한국어 복합문의 영 대용어 해결

김미진° 강보영 구상옥 박미성 이상조
경북대학교 컴퓨터공학과
{jean321,comeng99,tomato}@sejong.knu.ac.kr {mspark,sjlee}@knu.ac.kr

Zero Anaphora Resolution in Korean Complex Sentences

Mi-Jin Kim° Bo-Yeong Kang Sang-Ok Koo Mi-Sung Park Sang-Jo Lee
Dept. of Computer Engineering, Kyungpook National University

요 약

본 논문은 한국어 복합문에서의 영 대용어 해결을 위해 복합문 분해 알고리즘과 영 대용어 복원 규칙을 제안하고, 해결 방법을 제시한다. 복합문 분해를 위해서는 복합문 구성에 관여하는 활용 어미들을 이용하고, 영 대용어 복원을 위해서는 생략될 때 적용된 통사규칙을 역으로 이용한다. 제안한 방법을 이용한 결과 전체 영 대용어 중 83.53%가 해결 가능하며 11.52%는 부분적으로 해결 가능하다.

1. 서 론

자연언어에는 한 문장 속에 동일한 구조가 둘 이상 있을 때 이 중에서 하나만을 남기고 나머지 것은 생략을 하거나 더 짧은 다른 표현으로 대치하는 과정이 있다. 이러한 생략과 대치는 같은 유형의 언어 현상으로서 생략은 영 대치(또는 영 대용)라고 볼 수 있다[1,2]. 한국어에서의 영 대용은 복합문에서 빈번하게 발생하므로 자연어처리 시스템 구축을 위해서는 영 대용어 처리 능력이 필수적으로 요구된다. 영 대용 해결을 위한 연구로는 중심화이론(Centering)을 이용하여 호텔 예약대화에 나타나는 한국어 영 대명사의 선행사를 찾는 연구가 있었고[3], 또 중심화 이론과 개념 그래프를 이용하여 체언의 대용과 생략을 시도한 연구[4]와 HPSG 파서에 기반하여 한국어 조응대용어를 해결한 연구가 있었다[5]. 그러나 아직도 독립된 부분으로 영 대용을 다룬 연구는 거의 없는 실정이다.

본 논문에서는 복합문 구성에 관여하는 어미를 이용하여 복합문을 분해하고, 통사규칙을 역으로 이용하여 생략된 성분을 복원한다.

본 논문의 구성은 다음과 같다. 2장에서는 한국어 복합문의 통사규칙과 어미분류를 살펴보고 3장에서는 제안한 복합문 분해 알고리즘과 복원규칙에 대하여 설명하고, 4장에서는 실험결과 및 분석을, 5장에서는 결론 및 향후 과제를 제시한다.

2. 영 대용어 해결을 위한 한국어 복합문고찰

이 장에서는 한국어 복합문을 정의하고, 생략 시 적용되는 통사규칙과 복합문 구성에 관여하는 어미들을 분류한다.

2.1 복합문

본 논문은 보조용언 내포문을 제외한 복합문을 다루며 다음과 같이 정의한다[6].

정의 1 : 접속문(Conjunctive Sentences)
접속문은 선행절과 후행절로 구분되고, 접속어미 '고, 어/아(서), 며...'등에 의해 후행절에 연결된다.

정의 2 : 내포문(Embedded Sentences)
내포문은 상위문과 내포절로 구분되고, 명사화 어미 'ㅁ, 음, 기'와 함께 사용되는 명사화(nominalized) 내포문은 명사화 어미에 의해 술어에 연결되고, 관형화 어미 '은, 는, 을'과 함께 사용되는 관형화(adnominal) 내포문은 관형화 어미에 의해 피수식어에 연결된다.

정의 3 : 인용문(Quotations)
인용문은 상위문과 인용절로 구분되고, 인용어미 '(이)라고', '(하)고'에 의해서 술어에 연결된다.

2.2 복합문 어미 분류

한국어는 어미에 의해 문장의 문법적인 관념들이 다양하게 실현된다. [표 1]은 복합문 구성에 관여하는 어미 분류[7]에 본 논문 처리에 필요한 어미를 첨가한 어미분류 테이블이다. '...명사류, 명사류'와 '~등' 내부의 쉼표는 접속어미에서 제외시킨다.

본 논문은 [표 1]의 어미유형에 따라 서로 다른 분해 알고리즘을 적용시킨다

[표 1] : 어미 유형에 따른 어미분류 테이블

복합문	대등	접속어미(conj)		면서, 지만, 든지, 고, (으)며...
	종속	접속어미(conj)		니까, 어/아(서), 며, 쉼표(,)...
	관형화	관형화어미(adn)		-은, -는, -을, -ㄴ, -ㄹ
	명사화	명사화어미(nom)		-음, -기, -ㅁ, -ㄴ/ㄹ 것
인용문	직접·간접	인용어미(quot)		-고, -(이)라고, -하고
		종결어미	명령형	-(으/아/어)라
			의문형	-냐, -(느)ㄴ가, 나
			청유형	-자
			서술형	-다, -라, -구나(감탄형) -마(약속형), -ㄹ 것

2.3 변형(통사) 규칙[8]

국어의 생략현상을 다루는 변형규칙으로는 다음의 규칙들이 있다.

(1) 접속 삭감 규칙

접속문이 형성될 때 접속성분의 일부(동일 성분)가 탈락되는 규칙이다.

예) 영희는 자전거를 타고, 철수는Ø 밀었다.(Ø=자전거를)

(2) 관계대명사 탈락규칙

관계절의 수식을 받는 명사구와 관계절 내의 어떤 명사구와의 사이에 상호 지시성이 성립될 때 적용되는 규칙이다.

예)순경이 [Ø보석을 훔친] 도둑놈을 잡았다.(Ø=도둑놈이)

(3) 동일명사구 탈락규칙

일정한 조건 아래서 상위문의 주어를 삭제시키는 통사규칙이다.

조건1) 상위문장의 주어와 보문의 주어가 같을 때

예) 영희는 [Ø 성실한 사람이 되기를] 원한다.(Ø=영희가)

조건2) 상위문장의 간접 목적어와 보문의 주어가 같을 때

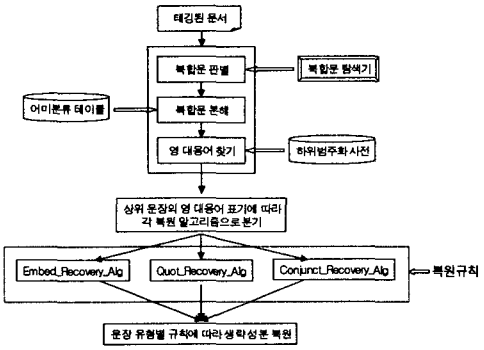
예) 선생님이 학생들에게 [Ø조용히 하라]고 말했다.(Ø=학생들이)

본 논문에서는 생략될 때 적용된 위의 규칙들을 역으로 문장의 생략성분 복원정보로 사용한다.

3. 복합문 영 대용어 해결 시스템

3.1 전체 시스템 구성도

복합문 영 대용어 해결을 위한 본 논문의 전체 시스템 구성도는 [그림 1]과 같다. 입력으로 태깅된 문서를 문장 단위로 받아들여 복합문 판별 과정과 복합문으로 판별된 문장을 EE(Elementary Event)[9]단위로 분해하는 복합문 분해 과정, 그리고 생략된 성분인 영 대용어 탐색 과정을 거친 후 각 문장 유형별 규칙에 따라 생략 성분을 복원한다. 이 들 과정에서 사용되는 정보는 명사 의미 사전, 어미 분류 테이블, 용언의 하위범주화 사전, 통사규칙, 영 대용어 복원 규칙이 있다.



[그림 1] : 영 대용어 해결을 위한 전체 구성도

3.2 복합문 분해 알고리즘

복합문 분해는 가장 상위문장부터 시작해서 내포된 문장으로 내려가며 수행하는데 인용문, 명사화 내포문, 접속문, 관형화 내포문 순으로 분해한다. 분해 시 [표 1]의 어미 분류 테이블을 참조한다.

다음은 복합문중 명사화 내포문 분해 알고리즘이다.

```
/* 한 어절씩 읽어가며 주어와 명사화어미를 찾고,
 주어 다음 어절부터 명사화 어미까지 분해한다. */

Function NE_Split_Alg(Sentence sent)
  Vector ejuls = sent.ejul
  For ( i=1; i<ejuls.size; i++ )
    If (sent.ejul[i] == IsSubject_JOSA)
      start = i+1
    Else If (sent.ejul[i] == IsNE_Ending)
          end = i-1
        End If
    End If
  End For
  return Make_SubSentence(sent, start, end)
End Function
```

3.3 복합문 영 대용어 복원 규칙

각 문장유형에 따라, 생략될 때 적용된 통사규칙을 역으로 이용하여 복원규칙을 생성한다. 아래의 복원 규칙들에서 '=>'는 왼쪽 성분이 오른쪽 성분으로 대치됨을 의미하고, MODee(Modifiee)는 피수식어를 나타낸다.

접속문은 '접속 삭감 규칙'을, 관형화 내포문은 '관계 명사구 탈락 규칙'을, 명사화 내포문은 '동일 명사구 탈락 규칙'을, 그리고 인용문은 '주어 인칭제약에 따른 동일 명사구 탈락 규칙'을 역으로 이용하여 생략 성분 복원 규칙을 생성한다.

[표 2]는 각 문장유형에 따른 복원규칙들이다.

[표 2] : 각 문장유형에 따른 복원규칙

접속문 : [Ante] conj [Cons$_n$]'		
C_1	Csub$_n$ 생략 시	Csub$_n$=>Csub$_{n-1}$
C_2	Cobj$_n$ 생략 시	Cobj$_n$=>Cobj$_{n-1}$
C_3	Csub$_n$&Cobj$_n$생략 시	Csub$_n$&Cobj$_n$=>Csub$_{n-1}$&Cobj$_{n-1}$
:		
관형화 내포문 : [Hsub] [[AE$_n$] adn] MODee Hpred		
E_1	AEsub 생략 시	AEsub=>MODee
E_2	AEobj 생략 시	AEobj=>MODee
E_3	AEadv 생략 시	AEadv=>MODee
E_4	AEsub&AEobj 생략 시	AEsub=>Hsub, AEobj =>MODee
E_5	AEsub$_n$&AEsub$_{n+1}$ 생략	AEsub$_n$&AEsub$_{n+1}$=>MODee
명사화 내포문 : Hsub [[NE$_n$] nom] Hpred		
E_6	NEsub$_n$ 생략 시	NEsub$_n$ => Hsub
:		

Ante : 선행절, Cons : 후행절, Csub : 후행절 주어
Cobj : 후행절 목적어, AE : 관형화 내포절, AEsub : AE 주어
AEobj : AE 목적어, AEadv : AE 부사어
NEsub : 명사화 내포절 주어

위의 규칙에서 문장성분들은 축약하여 표기하였고, 지면 관계상 모든 규칙을 표기하지 못했다.

4. 실험 결과 및 분석

실험 대상 문장은 증권, 금융, 부동산등 신문 기사문 1328문장이다. 이 중 단문이 302문장, 복합문이 1026문장이고 내포된 문장유형은 모두 1718번 출현하였다. 다음의 [표 3]은 각 문장유형별 출현빈도와 복합문의 인식 및 분해 정확도를 나타낸다. 실험 결과, 복합문의 인식 정확도는 91.56%, 분해 정확도는 96.76%로 매우 높은 것으로 나타났다.

복합문의 인식 및 분해 정확도는 다음 식에 의한다.

$$\text{인식 정확도} = \frac{\text{시스템이 인식한 내포된 문장 유형수}}{\text{말뭉치에 출현한 각 문장 유형수}}$$

$$\text{분해 정확도} = \frac{\text{시스템이 정확하게 분해한 문장 유형수}}{\text{시스템이 인식한 각 문장 유형수}}$$

[표 3] : 복합문 인식 정확도 및 분해 정확도

	출현 회수	출현 빈도(%)	인식 오류	인식 정확도(%)	분해 오류	분해 정확도(%)
인용문	602	35.04	0	100	0	100
명사화 내포문	109	6.34	0	100	0	100
접속문	537	31.25	26	95.17	2	99.68
관형화 내포문	470	27.36	79	83.16	34	91.22
계	1718	100	105	**91.56**	36	**96.76**

[표 4] : 문장 유형별 영 대용 분포

영대용 / 문장유형	접속문	인용문	명사화 내포문	관형화 내포문	전체
문장 수	364	524	100	184	1172
분포(%)	31.06	44.71	8.53	15.70	100

[표 5] : 전체 문장에서의 영 대용어 처리

영 대용어	문장 수	빈도(%)
해결	979	83.53
부분 해결	135	11.52
해결 불가능	58	4.95

복합문 분해 알고리즘을 통해 분해된 문장유형 중 영 대용어가 야기된 문장 수는 1172개였다. [표 4]는 문장 유형에 따른 영 대용어 출현빈도이고 [표 5]는 전체 문장에서의 영 대용어 처리 결과인데, 해결된 문장 수는 979문장이고 부분 해결 문장은 135문장, 해결 불가능한 문장은 58문장으로 각각 83.53%, 11.52%, 4.95%를 차지한다.

5. 결론 및 향후 과제

본 논문은 한국어 영 대용어 처리를 위해 복합문 분해 알고리즘과 영 대용어 복원 규칙을 제안하고, 해결방법을 제시했다. 이 때 복합문 구성에 관여하는 어미들과 생략될 때 적용된 통사규칙을 역으로 이용한다. 제안한 방법을 이용한 결과 전체 영 대용어 중 83.53%가 해결가능하며 11.52%가 부분해결 된다.

앞으로 보조 용언 내포문에서의 생략 성분 복원에 대한 연구가 더 필요하다.

참 고 문 헌

[1] 한재현, "생략과 대용현상", 한신문화사, 1984
[2] 양명희, "현대국어 대용어에 대한 연구", 서울대 박사논문, 1996
[3] Ik-Hwan Lee, Minhaeng Lee, "On the Anaphora Resolution in Korean Dialogues.", In : Harvard Studies in Korean Linguistics Vol. 7, 1999
[4] 한승연, "지식 기반을 이용한 대용어 해결 시스템" 연세대학교 전산과학과 석사학위논문, 1995
[5] 김정해, "HPSG 파서에 기반한 한국어 문맥 조응 대용어의 해결", 경북대학교 박사학위 논문, 1996
[6] 권재일, "국어의 복합문 구성연구", 집문당, 1985
[7] 김옥, "국어 접속문의 생략에 관한 고찰", 전남대학교 교육학과 국어교육전공 석사학위논문, 1998
[8] 남기심, "언어학 개론", 탑출판사, 1985
[9] Saliha Azzam, "Resolving Anaphor in Embedded Sentences", 1997