Two-level 한국어 형태소 해석에서의 복합명사 처리

이근용° 박기선 이용석
전북대학교 컴퓨터과학과
{cypher°, kspark}@cypher.chonbuk.ac.kr yslee@moak.chonbuk.ac.kr

A Compound Noun Processing in the Two-level Morphological Analysis of Korean

Keunyong Lee° Kiseon Park Yong-Seok Lee
Dept. of Computer Science, Chonbuk National University

요 약

Two-level 형태소 해석 모델은 단어들이 결합할 때 발생하는 철자변화를 처리하는 언어 독립적인 형태소 해석 모델이다. 그러나 한국어의 경우 활용과 첨용이 자유로운 교착어에 속하며 음절단위 표현법 때문에 two-level 모델을 이용한 형태소 해석 방법보다는 언어 종속적인 형태소 해석 방법을 사용하여 왔다. 한국어 용언과 다양한 변형을 처리하기 위한 two-level 규칙이 표현되었지만, 형태소 해석에서 사용하기 위해서 필요한 복합명사 처리와 미지어 처리에 대한 적절한 방법이 아직 제시되지 않았다. 본 논문은 어절 생성 규칙을 이용한 사전 구성을 이용하여 two-level 모델에서의 한국어 복합명사의 처리에 대해서 다루고, two-level 모델에서 한국어 복합명사 처리가 가능함을 보이고자 한다.

1. 서론

형태소 해석은 문장에 사용되는 단어의 구조를 파악하는 것으로서 단어의 원형에 대한 정의와 각 원형들의 결합관계 및 단어들이 결합할 때 발생되는 철자의 변형에 대해서 처리할 수 있어야 한다[1][2].

형태소 해석은 언어 독립적인 모델과 언어 종속적인 모델 두 가지로 나누어 볼 수 있으며 독립적인 모델로는 two-level형태론과 음절을 기반으로 처리하는 음절기반 형태론이 있다[3]. 한국어 형태소 분석은 한국어가 가지는 특성으로 인하여 언어 독립적인 형태소 분석 방법보다는 언어 종속적인 형태소 분석 방법을 이용하여 왔다[1][2].

언어 독립적인 형태소 분석 모델은 1983년 핀란드의 Kimmo에 의해서 제안된 two-level 형태소 해석 방법론[4]이 있다. Two-level 모델은 형태소 해석에서 철자의 변화를 처리하기 위한 목적으로 제안되었다. Two-level 모델에서는 단어들이 결합할 때 발생되는 철자의 변화에 대해서는 유한 상태 오토마타로 표현하고, 단어의 원형과 결합관계는 사전에 그래프 형태로 저장하는 방법을 사용한다[1][2][4][5][6].

한국어는 활용과 첨용이 자유로운 교착어에 속하며 음절단위 표현법 때문에 two-level 방법론에 의한 형태소 분석이 어렵다고 인식되어 왔지만 한국어 형태소의 가장 복잡한 문제인 용언의 원형복원 규칙과 조사의 결합 규칙, 모음조화 등의 규칙들이 [1][2]에 의해서 만들어졌다. 그러나 복합명사 처리나 미지어 처리에 대한 처리 방법이 제시되지 않았으며, 틀렸지만 맞았다고 분석되는 오분석이 다수 포함되어 있다. Two-level 한국어 해석이 복합명사에 대한 처리와 미지어 처리가 가능하다면 형태소 해석기를 요구하는 다양한 분야에서 이용할 수 있을 것이다.

본 논문은 공개 소프트웨어인 PC-KIMMO Version 2.1.8을 이용하기 위해서 [1][2]의 two-level 규칙을 참고하여 불규칙과 탈락에 관여하는 13개의 규칙과 조사 결합 규칙, 모음조화 규칙 등 21개 규칙, 총 34개의 규칙을 작성하였다. 또한 언어 종속적인 방법으로 많이 사용되고 있는 음절기반 형태소 분석에서 사용하고 있는 한국어 음절의 특성을 포함한 어절 형성 규칙을 이용한 사전을 구성함으로써 복합명사 처리에 대한 방법을 보이고자 한다.

2. 어절의 구성 및 사전의 구성

Two-level 규칙은 형태소 원형들이 결합할 때 발생하는 철자의 변화를 다루는 규칙이고, 형태소 열은 어절 생성 규칙에 의해서 결정된다[6]. 따라서 한국어 형태소 해석을 위해서는 two-level 규칙뿐만 아니라 어절의 형성 규칙을 사전에 같이 기술해줘야 한다.

그림1은 한국어 단어의 간단한 어절 형성 규칙이다. 표1은 그림1의 단어 형성규칙을 바탕으로 작성된 사전의 일부로서 단어의 처음 부분과 끝을 나타내는 것이다.

표1에서 사용된 ₩lf는 lexical form을 ₩lx는 품사를 ₩alt는 다음에 나타날 수 있는 형태소 열을 나타내고 있다. ₩gl은 형태소 열을 찾았을 때 출력으로 나타내는 값

을 지정하여 준다. ₩alt Isolate는 그림1의 Word = Isolate를 나타내며, ₩alt Root는 Word = Root를 나타내는 것으로 단어 형성 규칙의 처음에 나타날 수 있는 것을 가리킨다. ₩lx End는 단어의 마지막을 가리키는 것으로 ₩alt #으로 단어의 경계를 나타내는 #으로 이동하여 단어의 인식이 끝이 난다.

```
Word = Isolate | 부사 (조사)
Word = Root
Root = 체언 (일반조사) | 용언 선어말어미* 어미
Isolate = 감탄사 | 관형사
체언 = 대명사 | 수사 | 명사 | N1
용언 = 형용사 | 동사 | V1 | V2
V1 = 명사 [하|되]
V2 = 체언 (이)
N1 = 용언 선어말어미* 명사형어미
```

그림 1 한국어 단어의 어절 형성 규칙

<표1> 단어의 첫 부분과 마지막을 나타내는 사전 엔트리

₩lf 0	₩lf 0	₩lf 0
₩lx INITIAL	₩lx INITIAL	₩lx End
₩alt Isolate	₩alt Root	₩alt #
₩gl	₩gl	₩gl

표2는 Root에 해당하는 단어의 작성 예를 나타내고 있다.

<표2> 명사에 대한 사전 엔트리 작성 예

₩lf 사회	₩lf 복지	₩lf 위원회
₩lx N	₩lx N	₩lx N
₩alt AfterN	₩alt AfterN	₩alt AfterN
₩gl 사회/N	₩gl 복지/N	₩gl 위원회/N
₩lf 수원	₩lf 지방	₩lf 법원
₩lx N	₩lx N	₩lx N
₩alt AfterN	₩alt AfterN	₩alt AfterN
₩gl 수원/N	₩gl 지방/N	₩gl 법원/N
₩lf 방법	₩lf 원	₩lf 수원지
₩lx N	₩lx N	₩lx N
₩alt AfterN	₩alt AfterN	₩alt AfterN
₩fea ;	₩fea ;	₩fea ;
₩gl 방법/N	₩gl 원/N	₩gl 수원지/N

표2는 ₩alt AfterN은 다음에 나타날 수 있는 것은 AfterN의 값을 갖는 엔트리임을 나타내고 있다. AfterN = N JOSA의 값을 줌으로써 다음에 나타날 수 있는 것이 명사 또는 조사가 올 수 있음을 나타내고 있다. AfterN = N JOSA 만을 가지고 사전을 구성하게 되면, 한음절로 이루어진 모든 명사가 복합명사를 이룬다고 분석될 수 있는 문제점이 있다. 예를 들어서, "수원지방법원"에 대한 분석을 하게 되면 그림2와 같은 분석 결과를 볼 수 있다.

```
수/N원/N지/N방/N법/N원/N+ㄴ/Josa
수/N원/N지/N방/N법/N원/N
수/N원/N지/N방/N법원/N
...
수원/N지방/N법원/N
수원지/N방법/N원/N
...
```

그림 2 "수원지방법원"의 분석 결과

그림2의 분석 결과에서와 같이 조합 가능한 모든 수의 복합 명사가 분석되어 나온다. 따라서 명사를 복합명사를 이룰 수 있는 것과 이룰 수 없는 것으로 나누는 작업이 필요하다. 따라서 그림1의 명사 형성 규칙을 그림3와 같이 작성하고, 규칙에 따라서 명사 사전을 구성하였다.

```
명사 = 접두사 NN EN N
AfterN = EN N 접미사 JOSA
```

그림 3 복합명사 생성을 위한 단어 형성 규칙

그림2의 형성 규칙을 보면, NN, EN, N으로 분류된 모든 명사는 복합명사의 첫 부분에 나타날 수 있지만, NN으로 분류된 명사는 명사 다음에 따라 올 수 없다. 또한 ₩alt의 값을 조정하여 ₩lx가 EN인 경우는 ₩alt의 값을 AfterN이 아닌 Josa를 주어서 뒤에 더 이상의 명사가 결합되지 못하게 하였다. 그림3의 복합명사 생성을 적용하여 사전을 재구성하여 얻은 결과는 그림4와 같다.

```
수원/N지방/N법원/N
수원지/N방법/N원/N+이/si+ㄴ/Eomi
수원지/N방법/N원/N+ㄴ/Josa
수원지/N방법/N원/N
```

그림 4 "수원지방법원"의 분석 결과

그림4의 2, 3번째 결과 또한 "원"는 복합명사의 마지막에 추가되지 않는다는 점을 이용하여 ₩lx N을 ₩lx NN으로 변경하여 수정하면, 1번째와 4번째 결과만을 복합명사 결과로서 얻을 수 있다.

3. 실험 및 평가

3.1 실험

실험에 사용한 two-level 규칙 컴파일러는 Nathan Miles에 의해서 개발된 kgen[7]을 이용하였으며 형태소 해석기는 Evan L. Antworth에 의해서 개발된 PC-KIMMO[6]를 이용하여 이루어졌다.

본 논문에서는 원형복원의 결과에 대한 평가는 [1][2]에 의해서 이루어 졌기 때문에, 복합명사의 처리에 초점을 맞추어 실험 및 평가를 하였다. 실험을 위해서 사용한 명사사

전은 91249개의 엔트리를 가지고 있으며, 다섯 음절 이상의 단어는 단일어인 경우에만 사전에 추가하였다. 분석에 이용한 복합명사는 21세기 세종계획[8] 전자 사전 자료실의 복합어(명사) 500개 중에 475개와 사전을 구성하기 위해서 복합명사의 기준을 선정할 때 사용한 단어 1000개를 대상으로 하였다. 세종계획에서 제외된 25개의 복합명사는 사전을 구성할 때의 차이점으로 제외하였다. 사전을 구성할 때의 차이점은 <징검다리>, <질서정연>, <트레일러버스>와 같은 단어로서 본 실험에 사용할 사전에는 <징검>, <정연>, <트레일>과 같은 단어는 단일어로 사용하지 않는다.

<표 3> 복합 명사 분석 결과

단어	어절 수	결과1	결과2	실패	재현율
세종계획	475	823	622	11	97.68%
테스트셑	1000	1590	1478	0	100%

3.2 결과 분석

<표3>의 결과1은 본 논문에서 사용한 방법에 의해서 나타난 전체 결과수이다. 결과2는 결과1에서 어절을 분석하였을 때 사전에 단일어로 등록되어 있어서 단일어로도 분석된 것을 제외한 결과이다. 예를 들어서, <증조부>의 분석 결과는 {증/N조부/N} {증조/N부/N} {증조부/N}와 같이 세 개의 결과로 분석이 되지만, 마지막의 {증조부/N}와 같은 결과는 사전에 단일어로 등록이 되어 있어서 나온 결과이다. 이런 경우 결과의 수를 두개로 계산하였다. 세종계획의 경우 복합명사 한어절 당 평균 1.73개의 분석이 테스트셑의 경우는 1.59개의 분석이 이루어졌다. 이는 복합명사의 분석에 들어있을 수 있는 중의성의 결과이다. 예를 들어서, <수원지방법원>의 경우와 같은 경우로 다른 언어지식을 사용하지 않을 경우에 나타날 수밖에 없는 현상이다. 세종계획에서 나타난 실패 복합명사의 11개의 경우는 사이시옷이 들어가 있는 복합명사들이다. 예를 들어서 <치맛자락>, <칫솔>, <햅쌀>과 같은 경우이다. 이런 실패의 경우는 사전구성이나 단어 형성 규칙의 문제가 아닌 two-level 규칙에 이런 단어 변형 현상을 기술하지 않았기 때문에 나타난 결과이다.

4. 결론 및 향후 연구 과제

언어 독립적인 형태소 해석 모델인 two-level 모델을 한국어 형태소 분석에서 이용하기 위해서는 우선적으로 복합 명사의 처리와 미지어 처리가 해결되어야 한다. 규칙의 작성에 있어서는 언어 독립적인 모델만으로 충분하지만 사전의 구성이나, 어절 형성 규칙등과 같은 내용들은 언어 종속적인 모델을 이용하여 얻은 결과들을 이용할 때 분석에 들어있는 중의성을 줄이고 결과의 질을 높일 수 있을 것이다.

본 논문에서는 복합 명사의 처리에 대해서 다루고 있으며, 적절한 어절 형성 규칙과 사전의 구성에 의해서 복합명사를 처리할 수 있음을 보였다. 본 논문에서는 복합명사를 처리하기 위해서 어절 생성규칙을 사전에 표현하여 처리하였고, 다른 언어지식은 사용하지 않았기 때문에 결과적으로 분석의 결과가 많음을 알 수 있다. 사전을 구성함에 있어서 기존에 연구된 바 있는 한국어 특성을 반영한 연구 결과를 이용하여 사전을 구성한다면 복합명사의 분석에 더욱 도움이 될 것이다. 복합명사에 있어서 단순히 규칙에 의해서 처리된 만큼 복합명사 자체가 사전에 이미 수록되어 있는 경우 또한 있으므로 규칙에 의한 결과를 얻어낸 다음 적절한 처리를 거쳐서 불필요하다고 여겨지는 결과를 제거할 수 있어야 할 것이다.

한국어 형태소 분석을 위한 two-level 모델에 대한 연구가 많이 진행되지 않은 관계로 실제 응용분야에서 사용가능한 형태소 분석기가 되기 위해서는 앞으로도 개선되고 추가될 부분들이 많이 있다.

1. 사이시옷에 대한 규칙을 세우는 것이 쉬운 일은 아니지[9][10]만 복합명사 처리를 위한 사이시옷에 대한 규칙을 추가해야 할 것이다.
2. 미지어 처리에 대한 방법이 연구되어야 할 것이다.
3. Two-level 규칙을 실험에 의해서 계속해서 보강함으로써 two-level 모델의 양방향성을 이용한 한국어 생성에도 이용할 수 있도록 해야 할 것이다.
4. 틀렸지만 맞았다고 분석하는 오분석을 줄이기 위한 방법이 필요하다.

참고 문헌

[1] 이성진, Two-level 한국어 형태소 해석, 한국과학기술원 전산학과 석사학위 논문, 1992

[2] 이성진, 김덕봉, 서정연, 최기선, 김길창, Two-level 모델을 이용한 한국어 용언의 형태소 해석, 한국정보과학회 가을 학술발표논문집, Vol 19, No.2, pp 993~996, 1992

[3] 강승식, 음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 컴퓨터공학과 박사학위 논문, 1993

[4] Koskenniemi, Kimmo. A general computational model for word-from recognition and production. In Proceeding of COLING-84, pp 178-181. Association for Computational Linguistics. 1984

[5] Karttunen, Laurie. KIMMO: a general morphological processor. Texas Linguistic Forum 22:163-186, 1983

[6] Antworth, Evan L. User's Guide to PC-KIMMO Version2, Summer Institute of Linguistic, Inc, 1995

[7] Nathan Miles, PRELIMINARY DOCUMENTATION FOR KGEN: a rule compiler for PC-KIMMO, 1991

[8] http://www.sejong.or.kr/sejong_kr/index.html

[9] 이희승, 안병희, 한글 맞춤법 강의, 신구문화사, 1994

[10] 이익섭, 채완, 국어 문법론 강의, 학연사, 2000