자연스러운 텍스트 생성을 위한 추계적 텍스트 구조화

노지은[0] 이종혁
포항공과대학교 전자컴퓨터공학부 컴퓨터공학과, 첨단정보기술 연구센터
{jeroh, jhlee}@postech.ac.kr

A Stochastic Text Structuring using Simulated Annealing

Ji-Eun Roh[0] Jong-Hyeok Lee
Dep. of Computer Science and Engineering,
Div. of Electrical and Computer Engineering
Pohang University of Science and Technology
and Advanced Information Technology Research Center(AITrc)

요 약

언어가 아닌 다양한 지식원으로부터 그것을 설명하는 텍스트를 생성하는 텍스트 생성 (text generation)은 여러 가지 복합적이고 단계적인 과정을 거쳐 이루어진다. 자연스러운 텍스트를 생성하기 위한 여러 단계 중, 지식원으로부터 텍스트에 포함되기 위해 뽑힌 정보들간의 순서를 적절히 결정하는 과정을 텍스트 구조화(text structuring)라고 한다. 텍스트 구조화는 생성될 텍스트의 결속성(coherence)을 크게 좌우하므로, 양질의 텍스트를 생성하기 위해서는 텍스트 구조화를 다루기 위한 정교한 방법론이 요구된다. 본 논문에서는 SA(simulated annealing) 알고리즘을 이용해 추계적 텍스트 구조화 방안을 제안하며 특히, SA의 평가 함수(evaluation function)로서, 총 4가지의 방법론-중심화 이론(centering theory)을 이용한 센터 전이 유형의 선호도, 추론 비용에 근거한 전이 유형간의 선호도, 서두 문장을 결정하기 위한 가중치 할당에 따른 선호도, 인접한 문장간의 유사도에 따른 선호도-을 제안하고 실험을 통해, 그 효용성을 보였다.

1. 서론

자연어의 생성(natural language generation)은 자연어로 이루어지지 않은 기저의 정보들을 자연어로 사상하는 언어 처리의 한 분야로 일반적으로 두 가지의 범주로 나누어 진다. 기계 번역에 기반한 단문장의 생성(sentence generation)과 텍스트의 생성(text generation)이 바로 그것이다. 특히 텍스트의 생성은 여러 가지 측면에서 단문장의 생성과는 서로 다른 논점을 갖는다. 여러 문장이 긴밀히 결합되어 하나의 정보를 전달하는 단위를 텍스트라 볼 때, 높은 질의 텍스트를 생성하기 위해서는 문장간의 순서, 문장간의 결합, 각 문장들에서의 지시어 생성 등을 적절히 처리해 주어야 한다. 우리 나라에서는 한국어를 대상으로 한 텍스트 생성에 관한 연구가 거의 이루어 지지 않았으며, 노지은[2001]에서 제안한, 데이터베이스로부터 홈쇼핑 사이트의 각 상품 소개를 위한 텍스트 생성 시스템(XExplainer)이 그 첫 시도라 할 수 있다.

본 논문에서는 텍스트 생성을 위한 여러 가지 과정 중, 지식원에서 뽑혀진 정보를 대상으로 그것들 간의 순서를 정하는 텍스트 구조화의 효율적인 처리 방안을 제안하고자 하며, 최종적인 목표는 제안된 방법이 XExplainer 시스템에 효과적으로 적용되어 보다 낳은 양질의 텍스트를 생성하는 것이라 할 수 있다.

2. 관련 연구

텍스트 구조화는 최종적으로 생성된 텍스트의 관점에서 볼 때 문장의 순서를 결정하는 과정이라 볼 수 있고, 문장 순서의 결정은 생성될 텍스트의 결속성(coherence)을 크게 좌우한다. 담화 이론(discourse theory)에서, 오래 전부터 텍스트의 결속성에 관한 연구를 해 왔으며, 특히 문장 순서/배열의 관점에서 그 성질을 규명하고자 많은 노력이 있었다.

담화 이론에서 '결속력 있는 텍스트(coherent text)'를 정의하는 두 가지의 주된 견해 중, 그 첫번째는 텍스트의 모든 인접한 문장간에 결속 관계(coherent relation)가 존재해야 한다는 것이고, 다른 하나는, 텍스트를 이루고 있는 문장들간의 화제(topic)가 자연스럽게 유지되고 있다면 그 텍스트는 결속력이 있다고 보는 견해이다. 첫번째 견해는 텍스트를 구성하는 인접한 문장간의 관계에 초점을 두고 있는 반면, 두번째 견해는 문장이 대상이 아닌, 문장 내의 화제의 변화에 초점이 있다고 볼 수 있다. 텍스트의 결속성을 규명하는 이런 담화 이론을 토대로, 텍스트 구조화를 처리하기 위한 주요 이론들, RST[Mann88], Schema[McKeown85], CT[Grosz86,95]가 제안되어져 왔다. RST와 Schema는 첫

번째 견해에, CT는 두번째 견해에 이론적 토대를 두고 있다.

RST(Rhetorical Structure Theory) 총 23개의 문장간의 수사 관계(rhetorical relation)를 정의하고, 텍스트를 이루는 인접한 문장간의 관계가 23개의 수사 관계로 분석되느냐에 의해 텍스트의 결속성이 결정된다는 것이 핵심 내용이다. 텍스트 구조화 측면에서는 인접한 문장간의 수사 관계가, 정의된 23개 중의 하나에 적용되도록 문장 순서를 조정하게 된다. RST는 도메인에 독립적으로 적용 가능하므로 많은 텍스트 생성 시스템에서 RST를 이용해 텍스트 구조화를 처리하고 있지만, 수사 관계의 표준적 정의가 어렵고, 문장들 간의 수사 관계가 미리 정의되어져 있어야 하는 부담이 따른다. 또한, RSTree[Marcu96]를 만드는 것은 탐색 문제(search problem)로 시간 복잡도가 크다는 단점이 있다.

스키마 방식은 McKeown[85]의 TEXT 시스템에서 텍스트 구조화를 위해 제안된 것으로, 특정 도메인의 텍스트는 일정한 구조를 가지며, 스키마는 이런 텍스트 구조를 수사 술부(Rhetorical Predicate)를 이용해 정의해 놓은 하나의 틀이라고 볼 수 있다. 이 방식은 도메인에 의존적이지만, RST에 비해 비교적 간단하고 빠르며, 또 예측 가능한(deterministic) 생성 결과를 내므로 RST와 더불어 많은 텍스트 생성 시스템에서 스키마에 기반한 텍스트 구조화를 처리하고 있다.

마지막으로, CT (Centering Theory)는 Grosz와 Sidner의 연구에서 유래했고[Grosz86,95], 추론(inference)을 통제하기 위해 담화 내의 *attentional state*를 모형화 하는데 목적이 있다. CT는 자연어 해석(natural language interpretation) 특히, 지시어 처리(anaphora resolution)에 주로 이용되고 있으며 최근, 자연어 생성 연구자들은 CT를 텍스트 구조화[Cheng 2000], 문장 단위 계획(Sentence Planning)[Mittal98], 지시어 생성[Dale92]등 텍스트 생성의 여러 과정에 적용하여 성과를 거두고 있다. 본 논문에서는, 텍스트 구조화를 위한 SA의 평가 함수로써 CT의 적용을 제안한다.

3. 추계적 텍스트 구조화

기존의 텍스트 구조화를 처리하는 방법은, 관련 연구에서 언급된 대표적인 3가지 방식(특히, RST와 스키마 방식)등을 이용, top-down 방식의 규칙 기반 텍스트 구조화가 주로 제안되었다. 하지만, 최근 들어 추계적 텍스트 구조화(stochastic text structuring)가 몇몇 연구에 의해 시도되고 있다[Mellish98][Karamanis2002][Cheng2000][Pablo2002]. 텍스트 구조화가 최적의 문장 순서를 결정하는 일이라고 볼 때, 검색 문제라고 할 수 있으며, 추계적인 방법은 검색 공간을 크게 줄이고, 완전한 최적의 해답(global optimum)을 찾는다는 것을 보장할 수는 없지만, 임의의 시점에서 그 시점까지 발견되어진 최적의 해답을 찾을 수 있다는 큰 장점을 지닌다[Mellish98]. 특히 이것은 제한된 시간 내에 양질의 텍스트를 생성해야 하는 텍스트 생성 시스템의 특성상 아주 긍정적인 요소라고 할 수 있다. 추계적 검색(stochastic search)에 이용되는 최적화 기법으로 Hill-Climbing, SA(Simulated Annealing), GA(Genetic Algorithm)등이 있으며 기존 연구 [Mellish98], [Karamanis2002], [Cheng2000]에서는 텍스트 구조화를 위해 GA를 적용하고 있다. 본 논문에서는 SA를 이용한 추계적 텍스트 구조화 방안과, SA 적용에 필요한 여러 가지 평가 함수를 제안한다.

3.1 SA(Simulated Annealing)를 이용한 텍스트 구조화의 처리

SA는 물리학에 개념적 바탕을 두고 있으며, 어떤 물질의 온도를 서서히 낮추면 그 물질은 안정화 상태로 들어간다는 기본 원리에 따라, 엔트로피의 크기(degree of randomness)를 제어하는 인자(여기서는 온도)를 이용해 최상의 솔루션을 찾아내는 최적화 기법이다.

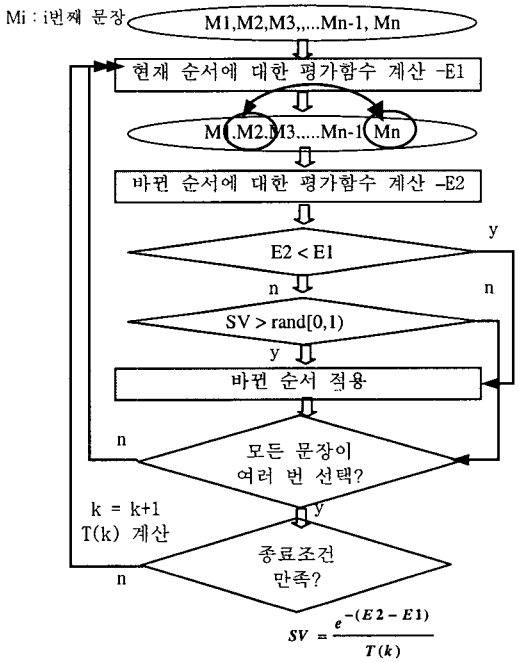


그림 1. SA를 이용한 텍스트 구조화

본 논문에서는, 그림1 처리도를 따라 SA를 텍스트 구조화에 적용한다. 텍스트 구조화의 입력값은 문장 순서가 결정되지 않은, 즉 순서가 없는 문장(그림 1에서 M으로 표현)들의 집합이다. (여기서는 이해의 편의를 위해 '문장' 이라고 표현했지만, 실제 텍스트 구조화의 입력들은 완성된 하나의 문장이 아니라, 내용 선택 과정에서 뽑혀진 가장 기초적인 정보의 단위들이다. 일반

적으로 fact, message라고 일컬으며, 노지은[2001]에서는 BIU라고 부르기도 했다.) 초기화 과정으로 각 문장들에 대해 임의로 순서를 부여하고 현재 부여된 순서에 대해 순서의 적절성을 판단하는 평가 함수(evaluation function)를 적용하여 E1값을 얻는다. 다음, 임의의 두 개의 문장을 선택해 그 둘의 순서를 바꾸고 E2를 계산하여 바꾸기 전의 것과 비교해서 값이 크면(즉, 순서를 바꾸는 것이 더 좋다고 판단되면) 순서를 바꾸고, 그렇지 않으면 확률적으로 바꿀 것인지를 결정한다. 즉, 현재 시점에서 순서를 바꾸지 않는 것이 좋다고 판단되더라도 무조건 바꾸지 않는 것이 아니라 SV(stochastic value)를 이용해 확률적으로 결정하게 되는데, 이런 확률성 때문에 SSA(Stochastic Simulated Annealing)라고 부른다. 이 과정을 모든 문장이 임의적으로 여러번 선택될 때까지 반복한 후, 온도(T(k)는 감소 함수)를 조금 낮춰서 다시 반복(annealing 과정), 종료 조건을 만족하면 종료하게 된다. T(k)가 높은 값을 가질 때(초기 annealing 과정)는 SV가 거의 1이 되어 문장 순서가 활발히 바뀌게 되며 안정화 단계에 들어 갈수록 확률적으로 바뀌게 되는 경우가 줄어든다. 이러한 확률성을 통해 지역적 최소값(local minima)에 빠지는 것을 막는다. 이렇게 텍스트 구조화에 SA를 적용하는데 있어 가장 큰 관건은 어떻게 문장 순서의 자연스러움 또는 적합성을 판단할 것인가, 즉, 문장들의 순서가 주어졌을 때 그 순서의 적합성을 판단할 평가 함수를 찾는 것이라 할 수 있다.

본 논문에서는 평가 함수로 다음 4가지 방법론을 제안한다.

◦방법론 1. CTP(Center Transition Preference)
➔ CT의 전이 유형의 선호도
◦방법론 2. ICP(Inference Cost Preference)
➔ 추론 비용에 기반한 전이 유형간의 선호도
◦방법론 3. SM(Similarity Maximization)
➔ 문장간의 유사도를 최대화
◦방법론 4. SWP(Sentence Weight Preference)
➔ 문장에 대한 가중치 부여를 통한 서두 문장의 결정

4. CTP(Center Transition Preference)

CTP는 CT(Centering Theory)의 센터 전이 유형의 선호도(제약2)를 이용한 것으로, CT는 각 발화에 대해 Cf(forward looking center)라는 담화 요소들과, 이 Cf 내에 Cb(backward looking center)라는 특별한 요소를 정의하고 다음의 3가지 제약과 2가지 규칙으로 이루어진다.

▣ 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 발화에서 실현된, 바로 전 발화의 Cf에서 가장 높은 순위의 담화 요소이다.

▣ 규칙(rules)
1. 앞 발화의 Cf의 어떤 요소가 현 발화에서 대명사화 되었다면, 현 발화의 Cb도 역시 대명사화 된다.
2. 발화간의 전이유형은 다음 순서로 선호된다.
continue > retain > smooth-shift > rough-shift

Cb는 하나의 발화에 있어서 가장 중점적인 담화 요소이고 현 발화에 선행하는 담화에 의해 결정된다(제약2). Cf목록의 요소들은 일정한 기준에 의해서 순위가 매겨지는데 그들 중 가장 우선 순위의 담화 요소가 Cp(preferred center)로 설정되며, 이 Cp가 Cf목록의 요소 중 다음 발화의 Cb가 될 가능성이 가장 높다는 것이 CT의 기본 전제이다. 발화간의 전이 유형은 다음과 같이 결정된다.

표 1. 발화간의 전이 유형

	$Cb(U_i) = Cb(U_{i-1})$	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i)=Cp(U_i)$	continue	smooth-shift
$Cb(U_i)\neq Cp(U_i)$	retain	rough-shift

표1에서 $Cb(U_i)$는 현 발화의 Cb, $Cb(U_{i-1})$는 바로 전 발화의 Cb, $Cp(U_i)$는 현 발화의 Cp를 나타낸다.

CT는 그 이론 자체의 단순함에도 불구하고 자연어 처리의 여러 분야에 응용되면서 다양한 이슈들을 만들어 내고 있는데 특히, Cf목록의 순위 결정은 언어에 따라 조금씩 다르며, 같은 언어에 대해서도 학자들마다 기준이 다르다.

한국어의 텍스트 구조화에 CT를 이용하기 위해서는, 한국어에 맞는, 또 도메인에 적합한 Cf 목록간의 우선 순위를 정하는 작업이 선행되어야 한다. 지금까지는 Cf 목록 순위 결정의 척도로써 구문 관계(syntactic role)가 가장 적합하다고 보고 있으며 영어에서 일반적으로 사용되는 Cf 목록의 우선 순위는 다음과 같다[Strube96].

주어 > 직·목 > 간·목 > 보어 > 수식어구 (영어)

김미영[94]에서는 한국어와 언어적인 특성이 유사한 일본어를 대상으로 제안된 Cf 목록 순서를 한국어에 그대로 적용할 수 있는지를 검토하고 다음과 같은 Cf목록의 순서를 제안하였다.

주제 > ID > 주어 > 직·목 > 간·목 > 그외 (한국어)

ID(Walker[90]에서 제안한 Empathy와 유사한 개념)는 화자가 자동적으로 지시체의 관점과 동일시하는 논항 위치로서, 이동이 있는 사건을 기술할 때 화자의 관점을 중시하느냐, 청자의 관점을 중시하느냐에 따라 특정 문장 성분에 감정 이입이 들어가는 현상을 반영한다. 김미영[94]은 우리말에 적용될 수 있는 대표적인 감정 이입 동사로 '오다', '가다'를 들고 있으며, '오다'는 간접 목적어에, '가다'는 주어에 감정 이입이 된다고 본다. 차건희[97]에서는 일본어의 Cf목록 순서를 한국어에 그대로 적용 하는 것과, 미리 고정된 Cf의 우선 순위를 정해 놓는 것이 문제점이 있다고 지적하고, 베이지안 확률을 이용해 Cf목록의 우선 순위를 확률적

으로 결정하는 방법을 제안하였다. 이 방법이 잘 적용되기 위해서는 $P(A/H_k)$값, 즉 증후(symptom)를 잘 찾아야 하지만 차건희[97]에서는 단 3개의 증후만을 언급함으로써, 실제 이 방법이 얼마나 효율적으로 적용될 수 있을지는 확인되지 않았으며, $P(H_k)$를 결정하는 것도 쉽지 않다.

본 논문에서는 김미영[94]이 제안한 내용에 한국어의 특징을 반영한 것(PV)과 도메인 의존적인 내용을 반영(DAE)해 Cf 목록의 순위를 다음과 같이 정의하였다.

주제 > ID> 주어 > DAE(directly-associate-entity) > 직·목 > 간·목 > 그외 > PV(pre-verbal)

ID를 포함시켰지만 우리말의 동사에서 ID를 갖는 동사는 '오다', '가다' 외에는 찾기 어려우며, 또 다른 동사가 있다 하더라도 각 동사마다 어떤 구문 관계에 ID가 있는지 자동적으로 결정하기 어렵기 때문에, 실질적으로 큰 역할은 하지 않는다. Cf 목록의 최하위 순위로 PV(Pre-Verbal)를 추가하였는데, 우리말의 화제와 초점에 관한 대부분의 연구에서 초점은 동사 앞에 위치하는 것이 가장 자연스럽다는데 거의 의견의 일치를 보고 있는 듯 하다[한나래97][박철우98]. 화제(given information, theme, topic, presupposition)는 이미 우리가 알고 있는 정보이므로 Cb에 대응된다고 볼 수 있고, 초점(new information, rheme, comment,focus)은 한 문장 내의 새로운 정보이므로 하나의 발화에서 Cb가 아닌 나머지 부분에 대응된다고 볼 수 있다. 이런 맥락에서, 초점은 동사 앞에 오는 것이 가장 자연스러우므로 역으로 동사 앞에 오는 성분, 즉 PV가 가장 Cb가 될 확률이 낮다고 결론 지을 수 있다. 또, DAE라는 새로운 Cf 목록을 추가하였는데 앞서 언급한 것처럼, 본 논문에서 제안하는 텍스트 구조화 방안은 최종적으로 노지은[2001]에서 제안한 텍스트 생성 시스템(XExplainer)에 적용될 것이다. XExplainer는 홈쇼핑 사이트의 DB로부터 각 상품에 대한 설명 및 묘사를 목적으로 한 텍스트를 생성한다. 이렇게 특정 대상(target)을 설명하고 묘사하는 텍스트에는 다음과 같은 문형 패턴이 많이 쓰이고, 이때 DAE는 설명하고자 하는 대상과 아주 밀접한 관련이 있는 또 다른 대상(entity)을 나타낸다.

- target은 DAE이다.
- target에는 DAE_1,$DAE2_2$,,DAE_n가 있다.
- target은 DAE_1,DAE_2,,DAE_n으로 구성되어 있다….

이런 DAE들이 그 다음 문장의 Cb가 됨을 많은 경우를 통해 확인할 수 있었으므로, DAE를 Cf목록에서 주어 다음에 위치시켰다.

본 논문에서 적용하는 전이 유형의 선호도(CTP)는 규칙 2에 Cheng[2001]에서 제안된 새로운 전이 유형을 첨가하여 다음과 같이 정의하였다.

continue(c)> associate-shift(as)> retain(r)> smooth-shift(ss) > rough-shift(rs) > resume(re) > no-Cb(n)

특정 대상에 대한 묘사 및 설명을 위한 목적의 텍스트에서는 그 대상에 대한 묘사, 설명으로 텍스트를 시작한 후, 그 대상과 밀접한 관련이 있는(본 논문에서 정의한 DAE같은) 대상으로 설명을 옮겨 가는 경우가 빈번히 발생한다. Cheng[2001] 역시 그런 점을 지적하고 그런 전이 유형을 associate-shift라 정의하여, associate-shift가 continue 다음으로 선호된다는 것을 밝혔다. 또, Oberlander[99]가 정의한 resume을 추가했는데 이것은 현재의 담화가 바로 직전의 담화에서는 언급되지 않았지만 그 전의 담화들에서 언급된 대상에 대해 진술하고 있을 때 적용되는 것으로 역시 묘사,설명문에서 빈번히 발생하는 전이 유형이라 할 수 있다.

표 2. 전이 유형의 선호도에 따른 점수

전이 유형	c	as	r	ss	rs	re	n
점수	6	5	4	3	2	1	0

이렇게 우리말의 특성과 도메인의 특성을 고려하여 Cf 목록의 순위와 전이 유형의 선호도를 새롭게 정의한 것을 바탕으로 CTP가 어떻게 적절한 문장 순서 결정에 도움을 주는지 다음 예를 통해 살펴보자.이는, 선호도가 높은 전이 유형이 많이 발생할 수록 더 좋은 텍스트라는 기본적인 가정에서 출발한다.

a) 아스피린은 백색의 결정성 분말이다.
b) 아스피린은 바이엘사에서 발명되었다.
c) 아스피린은 해열제, 진통제, 항류머티즘제로 쓴다.
d) 아스피린은 살리신산과 아세트산을 포함하고 있다
e) 왜냐면 살리신산은 해열,진통의 작용이 있기 때문이다.

a,b,c,d,e의 5문장이 순서 없이 주어 졌을 때, 다음과 같이 4개의 플랜이 제시될 경우,

(plan 1) a → b → c → d → e (c->c->c->ss) 21
(plan 2) a → d → b → c → e (c->c->c->n) 18
(plan 3) a → d → b → e → c (c->c->n->n) 12
(plan 4) a → b → e → d → c (c->n->rs->ss) 11

각 문장 순서에 대한 전이 유형을 구할 수 있고, 표2를 이용해 각 순서에 대한 평가를 할 수 있다. 이때, 플랜 1이 가장 높은 점수를 얻었으므로 CTP에 의해 가장 자연스러운 순서라고 판단하게 되며, 그것은 실제로도 그러하다. 따라서, 본 논문에서는 SA 적용을 위한 평가함수의 첫번째 방법론으로 CTP에 따라 부여된 점수표(표2)를 이용, 주어진 문장의 자연스러움을 판단하고자 한다.

CT을 이용해 텍스트 구조화를 시도한 모든 기존 연구들은, 본 논문의 방법론1과 같이 전이 유형의 선호도에 따라 선호도가 높은 것이 많이 나타날수록 좋다는 기본적인 가정에 충실했다. 하지만 이것만으로는 문장 순서를 판단하는 것이 충분하지 않은데, Strube[96]는, 그것이 잘 지켜지지 않아도 좋은 텍스트 일 수 있고, 역

으로 잘 지켜져도 어떤 경우에는 그다지 좋지 않은 텍스트 일 수 있다는 것을 지적했다. 또, 각각의 전이를 독립된 것으로 볼 것이 아니라 전이간의 상호 관계에 의존해, 특정 전이 다음에는 특정 전이가 오는 것이 더 자연스럽다는 사실을 주장했다. 따라서 본 논문에서는 Strube[96]의 제안을 부분적으로 받아들여, SA 적용을 위한 평가 함수의 두번째 방법론으로, CT의 기본적인 전이 유형의 선호도를 지키되 부족한 부분을 보완하고자 추론 비용에 기반한 상호 전이 유형간의 선호도를 제안하고자 한다.

5. ICP(Inference Cost Preference)

Strube[96]은 좋은 텍스트가 되기 위해, 항상 어떤 특정 전이 유형이 다른 것보다 더 선호되는 것이 아니라, 어떤 전이 유형간의 쌍(centering transition pairs)이 다른 쌍보다 더 선호된다는 사실을 주장했다. 선호되는 전이 유형 쌍이 실현될 경우, 담화가 훨씬 자연스럽고 이해하기가 쉬워, 이를 '추론 비용이 낮다(cheap)' 고 보고, 여러 개의 텍스트를 대상으로 각 전이 유형간의 비용(cost)을 구했다. 본 논문에서는 그것들 중 추론 비용이 낮은 전이유형 쌍을 일부 도입하여 우리말에 적용 가능한지를 검토하고, 확장하였다(표 3).

표 3. 추론 비용이 낮은 전이 유형 쌍

Strube (cheap)	추가 (cheap)
(continue, continue) **(retain, smooth shift)** (continue, retain) (smooth shift, continue) (rough shift, smooth shift)	(smooth shift, associate shift) (associate shift, associate shift)

아래의 예는 표3에 보이는 추론 비용이 낮은 쌍이 실현된 문장들로, 그것의 순서가 자연스러워 한국어에도 잘 적용될 수 있음을 보여 주며, 바람직한 문장 순서의 결정에도 도움을 준다는 것을 알 수 있다.

• 적용된 전이 유형 쌍 : (c,r),(r,ss),(ss,c)
a)철수는 여자 친구가 많다.
b)철수는 그 중 영희를 좋아한다.(c) Cb:철수
c)영희는 철수의 동창이다.(r) Cb:철수
d)영희는 얼굴이 매우 희고 예쁘다.(ss) Cb:영희
e)영희는 공부도 잘 한다 (c) Cb:영희
• 적용된 전이 유형 쌍 : (rs,ss)
a)철수는 여자 친구가 많다.
b)철수는 그 중 영희를 좋아한다.(c) Cb:철수
c)민수도 영희를 좋아한다.(rs) Cb: 영희
d)민수는 철수와 동창이다.(ss) Cb: 민수
• 적용된 전이 유형 쌍 : (ss,as),(as,as)
a)철수는 방학 동안 미국과 캐나다를 여행했다.
b)미국은 본토 48개의 주와 알래스카, 하와이 2개주로 구성된 연방 공화국이다.(c) Cb:미국
c)하와이는 아름다운 섬으로~~.(ss) Cb: 하와이
d)알래스카는~~ (as) no Cb
e)캐나다는~~~ (as) no Cb

텍스트 구조화에서 ICP(inference cost preference)를 이용해 CTP에 의한 방법론을 어떻게 보완할 수 있는지 다음 예를 통해 살펴보자.

a)철수는 오늘 휴가를 냈다.
b)철수는 영희에게 전화를 했다.(c) Cb:철수
c)철수는 영희와 여행을 가고 싶었다.(c) Cb:철수
d)영희는 철수의 전화가 반갑지 않았다.(r) Cb:철수
e)영희는 철수의 제의에 묵묵 부답이었다.(ss) Cb:영희
➔구조화 A의 CTP 점수: 19 + α((r,ss)쌍에 의한 ICP 점수)

a)철수는 오늘 휴가를 냈다.
b)철수는 영희에게 전화를 했다.(c) Cb:철수
d)영희는 철수의 전화가 반갑지 않았다.(r) Cb:철수
c)철수는 영희와 여행을 가고 싶었다.(c) Cb:철수
e)영희는 철수의 제의에 묵묵 부답이었다.(r) Cb:철수
➔구조화 B의 CTP 점수: 20

위 두개의 구조화 결과를 비교해 볼 때, CTP에만 의존하면 텍스트 B가 더 낫다고 판단하게 되지만, 'retain 다음에는 smooth-shift가 오는 것이 추론 비용이 낮다' 라는 ICP에 근거해, 텍스트 A가 더 자연스럽다고 평가할 수 있다. 그러나 이렇게 CTP와 ICP를 동시에 적용할 경우, 때로는 두개의 방법론이 충돌할 수 있고, 이를 적절히 해결하고자 하는 노력이 필요하다. 이 문제는 뒤에서 다시 언급하겠다.

6. SM(Similarity Maximization)

일반적으로 사람들이 글쓰기를 할 때, 어휘적으로 유사해 비슷한 내용인 문장을 인접하게 두기도 하고, 구조적으로 유사한 문장을 병렬식으로 인접해서 나열하기도 한다. 이런 기본적인 글쓰기 원칙을 텍스트 구조화에 적용하여, 인접한 문장간의 유사도를 크게 하는 문장 순서를 선호하고자 하는 것이 평가 함수의 3번째 방법론 SM(Similarity Maximization)이다. tf*idf를 이용해 용어에 대한 가중치를 계산하고, 인접한 문장간의 유사도를 구하기 위해 코사인 유사도(cosine similarity)를 적용하였다.

본 논문에서는, 각 문장을 구성하는 어휘 수준에서의 유사도만을 다루었지만, 텍스트 구조화에 영향을 미치는 것은 어휘 레벨 뿐만 아니라 구문 구조, 서법, 태 등도 포함되므로 고려 범주에 두어 확장할 수도 있겠다.

7. SWP(Sentence Weight Preference)

일반적인 텍스트 구성을 살펴보면, 특히 어떤 대상을 묘사, 설명하는 텍스트에서 서두에 위치하는 문장들의 특징을 파악할 수 있다. 예를 들어, 설명하고자 하는 대상을 정의하는 정의문이 온다든지, 그 대상의 또 다

른 명칭을 소개한다든지, 그 대상과 밀접이 관련이 있는 다른 대상(DAE)을 텍스트의 후반부에 상세히 설명하기 위해 먼저 간단히 언급할 수 있다. 본 논문에서는 이렇게 텍스트 내에서 일반적으로 앞쪽에 위치하는 정보들을 중요하다고 보고 이런 문장을 텍스트의 앞쪽에 위치시키고자 각 문장에 가중치를 부여하는 방법을 제안한다. 즉, 각 문장에 가중치를 부여해서 높은 가중치를 갖는 문장일수록 중요하다고 여기고 이를 텍스트의 앞쪽에 위치시키는 문장 순서를 선호하고자 하는 것이 평가 함수의 4번째 방법론, SWP(sentence weight preference)이다. 문장에 가중치를 부여하기 위한 휴리스틱은 다음과 같다.

(가정) 중요한 문장일수록 앞쪽에 위치한다.

H1.텍스트 내에서 자주 언급되는 용어(FME)들을 많이 포함하고 있는 문장일수록 더욱 중요하다.
H2.텍스트 내에서 자주 언급되는 용어들을 현저하게 (salient) 실현하는 문장일수록 더욱 중요하다.
H3.특정 역할(role)을 갖는 문장들은 다른 문장과는 상관없이 텍스트의 앞쪽에 위치하는 것이 좋다.
H4.어떤 문장들은 다른 문장의 위치에 영향을 받는다.

H1은 보다 포괄적이고 일반적인 문장을 앞쪽에 두자는 취지로, 텍스트 내에서 FME(Frequently Mentioned Entity)들을 많이 포함하고 있는 문장일수록 그 표현범위가 넓다고 본다. H2에서, FME가 그 문장에서 현저하게 표현된다는 의미는, FME가 Cf 목록 순위에서 상위에 위치한다는 의미로, 특정 대상을 설명하는 텍스트에서 그 대상을 문장의 주어나 주제로 삼고 있는 문장이, 그렇지 않은 문장보다 중요하다고 보는 것이다. H3의 적용을 위해 본 논문에서는 문장이 가질 수 있는 역할로 다음 3가지를 정의하였다.(표4)

표 4. 문장 역할

역할(flag)	문장 예
정의(DF)	씨앗그릇은 씨앗을 담기 위한 그릇이다.
다른이름소개(RF)	씨앗그릇은 씨앗망태라고도 부른다.
연관대상소개(AF)	씨앗그릇에는 다래끼, 종다래끼가 있다.

각 문장에 가중치를 부여하기 위해서는, 모든 문장을 대상으로 위 3 개 표지를 가질 수 있는지 미리 결정되어져 있어야 한다. 마지막으로 H4 는, AF 표지가 붙은 문장은 AF 에서 언급된 DAE 들을 주어 또는 주제로 갖는 문장보다 앞에 와야 하는 상황을 반영한다.

(1)식은 문장의 가중치를 구하기 위한 식으로 그 문장이 갖는 역할(r_w)의 가중치와 문장에 포함된 FME 의 가중치(fme_w), FME 의 Cf 목록의 순위의 가중치(gr_w)를 포함하고 있다. (2)는 문장의 3 가지 역할에 대한 가중치를 나타내며 H3 을 반영한다. (3)은 용어 가중치에 관한 식으로 H1 을 반영하고 있으며, (1)식에 gr_W 이 포함됨으로써 H2 를 반영하고 있다.

◦ $W_m = r_w + \sum(fme_w * gr_w)$ -(1)

◦ $r_w = DF*w_1 + RF*w_2 + AF*w_3$ -(2)

◦ if (tf > threshold) -(3) else

$$fme_w = \frac{tf}{\sum tf^2} * 1 \qquad fme_w = \frac{tf}{\sum tf^2} * (-1)$$

◦ gr_W : fme의 Cf 목록 순위에 부여된 가중치

정리하면, 문장의 가중치는 그 문장 자체의 성질에 의해 텍스트의 앞쪽에 위치할 수 있는 정도와, 문장 내에 FME들을 얼마나 많이 포함하고 있냐의 정도, 또 문장 내에서 FME가 얼마나 주제화 되어 표현 되느냐의 정도로 결정된다. 방법론 4 즉,SWP가 나머지 3개의 방법론과 구별되는 것은, 다른 방법론들은 전체 문장의 순서를 결정하는 것에 관여하지만 SWP는 전체 문장 중 단지 텍스트의 앞부분, 즉 서두에 위치할 문장을 결정하는 데만 영향을 미친다는 것이다.

8. 실험 및 평가

본 논문에서 제안한 내용을 검증하기 위해, 온라인 국립 민속 박물관 사이트 (http://www.nfm.go.kr/folk2002/)에서 각 전시물들을 설명하고 있는 300개의 텍스트를 모아서 각 텍스트를 SA의 입력형으로 바꾼 후, 본 논문에서 제안한 방식으로 텍스트 구조화를 시행해, 원문에서 문장의 순서와의 일치 정도를 평가의 기준으로 삼았다. 텍스트는 평균 13문장과 133개의 단어로 구성되어 있다. 먼저 텍스트의 각 문장을 구문 분석 과정과 별도의 전처리 과정을 거쳐 실험에 적용 가능한 형태, 즉 SA의 입력형으로 바꾼다. Cf 목록의 순위가 대부분 구문 관계에 의해 결정되기 때문에 포항공대 KLE연구실에서 개발 중인 구문분석기를 이용, 각 명사의 구문 관계를 획득하였다. 각 문장은 그 문장을 구성하고 있는 명사의 리스트로 표현되며 각 명사는 Cf 목록 중 하나의 값을 갖고, 그 중 가장 높은 순위를 갖는 명사를 Cp로 정하게 된다(그림 2). CT에서 발화의 단위에 대한 의견은 굉장히 분분하지만 본 논문에서는 간단히 한 문장을 하나의 발화로 정의하였다.

Index	DF	RF	AF	다래끼	그릇	씨앗
0	1	0	0	주어(Cp)	DAE	목적어

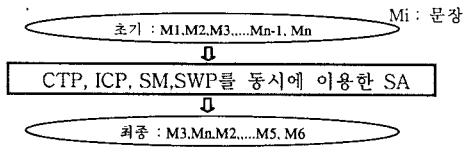
문장 : 다래끼는 씨앗을 담기 위한 그릇이다

그림 2. SA 입력형 예

첫번째 실험은, SA를 시행할 때, 제안된 4개의 방법론을 동시에 적용하여 텍스트를 구조화하는 것이다.(그림 3) 4가지 방법론을 SA에 동시에 적용할 때는, 각 방법론이 전체 텍스트 구조화에 미치는 중요도가 다르므로 각각에 대한 적절한 가중치 부여를 위해 여러 번의 실험을 반복, 최고의 성능을 내는 가중치를 적용하였다.

각 방법론을 따로 적용하여 텍스트 구조화를 시행 했을 때, SWP(0.3058), ICP(0.2726), SM(0.2354),

CTP(0.1979) 순으로 성능이 좋게 나왔다.



평가 함수: $CTP*w_1 + ICP*w_2 + SM*w_3 + SWP*w4$

그림 3. 평가 함수의 동시 적용

표 5에서 보듯이 각 방법론에 부여된 가중치도 같은 순으로 높게 할당 되었는데, 이 가중치를 이용해 각 방법론을 동시에 적용한 텍스트 구조화의 평균 성능은 0.4457로 기록되었다. 성능들이 기대한만큼 높지 않은 가장 큰 두 가지 이유는, 실험에 사용한 300개의 텍스트 자체의 문제와 각 방법론들의 충돌 현상 때문이라 볼 수 있다.

표 5. 각 방법론의 성능 평가

평가	동시적용					순차적용
		CTP	ICP	SM	SWP	SWP→CTP
	가중치	0.15	0.25	0.19	0.41	→ICP→SM
결과	0.4457 (41%~47%)					0.6487(60~68%)
	참고) %값은 입력 텍스트와의 유사도를 표현					

<< 씨앗그릇 >>
0. 씨앗을 담아놓는 그릇.
1. 씨앗 종자를 담아 보관하는 그릇으로서 씨앗망태, 씨앗통, 다래끼, 종다래끼 등이 있다.
2. 씨앗 그릇은 따로 만들어 사용하기보다는 씨앗을 오랫동안 상하게 하지 않고 ~~~~
3. 씨앗 그릇들 중 다래끼는 짚이나 싸리로 만들어 밭에 씨를 뿌릴 때 사용하고~~
4. 종다래끼는 다래끼보다 작은 것으로 콩, 팥, 감자 등을 심을 때 사용한다.
5. 씨앗망태는 씨둥기미, 씨부께둥으로도 부르며 주로 짚으로 만들었다.
6. 씨앗망태는 형태가 다양해서 물푸레나무를 말굽쇠 모양으로 구부려서 삼태기처럼 ~~~~
7. 씨앗망태나 종다래끼 외에도 씨앗 그릇이 따로 없이 소쿠리, 바구니, 뒤웅박 ~~~~~~~~

그림 4. 실험에 이용된 텍스트의 예

그림 4는 실험에 이용된 텍스트의 예이다. 이 텍스트를 SA의 입력형으로 바꿀 때 한 문장을 하나의 발화 단위로 가정하고 CTP와 ICP를 적용한다고 앞서 언급했다. CT에서의 발화는 일종의 단문이라 볼 수 있는데, 본 논문의 실험에 적용된 발화는 거의 3~4개의 단문이 결합되어 만들어진 복문을 하나의 발화로 취급하기 때문에 CT가 효과적으로 적용되지 못하는 문제가 발생한다. 또한 입력 텍스트에는 주어 등의 생략 현상이 빈번히 발생하는데, 생략된 성분을 복원해 주지 않고 CT를 적용하는 현재 실험에서는, no-Cb나 rough-shift가 대부분을 차지하는 것을 볼 수 있었다. 만약 단문 분할과 생략된 성분의 복원과 같은 전처리를 거친 후 CTP와 ICP를 적용한다면 훨씬 더 높은 성능을 가져오리라 기대한다.

본 실험에서 이미 완결된 텍스트를 이용한 가장 큰 이유는, 제안한 내용을 구현되어진 텍스트 생성 시스템에 적용했을 경우 객관적인 성능 평가가 어렵지만, 비교 가능한 텍스트가 존재함으로써 본래 텍스트 문장들의 순서 비교를 통해 텍스트 구조화 결과에 대한 성능 평가를 객관적으로 할 수 있는 용이함 때문이었다. 그러나 표5에서 동시적용 때의 낮은 성능은 대부분 완결된 텍스트를 쓰는 데서 기인하므로 본 논문에서 제안한 방법을 실제 텍스트 생성 시스템에 적용할 때는 텍스트 구조화의 입력인 발화의 단위가 노지은[2001]에서 언급한 것처럼, 단문 수준의 정보 형태(BIU)일 뿐 아니라, 생략된 성분들도 존재하지 않기 때문에 표5에서 보여지는 가중치와는 달리 CTP와 ICP가 가장 큰 역할을 수행할 것이라고 보고 성능 역시 크게 향상되리라 생각한다. 성능 저하의 또 다른 이유는 4개의 방법론이 동시에 적용될 때, 서로 간섭(interference) 현상을 일으킬 수 있다는 것인데, 각 방법론간의 충돌을 피하기 위해 4개의 방법론을 순차적으로 적용하여 총 3번의 SA를 시행하는 두번째 실험을 수행하였다.(그림 5)

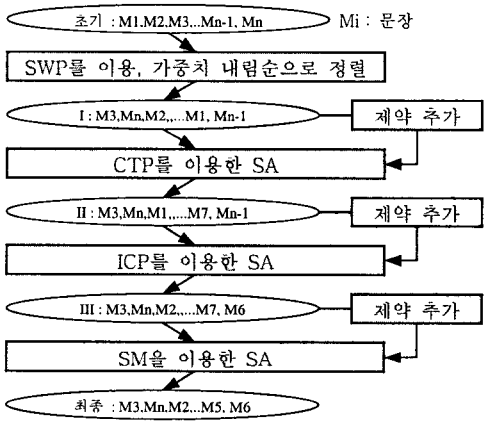


그림 5. 평가 함수의 순차 적용

각 방법을 순차적으로 적용할 때 각 방법론의 적용 순서에 따라 성능이 달라질 수 있고, 결과적으로 SWP, CTP, ICP, SM 순서로 적용하는 것이 가장 좋은 성능을 나타냈다. 이것은 우리의 직관과도 일치했다. 그림 6은, 각 방법론을 순차 적용한 SA를 이용해(그림 5) 그림 4의 텍스트를 구조화 하는 과정을 보여 준다. 먼저 각 문장에 부여된 문장의 가중치를 기준으로 문장을 내림순으로 정렬하여 가중치가 임계치 이상인 것들(여기서는 0,1,2의 인덱스를 갖는 문장이 해당된다)의 순서를 고정하는 제약을 설정한다(C1). 다음, 현재까지 설정된 순서상의 제약(C1)을 만족시키는 범위 내에서, CTP를

적용한 SA를 이용해 센터 전이를 최소화하는 문장 순서를 찾아내고 역시 임계치 이상의 조건을 만족하는 부분적인 문장 순서에 대해 그것을 계속 유지해 나갈 수 있도록 제약을 추가한다(C2,C3,C4). 마찬가지로, 현재까지 추가된 제약(C1,C2,C3,C4)을 지키는 범위 내에서 ICP를 적용한 SA를 이용해 추론 비용을 최소화하는 문장 순서를 찾아내어 제약(C5)을 추가 시키고, 마지막으로 모든 제약(C1,C2,C3,C4,C5)을 만족시키는 범위 내에서 문장간의 유사도를 최대화하는 문장 순서를 찾아내면 그것이 텍스트 구조화의 최종적인 결과가 된다.

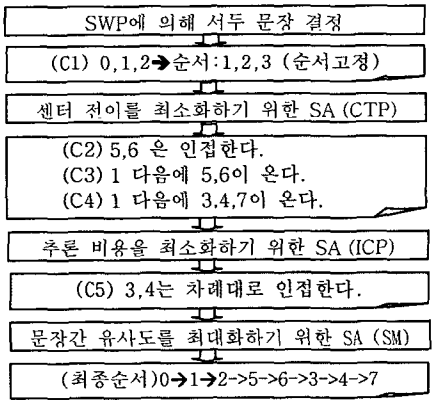


그림 6. 텍스트 구조화 처리 예(순차적용)

순차적용으로 텍스트를 구조화한 평균 성능은 0.6487 (60~68%)로, 동시적용보다 성능이 향상되었음을 알 수 있다.

9. 결론

본 논문에서는 SA를 이용해 추계적 텍스트 구조화 방안을 제안하고 동시에, SA적용에 필요한 평가 함수로서 4가지 방법론-센터 전이의 최소화(CTP),추론 비용의 최소화(ICP),인접한 문장간의 유사도의 최대화(SM), 서두 문장 결정을 위한 가중치 할당(SWP)-을 제안했다. 그리고 실제 텍스트 300개를 대상으로 텍스트 구조화를 시도하여 원래 텍스트와의 순서 비교를 통해 성능 평가를 하였으며 평균 60~68%의 유사도를 보였다. 물론, 원래 텍스트와 유사하지 않아도 충분이 결속성이 있는 텍스트일 수 있으므로 이 수치만으로 얼마나 텍스트 구조화가 잘 되는지 직관적으로 판단하기는 힘들다. 앞으로, 보다 신뢰할 만한 평가 방법을 통해, 본 논문에서 제안한 방법론을 실제 텍스트 생성 시스템에 적용해 봄으로써 그 실질적인 효용성을 검증해야 할 것이다.

감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

참고 문헌

[1]Ji-Eun Roh, Sin-Jae Kang, and Jong-Hyeok Lee (2001) "Korean Text Generation from Database for Homeshopping Sites", NLPRS 2001 (6th Natural Language Processing Pacific Rim Symposium), Tokyo, Japan, pp.419-426

[2]Mann,W.C. and Thompson,S.A.(1988) "Rhetorical Structure Theory:A Theory of Text Organisation",Text 8 (3), pp243-281

[3]McKeown,K.R.(1985) "Text Generation: Using discourse strategies and focus constraints to generate natural language text"," Cambridge University Press,

[4]Grosz, B.J. and Sidner,C.L.(1986) "Attention, Intentions and the Structure of Discourse",Computational Linguistics 12(3), pp175-204

[5]Grosz,B.J., Joshi,A.K. and Weinstein, S.(1995) "Centering: A Framework for Modeling the Local Coherence of Discourse", Computational Linguistics 21(2), pp203-225

[6]Hua Cheng (2000), "Experimenting with the Interaction between Aggregation and Text Planning", Proceedings of ANLP-NAACL 2000, USA

[7]V Mittal, J Moore, G Carenini and S Roth (1998),"Describing Complex Charts in Natural Language: A Caption Generation System". Computational Linguistics

[8]R Dale (1992), "Generating Referring Expressions",MIT Press

[9]C. Mellish, A. Knott, J. Oberlander and M. O'Donnell (1998) "Experiments using stochastic search for text planning", Proceedings of the Ninth International Workshop on Natural Language Generation, Niagara-on-the-Lake, Ontario, Canada

[10]Nikiforos Karamanis, Hisar Maruli Manurung (2002), "Stochastic Text Structuring using the principle of Continuity", International Natural Language Generation Conference 2002 (INLG'02), NewYork, USA, pp81-88

[11]Pablo A. Duboue, Kathleen R. McKeown (2002), "Content Planner Construction via Evolutionary Algorithms and a Corpus-based Fitness Function", International Natural Language Generation Conference 2002 (INLG'02), NewYork, USA, pp89-96

[12]Strube, Michael and Hahn, Udo (1996). "Functional Centering". Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, pp270-277

Marilyn Walker, Masayo Iida, and Sharon Cote. (1990) "Centering in Japanese Discourse" . Proceeding of the 13th International Conference on Computational Linguistics (COLING'90), Helsinki.

[14]Jon Oberlander, Alistair Knott, Mick O'Donnell, and Chris Mellish (1999), "Beyond elaboration: Generating descriptive texts containing it-clefts." In T Sanders, J Schilperoord and W. Spooren (eds.) Text representation: linguistic and psycholinguistic aspects. Amsterdam: Benjamins

[15]Marcu, D.(1996) "Building Up Rhetorical Structure Trees" in Proceedings of the 13th National Conference on Artificial Intelligence (AAAI'96), Portland,Oregon

[16]김미영 (1994)," 한국어의 담화와 중심화" , 서울대학교 대학원 언어학과 석사 학위 논문

[17]차건희, 송도규, 박재득 (1997), "한국어 대용어 생략 해결을 위한 센터링 이론의 적용" 제 9 회 한글 및 한국어 정보처리 학술대회 논문집, pp347-352 한나래 (1997), "한국어의 성분부정과 초점 해석에 관한 연구" , 서울대학교 대학원 언어학과 석사 학위 논문

[19]박철우 (1998) "한국어 정보 구조에서의 화제와 초점" , 서울대학교 대학원 언어학과 박사 학위 논문