Journal of Korean Language Education
한국어교육연구 제12권 1호, 2017.02. pp.148-165.

한국어 형태소 분석기 개발을 위한 문법 전략

이 영 민

한국어교육연구
Journal of Korean Language Education

『한국어교육연구』 제12권 1호, 2017.02. pp.148-165.

한국어 형태소 분석기 개발을 위한 문법 전략

이영민(세종대학교)

차 례

<국문 초록>

본 연구는 규칙 기반의 한국어 형태소 분석기 개발을 위하여 한국어 문법 규칙을 제시하고자 한다. 이는 형태소 분석기의 성능을 향상시키기 위한 방법으로서 먼저 한국어에서 실현 가능한 어절의 유형을 정리하고 이에 따라서 모든 유형을 하나씩 검색, 해당 유형에 해당하는 분석 결과를 도출해 내도록 한 것이다. 이를 위해서는 유형 자체가 간단하고 간략하게 정의되어야 하는데 이를 위하여 명사 유형과 동사 유형의 어절 유형을 조정할 필요가 있다. 명사 유형의 어절에는 명사에 조사가 통합되는데(명사+조사) 서술격 조사도 통합될 수 있다. 이럴 경우 대부분의 선어말어미 및 어미가 통합 가능하므로 유형의 수가 무척 많아진다. 또한 어미에는 다시 명사형 어미('-음', '-기')가 통합될 수 있는데 그럴 경우 다시 조사가 통합될 수 있으므로 그 경우의 수는 더 많아진다. 이를 조정하여 적절한 수의 어절 유형을 정하는 것이 중요하다.

이에 더하여 조사 복합체 및 선어말 어미 복합체의 숫자를 정의해 주는 것도 의미가 있다. 이를 위해서는 중복 실행이 가능한 조사와 선어말 어미의 목록을 정의해 주어야 한다. 그리고 가장 활용빈도가 높은 복합명사를 처리하기 위한 방안을 제시하였다. 이에 더하여 불규칙 활용의 문제를 처리함으로써 정확성을 높이고자 하였다.

주제어: 형태소 분석기, 한국어 문법, 문법 기반, 어절 유형 정의, 복합명사, 불규칙 활용

1. 서론

최근 인공지능(AI)의 기술 수준이 획기적으로 발전하면서 빅데이터(Big Data)의 중요성이 크게 대두하고 있다. 빅데이터는 매우 다양한 형태로 주어지지만 그 중에서도 언어 형태로 주어지는 데이터는 직접 정보를 처리할 수 있다는 점에서 특히 중요하다고 할 것이다.

이에 언어 정보를 분석, 처리할 수 있는 도구로 형태소 분석기는 매우 유용하다고 할 것인데, 한국어는 그 특성상[1] 한국어 형태소 분석기 개발이 쉽지 않다. 한국어 형태소 분석은 그 분석 처리단위와 분석방법으로 나뉜다. 처리단위는 기본적으로 한국어의 띄어쓰기 단위인 '어절'을 중심으로 단어와 조사, 어미 등을 분리해내고 이에 대해 품사 후보를 결정하는 것이다. 나아가 형태소, 어절, 자소 단위의 분석이 가능하다[2].

형태소 분석 방법은 초기에는 한국어의 규칙을 기반으로 하여 이루어졌는데 반하여 후기에는 통계에 기반하여 이루어졌다 (전자는 김성용(1987), 강승식(1993), 권오욱 등 (1999)이 후자는 이도길(2005), 이재성(2011), 나승

1) 교착어(agglutinative)

2) 서현민(2014) 참조. 세종 말뭉치(국립국어원(1998-2106))를 대상으로 기분석 부분 어절 사전을 구축하고 각 부분의 분석 내용으로 후보를 생성, 빈도 정보와 형태소 전이 빈도 등을 통해 평가 점수를 계산하는 방식을 따른다(신준철(2013)).

훈(2012) 참조). 규칙기반의 방법은 한국어 사전을 구축하고 한국어 문법을 연구, 규칙을 도출해내고 이를 통하여 한국어 형태소를 분석해 내는 방법이다[3]. 이러한 방법은 1) 사전을 구축하고 규칙을 정립하는 데 많은 시간과 노력이 요구되며, 2) 언어 변화에 따라 사전과 규칙을 주기적으로 관리해주어야 하는 어려움이 있다. 이러한 작업이 선/후행되지 않는 않는다면 그 성능을 보장받을 수 없다. 그렇지만 이러한 작업이 가능하다면 분석 속도가 빠르고 구조화된 문서에 대한 분석 성능을 높일 수 있다는 장점이 있다.

통계기반의 방법은 대용량의 말뭉치로부터 확률 정보를 학습하여 확률 규칙을 자동으로 생성하고 이를 형태소 분석에 활용한다. 이 방법은 품사가 미리 정해진 언어 자료(말뭉치)를 활용하므로 말뭉치만 확보하면[4] 시간과 노력이 절감하여 쉽게 구축할 수 있다는 장점이 있다. 그렇지만 말뭉치 학습을 통하여 규칙을 생성하므로 규칙을 정립, 수정하기가 어렵고 속도가 상대적으로 느리다는 단점이 있다. 또한 불규칙 활용의 처리에서 그 정확도를 보장받기가 쉽지 않다.

본 연구는 규칙기반에 의한 한국어 형태소 분석기 개발을 목표로 하며 이에 필요한 규칙을 제시하고자 한다.

2. 기본 구성

한국어는 그 특성상(주1) 참조) 어휘범주와 문법범주가 어절 단위로 통합되어 어절 단위로 분석이 가능하다. 이에 한국어에 실현 가능한 어절의 유형(type)을 전부 정의하고 이에 대한 규칙을 정립, 형태소 분석이 가능하도록 할 수 있다. 이에 더하여 복합명사는 상당수가 사전에 등재되어 있지 않으므로 별도의 문법적 처리가 필요하다. 그리고 규칙으로 처리하기 어려

3) 이에 더하여 기분석 사전을 구축해 두고 이를 활용하는 방법도 있다.
4) 말뭉치의 구축이 쉽지는 않지만 최근에는 개인과 국가기관 등에 의하여 상당히 많은 말뭉치가 구축되고 있다(주2) 참조)

운 일부 어절에 대한 별도의 처리를 한다. 이를 통하여 모든 한국어 어절을 규칙에 기반하여 분석할 수 있을 것이며 그 알고리즘을 설계할 수 있을 것이다(김현주 외(2017) 참조)

3. 어절의 유형 정의

형태소 분석은 무엇보다도 어절 내부에 대한 형태소 확인에 의존하는데, 한국어의 특성상 어절은 단어와 관련 문법 범주의 복합체로 구성된다. 단어는 명사가 대다수를 차지하고 있으며 동사, 형용사, 부사, 관형사가 주요 범주를 이룬다. 명사류로 구성된 어절의 기본적인 형태는 명사에 '조사 내지는 조사 복합체'가 결합된 형태, 동사와 형용사로 구성된 어절은 동사와 형용사에 '어미류'가 결합된 형태가 일반적이다. 그렇지만 명사는 이른바 서술격 조사, 동사는 전성어미에 의하여 각각 어미와 조사가 결합할 수 있으므로 어절 구성은 다소 복잡해진다. 먼저 명사 구성을 제시하면 다음과 같다.

3.1. 명사 유형

명사는 원칙적으로 조사를 요구하기에[5] 조사가 결합된 형태로 나타난다. 이에 가장 기본적인 어절 구성은 다음과 같이 실현난다.

5) 명사와 격범주와의 관계는 여러번 논의된 바 있다. 한편 한국어의 조사가 전적으로 격 범주를 담당한다고 하기는 어렵다. 그렇지만 격의 개념을 아주 폭넓게 상정, 특수 조사를 비롯한 모든 조사를 격조사로 처리한다.

(1) 명사 유형1

구분	유 형	구 성	예 시
1	NJ	명사-조사	학생-이 학생-도 시계-를 시계-만
2	NJJ	명사-조사복합체	학생-만이 학교-까지는 시계-에서도

(1)의 유형2에서 보듯이 조사는 중첩되어 실현될 수 있는데 '학교에서부터는'과 같이 3개까지도 가능하다. 그런데 유형의 수는 형태소 분석기의 성능(속도)과 관계가 있으므로 조사복합체를 전부 망라하여 목록을 정리, 하나의 조사로 처리한다. 이를테면 '-까지는', '-에서도', '-에서부터는'을 하나의 조사로 처리한다는 것이다. 이에 더하여 분석의 오류를 방지, 정확도를 높일 수 있다.

(2) ㄱ. 진주만을, 진주만도
ㄴ. 진주(N)-만을(J)/진주만(N)-을(J),
진주(N)-만도(J)/진주만(N)-도(J)
ㄷ. 진주만만, 진주만과
ㄹ. 진주만(N)-만(J), 진주만(N)-과(J)

(2ㄱ)의 '진주만을'은 명사 '진주'와 조사 '-만을', 명사 '진주만'과 조사 '-을'로, '진주만도'는 명사 '진주'와 조사 '-만도', 명사 '진주만'과 조사 '-도'로 분석된다. 반면, (2ㄷ)의 '진주만만'은 명사 '진주만'과 조사 '-만', '진주만과'는 명사 '진주만'과 조사 '-과'로 분석된다. 정리된 조사의 목록에 '-만만'과 '-만과'가 존재하지 않기 때문이다(최태성(1999)).

이렇게 되면 (1)에는 하나의 유형만 남게 된다. 물론 이를 위해서는 중첩되어 실현될 수 있는 조사들의 목록을 정확하게 정리하는 것이 필요하다.

(3) 명사 유형2[6]

이 유형은 명사에 서술격 조사와 어미 및 선어말 어미가 결합한 것이다.

구분	유 형	구 성	예 시
3	NCE	명사-서술격조사-어미	학생-이-다 시계-이-구나 학생-이-ㄴ
4	NCPE	명사-서술격조사-선어말어미-어미	학생-이-었-다 시계-이-겠-네
5	NCPPE	명사-서술격조사-선어말어미복합체-어미	학생-이-시었-다 시계-이-었겠-구나

(1)에서와 마찬가지로 유형의 수를 줄이는 것이 필요하므로 선어말어미 복합체를 별도로 정의하여 유형5 'NCPPE'를 제외한다. 그러면 여기는 2개의 유형만 남는다.

(4) 명사 유형3

전성어미 '-(으)ㅁ', '-기'에는 조사가 결합할 수 있으므로 (3)에 더하여 (4)의 유형이 필요하다.

구분	유 형	구 성	예 시
6	NCEJ	명사-서술격조사-어미-조사	학생-이-ㅁ-을 시계-이-기-에
7	NCPEJ	명사-서술격조사-선어말어미-어미-조사	학생-이-었-음-을 시계-이-었-기-에
8	NCPPEJ	명사-서술격조사-선어말어미복합체-어미-조사	학생-이-시었-음-을 시계-이-었겠-기-를

6) 문법적으로 이 유형을 '명사 유형'으로 처리하는 데는 논의의 여지가 있다. 그렇지만 명사를 우선적으로 분석하는 것이 필요한 경우가 많으므로 이를 '명사 유형'으로 분류한다. 이러한 설명은 (4)와 (5)의 유형에도 동일하게 적용된다. 어절 전체가 서술어로 기능하는 문제는 별도로 처리할 수 있을 것이다.

(1), (3)에서와 마찬가지로 선어말어미 복합체를 정리하면 유형8(NCPPEJ)를 제외할 수 있다. 또한 '학생임에도', '학생이었음에도'와 같이 조사가 중첩되어 실현되는 경우도 별도의 유형을 설정하지 않고 분석이 가능하다.

(5) 명사 유형4

명사에 서술격 조사가 결합한 유형과 비슷하게 명사에 용언화 접사가 결합하면 어미류가 결합할 수 있으므로 별도의 유형을 정의할 필요가 있다. 용언화 접사가 결합한 형태를 사전에 (형용사로) 등록시키기는 어렵다.

구분	유 형	구 성	예 시
9	NXE	명사-용언화접사-어미	학생-답-다
10	NXPE	명사-용언화접사-선어말어미-어미	학생-답-었-다
11	N X P P E	명사-용언화접사-선어말어미 복합체-어미	학생-답-시었-다
12	NXEJ	명사-용언화접사-어미-조사	학생-답-기-를
13	NXPEJ	명사-용언화접사-선어말어미-어미-조사	학생-답-었-음-을
14	NXPPEJ	명사-용언화접사-선어말어미 복합체-어미-조사	학생-답-시었-음-을

이 유형은 명사에 용언화 접사가 결합한 유형인데 선어말어미 복합체(유형11, 유형14)와 조사가 중첩되는 경우(학생-답-었-음-에도)는 (2), (3)의 설명으로 제외할 수 있다.

3.2. 동사 유형[7)]

한국어의 동사는 비자립 형태이므로 어말 어미를 반드시 요구한다. 어미가 결합된 형태가 기본이 되며 선어말 어미는 수의적으로 결합한다.

(6) 동사 유형

구분	유 형	구 성	예 시
15	VE	동사-어미	먹-다 예쁘-다
16	VPE	동사-선어말어미-어미	먹-었-다 예-뻤-다
17	VPPE	동사-선어말어미복합체-어미	먹-었겠-구나 예쁘-겠더-라

앞의 설명과 마찬가지로 이 유형에서도 유형17(선어말어미 복합체)은 제외할 수 있다. 이에 더하여 전성어미 '-(으)ㅁ', '-기'가 결합한 경우는 조사가 결합할 수 있으므로 별도의 유형이 더 필요하다. 아래의 유형에서도 선어말어미 복합체(유형20)와 조사가 중첩되는 경우(먹-기-만을)는 (2),(3)의 설명에 따라 제외할 수 있다.

구분	유 형	구 성	예 시
18	VEJ	동사-어미-조사	먹-기-를 예쁘-기-에
19	VPEJ	동사-선어말어미-어미-조사	먹-었-기-에 예-뻤-음-을
20	VPPEJ	동사-선어말어미복합체-어미-조사	먹-었겠-기-에 예쁘-었겠-기-에

7) '동사 유형'이라기보다는 '동사/형용사 유형'이라고 해야 한다. 따라서 유형15는 'VE/AE'가 되어야 할 것이다. 동사와 형용사를 구별하지 않고 '동사'로 처리한 것은 다만 편의를 위한 것이다.

3.3. 부사 및 기타 유형

(7) 부사 유형

부사는 단독으로도 쓰이지만 조사와 결합하여 쓰이기도 한다. 단독으로 쓰이는 경우는 사전에서 바로 검색이 된다.

구분	유 형	구 성	예 시
21	ADV	부사-조사	많이-는 잘-도

(8) 기타 유형

높임의 '-요'는 별도의 유형을 필요로 한다.

(9) ㄱ. 이제 밥을 먹어요.
ㄴ. 아이들이요 많이요 왔거든요.

(9)에서 보듯이 '-요'는 어말 어미 뒤와 쓰이며 명사 어절 뒤에 자유롭게 결합한다. 이에 '-요'를 처리하기 위한 별도의 유형이 필요하다.

구분	유 형	구 성	예 시
22	E요, J요	어미, 조사, 부사-조사	먹어요 제가요 많이요

이 유형은 '-요'가 쓰이면 먼저 선행하는 형태에 조사나 어미가 있는지를 확인한다.

이상의 유형을 정리하면 다음과 같다.

구분	유 형	구 성
1	NJ	명사-조사
2	NCE	명사-서술격조사-어미
3	NCPE	명사-서술격조사-선어말어미-어미
4	NCEJ	명사-서술격조사-어미-조사
5	NCPEJ	명사-서술격조사-선어말어미-어미-조사
6	NXE	명사-용언화접사-어미
7	NXPE	명사-용언화접사-선어말어미-어미
8	NXEJ	명사-용언화접사-어미-조사
9	NXPEJ	명사-용언화접사-선어말어미-어미-조사
10	VE	동사-어미
11	VPE	동사-선어말어미-어미
12	VEJ	동사-어미-조사
13	VPEJ	동사-선어말어미-어미-조사
14	ADV	부사-조사
15	-요	어미, 조사, 부사-요

4. 복합명사

복합명사는 파생명사와 합성명사[8]를 아우르는 개념으로서 사전에 등재된 다. 복합명사가 사전에 등재되어 있으면 해당 어절에서 조사만 분리해 낼 수 있으므로 일반 명사와 마찬가지로 형태소 분석의 알고리즘에 부담이 되지 않는다. 그런데 현실은 그렇지 않다. 많은 분야에서 복합명사를 명사 복합체의 개념으로 사용하고 있으며 어절별로 띄어쓰기를 하지 않는 경우가 대부분이다.

8) complex와 compound를 각각 복합과 합성으로 이해한 것이다(이익섭·임홍빈(1983), 고영근·남기심(1991), 이익섭·채완(2012)) 참조)

(10) ㄱ. 교양소설, 인공지능, 두꺼비집, 큰아버지
ㄴ. 질소화합물,
ㄷ. 남북국시대, 복소수평면
ㄹ. 연결재무제표
ㅁ. 중거리탄도유도탄

(11) ㄱ. 계좌번호, 거래실적, 연구성과, 인접과학, 남북분단
ㄴ. 한국인류학, 항공승무원, 자기효능감
ㄷ. 세계화시대, 성분별조사
ㄹ. 전국표본조사, 회계추정방법, 국제회계기준
ㅁ. 연차별도산출가액, 대중국무기수출가

(10)의 예들은 사전에 등재된 단어들로서 복합명사로 처리된다. 이를테면 '교양소설을', '중거리탄도유도탄은'은 '교양소설(명사)+을(조사)', '중거리탄도유도탄(명사)+은(조사)'로 분석되며 문제가 되지 않는다[9)10)].

반면에 (11)의 예들은 사전에 등재되어 있지 않으므로 문제가 복잡해진다. '계좌번호가'와 같이 다소 간단한 경우도 '계좌번호'가 사전에 등재되어 있지 않으므로 분석되지 않거나 조사를 분리해 낸다 하더라도 '계좌번호(UN[11))]+가(조사)'로 분석될 수밖에 없다. 따라서 별도의 분석 알고리즘이 필요한데, 대부분의 4음절 복합명사는 그 내부를 확인하기가 어렵지 않다. 그렇다 하더라도 조사를 분리해 내고 별도의 알고리즘을 사용하여 내부의 명사를 확인해야 한다. 이를테면 '계좌번호를'은 일단 조사를 분리해낸 상태에서('계좌번호(UN)+를(조사)') 미확인 부분인 '계좌번호'를 분석해내는

9) 물론 형태소 분석으로는 결과가 다르게 나타나야 하겠지만 명사를 정확히 분석해내는 것이 목적이므로 이와 같은 처리는 문제가 되지 않는다. 현재의 설명은 국립국어원의 표준국어대사전을 대상으로 한 것이다.

10) 사전에 등재된 명사의 숫자가 관건이 될 것이다. 그렇지만 언급한 바와 같이 많은 분야에서 복합명사를 명사복합체의 의미로 사용하고 있으며 어절별로 띄어쓰기를 하지 않는 경우가 대부분이므로 모든 '명사'를 등재하는 것 자체가 불가능할 것이다.

11) 확인할 수 없는(unknown) 단어나 어근.

알고리즘을 적용하는 것이다. 여기서는 4음절 이하의 복합명사는 간단한 알고리즘을 적용하기로 한다. 4음절 복합 명사의 경우 대부분이 '2+2'의 구성으로 되어있으므로 이를 확인하는 알고리즘을 적용하고 그렇지 않은 경우는 '3+1', '1+3'으로 분석하는 것이다. 이를테면 '일회용침'은 '2+2'의 구성으로 분석하면 '일회(명사)+용침(UN)'으로 분석되는데 이럴 경우는 다시 '3+1'의 알고리즘을 적용하여 '일회용(명사)+침(명사)'로 분석될 수 있도록 한다는 것이다.

5음절 이상의 복합명사는 모든 경우의 수를 따져서 분석한다. 5음절일 경우 가능한 경우의 수는 다음과 같다.

(12) '11111', '1112', '1121', '113', '1211', '122', '131', '14',
'2111', '212', '221', '23',
'311', '32'
'41',
'5'

이를 적용하면 5음절 복합명사는 24, 6음절은 25, 7음절은 26의 경우의 수가 산출된다. 이에 1음절로만 되는 경우('11111')와 전체가 하나의 단어로 등재된 경우(즉 (12)의 '5')를 배제하면 각각 24-2, 25-2, 26-2의 경우의 수가 산출된다. 이를 적용하여 14음절까지 분석가능한 알고리즘을 적용한다. 5음절 복합명사에 조사 '-이'가 결합된 경우를 예로 들면 다음과 같다.[12)]

(13) 총자산가액이
ㄱ. 총자(N)산(N)가액(N)이(N):100_UK
ㄴ. 총(N)자산(N)가액(N)이(N):100_UK
ㄷ. 총자산(N)가액(N)이(N):100_UK
ㄹ. 총자(N)산가(N)액(N)이(J):100:_NJ

12) 이에 반하여 통계 기반으로 복합명사를 분석하고자 하는 연구도 이루어졌다(강민규·강승식(2010), 윤보현, 임희석, 임해창(1995), 윤보현, 조민정, 임해창(1997))

ㅁ. 총(N)자산가(N)액(N)이(J):100:_NJ
ㅂ. 총자(N)산(N)가액(N)이(J):100:_NJ
ㅅ. 총(N)자산(N)가액(N)이(J):100:_NJ
ㅇ. 총자산(N)가액(N)이(J):100:_NJ
ㅈ. 총자(N)산가(N)액이(UN):65:_UK
ㅊ. 총자산(N)가(N)액이(UN):60:_UK
ㅋ. 총(N)자산가(N)액이(UN):60:_UK
ㅌ. 총자산가액(UV)이(E):50:_VE
ㅎ. 총자산가액(UN)이(J):50:_NJ ...[13]

(13ㄱ-ㅇ)은 확인되지 않은 형태가 없으므로 전부 수용가능하다(100점). 다만 (13ㄱ-ㄷ)은 유형에서 확인되지 않으므로(1장 참조) 배재할 수 있다. 그리고 (13ㄹ-ㅇ)은 일단 수용가능한 것으로 보는데 이들 중에서 음절 수가 가장 적은 것을 선택하도록 하면 (13ㅇ)이 선택된다. (13ㅈ-ㅎ)은 확인되지 않은 형태가 있으므로 배제된다. (13ㅎ)이 더 정확한 분석일 수도 있지만 '총자산가액'이 사전에 등재되어 있지 않으므로 확인이 되지 않은 것이다.

5. 불규칙 활용, 조사 복합체, 선어말어미 복합체

이상으로 어절 유형과 복합 명사를 분석하기 위한 알고리즘을 제시하였다. 이에 더하여 불규칙 어근의 처리, 조사 복합체 및 선어말 어미 복합체의 목록((1), (6) 참조)이 더 필요하다. 불규칙 어근은 어근의 유형을 다음과 같이 정의하고(허용 외(2005)),

(14)
ㄱ. 1군(자음으로 시작):-다, -고, -게, -는, -ㅂ니다/습니다 등
ㄴ. 2군(모음으로 시작):-아서/어서, -아라/어라, -아/어, -았/었-

13) 더 많은 분석 결과가 산출되나 동일한 설명이 적용되므로 생략함.

ㄷ. 3군('-으' 계열) : -(으)며, -(으)니, (으)니까, -(으)면 등

후행하는 어미에 따라서 각각 어간, 어미, 어간과 어미가 바뀌는 불규칙 활용을 해당 단어에 표시를 함으로써 처리할 수 있다. 이를테면 '짓다'는 사전에 'ㅅ' 불규칙 활용을 하는 단어로 등재해 두고 'ㅅ' 불규칙 활용을 하는 단어는 2, 3군 어미가 후행하는 경우 'ㅅ'이 탈락('ㅅ' 〉 ∅)하도록 지정해 둔다는 것이다. 이와 같은 방식으로 모든 불규칙 활용을 처리할 수 있다.

그리고 조사와 선어말 어미가 겹쳐서 실현되는 복합체의 경우, 이들의 목록을 정의해 둠으로써 처리가 가능하다.

6. 결론

이상의 논의를 통하여 한국어 형태소 분석을 위한 규칙 체계를 정리하였다. 본 연구는 한국어 형태소 분석에 있어서 규칙 기반을 활용함으로써 그 성능을 확보하고자 한 것으로서 한국어에서 실현 가능한 어절의 유형을 정리하고 이에 따라서 모든 유형을 하나씩 검색, 해당 유형에 해당하는 분석 결과를 도출해 내도록 한 것이다. 이를 위해서는 유형 자체가 간단하고 간략하게 정의되어야 하는데 이를 위하여 조사 복합체와 선어말어미 복합체를 정의하여 그 경우의 수를 줄이고자 하였다. 이에 더하여 불규칙 활용의 문제를 처리함으로써 정확성을 높일 수 있었으며 영문이나 숫자, 기호 등이 사용된 어절(특수어절)의 경우도 별도의 규칙을 제시하여 처리할 수 있도록 하였다(김현주 외(2015) 참조).

<참고문헌>

강민규·강승식(2010), 한국어 복합명사 분해 오류 교정 기법, 한국정보과학회 학술발표논문집 37-1, 한국정보과학회, 254-259쪽.

강승식(1993), 음절정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 박사학위논문.

강승식(1998), 한국어 복합명사 분해 알고리즘, 정보과학회논문지B 25-1, 한국정보과학회, 172-182쪽.

고영근·남기심(2014), 표준 국어문법론, 박이정.

권오욱 외(1999), 음절단위 CYK 알고리즘에 기반한 형태소 분석기 및 품사 태거, 제 11회 한글 및 한국어 정보학회 학술대회 및 제1회 형태소 분석기 및 품사 태거 평가 워크숍, 한국정보과학회언어공학연구회, 76-87쪽.

김성용(1987), Tabular parsing방법과 접속 정보를 이용한 한국어 형태소 분석기, KAIST 석사학위논문.

김현주 외(2015), 문장부호를 고려한 특수어절 알고리즘, 언어정보학회 논문집 22-2, 한국언어정보학회, 1-4쪽.

김현주, 이영상, 천승태(2017), 한국어 형태소 분석기 개발을 위한 알고리즘, 배재대학교 한국어교육연구소, 한국어교육연구 12-1, pp.56-69.

나승훈 외(2012), CRF에 기반한 한국어 형태소 분할 및 품사 태깅, 제24회 한글 및 한국어 정보처리 학술대회, 한국정보과학학회, 12-15쪽.

서현민(2014), 통계기반 기계번역 도구와 조건부 랜덤필드 도구를 이용한 한국어 형태소 분석, 충북대학교 석사학위논문.

신준철(2013), 기분석 부분 어절 사전 기반의 형태소 분석 및 음절-형태소 전이 확률 기반 품사·동형이의어 태깅, 울산대학교 박사학위논문.

윤보현, 임희석, 임해창(1995), 통계정보를 이용한 한국어 복합명사의 분석 방법, 봄학술발표논문집, 한국정보과학회, 925-928쪽.

윤보현, 조민정, 임해창(1997) 통계 정보와 선호 규칙을 이용한 한국어 복합 명사의 분해, 정보과학회논문지(B) 24-8, 한국정보과학회, 900-909쪽.

이도길(2005), 한국어 형태소 분석과 품사 부착을 위한 확률모형, 고려대학교 박사학위논문.

이익섭·임홍빈(1983), 국어문법론, 학연사.
이익섭·채완(2012), 국어문법론 강의, 학연사
이재성(2011), 한국어 형태소 분석을 위한 3단계 확률 모델, 정보과학회논문지 소프트웨어 및 응용, 제8권, 한국정보과학회, 257-268쪽.
최태성(1999), 조사 겹침에 대한 연구, 경희대학교 석사학위논문.
허용 외(2005), 외국어로서의 한국어교육학 연구, 박이정.

<Abstract>

Grammatical Strategy to develop a Korean morpheme analyser

This paper aims to present a grammatical rule for Korean morpheme analyser. Korean morpheme analyser needs a simple and compressive rule base to improve processing speed. We present 15 asal(lexical word + grammatical word(case markers or endings)) type. Noun type asal of Koran typically consist of noun+'josa'(case marker) and josa can be duplicated. To reduce the number of asal type, we present a list of duplicated josa(case marker). Predicate case marker('-i') can be integrated to a noun. In that case many (prefinal) endings can be integrated to predicate case marker(noun+predicate case marker+prefinal endings+endings). In addition another case markers can be integrated if endings were noun form('-m', '-ki'). As a result the numbers of asal type is increased and it's not good for algorithm.

We will compress and decide asal type, then reduce the number of asal to 15 : noun+josa, noun+predicate josa+ending, noun+predicate josa+prefinal ending+ending, noun+predicate josa+ending, noun+predicate josa+prefinal ending+ending+josa, noun+verbal suffix+ending, noun+verbal suffix+prefina ending+ending, noun+verbal suffix+ending+josa, noun+verbal suffix+prefinal ending+ending+josa, etc..

We present a rule to analyse noun compound of 14 syllables and a rule to analyse a irregular verb/adjective.

Key words: Korean morpheme analyser, rule system, 15 asal types, noun compound, irregular verb/adjective.

이영민(Lee youngmyn)
세종대학교 겸임교원
주　　소: 서울특별시, 강동구, 성내2동 600. 이편한세상 301-207
전자우편: 007-ymlee@daum.net
전화번 : 010-7726-4910.

■ 접　수　일: 2017.12.23.
■ 심사완료일: 2017.01.28.
■ 게재확정일: 2017.02.10.