(제 10회 한글 및 한국어 정보처리 학술대회)

한국어 영형 대명사의 식별 알고리듬

이 춘숙
충남대학교 언어학과
대전시 유성구 궁동 220.우: 305-764

노 용균
충남대학교 언어학과
대전시 유성구 궁동 220. 우 : 305-764
ynoling@hanbat.chungnam.ac.kr

An algorithm for identification of zero pronouns in Korean

Yi Chunsuk
Department of Linguistics,
Chungnam National University

No Yongkyoon
Department of Linguistics,
Chungnam National University

요약

이 논문은 대용어의 한 유형으로 인정되는 영형 대명사를 식별하기 위한 것이다. 이를 위해서는 한국어 통사 규칙들과 사전 항목들이 필요하다. 사전 항목들은 각각 자질과 값을 갖고, 통사 규칙 내부에는 이런 자질과 값들이 명세된다. 이 통사 규칙들을 토대로 하여, 발화체에 통사 구조들을 부여한다. 영형 대명사는 자질과 값을 명세한 통사 규칙을 씀으로써 식별이 가능하다.

영형 대명사는 주어와 보충어로 나뉘는데, 영형 주어는 동사가 머리인 S의 subj 자질 값이 cov(covert)일 때 식별된다. 영형 보충어는 다시 명사구와 동사구의 covc (covert complement) 자질 값이 0이 아닐 때 식별된다. 이러한 자질과 값으로 영형 대명사를 식별하는 하나의 알고리듬을 제안한다.

1. 서론

하나의 발화체 또는 발화체와 발화체 사이에서 앞서 나온 어떤 요소를 다른 간략한 요소로 대체하거나 그 요소를 드러내지 않는 경우가 많다. 한국어는 그런 요소들이 드러나지 않을 수 있기 때문에 많은 다른 언어들과는 다르다. 예를 들어, 영어에서 문장의 주어는 반드시 나타나야 함은 물론이고, 머리인 동사가 요구하는 요소도 대다수의 입지(registers)에서 필수적으로 나타나야 한다. 그러나, [10]에 따르면 한국어 발화체는 주어가 없는 경우가 55% ~ 65% 정도이다. 실제로 한국어에서는 동사가 요구하는 필수적인 요소도 나타나 있지 않은 경우가 많다.

(1) 처음엔 2천 불렀어.

예문 (1)에서 '불렀어'라는 동사의 주어가 나타나 있지 않다. 또, '불렀어'는 '~을' 또는 명사구와 '~으로'라는 후치사구를 요구한다. (1)에서는 보충어 두 개중 명사구만 나타났고, '~으로'라는 후치사구는 나타나지 않았다. MBC 텔레비전 연속극 "그대 그리고 나" 의 첫회 방송에서 동규부의 친구가 한 말로서의 (1)은 "민규한테 맞은 아이의 부모가 처음엔 2천 만원을 치료비로 불렀어"를 의미한다. 이 의미는 맥락에 의해 결정된다.

이와 같이 나타나지 않은 요소를 '형태가 없는 대명사(zero pronoun)'라고 일컫는다. 이러한 영형 대명사는 앞서 나온 요소를 대체한 대용어와 동등한 지위를 갖는다. 이 논문은 영형 대명사의 식별을 위한 것이다. 영형 대명사가 지칭하는 대상은 다루지 않는다.

영형 대명사가 들어 있는 문장을 그것의 식별 없이 컴퓨터로 자동 처리 하기에는 많은 어려움이 뒤따른다. 그러므로, 영형 대명사를 어떻게 찾을 것인가를 고려해야 한다. 구체적으로 고려해야 할 점은 다음과 같다.

우선 대부분의 동사는 주어를 요구한다. 그 주어가 주어 표지(이/가)를 갖는 경우와 표지가

없는 경우를 포착해야 한다. 또, 의무적으로 주어가 없어야만 정문이 되는 경우도 아울러 포착해야 할 것이다. 두 번째는 동사가 요구하는 보충어를 알아야 한다. 보충어에 관해서는 [7]을 참조하라. 보충어를 필요로 하지 않는 것, 보충어 하나를 필요로 하는 것, 보충어 두 개를 필요로 하는 동사들이 있다. 세 번째는 명사가 요구하는 보충어를 살펴보아야 한다. 대부분의 명사들은 보충어를 필요로 하지 않지만, 어떤 명사들은 보충어를 필요로 하기 때문이다.

이와 같은 문법관을 바탕으로 한국어 통사 규칙과 사전을 구축해 Allen[6]의 Parser로 자연스런 대화체 문장 (MBC 텔레비전 연속극 "그대 그리고 나"의 1997년 10월 11일 첫회 방송 대본의 대사) 106개 (평균 발화체의 길이 6.8)에 대한 올바른 분석을 찾고, 발화체 122개는 사전의 항목과 통사 규칙을 추가하여 그 결과를 보기로 한다[1]. 동사가 나타나 있지 않은 연쇄체는 분석 대상에서 제외한다.

2. 문법 규칙의 구성

2.1 개요

이 연구에서 채택한 문법 형식은 자질 통합 (feature unification)능력이 있는 맥락 자유 문법이다.

규칙의 구체적인 형식은 다음과 같다. 각각의 규칙을 괄호쌍에 넣고, 구범주 또는 어휘 범주가 갖는 자질과 값을 또 괄호안에 넣는다. 구범주를 포함한 규칙의 예는 (2)이다.

(2) ((s(vform relclause) (gap gap) (subj ov)) 15 (np) (vp(vform relclause) (gap gap)))

(2)는 vform 값이 relclause이고 gap 값이 gap이고 subj값이 ov인 문장이 np와 vform값이 relclause이고 gap 자질의 값이 gap인 vp를 직접 지배함을 나타낸다.

어휘 범주를 포함한 규칙의 예는 (3)~(5)이다.

(3) ((np (colloc colloc_ha) (covc 0) (gap nogap)) 16 (pp(plex 하고)) (dlm) (adv) (n(colloc colloc_ha) (subcat frame_i7)))

(4) ((vp(vform ?a) (covc 1) (gap nogap)) 1006 (v(vform ?a) (subcat frame_i5)))

(5) ((vp(vform ?a) (gap nogap) (covc ?b)) 1086 (pp(plex 부터)) (vp(vform ?a) (gap nogap) (covc ?b)))

(3)은 colloc의 값이 colloc_ha이고 covc의 값이 0이고, gap의 값이 nogap인 np가 pp와 dlm과 adv와, colloc 값이 colloc_ha이고 subcat 값이 frame_i7인 n을 직접 지배한다는 것을 나타내는 규칙이다. (4)는 covc의 값이 1이고, gap의 값이 nogap인 vp가 subcat자질의 값이 frame_i5인 v를 직접 지배 하며 v의 vform 값이 자신을 지배하는 vp의 vform 값을 결정한다는 사실을 나타내 준다. 규칙 (5)에 따르면, 머리인 후치사가 '부터'인 후치사구(pp)가 gap값이 nogap인 vp와 결합해 vp를 이루며, 이때 어머니인 vp의 vform 자질의 값과 covc 자질의 값은 딸인 vp의 vform 자질 값, covc 자질의 값과 같다.

사전의 각항목도 괄호쌍안에 낱말을 넣고 범주, 자질과 자질값을 쓴다.

(6) (같지 (v (vlex 같) (vform root) (subcat frame_i12)))
(울기(v (vlex 울) (vform nomclause) (subcat frame_i14)))
(하니까(v (vlex 하1) (vform advclause) (subcat frame_t11)))
(출발(n (subcat frame_i14) (colloc colloc_ha) (daa minus)))
(먹자구(v (vlex 먹) (vform iquote) (subcat frame_t11)))
(하셨어야(v (vlex 하1) (vform iquote) (subcat frame_t11)))
(죠(v (vlex zero_say) (vform root) (subcat frame_i10)))
(알아(v (vlex 알) (vform root) (subcat frame_t11)))
(써두(v (vlex 쓰) (vform advclause) (subcat frame_t11)))
(나온다는(v (vlex 나오) (vform relclause) (subcat frame_t8)))
(해(v (vlex 하1) (vform iquote) (subcat frame_t11)))
(주고(v (vlex cu3) (vform root) (subcat frame_i11)))

1 Allen의 parser의 source 파일은 "ftp://ftp.aw.com/cseng/authors/allen/NatLang2e/"에서 얻을 수 있다.

(이야(v (vlex copula) (vform root) (subcat frame_i12)))
(달래서(v (vlex 달래) (vform advclause) (subcat frame_t11)))
(깎은(v (vlex 깎) (vform nouncomp) (subcat frame_t5)))
(되(v (vlex 되1) (vform nouncomp) (subcat frame_i4)))

(6)은 사전 항목의 예이다. 각 항목은 (낱말꼴 (어휘범주명(자질1 값1)(자질2 값2)...(자질n 값n))) 의 꼴을 갖는다.

2.2 자질과 자질값

이 논문에 등장하는 subcat 자질 및 colloc 자질의 분류는 [2]를 바탕으로 한다.

자질의 값은 그 자체가 범주가 아니다. [9]의 AGR나 SLASH와 같이 값이 복잡한 범주일 수 있는 자질은 쓰이지 않는다. gap이라는 자질은 관계절 내부의 구속된 변인(bound variable)을, vform은 동사의 굴절을 표시하기 위해 쓰인다.

(7) 자질들과 그 값들

자질 명	값	영역
subcat	frame_t1 등 25개	N, V
colloc	colloc_ha,colloc_stat_ha, colloc_toi, colloc_nA, colloc_plain	N,NP, PP
gap	gap,nogap	NP,VP,PP, S
vform	root,nomclause,advclause, relclause,iquote,nouncomp govern,serial	V.VP, S
subj	ov, cov	S
covc	0, 1, 2, 3	NP, VP
uaa	plus, minus	N,NP
fst	fst, nfst	N,NP
vlex	'오' 등 415개	V
plex	'으로' 등 14개	P,PP
pos	pos, nonpos	N

이 연구에서 이용되는 문법 규칙들과 사전 항목들에 쓰이는 모든 자질이 표 (7)에 제시되어 있다.

2.3 영형 대명사 표시

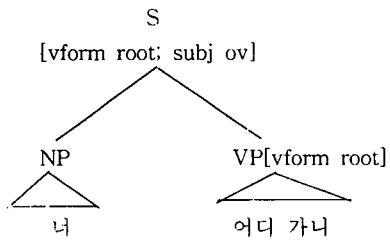
2.3.1 영형 주어

문장은 subj자질의 값으로 obligzero(의무적으로 주어가 영형), ov, cov를 갖는다. subj자질의 값이 obligzero (obligatory zero)인 문장의 예는 (8)의 최상위문이다. '같'이라는 동사는 머리가 '것'인 명사구를 보충어로 갖고, 이 '것'의 보충어인 S의 vform 자질의 값이 nouncomp일 때, 의무적으로 주어를 갖지 못한다. 계사(copula)도 마찬가지이다. subj 자질의 값으로 ov(overt)를 갖는 문장의 예는 (9)인데 '가니'의 주체가 되는 '너'가 나타나 있다. subj 자질의 값이 cov(covert)인 것의 예는 (10)인데 이 문장에는 '와요'라는 동사의 주어가 나타나 있지 않다.

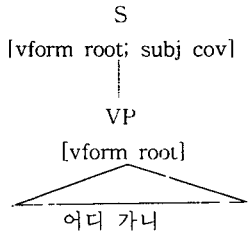
(8) 이제 오실 분은 거의 다 오신 것 같지?
(9) 너 어디 가니?
(10) 금방 와요.

subj값이 ov인 문장과 subj 값이 cov인 문장의 수형도는 각각 (11), (12)에 의해서 예시된다.

(11)



(12)



2.3.2 영형 보충어

동사가 요구하는 보충어의 종류를 기준으로 동사를 분류하면 25가지가 있다.

예를 하나 들면, (13)은 subcat 자질의 값으로 frame_t11을 갖는 동사를 포함한 예이다.

(13) 주말의 귀중한 시간을 내 (주셔서...)

'내다'라는 동사는 '~을/ ~를'이라는 후치사구 또는 명사구를 보충어로 요구하며, (13)에서는 그 보충어가 드러나 있다. (14)도 (13)과 마찬가지로 가능한 예이다. frame_t11 부류의 동사들은 항상 후치사가 없는 명사구를 허용한다. 이 변동은 통사 규칙으로 포착할 수 있다.

(14) 주말의 귀중한 시간 내 (주셔서...)

동사구는 covc라는 자질의 값으로 0, 1, 2를 갖는다. covc의 값이 0이면 동사가 요구하는 모든 보충어가 나타났다는 것을 의미하고, 1은 요구되는 보충어(들)중 하나가 나타나 있지 않았다는 것을, 2는 보충어 두 개가 나타나지 않았음을 표현한다. 또, covc 자질의 값이 3이면 나타나 있지 않은 보충어의 수가 3개임을 뜻한다.

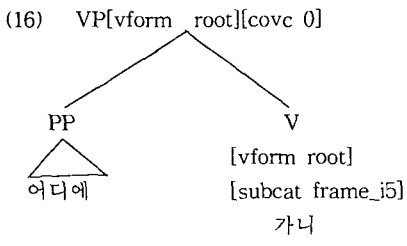
어떤 명사들은 보충어를 요구한다. 명사도 동사와 마찬가지로 subcat라는 자질의 값을 갖는다. 예를 들어 (15)에서 '통보'라는 명사는 '~을/를'의 후치사구를 보충어로 요구한다.

(15) 아우, 불참을 미리 통보해야 했는데...

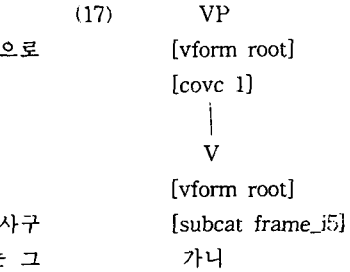
명사구도 covc값을 갖는데, 0이면 명사가 요구하는 보충어가 모두 나타난 것이고, 1이면 그 명사구안에는 보충어 하나가 결여되어 있다는 것이다.

covc의 값이 0인 동사구와 covc의 값이 1인 동사구를 다음의 (16)과 (17)이 예시한다.

(16)



(17)



3. 결과 및 논의

이 논문의 연장선상에 있는 중의성의 문제가 거의 대부분의 문장에서 나타났다. (18)은 발화체의 복잡성으로 인해 많은 양의 분석 결과를 낳는다. 분석 결과가 많으면 parser의 기억 용량의 한계 때문에 그 결과를 다 얻는 데에 실패할 수 있다. 그래서, '여기 ~ 힘 들어서' 가 부사절(S)를 구성하므로, (19a)와 (19b)로 나누어 parse했다. (19a)가 최종적으로 이루는 S는 상위에 있는 또 다른 S에 지배받는 S이다. 이 S를 대신하는 낱말로 vform 자질의 값이 advclause이고 gap 자질의 값이 nogap인 sad라는 가상적인 어휘를 사전에 추가했다. (19a)의 내부 중의성은 여기에서 설명하지 않겠다. (19b)에서 '입사 고시라구 해'의 발화체는 160가지 분석 결과가 나온다.

(18) 여기 입사 하기가 별 따기보다 힘 들어서 입사 고시라구 해

(19) a. 여기 입사 하기가 별 따기보다 힘 들어서
b. sad 입사 고시라구 해

이 160가지 의미해석의 가능성을 살펴보자. 먼저 (19b)가 이루는 최종적인 통사 범주는 VP, S 일 가능성이 있다. 두 번째 중의성의 요인은 '해'의 vform자질 값이 root , govern , iquote , advclause, serial일 가능성이다. 세 번째 중의성의 요인은 '입사' 가 S('고시라구 해')의 부가어, '라구'의 주어, 그리고 '해'의 주어일 가능성과, '고시'와 결합해 np를 이룰 가능성이 있다. 네 번째 요인은 '라구' 의 vform자질 값이 iquote, advclause 일 경우이다. 다섯 번째 요인은 부사절 sad(19a)가 '입사 고시 라구'의 부가어, '입사 고시라구 해'의 부가어 일 가능성이다.

세 번째 중의성의 요인 중에서 '입사' 가 S의

부가어일 경우는 (20)의 '아가씨' 와 같이 호격 명사구로 해석된 경우이다. 실제로 이런 해석은 불가능하나 적절히 제약을 가해서 이 해석을 제지하는 방법을 찾지 못했다.

(20) 아가씨! 뭐 해요?

그렇다면 거의 모든 한국어 발화체를 분석하는 데에 어느 정도의 통사 규칙이 필요할까? 이를 예측하기 위해 먼저 일정한 수의 발화체를 분석하고, 나중에 일정한 수의 발화체 분석을 통한 통사 규칙의 증가 추이를 본다.

최초 106개의 발화체를 분석한 결과 통사 규칙의 수는 225개이고, 사전 항목의 수는 405개였다. 그 다음 122개의 발화체에 대해 올바른 분석을 얻기 위해 추가해야한 통사 규칙은 46개, 사전 항목은 335개였다. 이 두 무리를 통틀어서 S를 구성하는 통사 규칙은 32개, VP를 구성하는 규칙은 168개, NP를 구성하는 규칙은 69개였다. PP를 이루는 규칙은 2개다.

(21) 통사 규칙과 사전 항목의 증가 추이

	통사 규칙의 수	사전 항목의 수
첫 106개 발화체	225 개	405 개
그 다음 122개 발화체	271 개 (46개 증가)	740 개 (335개 증가)

(21)에서 볼 수 있듯이 사전 항목의 증가율은 높지만 통사 규칙의 증가율은 낮다. 더 많은 자료를 분석하면 통사 규칙의 증가율은 (21)에서보다 훨씬 낮아 질 것으로 예상한다. 그렇게 되면 앞으로 기존의 통사 규칙들을 이용해 어떤 종류의 문장에 대해서도 거의 대부분 옳은 분석을 얻을 것이다.

대용어에 대한 기존의 연구 [1], [4], [5], [8]은 드러나 있는 대용어가 가리키는 것이 무엇인지 밝히는 데 목적을 둔다. 영형 대명사의 존재가 분명함에도 불구하고, 이것의 식별을 위한 논의는 거의 없다. 영형 대명사의 식별을 논의하기 위해서는 동사들과 명사들을 보충어의 종류로 구분, 즉 하위 범주화를 구축해야 하는데 여기에 어려움이 있기 때문인 것 같다. 낱말들에 하위 범주화의 틀을 부여하는 일은 [3]에서 지적되다시피, 자동화 될 수 없고, 전문가의 수작업을 필요로 하기 때문이다.

영형 대명사의 식별을 기반으로 영형 대명사가 가리키는 것이 무엇인가를 밝히는 방법에 관한 연구가 남은 과제라고 하겠다.

참고 문헌

[1] 김 정해, 조 준모, 이 상국, 이 상조. "중심어 주도 단방향 차트 파싱을 이용한 문맥 대용어 해결". 제 8회 한글 및 한국어 정보처리 학술대회 논문집. 386-392. 1996.

[2] 노 용균. "한국어 동사와 명사 사이의 하위범주화에 있어서의 평행성". 언어와 정보 제 1권. 한국언어정보학회 . pp. 27-65. 1997.

[3] 류 법모, 장 명길, 박 수준, 박 재득, 박 동인. "구문구조부착 말뭉치를 이용한 술어의 하위범주화 정보 구축". 제 9회 한글 및 한국어 정보처리 학술대회 논문집. pp.116-121. 1997.

[4] 이 익환. "Pronominal Anaphora in Korean". 어학 연구 14. pp. 63 - 99. 1978.

[5] 이 홍배. "Notes on Pronouns , Reflexives, and Pronominalisation", 어학 연구 12 : 2. pp.253 - 263. 1974.

[6] Allen, James. *Natural Language Understanding.* Redwood City, California : The Benjamin/ Cummings Publishing Company. 1995.

[7] Borsley, Robert D. *Syntactic Theory A Unified Approach.* London : Edward Arnold. 1991.

[8] Chang, Suk-Jin. "Anaphora in Korean". in J. Hinds, ed., *Anaphora in Discourse,* pp.223 - 278. Edmonton : Lingustics Research, Inc. 1978.

[9] Gazdar G, E Klein, G Pullum, I Sag. *Generalized Phrase Structure Grammar.* Oxford : Basil Blackwell. 1985.

[10] Kim, Young-joo. "Null subjects in Crosslingustic Acquisition Data and Theoretical Implications". 제 7회 한글 및 한국어 정보처리 학술대회 논문집. pp.264-280. 1995.