토픽-코멘트 구조에 기반한 한국어 표층 생성기

김정은 최기선
한국과학기술원 전산학과, 전문용어언어공학연구센터
{euni, kschoi}@world.kaist.ac.kr

Korean Surface Realizer Based on Topic-Comment Structure

Jung Eun Kim Key-Sun Choi
Department of Computer Science, KORTERM, KAIST

요 약

본 논문은 자연언어생성 기술을 이용하여 질병에 대한 기술문을 생성해 내는 시스템에서 사용되는 표층 생성기에 대해서 다루고 있다. 표층 생성기는 문장의 추상적인 표현으로부터 통사적으로, 형태론적으로 올바른 텍스트로 생성하여 내는 것을 목표로 한다. 질병에 관한 기술문에 있는 문장들은 두가지 특징을 가지고 있다. 첫번째로, 질병 기술문의 문장들은 토픽-코멘트 구조로 나타내어질 수 있다. 두번째로, 같은 의미 범주에 속하는 문장들은 같은 토픽을 가진다. 따라서, 토픽은 의미범주로부터 유추될 수 있으므로 표층 생성기의 입력인 구 명세 (phrase specification)에 표현될 필요가 없다. 본 논문에서는 이런 특징을 이용하여 효율적인 표층 생성기를 만들기 위하여 표층 생성의 단계를 내부 표현 생성과 외부 문장 생성의 두 단계로 나누었다. 내부 표현 생성 단계에서는 코멘트에 해당하는 부분을 생성하고 외부 문장 생성 단계에서 의미범주 태그에 따라 토픽을 첨가하여 최종 문장으로 생성하였다. 이런 방법으로 실험한 결과, 본 표층 생성기는 문법에 맞으면서 자연스러운 텍스트를 생성해 낸다는 것을 알 수 있었다.

1. 서론

자연언어생성 (Natural Language Generation)은 인공지능과 계산 언어학의 하부 분야로써, 정보의 비 언어적 표현으로부터 사람이 쓰는 언어로 된 이해하기 쉬운 텍스트를 생성하는 컴퓨터 시스템을 만드는 것을 목표로 한다[7]. 자연언어생성 시스템의 가장 일반적인 구조는 다음과 같은 세 단계의 파이프라인 구조이다[8].

문서 계획 (text planning): 문서의 내용과 구조를 결정한다.

문장 계획 (sentence planning): 문서 계획에서 선택된 내용과 구조를 표현하기 위하여 어떤 단어, 통사 구조 등이 사용될 것인가를 결정한다.

표층 생성 (surface realization): 문장 계획의 결과로 만들어진 추상적 표현을 실제의 텍스트로 생성해 낸다.

본 논문에서는 토픽-코멘트 구조를 이용한 표층 생성기를 제시한다. 이 표층 생성기는 여러 곳에 산재해 있는 질병에 관한 정보들을 종합하고 재구성하여 사용자에게 제시하는 시스템에서 사용되었다. 사용자가 필요로 하는 정보를 얻기 위한 지식 기반으로는 백과사전이 사용되었다. 백과사전은 정보들이 잘 조직되어 나타나 있어서 어떤 대상의 특징을 잘 드러내준다.

본 논문의 2장에서는 질병 기술문에 나타나는 문장의 특징을 바탕으로 표층 생성에 사용된 접근 방식과 그에 따른 표층 생성기의 구조를 설명한다. 3장에서 입력으로 사용된 텍스트 명세에 관하여 설명하고 4장에서 표층 생성기의 세부적인 모듈들에 관해 설명한다. 5장에서 토픽-코멘트 구조에 기반한 외부 문장의 생성에 대해서 기술하고 6장에서 실험과 평가를 제시한 뒤 7장에서 결론과 향후 계획을 제시한다.

2. 표층 생성에 대한 접근 방법과 구조

한국어는 토픽과 주제 중심의 언어이다[2]. 질병 기술문은 특정 질병에 대한 설명을 담고 있는 텍스트로, 한국어의 토픽 중심 언어적 특성이 두드러지게 나타난다. 질병기술문은 질병기술문이 가져야 할 전형적인 토픽들을 가지고 있고 이 토픽에 따라 내용이 조직된다. 질병 기술문의 대략적 구조는 그림 1과 같다.

<질병의 이름>

원인은 A이다.
증상은 B이다.
치료법은 C이다.

그림 1: 질병에 관한 기술문의 구조

'원인은", '증상은", '치료법은"은 각 문장의 토픽으로써 토픽 표지인 '은"을 가진다. 각 문장은 원인, 증상, 치료 등 어떤 의미를 가지고 있는데 이것을 나타내는 태그를 본 논문에서는 '의미범주 태그"라고 정의한다. 같은 의미범주 태그에 속하는 문장들은 같은 토픽을 가지므로 토픽은 의미범주 태그로부터 유추될 수 있다. 따라서 그림 1에서 A, B, C는 가변적이지만 나머지 부분은 의미범주에 따라 고정적이다. 즉, 질병에 관한 기술문에 포함된 한 문장은 '[X]는 [Y]이다"와 같은 구조로 나타내어질 수 있고, 각 의미범주 그룹에서 Y는 가변적이지만 나머지는 고정적이라고 할 수 있다. 이 점을 고려하면 표층생성시 얻고자 하는 문장을 표현할 때, 그 표현에 문장 구성 성분을 모두 포함시키지 않고 Y만 포함하여도 나머지 부분은 유추하여 생성할 수 있다. 따라서 표층 생성의 과정에서 Y의 생성은 나머지 부분의 생성과 분리될 수 있다. 본 논문에서는 Y를 생성하는 과정을 '내부 표현의 생성"으로, 나머지 부분을 생성하여 완성된 전체 문장으로 만들어 내는 과정을 '외부 문장의 생성"으로 칭한다. 내부 표현의 생성 단계에서는 내부 표현에 대한 추상적 의미 표현을 자연 언어로 생성해 내고 외부 표현의 생성 단계에서는 앞 단계에서 생성된 문자열에 주어 (토픽)와 서술격 조사 '이다"를 각각의 적절한 자리에 덧붙여주게 된다.

이러한 방식을 따르는 표층 생성기의 전체 구조는 그림 2와 같다. 표층 생성의 첫번째 단계는 문장 경계 결정이다. 다음으로는 내부 표현의 생성 단계를 거치게 된다. 이 단계에서는 단어들의 순서를 결정하고 적절한 조사 등을 선택하여 각 성분들을 생성한다. 그 후에 외부 문장의 생성을 하고 접속사와 연결어미를 결정한 뒤 문단을 나누어준다.

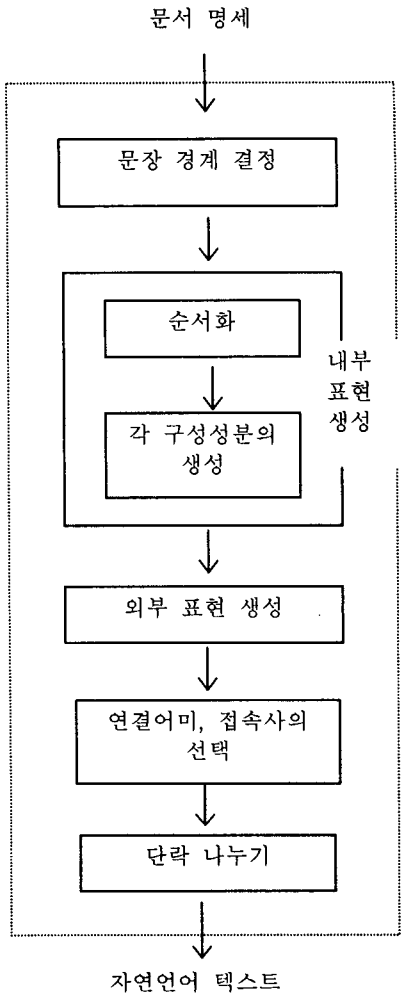


그림 2: 표층 생성기의 구조

3. 텍스트 명세

문장 계획에서는 문서 계획의 결과를 받아서 텍스트의 내용과 구조에 대한 추상적 표현이 만들어지는데 이것을 '텍스트 명세 (text specification)"라고 한다. 표층 생성기는 텍스트 명세를 받아서 실제의 문장으로 바꾸게 된다. 이것은 트리 형태인데, 내부 노드들은 문단이나 절 등 텍스트의 구조를 나타내고 단말 (leaf) 노드는 텍스트의 문장들을 나타내다. 이 단말 노드를 구 명세 (phrase specification)라고 부른다[8]. 구 명세는 의미범주 태그를 가지는데, 본 논문에서는 의미범주 태그로

써 원인, 증상, 치료법이 다루어진다. 본 표층 생성 시스템에서 텍스트 명세는 그림 3과 같다.

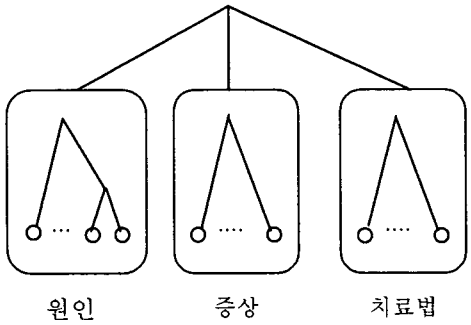


그림 3: 텍스트 명세

3.1. 구 명세

구 명세는 속성과 값으로 구성되는 특성 기술(feature description)의 형태를 가진다[6]. 구 명세의 예는 그림 4와 같다.

IE는 내부 표현에 해당하는 명세로써 기본적으로 SURGE[4]를 따르고 있다. IE는 주제 구조(thematic structure)로, 술어와 주제 역할들로 이루어진다. 주제 역할은 핵심 역할과 주변 역할로 나뉘어지는데, PROCESS와 PARTIC은 핵심 역할에 속하고 CIRCUM은 주변 역할에 속한다. 그림 4에는 시제, 법 등이 나타나지 않는데, 질병 기술문에 나타나는 문장들의 특징에 의해 이들은 각각 현재, 평서를 기본값으로 가진다.

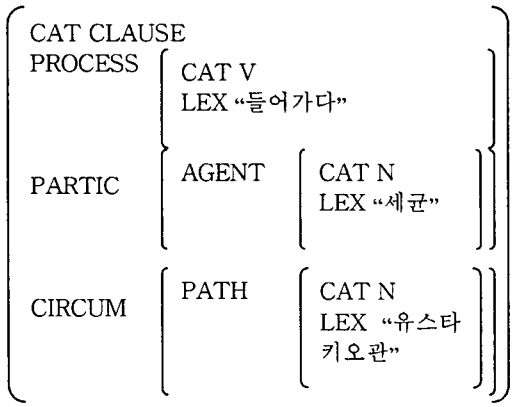
PROCESS : 문장의 술어가 된다. 이것은 하위범주화 정보에 따라 논항들을 가지게 된다.

PARTIC : PROCESS의 의무적 참여자(obligatory participants)로, 술어의 논항이 된다. 행위자, 착점 (goal), 수혜자 (beneficiary) 등 여러 주제 역할을 가진다.

CIRCUM : 술어 부속사 (adjunct) 또는 문장 부속사의 기능을 한다. 시간, 방향, 이유, 빈도 등을 나타내어[5] 부사구로 생성된다.

CAT : 절, 구, 단어 등 문법 단위의 범주를 나타낸다. 범주가 단어일 때 CAT은 품사 정보를 가지게 된다. 이는 한국어 표층 생성기에서는 올바른 형태론적 형태를 만들기 위해 품사 정보가 필요하기 때문이다. 예를 들어, 형용사와 동사는 현재시제의 표시방법이 다르다. 현재시제가 종결형에 의해 표시될 때 동사는 어간+ -'는-/-ㄴ- +'어말어미의 형태가 되지만 형용사에서는 현재시제를 표시하는 특수한 형태가 발견되지 않는다[1]. 동사인 "나오다"는 현재형이 "나온다"가 되는 반면 형용사인 "아프다"는 현재형도 마찬가지로 "아프다"이다.

입력 명세 : 원인(중이염, IE)
IE 명세



출력 문장: 원인은 유스타키오관을 통해 세균이 들어가는 것이다

그림 4: 입력과 출력의 예

LEX: 문장의 각 요소에 대해, 문장 계획에서 선택된 어휘이다.

4. 표층 생성기의 세부 모듈

표층 생성기는 주제 역할을 통사적 역할로 변환한다. 보통 PROCESS, AGENT, GOAL, BENEFICIARY는 각각 술어, 주어, 직접 목적어, 간접 목적어로 생성된다. 그런데, 어떤 동사의 종속자 (dependent)가 어떤 역할을 하는지를 구분하는 것이 항상 명확하지는 않다. 따라서 술어의 하위범주화 정보를 참조하여 이를 결정해 주게 된다. 주제 역할을 통사적 역할로 바꾸기 위해 표층생성기는 다음과 같은 일을 수행하게 된다.

4.1. 순서화

한국어는 기본적인 주어-술어의 단어 순서를 가지는 SOV 언어이다[2]. 한국어에서 주요 문장 성분들은 문장 내에서의 순서가 정해져 있다. 그러나 모든 경우에 순서가 늘 이렇게 명확히 정해지지는 않는다. 술어에 따라 논항들의 순서가 달라지기도 하고, 강조점에 따라 순서가 달라지기도 한다. 부사어의 경우에는 문장 내 위치가 더욱 자유롭다. 본 논문에서는 술어와 논항들의 순서를 정하기 위해 격틀을 이용하였다. 술어의 논항들은 주로 쓰이는 순서를 가지고 있는데, 이것은 격틀을 참조하여 알 수 있다. 이 때 술어의 의미가 무엇인가는 전혀 고려할 필요가 없으므로 격틀에는 술어들의 의미는

명시되어 있지는 않고, 그림 5처럼 논항과 논항들의 가능한 순서가 명시되어 있다.

나오다 :arg2에서 arg1이

그림 5 : 격틀의 예

이렇게 논항들의 순서를 정한 후 부속사의 위치를 정하게 된다. 부속사는 대개 부사어가 되는데, 쓰이는 위치가 유동적이므로 부사어들의 순서를 정하기 위해 각각의 부사어군들에 대해서 백과사전의 질병기술문에서 어순을 분석하였다. 이 결과, 각각의 부사어군들이 쓰이는 위치는 일정한 패턴을 가지고 있음을 알 수 있었다 따라서, 이 패턴을 따라 부사어들을 위치를 결정하였다.

4.2. 조사의 선택

체언은 문법적 기능을 발휘하기 위해 조사와 결합한다. 예를 들어, 문장성분이 주어냐 목적어냐에 따라 단어에 붙는 조사가 다르다. 주제 역할만으로는 조사를 선택하는 데 필요한 정보가 충분하지 않으므로, 격틀을 이용하여 적절한 조사를 선택한다.

4.3. 연결어미, 접속사의 선택

절단위의 표현들을 자연스러운 텍스트로 생성하기 위해서는 절들을 적절하게 잇는 것이 필요한데, 이것을 위해서 접속사와 연결어미가 사용되게 된다. 그런데, 한가지 수사관계에 대해서 선택될 수 있는 연결어미는 여러가지가 존재한다. 따라서 연결어미를 선택하기 위해서 기존의 질병 설명문에서 연결어미가 어떻게 쓰이고 있는지를 분석한 결과를 이용하였다. 각 의미군에 따라 연결어미들을 묶고 이중 대표적인 것을 선정하였다. 수사관계와 설정된 연결어미의 일부는 표 1과 같다.

수사관계	선택될 연결어미
부연설명	는데
나열	고
이유-결과	어서
대립	으나
전개	며

표 1 : 수사관계와 연결어미

4.4. 용언의 활용

한국어의 활용어에는 용언과 서술격 조사가 있고, 활용형에는 종결형, 연결형, 전성형 등이 있다. 따라서, 각각에 경우에 대해서 단어의 형태를 적절하게 바꾸어 주어야 한다. 예를 들어 "아프다"가 관형사형으로 활용될 때에는 "아픈"으로 형태를 변형시켜야 한다.

5. 의미범주 태그를 이용한 외부 문장의 생성

이미 언급했듯이 구 명세는 의미범주 태그를 가지고 있는데, 본 논문에서는 의미범주 태그로 원인, 증상, 치료를 가지는 경우를 다룬다.

5.1. 원인

입력 명세: CAUSE (황달, IE)
IE 명세

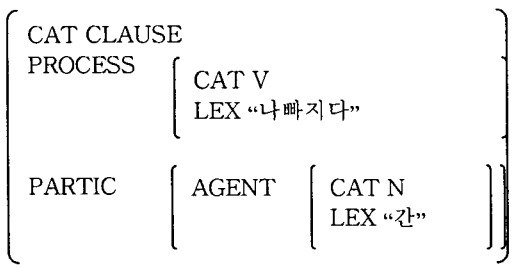


그림 6: 의미범주 태그로 CAUSE를 가지는 문장의 입력 명세의 예

질병의 원인에 해당하는 문장의 표현은 "원인은 X이다"의 형태를 가진다. X는 어떤 병의 원인을 설명하는 명사구이다. "원인은 간이 나빠지는 것이다"라는 문장의 입력 명세는 그림 6과 같다.

질병의 원인의 내용은 "정신적인 피로"와 같은 구의 형태로 나타나기도 하고 "간이 나빠지다"와 같은 절의 형태로 나타나기도 한다. 따라서 내부 표현이 절의 형태일 때는 명사구로 변환되어야 한다. 따라서 위의 예에서 내부 표현은 "간이 나빠지는 것"으로 변환된다. 그런 다음에 주어인 "원인은"과 서술격 조사 "이다"를 적절한 자리에 덧붙여서 "원인은 간이 나빠지는 것이다"로 최종적으로 생성되게 된다.

5.2. 증상

아래의 두 문장은 증상 범주에 속하는 문장의 예이다.

"증상은 척주가 옆으로 굽는다."
"증상은 두통, 현기증 등이다."

첫번째 문장의 입력 명세는 SYMPTOM (구루병,

척주가 옆으로 굽는다")이다. 따라서 내부 표현을 생성해 낸 뒤에 "증상은"을 주어 자리에 넣어주기만 하면 된다. 두번째 문장의 입력 명세는 SYMPTOM (귓병, 두통, 현기증 등")이므로 내부 표현을 생성한 뒤 주어 "증상은"과 서술격 조사 "이다"를 적절한 자리에 덧붙여준다.

5.3. 치료법

치료법에 속하는 문장의 처리는 토픽이 "증상은"이 아니라 "치료법은"이라는 것 외에는 증상에 속하는 문장의 처리와 유사하다. 다만, 다양한 형태를 가지는 자연스러운 문장을 만들어내기 위해서 치료법에 대해서는 내부 표현이 명사구일 때는 조사로 "은" 대신 "으로는"을 사용하였고 "이다" 대신 "이 있다"를 사용하였다.

TREATMENT(백혈병, 엑스선 용법, 부신피질호르몬 등")

예를 들어 입력이 위와 같을 때 결과 문장은 "치료법으로는 엑스선 용법, 부신피질 호르몬 등이 있다."이다.

6. 실험 및 평가

자연언어생성 시스템이 어떻게 평가되어야 하는가에 대해서는 아직 정리된 바가 없는데[3], 본 논문에서는 표층 생성기의 목적에 따른 평가 기준을 사용하였다. 표층 생성기의 목적은 문법 규칙들을 추상적 표현에 적용하여 통사적으로, 형태론적으로 올바른 텍스트를 생성하는 것이다[8]. 실험은 30개의 질병을 대상으로 하였고 계산 언어학을 전공하는 석사과정 학생 두명이 평가하였다.

본 실험에서는 결과로 나오는 텍스트가 통사적으로, 형태론적으로 얼마나 옳은지를 보기 위해서 어순, 연결어미 및 접속사, 조사, 용언 활용에 대해서 부적절하게 쓰인 횟수를 평가하였다. 그 결과는 표 3과 같다. 한 문장이 문법에 맞다고 해서 '좋은' 문장이라고 할 수는 없다. 문법에 맞지만 자연스럽지 못한 문장도 존재할 수 있다. 따라서, 위의 기준과 함께, 문장의 가독성도 평가하였다. 각 문장에 1~5의 점수를 주도록 하였는데, 1은 "전혀 이해할 수 없다"를 뜻하고 점수가 높을수록 가독성이 높다. 결과는 표 2와 같다.

기준	전체	A	B
어순	58(문장)	0	0
연결어미 및 접속사	60	0	0
조사	224	0	0
용언 활용	161	0	0

표 2: 형태론과 통사론의 관점에서의 평가 결과

표 3으로부터 30개의 질병에 대해서 어순, 연결어미 및 접속사, 조사, 용언의 활용 모든 면에서 어긋나는 경우가 전혀 없음을 볼 수 있다. 또한 표 2를 통해서 표층 생성기가 자연스러운 문장을 생성해 내고 있음을 알 수 있다. 5점을 받지 못한 문장들에 대해서는 평가자들에게 이유를 제시하도록 하고 분석해 본 결과, 외부 문장 생성 단계에서 같은 의미범주의 문장은 같은 형태로 생성되기 때문에 다소의 부자연스러움이 생긴 것으로 보인다. 그러나, 3미만의 점수를 받은 문장이 한문장도 없음을 볼 때 전체적으로 받아들일만한 수준의 가독성을 보이고 있다고 말할 수 있다.

7. 결론과 향후 연구

본 논문에서는 질병 기술문 생성 시스템에서의 표층 생성기를 제안하였다. 본 논문에서 제안된 표층 생성기는 질병에 관한 기술문에서 나타나는 두 가지 특성을 기반으로 하여 개발되었다. 첫째로, 질병기술문의 문장들은 토픽-코멘트 구조로 나타내어질 수 있다. 문장들은 대략적으로 "X는 Y이다"의 구조를 가지고 있다. 두번째로, 같은 의미범주에 속하는 문장들은 같은 주어 (토픽)를 가진다. 이런 특징들을 이용하여 효율적인 표층 생성기를 개발하기 위하여 표층 생성 과정을 내부 표현 생성과 외부 문장 생성의 두 단계로 분리하였다. 내부 표현 생성 단계에서는 Y가 생성되고 외부 문장 생성 단계

의미범주	가독성 점수										전체 문장 수	평균
	A					B						
	1	2	3	4	5	1	2	3	4	5		
원인	0	0	1	1	15	0	0	2	4	11	17	4.68
증상	0	0	2	4	28	0	0	1	6	27	34	4.76
치료	0	0	1	1	5	0	0	1	4	2	7	4.36

표 3: 문장의 가독성에 대한 평가 결과

에서는 나머지 부분의 생성을 통하여 전체 문장이 완성된다. 이 방법을 이용한 실험의 결과로 표층 생성기가 문법적으로 올바르면서 자연스러운 문장을 생성해 낸다는 것을 알 수 있었다.

본 논문에서는 질병 기술문의 구성요소로써 원인, 증상, 치료만을 다루었으나 질병 기술문에는 경과, 예방 등 더 다양한 구성요소들이 존재하고 있다. 따라서 앞으로는 본 논문에서 다루어진 것들 이외의 의미범주로 표층생성기를 확장하는 것이 필요하다. 또한, 실험을 통하여 같은 범주에 속하는 문장은 늘 같은 형태의 토픽을 가지게 되므로 부자연스럽게 느껴지는 경우가 있음을 알 수 있었으므로 각 의미범주 내에서도 토픽을 다양한 방법으로 표현하여 좀 더 유려한 문장을 만들 필요가 있다.

참고 문헌

[1] 남기심, 고영근 (1996) "표준 국어문법론", 탑출판사

[2] Cho See-Young (1991) "Focusing in English and Korean", Frankfurt am Main/ Bern/ New York/ Paris: Peter Lang.

[3] Dale Robert and Chris Mellish (1998) "Toward the Evaluation of Natural Language Generation", Proceedings of the First International Conference on Language Resources and Evaluation.

[4] Elhadad Michael and Jacques Robin (1996) "An Overview of SURGE: a Reusable Comprehensive Syntactic Realization Component", Technical Report 96-03, Ben gurion University, Department of Computer Science.

[5] Korkmaz Turgay (1996) "Turkish Text Generation with Systemic-Functional Grammar", Master's thesis, Bilkent university, Ankara Turkey.

[6] Marsi Erwin (1998) "A Reusable Syntactic Generator for Dutch", Computational Linguistics in the Netherlands 1997, Selected Papers from the Eight CLIN Meeting, Amsterdam: Rodopi.

[7] Reiter Ehud and Robert Dale (1997), "Building applied natural language generation systems", Natural Language Engineering Vol 3, Part 1, pp57-87.

[8] Reiter Ehud and Robert Dale (2000), "Building Natural Language Generation Systems", Cambridge University Press.