한글 문장의 단어들 묶음에 대한 ‘연구방법의 기본뼈대’ 적용

김명호(석궁김명호 출판사)

1. 요약

무수한 정보의 검색, 가공을 컴퓨터에 의지하고 있는 현실에서, 문장의 주어, 목적어, 보어, 서술어를 컴퓨터로 하여금 구별 저장하도록 하는 필요성이 발생하고 있다. 왜냐하면, 컴퓨터의 궁극적 목표는 인간의 말을 알아듣고 실행할 수 있는 로봇[1]이기 때문이다. 허나, 한글 문장은 영어와 비교할 때 어떤 절이나 구가 특정 단어를 수식한다는 것을 명시적으로 알려주는 관계대명사 등이 존재하지 않는다. 이 논문에서는 이러한 어려움을 극복할 수 있는 하나의 방법을 제시하고 그를 적용한 실행 프로그램을 소개함으로써, 그 가능성을 알리고자 한다.

실행 프로그램을 예를 들어 설명하면, “옆집에 사는 형이 빌려준 책은 내가 비디오방에서 본 영화들보다 재미있었다”라는 문장을 컴퓨터 프로그램에 입력하면 아래와 같이 출력되도록 하는 것이 현 단계 목표다. 나아가 확장된 의미의 번역[2]을 위한 사전 구축 등을 통해 외국어 번역에 응용될 수 있기를 기대한다.

```
 ================
옆집에 사는 형이 빌려준 책은
        옆집에 사는 형이
                옆집에
                사는
                형이
        빌려준
        책은
내가 비디오방에서 본 영화들보다
        내가
        비디오방에서
        본
        영화들보다
재미있었다.
```

2. 학문 연구 방법의 뼈대

현재까지 인간이 수행하여 온 학문 연구의 기본은 ‘뭉치화(분류; classification)’이고, 중요한 것은 뭉치화를 위한 기준을 설정하는 것이다. 인문학이든 자연과학이든 용어나 표현방법은 조금씩 다르지만, 모든 연구는 방법의 뼈대는 같고 변수를 다루는 도구가 다를 뿐이다. 문제는 학문에서 다루는 변수들의 규모이다. 분석학의 기본이라는 수학에서는 변수가 3개 이상이면 분석에 큰 어려움을 겪는 것이 현실이다 보니, 인문학과 의학은 자연과학보다 훨씬 많은 변수를 취급해야 하지만 수학과 같은 정밀한 도구를 사용할 수 없어 자의적이고 사기성이 농후한 통계 확률에 의존하고 있다.
물리학, 수학 등 자연과학의 뭉치화 기준은 숫자화(numericalization)가 대부분이다. 즉 연구 대상에 숫자를 대응시키고 그 숫자의 같고 다름을 기준으로 연구 대상들을 뭉치화한다. 예를 들면 키, 몸무게, 혈압, 키(사람)=168, 품사(단어)=명사 또는 부사, 동사 등에서 보듯이, 중고등학교에서 배우는 ‘함수’라는 개념이다.

이 논문에서는 함수의 개념을 문장 속의 단어를 어떻게 분석할 것인가 또한 단어 묶음에 어떻게 적용할 수 있는가를 중심으로 살펴보고자 한다. 보다 정확하게 말하면, 특정 단어에 대해 그 단어의 특성을 나타낼 수 있는 표현방법(representation)이 복수인 경우, 그 주변 단어들과

---

[1] 언젠가는 실현될 제임스 카메론 감독의 터미네이터, 스카이넷.

[2] 로봇을 만드는 과정 중, 음성을 문자로 번역(voice recognition), 문장은 기계언어로(machine code) 바뀌어야 한다는 사실은 번역의 의미가 외국어 번역만으로 한정 시킬 수 없다.

의 연관성 등을 고려하여 유일한 표현방법을 결정하는 프로그램을 도출하는 것이다. 이 프로그램의 개념적인 설명은 다음과 같다.

3. 한글 문장 속의 단어들 묶음에의 응용

이 프로젝트의 모든 문장은 표준어이자 맞춤법은 물론, 그에 따른 문장부호를(쉼표, 물음표 등) 올바르게 사용된 것으로 가정한다.[3]

(1) 각 단어의 특징을 표현하는 함수(표현 방법, representation)

우리가 어떤 단어를 접했을 때 그 문법적 특성을 나타내는 개념으로 손쉽게 활용하는 것이 '품사'이다. 그러나 여기에서는 함수의 개념을 품사에 한정하지 않고 일반화시킨다. 즉 품사의 변형과 타입을 지정하여 각 단어를 2개 기호의 묶음으로 표현한다.[4]

예1) 특징(마련할)=(V, E)
1번째 특징 'V'는 동사, 형용사 또는 조사('이다')라는 뜻이고, 2번째 특징 'E'는 자동사나 타동사가 될 수 있고 관형사형 전성어미를 가진다는 의미이다(참고로, 소문자 e는 자동사와 타동사가 될 수 있지만 관형사형 전성어미가 아니라는 의미이다).

예2) 특징(연말에)=(에, t),
'에'는 명사+조사인데 '에'로 표시되었고 't'는 명사가 시간을 나타냄을 의미함.

여기서 필요한 데이터와 도구는 한글 단어들에 대한 전자사전들, 예를 들면 단어 사전(total.dic), 동사 어미 사전 (eomi.dic), 조사 어미 사전(josa.dic) 등이고 분석은 현재 인터넷에 떠도는 형태소 분석기이다.

total.dic에 등록된 단어들을 보면, '마련할'의 용언 '마련'은 마련,e00ei0000X이고, '연말에'의 용언 '연말'은 연말,t00000000X라고 나타난다. total.dic의 단어 정보는 10자리 문자 또는 숫자로 표시된다. 1번째는 명사, 2번째는 동사, 3번째는 부사, 4번째는 '하다' 동사 여부, 5번째 는 '되다' 동사 여부 등을 의미한다. 여기서 '마련'은 명사이고, e는 보통명사 또는 의존명사이고 '하다' 동사로서 자동사 타동사로 쓰일 수 있으며 '되다' 동사는 자동사라는 의미이다[5].

josa.dic의 단어 정보는 2자리 문자로서 '에' 조사에 대한 특정값을 '에'로 정한다는 의미로서
에 에
와 같이 등록되어 있다.

그럼 이제부터 위의 특징 함수(?)를 주위 단어들과 어떻게 연관시킬 것인가의 문제가 남는다.

(2) 문장 속 단어들 간의 연관성 맺기(association)

지난 10여 년간 개발된 한글형태소 분석기를 실제로 돌려보면, 위의 예들은 [마련(N),하(t),ㄹ(e)] [연말(N),에(j)] 처럼 단 하나의 분석결과만을 가지고 있다. 반면에 '가는'의 경우에는 가늘(J),ㄴ(e) 품사=V, J 가(V),는(e) 품사=V, W 갈(V),는(e) 품사=V, T 가(N),는(j) 품사=는, 1과 같이 4가지 분석이 가능하다. 이들 중 하나를 컴퓨터가 결정하게 하기 위해서 우리 인간 자신이 문장을 읽으면서 어떻게 아는가에 대한 단순한 사실을 인식하기만 하면 된다.
'단어는 그 단어가 속한 문장 속에서 의미가 살아있다'는 말처럼 우리는 문장을 읽음과 동시에 앞뒤 단어들과 비교하고 연관성을 인식하며 단 하나의 분석결과를 선택한다. 이 과정을 컴퓨터 프로그램화하면 되는 것이다. 요즘 빅데이터(big data)가 많이 거론되지만 모든 단어들에 연관성을 부여하는 것은 비효율적이며 거의 불가능하다. 그러므로 단어들의 연관성을 찾는 일반적 규칙을 발견하여 적용하고, 특별한 경우에만 연관되도록 하는 것이 자연스럽다[6].

---

[3] 입력될 문장의 단어들에 대한 맞춤법 등 선행절차가 필요하다는 것

[4] 특징1, 특징2로 부르기로 하자

[5] total.dic에서 등록된 명사, 동사, 부사 등에 대한 표시표는 뒤에 첨부한다

[6] memory.txt라는 단어연관성을 기록하는 사전을 작성함. 예를 들어, 명사 '말'과 연관된 단어들을 말(N

예문: '저기 공을 차는 아이가 제 아들입니다'

이를 형태소 분석을 하면, 저기(D) 공을(O) 차는(가능한 분석들 2개 중: 차(V),는(e) 품사=V, N 차(N),는(j) 품사=는, e) 아이가(가) 제(가능한 분석들 3개 중: 적(N),에(j) 품사=에, e 제(P) 품사=D, m 제(N) 품사=N, E) 아들입니다(이)[7]라는 결과가 나온다.
여기에서 '차는'과 '제'에 대한 분석결과가 결정되지 않았다. 이를 해결하기 위하여 그 두 단어들의 앞과 뒤에 있는 것들의 특징들을 나열하면(특징들 나열한 것을 '구조단어'[8]라고 정의한다), '차는'의 '공을 차는' OEVN, OE는e의 2가지 경우 '제'의 '제 아들입니다'는 에e이p , Dm이p , NE이p의 3가지 경우가 발생한다. 여기서 프로그램은 OEVN과 Dm이p이 등록된 structureDic.txt(구조단어들을 등록한 사전)를 참조하여 '차는'과 '제'를 분석한다. 참고로 StructureDic.txt의 일부를 보면, 다음과 같다.
DjOx D,O
DjOz D,O
DjP1 D,P
VE에d V,에
VE에1 V,에
일단 위와 같이 문장 내의 모든 단어들의 용언과 타입을 정하게 되면, 주어, 보어, 서술어 등을 묶어주는 작업을 할 준비가 된 셈이다.

(3). 구나 절에 해당되는 단어들 묶기(recursive routine)

제1단계: 프로그램의 데이터 구조 중의 하나인 tree에 문장 내의 각 단어들을 저장한다.
제2단계: 각 단어의 그 자체, 특징 등을 고려하여 그 다음 단어와 묶어야 하는가 또는 명사, 명사+조사인 경우, 앞에 단어 동사인 경우, (동사, 명사 또는 명사+조사)에는 영어의 관계대명사 절 또는 구에 해당하는지 여부를 체크하여 필요시 적용하고 아닌 경우는 다음 단어로 나아간다.
제3단계: 제2단계를 반복적으로 적용한다[9].

가. 2단계 작업과정의 사례:

ㄱ. (부사, V) 는 V를 꾸미는 부사인 경우에 V로 묶어준다. 위의 표현으로 하자면, BV=>V로 바뀐다.

ㄴ. (명사+조사, 동사, 명사+조사)의 경우, 특징에서 특징1만을 모은 '로VS, 로V에, 로V은, 에V에, 엣V에'와 표현되는 경우들이 있고, 그러한 경우에 관계대명사 내지 구를 나타내는 구조 단어들의 사전을[10] 참조하여 단어들을 묶고 그 결과의 특징은 맨 뒤의 특징을 따른다.
예를 들어,
저기 공을 차는 아이가의 특징은 (가,p)
저기 공을
저기
공을
차는
아이가
제 아들입니다(이, p)
제
아들입니다.

고려(N 그대로(B 마리(N 신라(N 중얼(B 듣(V 표현(N 한마디(N 못(N 속셈(N 쓰(V 자르(V와 같이 나열함
[7] D: 관형사 또는 지시대명사 를 표시, O:목적어, N: 명사, P:인칭대명사 가:명사+조사이며 조사가 '가'
[8] 각 단어의 명사에 따른 조사와 동사의 어미들의 특징들을 나열한 구조단어들이 단어들 사이를 연관성을 보여주는 연결고리인 것.
[9] 프로그램에서 'if' loop로 실현
[10] althought.txt, after.txt 등이 있다

(4) 관계대명사 내지 구를 나타내는 사전 사용에 대한 구체적인 예
백미보다 섬유질과 영양은 풍부하지만, 딱딱하고 거칠어 아이들에게 먹이기 적합하지 않다.
 가. 등록되지 않은 경우
백미보다
섬유질과 영양은
섬유질과
영양은
풍부하지만
딱딱하고 거칠어
딱딱하고
거칠어
아이들에게 먹이기
아이들에게
먹이기
적합하지 않다
적합하지
않다
 나. although.txt[11]에 댄은V가 등록된 경우
백미보다 섬유질과 영양은 풍부하지만
백미보다
섬유질과 영양은
섬유질과
영양은
풍부하지만
딱딱하고 거칠어
딱딱하고
거칠어
아이들에게 먹이기
아이들에게
먹이기
적합하지 않다
적합하지
않다

여기에서 풍부하지만의 어미 '지만'이 영어의 although에 해당되고 although.txt를 참조하면 '댄은V'라는 구조단어가 등록된 것을 인식하여 그에 해당되는 단어들을 묶어 준다.

절 안에 '구'나 절이 있는 예:
'기예르모 델 토로'는 할리우드에서 가장 음울한 기운을 스크린에 투영하는 감독이다
'기예르모 델 토로'는
할리우드에서 가장 음울한 기운을 스크린에 투영하는 감독이다(절)
할리우드에서 가장 음울한 기운을(절 또는 구로 표현)
할리우드에서
가장 음울한

11 구조단어들을 등록한 structureDic.txt와 마찬가지로 이것도 단어들 사이의 연관성을 나타내는 일반적 형태의 구조단어들이 등록된 구조단어 사전

가장
음울한
기운을
스크린에
투영하는
감독이다

(4) 해결해야 할 과제- ‘뿐 아니라’, ‘얼마나’ 등이 포함된 관용적 표현에 대한 구조 단어들의 사전 구축[12]
예: ‘뿐 아니라’가 들어가는 문장, ‘소화가 잘 안 될 뿐 아니라 죽을 쑤어도 으깨기 어렵고 낱알이 남기 때문에 12개월 이후에 시도하는 게 좋다.

원하는 결과:
소화가
잘 안 될
  잘 안
    잘
    안
  될
뿐 아니라 죽을 쑤어도 으깨기 어렵고 낱알이 남기
    뿐
    아니라
    으깨기
    어렵고
    낱알이
    남기
때문에
12개월 이후에 시도하는 게
    12개월 이후에
      12개월
      이후에
    시도하는
    게
좋다

여기서 관건은 ‘뿐 아니라’ 이하를 묶는 방법으로는 먼저, ‘뿐 아니라’ 뒤에는 조사 ‘도’가 붙는 명사가 오거나 어미에 ‘도’를 포함하는 동사가 온다는 사실을 주목하고, 구조단어 사전 구축 내지 구조단어의 패턴으로 매칭시키는 방법을 생각할 수 있다.

4. 전망

단어 묶음이 완결되면, 묶음들과 묶음 내의 단어들의 순서를 외국어 어순에 맞추어 바꾸고 그 단어들에 해당되는 외국어 단어 내지 구를 바꿔 넣는다[13].

[12] 국어 연구 결과들을 구조단어들과 융합시키는 작업
[13] 기존의 한글 단어 대 단순 외국어 단어가 아닌 한글 구문에 대한 외국어 구문을 대응 시키는 사전이 필요하다.

참고문헌

1. 이수명(2013), arirang-analyzer, http://cafe.naver.com/korlucene
2. 다움 인터넷 한국어 사전

부록

total.dic 사전에 등재된 단어들의 10자리 정보들 표현하는 기호[14]

1. 부사(B): 3번째 위치

기호	설명	예
a	조사가 붙지 못하는 부사	
v(verb)	동사가 부사로 된 경우	그러다가
n(not)	동사 앞이나 명사 뒤에 씌여 부정을 나타내는 것	안, 못, 여간
f(first)	문장 처음에 나와야만 부사가 되는 것	물론
F	부사가 아니나 편리를 위하여 부사로 표시	작은, 막냇, 푸른, 푸짐
j	형용사 같은 부사	아메리칸 리그의 아메리칸
t	시간을 나타냄	그때
s	소리를 나타냄	쩍쩍
g	강조의 뜻	잘, 매우
l(link)	문장을 연결시켜 주는 것	한편, 그러나, 허나
c(connect)	단어, 구를 연결시켜 주는 것	및
w(where)	장소나 위치를 나타냄	거기다가
r(refer)	앞의 것을 언급	그렇게, 그야
d(degree)	정도를 나타냄	그만큼, 그토록
y	행태 모양을 나타냄	가득가득
b=(s, y)	소리와 모양을 동시에 나타냄	
I,i, j, t, e	부사에 하다, 거리다, 대다가 붙어 서술어가 되는 경우, 동사의 구분을 따른다. I는 I와 같되 조사가 붙어 부사가 되는 것. 이와 마찬가지로 대문자로 표시된 거은 조사가 붙는다.	느긋하다, 느긋거리다, 느긋대다 히죽이
m	명사, 부사 가능성이 있으나 조사 없이는 부사로 쓰이는 것	무작정, 잠깐
와=wa=where+add		거기다

2. 서술어(V, 동사, 형용사) 2번째 위치

기호	설명	예
i, j, t, k, l	자동사, 형용사, 타동사, 보조동사, 보조 형용사	
a	i, l	터지다
b	i,l,t	빠지다
c	앞에 것을 언급하며 연결하는 것	그러다
d	i,j, l	성하다
e	i, t	
f	j, l	뻔하다
g	j,t	걸다, 싫다
h	i, j 즉 자동사 형용사 되는 것	

[14] 여기 기호는 필자가 기본 원리 및 structureDic.txt, although.txt 등 구조단어 사전들을 만들기 실험적인 것뿐이다. 따라서, 일관성만 있다면 얼마든지 마음대로 정할 수 있다. 일관성 있고 와 닿는 기호를 만드는 것은 중요한 일이다.

m	i, j, k	있다, 들다
n	i,j, t, 즉 자동사, 형용사 타동사가 되는 것	마르다
q	i,k	죽다
r	k, l	
s	k, t	제치다
u	k, l, t	아니하다
v	i,k,l,t	하다, 보다
w	i,k,t	먹다
x	j, k, l, t	못하다
y	i, j, k, t	보이다, 되다
이	명사+이다 조사가 붙는 것	

3. 명사(N): 1번째 위치: 명사(N)

기호	설명	예
a	동물	
b	피부 접촉과 관계되는 명사	
c	운반수단, 대문자 C는 DictionaryUtil에서 복합명사로 등록됨	차, 배
d	의존명사 ㄱ. 보편성: 것, 분, 이, 데 ㄴ. 주어성: 지, 수, 리, 나위 ㄷ. 서술형: 때문, 나름 뿐, ㄹ. 부사성: 대로, 듯, 뺀, 통 ㅁ. D: 단위성: 마리, 대, 자, 분, 특별한 단위성에 대하여 대문자로 표시한다(2014.8.9일 수정) ㅂ. 의존명사 것+을=걸을 의존명사 취급한다	게, 것, 셈
e	d, 1 의존명사, 보통명사	수
f	음식 명사	
F	존재하지 않으나 편리를 위한 명사	일루수의 루수
g	기관 또는 단체를 나타냄	
j	관형적 성격을 갖는 명사	당면
하다(4번째), 되다(5번째) 위치의 i, j, t, e	명사들 중 하다, 되다가 붙어 동사가 된다. 그런 경우 4번째, 5번째 위치에 표시되며 동사의 규칙을 따른다	화려하다(형용사), 공부되다(자동사)
k	코 냄새에 관계되는 명사	
K	화학 물질	
l(알파벳)	식물을 나타냄	
m	d, 1, n 의존명사, 보통명사, 수사	이
M	1, w 보통명사, 장소 명사	
n	수사	삼십, 몇몇
o	나라를 나타냄	미국, 스페인
O	외래어	마스카라
p	사람을 나타냄	
P	특별한 사람을 나타내는 고유명사	

q	w, 1, n 장소, 보통, 수사	
Q	w, d, 1, n 장소, 의존, 보통, 수사	구
r	귀 또는 듣는 것과 관계되는 명사	
s	성질을 나타냄	개연성
t	시간을 나타냄	순식간
T	특별한 시간을 나타냄	아침, 저녁
u	n,p 수사, 인칭명사	
U	측정단위 표시	km, mm, mg
v	d, n 의존명사, 수사	투
w	장소를 나타냄	
W	장소를 나타내는 고유명사	
x	1, n 보통, 수사	
y	명사가 접미사 ㅁ이 붙어 명사가 된 경우	
인	사람이름 표시	
잡	직위 내지 신분을 나태내는 명사, 보통명사(ㅈ,ㅂ)	
잣	장소, 직위를 나타냄	시장
퓨	음식과 장소를 동시에 나타냄	
6번째 위치(IDE_NE=5)에 있는 i,j,e,t	명사들 중 '나다, 내다'가 붙어 동사를 만든다. 그 품사를 나타낸다. * (나, 내)= ㄱ.(i, i), (i, 0), (0, i)=i ㄴ.(i, e)=j ㄷ.(i, t)=e ㄹ.(t, t)=t ㅁ.(j, 0), (0, j), (j, j)=k	결딴나다, 결딴내다
1(숫자)	보통명사	
2	도서상 처럼 상 등은 복합명사로 취급하지 않는다. 예외적으로 암달러상을 복합명사로 취급하기 위하여 2라 표시	
4	한자성어	

4. 대명사(P): 8번째 위치 IDX_NPR=7

기호	설명	예
a	f, j, p, w	그
b	f, j, w 보통명사	예
e	f, p	
f	지시 대명사	
j	관형사	그깟
J	수 관형사	서, 두
p	인칭	소비
P	우리, 나 같은 특수한 인칭 대명사	
q	j, p	모
r	p, w	자
t	f,j 지시대명사, 관형사	그따위

v	f, p, w	여
w	감탄사	어쩜
x	f, w	어, 뭐
y	j, w	온
z	j, p, w	네

5. 복합명사(C): 9번째 위치(IDX_IDX_CNOUNX=8)

기호	설명	예
i	사람을 나타내는 '인'을 복합명사 구성원으로 택한다	외국인
b	복합명사의 앞 단어를 구성할 수 없다	
f	복합명사의 뒷 단어 구성 성분일 수 없다	
n	복합명사를 취하지 않는다	

6. 10번째 위치

기호	설명	예
F	외래어로서 명사의 'O'와 달리, 조사가 붙지 못한다	
H	주로 한자어로서 조사나 어미가 붙어야만 하는 명사들	간의, 해도
N	ㄴ, ㄹ이 붙을 수 없는 명사	찬찬이 '찬차+ㄴ' 될 수 없다