중심화 이론을 이용한 텍스트 구조화
(Text Structuring using Centering Theory)

노 지 은 † 나 승 훈 † 이 종 혁 ††
(Ji-Eun Roh) (Seung-Hoon Na) (Jong-Hyeok Lee)

요 약 본 논문에서는 자연스러운 텍스트 생성을 위한 여러 과정 중, 문장 순서를 결정하기 위한 텍스트 구조화(text structuring)에 관한 것으로, 중심화 이론(centering theory)에 기반하여 문장 순서의 자연스러움을 판단할 수 있는 다양한 평가 척도를 논의한다. 먼저, 기존 연구들에서 중심화 이론에 기반한 문장 순서의 평가 척도들 중 가장 효과적이라고 알려진 MIN.NOCB를 텍스트 구조화에 적용할 때 발생할 수 있는 문제점을 지적하고, 대안이 될 수 있는 새로운 평가 척도인 MAX.CPS를 제안한다. 또, 임의의 평가 척도가 주어진 문장들에 대해 가질 수 있는 기대치를 먼저 예측하고, 그것에 따라 다른 평가 척도를 적용하게 하는 프레임워크를 제안하여, 중심화 이론 안에서 최상의 문장 순서를 찾기 위한 새로운 방법론을 모색한다. 또한, 중심화 이론의 적용에 있어 핵심이라 할 수 있는, 명사들의 돋보임성(salience)을 서열화(cf-ranking) 하는 다양한 방식을 중심화 기반 문장 순서 평가 척도의 관점에서 분석하였다. 그 결과, 텍스트 구조화에 관한 한, 단순히 문장에서 실현된 순서에 따라 명사들의 돋보임성의 서열을 정하는 것이 한국어의 특성상 가장 간단하면서도 효율적임을 입증하였다.

키워드 : 텍스트 구조화, 문장 순서, 중심화 이론, 문장 순서의 평가 척도, 명사의 돋보임성, 서열화, MIN.NOCB, MAX.CPS

Abstract This paper investigates Centering-based metrics to evaluate ordering of utterances for text structuring. We point out a problem of MIN.NOCB metric which has been regarded as the simplest and best measure to evaluate coherence of ordering within Centering framework, and propose a new Centering-based metric, MAX.CPS as an alternative or supplementary one. This paper introduces a framework which pre-estimates the effectiveness of a metric on a given input ordering, and selects an applicable metric according to the pre-estimation result. Using this framework, we propose a new policy which can generate more optimal ordering within Centering framework. Moreover, we evaluate several kinds of Cf-ranking methods in terms of Centering-based metrics, and find that simply ranking entities by their linear order is generally the most suitable because of characteristics in Korean.

Key words : text structuring, ordering, Centering Theory, Centering-based metrics, Cf-ranking, MIN.NOCB, MAX.CPS

1. 서 론

텍스트 생성은 자연어로 이루어지지 않은 기저의 정보들을 자연어로 사상하여 텍스트를 생성해 내는 언어처리의 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.

이런 텍스트 생성의 여러 과정 중 본 논문은, 텍스트의 국소적 응집성(local coherence)에 큰 영향을 미치는 자연스러운 문장[1] 순서의 결정에 관한 것으로, 이를 일반적으로 텍스트 구조화(text structuring)과정이라고 일컫는다. 그간 텍스트의 국소적 응집성을 모델링 하기 위한 많은 연구가 이루어져 왔는데, 그 중심에 중심화 이론(centering theory)[1]이 있다고 볼 수 있다. 중심화 이론은 발화를 구성하는 명사들로 인접한 발화 사이의

† 학생회원 : 포항공과대학교 컴퓨터공학과
jeroh@postech.ac.kr
nsh1979@postech.ac.kr
†† 종신회원 : 포항공과대학교 컴퓨터공학과
jhlee@postech.ac.kr
논문접수 : 2005년 8월 17일
심사완료 : 2007년 4월 11일

1) 본 논문에서 '문장'은 중심화 이론이 적용되는 단위인 '발화'와 같은 의미이며, 발화는 통상 '시제절(tensed clause)'로 정의된다[8-11].

전이 유형을 구하여 텍스트의 응집성을 모델링한 이론으로, 담화를 객체 기반 표현(entity-based representation)으로 분석할 수 있는 출발점을 제공하였다. 초기에는 대용 해석을 위해 주로 활용되었지만, 최근에는 텍스트 생성의 여러 과정 - 텍스트 구조화[2-5], 문장 계획(sentence planning)[6], 지시어 생성(referring expression generation)[4,7] - 에 널리 적용되고 있다. 자연스러운 문장 순서를 결정하기 위해 많은 기존 연구들 역시, 중심화 이론에 기반해 문장 순서에 따른 응집성을 평가할 수 있는 척도를 제안해 왔는데[2-5], 본 논문은 기존의 중심화 이론에 기반한 문장 순서 평가 척도에 대한 2가지의 문제점에서 출발한다. 최근 [2,3]의 연구에서 기존의 중심화 기반 문장 순서 평가 척도 4가지 (*MIN.NOCB, MIN.V_CHP, MIN.KP, MIN.MIL*)를 코퍼스 분석을 통해 비교 평가한 결과, MIN.NOCB가 가장 단순하면서도 가장 강력한 척도라고 주장하였다. 향후 MIN.NOCB에 대해서 자세히 설명하겠지만, 간단히 설명하면 이는, 서로 인접한 문장 사이에 명사를 적어도 하나 이상 공유하지 않는 발화의 수를 최소화 하는 문장 순서가 가장 좋은 순서임을 의미한다. 다시 말하면, 인접한 발화 사이에 적어도 하나 이상의 명사를 공유하게 하는 문장 순서를 찾는 것이 다른 복잡한 척도나 그것들의 결합보다 중심화 이론으로 할 수 있는 가장 최선의 문장 순서 결정 방법이라는 것이다. 그러나, 그들은 실제 MIN.NOCB를 적용하는데 발생할 수 있는 중요한 문제를 하나 간과하였다. 그것은 모든 문장이 어떻게 인접하든지 간에 그것의 순서와 상관없이 적어도 하나의 명사 이상을 공유하게 되는 경우이다. 이때는 MIN.NOCB는 임의의 문장 순서로부터 이상적인 문장 순서를 찾아 낼 수 있는 척도의 기능을 상실하게 된다. 이것이 본 논문에서 제기하는 기존 연구에 대한 첫 번째 문제 제기로, 본 논문에서는 이를 해결하기 위한 방법론을 제시할 것이다.

중심화 이론에서는 하나의 문장에 실현되어 있는 명사들은 그것들이 갖는 돋보임성의 정도가 상대적으로 다르며, 돋보임성의 정도에 따라 명사들을 서열화 할 수 있다고 가정한다. 따라서, 중심화 이론을 적용할 때 명사의 서열화를 어떤 기준으로 정할 것인지가 핵심이라 할 수 있다. 서열화의 기준은 일반적으로 언어마다 다른데[12], 구문 관계(grammatical role)[13-18], 의미 관계(thematic role)[19], 단어 순서(linear order)[20], 정보 위상(information status)[21] 등이 논의 되어 왔으며, 그 중 구문 관계에 따른 서열화가 가장 보편적이다. 언어별로 또 도메인 별로 이러한 서열화의 기준이 중심화 이론의 적용에 미치는 영향을 다각도로 분석한 논문이 많은데, 그런 대부분의 논문들은 서열화의 기준이 대용형의 해결에 어떤 영향을 미치는지에 대해 초점을 둔 반면, 중심화 기반 문장 순서 평가 척도의 관점에서는 다루어 진 적은 없다. [2,3]에서도 네 가지의 문장 순서 평가 척도를 비교 할 때, 구문 관계에 따른 서열화에서만 비교가 되었을 뿐 다른 서열화에서도 MIN.NOCB가 가장 좋은 결과를 가져올 수 있다는 것을 보장하지 못한다. 이것이 본 논문에서 제기하는 기존 연구에 대한 두번째 문제 제기로, 본 논문에서는 다양한 서열화의 관점에서 문장 순서 평가 함수를 논의하고, 한국어에서 문장 순서를 결정하는데 가장 적합한 명사의 서열화가 무엇인지를 밝히고자 한다. 본 논문에서는, 중심화 기반 문장 순서 평가 척도의 관점에서만 문장 순서의 자연스러움을 다루고, 다른 척도들은 배제한다.

2. 중심화 이론

중심화 이론[1]은 텍스트를 구성하고 있는 발화의 각 명사(구) 관점에서 응집성(cohesion)과 돋보임성(salience)의 상호작용을 통해, 텍스트의 국소적 결속성(local coherence)을 모델링한 담화 해석의 계산 모델이다.

중심화 이론에서 분석의 최소 기본 단위는 발화(utterance)로, 한 개의 발화는 세 개의 중심구조 - *Cf(forward-looking center)*, Cb(backward-looking center), Cp(preferred 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(Ui)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 대명사화되었다면, 현 발화(U_i)의 Cb도 역시 대명

표 1 발화간의 전이유형(transition type)

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

표 2 추론 비용을 고려한 발화간의 전이유형[21]

	$Cb(Un) = Cb(Un-1)$	$Cb(Un) \neq Cb(Un-1)$
$Cb(U_n) = Cp(U_n)$ and $Cb(U_n) = Cp(U_{n-1})$	Cheap-CONTINUE (C-CON)	Cheap-SMOOTH-SHIFT (C-CON)
$Cb(U_n) = Cp(U_n)$ and $Cb(U_n) \neq Cp(U_{n-1})$	Expensive-CONTINUE (E-CON)	Expensive-SMOOTH-SHIFT (E-SSH)
$Cb(U_n) \neq Cp(U_n)$	RETAIN (RET)	ROUGH-SHIFT (RSH)

사화 된다.

2. 발화간의 전이유형은 다음 순서로 선호된다.

CONTINUE > RETAIN > SMOOTH-SHIFT > ROUGH-SHIFT

전이유형은 $Cb(U_i)$와 $Cb(U_{i-1})$의 일치 여부와, $Cb(U_i)$와 $Cp(U_i)$의 일치 여부에 의해 결정된다(표 1). CON은 화자가 특정 지시물에 대해 이야기하고 있으면서, 다음 발화에서도 계속 그 지시물에 대해 이야기 하겠다는 의도를 표시하며, 그 지시물은 현 발화에서 Cb인 동시에 Cp로 표현된다. RET는 화자가 다음절에서 의식의 중심을 새로운 대상으로 옮기고 싶다는 의도를 표시하는 것으로, Cb는 그대로 유지되지만, 현재의 중심을 Cf-list에서 낮은 서열에 배치함으로써 Cp가 바뀌는 경우이다. SSH는 이전 발화와 비교해 중심은 변했지만, 새로운 중심에 대해 이야기하며 다음 발화에서도 계속 새롭게 바뀐 현재의 중심에 대해 이야기 하겠다는 의도를 표시한다. RSH는 중심도 변하고 새로운 중심이 Cf-list에서 낮은 서열에 배치되는 경우이다.

[21]에서는 추론 비용(inference cost)을 고려하여 중심화 이론을 재고안 하였는데, $Cb(U_i)$와 $Cp(U_{i-1})$의 일치 여부에 따라 기존의 네 가지 전이유형을 여섯 가지로 확대하였다(표 2). 또한 [21]에서는 모든 전이 유형 쌍에 대해 추론 비용이 싼, 즉 흐름이 자연스러운 전이 유형을 구했는데, (CON, CON), (CON, RET), (RET, SSH), (RSH, SSH)등이 그것이다. [22]에서는 이렇게 여섯 가지의 전이유형으로 확장된 중심화 이론을 '비용기반 중심화 이론'이라고 명명하고 한국어의 영형을 설명하는데 적용하였다. 본 논문에서는 추론 비용이 싼 전이 유형과 유형 쌍을 문장순서 평가 척도에 활용할 것이다.

다음 예제 텍스트를 통해 비용기반 중심화 이론이 어떻게 적용되는지 살펴보자. 이 예제 텍스트에서는 발화의 단위를 *시제절로 정의하고*[8-11], [10]에서 설정한 다음과 같은 순위로 명사의 서열화를 결정한다고 가정하였다.

주제[2] > 주어 > 직접 목적어 > 간접 목적어 > 관형어 > 부사어

실제 [10]에서는 관형어와 부사에 대해서는 순위를 명시하지 않았지만, 본 논문에서는 직관적으로 관형어를 부사어보다 높은 순위에 두었으며, 각각의 동일 순위에 여러 명사가 존재할 경우, 문장에서 먼저 실현된 것을 우선 순위에 두어 순서를 매겼다.

(텍스트 1)

U_1: 쇠줄은 쇠나 나무의 거친 표면을 갈아내는데 사용하는 도구이다.
⇒ Cf: 쇠줄 > 표면 > 쇠 > 나무 > 도구, Cp: 쇠줄

U_2: 쇠줄은 강철로 만드는데,
⇒ Cf: 쇠줄 > 강철, Cp: 쇠줄, Cb: 쇠줄, Trans: C-CON

U_3: 쇠줄은 줄칼, 좌도라고도 한다.
⇒ Cf: 쇠줄 > 줄칼 > 좌도, Cp: 쇠줄, Cb: 쇠줄, Trans: C-CON

U_4: 쇠줄에는 줄과 환이 있다.
⇒ Cf: 줄 > 환 > 쇠줄, Cp: 줄, Cb: 쇠줄, Trans: RET

U_5: 줄과 환은 그 모양과 쓰임이 거의 비슷하나,
⇒ Cf: 줄 > 환 > 모양 > 쓰임, Cp: 줄, Cb: 줄, Trans: C-SSH

U_6: 줄은 주로 금속을 연마할 때 쓰고,
⇒ Cf: 줄 > 금속 > 연마, Cp: 줄, Cb: 줄, Trans: C-CON

U_7: 환은 목재를 마름질할 때 쓴다.
⇒ Cf: 환 > 목재 > 마름질, Cp: 환, Cb: NULL, Trans: NOCB

U_8: 환은 일반적으로 쇠붙이 대신 상어껍질로 만든다.
⇒ Cf: 환 > 쇠붙이 > 상어껍질, Cp: 환, Cb: 환, Trans: C-CON

위의 텍스트는 한국의 전통적인 생활 기구 '쇠줄'에 대한 설명으로, 텍스트 전체의 주제는 '쇠줄'이다. U_1은

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

중심화 이론의 정의에 따라 Cb가 없다. U_2에서는 $Cb(U_1)$이 정해지지 않았고, $Cb(U_2) = Cp(U_2)$이므로 C-CON이 된다. U_3은 이전 발화의 주제였던 '쇠줄'이 계속 주제로 유지되어 $Cb(U_3) = Cp(U_3) = Cb(U_2) = Cp(U_2)$이 되고 C-CON이 된다. U_4에서는 이전 발화들의 주제였던 '쇠줄'의 돋보임성의 강도가 떨어진 반면, '줄'과 '환'이 새로운 주제로 된다. 중심화 이론의 규칙상 Cb와 Cp는 하나이므로 순서상 '줄'을 Cp로 가정하고 RET이 된다. 앞서 설명했듯이, RET은 화자가 다음절에서 의식의 중심을 새로운 대상으로 옮기고 싶다는 의도를 표시하는 것으로, 여기서는 Cb, 즉 '쇠줄'은 그대로 유지되지만, 현재의 중심을 Cf-list에서 낮은 서열에 배치함으로써 다음절에서 의식의 중심을 새로운 대상, 즉 '줄'로 옮기고 싶다는 의도를 표시하는 것이다. U_5에서는 '줄'이 화제로 자리 잡아서 C-SSH가 실현되고, 이는 화제가 '쇠줄'에서 '줄'로 자연스럽게 옮겨 갔음을 의미한다. U_6에서는 화제가 '줄'로 C-CON이 실현되었다. U_7은 U_6과 공유하는 명사가 없으므로 NOCB가 되고, U_8은 새로운 화제 '환'에 관한 것으로 C-CON이 실현되었다. 위의 텍스트에서는 U_7, U_8을 제외하고는 다 전이 유형이 싼, 자연스러운 전이의 흐름이 실현되었다고 볼 수 있다.

3. 중심화 이론 기반 문장 순서 평가 척도

3.1 기존 연구

이 절에서는 어떻게 중심화 이론이 문장 순서를 평가하는 척도로 사용될 수 있는지를 설명하고, 기존에 제안되었던 방법을 소개한다.

3.1.1 중심화 이론 기반 문장 순서 평가 척도의 종류

총 10가지의 평가 척도가 표 3에 정리되어 있는데, 이름에 'MIN'이 포함되어 있는 평가 척도는 해당되는 조건을 최소화하는 문장 순서를 선호하는 것이고 'MAX'는 반대로 해당 조건을 최대화 하는 문장 순서를 선호하는 것이다. 이 중 (1)에서 (6)까지의 6가지 방법이 영어를 대상으로 기존의 여러 논문에서 각기 제안되었던 것으로 [2], [3]에서 코퍼스 분석을 통해 비교 분석되었다. MIN.NOCB는 가장 간단한 방법으로 NOCB를 최소화 하는 문장 순서를 선호하는 것이다. MIN.MIL은 [5]에서 제안되었으며 MIN.KP는 [4]에서 제안되었다.

3.1.2 가능한 문장 순서들의 탐색

[2], [3]에서는 표 10의 (1)-(6)까지의 평가 척도를 평가하기 위해 분류율(classification rate)과 BfC(Basis for Comparison)의 개념을 정의하였다. BfC는 텍스트의 본래 문장 순서를 의미하며, 각 문장의 Cf-list로 표현된다. 다음은 2장의 텍스트 1에 대한 BfC이다.

(텍스트 1의 BfC B_1)

U_1: 쇠줄 > 표면 > 쇠 > 나무 > 도구
U_2: 쇠줄 > 강철
U_3: 쇠줄 > 줄칼 > 좌도
U_4: 줄 > 환 > 쇠줄
U_5: 줄 > 환 > 모양 > 쓰임
U_6: 줄 > 금속 > 연마
U_7: 환 > 목재 > 마름질
U_8: 환 > 쇠붙이 > 상어껍질

[2], [3]에서는 텍스트 구조화의 입력은 위의 BfC B_1와 같이 Cf-list로 표현된 문장들의 집합이라 가정한다. BfC B에 대한 평가 척도 M의 분류율은, M이 텍스트 구조화의 결과로 B를 생성할 수 있는 가능성을 예측하는 것으로 다음 시나리오에 따라 예측된다.

1. B를 구성하고 있는 문장들로 만들어 낼 수 있는 가능한 문장 순서들을 탐색한다.
2. 각각의 문장순서들에 대해 M을 적용하여 B보다 점수가 높은 것, 같은 것, 낮은 것들을 분류하고 이 세 그룹의 비율을 조사한다.

문장 순서의 점수가 높을수록, 텍스트 구조화의 결과로 채택될 가능성이 높다는 가정하에, 분류율 v는 다음 식 (1)에 의해 구해진다.

$$v(M,B) = 1 - (Better(M) + \frac{Equal(M)}{2}) \quad (1)$$

표 3 중심화 기반 문장 순서 평가 척도

이름	설명 (다음 조건을 만족하는 문장 순서를 선호)
(1) MIN.NOCB	NOCB의 개수를 최소화
(2) MIN.V_CHP	CHEAPNESS ($Cb(U_i) = Cp(U_{i-1})$)의 위반 횟수를 최소화
(3) MIN.V_SAL	SALIENCE ($Cb(U_i) = Cp(U_i)$)의 위반 횟수를 최소화
(4) MIN.V_COH	COHERENCE ($Cb(U_i) = Cb(U_{i-1})$)의 위반 횟수를 최소화
(5) MIN.KP	NOCB + V_CHP+V_SAL+V_COH를 최소화
(6) MIN.MIL	NOCB 와 RSH를 최소화
(7) MAX.C-CON	C-CON의 개수를 최대화
(8) MAX.E-CON	E-CON의 개수를 최대화
(9) MAX.CPS	인접한 문장의 Cp를 같게 하는 문장쌍의 개수를 최대화
(10) MAX.P_SEQ	추론 비용이 싼 전이 유형의 개수를 최대화

Better(M)은 M을 적용했을 때 B보다 높은 점수를 갖는 문장 순서의 비율을 의미한다. Equal(M)/2는, B와 같은 점수를 얻은 문장 순서는 M에 의해 텍스트 구조화의 결과로 선택될 가능성이 반반이라는 의미에서 더해졌다. $\upsilon(M_x, B)$가 M_x를 B에 적용했을 때의 분류율이라 하고 $\upsilon(M_y, B)$가 M_y를 B에 적용했을 때의 분류율이라고 할 때, $\upsilon(M_x, B)$가 $\upsilon(M_y, B)$보다 클 경우, M_x가 M_y보다 B를 생성하는데 더 적합하다고 볼 수 있으며, 분류율이 큰 M일수록 더 좋다. 두 개의 평가 척도 M_x와 M_y를 비교할 때 코퍼스 C의 여러 개의 BfC(B1, …, Bm)에 대해 비교 되어야 하므로 M의 평균 분류율 Y는 다음 식 2(와) 같이 계산된다.

$$Y(M,C) = \frac{\upsilon(M,B_1) + ... + \upsilon(M,B_m)}{m} \quad (2)$$

3.2 제안하는 중심화 이론 기반 문장 순서 평가 척도의 종류

본 논문에서 제안하는 문장 순서 평가 척도를 표 3의 (7), (8), (9), (10)에 정리하였다. MAX.C-CON과 MAX.E-CON은 C-CON과 E-CON이 문장 순서의 자연스러움을 판단하는 데 어떤 차이를 보이는지를 알기 위함이다. 각각은, C-CON과 E-CON 각각을 최대화 하는 문장 순서를 찾고, 그런 문장 순서가 여러 개 일 때는 중심화 이론의 규칙 2에 따라 RET가 최대인 것을, 그 다음은 SSH, RSH의 순으로 많은 것을 선호하게 된다.

중심화 이론을 통해 텍스트의 응집성, 대용 해결, 대용어 생성들을 설명할 때 대부분의 논문들은 인접한 두 발화(U_{i-1}, U_i)의 전이를 함께 고려하는 것이 유용하다는 것에 의견의 일치를 보인다[1], [4], [19], [10], [12], [13], [15], [21]-[26]. MAX.CPS와 MAX.P_SEQ는 이런 기존 연구 결과에 바탕을 두었다. MAX.CPS와 관련해, Cp는 현행 발화에서 가장 두드러진 명사를 의미하는 것으로 인접한 두 문장에서 각각 가장 두드러진 명사가 일치하면 그 두 문장은 국소적으로 응집되어 있다고 볼 수 있다. MAX.P_SEQ는 [21]의 값싼 전이 유형이 많은 문장 순서를 선호하는 척도이다.

본 논문에서는 각각의 중심화 기반 문장 순서의 평가 척도를 비교하기 위해, 코퍼스에서 추출한 텍스트를 활용하였으며, 평가 척도 (1)-(6)과 함께 제안한 평가 척도 (7)-(10)을 한국어에 적용하고 그 효과를 비교하기 위해 3.1.2절에서 소개한 분류율(classification rate)의 개념과 BfC(Basis for Comparison)의 용어를 빌려왔다.

4. 중심화 기반 문장 순서 평가 척도의 기대치

4.1 MIN.NOCB의 문제점

[2], [3]에서 네 가지 평가 척도 - MIN.NOCB, MIN.V_CHP, MIN.KP, MIN.MIL을 식(1)의 분류율의 관점에서 비교한 결과, MIN.NOCB가 나머지 세 개에 비해 가장 단순하면서도 효율적이라고 주장하였다. 즉, 단순히 NOCB의 전이 유형을 최소화 하는 문장 순서를 찾는 것이 중심화 이론 안에서 최상의 문장 순서를 찾는 방법이라는 결론을 도출하였다. 3.1.2절의 BfC B_1을 살펴보자. B_1에는 NOCB가 1개가 존재하지만, B_1을 구성하고 있는 문장으로 임의의 문장 순서를 만들었을 때는 NOCB가 더 많이 생길 수 있다. 가령, $U_1 \rightarrow U_4 \rightarrow U_7 \rightarrow U_3 \rightarrow U_5 \rightarrow U_6 \rightarrow U_2 \rightarrow U_8$의 순서에 대해서는 4개의 NOCB가 발생한다. 따라서 MIN.NOCB는 B_1에 대해서는 효과적으로 적용될 수 있다. 하지만, MIN.NOCB를 적용할 때 발생할 수 있는 중요한 문제가 있는데, 먼저 '어리'를 설명하고 있는 다음 텍스트와 그것에 해당되는 BfC B_2를 살펴보자.

(텍스트 2)

U_1: 어리는 병아리를 기르는 우리이다.
U_2: 어리는, 종두리, 가리, 닭장이라고도 한다.
U_3: 어리와 같은 가축용 우리는 삼국시대부터 제작된 것으로 추정된다.
U_4: 어리의 재질 및 형태는 지역에 따라 다양하다.
U_5: 어리는 대나무나 싸리 등으로 성글게 만드는데,
U_6: 영남지방에서는 어리를 둥글고 갸름하게 짠 반면,
U_7: 호남지방에서는 어리를 길고 넓게 짰다.
U_8: 주로 행랑채나 헛간에 어리를 매달아 두었다.

(텍스트 2의 BfC B_2)

U_1: 어리 > 병아리 > 우리
U_2: 어리 > 종두리 > 가리 > 닭장
U_3: 우리 > 어리 > 가축용 > 삼국시대
U_4: 재질 > 형태 > 어리 > 지역
U_5: 어리 > 대나무 > 싸리
U_6: 어리 > 영남
U_7: 어리 > 호남
U_8: 어리 > 행랑채 > 헛간

BfC B_2에서 NOCB는 존재하지 않으며, 8개의 모든 문장이 '어리'를 함께 공유하고 있어 8개의 문장으로 만들어 낼 수 있는 어떠한 문장 순서에서도 NOCB는 발생하지 않는다. 즉 B_2의 경우는 MIN.NOCB가 텍스트 구조화 과정에 있어 B_2를 생성할 분류율은 0.5이며, 실상 모든 문장 순서에 있어서 B_2를 구별할 수 있는 가능성은 없다.[3] 따라서, 임의의 BfC B에 대해 MIN.NOCB

3) 마찬가지로, 어떤 명사도 전혀 공유하지 않아서 임의의 문장순서에서도 NOCB가 똑같이 생기는 BfC에서도 MIN.NOCB는 아무런 효력을 발휘하지 못한다. 하지만, 어떤 대상을 설명하는 텍스트에서 이런 *BfC*는 없다고 가정할 수 있다.

를 적용해 문장 순서의 적절성을 판단하는 것은 그 한계가 명확하다. 이것이 본 논문에서 다루고자 하는 첫 번째 사안이다. 기존 연구에 따라 일반적으로 MIN.NOCB가 다른 평가 척도에 비해 높은 성능을 낸다고 가정할 때, MIN.NOCB의 적용에 따른 해결책은 다음과 같이 정리될 수 있다.

(1) MIN.NOCB가 BfC B에 적용되었을 때의 기대치를 먼저 예측하라. 그 기대치가 높으면 MIN.NOCB를 적용하고 그렇지 않으면 다른 평가 척도를 적용하라.

위의 (1)의 정책이 중심화 이론 안에서 최상의 문장 순서를 찾을 수 있는 방법이라 생각할 수 있는데, (1)의 정책을 실현함에 있어 2가지 사안을 고려해야 한다. 첫 째는 MIN.NOCB의 BfC B에 대한 기대치를 어떻게 예측할 것인가 이고, 둘째는, MIN.NOCB를 대체할 수 있는 다른 평가 척도는 무엇인가를 밝히는 것이다.

4.2 MIN.NOCB의 기대율

평가 척도 M이 BfC B에 대해 갖는 기대율 e는, M을 적용했을 때 B를 생성할 가능성을 *미리 예측*하는 것으로 식 (3)으로 표현된다. 먼저 M이 MIN.NOCB인 경우만을 고려해 보자.

$$e(M,B) = avg(M, permB) - ori(M,B) \quad (3)$$

permB는 B로 만들어 낼 수 있는 모든 가능한 문장 순서를 의미한다. avg(M, permB)는 permB에서 NOCB의 평균 개수를, ori(M, B)는 B에서 NOCB의 개수를 의미한다. Avg(M, permB)는 다음의 간단한 유도 과정으로 구할 수 있다.

<Avg(M, permB) 유도>

$N(k)$를 k개의 NOCB를 갖는 문장 순서의 개수라고 하고, n은 B의 문장 개수라고 하자. 그러면 avg(M, permB)를 다음과 같이 표현할 수 있다.

$$avg(M, permB) = \frac{\sum_{k=1} N(k)k}{n!} \quad (4)$$

두 문장이 인접했을 때 어떤 명사도 공유하지 않는, 즉 인접했을 때 NOCB를 생성하는 문장쌍의 개수를 p라고 두고 이런 문장쌍을 *끊긴 문장쌍*이라 부르자. 그러면, $\sum_{k=1} N(k)k$ 는 각각의 끊긴 문장쌍을 포함하는 문장 순서의 개수의 합과 같으며, 그 합은 $2p(n-1)!$이다. 문장쌍을 이루는 각 문장을 바꾼 경우도 함께 고려해 주어야 하기 때문에 2가 곱해졌다. 따라서, avg(M, permB)는 다음과 같이 정리될 수 있다.

$$avg(M, permB) = \frac{\sum_{k=1} N(k)k}{n!} = \frac{2p(n-1)!}{n!} = \frac{2p}{n} \quad (5)$$

3.1.2절의 BfC B_1의 경우를 살펴보자. B_1에서 n = 8이고, 다음과 같은 14개의 끊긴 문장쌍 - (U_1, U_5), (U_1, U_6), (U_1, U_7), (U_1, U_8), (U_2, U_5), (U_2, U_6), (U_2, U_7), (U_2, U_8), (U_3, U_5), (U_3, U_6), (U_3, U_7), (U_3, U_8), (U_6, U_7), (U_6, U_8) - 이 존재하므로 p = 14이다. 따라서 이 14개의 각각의 끊긴 문장쌍을 포함하는 문장 개수의 합은 2*14*7!개이고, permB$_1$의 평균 NOCB의 개수는 2*14/8 = 3.5가 된다. 최종적으로, e(M, B_1)는 3.5-1 = 2.5이다. 반면, 4.1절의 BfC B_2의 경우는 p = 0이 되므로 avg(M, PermB_2) = 0이고 따라서, MIN.NOCB가 B_2에 적용되었을 때 B_2를 생성할 예상 기대율은 0이다. B_1과 B_2의 기대율의 비교를 통해서도 알 수 있듯이, e(M, B_1)이 e(M, B_2)보다 크면 M은 B_2를 생성하는 것보다 B_1를 생성하는데 더 적합하며, 이는 v(M, B_1)이 v(M, B_2)보다 크다는 것을 예측 가능하게 한다. 그림 1은 M이 MIN.NOCB일 때, avg(M, permB), ori(M, B), e(M, B)와 Better(M)의 관계를 설명해 준다. MIN.NOCB일 때 일반적으로 ori(M, B)가 avg(M, permB)보다 작은데, ori(M, B)의 위치가 왼쪽으로 갈수록 e(M, B)가 커지고 그 결과 Better(M)가 작아져서 v(M, B)가 커지게 됨을 알 수 있다. e(M, B)와 v(M, B)의 관계는 6.2절에서 다시 설명하겠다.

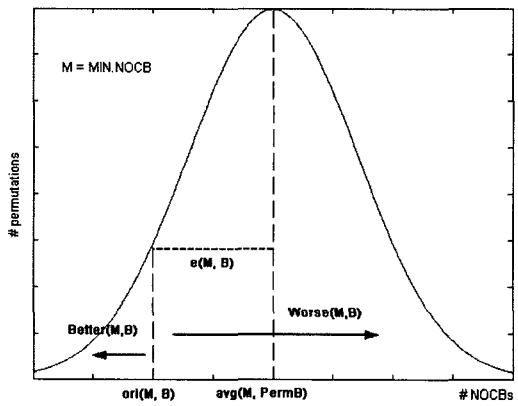

그림 1 Avg(M, permB), ori(M, B), e(M, B)와 Better(M)의 관계

4.3 다른 평가 척도에 대한 기대율의 일반화

4.2절에서 MIN.NOCB에 대한 기대율은, MIN.V_SAL, MIN.V_CHP, MAX.CPS에는 그대로 적용될 수 있다. MIN.NOCB에 대한 기대율이 그대로 적용될 수 있으려면, MIN.NOCB에서 끊긴 문장쌍처럼 주어진 평가 척도의 조건에 대한 결과를 발화쌍 내에서 구할 수 있는 것들이어야 하는데, MIN.V_SAL, MIN.V_CHP, MAX.CPS은 그것이 가능하다. 하지만 발화쌍내의 발화에 대해 조건이 대칭성을 가지는지의 여부는 다른데, MAX.CPS는 대칭성을 가지기 때문에 MIN.NOCB와 똑같이 계산되지만, MIN.V_SAL과 MIN.V_CHP는 대

칭성을 갖지 않으므로 p개에는 (U_i, U_j)와 (U_j, U_i)가 다 포함될 수 있어 avg(M, permB) = p/n으로 계산된다. 나머지 평가 척도에 대해서는 (U_i, U_j)에 대해 해당 조건 값을 구하기가 힘든데, 예를 들어 MIN.V_COH의 경우 U_j가 U_i의 Cb와 같은지를 확인하기 위해서는 U_i의 Cb를 알아야 하는데 U_i의 Cb는 U_i에 선행하는 문장에 의해 결정된다. 따라서 나머지 평가 척도는 MIN.NOCB의 경우보다 시간적으로 높은 복잡도를 갖거나, 일반화되기 힘들다.

5. 명사 서열화의 다양한 방법

문장에서 실현된 명사들의 상대적인 돋보임성을 결정하는 것은 중심화 이론을 적용하는 데 있어 가장 핵심적인 부분이라 할 수 있다. 중심화 기반 문장 순서 평가 척도의 대부분은 중심 전이와 관련이 있는데, 이런 중심 전이들은 명사들의 상대적인 돋보임성의 순위를 어떻게 결정하느냐에 따라 완전히 달라질 수 있기 때문이다.

서열화의 기준은 일반적으로 언어마다 다른데[12], 구문 관계(grammatical role)[13-18], 의미 관계(thematic role)[19], 문장 순서(linear order)[20], 정보 위상(information status)[21] 등이 논의 되어 왔으며, 그 중 구문 관계에 따른 서열화가 가장 보편적이다. 한국어에서는 [10]에서 적용한 다음 기준에 따른 서열화가 보편적이다.

주제 > 주어 > 직접 목적어 > 간접 목적어 > 그 외

영어권에서는 위의 서열화에서 주제를 제외한 나머지 부분의 서열이 일반적이다.

언어별로 또 도메인 별로 이러한 서열화의 기준이 중심화 이론의 적용에 미치는 영향을 다각도로 분석한 논문이 많은데[25], 그런 대부분의 논문들은 서열화의 기준이 대용형의 해결에 어떤 영향을 미치는지에 대해 초점을 둔 반면, 중심화 기반 문장 순서 평가 척도의 관점에서는 다루어 진 적은 없다. [2], [3]에서도 네 가지의 문장 순서 평가 척도를 비교 할 때, 구문 관계에 따른 서열화에서만 비교가 되었기 때문에 다른 서열화에서도 M.NOCB가 가장 좋은 결과를 가져올 수 있다는 것을 보장할 수 없다.

표 4에서 네 가지의 기본 서열화 방식 - 구문 관계, 단어 순서, 정보 위상, 명사의 대용형을 정리하였다. 정보 위상은, 구정보를 신정보보다 돋보임성을 높게 두는 것으로, 이전 문장에서 실현되었던 명사가 현 문장에서 구정보가 되며, 그렇지 않은 명사가 신정보가 된다. 명사의 대용형과 관련해 본 논문에서는 영형 대용형, 비영형 대용형의 순으로 위계를 설정하였는데, 영형의 돋보임성은 기존 연구에서 여러 차례 다루어져 왔다. [27]에서는 한국어에서 돋보이는 요소로 '는'으로 실현된 화제뿐만 아니라 영형도 해당한다고 주장한다. 한국어 외에도 일본어나 영어에서도 영형의 돋보임성은 확인된다. [12]는 일본어에서 영형의 돋보임을 보이고 Cp에 영형이 위치할 수 있음을 제시하고 있다.

명사의 돋보임성의 위계를 결정하는데 대용형의 활용은 문장 순서를 결정하는 텍스트 구조화 단계에서는 적용이 어려울 수 있다. 그 이유는, 일반적인 텍스트 생성 과정에서 문장 순서를 결정하는 텍스트 구조화 단계 다음에 대용화(pronominalization) 과정을 처리하므로, 문장 순서를 결정할 때는 명사들의 대용형이 결정되지 않은 시점이기 때문이다. 하지만 텍스트-텍스트 생성 시스템(text-to-text generation)에서는 대용형이 텍스트 구조화에서 활용 가능하므로, 본 논문에서는 명사의 대용형에 따른 위계 설정을 다른 서열화 방법과 함께 논의하도록 한다.

중심화 이론의 제약 1에 따라, 모든 문장에는 Cb가 하나여야 하므로 명사의 서열화에 있어 같은 서열을 갖

표 4 기본 서열화 방식

이름	설명
구문 관계 (G)	주제 > 주어 > 간.목 > 직.목 > 나머지
단어 순서 (O)	문두에 실현된 명사 > … > 문미에 실현된 명사
정보 위상 (I)	구정보 > 신정보
명사의 대용형 (F)	영형 > 비영형 대용형 > 대명사가 아닌 명사류

표 5 기본 서열화 방식의 조합

이름	설명
G+O	G에서 동일 서열의 명사들에 대해, O 적용
I+O	I에서 동일 서열의 명사들에 대해, O 적용
I+G+O	I에서 동일 서열의 명사들에 대해, G+O 적용
F+O	F에서 동일 서열의 명사들에 대해, O 적용
F+G+O	F에서 동일 서열의 명사들에 대해, G+O 적용
I+F+O	I+F에서 동일 서열의 명사들에 대해, O 적용
I+F+G+O	I+F에서 동일 서열의 명사들에 대해, G+O 적용

는 명사들을 허락하지 않는다. 표 4에 있는 서열화 방식 중 단어 순서에 따른 서열화를 제외한 나머지는 동일 서열의 명사들을 허락하기 때문에 이를 해결하기 위한 서열화의 조합을 표 5에 제안하였다. 구문 관계와 문장 순서의 조합은 아주 기본적인 방식으로 기존 연구에서 여러 차례 활용되었다[17,21,25].

6. 실험 및 평가

6.1 실험 준비

[22]에서 사용되었던 온라인 박물관의 전시물들을 설명한 텍스트 51개를 이용하였다. 수작업으로 텍스트의 모든 문장을 중심화 이론을 적용할 수 있도록 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 구문 분석기를 통해 각 단어들의 구문 관계를 획득하고 중심화 이론을 적용하였다.

임의의 한 텍스트가 n개의 발화로 구성되었을 때, n개의 발화로 만들어 낼 수 있는 문장 순서는 $n!$개이다. n이 작은 경우에는 $n!$ 즉, 전체 탐색 공간을 다 조사할 수 있지만, n이 큰 경우에는 모든 탐색 공간을 조사하는 것은 현실적으로 불가능하다.

본 논문에서 사용한 코퍼스에서 텍스트당 평균 발화 개수는 13.18개이고 최소 6개에서 최대 23개의 발화로 구성되어 있다. 발화의 개수가 7이하인 경우 전체 탐색 공간을 다 조사하였다. 그 이상인 경우는 발화의 수와 상관없이 임의의 10000개의 문장 순서만을 조사하였으며, 이를 3회 시행하여 M에 대한 분류율을 구할 때 평균값을 취했다. 이 때, 동일 텍스트에 대해 3회 실행한 실험에서 10000개의 문장 순서에 대한 각 실험에서 분류율의 차이는 0.001 미만으로, 발화의 개수와는 상관없이 10000개 정도의 문장 문서에 대한 검색 만으로도 M에 대한 분류율을 안정적으로 구할 수 있었다. M_x와 M_y에 대한 분류율의 차이가 통계적으로 유의한지를 알아 보기 위해 t-test를 적용하였다.

6.2 중심화 기반 문장 순서 평가 척도의 비교

기대율 vs. 분류율 그림 2는 MIN.NOCB의 기대율(x축)과 분류율(y축)의 관계를 보여준다. 기대율이 약 4보다 작을 때까지는 기대율이 커질수록 분류율도 커지고, 기대율이 4보다 큰 경우는 분류율이 거의 1에 근접한다. 기대율이 0.5~2사이에 분류율의 변동이 심한 부분이 있는데, 이는 해당되는 텍스트의 n!개의 문장 순서에서 NOCB의 개수에 대한 분산(variance)의 차이으로 설명할 수 있다. 그림 3은 같은 기대율에서 분산이 분류율에 미치는 영향을 보여준다. 두 그래프의 면적이 동일하다

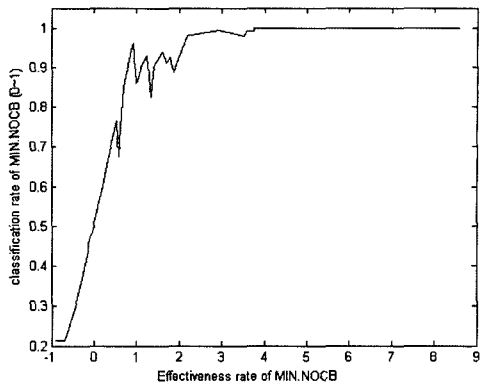

그림 2 MIN.NOCB의 기대율(x축)과 분류율(y축)의 관계

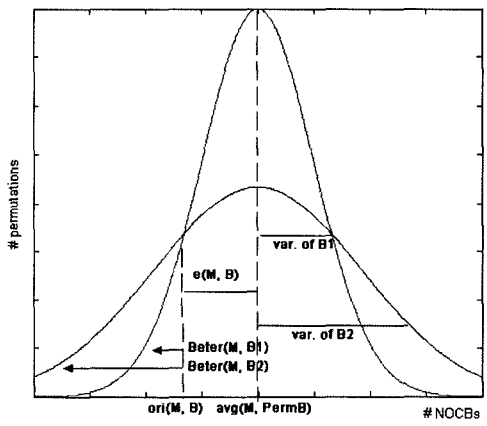

그림 3 분류율에 대한 분산의 영향

고 가정할 때, 분산이 큰 B_2의 경우 분산이 커서 Better(M)의 면적이 넓어지고 따라서 M의 B_2에 대한 분류율은 낮아진다. 반면, 분산이 작은 B_1의 경우, 분산이 작아서 Better(M)의 면적이 좁아지고 따라서 M의 B_1에 대한 분류율이 커진다. 이런 사실을 통해, n!개의 문장 순서에서 M의 적용에 따른 값들에 대한 분산이 분류율에 영향을 끼치고, 기대율을 구할 때 ori(M, B)와 avg(M, permB)의 값의 차이뿐만 아니라 variance(M, permB)에 대한 값도 함께 고려해 주는 것이 더 타당하다는 사실을 알 수 있다. 하지만, 기대율을 구할 때 variance(M, permB)를 함께 고려할 수 없는 이유는, avg(M, permB)와는 달리 variance(M, permB)값은 n!개의 탐색 공간을 조사하지 않고서는 쉽게 예측될 수 없기 때문이다. 하지만 그림 2를 통해 ori(M, B)와 avg(M, permB)의 값이 차이에 의한 기대율이 전반적으로 실제 분류율을 잘 설명해 준다는 사실을 확인할 수 있다.

그럼, 지금부터는 먼저 모든 텍스트의 BfC에 대해서 각 중심화 기반 문장 순서 평가 척도에 대한 실험적 비

교 결과를 논의한 다음, 기대율이 0.5이하(이때 분류율은 대략 0.7이하가 된다)인 BfC만을 따로 분류하여 BfC.LN이라고 명명하고 이들에 대한 평가 척도의 우위를 따로 분석할 것이다. 평가 척도간의 비교는 주로 MIN.NOCB와 나머지 것들의 비교에 초점을 맞춘다.

MIN.NOCB vs. 다른 평가 척도 표 6은 서열화 방법에 따른 각 평가 척도에서 51개 BfC의 평균 분류율을 보여준다. 볼드체로 표현된 4가지의 경우에서 (MAX.CPS-O, MAX.CPS-F+G+O, MIN.MIL-O, MIN.MIL-F+O) 분류율이 MIN.NOCB보다 높았으나 그 차이는 t-test에 의해 통계적으로 유의하지 않았다. 총 17개의 BfC.LN에 대한 평균 분류율을 표 7에서 정리하였다. 이 결과로부터 얻어지는 몇 가지 흥미로운 결과는 다음과 같다. 첫째, MIN.NOCB의 분류율과 비교했을 때, 다른 평가 척도에서 서열화 방법과 상관없이 MIN.NOCB의 분류율보다 높은 경우가 많다. 또한 이탤릭체로 표기된 26가지의 경우에 해당하는 분류율은 MIN.NOCB의 분류율과의 차이가 t-test에 의해 통계적으로 유의함이 밝혀졌다 ($p < 0.03$). 특히 MAX.CPS-O에서의 분류율(볼드체로 표기)이 가장 높으며 MAX.NOCB와의 차이는 $p = 0.0031$로 통계적으로 역시 유의하다. 이는 MIN.NOCB의 기대율이 낮은 BfC에 대해서는 MAX.CPS-O가 가장 효율적이라는 의미이다. 단어 순서에 기반한 명사의 서열화에서 $Cp(U_i)$는 현 문장에서 순서상 처음으로 실현된 명사를 의미한다. 따라서 다시 정리하면, n!개의 문장순서에서 NOCB가 거의 발생하지 않고 그래서 avg(MIN.NOCB, permB)와 ori(MIN.NOCB, B)의 차이가 작아 MIN.NOCB가 적합하지 않은 BfC B에 대해서는, 인접한 문장들에서 단어 순서상 처음으로 실현되는 명사들을 최대한 같게 만드는 문장 순서를 찾는 것이 중심화 이론으로서 할 수 있는 가장 효과적인 방법이라고 볼 수 있다. 따라서 4.1절에서 제안되었던 정책 (1)은 다음과 같이 최종 정리될 수 있다.

(2) MIN.NOCB가 BfC B에 적용되었을 때의 기대치를 먼저 예측하라. 그 기대치가 높으면 MIN.NOCB를 적용하고 그렇지 않으면 문장 순서에 기반한 명사의 서열화로 MAX.CPS를 적용하라. 즉, NOCB를 최소화 하는 문장 순서를 찾는 것이 의미가 없으면, 인접한 두 문장에서 단어 순서상 맨 처음 실현되는 명사가 같은 문장을 최대한으로 많게 하는 문장 순서를 찾아라.

하지만, 위의 정책을 실제 텍스트 생성 시스템에 바로 적용하기는 힘들다. 그 이유는, 위에서 행해진 실험과는 달리 ori(M, B)가 텍스트 구조화 과정에서는 알려져 있지 않기 때문이다. 하지만 BfC B에 대해 NOCB가 가장 작은 문장 순서가 이상적이라고 가정할 때, Viterbi 알고리즘[28]과 같은 최소 경로 찾기를 위한 알고리즘의 적용을 통해 주어진 B에서 n!개의 문장 순서 중 가장 작은 수의 NOCB를 갖는 문장 순서가 어떤 것인지, 또 그 개수는 몇 개인지는 쉽게 찾을 수 있다. 따라서, 이런 알고리즘을 이용해 텍스트 구조화 과정에서 ori(MIN.NOCB, B)를 먼저 예측한 다음 위의 정책을 적용할 수 있다. 또 다른 가능한 정책으로는 다음을 들 수 있다.

(3) 문장 순서에 기반한 서열화를 통해 MAX.CPS를 적용하라. 즉, 인접한 두 문장에서 단어 순서상 맨 처음 실현되는 명사가 같은 문장을 최대한으로 많게 하는 문장 순서를 찾아라.

위 정책에 대한 근거는 다음과 같다. 먼저, 표 6에서

표 6 서열화 방법에 따른 각 평가 척도에서 51개의 BfC의 평균 분류율

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
O	0.756	0.425	0.516	0.408	0.593	**0.764**	0.730	0.513	**0.766**	0.739
G+O	〃	0.477	0.401	0.374	0.520	0.708	0.684	0.581	0.739	0.693
I+O	〃	0.360	0.440	0.392	0.538	0.756	0.641	0.456	0.644	0.697
I+G+O	〃	0.399	0.395	0.394	0.534	0.740	0.617	0.520	0.632	0.681
F+O	〃	0.412	0.551	0.405	0.602	**0.767**	0.712	0.466	0.738	0.726
F+G+O	〃	0.464	0.470	0.376	0.575	0.738	0.710	0.569	**0.772**	0.683
I+F+O	〃	0.363	0.451	0.391	0.544	0.756	0.648	0.460	0.652	0.696
I+F+G+O	〃	0.402	0.405	0.391	0.541	0.747	0.625	0.522	0.638	0.678

표 7 서열화 방법에 따른 각 평가 척도에서 17개의 BfC.LN의 평균 분류율

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
O	0.425	*0.610*	*0.581*	*0.550*	*0.612*	*0.468*	*0.628*	*0.611*	***0.637***	*0.631*
G+O	〃	*0.532*	*0.538*	0.486	0.494	0.451	0.507	0.464	0.490	0.512
I+O	〃	0.484	0.466	*0.514*	0.445	0.415	0.452	0.423	0.421	0.490
I+G+O	〃	0.410	0.405	0.491	0.358	0.408	0.408	0.336	0.344	0.392
F+O	〃	*0.594*	*0.559*	*0.540*	*0.578*	0.456	*0.546*	*0.539*	*0.540*	*0.565*
F+G+O	〃	*0.568*	*0.522*	0.453	0.445	0.412	*0.620*	*0.524*	*0.571*	0.625
I+F+O	〃	0.503	0.480	*0.512*	*0.466*	0.416	0.472	0.441	0.441	0.480
I+F+G+O	〃	0.441	0.441	0.486	0.398	0.405	0.436	0.373	0.379	0.456

단어 순서에 의한 서열화에 있어서는 MIN.NOCB의 기대율과는 상관없이 MAX.CPS의 평균 분류율이 MIN.NOCB보다 높으며 게다가 MAX.CPS의 모든 BfC에 대한 분류율의 분산이 MIN.NOCB의 분산보다 작기 때문에 더 안정적이라고 볼 수 있다. 또한 표 7에서 MIN.NOCB의 기대치가 낮은 경우에 대해서도 MAX.CPS-O가 가장 높은 분류율을 보이고 있다. 하지만 정책 (2)와 (3)중 어떤 것이 더 나을지에 대한 검증은 본 논문에서 이루어 지지 않았으며 향후 연구 계획으로 남겨 둔다.

추론 비용을 고려한 중심화 이론에 기반한 평가 척도 표 6에서 일곱 번째와 여덟 번째의 열을 비교해 볼 때, MAX.C-CON이 MAX.E-CON보다 분류율이 더 높으며 이는 t-test에 의해 유의하다고 판단된다. 이런 특징은 비록 통계적으로는 유의하지 않지만 표 7에서도 드러난다. MAX.P_SEQ에 대해서는, 표 6과 표 7에서 각각 F+G+O와 I+G+O를 제외한 나머지 서열화 방법에서 MAX.P_SEQ가 MAX.C-CON보다 분류율이 높다. 이런 결과들을 바탕으로, 추론 비용을 기반한 중심 전이의 세분화와, 단일 전이의 발생 횟수 보다는 추론 비용이 싼 전이쌍의 실현 횟수가 좋은 문장 순서를 찾는데 효과적임을 알 수 있다.

6.3 서열화 방식의 비교

표 6과 표 7에서 각 열의 성능 차이를 비교함으로써 8가지의 서열화 방식에 대한 효과를 분석할 수 있다. 표 6과 표7에서 밑줄이 그어진 분류율은 각 열에서 가장 높은 성능을 의미한다. 표 6에서는 O와 F+O에 의한 서열화에서 가장 높은 성능을 낸 것이 3회로 가장 자주 높은 성능을 내었으며 그 다음은 G+O이고 그 다음은 F+G+O이다. BfC.LN에 대해서는 표 7에서 보듯이 모든 평가 척도에서 O가 항상 가장 높은 분류율을 보였다.

단어 순서 (O) vs. 구문 관계와 단어 순서의 결합 (G+O) 대부분의 기존 연구에서는, 단어 순서에만 의존에서 서열화를 결정하는 경우는 거의 없었던 반면, 다른 주 서열화 방식을 보완하기 위해 주로 단어 순서를 활용하였다. [25]에서는 영어에서 O에 의한 서열화와 G+O에 의한 서열화를 비교한 결과, 규칙 1과 규칙 2의 위반 횟수 관점에서 O가 G+O보다 좋지 않다는 결론을 내렸다. 이런 기존의 결과와는 반대로, 본 논문에서는 한국어에서 문장 순서를 결정하기 위해 중심화 이론을 적용할 때는 O가 G+O보다 좋은 서열화 방식이라고 본다. 특히, 표 7의 MIN.KP (5), MAX.C-CON (7), MAX.E-CON (8), MAX.CPS (9), MAX.P_SEQ (10)에서 O에 의한 분류율과 G+O에서의 분류율의 차이는 통계적으로 유의하다 ($p < 0.01$). O가 G+O보다 좋을 수 있는 몇 가지 이유를 살펴보자. 첫째, 한국어가 비교적 어순이 자유로운 언어라 할지라도, G에 의한 서열화에서 간접 목적어와 직접 목적어를 제외한 나머지 것들의 어순은 비교적 잘 지켜지는 편이다. 즉, G+O에 의한 서열화에서 높은 서열에 위치하는 명사들은 역시 O에 의해서도 높은 서열에 위치할 수 있는 것들로, 이는 G+O와 O가 별개의 것이 아니라 O에 의한 서열화에서도 G+O에 의한 서열화가 상당 부분 반영된다는 의미이다. 둘째, G에 의한 서열화 방식에서는 5장에서 보여지는 것처럼 주제, 주어, 목적어를 제외한 나머지 구문 관계(예, 관형사류, 부사류)에 대한 서열화는 정해지지 않았다. 따라서 이런 명사류에 대해서는 G+O에 의해 그 서열의 결정이 힘들지만 O에 의해서는 효과적으로 결정될 수 있는데, 이는 비교적 어순이 자유로운 언어에 있어서 먼저 실현되는 단어가 그렇지 않은 단어보다 비교적 돋보임성이 더 강할 수 있기 때문이다. 셋째, O에 의한 서열화는 문중에 위치하는 대조 화제(contrastive topic)의 서열을 낮추어 보다 정확하데 다음 문장의 Cb를 예측할 수 있게 한다. 5장에서 보여지는 서열화에서 '주제'를 Cp로 두는 것은 언어학적인 관점에서 다소 문제의 소지가 있을 수 있는데, 그 이유는 주제를 나타내는 표지 '-은/-는'은 주로 두 가지 종류의 화제의 표지로 같이 쓰이기 때문이다. 두 종류의 화제란, 첫째는 정보 구조(information structure)측면에서 '초점'(focus)과 반대되는 개념의 '화제'이고(화제 1), 둘째는 대조의 의미를 갖는 대조 화제(contrastive topic)이다(화제 2). 중심화 이론에서 명사의 돋보임성을 위한 서열로 '주제'를 가장 높게 두었을 때 그 '주제'는 화제 2가 아닌 화제 1만을 포함해야 정확하지만, 기존의 서열에서는 화제 1과 화제 2를 구별하지 않고 화제 표지 '-은/-는'의 부착만으로 해당 명사를 Cp로 설정하였다. 이런 두 종류의 화제를 구별할 수 있는 방법에 대해서는 언어학자마다 의견이 분분하나 문두의 화제를 화제 1이라 보고, 문중의 화제를 화제 2라고 보는 것이 지배적이다[29-31]. 따라서 G+O에 의한 서열화에서는 문중의 화제 2를 Cp로 잘못 인식할 수 있으나, O에 의한 서열화에서는 문중의 화제 2의 서열을 낮추어 줄 수 있어 보다 정확하게 Cp를 인식할 수 있게 한다.

반면에, O에 의한 서열화에서 발생할 수 있는 문제점은 한국어에서 수식어-피수식어 구조에서 기인한다. 한국어는 영어와 달리 수식어가 피수식어 앞에 위치해서 실제 Cp가 되어야 하는 명사를 수식하는 명사가 있을 경우, 그 수식어가 Cp가 되는 경우가 발생한다. 이를 해결하기 위해서는 문두에 위치한 수식어의 명사의 서열을 낮추는 방법이 보완되어야 할 것이다.

정보 위상 vs. 명사의 대용형 표 6과 표 7에서 I+O와 F+O, I+G+O와 F+G+O를 비교해 본 결과, F를 고려하는 것이 I를 고려한 것보다 더 좋음을 알 수 있다.

7. 결 론

본 논문에서는, 기존 연구들에서 가장 간단하고도 효과적이라고 알려진 MIN.NOCB를 텍스트 구조화에 적용할 때 발생할 수 있는 문제점을 지적하고 대안이 될 수 있는 새로운 평가 척도인 MAX.CPS를 제안하였다. 또한 텍스트 구조화에 주어진 입력에 대해 평가 척도의 기대치를 먼저 예측하고, 그것에 따라 다른 평가 척도를 적용하게 하는 프레임워크를 제안하였다. 이를 이용해 중심화 이론에서 보다 바람직한 문장 순서를 찾을 수 있는 방법론 - NOCB를 최소화 하는 문장 순서를 찾는 것의 의미가 없으면, 인접한 두 문장에서 단어 순서상 맨 처음 실현되는 명사가 같은 문장을 최대한으로 많게 하는 문장 순서를 찾아라 - 을 제안하였고 또 다른 대안으로써 단어 순서에 기반한 서열화에서 MAX.CPS의 적용을 제안하였다.

또한, 중심화 기반 문장 순서 평가 척도의 관점에서 8가지의 명사의 돋보임성의 서열화 방식을 평가한 결과 텍스트 구조화에 관한 한, 단순히 문장에서 실현된 순서에 따라 명사들의 서열을 정하는 것이 한국어의 특성상 가장 간단하면서도 효율적임을 입증하였다.

향후 연구로는, 텍스트 구조화를 위해 제안된 2가지 정책을 실제 텍스트 생성 시스템이나 문서 요약 시스템에 적용해서 그 효과를 입증하고 어떤 것이 더욱 효과적인지 비교해 볼 필요가 있다. 또 2가지 정책이 기존의 MIN.NOCB만을 이용하는 것보다 얼마나 더 효과적인지 판단하기 위해 텍스트 구조화를 통해 정해진 문장 순서와 실제의 문장 순서를 비교하는 [32]의 방법론을 적용해 볼 수 있을 것이다. 또한 6.3절에서 논의된 단어 순서에 의한 서열화 방식의 문제점을 해결하기 위한 보완책이 필요하다.

참 고 문 헌

[1] Grosz, B.J., Joshi, A.K., and Weinstein, S., "Centering: a framework for modeling the local coherence of discourse," Proc. Computational Linguistics 21(2): 203-225, 1995.

[2] Karamanis, N., Poesio, M., Mellish, C., and Oberlander, J., "Evaluating Centering. based metrics of coherence using a reliably annotated corpus," Proc. the 38th Annual Meeting of the Association for Computational Linguistics, pp.391-398, 2004a.

[3] Karamanis, N., Mellish, C., Oberlander, J., and Poesio, M., "A Corpus-based Methodology for Evaluating Metrics of Coherence for Text Structuring," Proc. of INLG-04, pp. 90-99. Brockenhurst, UK, 2004b.

[4] Kibble, R., and Power, R., "An integrated framework for text planning and pronominalization," Proc. 1st International Natural Language Generation, Mitzpe Ramon, Israel, pp.77-84, 2000.

[5] Miltsakaki, E., and Kukich, K., "The role of Centering theory's rough shift in the teaching and evaluation of writing skills," Proc. the 38th Annual Meeting of the Associ-ation for Computational Linguistics, Hong Kong, pp. 408-415, 2000.

[6] V. Mittal, J. Moore, G. Carenini, and S. Roth, "Describing complex charts in natural language: a caption generation system," Proc. Computational Linguistics, Special issue on Natural Language Generation, vol.24, no.3, pp.431-467, 1998.

[7] Roh, J.E., and Lee, J.H., "An empirical study for generating zero pronoun in Korean based on Cost-based Centering Model," Proc. Australasian Language Technology Association, Melbourne, Australia, pp.90-97, 2003.

[8] Kim, M.K., "Conditions on deletion in Korean based on information packaging," Proc. Discourse and Cognition, 1(2): 61-88, 1999.

[9] Kim, M.K., "Zero vs. overt NPs in Korean discourse: a centering analysis," Korean Journal of Linguistics, 28(1): 29-49, 2003.

[10] Kim, M.Y., "The centering of Korean discourse," Seoul National University, M.S. Thesis, 1994.

[11] Ryu, B.R., "Centering and zero anaphora in the Korean discourse," Seoul National University, M.S. Thesis, 2001.

[12] Walker, M., Iida, M., and Cote, S., "Japanese discourse and the process of centering," Proc. Computational Linguistics, 20(2): 193-232, 1994.

[13] Kameyama, M., "Intra sentential centering: a case study," In Walker, M.A., Joshi, A.K., and Prince, E.F., editors, Centering Theory in Discourse, chapter 6, pp.89-112, Oxford, 1998.

[14] Hudson, S.D., and Tanenhaus, M.K., "Assigning antecedents to ambiguous pronouns: The role of the center of attention as the default assignment," In M.A. Walker, A.K. Joshi, and E.F. Prince, editors, Centering in Discourse. Oxford University Press, pp.199-226, 1998.

[15] Eugenio, B. D., "Centering in Italian," In Walker, M.A., Joshi, A.K., and Prince, E.F., editors, Centering Theory in Discourse, chapter 7, pp. 115.138, Oxford, 1998.

[16] Hoffman, B., "The Computational Analysis of the Syntax and Discourse Use of Free Word order in Turkish," University of Pennsylvania, Ph.D. dissertation, 1995.

[17] Gordon, P.C., Grosz, B.J. and Gillion, L.A. "Pronouns, names, and the centering of attention in discourse," Proc. Cognitive Science, 17: 311-348, 1993.

[18] Brennan, S.E., Friedman, M.W. and Pollard. C.J. "A centering approach to pronouns," Proc. The 25th Annual Meeting of the Association for

Computational Linguistics, pp.155-162, 1987.
[19] Cote, S., "Ranking forward looking centers," In M.A. Walker, A.K. Joshi, and E.F. Prince, editors, Centering Theory in Discourse. Oxford, chapter 4, pp. 55-70, 1998.
[20] Rambow, O., "Pragmatics aspects of scrambling and topicalization in German. Proc. Workshop on Centering Theory in Naturally Occurring Discourse," Philadelphia. Insti-tute for Research in Cognitive Science (IRCS), 1993.
[21] Strube, M., and Hahn, U., "Functional centering: grounding referential coherence in information structure," Proc. Computational Linguistics, 25(3): 309-344, 1999.
[22] Roh, J.E., and Lee, J.H., "Generation of natural referring expressions by syntactic information and Cost-based Centering Model," Journal of KISS: Software and Applications, vol.21, no.12, pp.1649-1659, 2004.
[23] Byron, D., and Stent, A., "A preliminary model of centering in dialog," Proc. 36th Annual Meeting of the Association for Computational Linguistics, Montreal, Canada, pp. 1475-1477, 1998.
[24] Passonneau, R.J., "Getting and keeping the center of attention," In Bates, M. and Weischedel, R.R., editors, Challenges in Natural Language Processing, Cambridge Uni-versity Press, pp.179-227, 1993.
[25] Poesio, M., Stevenson, R., Cheng, H., Eugenio, B.D., and Hitzeman, J., "Centering: a parametric theory and its instantiations," Proc. Computational Linguistics, 30(3): 309-363, 2004.
[26] Tetreault, J.R., "A corpus based evaluation of centering and pronoun resolution," Proc. Computational Linguistics, 2(4): 507-520, 2001.
[27] Kim, M.Y., "An optimality approach to the referential interpretation of zero anaphors in Korean," Seoul National University, PhD. Thesis, 2003.
[28] Viterbi, A. J., "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," IEEE Trans. Information Theory, IT-13:260-269, 1967.
[29] Bak, S.Y., "Topic in Korean discourse, Korean Journal of Linguistics," 11(2): 1-15, 1986.
[30] Yang, D.W., "Topicalization and Relativization in Korean," Pan Korean Book Cor., 1975.
[31] Chae, W., "Meaning of topic marker -nun, Korean Journal of Linguistics," 4: 93-111, 1976.
[32] Lapata, M., "Probabilistic text structuring: Experiments with sentence ordering," Proc. the 37th Annual Meeting of the Association for Computational Linguistics., pp.545-552, 2003.

노 지 은
2000년 2월 부산대학교 컴퓨터공학과 학사. 2006년 2월 포항공과대학교 컴퓨터공학과 박사. 2006년 3월~국방과학연구소 선임연구원. 관심분야는 텍스트 생성, 기계 번역, 자연언어처리, 한국어처리

나 승 훈
2001년 2월 아주대학교 컴퓨터공학과 학사. 2003년 2월 포항공과대학교 컴퓨터공학과 석사. 2003년 3월~포항공과대학교 컴퓨터공학과 박사과정. 관심분야는 정보검색, 자연언어처리, 한국어처리, 기계 번역

이 종 혁
정보과학회논문지 : 소프트웨어 및 응용
제 34 권 제 2 호 참조