지 않은 한국어 형태소 배열규칙

| 왼쪽 품사 | 오른쪽 품사 | 왼쪽 품사 | 오른쪽 품사 |
|---|---|---|---|
| <a, | jca> | <xpa, | exa> |
| <a, | jcm> | <xpa, | exm> |
| <a, | jcp> | <xpa, | exn> |
| <a, | jx> | <xpv, | ecq> |
| ... | ... | <xpv, | exn> |

표 2: 어휘된 한국어 형태소 배열규칙

| 왼쪽 품사 | 오른쪽 품사 | 왼쪽 품사 | 오른쪽 품사 |
|---|---|---|---|
| <*/a, | */s'> | <화/xn, | 는/jx> |
| <*/a, | */s,> | <화/xn, | 도/jx> |
| <*/a, | */s.> | <화/xn, | 되/xpv> |
| <*/a, | */s'> | <화/xn, | 로/jca> |
| ... | ... | <히/xa, | 들/jx> |

정보는 품사들의 쌍에 의해서 정의되는 이진관계(binary relation)이다. 앞에서 언급한 바와 같은 문제로 본 논문에서는 이와 같은 접속정보에 특별한 형태소를 첨가시킨다. 이를 본 논문에서는 **어휘화된 (형태소) 배열규칙** 혹은 **어휘화된 (형태소) 접속정보**라고 하며, 표 2와 같은 구조를 갖는다. 표 2에서 $*/t_i$는 품사 $t_i$를 갖는 모든 형태소를 의미한다. 어휘화된 접속정보는 어휘화되지 않은 접속정보와 마찬가지로 이진순서관계(binary ordered relation)로 표현될 수 있다.

### 2.2 형태소 배열규칙에 포함될 형태소의 선택

어휘화된 형태소 배열규칙을 구성하는 데 있어서 가장 큰 문제는 어떤 형태소를 배열규칙에 포함시킬 것인가 하는 것과 선택된 형태소의 배열규칙으로 어떻게 만들 것인가 하는 것이다. 본 논문에서는 비교적 간단한 방법으로 이 문제를 해결한다. 첫째 문제는 어휘의 수가 제한적인 일부의 기능어는 모두 포함시켰다. 본 논문에서 선택된 기능어는 크게 어미와 조사 그리고 접사들로서 아래와 같다.

```
ecq, ecs, ecx, ef, efp, exa, exm, exn
jc, jca, jcm jcp, jcv, jj, jx,
xa, xn, xpa, xpv
```

둘째 문제는 품사 태깅 말뭉치[3]로부터 모든 형태소/품사 쌍을 구한다. 그리고 나서, 구해진 모든 형태소/품사 쌍$(m_i/t_i)$에서 위에서 선택된 품사를 제외한 모든 품사에 대해서는 형태소 부분을 일반화시킨다$(*/t_j)$.

이와 같은 접근방법의 가장 큰 문제는 접속정보 내에 포함되지 않은 관계가 문장에서 나타나는 경우이다. 이는 미등록어 처리부에서 어휘화된 관계를 완화해서 관계를 검사해 봄으로써 문제를 해결할 수 있다.

### 2.3 어휘화된 접속정보를 이용한 형태적 모호성 축소

어휘화되지 않은 접속정보를 사용할 경우의 형태소 해석 방법[5]과 어휘화된 접속정보를 사용한 형태소 해석 방법은 거의 같다. 단지, 접속정보를 검사하는 부분(morphotactics checking)에서 어휘화된 관계에 해당하는 형태소가 속하는지를 확인하면 된다. 그리고 미등록어 처리부(unknown word processing)에서 어휘화된 접속정보에 의해서 실패되었을 경우에는 이를 완화하여(어휘화 부분을 검사하지 않음) 처리할 수 있도록 하는 기능만 추가하면 된다.

## 3. 형태적 포섭관계을 이용한 형태적 모호성 축소

한국어 단어의 많은 경우는 단어의 형성 과정을 통해서 만들어진 복합어들이다. 복합어의 일부는 형태소 해석의 처리대상에 포함되고, 또 다른 일부는 하나의 단어(단일어, simple word)로 간주된다. 복합어를 단일어로 간주할 경우에도 형태소 해석에서는 복합어에 포함된 어근들의 결합으로 해석할 수 있다. 본 논문에서는 이와 같이 두 형태소의 해석 결과들 사이의 복합어 관계를 찾아서 어근들의 결합으로 해석된 결과를 최종적인 형태소 해석 결과로부터 제거하고자 한다. 복합어에 대한 형태소 해석 결과를 **구체적인 형태소 해석 결과**라고 하고, 어근들의 결합에 의한 해석 결과를 **일반적인 형태소 해석 결과**라고 한다.

### 3.1 포섭관계의 정의 및 성질

**정의 1**
어절 $\Xi$에 대한 형태소 해석 결과 $A = \{A_1, A_2, \ldots, A_n\}$가 있다고 가정하자. 이때 해석 결과 $A_i$가 해석 결과 $A_j$보다 더 일반적인*(more general)* 해석이라면, $A_i$가 $A_j$를 **포섭한다***(subsume)*라고 하며, $A_i \sqsubseteq A_j$로 표기한다. 여기서, $A_i$를 **포섭하는 해석***(subsuming analysis)*이라고 말하고, $A_j$를 **포섭되는 해석***(subsumed analysis)*이라고 말한다. □

예를 들면, 어절 '**날아오다**'에 대한 형태소 해석 결과는 예 (1)과 같으며, 예 (1ㄱ)은 복합어이다. 정의에 따라서, 예 (1ㄴ)은 예 (1ㄱ)을 포섭한다고 말할 수 있다. 즉, 예 (1ㄴ)이 예 (1ㄱ)보다 더 일반적인 형태소 해석 결과이다.

(1) ㄱ. "날아오/pv + 다/ef"
ㄴ. "날/pv + 아/ecx + 오/px + 다/ef"

**정의 2** 어절 $\Xi$에 대한 형태소 해석 결과를 $A = \{A_1, A_2, \ldots A_n\}$라고 하자. 형태소 해석 결과 $A$에 관한 **포섭관계***(subsumption relation)* $S = (A, \sqsubseteq)$는 식 *(1)*와 같이 정의된다.

$$S = \{(A_i, A_j) | A_i \sqsubseteq A_j\} \square \quad (1)$$

**성질 1** 포섭관계는 **부분순서 관계***(partial ordered relation)*이다. □

(2) ㄱ. "소나무/명사"
ㄴ. "소/명사 + 나무/명사"
ㄷ. "소/명사 + 나/명사 + 무/명사"

예를 들면, 어절 '소나무'에 대한 형태소 해석 결과인 예 (2)로부터 포섭관계 $S$(표 3)을 얻을 수 있다.

일반적으로 부분순서 관계는 Hasse 다이그램으로 표현될 수 있으며, 포섭관계 $S$의 Hasse 다이그램은 그림 1과 같다. Hasse 다이그램에서는 이행관계(transitive relation)는 나타나지 않는다.

**정의 3** 어절 $\Xi$에 대한 형태소 해석 결과 $A$에 대한 포섭관계 $(A, \sqsubseteq)$가 부분순서 관계라고 하면, 어떤 해석 $A_j$도 포섭

표 3: 어절 '소나무'에 대한 형태소 해석 결과인 예 (4.2)로부터 포섭관계 $S$

| 포섭하는 해석 결과 | 포섭되는 해석 결과 |
|---|---|
| (소/nc + 나무/nc, | 소나무/nc) |
| (소/nc + 나/nc + 무/nc, | 소나무/nc) |
| (소/nc + 나/nc + 무/nc, | 소/nc + 나무/nc) |

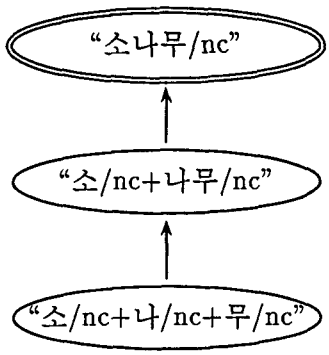


그림 1: 어절 '소나무'에 대한 포섭관계 $S$의 Hasse 다이그램

하지 않는 해석 $A_i$를 **최대해석**(*maximal analysis*)이라고 한다. □

그림 1에서 최대해석은 '소나무/nc'이고, 복선으로 표현되었다.

**성질 2** 어절 $\Xi$에 대한 형태소 해석 결과 $A$에 대한 포섭관계 $(A, \sqsubseteq)$가 부분순서 관계일 때, $\Xi$에 대한 형태소 해석 결과로는 최대해석들만 있으면 충분하다. □

최대해석을 포섭하는 모든 해석들은 최대해석보다 더 일반적인 해석(more general analysis)이기 때문에 최종적으로는 최대해석만 이용하면 된다.

### 3.2 포섭조건

두 형태소 해석 결과가 포섭관계에 포함되는지를 검사하기 위한 조건이 필요하다. 이 조건을 본 논문에서는 **포섭조건**(subsumption condition)이라고 한다. 포섭조건은 아래 항목들의 부울 표현식(Boolean expression)으로 표현된다.

1. 어절 $\Xi$
2. 어절 $\Xi$의 형태소 해석 결과, $\{A_1, A_2, \ldots, A_n\}$
3. 문장 상에서 어절 $\Xi$의 위치, $loc(\Xi)$
4. 예외적인 해석, $exception(\Xi_{sub})$

여기서, $\Xi_{sub}$는 $\Xi$의 부분 문자열(substring)에 의해서 만들어진 정합 패턴이다. 표 4에서는 포섭조건을 만족하는 포섭관계에 대한 예를 보여 주고 있다. 표에서 $T_i$와 $T_j$는 각각 포섭하는 해석 $A_i$와 포섭되는 해석 $A_j$의 품사열이다. 대부분의 포섭조건은 품사열에 의해서 구성된다. 그러나, 다섯번째 조건

표 4: 포섭관계의 예

| 번호 | 포섭하는 해석($A$) | 포섭되는 해석($A_j$) |
|---|---|---|
| 1 | (매일/nc, 우유/nc) | (매일우유/nq) |
| 2 | (서울/nq, 우유/nc) | (서울우유/nq) |
| 3 | (호텔/nc, 신라/nq) | (호텔신라/nq) |
| 4 | (소/nc, 나무/nc) | (소나무/nc) |
| 5 | (따르/pv, 아서/ecs) | (따라서/ajs) |
| 6 | (날/pv, 아/ecx, 오/px) | (날아오/pv) |
| 7 | (즐겁/pa, 어/ecx, 지/px) | (즐거워지/pv) |
| 8 | (동시/nc, 에/jca) | (동시에/a) |
| 9 | (이/npd, 대로/jca) | (이대로/ad) |

시작상태 : { 0 }
종결상태 : { 2 }
시작조건 : { }
정규규칙 : 0 → */nc 1
0 → */nq 1
1 → */nc 2
1 → */nq 2
2 → */nc 2
2 → */nq 2

그림 2: 고유명사(nq)의 포섭조건을 표현한 오토마타, $R^{nq}$

의 경우에는 주어진 어절의 위치가 추가되어 있다. 구체적인 예를 보면, 형태소 해석 '따라서/ajs'는 문장의 시작위치에 있을 경우에만 형태소 해석 '따르/pa+아서/ecs'에 비해 더 구체적인 해석이 될 수 있다. 포섭조건은 포섭되는 해석에 따라 서로 구별되어 질 수 있다. 따라서, 각 포섭되는 해석에 해당하는 포섭조건들은 하나의 오토마타에 의해서 표현될 수 있다. 즉, 정규문법(regular grammar)에 의해서 표현될 수 있다. 포섭조건을 표현하는 정규표현의 단말기호(terminal symbol)는 '$m/t$'와 같은 형태로 표현된다. 여기서 $m$은 형태소이거나 '*'(어떤 형태소와도 정합이 가능함을 의미함)이고, $t$는 품사 태그 중의 하나이다. 시작상태(starting state)에는 $loc(\Xi)$, $exception(\Xi_{sub})$이 부가되며, 이를 **시작조건**이라고 한다. 그림 2는 포섭되는 해석이 고유명사(nq)일 경우의 포섭조건을 표현하는 정규문법이다. 시작상태와 종결상태는 각각 0과 2이고, 시작조건은 아무런 제약조건을 가지지 않는다.

### 3.3 포섭관계를 이용한 형태적 모호성 축소

포섭관계를 이용한 형태적 모호성 축소 방법은 형태소 해석 결과에 대한 포섭관계를 찾고, 그 포섭관계로부터 최대해석을 찾는 것이다[6]. 두 형태소 해석 사이의 포섭관계가 있는지 없는지에 관한 결정은 3.2절에서 설명한 포섭조건을 검사함으로써 이루어진다. 포섭조건을 효율적으로 검사하기 위해서, 먼저 두 해석의 상대적인 차를 구한다. 얻어진 상대적인 차에 대해서 포섭조건을 검사한 후, 그 조건을 만족할 경우, 포섭관계가 성립하는 것으로 간주한다. 예를 들면, 어절 '중국의'를 형태소 해석 결과는 예 (3)과 같다. 예 (3ㄱ)과 (3ㄴ)의 상대적인 차는 예 (4)와 같으며, 예 (4)에 대해서 포섭조건을 검사하면, 예 (4ㄴ)이 예 (4ㄱ)을 포섭하므로 예 (4ㄱ)은 어절 '중국의'에 대한 형태소 해석의 최대해석 중 하나가 될 수 있다. 결과적으로 예 (4ㄴ)은 최종적인 해석에서 제거되게 된다.

(3) ㄱ. "중국/nq + 의/jcm"

표 5: 어휘화된 형태소 배열규칙과 포섭관계를 함께 적용했을 때, 형태소 해석의 모호성 감소

| 형태적 모호성 축소 방법 | 어휘화된 배열규칙 (O) 해석 수 | 어휘화된 배열규칙 (X) 해석 수 | 감소율 |
|---|---|---|---|
| 포섭관계(X) | 353,903개 | 162,874개 | 53.98% |
| 포섭관계(O) | 211,069개 | 115,110개 | 45.46% |
| 감소율 | 40.36% | 29.32% | 67.47% |

ㄴ. "중/nc + 국/nc + 의/jcm"

(4) ㄱ. "중국/nq"
ㄴ. "중/nc + 국/nc"

## 4. 실험 및 평가

### 4.1 실험 환경

학습과 시험을 위해서 KAIST 말뭉치를 사용하였으며, 학습 및 시험 말뭉치는 각각 131,581개와 41,122개의 어절로 구성되었다. 평가에 기본이 되는 형태소 해석기는 [5]를 사용하였다. 포섭관계를 이용한 방법에서 포섭조건은 자동적으로 추론할 수도 있고[6], 수동으로 추론할 수도 있으나, 본 논문에서는 수동으로 추론된 포섭조건을 사용하였다.

#### 4.1.1 성능평가

표 5는 어휘화된 형태소 배열규칙과 포섭관계를 동시에 사용했을 경우, 형태소 해석의 모호성 감소를 보이고 있다. 표 5에서 보아 알 수 있듯이 어휘화된 형태소 배열규칙을 사용하고, 동시에 포섭관계를 이용할 경우에 약 67%이상을 감소할 수 있었다.

### 4.2 토의

어휘화된 접속정보는 비교적 간단한 정보이지만 약 54%의 형태소 해석의 모호성을 감소시킬 수 있었다. 이 실험을 통해서 어휘화된 접속정보는 형태소 해석의 모호성 감소를 위한 매우 유용한 정보임을 알 수 있었다. 그러나,어휘화된 접속정보가 완전한 형태소 배열규칙으로 사용되기 위해서는 충분히 많은 양의 품사 태깅 결과가 요구되나, 이는 미등록어처리부에서 충분히 이를 지원할 수 있기 때문에 큰 문제가 되지 않는다.

형태소 해석에서 최장일치법과 포섭관계를 이용한 방법의 차이점에 관해서 살펴본다. 그 차이를 한 마디로 요약하면, 포섭관계를 이용한 방법은 조건(포섭조건)을 가진 최장일치법이라고 말할 수 있다. 물론 최장일치법의 경우에도 잘못된 해석이 발생될 경우, 여러 가지의 경험(heuristics)에 의해서 이를 제거하나[9], 자동 추론과 같은 방법에 의해서 그와 같은 조건을 구하는 경우는 없었다. 한국어에서 형태소 과잉해석의 줄이는 연구로서 음소, 음절, 그리고 문자열 단위의 배제 정보를 이용한 연구[8]가 있었다.

## 5. 결론

본 논문에서는 형태소 과잉해석으로 발생되는 형태적 모호성을 축소하는 방법에 대해서 기술하였다. 과잉해석은 간단한 형태소 배열규칙과 불규칙 현상 처리에 의해서 발생되는데, 이를 줄이기 위해서 단어의 형성과정에서 이용되는 언어적 지식을 이용하였다. 이를 효과적으로 처리하기 위해서 포섭관계를 정의하였으며, 이 포섭관계를 이용해서 형태적 모호성을 축소할 수 있었다. 이들 두 언어지식을 동시에 사용했을 경우는 약 67%이상의 해석을 감소시킬 수 있었다. 자동으로 포섭조건을 추출하는 방법은 정규화 문법이나 구 구조 문법에 어휘화 혹은 구체화 방법으로 응용될 수 있을 것이다. 또한 형태적 모호성 축소 방법은 구문 해석이나 자연언어 처리 전반에 응용될 수 있을 것이며, 특히 오류 입력을 요구하지 않는 자연언어 처리 시스템에 매우 유용한 도구로 이용될 수 있을 것이다.

## 참고 문헌


[1] 김성웅, *Tabular Parsing* 방법과 접속정보를 이용한 한국어 형태소 해석기, 한국과학기술원, 전산학과, 석사학위 논문, 1987.

[2] 김재훈, 어류-보정 기법을 이용한 어휘모호성 해소, 한국과학기술원, 전산학과, 박사학위 논문, 1996.

[3] 김재훈·김길창, 한국어 품사 부착 말뭉치의 작성요령: KAIST 말뭉치 한국과학기술원, 전산학과, 기술문서(CS-TR-95-99), 1995.

[4] 김재훈·서정연, 자연언어 처리를 위한 한국어 품사 태그, 한국과학기술원, 인공지능 연구센터, 기술문서(CAIR-TR-94-55), 1994.

[5] 김재훈·서정연·김길창, 실용적인 한국어 형태소 해석, 한국과학기술원, 전산학과, 기술문서(CS-TR-95-98), 1995.

[6] 김재훈·장병규·김길창·서정연, "형태소의 모호성을 축소하기 위한 포섭조건의 자동 추론," 제7회 한글 및 한국어 정보처리 학술대회 발표논문집, pp. 175-180, 연세대학교, 서울, 1995.

[7] 이은철, CYK법에 기반한 한국어 형태소 분석에서의 개선 기법, 포항공과대학 대학원, 전자계산학과, 석사학위 논문, 1992.

[8] 임희석·윤보현·임해창, "배제 정보를 이용한 효율적인 한국어 형태소 분석기," 한국정보과학회 논문지, 제22권, 제6호, pp. 987-964, 1995.

[9] Oi, K., Yumura, T., and Nishida, Y., "A Method of Japanese Morphological Analysis Using Longest Matching Method," *Proceedings of the 43th Conference on Information Processing*, vol. 3, 119-120, 1991(in Japanese).