문장 내 영 조응어 해석을 위한 영대명사의 조응성 결정

(Anaphoricity Determination of Zero Pronouns for Intra-sentential Zero Anaphora Resolution)

김 계 성 † 박 성 배 †† 박 세 영 †† 이 상 조 ††
(Kye-Sung Kim) (Seong-Bae Park) (Seyoung Park) (Sang-Jo Lee)

요 약 문서에서 생략된 요소가 지시하는 대상을 식별해 내는 작업은 기계 번역, 정보추출 등과 같은 자연언어처리 분야의 다양한 응용들을 위해 필요하다. 문장에서 생략된 요소들은 영조응사, 영대명사 등으로 불리며, 지시(reference)의 한 유형으로 간주되고 있지만, 모든 영형이 문서에서 명확하게 언급된 지시 대상을 지시하지는 않는다. 이에 영형의 조응성을 결정하려는 연구가 최근 진행되고 있으며, 본 논문에서는 한국어에서 가장 빈번하게 나타나는 영형 주어(subject zero pronouns)의 문장 내 조응성 결정에 초점을 맞춘다. 주어진 영형과 선행사 후보들 간의 쌍대 비교(pairwise comparison)에 기반한 기존 연구와 달리, 본 논문은 비조응적 혹은 문장 간에서 해결 가능한 영형이 나타난 절의 구조를 직접 학습함으로써 영형의 문장 내 조응성을 결정한다. 실험에서 제안한 방법은 베이스라인보다 나은 성능을 보였으며, 영형의 조응성 결정은 향후 영형 조응어 해석에 긍정적인 영향을 줄 수 있을 것으로 기대된다.

키워드 : 영형, 영대명사, 문장 내 조응성 결정, 문장 내 영 조응어 해결

Abstract Identifying the referents of omitted elements in a text is an important task to many natural language processing applications such as machine translation, information extraction and so on. These omitted elements are often called zero anaphors or zero pronouns, and are regarded as one of the most common forms of reference. However, since all zero elements do not refer to explicit objects which occur in the same text, recent work on zero anaphora resolution have attempted to identify the anaphoricity of zero pronouns. This paper focuses on intra-sentential anaphoricity determination of subject zero pronouns that frequently occur in Korean. Unlike previous studies on pair-wise comparisons, this study attempts to determine the intra-sentential anaphoricity of zero pronouns by learning directly the structure of clauses in which either non-anaphoric or inter-sentential subject zero pronouns occur. The proposed method outperforms baseline methods, and anaphoricity determination of zero pronouns will play an important role in resolving zero anaphora.

Key words : zero pronoun, intra-sentential anaphoricity determination, intra-sentential zero anaphora resolution

· 이 논문은 2009년도 경북대학교 학술연구비에 의하여 연구되었음

† 학생회원 : 경북대학교 컴퓨터공학과
kskim@sejong.knu.ac.kr
†† 종신회원 : 경북대학교 IT대학 컴퓨터학부 교수
seongbae@knu.ac.kr
seyoung@knu.ac.kr
sjlee@knu.ac.kr
(Corresponding author임)

논문접수 : 2010년 9월 16일
심사완료 : 2010년 10월 14일

정보과학회논문지 : 소프트웨어 및 응용 제37권 제12호(2010.12)

1. 서 론

한국어, 일본어, 중국어 등의 언어에서는 반복된 요소들의 생략이 빈번하다. 문장에서 나타난 생략된 요소들의 지시(reference)는 영 조응(zero anaphora), 영대명사(zero pronoun) 등의 문제로 알려져 있으며, 최근 영형의 지시 해석에 대한 관심이 높아짐에 따라 관련 연구들이 활발히 진행되고 있다[1-4]. 영 조응어 해석은 자연어 이해에서 매우 중요한 문제 중 하나이며, 문서에 나타난 영형이 지시하는 대상을 식별하는 작업은 기계 번역, 정보추출 등과 같은 다양한 자연어 처리 응용을 위해 필요하다.

영형을 포함한 지시 해석에 관한 기존 연구들은 대부분 지시적 표현과 그 선행사 사이의 성, 수 인칭 등에 관한 일치를 중요한 요소로써 평가한다. 하지만, 한국어에서 이같은 실마리 정보를 분석하는 것은 쉽지 않으며, 특히 지시적 표현이 생략된 형태로 나타나는 영형에 관한 국내 연구는 아직 미흡한 실정이다.

영 조응어 해석은 크게 영형의 조응성 결정과 선행사 식별의 단계로 나누어볼 수 있다. 조응어 해석에 관한 기존 연구들은 대부분이 영형의 선행사가 문서 상에 존재한다고 가정하면서, 영형의 선행사를 식별하는 데에 그 초점을 맞춘다. 하지만, 표 1에서 보듯이, 생략된 요소의 지시 대상이 항상 문서상에 존재하는 것은 아니므로 이들을 구분할 필요가 있다. 표 1은 영형이 나타난 문장의 예를 보여준다. 표 1에서 네 개의 영형($ø_1$, $ø_2$, $ø_3$, $ø_4$)은 모두 지시적 표현으로 간주될 수 있지만, 그들의 선행사 식별에는 차이가 있다.

표 1 조응적 영형과 비조응적 영형의 예

구 분	예 문
조응적 (anaphoric)	철수가 내기에 지니까 ($ø_1$-NOM) 심술을 부린다. ($ø_2$-NOM) 집에 들어서면 아이가 엄마부터 찾는다.
비조응적 (non-anaphoric)	($ø_3$-NOM) 2시경이면 태양이 산허리 아래로 내려간다. ($ø_4$-NOM) 유심히 살펴보았더니 스쿠아가 코모란트를 공격하고 있었다.

다시 말해, $ø_1$과 $ø_2$는 같은 문장에 있는 "아이"와 "철수"를 각각 지시하고 있지만, $ø_3$와 $ø_4$의 지시 대상은 문서 내에 존재하지 않는다. 따라서 $ø_3$, $ø_4$와 같은 비조응적 영형은 영 조응어 해석의 선행사 식별 과정에서 구분되는 것이 바람직하다. 하지만 대부분의 연구들은 이들을 구분하지 않은 채 문서 상에서 그 선행사를 발견하지 못한 경우를 비조응적이라 판단한다. 그러나, 이러한 결과에는 여러 오류들도 함께 포함될 수 있으므로, 그들을 모두 비조응적으로 간주하는 것에는 문제가 있다. 이에 영형의 지시 해석에 관한 최근 연구들은 조응어 해석의 전체 성능을 향상시키기 위하여 영형의 조응성을 구분하려는 시도를 하고 있다[3].

조응성 결정에 관한 기존 연구들은 대부분 지시적 표현과 그 선행사 후보들 간의 쌍대 비교(pairwise comparison)를 통해 조응성을 측정하려고 하며, 많은 경우 비조응적 훈련 예들로부터 비조응적 클래스를 직접 학습하지 않는다. 이러한 접근 방법은 대부분이 조응적 영형의 선행사 식별 모델에 그 기반을 두고 있음을 반영하는 것이며, 그들의 선행사가 문서상에 존재하지 않는 비조응적 영형을 직접 구분하는 데에는 사실상 한계를 가진다.

영 조응 현상은 최근 연구들에서 문장 내(intra-sentential)와 문장 간(inter-sentential) 조응으로 구분되어 다루어지고 있다[3,6]. 이러한 관점에서 보면 대다수의 영형은 문장 내에서 해결이 가능하며, 그 비율은 약 78% 정도이다(표 5 참조). 따라서 본 논문에서는 문서에서 가장 높은 출현 빈도를 보이는 영형, 즉 주어 자리에 나타난 영형의 문장 내 조응성 결정 문제에 초점을 맞춘다. 쌍대 비교에 기반한 기존 연구와 달리, 제안한 방법은 영형 주어의 조응성 결정 문제를 비조응적 혹은 문장 간 영형 주어가 있는 절(clause)들의 식별 문제로 다룬다. 즉, 조응적, 비조응적 영형이 나타난 절들의 구조가 서로 유사할 것이므로 이들이 나타난 절의 구조를 학습함으로써 문장 내 비조응적 예들을 직접 구분하려고 한다. 이를 위해 구조 정보를 잘 다루는 파스 트리(parse tree) 커널[7,8]과 분류 문제를 해결하는데 있어 가장 강력하고 널리 쓰이는 모델 중 하나인 SVM(support vector machine)을 이용한다. 따라서 본 논문의 목적은 파스 트리로부터 추출한 구조 정보가 영형의 조응성 결정에 도움을 주는가, 그리고 조응성 결정 단계가 향후 영형 조응어 해석의 성능에 영향을 미치는가를 살펴보는 것이다.

2장에서 영대명사에 관한 기존 연구들을 살펴보고 3장에서 문장 내 영 조응어 해석을 위한 영대명사의 조응성 결정에 관해 설명한다. 4장에서 실험 및 결과를 분석하고 마지막 5장에서 결론 및 향후 연구를 다룬다.

2. 관련 연구

영대명사를 포함한 지시 해결에 관한 기존 연구들은 규칙에 기반한 방법과 기계 학습에 기반한 방법으로 크게 나누어진다. 먼저, 규칙에 기반한 접근 방법은 주로 휴리스틱 룰과 센터링 이론(Centering Theory)[9]에 기반을 두고 있다. 지역적 결속성(local coherence)에 기반을 두고 있는 센터링 이론은 대명사의 지시 해석을 위한 모델로써 영어권에서 주로 사용되어져 왔다. 하지만 센터링의 중요한 표기들인 발화(utterance), 선행발화, 순위, 실현(realization) 등에 대한 이해가 언어마다 조금씩 다를 수 있으며, 특히 일본어, 한국어와 같이, 조응적, 비조응적 영형의 출현이 빈번한 언어에서는 센터링 이론만으로 모든 영형들을 해석하기가 쉽지 않다. Roh and Lee(2006)는 한국어 영대명사를 생성하기 위한 비용 기반의 센터링 모델을 제안하였지만[10], 영대명사 해결과 생성은 그 문제의 복잡도가 다르기 때문에 제안한 모델을 그대로 영형 조응어 해석에 적용하는 것에는 문제가 있다.

다음으로 기계 학습에 기반한 방법들이 있다. 이들은 조응성 결정 단계의 분리 여부에 따라 다시 구분될 수

있다. 먼저, 선행사 식별에 초점을 맞추는 이전 연구들은 지시적 표현과 그 선행사 사이에 나타난 명사구들, 혹은 공지시 체인(coreference chain) 안에 포함되지 않은 명사구들을 부정적 실례로써 사용하고 있다[11,12]. 이들은 문서에서 그 선행사를 결정하지 못한 경우를 비조응적이라고 판단하지만, 그 결과에는 다른 오류들도 함께 포함될 수 있으므로 이러한 경우를 모두 비조응적으로 인정하는 것에는 무리가 있다.

최근에는 조응성 결정을 선행사 식별과 분리된 단계에서 수행하려는 시도들이 나타나고 있다[3,4]. 문장 내 영 조응어 해석에 관한 Iida et al.(2007)의 연구는 토너먼트 모델에 의해 영형의 가장 두드러진 선행사 후보를 문서에서 식별한 뒤, 영형과 그 선택된 후보 사이의 조응성 평가를 통해 문장 내 비조응적 영형을 구분하려고 한다[3]. 하지만, 이들의 조응성 결정 모델은 영형의 선행사 식별 모델에 의존하고 있는 매개변수적 접근 방식을 취하고 있다. 한국어의 영대명사 문제를 다룬 Han (2006)의 연구는 조응적, 비조응적 영형을 모두 대상으로 하며, 특히 비조응적 영대명사를 유형에 따라 분류하고 그들을 각각 식별하였다[2]. 하지만 조응적, 비조응적 영형을 단일한 접근 방법으로 함께 해결하려 함으로써, 영대명사의 조응성 결정을 위한 새로운 관점의 자질을 제시하지 못하였다.

조응성 결정에 대한 관심은 최근 영어권으로도 다시 확대되고 있으며, Bergsma(2008)는 영어의 3인칭 대명사 'it'의 지시성을 구분하기 위한 연구를 진행하였다[13]. 또한 Jo et al.(2007)은 한국어를 대상으로 대화에서 자주 등장하는 '~것'의 지시적, 비지시적 쓰임을 구별하기 위한 방법을 제안하였다[14]. 이처럼 다양한 지시적 표현들의 조응성 결정 문제는 지시 해결을 위한 중요한 이슈로써 부각되고 있다.

3. 한국어 영대명사의 조응성 결정

3.1 영 조응어 해석과 조응성 결정

본 논문은 한국어 복합문에서 가장 많은 출현 빈도를 보이는 영형 주어의 문장 내 조응성 결정 문제를 다룬다. Han(2006)은 한국어 영대명사의 대다수가 주어 위치에서 관찰되었다고 보고하였다[2].

영 조응어 해석은 선행사 후보의 그 탐색 범위에 따라 문장 내(intra-sentential)와 문장 간(inter-sentential)으로 구분된다[3,6]. 이러한 관점에서 본다면 영형 주어의 선행사는 주어가 생략된 그 문장 안에 있거나, 문장 간, 혹은 문서 상에서 발견할 수 없는(extra-sentential) 경우도 있다. 따라서, 문장 내 영 조응어 해석에서 영형 주어의 선행사가 같은 문장 내에 존재하는지의 여부를 판단하는 것은 중요하다.

Ex-1. 실내에서 수분이 응결하여 ($Ø_1$-이) 물방울이 되어 ($Ø_2$-이) 떨어진다거나 곰팡이가 핀 것을 ($Ø_3$-이) 염려하여 ($Ø_4$-이) 습도는 40 퍼센트 정도로 맞추어 놓았다.

Training Instances	*Label*
(실내, $Ø_1$)	0
(수분, $Ø_1$)	1
(실내, $Ø_2$)	0
(수분, $Ø_2$)	0
(물방울, $Ø_2$)	1
(실내, $Ø_3$)	0
(수분, $Ø_3$)	0
(물방울, $Ø_3$)	0
(곰팡이, $Ø_3$)	0
(것, $Ø_3$)	0
(실내, $Ø_4$)	0
(수분, $Ø_4$)	0
(물방울, $Ø_4$)	0
(곰팡이, $Ø_4$)	0
(것, $Ø_4$)	0

그림 1 기존 연구에서의 문장 내 선행사 식별 모델

그림 1은 기계 학습에 기반한 최근 연구들에서 보여지는 선행사 식별의 접근 방법을 보여준다. 그림 1에서 $ø_1$과 $ø_2$는 문장 내에서 해결할 수 있는 조응적 영형으로, $ø_3$과 $ø_4$는 비조응적 기능의 직시적 영형으로 파악할 수 있다.

문장 내 선행 조응어 해석에 관한 기존 연구들은 영형과 선행사 후보 간의 조응적 관계성을 평가함으로써 영형의 선행사를 식별한다[2,3]. 이들은 (수분, $ø_1$), (물방울, $ø_2$)의 관계를 긍정적(1)으로, 영형을 선행하는 다른 모든 명사구들과의 관계는 그림 1과 같이 부정적인 관계(0)로 파악하며, 지시적 표현과 선행사 후보 간의 관계성을 평가하기 위하여 거리, 일치 등의 여러 자질들을 제안하고 있다.

하지만, 기존 연구들의 선행사 식별 모델에서 나타나는 중요한 문제 중 하나는 영형의 조응성을 결정하지 않음으로써 생기는 부정적인 관계들의 증가가 조응적 영형의 선행사 식별에 영향을 미친다는 것이다. 따라서 실제적인 활용을 위해서는 영대명사의 조응성에 대한 직접적인 구분이 필요하며, 영형의 선행사 식별과 조응성 결정은 상호 보완적 관계로 파악되어야 한다. 특히, 한국어에서 영형의 조응성은 문장 간으로 그 선행사 탐색 범위가 넓혀지기 전에 결정되어질 필요가 있으며, 그렇지 않으면 최악의 경우에 영대명사를 선행하는 모든 명사구를 선행사 후보로써 평가해야 하는 결과를 가져올 수 있다. 또한 생략된 요소와 그 선행사 사이의 성, 수, 인칭 등에 대한 일치 정보가 표층에 나타나지 않는 한국어와 같은 언어에서는 선행사 후보와 영형 사이의 후보 제약이나 그들 사이에서 나타나는 선호도(preference) 분석 등의 접근 방법을 통해 영형의 조응성을 밝혀내기에 부족하다.

이에 본 논문에서는 구조 정보를 활용한 영대명사의 조응성 결정을 제안하며, 영형이 나타난 절의 구조 정보

를 영형의 문장 내 조응성 결정에 이용하고자 한다. 복합문을 구성하고 있는 절은 담화 분석의 기본 단위로 파악할 수 있으며, 절 간의 구조 정보는 영형의 조응적 기능을 분석하는 데에 도움을 줄 수 있다. 파스 트리로부터 추출한 구조 자질의 중요성은 영형 해석에 관한 최근 연구에서 보고되고 있다[3,15]. Iida et al.(2007)은 문장 내(intra-sentential) 영대명사의 지시 해결에서 구조 정보가 유용한 자질임을 소개하였다[3]. 하지만, Iida et al.(2007)의 조응성 결정은 문장 내 선행사 식별 모델이 선택한 후보가 주어진 영형과 조응적인지를 평가한 것이므로, 그들이 제안한 토너먼트 기반의 선행사 식별 모델에 의존하는 한계를 가진다. 본 논문은 선행사 식별 전에 영형의 조응성을 구분하는 문장 내 영 조응어 해석에 대한 접근 방법을 제시한다.

(1) 문장 내 조응성 결정
(Intra-sentential anaphoricity determination)

(2) 문장 내 선행사 식별
(Intra-sentential antecedent identification)

이를 위해 문장 내 비조응적 영형들을 직접 구분하기 위한 영대명사의 조응성 결정 방법을 제안한다.

3.2 영대명사의 문장 내 조응성 결정

조응성 결정은 선행사 식별과 더불어 영 조응어 해석의 중요한 문제 중 하나이다. 본 논문은 문장 간에서 해결되어야 하는 영형이나 문서 상에서 그 선행사를 발견할 수 없는 비조응적 영형들을 문장 내 조응성 결정을 통해 구분하고자 한다. 이 때 직시적, 상황적 기능을 수행하는 영대명사와 부정 인칭 영대명사를 비조응적 기능의 영형으로 분류한다[2].

(1) 직시적(deictic) 영대명사: 화자(speaker), 청자(hearer) 등을 비롯한 담화의 참여자들을 지시하는 영대명사가 주로 여기에 속한다.

(2) 일반 상황적(general situational) 영대명사: 시간, 날씨, 일반적 상황 등을 지시하는 영대명사를 말하며, 허사적(expletive) 기능을 수행하는 영어의 대명사 'it'이 가장 대표적인 예이다. 또한 Han(2006)의 작업과 같이, "~에 따라서, ~에 관해서, ~에 대해서, …" 등과 같은 숙어적 표현들을 같은 유형으로 함께 분류한다.

표 2 비조응적 영대명사(ZP)의 분류 예

구 분	예 문
Deictic	ø 점심 먹어라. 한국팀이 이겨서 ø 기쁘다.
General Situational	ø 벌써 열시다. 철수는 온난화에 대해서 설명했다.
Indefinite Personal	ø 호랑이를 잡으려면 ø 산에 가야 한다.

(3) 부정 인칭(indefinite personal) 영대명사: 대부분 불특정한 일반인(generic person)을 지시하는 영대명사들이 여기에 속한다.

이외에도 명사적(nominal) 표현을 선행사로 가지지 않는 영형을 비조응적으로 함께 구분한다.

따라서, 본 논문에서 영형 주어의 조응성 결정 문제는 절이 문장간 영 조응 혹은 비조응적 기능의 영형 주어를 가지는지 아닌지의 이진 분류 문제로 다루어진다.

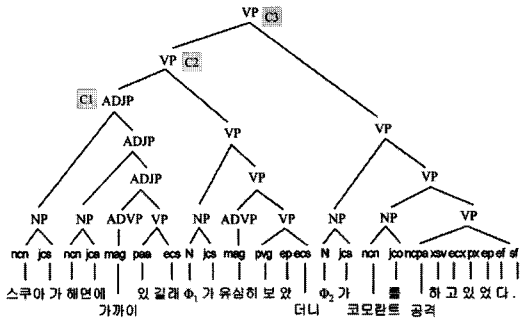

그림 2 영형이 나타난 문장의 파스 트리 예

그림 2는 3개의 절(C1, C2, C3)로 이루어진 문장의 파스 트리의 예를 보여준다. 그림 2에서 C2절은 담화 참여자를 지시하는 비조응적 영형 주어(Φ_1)를 가지고 있으며, C3절은 다른 절(C1)의 주어 '스쿠아'를 지시하는 조응적 영형 주어(Φ_2)를 가지고 있다. 이들을 구분하기 위하여 본 논문은 비조응적 영형이 나타난 C2 절에 대한 구조 정보는 긍정적인(1) 예로, C3 절의 구조 정보는 부정적인(0) 학습 예로 이용한다. 그림 3은 조응성 결정을 위해 파스 트리로부터 추출하는 구조 정보의 예를 보여주며[1], 파스 트리의 구조를 그대로 사용함으로써 정보의 손실을 최소화한다.

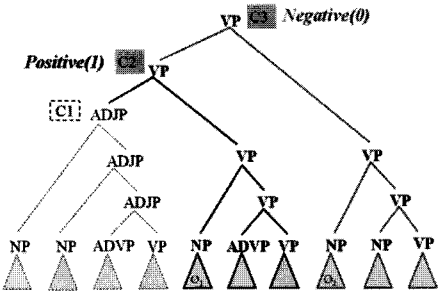

그림 3 파스 트리로부터 추출하는 구조 정보

1) 그림 3에서 구(phrase)들에 대한 세부 정보가 생략되어 있지만, 실제 사용하는 구조 정보의 단말 노드들은 문장을 이루는 어휘들이다(그림 2를 참조).

문서에서 나타난 지시적 표현은 문서의 응집성(coherence)을 높이는 요소로 파악할 수 있다[5]. 즉, 절 사이에서 형성된 관계적(relational), 참조적(referential) 응집성은 영형의 조응성을 평가하는데 도움을 줄 수 있다. 표 3은 조응성 결정을 위해 절의 구조 정보와 함께 사용하는 자질들을 보여주며, 술어와 어미, 그들의 의미부류, 영형(ZP)의 출현 여부 등이 조응성 결정에 기여할 수 있는지를 살펴본다. 표 3에서 C_i는 조응성 결정의 대상이 되는 절을 의미하며, C_{i-1}과 C_{i+1}은 C_i의 앞 뒤에 나타난 절의 정보를 반영한다. 이 때 C_i 절이 문장에서 나타난 첫번째 혹은 마지막 절이라면, C_{i-1} 혹은 C_{i+1}에 대한 정보가 NULL(0)의 값을 가지도록 하였다.

그리고 영대명사를 다루는 대다수의 연구가 영형이 출현한 위치를 미리 파악하고 있는 것과 같이[2,3,6], 본 논문에서도 영형 주어의 위치가 이미 알려져 있다고 가정한다.

표 3 조응성 결정을 위해 함께 사용하는 자질들

절	No.	의미	값
C_i	1	ZP의 유무	0/1
	2	용언	Word
	3	어미	Word
	4	용언의 POS정보	POS
	5	어미의 POS정보	POS
	6	보조용언	Word
	7	인용절 내포 유무	0/1
	8	용언의 의미부류	Semantic class[16][2)]
	9	어미의 의미기능	Grammatical meaning[16][2)]
C_{i-1}	10	ZP의 유무	0/1
	11	용언	Word/NULL
	12	어미	Word/NULL
	13	용언의 POS정보	POS/NULL
	14	어미의 POS정보	POS/NULL
	15	보조용언	Word/NULL
	16	인용절 내포 유무	0/1
	17	용언의 의미부류	Semantic class[2)]/NULL
	18	어미의 의미기능	Grammatical meaning[2)]/NULL
C_{i+1}	19	ZP의 유무	0/1
	20	용언	Word/NULL
	21	어미	Word/NULL
	22	용언의 POS정보	POS/NULL
	23	어미의 POS정보	POS/NULL
	24	보조용언	Word/NULL
	25	인용절 내포 유무	0/1
	26	용언의 의미부류	Semantic class[2)]/NULL
	27	어미의 의미기능	Grammatical meaning[2)]/NULL

2) 용언, 어미의 의미 모호성에 해소하는 것은 자연어처리 분야의 중요한 연구과제 중 하나이므로, 본 논문에서는 21세기 세종계획 성과물인 전자사전에 기입된 첫번째 의미부류 및 의미기능을 해당 용언, 어미의 속성값으로 각각 사용하였다.

3.3 파스 트리 커널(Parse Tree Kernel)

Collins and Duffy(2001)에 의해 제안된 파스 트리 커널은 본 논문에서 절의 구조 정보를 모델링하기 위해 사용한다[7]. 파스 트리 커널에서 벡터의 자질들은 각 파스 트리에 나타날 수 있는 모든 부분 트리(subtree)들로 이루어지며, 각 자질의 값은 부분 트리의 빈도수를 나타낸다. 그러나, 이러한 부분 트리를 명시적으로 구하는 것은 불가능하므로, Collins and Duffy(2001)는 아래 재귀 규칙을 두 파스 트리의 모든 노드에 대해 적용함으로써 명시적인 열거없이 내적을 구하는 방법을 제시하였다[7].

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 파스 트리 T1과 T2의 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 논문은 STEP 2000 과제의 결과물인 구문 부착 말뭉치에서 추출한 데이터를 이용하였으며, 영형 주어의 조응성에 대한 분류 작업은 모두 수작업으로 진행하였다. 실험 말뭉치로부터 하나 이상의 영형 주어를 가지고 있는 5,221개의 문장을 추출하였으며, 추출된 문장은 평균 3.97개의 절과 각 절은 평균 7.67개의 단어를 포함하고 있다.

표 4 실험 데이터에 대한 간단한 통계(statistics)

정 보	값
문장 수	5,221
총 절의 수	20,748
영형 주어가 나타난 절의 수	13,172

표 5 데이터셋에서 관찰된 영형 주어의 분포

문장 내 (Intra-sentential)	문장 간 (Inter-sentential)	문장 외 (Extra-sentential)
10,375 (78.77%)	666 (5.06%)	2,131 (16.18%)

표 4는 STEP 2000 말뭉치에서 추출한 실험 데이터에 대한 통계 정보이며, 표 5는 데이터셋에서 관찰된 영형 주어의 분포이다. 모든 실험은 5회 교차 검증(cross validation)을 통해 평가하였으며, 분류기로 SVM_{light} [17]를 사용하였다. 영형 주어의 문장 내 조응성 결정에 대한 성능 평가는 정확도(accuracy)와 F1 척도(F1 measure)를 이용하여 분석하였다.

$$Accuracy(A) = \frac{\text{조응성이 올바르게 결정된 절의 수}}{\text{영형 주어가 나타난 절의 수}}$$

$$Precision(P) = \frac{\text{올바르게 식별된 문장간 조응, 혹은 비조응적 영형의 수}}{\text{문장간 조응, 혹은 비조응적이라고 식별된 영형 주어의 수}}$$

$$Recall(R) = \frac{\text{올바르게 식별된 문장간 조응, 혹은 비조응적 영형의 수}}{\text{문장간 조응, 혹은 비조응적 영형 주어의 수}}$$

4.2 영형의 조응성 결정에 대한 실험 결과

표 6은 영형 주어의 조응성 결정에 대한 실험 결과이다. 본 논문은 두 개의 베이스라인 시스템을 사용한다. 표 6에서 'BOW'(bag-of-words)는 절을 구성하는 어휘들을 벡터로 표현한 모델의 실험 결과이며 'SEM'은 표 3에서 소개한 자질만을 이용한 모델의 실험 결과이다. 'STRUC'는 파스트리 커널을 이용하여 절의 구조 정보를 학습한 제안한 모델의 실험 결과이며, 'STRUC+'는 구조 정보와 표 3에서 소개한 자질들을 함께 사용한 복합 커널의 결과를 보여주고 있다. 실험에서 사용한 복합 커널 $K = \alpha \cdot K_1 + (1-\alpha) \cdot K_2$ 이다. 여기서 K_1은 파스트리 커널을, K_2는 degree가 3인 다항식(polynomial) 커널을 의미하며 α는 혼합(mixing) 파라미터이다($0 \le \alpha \le 1$). 실험 결과, 구조 정보를 활용한 제안한 모델이 베이스라인 시스템들보다 나은 성능을 보여줌을 알 수 있다.

표 6 조응성 결정에 대한 실험결과

Methods	A	P	R	F1
BOW_{clause}	0.8382	0.8219	0.2979	0.4369
SEM	0.8370	0.8109	0.3026	0.4402
STRUC	0.8450	0.7884	0.3690	0.5023
STRUC + (α=0.7)	0.8493	0.8143	0.3753	0.5132

그림 4는 혼합 파라미터 α의 변화에 따른 문장 내 조응성 결정 모델의 성능 변화를 보여준다. α 값이 커질수록 구조 정보의 기여도가 증가하는 것을 나타내며, 실험 결과 α의 값이 0.7일 때 'STRUC+'가 가장 좋은 성능을 보여주었다. 이는 파스 트리로부터 추출한 절의 구조 정보가 조응성 결정에 유용한 자질임을 반영하는 것이다.

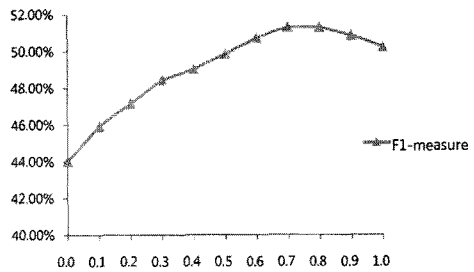

그림 4 혼합 파라미터 α 값의 변화에 따른 조응성 결정에 대한 성능 변화

그림 5는 'STRUC+(α=0.7)'의 실험 결과를 통해 살펴본 조응성 결정의 세부 유형별 인식 결과를 보여준다. 그림 5에서 "total"은 실험 데이터셋에서 해당 유형의 영형이 차지하고 있는 분포를 보여주며, "correct"는 본 논문에서 제안한 모델이 이들을 제대로 식별했는가에 대한 결과이다. 즉, 실험 데이터셋에서 나타난 문장 내 비조응적 영형의 분포는 일반 상황적(37.2%), 직시적(24.7%), 문장 간(24.6%), 부정 인칭(9.4%), 비명사적(4.1%) 영형 주어의 순으로 나타났으며, 그림 5에서 보여주듯이 절의 구조 정보가 특히 일반 상황적 기능의 비조응적 영형을 분석하는 데에 많은 도움이 됨을 알 수 있다.

하지만, 제안한 방법이 영형의 조응성 결정에 대해 아직 만족할 만한 성능을 보여주지 못하고 있다. 조응성 결정에 대한 실험 결과, 정확률은 높지만 재현율이 낮은 것을 확인할 수 있다. 이는 불균형 데이터 문제가 주요한 원인 중 하나로 파악되며, 실험에 사용된 데이터셋의 경우 문장 내 영형 주어의 분포가 문장 간, 또는 문장 외에서 해결되어야 하는 영형에 비해 약 4배 정도 많음을 알 수 있다. 이러한 문제를 해결하기 위해서 배깅(bagging), 부스팅(boosting) 등의 앙상블(Ensemble) 기법이나 샘플링(sampling) 방법 등에 관한 연구를 계속 진행할 예정이다.

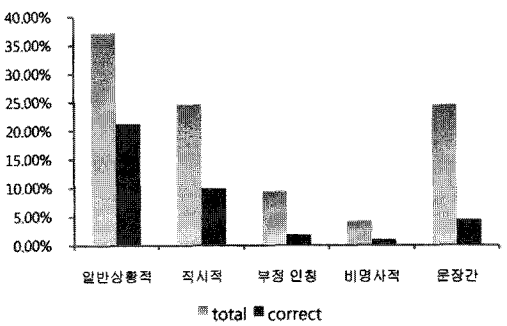

그림 5 문장 내 비조응적 영형 주어의 세부 유형별 인식 결과

4.3 영형 조응어 해석으로의 확장을 위한 주어 공유 식별과 조응성 결정

조응성 결정이 실제 영형의 지시 해결에 어느 정도 영향을 미치는가를 살펴보기 위해 본 논문에서는 절 간 주어 공유 식별 문제에 제안한 조응성 결정 모델을 적용하고자 한다. 한국어 복합문에서 생략된 절의 주어는 대개의 경우 그 문장 안에서 원형을 복원할 수 있다. 이는 생략된 주어가 다른 절의 주어를 공유하는 문제로 살펴볼 수 있으며, 이를 절 간 주어 공유 문제라 부른다 [15]. 따라서 Kim et al.(2008)의 방법론에 제안한 조응성 결정 모델이 어떠한 영향을 미치는 지를 실험을 통해 살펴봄으로써 조응성 결정이 향후 영형 조응어 해석에 도움을 줄 수 있는지를 살펴보았다.

본 논문에서 제안한 조응성 결정 모델을 통해 올바르게 식별된 문장 간 조응과 비조응적 영형 주어들을 미리 제거한 경우와 영형의 조응성을 반영하지 않은 경우를 주어 공유 식별을 통해 각각 실험하였다. 실험 결과, 표 7과 같이, 영형 주어의 조응성 결정 모델이 문장 내 주어 공유 식별 문제에 긍정적인 영향을 줄 수 있음을 실험을 통해 확인할 수 있었다. 앞으로 조응성 결정 모델의 전체적인 성능이 보다 만족할 만한 수준에 도달한다면, 문장 내 영형 주어, 더 나아가 영대명사의 지시 해결에 조응성 결정 모델이 기여할 수 있을 것으로 기대된다.

표 7 조응성 결정에 따른 절 간 주어 공유 식별의 성능 변화

	A	P	R	F1
주어공유 식별(SSI)	76.34	69.55	61.58	65.30
문장 내 조응성 결정+SSI	80.77	75.19	69.85	72.42

5. 결론 및 향후 연구

영형 또는 대명사, 한정적(definite) 명사구 등과 같이 다양한 지시적 표현들은 문서상에서 자주 관찰되며, 이들이 가리키는 대상을 식별하는 작업은 자연어처리 분야의 중요한 연구 과제 중 하나이다.

본 논문에서는 한국어 복합문에서 빈번하게 나타나는 영형 주어들의 조응성 결정 문제를 다루었다. 제안한 모델은 문장 내 영 조응어 해석을 위해 영형 주어가 나타난 절의 구조 정보를 학습하여 영형의 조응성을 구분하였으며, 비조응적 훈련 예들을 이용하여 이들을 직접 구분하려고 시도하였다. 제안한 방법은 영형의 선행사 식별 전에 조응성을 결정함으로써 선행사 식별 모델의 결과에 의존하지 않는다. 또한 주어 공유성 식별의 실험을 통해 조응성 결정 모델이 영대명사 해결의 성능에 긍정적인 효과를 줄 수 있음을 확인할 수 있었다.

하지만 영형의 조응성 결정은 어려운 작업이며, 앞으로 보다 정제된 의미 자질의 사용이나 불균형 데이터(imbalanced data) 문제의 해결 등에 대한 연구가 계속되어야 한다. 또한 향후 연구로는 제안한 방법을 주어 뿐만 아니라 다양한 위치에서 나타나는 영형의 조응어 해석 문제로 적용하여 연구의 범위를 확대해 나갈 예정이다. 제안한 방법은 영형 뿐만 아니라 대명사, 명사구들의 조응성 결정 모델로도 확장이 가능할 것으로 기대된다.

참 고 문 헌

[1] D.-S. Wu and T. Liang, "Zero Anaphora Resolution by Case-based Reasoning and Pattern Conceptualization," *Expert Systems with Applications*, vol.36, no.4, pp.7544-7551, May 2009.

[2] N.-R. Han, Korean Zero Pronouns: Analysis and Resolution, Doctoral dissertation, Department of Linguistics at the University of Pennsylvania, 2006.

[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] 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.

[5] Halliday, M.A.K and Hasan R., *Cohesion in English*, London:Longman, 1976.

[6] R. Iida, K. Inui, and Y. Matsumoto, "Capturing salience with a trainable cache model for zero-anaphora resolution," *In Proceedings of the Joint Conference of the ACL-IJCNLP*, pp.647-655, 2009.

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

[8] A. Moschitti, "Making Tree Kernels Practical for Natural Language Learning," *In Proceedings of EACL*, pp.113-120, 2006.

[9] 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.

[10] J.-E. Roh and J.-H. Lee, "Generation of Zero Pronouns Based on the Centering Theory and Pairwise Salience of Entities," *IEICE Transactions on Information and Systems*, vol.E89-D, no.2, pp.837-846, February 2006.

[11] V. Ng and C. Cardie, "Improving Machine Learning Approaches to Coreference Resolution," *In Proceedings of ACL*, pp.104-111, 2002.

[12] W. M. Soon, H. T. Ng, and D. C. Y. Lim, "A Machine Learning Approach to Coreference Reso-

lution of Noun Phrases," *Computational Linguistics*, vol.27, no.4, pp.521-544, 2001.
[13] S. Bergsma, D. Lin and R. Goebel, "Distributional Identification of Non-Referential Pronouns," *In Proceedings of ACL-HLT*, Columbus, Ohio, pp.10-18, 16th-18th, 2008.
[14] E. Jo, H. Kim, and J. Seo, "Distinguishing Referential Expression 'Geot' Using Decision Tree," *Journal of KIISE : Software and applications*, vol.34, no.9, pp.880-888, Sep. 2007. (in Korean)
[15] K.-S. Kim, S.-B. Park, H.-J. Song, S.-Y. Park, and S.-J. Lee, "Identification of Subject Shareness for Korean-English Machine Translation," *In Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence*, pp.211-222, 2008.
[16] Korean electronic dictionaries in the 21st Century Sejong Project (http://www.sejong.or.kr)
[17] 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
1996년 부산여자대학교 전자계산학과 졸업(학사). 1998년 경북대학교 대학원 컴퓨터공학과 졸업(석사). 2003년~2007년 경일대학교 교양학부 강의전담교수. 2000년~현재 경북대학교 대학원 컴퓨터공학과 박사과정. 관심분야는 자연어처리, 정보추출, 기계번역

박 성 배
1994년 한국과학기술원 컴퓨터과학과 졸업(학사). 1996년 서울대학교 대학원 컴퓨터공학과 졸업(석사). 2002년 서울대학교 대학원 컴퓨터공학과 졸업(박사). 2004년~현재 경북대학교 IT대학 컴퓨터학부 교수. 관심분야는 기계학습, 자연어처리, 텍스트마이닝, 정보추출, 생명정보학

박 세 영
1980년 경북대학교 전자공학과 졸업(학사). 1982년 한국과학기술원 전산학과 졸업(석사). 1989년 프랑스 파리 7대학 전산학(박사). 1982년~2000년 한국전자통신연구원 자연어처리 연구부장, 지식정보연구부장. 2000년~2003년 서치캐스트㈜ 대표이사, 2003년~2005년 한국정보통신연구진흥원 전문위원. 2005년~현재 경북대학교 IT대학 컴퓨터학부 교수. 관심분야는 한국어정보처리, 시맨틱 웹, 인공지능, 정보검색

이 상 조
1974년 경북대학교 수학교육과 졸업(학사). 1976년 한국과학기술원 전산학과 졸업(석사). 1994년 서울대학교 컴퓨터공학과 졸업(박사). 1976년~현재 경북대학교 IT대학 컴퓨터학부 교수. 관심분야는 자연어처리, 기계번역, 정보검색, 정보추출