구문 정보와 비용기반 중심화 이론에 기반한 자연스러운 지시어 생성

(Generation of Natural Referring Expressions by Syntactic Information and Cost-based Centering Model)

노 지 은 † 이 종 혁 ††
(Ji-Eun Roh) (Jong-Hyeok Lee)

요 약 텍스트 생성(text generation)은 언어가 아닌 다양한 지식원으로부터 텍스트를 생성해 내는 언어 처리의 한 분야로, 여러 가지 복합적이고 단계적인 과정을 통해 이루어진다. 본 논문에서는 자연스러운 텍스트 생성을 위한 여러 과정 중, 한번 언급된 대상(entity)을 자연스럽게 지시(refer)하기 위한 지시어 생성(referring expression generation), 특히 한국어에 두드러진 영형(zero pronoun)에 의한 대용화(pronominalization) 과정에 초점을 맞춘다. 이를 위해, 구문 정보와 비용기반 중심화 이론(cost-based centering model)을 바탕으로, 한국어에 적합한 지시어 특히 영형의 생성에 영향을 미치는 다양한 자질(feature)들을 규명하고, 기계 학습을 통해 지시어 생성 모델을 구축하였다. 세 개의 장르 - 묘사문(설명문), 뉴스, 짧은 우화 - 에서 총 95개의 텍스트로부터 학습이 이루어 졌으며 이를 대상으로, 제안된 자질들이 지시어의 생성, 특히 영형의 생성에 효율적으로 적용될 수 있음을 보였다. 또한, 지시어 생성과 관련된 기존의 방법론들과 본 논문에서 제안한 모델을 비교하여 성능이 크게 향상되었음을 보이고, T-test를 통해 99.9%의 신뢰 구간에서 그 성능 향상이 통계적으로 의미가 있음을 확인하였다.

키워드 : 텍스트 생성, 지시어 생성, 대용화, 영형, 자질, 비용기반 중심화 이론, 기계 학습

Abstract Text Generation is a process of generating comprehensible texts in human languages from some underlying non-linguistic representation of information. Among several sub-processes for text generation to generate coherent texts, this paper concerns *referring expression generation* which produces different types of expressions to refer to previously-mentioned things in a discourse. Specifically, we focus on pronominalization by zero pronouns which frequently occur in Korean. To build a generation model of referring expressions for Korean, several features are identified based on grammatical information and cost-based centering model, which are applied to various machine-learning techniques. We demonstrate that our proposed features are well defined to explain pronominalization, especially pronominalization by zero pronouns in Korean, through 95 texts from three genres - Descriptive texts, News, and Short Aesop's Fables. We also show that our model significantly outperforms previous ones with a 99.9% confidence level by a T-test.

Key words : text generation, referring expression generation, pronominalization, zero pronoun, feature, cost-based centering model, machine learning

1. 서 론

텍스트 생성(text generation)은 자연어로 이루어지지 않은 기저의 정보들을 자연어로 사상하여 텍스트를 생성해 내는 언어 처리의 한 분야로, 여러 문장이 긴밀히 결합되어 하나의 정보를 전달하는 단위를 텍스트라 볼 때, 양질의 텍스트를 생성하기 위해서는 문장간의 순서, 문장간의 결합, 각 문장들에서의 지시어 생성 등을 적절히 처리해 주어야 한다. 국내에서는, 한국어를 대상으로 한 텍스트 생성에 관한 연구가 거의 이루어 지지 않았으며, [1]에서 제안한 데이타베이스로부터 홈쇼핑 사이트의 상품 소개를 위한 텍스트 생성이 그 첫 시도라 할 수 있다.

· 본 연구는 첨단정보기술연구센터를 통한 과학재단 및 2002년도 두뇌한국21사업에 의하여 지원 되었음
† 비 회 원 : 포항공과대학교 컴퓨터공학과
jeroh@postech.ac.kr
†† 종신회원 : 포항공과대학교 컴퓨터공학과 교수
jhlee@postech.ac.kr
논문접수 : 2004년 6월 19일
심사완료 : 2004년 10월 4일

본 논문에서는 자연스러운 텍스트 생성을 위한 여러 과정 중, 지시어의 생성(referring expression generation), 특히 대용화(pronominalization) 과정에 초점을 맞춘다. 텍스트나 일상적인 담화에서 이전 문장 또는 발화에 한번 언급된 대상(entity)을 다시 언급하는 경우, 동일한 표현을 반복해서 쓰기 보다는 적절한 조응 표현(anaphoric expression)으로 바꿔 쓰는 것이 일반적이다. 동일한 표현을 반복해서 쓸 경우, 정보의 잉여성(redundancy)으로 인해 정보 전달의 효율성(effectiveness)과 가독성(readability)이 떨어지고, 전체적으로 텍스트의 결속성(coherence)이 저하된다.

이러한 지시어의 생성은 언어마다 서로 다른 특징을 갖는데, 한국어는 영어와 달리 조응 표현으로 영형(zero pronoun)의 사용이 두드러진다. 한국어는 문맥 의존 언어(context-dependent language)로, 주어진 문맥에서 회복 가능한(recoverable)한 임의의 논항(argument)들은 기본적으로 모두 생략 가능하다.

본 논문에서는, 이러한 한국어의 특징을 반영해 자연스러운 지시 표현, 특히 적절한 영형의 생성에 그 목적이 있으며, 이를 위해 다양한 자질(feature)들을 규명하고 이러한 자질들을 바탕으로 학습 데이타를 이용, 여러 가지 기계학습(machine learning)을 적용해 한국어에 적합한 지시어 생성 모듈을 구축하고자 한다.

2. 중심화 이론

중심화 이론[2](Centering Theory)은, 자연스러운 지시어 생성을 위해 본 논문에서 제안하는 자질과, 지시어 생성을 위한 기존 방법론들의 배경 이론이 되므로 본 절에서 간단히 설명한다.

중심화 이론은 텍스트를 구성하고 있는 발화의 각 명사(구) 관점에서 응집성(cohesion)과 현저성(salience)의 상호작용(interaction)을 통해, 텍스트의 국소적 결속성(local coherence)을 모델링한 담화 해석의 계산 모델이다.

중심화 이론에서 분석의 최소 기본 단위는 발화(utterance)로, 한 개의 발화는 두 개의 중심구조 - Cf (forward-looking center), Cb(backward-looking center) - 를 가진다. 이때 중심(center)은 발화 시점에서 화자의 의식이 활성화되고 집중되어 있는 대상물들을 말한다. Cf는 현 발화에서 실현된 객체 지시물들이고, Cf-list는 발화에 실현된 객체 지시물들을 화자의 의식 내에서 활성화된 정도에 따라 서열을 매긴 것으로, 다음 발화에 나타나게 될 지시물에 대한 선행사(antecedent)의 집합이다. Cf-list에 있는 지시물 중에서 가장 높은 서열에 있는 지시물은 Cp(preferred center)가 되며, Cp는 다음절에서 주제로 논의될 가능성이 가장 높은 후보자이다. Cb는 문장의 주제(topic)와 유사한 개념으로, 많은 경우에 바로 앞의 발화의 Cp가 다음 발화에서 Cb가 된다.

중심화 이론에서 가정하는 세 가지 제약과 두 가지 규칙은 다음과 같다.

▣ 제약(constraints)

1. 각 발화 내에 하나의 Cb가 있다.
2. 각 발화의 Cf 목록의 모든 요소는 반드시 현 발화 안에서 실현되어야 한다.
3. 각 발화의 Cb는 현 발화(U_i)에서 실현된, 바로 전 발화(U_{i-1})의 Cf에서 가장 높은 순위의 담화 요소이다.

▣ 규칙(rules)

1. 앞 발화(U_{i-1})의 Cf의 어떤 요소가 현 발화(U_i)에서 대명사화 되었다면, 현 발화(U_i)의 Cb도 역시 대명사화 된다.
2. 발화간의 전이 유형은 다음 순서로 선호된다.
   CONTINUE > RETAIN > SMOOTH-SHIFT > ROUGH-SHIFT

전이 유형은 Cb(U_i)와 Cb(U_{i-1})의 일치 여부와, Cb(U_i)와 Cp(U_i)의 일치 여부에 의해 결정되는데, 전자의 조건이 만족될 경우 현 발화는 응집성(cohesion)이 있다고 보며, 후자의 조건이 만족될 경우, 현 발화는 현저성(salience)이 있다고 본다. CONTINUE는 화자가 특정 지시물에 대해 이야기 하고 있으면서, 다음 발화에서도 계속 그 지시물에 대해 이야기 하겠다는 의도를 표시하며, 그 지시물은 현 발화에서 Cb인 동시에 Cp로 표현된다. RETAIN은 화자가 다음절에서 의식의 중심을 새로운 대상으로 옮기고 싶다는 의도를 표시하는 것으로, Cb는 그대로 유지되지만, 현재의 중심을 Cf-list에서 낮은 서열에 배치함으로써 Cp가 바뀌는 경우이다. SMOOTH-SHIFT는 이전 발화와 비교해 중심은 변했지만, 새로운 중심에 대해이야기하며 다음 발화에서도 계속 새롭게 바뀐 현재의 중심에 대해 이야기 하겠다는

표 1 발화간의 전이유형(transition type)

	Cb(U_i) = Cb(U_{i-1}) or C(U_{i-1}) is undefined	Cb(U_i) != Cb(U_{i-1})
Cb(U_i) = Cp(U_i)	CONTINUE	SMOOTH-SHIFT
Cb(U_i) != Cp(U_i)	RETAIN	ROUGH-SHIFT

의도를 표시한다. ROUTH-SHIFT는 중심도 변하고 새로운 중심이 Cf-list에서 낮은 서열에 배치되는 경우이다.

중심화 이론에서 발화의 단위와 Cf-list에서 순위 결정은 언어에 따라 조금씩 다르며, 같은 언어에 대해서도 학자들마다 그 설정 기준이 다르다. 본 연구에서는 발화의 단위를 시제절로 정의하고,[1] [6]에서 설정한 Cf-순위의 랭킹과 전이 유형 선호도를 따른다.[2]

▣ 본 논문에서 따르는 Cf-list의 순위

주제[3] > 주어 > DAE > 직접 목적어 > 간접 목적어 > 나머지 논항

▣ 전이 유형 선호도

CONTINUE > EXP-CONTINUE > ASSOCIATE-SHIFT> RETAIN > SMOOTH-SHIFT > EXP-SMOOTH-SHIFT > ROUGH-SHIFT > RESUME > NO-Cb

중심화 이론은 이론 자체의 단순함에도 불구하고, 자연어 처리의 여러 분야에 응용되는 강력한 계산 모델로써, 초기에는 대용 해석을 위해 주로 활용되었지만, 최근에는 텍스트 생성의 여러 과정 텍스트 구조화[6,7] (text structuring), 문장 계획[8] (sentence planning), 지시어 생성[9-11] (referring expression generation)에 널리 적용되고 있다.

3. 지시어 생성에 관한 기존 연구

지시어 생성에 있어 가장 원시적인 방법은 [12]에서 논의된 것으로, '이전 문장에서 언급된 지시물이 현재 문장에서 다시 언급되는 경우, 그것들은 다 대명사화 하라'는 것이었다. 하지만, 이는 대명사를 과다생성(over-generation)하게 하여 각 대명사의선행사를 모호하게 하고 전체적으로 텍스트의 애매성(ambiguity)을 가중시키는 문제점을 안고 있다.

(R1-McKeown) If U_i contains the same word mentioned in U_{i-1}, use a pronoun to refer to the word

중심화 이론을 지시어 생성에 적용한 기존 연구 중, [10]에서는 대명사를 생성하기 위한 여러 가지 전략들 가운데, 'CONTINUE에서 Cb를 대명사화'하는 것이 가장 적합하다고 본다.

(R2-Kibble) If tran(U_i) (transition of U_i) is CONTINUE, pronominalize the Cb(U_i)

[11]에서는 일본어에서, 모든 CONTINUE와 SMOOTH-SHIFT에서 Cb는 영형으로 생성될 수 있다고 주장한다.

(R3-Yamura) If tran(U_i) is CONTINUE or SMOOTH-SHIFT, pronominalize Cb(U_i) as a zero pronoun

[13]에서는, 힌두어에서 영형의 발생은 빈번하나 그 발생에 있어 다소 엄격한 제약을 받으며, U_i가 CONTINUE이고 U_{i-1}이 CONTINUE이거나 SMOOTH-SHIFT일 때 U_i에서 Cb는 생략 가능하다고 본다.

(R3-Prasad) If tran(U_{i-1}) is CONTINUE or SMOOTH-SHIFT and tran(U_i) is CONTINUE, pronominalize Cb(Ui) as a zero pronoun

이상의, 중심화 이론에 기반한 기존 연구들의 문제점들은 4.1절과 4.2.2에서 자세히 논의된다.

[14]에서는 일본어에서 영형을 포함한 다양한 지시어를 생성하기 위해 뉴스기사를 분석하고 수동으로 결정트리(decision tree)를 구축하였지만, 뉴스 장르에 의존적일 뿐만 아니라, 테스트 집합 또한 아주 작아 그 유용성을 입증하기가 힘들다.

[15]에서는 중국어를 대상으로 다섯 가지 조건에 근거하여 지시어 생성을 위한 결정트리를 제안하였는데, 그들이 고려한 자질 중, 지역성(locality)과 주제의 연속성(topic continuity)은 지시어 생성에 있어 새로운 개념이라 할 수 없고, 통사적 제약은 중국어에만 적용 가능한 것이라 한국어 지시어 생성에는 적합하지 않다.

4. 지시어 생성을 위한 자질

4.1 지시어 생성 대상

중심화 이론에 기반해 지시어 생성을 다루는 대부분의 기존 연구들 (예, [10], [11], [13])은, 실제 많은 영형 및 그 외의 대용형이 Cb 뿐만 아니라 Cb를 제외한 나머지 구정보(old information)[4]에서 빈번히 발생함에도 불구하고, 지시어 생성의 대상(target)으로 오직 Cb만을

1) 중심화 이론에서는, 이 이론을 구성하는 여러 개의 인자(parameterer)들 - '발화(utterance)', '실현(realization)', 'Cf-list의 랭킹(ranking)' - 에 대한 구체적인 정의가 명시되어 있지 않다. 따라서, 중심화 이론을 적용하는 많은 논문들이 임의로 이러한 인자들을 정의해 왔으며, 특히 [3]에서는 각 인자의 다양한 정의에 따라 중심화 이론이 실제 텍스트에 얼마나 효과적으로 적용될 수 있는지에 관해 논의하였다. [4]에서는 발화를 *시제절*로 정의하였으며, [5]에서는 *문장*으로 정의하였다. 본 논문에서는 발화의 단위를 [4]에서와 마찬가지로 시제절로 정의 하였는데, 발화의 단위를문장으로 정의하는 것보다 시제절로 정의하는 것이 중심화 이론에서 제약1의 위반 횟수를 줄일 수 있기 때문이다.

2) 추가적으로 고려된전이유형 중 EXP-CONTINUE와 EXP-SMOOTH-SHIFT에 대한 상세한 설명은 절 4.2.2를 참고, 나머지 전이 유형과 Cf-list의 순위 설정에 대한 자세한 설명은 [6]을 참고.

3) 본 논문에서 '주제(topic)'는 주제 표지(topic marker) '은/는'을 조사로 갖는 단어를 의미한다.

4) 본 논문에서 '구정보(old-information)'는 직전 발화 U_{i-1}에서 언급된 객체 중, 현 발화 U_i에서 다시 언급된 객체들로 정의하며, 계산의 편이성을 위해 문맥이나 상황에 의해 청자와화자가 추론할 수 있는 객체들은 배제한다.

$oldA(U_i) = Cf(U_i) \cap Cf(U_{i-1})$ --- $oldA(ll)(U_i)$: U_i의 구정보들, $Cf(U_i)$: U_i에 실현된 객체들
$oldCp(U_i) = oldA(U_i) \cap Cp(U_i) \cap \neg Cb(U_i)$
$oldR(U_i) = oldA(U_i) - Cb(U_i) - oldCp(U_i)$

$y = f_{oldR}(x_{oldR})$ --- $f_{oldR}(x_{oldR})$: oldR를 위한 지시어 생성 함수
$y = f_{oldCp}(x_{oldCp})$ --- $f_{oldCp}(x_{oldCp})$: oldCp를 위한 지시어 생성 함수
$y = f_{Cb}(x_{Cb})$ --- $f_{Cb}(x_{Cb})$: Cb를 위한 지시어 생성 함수

$x_{oldR} \in oldR(U_i)$, $x_{oldCp} = oldCp(U_i)$, $x_{Cb} = Cb(U_i)$
$y \in$ {NP, NP 변이형, 영형이 아닌 대명사, 영형}

다루고 있다는 데 가장 큰 한계가 있다. 본 논문에서 수집된 텍스트 집합에서, Cb에서 발생한 대용 표현은 전체 구정보에서 발생한 대용표현의 61%에 해당되며, 나머지 39%는 Cb가 아닌 나머지 구정보에서 실현되었음을 확인하였다(표 4 참고).

또한, 중심화 이론에 이론적 근거를 두지 않는 대부분의 연구들(예, [14], [15])은, 지시어 생성의 대상으로 모든 구정보들을 다 포함시키되, 각 구정보들이 갖는 위상 차이에 대한 고찰 없이 동등하게 인식함으로써, 지시어 생성의 정교한 모델을 만들지 못했다는 점에 한계가 있다. 실제, 구정보라는 큰 범위 안의 각각의 정보들은 중심화 이론에 비추어 볼 때, 텍스트의 응집성(cohesion)을 유지하는 지시물들이 있는가 하면, 주제를 명확하게 드러내 현저성(salience)을 유지하는 지시물이 있기도 하고, 단지 정보 기능 면에서 새롭지 않은 정보일 수도 있다. 따라서, 본 논문에서는 구정보라는 범위 안에서 각 정보들이 갖는 위상 차이는 지시어 생성에 서로 다른 경향(tendency)를 가진다는 가정하에, 모든 구정보들을 지시어 생성의 대상으로 포함시키고, 다음과 같이 세 가지로 분류하여 지시어 생성 모듈을 독립적으로 구축하고자 한다.

▣ U_i에서 지시어 생성의 대상 $oldA(U_i)$
① $Cb(U_i)$: 중심화 이론의 Cb
② $oldCp(U_i)$: 구정보이면서 Cb가 아닌Cp
③ $oldR(U_i)$: 구정보 중 Cb와 oldCp를 제외한 나머지

중심화 이론에서 Cb는 현 발화에서 언급된 지시물 가운데 이전 발화에 대해 가장 응집적인(cohesive) 객체로, 텍스트의 응집성(cohesion)의 지표로 삼을 수 있는 반면, Cp는 현 발화에서 언급된 지시물 가운데 가장 현저하게 두드러지는(salient) 객체로 텍스트의 현저성(salience)의 지표로 삼을 수 있다. 따라서, oldCp는 구정보인 동시에 응집적이지는 않지만 현저한 지시물을, oldR은 응집적이지도 현저하지도 않고 정보기능상 구정보의 역할만을 하는 지시물을 의미한다. 응집성과 현저성이 대용 해결(anaphor resolution)과 지시어 생성에 있어 갖는 영향력은 전통적으로 많은 논의가 되어 왔는데, 본 논문에서는 구정보의 이런 세가지 분류를 통해 응집성과 현저성이 지시어 생성에 미치는 영향을 관찰하고자 한다.

4.2 선택된 자질들

지시어 생성을 위해 본 논문에서 고려하는 자질들이 표 2에 정리되어 있다. 자질 1부터 4까지는 구정보의 통사적 정보(syntactic information)들이고, 자질 8부터 12까지는 중심화 이론에 이론적 근거를 둔 것이며, 나머지 자질 5에서 7까지는 대용형 생성에 유용하다고 알려진 자질들이다. 자질 1부터 자질 7까지는 모든 구정보(Cb, oldCp, oldR)에 대해 적용 가능한 반면, 중심화 이론에 기초한 자질들(8~13)은, Cb와 oldCp에 대해서만 적용 가능하며 oldR은 중심화 이론과 관계가 없는 지시물들이므로 배제된다.

4.2.1 일반적인 구정보(oldA)를 위한 자질들

[4]에서는 한국어에서 영형의 동기를 중심화 이론으로 설명할 수 있는 가를 구전으로 내려오던 민담을 필사해 놓은 자료에 대한 통계적 분석을 바탕으로 검증해 보았다. 특히, 영주어(zero subject)와는 달리 영목적어(zero object)는 중심화 이론만으로 설명이 부족하다고 보고, 영목적어의 분포를 선행사와의 근접성과 구문 관계(grammatical role)의 변화를 중심으로 재 관찰하였다. 그 결과, 영목적어의 58.3%는 생략된 목적어의 선행사가 바로 앞 절에 위치하면서 똑같이 목적어의 역할을 한 경우임을 밝혔다. 따라서 본 논문에서는, [4]에서의 영목적어 발생 동기를 반영하고, 다른 영형에서도 인접한 절 사이에서 구문 기능의 변화가 대용형 생성에 영향을 끼칠 수 있다고 판단, 현 발화 Ui에 실현된 구정보들의, 앞선 두 발화 Ui-1, Ui-2 에서의 구문 관계를 고려하였다(자질 1과 자질 2).

본 논문에서는 현 발화에서 언급되는 모든 지시물의 현저성은, Cf-list의 서열을 정의할 때와 마찬가지로 구문 관계에 의해 가장 잘 인식된다고 보고, 지시어를 생성하고자 하는 구정보의 구문 관계를 고려하는데(자질

표 2 U_i의 구정보 e의 지시어 생성을 위해 고려하는 자질들 ($e \in oldA(U_i)$)

	자질들	설명	자질들이 갖는 값
1.	gr(e, U_{i-2})	U_{i-2}에서 e의 구문 관계	주제, 주어, 직접 목적어, 간접 목적어…
2.	gr(e, U_{i-1})	U_{i-1}에서 e의 구문 관계	〃
3.	gr(e, U_i)	U_i에서 e의 구문 관계	〃
4.	modifee(e)	e의 피수식어 여부	true, false
5.	animacy(e)	e의 생물성(animacy) 여부	true, false
6.	pro(e,U_{i-2})	U_{i-2}에서 e의 대용형	절 5의 7가지 대용형
7.	pro(e,U_{i-1})	U_{i-1}에서 e의 대용형	〃
8.	$e \in Cf(U_{i-2})$, trans(U_{i-2})	e를 가지는 U_{i-2}의 전이유형	절 2의 9가지 전이유형
9.	$e \in Cf(U_{i-1})$, trans(U_{i-1})	e를 가지는U_{i-1}의 전이유형	〃
10.	trans(U_i)	U_i의 전이유형	〃
11.	equality(Cb(U_i), Cp(U_{i-1}))	Cb(U_i)와 Cp(U_{i-1})의 일치 여부	같다, 다르다
12.	cost(U_i, U_{i-1})	U_i와 U_{i-1}사이의 추론 비용	싸다, 비싸다

3), Cf-list에서 구문 관계의 서열이 높을수록 현저(prominent)하다고 본다.

일반적으로, 구정보가 수식어(modifier)를 가지면, 즉 구정보가 피수식어(modifee)인 경우, 영형으로 실현되지 않으므로 구정보가 피수식어인지의 여부를 표시하기 위해 자질 4를 고려하였다.

주제 돌출 언어(topic-prominent language)를 대상으로 생물성(animacy)과 영형의 관계를 규명한 기존 연구들(예, [15], [16])은, 대용형을 취한 지시물들 대부분이 생물성을 갖는 것들이라는데 의견의 일치를 보고 있다. 즉, 생물성을 가진 지시물이 그렇지 않은 지시물보다 자주 대용화된다는 사실에 동의한다. [15]에서는 중국어에서 대용형을 생성할 때, 영형 생성 조건에 해당되지 않는 일부 지시물에 대해 그것이 생물성을 가지면 영형이 아닌 대명사에 의한 대용형을 생성하도록 하였다. 한국어도 중국어와 마찬가지로 주제 돌출 언어에 속하기 때문에 지시물의 생물성이 대용형 생성에 미치는 영향을 관찰하고자 자질 5를 고려하였다.

구정보의 지시어 표현은 연속하는 문장에서 동일한 형태로 반복되어 사용되는 경향이 있다. 뉴스 기사에서 발췌한 다음 기사를 살펴보자.

(U_1) 전국경제인연합회는 11월 중 기업경기지수가 최저로 떨어졌다고 밝혔다.
(U_2) 전경련은 기업의 자금 사정 악화가 경기 침체를 주도하고 있는 것으로 분석하였다.
(U_3) 전경련은 기업의 투자심리도 역시 위축될 것으로 내다봤다.

U_1에서 단체명 '전국경제인연합회'는 전체 이름(본래의 NP표현)으로 기술되었고 그 다음 연속하는 두 발화 U_2, U_3에서 '전경련'이라는 축약형으로 동일하게 지시되었다.[5] 이렇듯, NP 변이형 또는 대명사에 의한 대용형의 반복은 뉴스 기사에서 단체명이나 인명 등이 지시될 때 자주 나타난다. 따라서, 본 논문에서는 현 발화에 앞선 두 발화 U_{i-1}, U_{i-2}를 대상으로, 현 발화 U_i에서 지시어 생성의 대상이 되는 구정보들의 지시어 형태를 고려하고(자질 6과 자질 7), 현 발화에서 적합한 대용형을 생성할 때, 연속하는 이전 발화에서 동일한 지시물에 대한 대용형을 참고하는 것이 도움이 될 것이라 가정한다.

4.2.2 중심화이론에 기초한 구정보(Cb와 olcCp)를 위한 자질들

중심화 이론에 기반한 기존 연구들은 중심화 이론에 대한 깊은 고찰 없이 지시어 생성의 대상으로 특정 전이유형(CONTINUE 또는 SMOOTH-SHIFT)에서 Cb만을 다루었다는 점에서 그 한계가 명백하다. 본 논문에서는, Cb와 oldCp의 자연스러운 지시어 생성을 위해, 연속하는 세 개의 발화에 대한 전이유형의 변화와 인접한 두 발화간의 추론 비용을 고려하였다.

한국어에서 대용형, 특히 영형을 해석하는 관련 연구에서(예, [4], [5], [17], [18]) 대부분의 언어학자들은, 영형은 CONTINUE의 Cb로부터 실현된다는 사실에 동의하는 듯 하다. 그러나, 본 논문에서 관심이 있는 것은 이러한 분석의 관점이 아니라, 역으로 '모든 CONTINUE의 Cb는 영형으로 실현되는가?'하는 생성의 관점이다. 본 논문에서 수집된 텍스트를 분석한 결과, 모든 CONTINUE의 Cb중 46%만이 영형으로 실현되었다는

5) 지시어 생성 과정을 바꾸어 쓰기(paraphrasing)의 좁은 범주로 볼 때, 관점에 따라서는 '전국경제인연합회'를 '전경련'으로 대용하는 문제는 지시어 생성(referring expression generation)의 범주를 벗어난다고 볼 수 있지만, 본 논문에서는 영형과 대명사에 의한 대용형 외에, 본래 NP 표현에 대한 다양한 NP변이형의 표현도 지시어 생성 문제로 광범위하게 해석하되, 영형의 생성에 치중하여 다양한 자질들을 규명하였다.

표 3 추론 비용을 고려한 발화간의 전이 유형 [19]

	$Cb(U_n)=Cb(U_{n-1})$	$Cb(U_n)!=Cb(U_{n-1})$
$Cb(U_n)=Cp(U_n)$ and $Cb(U_n)=Cp(U_{n-1})$	*CHEAP-CONTINUE*	*CHEAP-SMOOTH-SHIFT*
$Cb(U_n)=Cp(U_n)$ and $Cb(U_n)!=Cp(U_{n-1})$	*EXP-CONTINUE*	*EXP-SMOOTH-SHIFT*
$Cb(U_n)!=Cp(U_n)$	RETAIN	ROUGH-SHIFT

사실을 확인하였다. 따라서, 기존의 영형의 생성 방식 [10], [11] - 모든 CONTINUE의 Cb는 영형으로 실현하라 은 영형의 과대한 생성(overgeneration)을 유발하고, 이로 인해 텍스트 전체의 애매성을 가중시키는 문제점을 안고 있다.

이러한 기존 문제를 해결하기 위해, 본 논문은 '비용기반 중심화 이론(Cost-based Centering Model)'의 적용을 제안한다. [19]에서는 추론 비용(inference cost)을 고려하여 중심화 이론을 재 고안 하였는데, 본 논문에서는 이를 비용기반 중심화 이론이라 부른다. [19]에서, 텍스트를 구성하는 발화의 흐름을 이해하기 위해서는 추론 비용이 든다고 가정하고, 그 추론의 비용을 '$Cb(U_n)$와 $Cp(U_{n-1})$의 일치' 여부로 계산하였다. $Cb(U_n)$와 $Cp(U_{n-1})$가 일치하면 인접한 두 발화를 이해하는데 드는 추론 비용은 '싸며(cheap)', 그렇지 않으면 '비싸다(expensive)'. 또한 $Cb(U_n)$와 $Cp(U_{n-1})$의 일치 여부에 따라 기존의 네 가지 전이 유형을 여섯 가지로 확대하였다(표 3). $Cb(U_n)$와 $Cp(U_{n-1})$의 일치 여부에 의한 추론 비용의 설정은, SMOOTH-SHIFT 직전에 CONTINUE가 실현되는 것보다 RETAIN이 실현되는 것이 더 자연스럽다는 것을 설명할 수 있다. RETAIN은 중심의 전이를 예측하게 하므로, RETAIN 다음에 RETAIN의 Cp와 SMOOTH-SHIFT에서 Cb가 일치하는 SMOOTH-SHIFT가 실현되는 것은 중심의 흐름이 자연스럽게 느껴지는 반면, CONTINUE 다음의 SMOOTH-SHIFT는 CONTINUE의 Cp와 SMOOTH-SHIFT의 Cb가 같을 수 없으므로 추론 비용이 비싼데, 실제로도 주제의 전이가 전자에 비해 자연스럽지 않다.

그럼, $Cb(U_n)$와 $Cp(U_{n-1})$의 일치 여부를 고려한 비용기반 중심화 이론이 한국어에서 대용형, 특히 영형의 생성에 어떠한 도움을 줄 수 있는지 살펴보자. 생략의 기본적인 원칙을 '회복성'에 있다고 볼 때, 영형이 자연스럽다는 것은 해당 문맥에서 생략된 논항의 회복이 용이하다는 것을 의미한다. 이전 발화에 대해 현 발화의 추론 비용이 싸다는 것은, 현 발화에서 주제의 전이가 없거나, 있다 하더라도 그것이 자연스럽다는 것을 의미하고 이런 상황에서 주제 관련 논항의 생략은 그렇지 않은 상황에서의 생략보다 훨씬 회복이 용이하다고 볼 수 있다. 따라서, 추론 비용이 싼 발화는 그렇지 않은 발화에 비해 주제 관련 논항이 생략되었을 경우 회복 가능성이 높다는 점에서, 추론 비용이 싼 발화의 주제 관련 논항은 그렇지 않은 발화에서의 주제 관련 논항보다 더 생략적이라 볼 수 있다. 구체적으로, 비용기반 중심화 이론의 관점에서 CHEAP-CONTINUE와 CHEAP-SMOOTH-SHIFT의 각각의 Cb가 EXP-CONTINUE와 EXP-SMOOTH-SHIFT의 각각의 Cb보다 더 생략적일 것임을 의미한다. 다음 텍스트는 본 논문에서 실험 대상으로 삼은 텍스트 중 우리나라 전통 농기구 '자귀'를 묘사한 텍스트의 일부를 발췌한 것이다.

(U_1) 자귀는 우리나라의 전통적인 농기구이다.
→ Cp: 자귀, Cb: undefined
(U_2) 자귀는 도끼나 톱으로 잘라낸 원목을 가공하는 데 사용한다.
→ Cp: 자귀 Cb: 자귀, CHEAP-CONTINUE
(U_3) 자귀의 **날**은 절삭날이라 불린다.
→ Cp: 날, Cb: 자귀, RETAIN
(U_4) 자귀는 도끼와 비슷하게 생겼다.
→ Cp: 자귀, Cb: 자귀, EXP-CONTINUE
(U_5) (자귀는, 영형) 그 크기에 따라 대자귀, 중자귀, 소자귀로 나뉜다.
→ Cp: 자귀 Cb: 자귀, CHEAP-CONTINUE

U_2로부터 U_3이 실현되면서 화제의 중심이 U_2의 '자귀'에서 '날'로 바뀌어 RETAIN이 실현되었다. 이는, U_3에서 청자(독자)가 다음 발화의 화제의 중심은 '날'이 될 것이라고 예측하게 하며, 청자의 의식 상태에서 그간의 발화 주제였던 '자귀'의 현저성은 감소된다. 하지만 U_4에서 화제의 중심이 '날'에서 '자귀'로 복귀되며 EXP-CONTINUE가 실현되고, U_5에서 '자귀'는 화제의 중심으로 유지되어 CONTINUE가 실현된다.

위의 텍스트에서 U_4, U_5의 화제의 중심 Cb는 '자귀'로 동일하지만 U_4의 '자귀'가 U_5의 '자귀'보다 발화의 흐름 관점에서 덜 주제적이다. 또, U_4에서 '자귀'가 영형으로 실현되었다면 '비슷하게 생겼다'의 주제는 '자귀'가 아닌 '날'로 잘못 이해될 될 수 있다. 즉, CHEAP-CONTINUE의 Cb는 EXP-CONTINUE의 Cb에 비해 주제로서 더욱 현저하고 잉여적이며, 생략되었을 경우 회복이 용이하여 영형으로 실현될 가능성이 높고, 반대로 EXP-CONTINUE에서Cb가 영형으로 실현될 경우

표 4 학습 데이타의 통계 자료

장르	묘사문			뉴스			이야기			총계		
지시어 생성 대상	Cb	oldCp	oldR	Cb	oldCp	oldR	Cb	oldCp	oldR	Cb	oldCp	oldR
NP와 NP 변이형 개수	358	81	338	119	77	85	234	60	182	711	218	605
영형이 아닌 대용형의 개수	14	8	6	16	15	7	58	14	26	88	37	39
영형의 개수	189	40	47	59	43	10	110	35	32	358	118	89
총계	561	129	391	194	135	102	402	109	240	1157	373	733
텍스트 개수		40			35			20			95	

곡해(mis-interpretation)의 위험이 있음을 알 수 있다. 따라서, EXP-CONTINUE와 CHEAP-CONTINUE를 구별하는 $Cb(U_n)$와 $Cp(U_{n-1})$의 일치 여부(자질 11)는 영형의 실현에 영향을 미치는 중요한 자질이라 볼 수 있다.

중심화 이론을 연구하는 대부분의 학자들은, 대용 해결과 텍스트의 결속성을 규명하는데 있어 특정 전이에 대한 기존의 선호도보다는 인접한 발화간의 전이쌍을 고려하는 것이 훨씬 효과적이라고 본다(예, [2], [9], [10], [19]). 특히, [19]에서는 추론 비용에 기반하여 모든 전이쌍을, 선호되는 것과 그렇지 않은 것으로 분류하였다. 따라서, 본 논문에서는 '선호되는 전이쌍에서 실현된 발화가 선호되지 않는 전이쌍에서 실현된 발화에 비해, 중심화 관련 논항의 생략 가능성이 높다'고 보고 인접한 두 발화 사이의 추론 비용을 고려하였다(자질 12). 또한, 본 논문에서는 [19]에서 확장된 여섯 가지 전이유형 외에 두 개의 전이(associate-shift, resume)를 추가적으로 고려하였으므로, 알려지지 않은 전이쌍에 대한 추론 비용을 인식하고자 현 발화의 전이유형(자질 10)외에 인접한 이전 두 개 발화의 전이유형(자질 8, 자질 9)을 고려하였다.

5. 지시어 생성 모델의 구현

본 논문에서는 생성할 대용형 종류를 2가지(명사 표현 그대로 NP, 영형- ZERO), 3가지(명사표현 그대로, 영형이 아닌 대명사 - PRO, 영형), 7가지(명사 표현 그대로, 축약형, 상위어, 동의어, 지시관형사 동반 명사구, 영형이 아닌 대명사, 영형)로 복잡화 하면서 각각을 위해 코퍼스를 분석, 대용형 생성 모델을 구축하였다.[6]

먼저, 학습을 위해 세 개의 장르 - 묘사문, 뉴스 기사, 이야기(짧은 이솝 우화) - 로부터 총 95개의 텍스트를 수집한 후, 수작업으로 모든 문장을 발화 단위인 시제절 기준으로 분할하고, 영형 및 영형 외의 대명사에 대한 선행사(antecedent)를 찾는 대용 해결(reference resolution)을 처리하였다. 다음, 대용 해결된 각각의 일련의 시제절에 대해 포항 공대 지식 및 언어 공학 연구실의 구문 분석기를 통해 각 단어들의 구문 관계를 획득하고, 이 중 체언을 대상으로 단락 2에서 정의한 rank에 따라 *cf-list*를 구성한다. 다음, *cf-list*를 가지는 인접한 시제절들을 대상으로 비용기반 중심화 이론을 적용하여, 각 발화에 대해 Cb, Cp, 전이 유형을 자동으로 구하고, 중심화 이론을 제외한 나머지 자질들을 반자동으로 기술해 학습 데이타를 완성하였다. 학습 데이타의 여러 가지 통계치는 표 4에 정리되어 있다. Cb와 oldCp에서 영형 발생 빈도가 31~32%로 거의 비슷하다는 사실에서, Cb가 아닌 구정보 Cp에서도 영형이 빈번히 실현됨을 알 수 있고, oldR에서는 12% 정도로 Cb와Cp에 비해서 빈번하지는 않지만 역시 영형이 실현된다는 사실을 알 수 있다.

5.1 다양한 기계 학습의 적용에 의한 지시어 생성 모델

지시어 생성 관련 대부분의 기존 논문들은 규칙 기반의 생성 모델을 제안한 반면, 본 논문에서는 WEKA 3.0[7]에서 지원하는 다음 네 가지 학습 모델과 그것들의 변이형들을 적용하여, 지시어 생성에 있어 주어진 자질에 가장 적합한 기계 학습 방법을 찾는다.

▣ 지시어 생성에 적용된 기계 학습

- C4.5를 이용한 결정트리 (DT)
- 개체중심학습 (Instance-based Learning): K*
- 베이지안 분류기 (Nave-Bayes classifier, NB)
- 메타학습: 스택킹(Stacking), 부스팅(Boosting) AdaBoosting(AB), LogitBoosting(LB)

각 기계 학습에 대한 성능 평가는, 10-분할 교차 검증(10-fold Cross-Validation)을 총 10회 시행한 평균값으로 이루어 졌다. 표 5는 학습 데이타의 지시어 생성

6) 본 논문은, 본 논문에서 추출한 여러 가지 자질들로부터 다양한 기계 학습을 적용하여 만들어진 지시어 생성 모델이, 세 개의 도메인에서 수집된 텍스트에서의 지시어 표현을 얼마만큼 정확하게 재현할 수 있는가를 검증하는 것까지만 연구 범위로 한정하고 있다. 본 논문에서 구축된 지시어 생성 모듈이 텍스트 생성 시스템에 적용 되었을 때, 실제 생성된 지시어가 얼마나 자연스러운지를 판단하는 것은 또 다른 중요한 이슈로, 이는 향후 과제로 남겨 둔다.

7) WEKA는 뉴질랜드의 Waikato 대학에서 java로 개발된 기계 학습 소프트웨어로, 자연언어처리의 다양한 분야에 편리하게 적용할 수 있어, 기계 학습을 이용한 많은 논문에서 활용되고 있다(http://www.cs.waikato.ac.nz/~ml/weka/index.html).

표 5 각 기계 학습의 적용에 따른 지시어 생성 대상별, 대용형 종류별, 텍스트 장르별 평균 성능

대상	Cb									oldCp									oldR								
종류	2			3			7			2			3			7			2			3			7		
장르	D	N	S	D	N	S	D	N	S	D	N	S	D	N	S	D	N	S	D	N	S	D	N	S	D	N	S
NB	78	91	70	77	81	60	75	70	**51**	74	94	84	72	83	44	77	75	44	81	91	73	80	83	**68**	77	77	**59**
DT	83	93	73	84	89	59	80	76	45	81	96	88	81	88	**68**	70	79	**60**	86	93	77	84	90	62	83	83	57
K*	83	89	67	85	80	55	81	71	42	88	88	72	**88**	81	44	**87**	71	40	87	93	77	**86**	89	62	84	83	58
LB+DS	82	**94**	**76**	82	92	**61**	79	80	49	83	95	**92**	83	93	**68**	85	82	36	85	93	**80**	84	88	66	81	82	**59**
AB+DT	**87**	**94**	75	**87**	**93**	60	**85**	**83**	49	**92**	**98**	90	87	**95**	66	**87**	**88**	**60**	**88**	**94**	79	**86**	**92**	**68**	**85**	**84**	58
Stacking	83	93	72	84	89	**61**	80	76	47	81	95	76	79	**95**	60	62	76	44	85	93	74	83	88	60	83	81	53

D: 묘사문, N: 뉴스, S: 이야기, DS(decision stump): 1-level binary tree (단위: %)

대상별, 대용형 종류별, 텍스트 장르별로 적용된 여섯 가지의 기계 학습 성능을 보여준다.[8)]

지시어 생성의 대상 측면에서 성능을 살펴보면, 평균적으로 oldCp의 성능이 Cb보다 높고, 대부분의 학습 모델에서 oldCp의 대용형이 Cb의 대용형보다 잘 예측 되었음을 알 수 있다. 이 결과에 대한 두 가지 가능한 해석은, 제안된 여러 자질들이 Cb보다 oldCp의 대용형을 파악하는데 더 효과적일 수 있다는 것과, oldCp가 Cb보다 대용형을 선택하는데 있어 본래 보다 뚜렷한 경향을 가질 수 있다는 것이다. 하지만, oldCp와 Cb 둘 다 중심화 이론에 기반한 논항이고, 고려되는 자질 역시 중심화 이론을 벗어나지 않으므로, 자질의 적용 측면 보다는 oldCp가 Cb보다 대용형을 선택하는데 있어 보다 뚜렷한 경향을 가진다고 해석하는 것이 타당하다. 이를 대용화에 관한 현저성(salience)과 응집성(cohesion)의 관점에서 재해석해 보자. oldCp는 현 발화의 구정보 중 가장 현저한 지시물이고, Cb는 구정보 중 가장 화제의 중심과 가까운, 화제에 응집적인 지시물이라는 관점에서 oldCp의 대용형이 Cb의 대용형보다 결정적(deterministic)이라는 것은, 현저성과 구정보성을 동시에 가진 지시물이 응집성을 가진 지시물보다 대용화에 대해 뚜렷한 경향을 가진다고 볼 수 있다. 그러나 이것이 oldCp가 Cb에 비해 더 자주 영형 또는 영형이 아닌 대명사에 의해 대용화된다는 사실을 의미하지는 않는다.

oldR과 다른 정보간의 성능 차이는 표 5에서 명확히 드러나지 않는다.

텍스트 장르의 관점에서 성능을 살펴보면, 이야기 장르가 나머지 두 장르에 비해 성능이 낮음을 알 수 있다. '이야기'는 다른 장르에 비해 대화체와 문맥 전환(context switch)이 자주 발생하는 등 반-문어체(quasi-written)의 성격을 띄고, 대명사들이 지시하는 선행사가 인접한 발화에 존재하지 않는 경우가 흔히 발생한다. 그 결과, 이야기 장르에 중심화 이론이 잘 적용되지 못하고, 중심화 이론에 기반한 여러 자질들 역시 대용형을 판단하는데 있어 효과적으로 적용되지 못했다고 볼 수 있다.

대용형 생성 종류가 많을수록 대용형을정확하게 예측하는 것이 어려운데, 특히 네 가지의 NP 변이형을 포함한 일곱 종류의 대용형 생성에서, 다양한 NP 변이형 가운데 단 하나의 NP 변이형을 찾는 것은 본 논문에서 제안된 자질 외에 더 많은 지식이 요구되는 문제이므로 그 성능이 만족스럽지 않다고 볼 수 있다.

학습 모델의 측면에서, 결정트리를 이용한Ada-Boosting이 가장 높은 성능을 보였다. 여러 개의 약한 선택 기준(weak learner)들을 결합하여 점차적으로 강한 선택 기준(strong learner)을 만들어 내는 부스팅 기법은, 매번 반복되는 학습 단계에서 새로운 모델이 이전 모델에서 분류하지 못했던 데이타를 처리할 수 있도록 학습되므로, 메타 학습은 일반적으로 단일 기계 학습(stand-alone classifier)보다 성능이 좋다.

표 6은 각 장르에서 지시어 생성 대상(Cb, oldCp, oldR) 별로 각 대용형의 종류(NP-명사 표현 그대로, PRO-영형이 아닌 대명사, ZERO-영형)에 대한 예측 성능을 보여 준다. 평균적으로, 모든 장르에서 NP 대용형을 예측할 때, 즉 영형이나 영형이 아닌 대용형으로 대용화 되지 않고 본래의 명사 표현 그대로 지시어를 생성할 경우, oldCp가 Cb보다 정확률이 높은 반면, ZERO 대용에서는 Cb가 oldCp보다 정확률이 높다. NP 대용형에 있어 oldCp가 Cb보다 정확률이 높은 것은 oldCp가 NP 그대로 표현되는 것에, 보다 뚜렷한 경향을 가지는 것이라 볼 수 있고, Cb가 영형 대용에 있어 정확률이 높은 것은 Cb가 영형으로 대용되는 것에, 보다 뚜렷한 경향을 가진다고 볼 수 있다. 이를 현저성과 응집성의 관점에서 재해석하면, 응집성이 현저성에 비해

8) 본 논문에서는 표 5에서 보여지는 여섯 가지의 기계 학습 외에, SVM, K- NN, Neural Network등을 적용해 보았으나, 표시된 여섯 가지의 기계 학습에 비해 한번도 최고의 성능을 내지 못했으므로 표 5에 기록하지 않았다.

표 6 대용형 종류별과 텍스트 장르별에서 평균 성능

		묘사문			뉴스			이야기		
		Cb		oldCp	Cb		oldCp	Cb		oldCp
2종류	NP	87	<	95	98	<	99	79	<	89
	ZERO	87	>	60	88	>	62	50	>	37
3종류	NP	87	<	97	98	<	99	67	<	73
	PRO	83	<	100	62	<	82	44	>	15
	ZERO	86	>	40	88	>	80	42	>	33

(단위: %)

영형의 생성에 더 많은 영향을 끼친다고 볼 수 있다. 이는, CONTINUE, RETAIN이 응집성이 유지되는 전이 유형이라는 점에서, 'Cb의 생략은 CONTINUE, RETAIN의 순으로 많이 발생한다'는 [5]의 실험결과를 같이 설명할 수 있다.

이야기 장르를 제외한 나머지 장르에서 NP 대용형과 마찬가지로 PRO 대용형에서도 oldCp가 Cb보다 성능이 높은 이유는, 영형이 NP로 실현되는 것보다 영형이 아닌 대명사가 NP로 실현되는 것이 더 자연스러울 수 있기 때문이다.

5.2 대용형 생성에 있어 각 자질의 특징 분석

본 논문에서는 한국어의 대용형, 특히 영형을 분석하고, 자연스러운 대용형을 생성하기 위한 총 12개의 자질을 제안했는데, 실제 이 각각의 자질들이 대용화에 미치는 영향은 다르다. 본 논문에서는 각 지시어 생성 대상별 핵심 자질(central feature)과 장르에 독립적인 핵심 자질을 규명하고자 총 4가지의 탐색 기법 - 최상우선 탐색(best-first search), 유전자 탐색(genetic search), 임의 탐색(random search), 전역 탐색(exhaustive search) - 에 의한 상관관계 분석 기반의 자질 선택론[20](Correlation-based feature selection)을 적용하였다.

자질 2, 3, 4, 10, 11은 장르에 독립적인 핵심 자질로 선택되었다. 자질 11, Cb(U_n)와 Cp(U_{n-1})의 일치 여부는 영형과 나머지 대용형을 구별하는데 중요한 역할을 하는 것으로 분석 되었는데, 특히, 묘사문에서 CONTINUE의 Cb중 61%가 영형으로 실현되었고, 61%중 95%가 CHEAP-CONTINUE의 Cb로부터 실현된 반면, EXP-CONTINUE의 Cb로부터 영형이 실현된 것은 61%중 5%에 불과했다. 즉, CONTINUE의 Cb에서 실현된 영형의 거의 대부분은 사실, EXP-CONTINUE가 아닌 CHEAP-CONTINUE의 Cb였다는 사실을 알 수 있었다. 이는, 비용기반 중심화 이론이 한국어의 영형 분석 및 생성에 주효하며, 본 논문의 초기 가정 CHEAP CONTINUE의 Cb가 EXP-CONTINUE의 Cb보다 훨씬 자주 영형으로 실행된다 이 옳았음을 보여준다. 이로부터, R2-Kibble과 R3-Yamura의 지시어 생성 규칙에서 'CONTINUE의 Cb'를 'CHEAP-CONTINUE의 Cb'로 조건을 바꾸는 것만으로도 CONTINUE에서 Cb를 영형으로 실현할 때 발생하는 과다생성 문제를 해결할 수 있을 것이다.

자질 12, U_i와 U_{i-1} 사이의 추론 비용과 관련해 학습 데이타의 모든 장르를 분석한 결과, 선호되는 전이쌍에 속한 발화의 논항 중 52%가 영형으로 실현된 반면, 선호되지 않는 전이쌍에 속한 발화의 논항 중 영형으로 실현된 것은 21%에 그쳤다. 따라서, 비록 자질 12가 장르에 독립적인 핵심 자질로 선택되지는 않았지만, 영형과 나머지 대용형을 구별하는데 중요한 역할을 한다고 볼 수 있다.

U_{i-1}, U_{i-2}에서 대용형의 형태(자질 6, 7)를 고려하는 것은 뉴스 기사에 한해 핵심 자질로 선택 되었고, U_{i-1}, U_{i-2}에서 구정보의 구문 관계(자질 1, 2)와 알려지지 않은 전이 유형간의 추론 비용을 알기 위해 고려되었던 U_{i-1}의 전이 유형(자질 9)도 유효함을 알 수 있다. 반면, U_{i-2}의 전이유형(자질 8)은 대용형의 생성에 크게 영향을 끼치지 않음을 관찰했다.

지시어 생성 대상별 핵심 자질에서 자질 2, 3, 4는 세 가지 대상에 공통적인 핵심 자질로 선택되었으며, 중심화 이론과 관련 있는 Cb, oldCp에 대해서는 추가적으로 자질 10이 핵심 자질로 선택되었다. 특히, Cb에 있어 자질 11이 선택 된 것은, 앞에서 논의된 바와 같이 Cb의 영형 생성 여부가 이전 발화의 Cp와의 일치 여부에 의해 크게 좌우되기 때문이다.

생물성(animacy)과 관련해, 모든 장르에서 생물성을 가진 구정보의 29%가 대용화된 반면, 생물성을 갖지 않은 구정보 중 오직 1%만 대용화 되었다. 특히 이야기 장르에서 사람, 동물과 같은 생물성을 갖는 지시물들이 자주 등장하는데, 대용화 된 것 중 71%가 생물성을 가진다는 사실을 관찰할 수 있었다. 따라서, 대부분의 대용형들이 지시하는 선행사들은 생물성을 갖는 지시물이라는 기존의 주장은 대체로 의미가 있다고 볼 수 있다.

5.3 기존 방법론과의 성능 비교

표 7 텍스트 장르별, 지시어 생성 대상별 핵심 자질들

장르	장르-독립적	장르-의존적
묘사문	2,3,4,10,11	9,12
뉴스		6,7,9
이야기		1

대상	대상-독립적	대상-의존적
Cb	2,3,4,10	11,1,7
oldCp		9
oldR	2,3,4	7

표 8 기존 방법론과의 성능 비교

장르	묘사문				뉴스				이야기			
지시어 생성 대상	Cb		oldA		Cb		oldA		Cb		oldA	
대용형 종류	2	3	2	3	2	3	2	3	2	3	2	3
Yeh (1997) [15]				**77**				**68**				**68**
Kibble (2000) [10]	62	**66**			55	57			65	**65**		
Yamura-Takei (2001) [11]	62	62			59	56			59	53		
Prasad (2003) [13]	**67**	**66**			**62**	**59**			**70**	59		
본 논문의 생성 모델	**87**	**87**	**88**	**86**	**94**	**93**	**94**	**93**	**76**	**61**	**80**	**67**

(단위: %)

본 논문에서 제안한 자질들로 학습된 지시어 생성 모델과 기존의 방법론들의 성능을 비교한 결과가 표 8에 정리되어 있다. 기존 방법론들은 본 논문의 95개의 텍스트에 기술된 자질들을 이용해 직접 구현 가능하였으며, 각 방법론을 95개의 텍스트에 적용하여 지시어를 예측한 결과에 대한 정확률을 표기하였다.[9] [15]에서는 지시어 생성의 대상으로 모든 구정보들을 고려하였고, 나머지 방법론들은 오직 Cb만을 그 대상으로 삼았다. 이야기 장르의 대용형 종류 3을 제외한 나머지 부분에서, 본 논문에서 구축한 모델이 기존의 방법론보다 성능이 높음을 알 수 있으며, T-test를 통해 99.9%의 신뢰 구간에서 본 논문의 모델이 이룬 성능 향상이 통계적으로 의미 있음을 확인하였다. 특히, 일본어는 언어 특징상 영형의 발생이 한국어와 비슷하다는 점에서, 본 논문의 모델이 일본어 대상의 [11]과 비교해 성능 향상을 이루었다는 것은 의미가 있다.[10]

6. 결 론

본 논문에서는 자연스러운 대용형 생성을 위해 구문 정보와 비용기반 중심화 이론으로부터 다양한 자질들을 규명하고, 여러 가지 기계 학습을 통해 지시어 생성 모델을 구축하였다. 특히, 지시어 생성 대상으로 모든 구정보를 고려하되, 각각의 위상에 따라 Cb, oldCp, oldR로 분류하고 이들 각각이 대용형에 대해 서로 다른 경향을 보임을 확인하였으며, 각각에 대한 독립적인 생성 모델을 구축하였다. 또한, 제안된 자질이 한국어에서 영형의 발생을 설명하는데 효과적으로 적용될 수 있음을 보였으며, 장르에 독립적인 핵심 자질들도 규명하였다. 본 논문에서 제안한 대용형 생성 모델의 객관적인 성능 판단을 위해, 영어/일본어/중국어/힌두어 대상의 기존 방법론과 성능을 비교하여 성능 향상이 있었음을 확인하였다. 특히, 일본어와 중국어는 언어 특징상 영형의 발생이 한국어와 비슷하다는 점에서 이러한 성능 비교는 의미가 있다고 볼 수 있다.

일반적으로 텍스트 생성의 각 프로세스에 대한 객관적인 성능 평가는 어려운 문제로 알려져 있다. 지시어 생성 역시 마찬가지로, 지시어 선택의 문제는 정답/오답의 문제가 아니라 자연스러움의 문제라는 점에서, 본 논문에서 모델이 생성한 대용형과 텍스트에서 쓰인 대용형을 비교하는 현재의 평가 방법은 다소 엄격한 측면이 있다. 모델에서 생성한 대용형이 텍스트에서 쓰인 대용형과 다르다 하더라도 실제 자연스러울 수 있기 때문이다. 향후, 여러 명의 피실험자(subject)들이 모델에서 제시한 대용형의 자연스러움을 판단하는 추가 실험을 수행할 필요가 있다. 또한, 본 논문에서 제안된 자질들을 한국어와 비슷한 언어 유형을 갖는 중국어, 일본어에 적용하여 세 언어에 독립적인 자질들을 분석해 보고, 제안된 모델이 실제 텍스트 생성 시스템에 효과적으로 적용될 수 있음을 보일 것이다.

9) 본 논문에서 제안한 지시어 생성 모델이 실제 텍스트 생성 시스템에 적용되었을 때, 다른 방법론에 비해서 얼마만큼 정보전달의 효율성, 가독성, 텍스트의 결속성을 향상시켰는지에 대한 객관적인 평가는 어렵다. 따라서, 본 논문에서는 각 방법론이 실제 텍스트의 지시어 유형을 얼마나 잘 예측할 수 있느냐의 여부로 그 성능을 평가하고자 하며, 이런 방법이 성능 비교의 객관적인 지표가 될 수 있을 것이라 생각한다.

10) 본 논문에서는 영어, 중국어, 힌두어, 일본어를 대상으로 중심화 이론 등에 기반하여 자연스러운 지시어를 생성하기 위한 기존의 네 가지 방법론을 구현한 다음, 한국어 텍스트에 적용해 본 결과, 실제 한국어에서 지시어를 생성을 하는데 있어 본 논문에서 제안한 자질들로 구축된 생성 모델이 기존의 다른 언어에 대해 제안되어 온 방법론보다 훨씬 효과적임을 보였다. 특히, 일본어는 언어 특질이 한국어와 아주 유사할 뿐만 아니라, 한국어와 지시어의 쓰임 또한 매우 유사하다고 볼 수 있는데, [11]에서는 일본어는 한국어와 마찬가지로 문맥(context)에 의존적인 언어로 문맥상에서 회복 가능한(recoverable)한 요소들은 모두 삭제(영형)가능하다고 밝히고 있다. 이는 영-한/한-영 또는 일-영/영-일 번역에서와는 달리 일-한/한-일 양 언어간의 번역에 있어 대용 표현에 대한 별도의 처리 없이 직접 번역 가능하다는 점에서도 간접적으로 알 수 있다. 따라서, 일본어를 대상으로 제안된 지시어 생성 알고리즘과 본 논문에서 제안한 지시어 생성 모델을 동일한 한국어 텍스트에 적용하여 얻은 정확률을 비교하는 것은 의미가 있다고 볼 수 있다.

참 고 문 헌

[1] Roh, J.E., Kang, S.J. and Lee, J.H., "Korean Text Generation from Database for Home shopping Sites," NLPRS, Tokyo, Japan, pp. 419-426, 2001.

[2] Grosz, B.J., Joshi, A.K. and Weinstein, S., "Cen-

tering: A Framework for Modeling the Local Coherence of Discourse," Computational Linguistics 21(2), pp. 203-225, 1995.
[3] Poesio, M., Stevenson, R., Eugenio, B. D., Hitzeman, J., and Cheng, H., MS, "Centering: A Parametric Theory and its Instantiations," to appear in Computational Linguistics, 2004.
[4] 김미경, "중심화 이론에서 본 한국어 논항의 생략현상", 언어, 28권, 제1호, pp. 29-49, 2003.
[5] 류병률, "한국어 담화상의 중심화와 영형 조응 현상", 서울 대학교 언어학과 석사 학위논문, 2001.
[6] Roh, J.E. and Lee, J.H., "Coherent Text Generation using Entity-based Coherence Measures," ICCPOL, Shen-Yang, China, pp. 243-249, 2003.
[7] Cheng, H., "Experimenting with the Interaction between Aggregation and Text Planning," Proceedings of ANLP-NAACL, USA, 2000.
[8] Mittal, V., Moore, J., Carenini, G., and Roth, S., "Describing Complex Charts in Natural Language: A Caption Generation System," Computational Linguistics, 1998.
[9] Kibble, R. and Power, R., "Using centering theory to plan coherent texts," In Proceedings of the 12th Amsterdam Colloquium., 1999.
[10] Kibble, R. and Power, R., "An integrated framework for text planning and pronominalization," INLG, Mitzpe Ramon, Israel, pp. 77-84, 2000.
[11] Yamura-Takei , M., Fujiwara M., and Aizawa, T., "Centering as an Anaphora Generation Algorithm: A Language Learning Aid Perspective," NLPRS, Tokyo, Japan, pp. 557-562, 2001.
[12] McKeown, K.R., "Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text," Cambridge, U.K.: Cambridge University Press, 1985.
[13] Prasad, R., "Constraints on the generation of referring expressions, with special Reference to Hindi", U of Pennsylvania, PhD Thesis, 2003.
[14] Hashimoto Sachie, "Anaphoric Expression Selection in the Generation of Japanese," Information Processing Society of Japan, No.143, 2001.
[15] Yeh, Ching-Long, Mellish, Chris, "An Empirical Study on the Generation of Anaphora in Chinese," Computational Linguistics, 23-1, pp. 169-190, 1997.
[16] Artstein, R., "Animacy and null subjects," Proceedings of Console VII, pp. 1-15, 1999.
[17] 김미영, "한국어 담화의 중심화", 서울 대학교 언어학과 석사 학위 논문, 1994.
[18] 김미경, "정보구조화 관점에서 본 생략의 의미와 조건", 담화와 인지, 제6권, 2호, pp. 61-88, 1999.
[19] Strube, M. and Hahn, U., "Functional Centering: Grounding Referential Coherence in Information Structure," Computational Linguistics 25(3), pp. 309-344, 1999.
[20] Hall, M. A., "Correlation-based Feature Subset Selection for Machine Learning," PhD Thesis at the University of Waikato, 1998.

노 지 은
2000년 2월 부산대학교 컴퓨터공학과 학사. 2000년 3월~포항공과대학교 컴퓨터공학과 석박사통합과정. 관심분야는 텍스트 생성, 기계 번역

이 종 혁
정보과학회논문지 : 소프트웨어 및 응용
제 31 권 제 4 호 참조