결정 트리를 이용한 지시 표현 '것'의 구별
(Distinguishing Referential Expression 'Geot' Using Decision Tree)

조 은 경 † 김 학 수 †† 서 정 연 †††
(Eunkyoung Jo) (Harksoo Kim) (Jungyun Seo)

요 약 지시 표현 '것'은 한국어 대화에서 자주 등장하는 표현이지만, 그 자체로서 대명사나 한정 명사구와 같은 지시 표현이 아니고, 비지시적인 표현으로 쓰이는 '것'과 구별되지 못했기 때문에 지시 해석(reference resolution)에 관한 기존 연구에서 제대로 다루어지지 못했다. 이러한 문제를 해결하기 위해 '것'이 가지고 있는 언어학적 속성과 담화 상의 속성을 기반으로 하여 자질 집합을 설정하고, 결정트리를 이용하여 '것'을 구별하는 방법을 제안한다. 이 방법에 의한 시스템은 비지시 표현의 것에 대해 92%, 지시 표현의 것에 대해 82%의 F-measure를 보였으며, 전체적인 분류 성능은 89%였다. 이는 패턴에 따른 규칙을 적용한 분류 성능에 비해 약 15% 가량 향상된 결과이다.

키워드 : 지시 표현, 조응어 구별, 조응성, 지시 표현 해석

Abstract Referential expression 'Geot' is often occurred in Korean dialogues. However, it has not been properly dealt with by the previous researchers of reference resolution, since it is not by itself the referential expression like pronoun and definite noun phrases, and it has never been discriminated from non-referring 'geot'. To resolve this problem, we establish a feature set which is based on the linguistic property of 'geot' and the discourse property of its text, and propose a method to identify referential 'geot' from non-referring 'geot' using decision tree. In the experiment, our system achieved the F-measures of 92.3% for non-referring geot and of 82.2% for referential geot and the total classification performance of 89.27%, and outperformed the classification system based on pattern rules.

Key words : referential expression, anaphora identification, anaphoricity, reference resolution

1. 서 론

지시 해석(reference resolution)은 의미 해석과 담화 이해를 위해 필수적인 자연어처리 과정의 하나이며, 지시 표현 구별은 지시 해석의 대상을 결정짓는 기본 과정이다. 본 논문에서는 대화에서 종종 쓰이는 지시 표현 '것'을 비지시 표현으로 쓰이는 '것'과 구별하여 인식할 수 있는 방법을 제시함으로써 지시 해석의 대상을 넓히고, 담화 이해의 수준을 높이고자 한다. 즉, 표 1과 같이 '비행기'를 지시하는 '출발하는 것'의 '것'과 무엇인가를 지시하는 표현이 아닌 '투숙하실 것'의 '것'을 구별하여 지금까지 지시 표현으로 처리되지 못한 '것'을 지시 해석의 대상으로 삼을 수 있게 한다.

표 1 지시 표현과 비지시 표현의 예

분류	문장
지시 표현	죄송한데요, 히로시마행 비행기를 취소하구, 교토행 같은 날 출발하는 **것**으로 예약하고 싶은데요.
비지시 표현	네, 기꺼이 도와드리겠습니다. 몇 분이 투숙하실 **것**입니까?

2. 관련 연구

지시 표현의 구별 문제는 대용량의 말뭉치를 기반으로 한 언어 처리가 이루어지고 있는 시점에서 부각되었다[10].

· 이 연구(논문)는 산업자원부 지원으로 수행하는 21세기 프론티어 연구개발사업(인간기능 생활지원 지능로봇 기술개발사업)의 일환으로 수행되었습니다. 또한 김학수의 이 연구는 부분적으로 강원대학교 정보통신연구소의 지원을 받았습니다.
† 정 회 원 : ㈜다음커뮤니케이션
jek@lex.yonsei.ac.kr
†† 정 회 원 : 강원대학교 IT특성화학부(대학) 컴퓨터정보통신공학전공 전임강사
nlpdrkim@kangwon.ac.kr
††† 종신회원 : 서강대학교 컴퓨터학과 교수
seojy@mail.sogang.ac.kr
논문접수 : 2005년 7월 11일
심사완료 : 2007년 6월 7일

영어 지시 표현의 구별에 관한 연구로는 대명사 'it'과 'the+명사'의 한정명사구, 모든 명사구를 대상으로 한 것들이 있다. Lappin and Leass(1994)는 'it'이 지시 표현인지 아닌지를 구별하기 위해 문장 패턴을 이용하였다[8]. 예를 들어 'It+(any form of the verb 'be')+(a cognitive verb(past tense))+that'과 같은 패턴에서 'it'은 허사적(pleonastic) 용법임을 구분하는 방법이었다. 그러나 이러한 방법은 다른 성분이 끼어든 구조를 만나면 실패하게 되는 등 그 적용 범위가 매우 제한적이었다. 이러한 문제를 해결하기 위하여 Evans(2001)은 3,171개의 'it'을 포함하고 있는 37만 어절의 SUDDANE, BNC 말뭉치에서 77개의 텍스트를 대상으로 기계 학습 방법의 하나인 kNN(K-nearest neighbor) 방법을 이용하여 'it'의 쓰임을 분류하였다[11]. 한정 명사구를 만드는 'the'는 구정보(old information)임을 나타내기 때문에 담화 상에 언급된 개체를 표현할 수도 있고, 관용적인 표현에 쓰일 수도 있고, 누구나가 인지하고 있는 유일물이나 사회적인 사건 등을 가리키는 데 쓰일 수도 있다. Vieira and Poesio(2000)은 월 스트리트 저널(Wall street journal)의 기사 말뭉치에서 한정 명사구의 이러한 쓰임에 대해 조응어 분류(anaphora classification) 실험을 하여 재현율 78%, 정확률 89%의 결과를 보였다[12]. Ng and Cardie(2002)는 텍스트 내에 등장한 대명사, 한정 명사구, 비한정 명사구, 고유 명사 등 모든 명사구를 대상으로 두 명사구가 서로 조응되는지를 결정 트리를 이용하여 식별하는 방법을 제안하였다[13,14].

한국어처리에서는 지시 표현 구별의 문제가 다루어진 바가 없고 다만, 지시 해석에 관한 연구에서 구별의 문제점이 제시된 바 있다. 한국어 대화에서 지시 해석에 관한 실험적 결과를 낸 연구로는 (김학수, 1997), (노현철외, 1998) 등이 있다. 김학수(1997)은 지시사가 없는 '다른 것', '옆의 것'과 같은 지시 표현 '것'이 대화에서 종종 등장하며, 지시사가 없기 때문에 문맥에 따라 지시하는 대상이 달라짐을 보였다[1]. 노현철외(1998)은 표면적으로 지시사가 없는 모든 담화 개체가 실제는 대용어가 아닌지 고려해야 함을 말했다[3].

3. '것'의 쓰임과 분류

'것'의 쓰임에 대해 사전[1]의 기술 내용을 참고하여 나열하면 다음과 같다.

첫째, 명사절 보문을 만드는 보문소

둘째, 사물, 현상, 사실 등 어떤 개체를 가리키는 표현.

셋째, 관용 표현의 요소.

첫 번째, 비지시 표현으로서 '것'은 대부분이 보문소[2]로서 표 2와 같이 명사절을 만드는 문법적 기능 외에 어떤 의미를 갖지 않는다.

표 2 비지시 표현 '것의 예

비지시 표현의 '것'을 포함하는 문장	동일 의미 다른 표현
그러면, 현금으로 내실 **것**인가요?	그러면, 현금으로 내시겠습니까?
환전을 하실 때 미국 달러로 교환하는 **것**이 좋습니다.	환전을 하실 때 미국 달러로 교환하시면 좋겠습니다.

두 번째, 지시 표현으로 쓰인 '것'의 예를 보면 표 3과 같다. 표 3에서 보는 것과 같이 지시 표현으로 쓰인 '것'은 실제로 가리키고 있는 문맥상의 개체나 추상적인 개체, 혹은 추론 가능한 어떤 개체를 이름하는 다른 명사로 바꿔 쓸 수 있다.

표 3 지시 표현 '것의 예

지시 표현 '것'을 포함하는 문장	동일 의미 다른 표현
예, 한 시 반 걸로(**것**으로) 부탁하고요.	예, 한 시 반 **비행기**로 부탁하고요.
한국에 여행 상품에 어떤 **것**이 있는지 알고 싶어요.	한국에 여행 상품에 어떤 **상품**이 있는지 알고 싶어요.
태안반도에서 주왕산으로 이동하는 **것**은 버스가 있습니다.	태안반도에서 주왕산으로 이동하는 **차량/교통편**은 버스가 있습니다.
당시는 일제 치하였기 때문에 읽은 책들은 일본말로 된 **것**일 수밖에 없었다.	당시는 일제 치하였기 때문에 읽은 책들은 일본말로 된 **책**일 수밖에 없었다.

세 번째 사전적 정의인 관용 표현이란 '고정된 형태나 구문'으로 쓰이는 것으로 예를 들면 표 4와 같다. 관용 표현은 고정된 구문으로 출현하는 것이므로 사전으로 매핑(mapping)할 대상이며 구별의 대상이 되지 못한다.

표 4 관용 표현 '것의 예

관용 표현 '것'을 포함하는 문장
어른 아이 **할 것 없이** 무차별하게 살상할 수 있는 집단은 그들뿐이라며 치를 떨었다.
너나 **할 것 없이** 아무나 나서서 해야 할 일이다.

1) 연세대학교 언어정보개발연구원 편. 연세한국어사전. 두산동아. 1998. 국립국어연구원 편. 표준국어대사전. 두산동아. 1999.

2) 보문소(complementizer)란 영어 'that과 같이 명사절 보충어(complement)를 만드는 것.

표 3에서 보인 바와 같이 지시 표현 '것'의 세부 종류는 지시 대상의 분포에 따라 달라지며 지시 해석 이후에 결정된다. 그러므로 본 논문에서는 지시 해석의 전 단계인 지시 표현 구별의 문제에 중점을 두어 지시 대상에 따른 세부적인 분류를 하지 않고 지시 표현으로서의 '것'과 그렇지 않은 비지시 표현으로서의 '것'만을 분류한다. 말뭉치 태깅 과정에서 '사실', '현상' 등의 의미를 가지면서 보문소 '음/기' 등으로 대체되는 경우들은 연세한국어사전과 달리 비지시 표현으로 보았다[6].

4. 지시 표현 '것'의 구별

4.1 기계 학습 기법의 이용 배경

'것'의 쓰임에 대해 사전의 기술 내용을 보면 '발견하다', '생각하다' 등의 인지나 감성 동사와 함께 쓰이는 경우나 '~ㄹ 것'의 꼴로 쓰이는 경우에 주로 보문소라는 것 이외에 지시 표현을 구별할 수 있는 특정 패턴이나 특징은 기술되어 있지 않다. 이 사실은 지시 표현이 되게 하는 언어적 규칙을 찾아내기가 어렵기 때문이며, 기계학습이나 통계적 방법을 이용하여 귀납적으로 패턴이나 규칙을 찾아내는 방법이 요구된다고 할 수 있다[6].

4.2 지시 표현 결정을 위한 자질 설정

지시 표현 '것'이 담화 상에서 지시 표현으로 인식되기 어려운 이유는 '것' 자체가 대명사나 한정 명사구와 같이 독립적인 지시 표현이 아니며, 언어적 속성과 주위 문맥에 따라 쓰임이 결정되기 때문이다. 이러한 문제를 해결하기 위하여 본 논문에서는 비교적 신뢰도가 높은 형태소 분석 결과만을 바탕으로 '것'의 언어문맥적 자질, 담화적 자질, 위치적 자질을 정의하고, 기계학습 방법의 하나인 C4.5 결정 트리(decision tree) 알고리즘을 이용하여 '것'의 쓰임을 구별한다[20]. 언어문맥적 자질이란 '것' 자체가 가지고 있는 언어적 속성과 '것'이 쓰이고 있는 문장 내에서의 앞뒤 문맥을 이용하여 추출된 자질을 의미하며, 담 화적 자질이란 '것'이 쓰인 문장의 속성이나 대화상에서 '것'이 쓰인 문장과 이전 문장과의 관계를 이용하여 추출된 자질을 의미한다. 위치적 자질은 '것'이 대화 내에서 쓰인 위치를 이용하여 추출된 자질을 의미한다.

언어문맥적 자질을 설정하기 위해 '것'이 의존 명사라는 품사적 정의를 이용한다. '것'의 품사적 정의는 그림 1과 같은 두 가지 언어학적 속성과 그에 따른 가설을 갖출 수 있게 한다.

그림 1의 '가설 1'을 이용하면 '것'의 앞에서 수식하는 요소에 관하여 다음과 같은 4가지 언어문맥적 자질을 설정할 수 있다.

(1) 어휘 기본형, (2) 어휘의 품사 태그, (3) 어미, (4) 어미의 품사 태그

마찬가지로 '가설 2'를 이용하면 '것'의 뒤에서 통어하는 요소에 관하여 다음과 같은 4가지 언어문맥적 자질을 설정할 수 있다.

(5) 격조사, (6) 격조사의 품사 태그, (7) 술어, (8) 술어의 품사 태그

품사 태그(part-of-speech tag)라는 자질이 어휘나 어미, 조사 등의 자질과 쌍을 이루는 이유는 어휘나 어미, 조사에 대해 보다 일반화된 정보를 분류 자질로 이용하여 지시 표현 구별의 효율성을 높이기 위해서이다.

담화적 자질을 추출하기 위해 '것'이 담화 상에서 초점화되거나 추론 가능한 어떤 개체를 대신하여 쓰인다는 사실을 이용한다. 그러므로 '것'이 쓰인 발화 안에 화제(topic)가 되고 있는 어떤 개체가 있는지 볼 필요가 있고, 화자 자신의 의지나 확신을 표현하거나 개체에 관한 단언을 하는 평서문인지, 청자에게로 의지를 돌리거나 개체에 관한 물음을 표현하는 의문문인지를 볼 필요도 있다. 이러한 성질을 이용하면 다음과 같은 2가지 담화적 자질을 설정할 수 있다.

의존 명사의 언어학적 속성
속성 1. '것'은 앞에서 수식하는 요소에 의존한다.
속성 2. '것'은 명사로서 격조사와 결합할 수 있다.
의존 명사의 언어학적 속성을 바탕으로 한 가설
가설 1. '것'은 수식하는 요소에 따라 지시적이거나 비지시적인 구분이 될 수 있다.
가설 2. '것'에 결합된 격조사와 성분 통어하는 술어에 따라 지시적이거나 비시적인 구분이 될 수 있다.

그림 1 의존 명사의 언어학적 속성과 지시 표현 구별을 위한 가설

표 5 지시 표현 구별을 위한 자질

자질 유형	자질명	설명
언어문맥	preprelem	'것'의 앞 실질 형태소로 용언의 기본형, 명사, 관형사 등.
	preprecat	preprelem의 품사 태그.
	prelem	'것'의 앞 형식 형태소로 어미나 접사이며, 선어말어미와 어말어미가 분리되지 않음. 어미나 접사가 없으면 null.
	precat	prelem의 품사 태그.
	postlem	'것'의 뒤 형식 형태소로 조사, 접사 등이 해당되며, 조사나 접사 없이 '것'의 다음이 공백이면 null.
	postcat	postlem의 품사 태그.
	postpostlem	'것'의 뒤 실질 형태소로 용언의 기본형, 명사, 관형사 등.
	postpostcat	postpostlem의 품사 태그.
담화	topichood	'것'의 앞에 체언에 결합한 '은/는' 보조사 여부.
	smark	의문문인지 평서문인지 여부(종결부호만 고려).
위치	worder	'것'이 몇 번째 어절인가/발화의 어절 길이.
	sorder	'것'을 포함한 문장이 몇 번째 발화인가/대화 내 위치.

(9) '것' 앞에 체언에 결합된 '은/는'의 존재 여부
(10) 의문문, 평서문 여부

위치적 자질은 '것'이 문장과 대화 내에서 출현하는 위치적 속성을 정의한 것으로 문장 내에서의 위치와 대화 내에서의 위치로 나누어지며 다음과 같은 2가지 자질로 나뉘어질 수 있다.

(11) 문장의 길이 대비 '것'의 출현 위치
(12) 대화 내에서의 '것'이 출현한 문장의 위치

지금까지 서술한 언어문맥적 자질, 담화적 자질, 위치적 자질을 정리하면 표 5와 같다.

본 논문에서는 표 5에서 정의한 12가지 자질의 값들을 '것'이 나타난 대화 문장에서 추출하여 C4.5 결정 트리의 입력으로 사용한다. 그림 2는 C4.5 결정 트리의 입력으로 사용된 인스턴스의 예이다. 그림 2에서 'pv', 'exm', 'pj', 's.'은 본 논문에서 사용한 형태소 분석기의 품사 태그들을 의미한다[5].

입력 문장	네, 그 패키지는요 숙식과 비행기 요금이 포함된 것이고 날짜는 아무 날짜나 원하시는 대로 할 수가 있습니다.
자질 패턴	2.29, 10, 포함되, pv, ㄴ, exm, null, pj, 이, pj, topic, s., ref

그림 2 자질 패턴의 예

5. 실험 및 결과 분석

5.1 실험 데이타 분석

말뭉치는 호텔 예약, 항공 예약, 여행 정보 분야에서 여행자와 여행사 직원, 여행자와 항공사 직원, 여행자와 그의 지인, 여행자와 호텔 직원이 전화로 주고받는 대화를 전사한 것으로 528개의 대화로 구성되었으며, 이 대화들의 발화 수는 모두 10,285 개이다. 이러한 말뭉치에서 '것'은 550번 출현하였고, 그 중 174개가 지시 표현이었다. 표 6은 말뭉치에서 나타난 여러 가지 지시 표현의 출현 빈도이다. 표 6에서 보는 것과 같이 지시 표현으로 사용된 '것'은 지시 관형사와 간투사를 포함한 '그' 다음으로 많이 쓰이는 표현으로 올바른 지시 해석을 위해 반드시 구별해야 하는 언어 표현이다.

표 6 지시 표현의 빈도

분류	빈도수(비율)	분류	빈도수(비율)
대명사 '그'	64개(11%)	대명사 '그녀'	0(0%)
대명사 '이것'	23개(4%)	지시 관형사 '이'	0(0%)
대명사 '저것'	0개(0%)	지시 관형사 '저'	0(0%)
대명사 '그것'	83개(14%)	대명사 '이'	0(0%)
지시 표현 '것'	174개(29%)	지시 관형사+간투사[3]	255(42%)

5.2 실험 설정 및 성능 평가

결정 트리는 학습 데이타에 담긴 규칙을 위계적인 구조의 'if then' 규칙으로 추출하는 기계 학습 기법으로 국외 연구에서 지시 해석 및 구별에 자주 이용되었다[13,14,16]. 본 논문에서는 기계 학습을 위한 공개소스인 Weka 3.0에 포함된 C4.5 알고리즘을 이용하여 학습하였으며[19], 성능 평가를 위해 10-fold 교차 검증(10-fold cross validation)을 수행하였다. 그리고, 기계 학습에 의해 도출된 규칙에 의한 시스템이 패턴으로 추출된 규칙에 의한 시스템에 비해 얼마나 향상된 성능을 보이는지를 비교하도록 하였다.

3) "그 사람 전화 번호가 뭐죠?"에서와 같은 관형사 '그'와 "저희가 그 종합된 서비스를 하고 있는데요?"에서와 같은 간투사 '그'가 형태소 분석기에 의해 정확하게 구분되지 않으며 그 구분이 모호한 측면이 있다.

모든 자질을 사용한 결정 트리	개별 어휘 자질을 제외한 결정 트리
prelem = 었던: com (2.0) prelem = 라는: com (4.0) prelem = 다는: com (2.0) prelem = 는 | postpostlem = 하: ref (3.0) | postpostlem = 좋 | | sorder <= 8 | | | smark = ?/s.: ref (2.0) | | | smark = ./s.: com (5.0/1.0) | | sorder > 8: com (16.0) | postpostlem = 있: ref (16.0) | postpostlem = 이: ref (43.0/13.0) | postpostlem = 욕: ref (2.0) | postpostlem = 예약하: ref (8.0) | postpostlem = 예약: ref (6.0) | postpostlem = 어떻: com (5.0) | postpostlem = 사: ref (2.0) | postpostlem = 라스베가스: ref (2.0) | postpostlem = 더: com (4.0) | postpostlem = 기대하: com (2.0) | postpostlem = 그랜드: ref (2.0) | postpostlem = 가능하: com (13.0) prelem = ㄹ: com (141.0/1.0) prelem = ㄴ: com (70.0/26.0) prelem = 시는: com (52.0) prelem = 신: com (7.0) prelem = 실: com (55.0) prelem = null: ref (49.0/6.0)	precat = exm | prelem = 었던: com (2.0) | prelem = 는 | | postpostcat = pv | | | postcat = jo: com (8.0/2.0) | | | postcat = ja: ref (12.0/1.0) | | postpostcat = pj | | | sorder <= 4: com (4.0) | | | sorder > 4 | | | | smark = ?/s. | | | | | worder <= 1.22: com (7.0/1.0) | | | | | worder > 1.22: ref (2.0) | | | | smark = ./s.: ref (28.0/3.0) | | postpostcat = pa | | | sorder <= 7: ref (16.0/2.0) | | | sorder > 7: com (44.0/5.0) | | postpostcat = nqc: ref (7.0) | | postpostcat = npp: com (2.0) | | postpostcat = nn: ref (10.0/1.0) | | postpostcat = nc: ref (13.0/2.0) | | postpostcat = jx: ref (2.0) | | postpostcat = a: com (6.0/1.0) | prelem = ㄹ: com (141.0/1.0) | prelem = ㄴ | | topichood = topic: ref (8.0/1.0) | | topichood = none | | | worder <= 1.27: ref (6.0/2.0) | | | worder > 1.27: com (56.0/15.0) | prelem = 시는: com (52.0) | prelem = 신: com (7.0) | prelem = 실: com (55.0) precat = eqm: com (6.0) precat = null: ref (49.0)

그림 3 자질에 따라 학습된 결정 트리[4)]

표 7은 4.2절에서 제안한 12개의 자질을 사용한 시스템의 성능을 보여주며, 그림 3의 첫 번째 열은 모든 자질을 사용하여 학습된 결정 트리의 모습이다. 그림 3에서 보듯이 모든 자질을 사용한 경우에 개별 어휘 자체를 의미하는 'preprelem'과 'postpostlem'이 지시 표현 구별의 주요 결정 요인으로 작용함을 알 수 있었다. 그러나 이러한 개별 어휘 자질들은 응용 영역과 데이타의 변화에 의존하는 경향이 있으므로 시스템의 성능을 불안정하게 만들 가능성이 매우 크다. 즉, 항공 예약, 호텔 예약 등에 관한 대화가 아닌 경우에 그림 3에 나타난 '출발하~', '예약하~', '라스베가스' 등의 어휘는 주요 어휘로 취급되지 못할 것이며, 시스템의 성능을 저하시키는 원인으로 작용할 것이다. 그러므로, 본 논문에서는 상기한 두 어휘 자질을 제외한 실험을 수행하여 그 결과를 살펴보았다. 표 8은 상기한 어휘 자질을 제외하고 10개의 자질만을 이용하여 학습한 시스템의 성능을 보여준다. 표 8가 보이는 바와 같이 개별 어휘 자질을 제외한 시스템의 성능이 모든 자질을 사용한 것에 비해 크게 떨어지지 않음을 알 수 있었다. 그림 3의 두 번째 열은 개별 어휘 자질을 제외하고 학습된 결정 트리의 모습이다. 그림 3에서 보는 것과 같이 모든 자질을 사용한 결정 트리에 비해 개별 어휘 자질을 대신한 품사들이 지시 표현의 구별에 주요한 요인으로 작용함을 알 수 있었다.

표 7 모든 자질을 사용한 시스템의 성능

Precision	Recall	F-Measure	Target Class
0.917	0.931	0.924	Com (비지시 표현)
0.843	0.814	0.828	Ref (지시 표현)

표 8 개별 어휘 자질을 제외한 시스템의 성능

Precision	Recall	F-Measure	Target Class
0.908	0.939	0.923	Com (비지시 표현)
0.855	0.791	0.822	Ref (지시 표현)

그림 4는 지시적 것과 비지시적 것을 분류하기 위해 말뭉치에서 추출된 패턴 규칙이다.

4) 가지 잘린 트리(Pruned tree)이며, 하나 이하의 인스턴스(instance)를 가지는 단말 노드(leaf nod)는 편의상 삭제하였다.

패턴에 의한 '것' 분류 규칙
1. '것'에 체언, 수식언 범주가 있으면 'ref'이다. 2. '것 같다'는 'com'이다. 3. '~ㄹ 것이다'는 'com'이다. 4. '~라 하는 것'은 'ref'이다 5. '~이라는 것'은 'ref'이다. 6. '~ 것이 있다'는 'ref'이다. 7. '~같은 것'은 'ref'이다.

그림 4 패턴에 따라 작성된 규칙

표 9는 개별 어휘 자질을 제외한 자질 집합으로 '것'을 분류하는 시스템의 성능과 패턴 규칙에 의한 성능을 보여준다. 기계 학습에 의한 규칙으로 '것'을 분류했을 때, 패턴 규칙에 의해 분류했을 때보다 약 15% 정도의 성능 향상이 있음을 보인다.

표 9 기계 학습에 의한 시스템의 성능과 패턴 규칙에 의한 시스템 성능

시스템	틀리게 분류된 개수	맞게 분류된 개수	전체 성능 (%)
기계 학습에 의한 규칙	59	491	89.27
패턴 규칙	140	410	74.5%

5.3 오류 분석

그림 3의 결정 트리는 규칙과 동시에 괄호 안에 '왼쪽 숫자/오른쪽 숫자' 제시를 통해 해당 규칙이 적용되었을 때의 '맞는 분류 개수/틀린 분류 개수'를 보이고 있다. 표 9에서 기계 학습에 의한 규칙은 그림 3의 오른 쪽 결정 트리 규칙을 적용했을 때 잘못 분류된 개수가 59개임을 보인다. 이 59개 중 결정 트리 규칙으로 잘못 분류된 개수는 37개이고 나머지는 규칙으로 도출되지 않은 경우들이다. 표 10은 지시 표현과 비지시 표현으로서 '것'을 구별하도록 도출된 규칙에 대해서 오류를 보이는 경우에 대한 정리표이다. 대부분의 오류가 개별 어휘 자질의 이용과 담화 맥락에 대한 의미적 처리를 하지 않았기 때문임을 보인다.

5.4 자질 조정에 따른 성능 변화

우리는 지시 표현 구별에 우수한 자질들을 보기 위해 자질 조정에 따른 성능의 변화를 실험하였다. 표 11은 언어문맥 자질만을 사용한 시스템의 성능을 보여준다. 이는 언어문맥적 자질만으로도 비지시 표현 분류와 지시 표현 분류에 있어서 각각 91.5%와 79.4%의 F-measure를 얻을 수 있음을 보인다. 이러한 실험 결과를 바탕으로 지시 표현 구별 문제에 있어서 언어문맥 자질의 중요성과 언어문맥 자질을 도출한 가설들의 정당함을 알 수 있었다.

표 11 언어 문맥 자질만을 사용한 시스템의 성능

Precision	Recall	F-Measure	Target Class
0.89	0.94	0.92	Com (비지시 표현)
0.86	0.74	0.79	Ref (지시 표현)

표 12는 언어 문맥 자질에 담화 자질을 추가한 시스템의 성능을 보여준다. 표 12에서 보듯이 담화 자질을 추가한 경우에 지시 표현 분류와 비지시 표현 분류 모두에서 그렇지 않은 경우보다 높은 재현율과 정확률을 보였다. 표 11에 위치 자질을 추가한 경우는 표 8과 같으며, 위치 자질의 추가는 정확률을 떨어뜨렸지만 재현율을 상대적으로 향상시킴으로써 보다 높은 F-measure를 얻게 하였다. 하지만 그 성능의 변화가 크지 않아 위치 자질의 유용성이 다른 자질들에 비해 떨어짐을 알 수 있었다.

자질 추가에 따른 성능의 변화를 그래프를 통하여 살펴 보면, 비지시 표현 구별의 경우 그림 5와 같고, 지시 표현 구별의 경우 그림 6과 같다.

표 12 언어문맥 자질과 담화 자질을 사용한 시스템의 성능

Precision	Recall	F-Measure	Target Class
0.89	0.96	0.92	Com (비지시 표현)
0.88	0.75	0.81	Ref (지시 표현)

표 10 분류 오류 표

1		
precat = exm | prelem = 는 | | postpostcat = pv | | | postcat = jo: **com** (8.0/2.0)	com	네 저가 요금을 여쭈어 보는 **것**을 잊어 버리었군요.
	ref	그럼 사월달에 출발하는 것을 예약하고 싶은데요.
2		
precat = exm | prelem = 는 | | postpostcat = pv | | | postcat = ja: **ref** (12.0/1.0)	ref	예 그럼 사월 삼일날 출발하는 **것**으로 예약하고 싶습니다.
	com	여행 일정은요 팔월 십구일 오전 일곱시에 광주 사직공원 후문에서 출발해서 팔월 이십일 오후 여섯시 삼십분에 동일 장소에 도착하는 것으로 되어 있습니다.
3		

precat = exm | prelem = 는 | | postpostcat = pj | | | sorder > 4 | | | | smark = ?/s. | | | | | worder <= 1.22: **com** (7.0/1.0)	com	네 그러면 식사는 어떻게 되는 **것**이지요?
	ref	그럼 식비는 연수비에 포함이 되어 있는 **것**인가요?
4		
precat = exm | prelem = 는 | | postpostcat = pj | | | sorder > 4 | | | | smark = ./s.: **ref** (28.0/3.0)	ref	오전 열시 삼십분은 뉴욕을 출발해서 하와이를 거치어서 가는 것이고 그 다음 **것**은 직행입니다.
	com	그것 끝나고 나아서 에트리를 방문할 수 있을지 싶어서 전화를 하는 **것**입니다
5, 6 sorder 자질 값은 대화 전체를 보이지 않아서 드러나지 않음.		
precat = exm | prelem = 는 | | postpostcat = pa | | | sorder <= 7: **ref** (16.0/2.0) | | | sorder > 7: **com** (44.0/5.0)	ref	십사일 출발하는 것이 요세미티로 가는 **것**이 있는데 그것은 오 일간이고요.
	com	환전은 미국 달러로 하는 **것**이 좋겠습니다.
7 이 규칙에서 com에 해당하는 오류의 경우는 아래 예에서처럼 작업자의 태깅 오류였다.		
precat = exm | prelem = 는 | | postpostcat = nn: **ref** (10.0/1.0)	ref	팔일이나 구일에 출발하는 상품은 라스베가스로 가는 **것** 하나밖에 없는데요.
	com	네 저희가 취급하고 있는 것은 십사일짜리인데요.
8		
precat = exm | prelem = 는 | | postpostcat = nc: ref (13.0/2.0)	ref	예 태안반도에서 주왕반 주왕산으로 이동하는 **것**은 버스가 있습니다.
	com	팜플렛 보내어 드리는 **것** 팩스로 보내어 드리어도 괜찮면은 팩스번호 좀 알리어 주시겠어요?
9		
precat = exm | prelem = 는 | | postpostcat = a: com (6.0/1.0)	com	손님한테 연락이 오는 **것**이 너무 늦어지어 가지구요 취소가 되어 버리었는데요.
	ref	설악산에 가는 **것**으로 좀 더 알아보고 싶은데요.
10		
precat = exm | prelem = ㄹ: com (141.0/1.0)	com	네 저 혼저 투숙할 것인데요.
	ref	다른 것 뭐 도와 드릴 **것**이 있을까요?
11		
precat = exm | prelem = ㄴ | | topichood = topic: ref (8.0/1.0)	ref	이 스키장은 사람들에게 아주 널리 알리어진 **것**으로요 상당히 권유드릴 만한 곳입니다.
	com	네 샌프란시스코는 경유하는데 시애틀은 아닌 **것** 같습니다.
12		
precat = exm | prelem = ㄴ | | topichood = none | | | worder <= 1.27: ref (6.0/2.0)	ref	그 여행 요금에요 입장료와 비행기 요금이 다 포함이 된 **것**입니까?
	com	잘못 말한 **것** 같은데요.
13		
precat = exm | prelem = ㄴ | | topichood = none | | | worder > 1.27: com (56.0/15.0)	com	네 구월 이십오일이면은 시간이 괜찮은 **것** 같군요.
	ref	예 오월 삼일에서 오월 사일에 출발하는 **것**이 좋은 것이 없을까요?

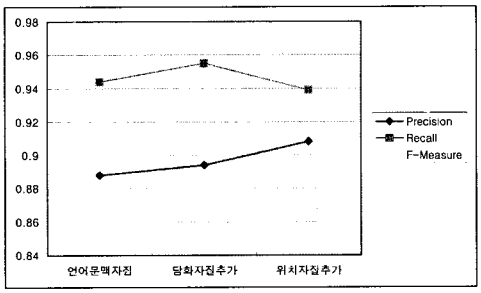

그림 5 자질 조정에 따른 비지시 표현 분류 성능의 변화

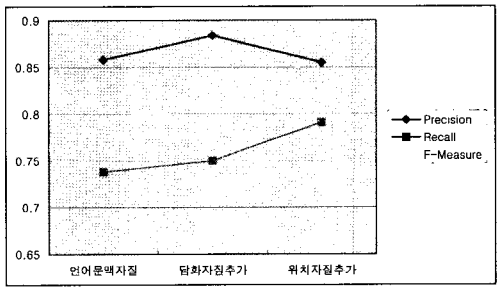

그림 6 자질 조정에 따른 지시 표현 분류 성능의 변화

6. 결론 및 향후 과제

한국어 대화에서 '~ 것'과 같은 지시 표현의 출현이 빈번하지만 비지시적인 쓰임의 경우들과의 구별이 어렵기 때문에 기존의 지식 해석에서 그 대상이 되지 못하였다. 이러한 문제를 해결하기 위하여 본 논문에서는 결정 트리를 이용하여 지시 표현 '것'을 구별하는 방법을 제안하였다. 제안한 방법은 형태소 분석 결과를 바탕으로 '것'이 가지는 언어적 속성, 문맥적 속성, 담화적 속성 그리고 위치적 속성을 자질로 추출하고 이것을 결정 트리의 입력으로 사용하여 '것'의 쓰임을 선택한다. 실험을 통하여 '것'의 구별에 가장 유용한 자질은 언어문맥적 자질임을 알 수 있었으며, 담화적 자질과 위치적 자질도 성능 향상에 기여함을 알 수 있었다. 또한 응용 영역에 의존적인 개별 어휘 자질을 제외한 언어문맥 자질을 사용한 경우가 그렇지 않은 경우에 비해 성능 하락이 크지 않아 일반화하기에 가장 좋은 자질임을 알 수 있었다.

향후에는 제안된 시스템을 지시 해석 시스템에 통합하여 지시 해석의 성능 향상에 미치는 영향을 실험해야 할 것이다. 그리고, 보다 많은 사용 예를 담은 말뭉치를 확보하여 지시 표현 '것'의 다양한 용법과 쓰임에 대해 연구해야 할 것이다. 또한, 지시 해석의 성능 향상에 기여할 수 있도록 지시 표현의 범주를 세분화하고 구별하는 방법을 연구해야 할 것이다.

참 고 문 헌

[1] 김학수, 다중모드 대화 시스템에서의 명사 대용어구 처리, 석사학위논문, 서강대학교, 1997.

[2] 남기심, "불완전명사 '것'의 쓰임", 국어의 이해와 인식, 한국문화사, 1991.

[3] 노현철, 이근배, 이종혁, 박재득, "한국어 담화 특성에 기반한 영역 독립 생략 및 대용 처리", 정보과학회논문지(B) 제25권 제12호, pp. 1845-1857. (1998).

[4] 양명희, 현대국어 대용어에 대한 연구, 국어학총서33, 국어학회, 태학사, 1998.

[5] 이상호, 미등록어를 고려한 한국어 품사 태깅 시스템 구현, 석사학위논문, 한국과학기술원, 1995.

[6] 조은경, 이민행, "지시 해석을 위한 것의 구별과 쓰임에 관한 연구", 한국어학 제 31집, 한국어학회, 2006.

[7] MAK Halliday and Ruqaiya Hasan, Cohesion in English, Longman, 1976.

[8] Shalom Lappin and Herbert J. Leass, "An Algorithm for Pronominal Anaphora Resolution," Computational Linguistics, volume 20, number 4. (1994).

[9] Lluís Màrquez, "Machine Learning and Natural Language Processing," Technical Report LSI00-45-R, Departament de Llenguatges i Sistemes Informatics (LSI), Universitat Politecnica de Catalunya (UPC), Barcelona, Spain (2000).

[10] David L. Bean and Ellen Riloff., "Corpus-Based Identification of Non-Anaphoric Noun Phrases", In the proceedings of ACL. (1999).

[11] Richard Evans, "Applying Machine Learning Toward an Automatic Classification of It," Literary and Linguistic Computing (2001).

[12] Renata Vieira and Massimo Poesio, "Processing definite descriptions in corpora," Corpus-based and computational approaches to discourse anaphora. Simon Botley and Anthony Mark McEnery.(ed.) Benjamins Pub. 2000.

[13] Vincent Ng and Claire Cardie, "Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution," In the proceedings of COLING. (2002).

[14] Vincent Ng and Claire Cardie "Learning Noun Phrase Anaphoricity to Improve Coreference Resolution-Issues in Representation and Optimization," In the proceedings of ACL. (2004).

[15] Antonio Ferrández and Lidia Moreno "A computational approach to pronominal anaphora, one-anaphora and surface count anaphora," In the proceedings of Discourse Anaphora and Anaphora Resolution (1998).

[16] Michael Strube and Christoph Muller, "A Machine Learning Approach to Pronoun Resolution in Spoken dialogue," In the proceedings of ACL. (2003).

[17] Didier Baltazart and Laurence Kister, "Is it possible to predetermine a referent included in a French N De N structure?," Corpus-based and computational approaches to discourse anaphora. Simon Botley and Anthony Mark McEnery.(ed.) Benjamins Pub, 2000.

[18] Joseph F. McCarthy and Wendy G. Lehnert, "Using Decision Trees for Coreference Resolution," In the proceedings of International Joint Conference on Artificial Intelligence (1995).

[19] Ian H. Witten and Eibe Frank, Morgan Kaufmann. Data Mining: Practical machine learning tools and techniques, San Francisco, 2005.

[20] Quinlan R. J., C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

조 은 경

1997년 2월 연세대학교 국어국문학과 졸업. 2001년 2월 연세대학교대학원 언어정보학과 석사. 2005년 8월 서강대학교 컴퓨터학과 석사. 2007년 8월 연세대학교 대학원 언어정보학과 박사. 2006년~현재 (주)다음커뮤니케이션 검색포털본부 데이터마이닝팀. 관심분야는 한국어정보처리, 대용어 처리, 질의응답시스템

김 학 수

1996년 건국대학교 전자계산학과 학사 1998년 서강대학교 컴퓨터학과 석사 2003년 서강대학교 컴퓨터학과 박사 2004년 University of Massachusets, Amherst 박사후연구원. 2005년 한국전자통신연구원 선임연구원. 2006년~현재 강원대학교 컴퓨터정보통신공학전공 전임강사. 관심분야는 한국어정보처리, 생략 및 대용어 처리, 대화 인터페이스 시스템, 정보검색 시스템, 질의응답 시스템

서 정 연

1981년 서강대학교 수학과 학사. 1985년 미국 Univ. of Texas, Austin 전산학과 석사. 1990년 미국 Univ. of Texas, Austin 전산학과 박사. 1990년~1991년 미국 Texas Austin, UniSQL Inc. Senior Researcher. 1991년 한국과학기술원 인공지능 연구센터 선임연구원. 1991년~1995년 한국과학기술원 전산학과 조교수. 1995년~1996년 서강대학교 전산학과 조교수. 1996년~2001년 서강대학교 컴퓨터학과 부교수. 2001년~현재 서강대학교 컴퓨터학과 교수. 관심분야는 한국어정보처리, 대화 인터페이스 시스템, 지능형 로봇 상호작용, 웹기반 정보검색, 문서요약, 문서분류, 기계번역