KIPS Tr. Software and Data Eng.
Vol.2, No.12 pp.873~880 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2013.2.12.873

Antecedent Identification of Zero Subjects using Anaphoricity Information and Centering Theory

Kye-Sung Kim† · Seong-Bae Park†† · Sang-Jo Lee†††

ABSTRACT

This paper approaches the problem of resolving Korean zero pronouns using Centering Theory modeling local coherence. Centering Theory has been widely used to resolve English pronouns. However, it is much difficult to apply the centering framework for zero pronoun resolution in languages such as Japanese and Korean. Since in particular the use of non-anaphoric zero pronouns without explicit antecedents is not considered in the Centering Theory of Grosz et al., the presence of non-anaphoric cases negatively affects the performance of the resolution system based on Centering Theory. To overcome this, this paper presents a method which determines the intra-sentential anaphoricity of zero pronouns in subject position by using relationships between clauses, and then identifies antecedents of zero subjects. In our experiments, the proposed method outperforms the baseline method relying solely on Centering Theory.

Keywords : Zero Pronoun, Centering Theory, Anaphoricity Determination, Non-Anaphoric, Local Coherence

조응성 정보와 중심화 이론에 기반한 영형 주어의 선행사 식별

김 계 성† · 박 성 배†† · 이 상 조†††

요 약

본 논문은 지역적 응집성을 모델링하는 중심화 이론을 이용하여 한국어 영형대명사의 지시해결에 접근한다. 중심화 이론은 영어 대명사의 해결을 위해 널리 사용되고 있지만, 일본어, 한국어 등의 언어에서 나타나는 영형대명사 해결에 중심화의 프레임워크를 적용하는 데에는 많은 어려움이 따른다. Grosz et al.의 중심화 이론은 지시적 표현들의 비조응적 사용을 고려하지 않으므로, 문서에 나타나는 비조응적 기능의 영형대명사가 중심화 이론을 이용한 영형대명사의 선행사 식별에 중요한 영향을 미친다. 본 논문은 이를 위해 먼저 절 간의 결속 관계를 이용하여 영형대명사, 특히 영형주어의 문장 내 조응성을 결정하고, 다음으로 중심화의 순위를 이용하여 그 영형의 선행사를 식별하는 방법을 제안한다. 실험을 통해 조응성 결정을 이용하는 제안한 방법이 이를 이용하지 않는 베이스라인 시스템보다 우수함을 알 수 있었다.

키워드 : 영형대명사, 중심화 이론, 조응성 결정, 비조응적, 지역적 응집성

1. 서 론

영어와 달리, 한국어, 일본어, 중국어 등의 언어에서는 문장의 필수 성분들의 생략이 빈번하다. 생략된 문장요소의 지시(reference)는 영 조응(zero anaphora), 영형대명사(zero pronoun) 등의 문제로 알려져 있다. 자연언어 이해 및 처리를 위해서는 지시적 표현의 지시대상을 제대로 찾아내야만 그 지시어의 의미를 명확히 할 수 있으며, 이러한 과정이 지시 해결이다. 한국어 정보처리를 위해 영형대명사의 지시는 반드시 해결해야 할 과제이며 기계번역, 문서요약, 음성인식 후처리 등과 같은 여러 응용을 위해서도 중요한 문제이다.

영형대명사 해결에 관한 최근의 연구들은 기계 학습을 이용하여 영형과 선행사 간의 조응 관계를 분석하고자 하였다[1-3]. 이들은 영형대명사의 선행사 식별을 위한 자질집합을 정의하고 성능 향상에 기여하는 자질의 유형을 파악하는 데 중점을 두었다. 하지만 다른 언어와 달리, 한국어, 일본어 등의 언어에서는 영형에 대한 형태적, 지시적 특성이 문장에 명시적으로 드러나지 않기 때문에 이들에 대한 정보를 영형대명사 해결에 이용하기가 쉽지 않다.

텍스트 분석적 관점에서 보면 생략된 형태로 나타난 영형은 문장들을 서로 연결시켜주는 언어적 장치로서의 역할을

※ 본 연구는 2012(2013)학년도 경북대학교 학술연구비에 의하여 연구되었음.
미래창조과학부 및 한국산업기술평가관리원의SW컴퓨팅산업원천기술개발사업(SW)의 일환으로 수행하였음[10044494, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발].
† 준 회 원 : 경북대학교 소프트웨어기술연구소 연구원
†† 비 회 원 : 경북대학교 IT대학 컴퓨터학부 교수
††† 정 회 원 : 경북대학교 IT대학 컴퓨터학부 교수
논문접수 : 2013년 8월 8일
수 정 일 : 1차 2013년 10월 23일
심사완료 : 2013년 10월 24일
* Corresponding Author : Sang-Jo Lee(sjlee@knu.ac.kr)

한다. 이것은 텍스트의 응집성을 강화시키는 중요한 요소라 할 수 있다. 따라서 영형대명사의 해결은 두 명사구, 즉, 영형과 지시대상(선행사)의 관계뿐 아니라 그들이 나타난 발화와 발화 간의 관계를 통해 형성된 응집성을 바탕으로 접근할 필요가 있다. 중심화 이론은 텍스트의 지역적 응집성(local coherence)을 모델링한 담화해석의 계산모델로 담화 상에서 대화 참여자들의 관심의 대상이 어떻게 움직이는지를 모형화한 것이다[4]. 이 이론에서 "중심(center)은 한 발화를 인접한 다른 발화들과 연결하는 담화 상의 개체로 파악하며 이들이 지역적 응집성의 주요 요소가 된다고 본다. 이 때문에 중심화의 토대 위에서 영어 대명사의 지시를 분석하기 위한 연구가 주로 이루어져 왔다. 하지만 영어와 다른 특성을 갖는 언어에 중심화의 프레임워크를 적용하기 위해서는 고려해야 할 요소가 여러 가지 있으며[5], 그 중 한국어에서 빈번하게 일어나는 생략 현상은 그것의 적용을 어렵게 하는 주된 요인이 되고 있다.

본 연구는 중심화 이론을 한국어 영형대명사 해결에 적용하기 위해 먼저 절(clause)을 하나의 문장(발화)으로 간주한다. 이것은 영형대명사의 사용이 여러 개의 절들로 구성된 복합문에서 자주 등장하기 때문이다. 절을 발화의 기본단위로 살피고자 하는 연구들은 이미 진행된 바가 있다[6]. 하지만 어떤 유형의 절들이 중심의 갱신 단위인가에 대한 논의는 지금까지도 계속되고 있는 실정이므로, 본 연구는 절의 유형을 구분하지 않고 모든 절들을 각각의 발화로 고려한다. 대신 중심의 갱신에 영향을 미치지 않는 영형대명사를 미리 식별하여 발화들을 계층적으로 구성함으로써 중심화의 개념을 그대로 적용할 수 있다. 이를 위해 본 연구는 영형대명사의 조응성 결정을 이용하며, 이 때 영형대명사의 지시의 방향을 함께 파악한다. 이 조응성 결정을 통해 주어진 영형대명사가 명시적인 지시를 수행하고 있는가를 분석하며, 조응적이라 판단된 영형대명사에 대해 지역적 결속성을 모델링하는 중심화의 개념을 적용함으로써 한국어 영형대명사의 지시 해결을 수행할 수 있다.

성능 평가를 위해 본 연구는 영형대명사 중에서 가장 높은 비중을 차지하는 영형주어(zero subject)의 문장 내지시 해결에 제안한 방법을 적용해 보았으며, 실험을 통해 조응성 결정과 중심화 이론을 이용한 제안한 방법이 기존의 중심화 이론을 채택한 방법보다 우수함을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구들을 살펴보고 3장에서 제안하는 방법에 대해서 기술한다. 4장에서는 성능 평가 및 분석에 대한 설명을 하고 마지막으로 5장에서는 결론 및 향후 연구 과제에 대하여 기술한다.

2. 관련 연구

영형을 포함한 대명사의 지시 해결에 관한 연구는 크게 규칙 혹은 이론에 기반한 방법과 기계 학습에 기반한 방법으로 나뉘어진다. 먼저, 규칙이나 이론에 기반한 방법은 주로 중심화 이론에 그 기반을 두고 있다[4]. 지역적 결속성에 기반을 두고 있는 중심화 이론은 대명사의 지시해석을 위한 모델로써 영어권에서 주로 사용되어져 왔다. 기존 중심화 프레임워크에서는 한 문장을 발화의 기본 단위로 보며 문장 단위로 중심이 갱신된다. 하지만, 일본어, 한국어 등의 언어에서는 복합문에서 나타나는 문장요소들의 생략이 빈번하다. 이는 한 문장에서 둘 이상의 영형이 나타날 수 있음을 의미하며 기존 프레임워크에서는 이를 모두 해결하기가 어렵다. 이 때문에 여러 연구들은 기존의 센터링 모델 안에서 복합문을 어떻게 분석할 것인가를 고려하고 있지만[7], 아직까지 명확한 기준이 정립되어 있지 않은 상태이다. 최근 몇몇 연구들은 생략된 형태로 나타난 영형대명사 해결에 중심화 이론을 이용하고 있지만[8], 대부분 후행조응적 영형과 비조응적 영형의 출현을 고려하지 않고 있다. Yeh and Chen[9]의 연구는 영형대명사의 선행사 식별 전에 후행조응적 기능의 영형을 제거하기 위한 휴리스틱 룰을 제안하였지만, 관형절에서 나타난 영형대명사를 모두 후행조응적 영형으로 간주하는 데 그치고 있다.

다음으로 기계학습에 기반한 방법들이 연구되고 있다. 이들은 영형대명사와 선행사 후보 사이의 선호도를 측정하기 위하여 다양한 유형의 자질들을 제안하고 있다. Zhao and Ng[1]은 영형대명사, 선행사 후보, 그리고 영형대명사와 선행사 후보 사이에서 나타나는 구문 정보를 추출하여 선행사 식별을 위한 자질 집합을 구성하였으며, 최근에 여러 연구들은 지시적 표현이 나타난 문장의 파스트리(parse tree) 구조를 직접 사용하여 선행사 식별에 이용하고 있다[2,3]. 하지만 지시대상이 명확하게 나타나지 않는 비조응적 영형의 존재는 영형대명사 해석에 부정적인 영향을 미치게 된다. 이에 최근 여러 연구들은 영형대명사 해결의 성능을 향상시키기 위해 영형의 조응성 정보를 이용하려고 한다[2,3]. 그들은 선행사 식별 전에 비조응적 영형을 미리 제거시킴으로써 영형대명사 해결의 전체적인 성능을 높이고자 하였다. 하지만 제안된 방법들의 대부분이 사용자에 의해서 설정된 파라미터 값이나 휴리스틱한 정보들에 의존하고 있는 실정이다. 좀 더 최근에는 조응성 결정과 공지시(coreference) 모델의 공동 결정에 의해 영형대명사를 해결하려는 노력들이 진행되고 있지만[10,11], 조응성 결정과 선행사 식별 사이의 의존관계를 파악하는 것이 쉽지 않다.

생략된 요소가 지시하는 대상을 식별하는 작업은 언어분석을 위해 매우 중요하다. 하지만 국내에서는 아직 이에 대한 연구가 부족한 실정이다. 본 연구는 담화분석적 접근방법으로 널리 알려져 있는 중심화 이론을 활용하여 한국어 영형대명사 해결에 접근하고자 한다.

3. 영형주어의 선행사 식별

영형대명사 해결은 문장에 나타난 생략된 영형이 지시하는 대상을 복원해내는 작업이다. 본 연구는 한국어에서 가장 높은 빈도를 나타내는 영형대명사, 즉 영형주어의 선행사 식별에 영형의 조응성 정보와 중심화 이론을 적용하여

그 유용성을 검증한다. 다음은 영형대명사가 나타난 문장의 예를 보여준다.

(1) (Φ_1) 스승의 충고를 들으며 헤밍웨이는 노트르담의 초라한 셋방에서 양파 수프와 값싼 포도주와 물만 마시며 (Φ_2) 정진을 계속하였다.

(2) 바다가 얼면 해빙 위에 큰 운동장이 생기므로 (Φ_3) 이번 겨울에도 기지 앞 바다가 두껍게 얼기를 기다린다.

먼저, 문장 (1)에서 영형대명사 Φ_1 과 Φ_2는 모두 같은 문장 안에 있는 '헤밍웨이'를 각각 지시하고 있다. 하지만 이들은 지시의 방향에서 서로 구분이 되며, Φ_2의 지시대상은 영형대명사보다 앞에서 언급되으므로 Φ_2를 선행조응적 영형이라 한다. 반면 Φ_1은 후행조응적 영형대명사의 한 예를 보여주고 있으며, 그것의 지시대상은 뒤따르는 문맥을 통해 발견할 수 있다. 하지만 이러한 선행, 후행 조응 현상 외에도, 영형대명사는 때때로 문서 내에서 그것이 지시하는 대상을 발견하지 못할 수 있다. 문장 (2)에 나타난 Φ_3는 불특정 다수를 가리키는 비조응적 영형대명사의 한 예라 할 수 있다. 이와 같이, 한 문장은 다른 유형의 여러 영형대명사를 포함할 수 있으므로 모든 유형의 영형대명사를 동일한 유형으로 가정하고 접근하는 것은 실제 적용에 적합하지 않다. 본 연구는 다양한 영형대명사의 사용을 고려하면서 한국어 문서에서 나타나는 영형주어의 선행사를 식별하는 데 초점을 맞춘다.

영형대명사 해결에 대한 대부분의 연구들은 영형대명사의 위치가 미리 알려져 있다고 가정한다. 기존 연구들과 마찬가지로 본 연구에서는 영형주어의 위치가 미리 알려져 있다고 가정한다.

3.1 중심화 이론과 고려할 요소들

중심화 이론에서 한 개의 발화는 두 개의 중심을 가지고 있다. 하나는 전향적 중심(forward-looking center: Cf)이며 다른 하나는 후향적 중심(backward-looking center: Cb)이다. 여기서 중심은 발화 시점에서 화자의 의식이 활성화되고 집중되어 있는 대상물들을 말한다. 전향적 중심은 현 발화에 실현된 지시물들을 전향 중심 목록(Cf-list)에 서열에 따라 정렬을 해 둔 것으로 다음 발화에 나타나게 될 지시물에 대한 잠정적인 선행사의 집합이다. Cf-list에 있는 지시물 중에서 가장 높은 서열에 있는 지시물은 선호 중심(preferred center: Cp)이 되며, 선호 중심은 다음 발화에서 주제로 논의될 가능성이 가장 높은 후보자로 선출된다. 후향적 중심(Cb)은 문장의 토픽과 유사한 개념으로 많은 경우 바로 앞 발화의 선호 중심(Cp)이 다음 발화에서 후향적 중심이 된다.

여기서, Cf-list의 순위 결정은 언어에 따라 조금씩 다르며 같은 언어에 대해서도 학자들마다 그 설정 기준이 다르다[8]. 본 연구에서 사용한 전향적 중심의 순위는 topic > subject > object > object2 > others의 순이다. 여기서 주제는 주제 표지 '-은/-는'을 조사로 갖는 단어를 의미한다. 또한 중심의 전이 유형은$Cb(U_i)$와 $Cb(U_i-1)$의 일치 여부, 그리고 $Cb(U_i)$와 $Cp(U_i)$의 일치 여부에 의해 결정되며 그 내용은 Table 1과 같다.

Table 1. Types of transition relations across pairs of utterances

	$Cb(U_i)=Cb(U_{i-1})$ 또는 $C(U_{i-1})=NULL$	$Cb(U_i)\neq Cb(U_{i-1})$
$Cb(U_i)=Cp(U_i)$	CONTINUE	SMOOTH-SHIFT
$Cb(U_i)\neq Cp(U_i)$	RETAIN	ROUGH-SHIFT

Fig. 1은 중심화의 Cf-list 순위에 기반한 영형대명사 해석의 예를 보여준다. 중심화의 토대 위에서 영형대명사의 사용이 중심화의 'CONTINUE' 전이 유형에서 선호된다는 점을 고려하면, Fig. 1과 같이 영형대명사 Φ_1과 Φ_2의 선행사는 동일하게 "공군기"로 결정될 수 있다. 하지만U_1과 U_2사이에 "시계가 나빠"라는 새로운 발화 U^+가 삽입되는 문장을 고려해 본다면, Φ_1의 지시대상은 "시계"로 잘못 판단될 수 있으며 이것이 Φ_2의 지시대상을 결정하는 데에도 마찬가지로 영향을 미치게 된다. 즉 인접한 두 발화 U_{i-1}, U_i의 전이에 초점을 맞추고 있는 중심화의 경우 이전 발화의 잘못된 선행사 식별의 결과가 다음 발화로 계속 전파될 수 있다는 문제가 있다.

(U_1) 공군기가 들어왔으나
  ⇒ Cb={null}; Cf={공군기};
  ⇒ Cp={공군기}
(U_2) (Φ_1) 착륙을 못하고
  ⇒ Cb={공군기}; Cf={Φ_1,착륙};
  ⇒ Cp={Φ_2(=공군기)} // *(CONTINUE)*
(U_3) (Φ_2) 돌아 갔다.
  ⇒ Cb={공군기}; Cf={Φ_2};
  ⇒ Cp={Φ_2(=공군기)} // *(CONTINUE)*

Fig. 1. Antecedent identification of zero pronouns based on the ranking of Cf

본 연구는 이러한 문제를 해결하기 위해 다음의 세 가지 요소를 고려한다. 첫째, 중심화에서는 하나의 문장을 하나의 발화로 간주하기 때문에 한 문장에 여러 개의 영형대명사가 나타날 경우 이들을 함께 다루기가 쉽지 않다. 하지만 한국어 복합문의 경우에 여러 유형의 영형이 같은 문장 안에 함께 나타날 수 있다. 둘째, Grosz et al.[4]의 중심화는 대명사의 지시대상이 선행하는 문맥 상에 존재한다고 본다. 그러나 한국어와 같이 어순이 비교적 자유롭고 생략이 빈번한 언어에서는 그 지시대상이 항상 선행하는 문맥에 존재한다고 가정하기가 어렵다. 마지막으로 중심화에서는 대명사의 지시대상이 문서에 분명하게 나타난다고 인식한다. 하지만, 영형대명사 중에는 명시적인 지시대상을 가리키지 않는 것들도 상당수 존재한다. 따라서 중심화에 기반한 영형대명사

해결의 성능을 향상시키기 위해서는 이에 대한 고려가 반드시 이루어져야 한다.

이러한 문제는 특히 한국어 복합문에서 두드러지게 나타난다. 따라서 본 연구에서는 한국어 복합문에서 나타나는 영형주어와 선행사의 관계를 중심화 이론을 적용해 분석해 보고자 한다. Fig. 2는 본 논문에서 제안한 영형대명사 해결의 전체적인 개요를 보여준다. 제안한 방법은 크게 세 가지 작업으로 나뉘어진다. 먼저, 영형의 조응성 결정 모델을 이용하여 영형의 조응성을 평가한다. 그리고 후행 영형대명사의 필터링을 선행사 식별 전에 고려해 볼 수 있다. 특히 후행조응적 기능의 영형대명사는 관형절에서 자주 나타나므로 이들을 탐지하기 위해 간단한 휴리스틱을 이용하거나 본 연구에서 제안한 조응성 결정에 기반하여 식별하는 방법을 이용할 수 있다. 마지막으로 조응성 결정과 후행조응적 영형의 필터링을 통해 최종적으로 조응적이라 식별된 영형대명사를 대상으로 중심화의 프레임워크 안에서 그것이 지시하는 대상을 식별한다.

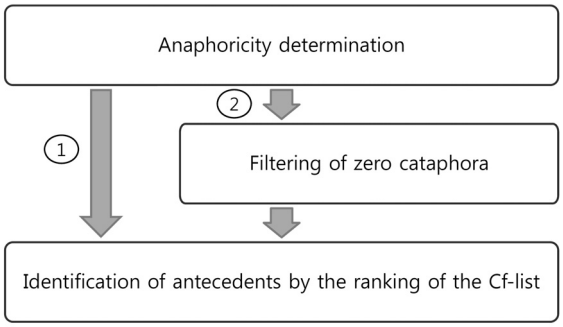

Fig. 2. Overview of zero pronoun resolution

3.2 영형주어의 조응성 결정

조응성 결정(anaphoricity determination)은 주어진 영형대명사가 명시적인 지시대상을 가리키고 있는가를 결정하는 작업이다. 최근 [12]의 연구가 영형대명사가 나타난 절의 구조 정보에 초점을 두고 있는 것과 달리, 본 연구에서는 영형과 지시대상을 가지고 있는 절과 절 사이에서 추출한 구조 정보를 조응성 결정을 위해 사용한다.

텍스트에서 형성된 응집성의 토대 위에서 영형주어와 지시대상이 나타난 절들 사이의 관계성은 영형의 조응성을 구분하는 데 중요한 단서를 제공한다. 주어진 영형에 대해 동일 문장에 그것의 명시적인 선행사가 존재한다면, 영형대명사가 나타난 절과 그 선행사가 나타난 절 사이의 응집성은 문장 내 다른 절들과의 관계보다 높게 나타날 것이다. 따라서 영형대명사와 후보 선행사를 포함한 절 쌍(clause-pair) 후보가 강한 응집성을 나타낸다면 그 영형을 조응적이라 인식할 수 있으며, 그 쌍은 “영형대명사와 지시대상”의 관계를 가지고 있다고 판단할 수 있다. 조응성 결정에 대한 이 작업은 주어진 절 쌍이 “영형과 선행사”의 관계를 가지고 있는지를 판별하는 이진(binary) 분류 문제로 간주할 수 있다. 이를 위해 본 연구는 문장을 여러 개의 절들로 구분하며 이 과정은 한국어 연결어미 정보를 이용함으로써 수행할 수 있다.

Fig. 3은 다음 문장에 대한 파스트리의 예를 보여준다.

U_1 : 공군기가 들어왔**으나**
U_2 : 시계가 나**빠**
U_3 : (zp_1) 착륙을 못하**고**
U_4 : (zp_2) 돌아갔다.

이 문장은4 개의 절(발화)로 이루어져 있으며, 영형주어 zp_1에 대해 세 개의 절 쌍 후보 (U_1, U_3), (U_2, U_3), (U_3, U_4)가 생성된다. 조응성 결정을 위해 본 논문은zp1을 포함하고 있는 U_3와 그것의 명시적인 선행사를 포함하고 있는 U_1 사이의 구조 정보는 긍정적인(1) 학습 예로, 나머지는 모두 부정적인(0) 학습 예로 사용한다.

이 때 두 절을 연결하는 최단 경로 상에 있는 모든 노드와 에지들을 포함하는 경로포함트리(path-enclosed tree)[13]

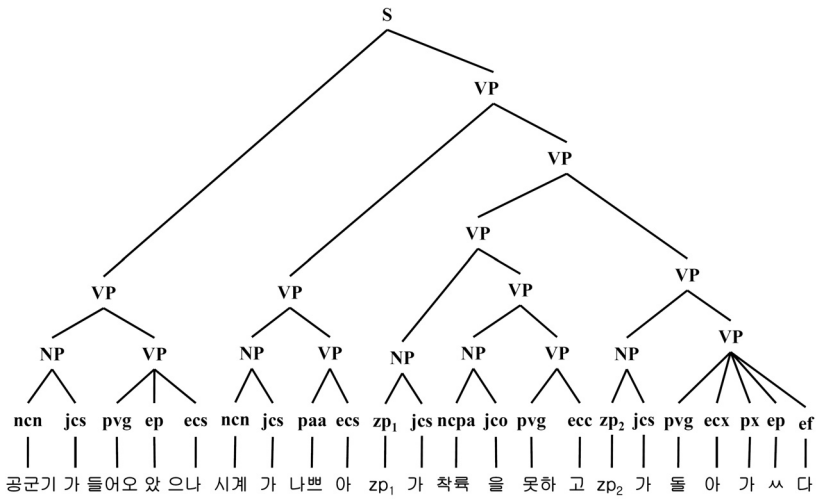

Fig. 3. The parse tree for the example sentence

가 조응성 결정을 위한 구조화된 자질로써 사용되며, Fig. 4는 U_1과 U_3쌍으로부터 얻어진 경로포함트리의 예를 보여준다. Fig. 4에는 파스트리의 단말노드들이 명시적으로 표현되어 있지 않지만, Fig. 3에 나타난 각각의 어휘들이 트리의 단말노드들로 구성된다.

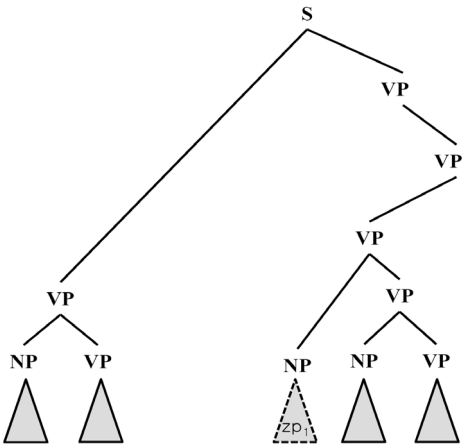

Fig. 4. The path-enclosed tree of the U_1 and U_3

본 논문은 추출된 경로포함트리가 영형대명사의 선행사를 포함할 가능성에 대한 점수를 부여하며 이를 위해 지지벡터기계(Support Vector Machines)를 이용한다. 그리고 절 간의 구조정보를 모델링하기 위해 트리구조를 다루는데 특화된 파스트리 커널[14]을 사용한다. 즉, SVM의 마진값이 모두 0보다 작다면, 그 영형을 비조응적이라 판단하며, 마진값이 1보다 큰 쌍들을 가진다면 그것을 조응적이라 결정하여 그 중에서 최대값을 가지는 절 쌍을 "영형과 선행사"의 관계를 가지고 있다고 판별한다. 조응적이라 결정된 절 쌍은 선행사 식별을 위해 사용되어지며, 이 때 영형을 가지고 있는 절이 상대절보다 앞서 있다면 그 영형을 후행조응적이라 하며, 그렇지 않은 경우에 선행조응적이라 판단한다.

이러한 조응성 결정 과정을 통해 중심의 갱신에 영향을 미치지 않는 영형들은 선행사 식별에서 제외시킬 수 있다. 다시 말해 조응성 결정의 결과를 토대로 Fig. 5와 같이 발화를 계층적으로 구성함으로써 영형주어 zp_1에 대해 현재 발화 U_3와 이전 발화 U_1의 관계를 분석하여 그것의 선행사를 식별하는 것이 가능하다.

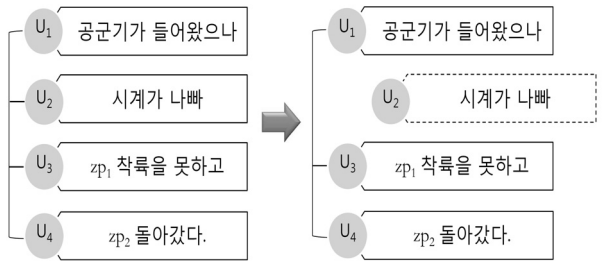

Fig. 5. The hierarchy of utterances in the example sentence after anaphoricity determination

3.3 Cf-list 순위에 기반한 선행사 식별

조응성 결정을 통해 조응적이라 식별된 영형을 대상으로 중심화의 프레임워크 안에서 그것의 선행사를 식별한다. Fig. 6은 선행사 식별의 알고리즘을 나타낸다[9]. 본 논문은 복합문에서 나타나는 문장 내 영형주어의 해결에 초점을 맞추고 있으므로 문장 내에서 그 선행사를 찾지 못한 영형은 문장 간의 지시 해결 문제로 남겨둔다.

즉 조응성 결정을 통해 주어진 영형의 선행사를 가지고 있다고 판단된 절 내에서 중심화의 cf-list와 그 요소들의 순위에 기반하여 그것의 선행사를 식별한다. 이를 통해 인접하지 않은 두 개의 발화는 동일한 수준에 놓이게 되며 다양한 영형의 존재를 고려하면서 중심화의 개념을 이용할 수 있다.

For each zero pronoun z in the set of the utterances of a sentence (*Ui*, ..., *Um*),
1: if the utterance with the antecedent of z is *Ui-k*, then
Choose the noun phrase with the most highest ranking in the Cf-list of *Ui-k* as the antecedent according to grammatical role criteria
topic > subject > object > object2 > others
2: Else
Regard z as inter-sentential (or not intra-sentential).

Fig. 6. An algorithm for antecedent identification of zero pronouns

3.4 파스 트리 커널

Collins and Duffy[14]에 의해 제안된 파스트리 커널은 절의 구조 정보를 모델링하기 위해 사용한다. 파스트리 커널에서 벡터의 자질들은 각 파스트리에 나타날 수 있는 모든 부분 트리(subtree)들로 이루어지며, 각 자질의 값은 부분 트리의 빈도수를 나타낸다. 그러나, 이러한 부분 트리를 명시적으로 구하는 것은 불가능하므로 [14]는 아래 재귀 규칙을 두 트리의 모든 노드에 대해 적용함으로써 명시적인 열거없이 내적을 구하는 방법을 제시하였다.

규칙 1. n_1과 n_2가 다르면
$C(n_1, n_2) = 0$
규칙 2. n_1과 n_2가 단말 노드(leaf node)라면
$C(n_1, n_2) = 1$
규칙 3. 그 외

$$C(n_1, n_2) = \prod_{i}^{nc(n_1)} (1 + C(ch(n_1, i), ch(n_2, i)))$$

이 때, $ch(n_1, i)$는 노드 n_1의 i번째 자식노드를 의미한다. 함수 $nc(n_1)$는 n_1의 자식 노드 수를 반환한다. 위의 알고리즘을 이용하여 파스트리 T_1과 T_2의 내적은 다음과 같이 계산한다.

$$< V_{T_1}, V_{T_2} > = \sum_{n_1 \in N_{T_1}} \sum_{n_2 \in N_{T_2}} C(n_1, n_2)$$

4. 실 험

4.1 실험 데이터

본 연구의 모든 실험은 STEP 2000 과제의 결과물인 STEP 2000 한국어 구문부착말뭉치에 대해서 평가되었다[12]. Table 2는 실험에 사용한 데이터셋에 대한 간단한 통계정보이며, 여기서 하나의 문장은 평균 3.97개의 절로 이루어져 있다.

Table 2. Summary of the dataset used in our experiments

Dataset	Number
Sentences	5,221
Clauses	20,748
Clauses containing zero subjects	13,171

Table 3은 한국어에서 나타난 영형주어의 분포를 보여준다. Table 3에서 볼 수 있듯이, 영형대명사의 많은 수는 그들이 나타난 동일 문장 내에서 해결이 가능하며, 이것은 많은 수의 영형이 다수 개의 절로 구성된 복합문에서 자주 사용되고 있음을 시사한다. 따라서 복합문에서 나타나는 문장 내 영형대명사 해결의 실제적인 응용을 위해서는 계속해서 문장 내에 명시적인 선행사를 가지고 있지 않은 영형들에 대한 보다 많은 논의가 이루어져야 할 것으로 보인다.

Table 3. The distribution of zero subjects

Intra-sentential	Inter- sentential	Extra- sentential
10,371 (78.74%)	666 (5.06%)	2,134 (16.20%)

실험을 위하여 수집한 데이터 집합을 학습(training) 데이터 집합과 실험(test) 데이터 집합으로 나누어 사용하였다. 모든 실험은 5 번의 교차 검증(cross validation)을 통해 수행되었으며, Joachims의 SVM_{light}[15]가 분류기로 사용되었다. 그리고 제안한 모델의 효과를 관찰하기 위해 정확도(accuracy), 정확률(precision), 재현율(recall)의 평가 척도가 사용되었다.

4.2 조응성 결정에 대한 결과 및 분석

제안한 절 간의 구조적 관계성을 이용한 조응성 결정에 대한 성능을 평가하기 위해Kong and Zhou[2]의 모델과 비교하였다. 그들은 조응성 결정을 위해 영형의 주변 문맥의 구조 정보를 이용하였다. 즉, 영형의 가장 가까운 선행하는 술어구 노드와 영형을 후행하는 문맥상에서 가장 먼저 만나는 술어구 노드 사이의 구조 정보를 조응성 결정을 위해 사용하였으며, 추출된 두 술어구 사이에 놓여있는 동사와 명사구에 대한 구조 정보도 함께 이용하였다.

$$Precision = \frac{\text{올바르게식별된비조응적영형대명사의수}}{\text{문장내비조응적이라고식별된영형대명사의수}}$$

$$Recall = \frac{\text{올바르게식별된비조응적영형대명사의수}}{\text{문장내비조응적영형대명사의총수}}$$

Table 4는 조응성 결정에 대한 결과를 보여준다. [2]의 조응성 결정 모델은 정확률은 높은 반면에 재현율이 상당히 낮게 나타났다. 이것은 [2]의 연구에서 사용한 구조 정보가 다양한 문맥 속에 나타난 비조응적 영형대명사의 특징을 표현하기에는 부족함이 있음을 의미한다. 반면 제안한 방법은 [2]의 모델에 비해 재현율을 상당히 향상시키고 있음을 볼 수 있다. 이것은 절을 이루는 구성요소들 간의 통사적 결속 관계가 조응적, 비조응적 영형의 특징을 표현하기에 좀 더 효과적임을 시사하는 것이다. 하지만, 생략된 영형의 조응성 결정은 여전히 어려운 문제이다. 제안한 모델이 [2]에 비해 F1의 관점에서 약 53%의 성능 향상을 가져옴을 확인하였지만, 계속해서 정확률과 재현율을 높일 수 있는 부분에 대한 연구가 이루어져야 할 것이다.

Table 4. Result of anaphoricity determination

	Precision	Recall	F1
Kong and Zhou[2]	0.7397	0.2062	0.3225
Our method	0.5169	0.7364	0.6074

4.3 영형주어의 선행사 식별에 대한 결과 및 분석

본 절에서는 조응성 결정과 Cf-list 순위에 기반한 선행사 식별의 2 단계로 구성된 영형주어의 선행사 식별에 대한 성능을 살펴본다. Table 5는 그 결과를 보여준다.

Table. 5 Result of antecedent identification of zero subjects

	Accuracy	$Accuracy^{EM}$
Baseline	0.4425	0.4682
CT	0.4516	0.4874
CT+Anaphoricity	0.6038	0.7381

여기서, “baseline”은 영형이 발화 U_i에서 나타났다면U_{i-1}에 있는 영형대명사와 가장 멀리 떨어져 있는 명사구를 그것의 선행사로 결정하는 방법이다[9]. “CT”는 Grosz et al.[4]의 중심화를 한국어 복합문에 나타난 영형주어의 해결에 그대로 적용한 결과이다. “CT+Anaphoricity”은 “CT”의 방법에 제안한 조응성 결정 모델을 결합한 결과를 보여준다. Table 5에서 “accuracy”는 문장에 나타난 모든 선행조응적 영형을 대상으로 한 결과이며, “$accuracy^{EM}$”은 내포절에서 나타난 영형들을 모두 후행조응어로 간주한 지시 해결의 결과를 나타낸다.

$$\text{Accuracy} = \frac{\text{올바르게선행사가식별된영형대명사의수}}{\text{문장내선행조응적영형대명사의총수}}$$

Table 5에서 알 수 있듯이 제안한 방법이 첫번째 컬럼과 두번째 컬럼 모두에서 다른 방법들에 비해 높은 정확도를 보여주었다. 이것은 한국어 복합문에서 영형주어의 선행사가 영형이 나타난 발화의 바로 이전 발화에서 발견되지 않는 경우가 많음을 시사하는 것이다. 또한 "CT"는 "baseline"에 비해 더 높은 결과를 보여주었지만 그 차이가 크지 않다. 이는 "CT"에서 1 순위로 보고 있는 토픽에 대응하는 명사구가 문장(혹은 절)의 시작부에서 나타나는 경우가 많다는 것을 말해준다. 즉, 한국어에서는 비조응적 영형의 출현이나 연속적인 주어 생략 등이 자주 발견되기 때문에, 기존의 중심화 프레임워크를 그대로 적용하기에는 문제가 있음을 확인할 수 있었다. 앞으로 조응성 결정에 대한 성능을 만족할 만한 수준으로 높인다면, 영형주어 해결에 대한 결과는 지금보다 더 나아질 것으로 기대한다.

5. 결론 및 향후 연구

문서에서 나타나는 다양한 지시적 표현의 사용은 문장(절) 간의 지역적 결속성 및 현저성과 밀접한 관련이 있다. 본 논문은 중심화 이론을 한국어 영형대명사 해결에 적용하였다. 하지만 기존 중심화 프레임워크에서 고려되지 않던 비조응적, 후행조응적 영형의 출현은 중심화에 기반한 영형대명사 해결의 성능을 저하시키는 주요한 원인이 되며, 본 연구에서는 절 간의 구조적 관계성을 고려하는 조응성 결정을 통해 이러한 문제를 해소하고자 하였다. 실험을 통해 제안한 방법이 영형대명사 해결에 기여할 수 있음을 알 수 있었다.

향후 연구과제로는 조응성 결정에 대한 성능을 보다 개선시킬 필요가 있다. 절 간의 결속관계는 통사뿐 아니라 의미적 접속관계에 의해서도 영향을 받으므로 앞으로 이들을 함께 추론할 수 있는 자질들을 설정하고 그 성능을 보다 향상시키기 위한 연구가 계속적으로 진행되어야 한다. 또한 계속해서 문장 간에서 나타나는 영형과 선행사의 관계를 고려하기 위한 후속 연구가 이루어져야 할 것이다.

참 고 문 헌

[1] S. Zhao and H. T. Ng, "Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach", In Proceedings of the Joint Conference on EMNLP-CoNLL, pp.541-550, 2007.

[2] F. Kong and G. Zhou, "A tree kernel-based unified framework for Chinese zero anaphora resolution," In Proceedings of Empirical Methods in Natural Language Processing, pp. 882-891, October, 2010.

[3] R. Iida, K. Inui, and Y. Matsumoto, "Zero-Anaphora Resolution by Learning Rich Syntactic Pattern Features", ACM Transactions on Asian Language Information Processing, Vol.6, No.4, article 12, December, 2007.

[4] B. J. Grosz, S. Weinstein, and A. K. Joshi, "Centering: A Framework for Modeling the Local Coherence of Discourse", Computational Linguistics, Vol.21 No.2, pp.203-225, June, 1995.

[5] M. Poesio, R. Stevenson, B. D. Eugenio, and J. Hitzeman, "Centering: A parametric theory and its instantiations," Computational Linguistics, Vol.30, No.3, pp.309-363, September, 2004.

[6] Y. Cui, Q. Hu, H. Pan, and J. Hu, "Zero anaphora resolution in Chinese discourse," In Proceedings of Computational Linguistics and Intelligent Text Processing, pp.245-248, 2006.

[7] Y. Sakurai, "Centering in Japanese: How to rank forward-looking centers in a complex sentence," In Proceedings of the 22nd Northwest Linguistic Conference, pp. 243-256, 2006.

[8] Q. Hu, "A corpus-based study on zero anaphora resolution in Chinese discourse," Ph.D. dissertation, Department of Chinese, Translation and Linguistics, City University of Hong Kong, 2008.

[9] C.-L. Yeh and Y.-C. Chen, "Zero anaphora resolution in Chinese with partial parsing based on centering theory," In Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE'03), pp.683-688, 2003.

[10] R. Iida and M. Poesio, "A cross-lingual ILP solution to zero anaphora resolution," In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp.804-813, 2011.

[11] A. Rahman and V. Ng, "Syntactic Parsing for Ranking-Based Coreference Resolution," In Proceedings of IJCNLP, pp. 465-473, 2011.

[12] K.-S. Kim, S.-B. Park, S.-Y. Park and S.-J. Lee, "Anaphoricity determination of zero pronouns for intra-sentential zero anaphora resolution", Journal of KIISE : Software and applications, Vol.37, No.12, pp.928-935, Sep., 2010. (in Korean)

[13] M. Zhang, J. Zhang, and J. Su, "Exploring syntactic features for relation extraction using a convolution tree kernel," In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp.288-295, 2006.

[14] M. Collins and N. Duffy, "Convolution Kernels for Natural Language", In Proceedings of Neural Information Processing Systems, pp.625-632, 2001.

[15] T. Joachims, "Making large-Scale SVM Learning Practical", Advances in Kernel Methods - Support Vector Learning, B.Scholkopf and C.Burges and A.Smola (ed.), MIT-Press, 1999.

김 계 성

e-mail : kskim@sejong.knu.ac.kr
2012년 경북대학교 컴퓨터공학과(박사)
2013년~현 재 경북대학교 소프트웨어 기술연구소 연구원
관심분야 : 자연어처리, 정보추출, 기계학습

박 성 배

e-mail : sbpark@sejong.knu.ac.kr
2002년 서울대학교 컴퓨터공학과(박사)
2004년~현 재 경북대학교 IT대학 컴퓨터학부 교수
관심분야 : 기계학습, 자연어처리, 텍스트 마이닝, 정보추출, 생명정보학

이 상 조

e-mail : sjlee@knu.ac.kr
1994년 서울대학교 컴퓨터공학과(박사)
1976년~현 재 경북대학교 IT대학 컴퓨터학부 교수
관심분야 : 자연어처리, 기계번역, 정보검색, 정보추출