온톨로지 기반의 문서 생성 시스템

류재현, 박성배
경북대학교
{jhryu, sbpark}@sejong.knu.ac.kr

A Document Generation System Based on an Ontology

Jae-Hyun Ryu, Seong-Bae Park
Kyungpook National University

요 약

온톨로지란 사물이나 개념의 속성이나 관계를 사람과 컴퓨터 모두 이해할 수 있는 형태로 표현한 모델로 정보검색, 인공지능, 소프트웨어 공학 등의 분야에서 많이 활용된다. 온톨로지에는 다양한 정보가 구조화되어 저장되어 있지만 일반적으로 온톨로지가 제공하는 그래프 형태의 데이터들은 사용자들이 직관적으로 이해하기가 힘들다. 따라서 본 논문에서는 온톨로지의 정보를 문장화하여 한국어 문서를 생성하는 시스템을 제안한다. 제안하는 시스템은 주제와 관련된 트리플을 추출하고 이를 문장정렬, 결합, 생성을 위한 정보가 담긴 템플릿을 생성한 뒤 한국어 문법에 맞게 문장을 생성한다. 또한 기존 연구에서 다루지 않았던 이벤트 온톨로지의 내용을 포함하여 문장을 생성한다. 두 온톨로지로부터 생성된 문장을 연결하여 주제어를 설명하는 하나의 문서를 작성한다.

주제어: 온톨로지, 문서 생성, 트리플기반 문장 생성

1. 서론

온톨로지란 어떠한 사물이나 개념의 속성이나 관계를 사람과 컴퓨터 모두 이해할 수 있는 형태로 표현한 모델이다. 온톨로지는 클래스, 인스턴스, 속성(관계)으로 구분할 수 있고 클래스나 인스턴스 사이에 특정한 속성이나 관계가 있음을 표현할 때 주로 <주어, 속성(관계), 목적어> 형태의 트리플로 나타낸다. 온톨로지는 형태나 목적에 따라 다양한 종류로 표현될 수 있으며 인공지능, 정보검색, 소프트웨어공학 등의 분야에서 활용된다. 예를 들어 정보검색에서는 잘못 입력한 키워드를 동의어 사전형태의 온톨로지를 사용하여 키워드를 바로 잡거나, 같은 개념에 대한 다른 어휘를 연결하여 해당 개념에 대한 풍부한 검색 결과를 제공할 수 있도록 돕는다. 그러나 온톨로지는 실제로 그래프 형태로 표현되어 있을 뿐만 아니라 그의 방대한 규모로 인해 사용자들에게 정보를 직관적으로 전달하기 힘들다. 따라서 온톨로지로부터 핵심 정보만 추출하여 문서화할 필요가 있다.

Androutsopoulos et al. [1]은 온톨로지를 기반으로 자연어 문서를 생성하는 방법을 제안 하였다. 위 논문은 문서를 생성하는 순서를 세 가지의 단계에 맞춰 순차적으로 진행되도록 나누었다. 온톨로지 내 주제를 선택하고 문서가 될 트리플을 선정하는 문서 계획 단계를 서두로 하여 문장을 결합하거나 순서를 재배치시키는 세부계획 단계를 진행한다. 최종적으로 문서 생성 단계를 거쳐 주어진 문장을 실체화하여 문서로 생성될 수 있도록 한다. 이 방법은 영어 기반의 문법과 단어 사전을 활용하기 때문에 한국어에서 그대로 적용하기 힘들다.

본 논문에서는 [1]의 방법을 한국어에 맞게 각 단계를 적용하여 온톨로지를 바탕으로 문서를 생성하는 시스템을 제안한다. 제안하는 시스템은 도메인 온톨로지의 트리플 중 문장이 될 트리플을 선정하고 각 트리플의 속성별로 문장의 형태를 지정하여 템플릿화 한다. 이후 각 템플릿의 주어와 속성(관계)에 따라 순서를 정렬하고 인접한 두 템플릿 결합한 뒤 이를 문장으로 실체화 하여 특정 주제에 대한 문서를 생성한다. 또한 기존 연구에서는 다루지 않았던 이벤트 온톨로지를 바탕으로 트리플을 추출하여 자연어 문장을 생성한다. 최종적으로 도메인 온톨로지로부터 생성된 자연어 문장과 결합하여 하나의 문서를 생성한다. 이벤트 온톨로지란 사실에 기반한 사건을 육하원칙에 맞춰 구조화한 온톨로지를 의미한다 [2]. 이벤트 온톨로지가 문서에 포함되면 주제에 대한 정의뿐 아니라 배경, 관련된 사실 등과 같은 더욱 풍부한 정보를 얻을 수 있다.

2. 온톨로지 기반의 문서 생성 시스템

본 연구에서 제안하는 시스템은 도메인 온톨로지와 이벤트 온톨로지에서 주제어와 인스턴스 간의 거리를 측정하여 문장이 될 트리플을 선택하고, 트리플의 속성에 따라 문장형태를 결정해 문장정보를 담은 템플릿을 만든다. 다음으로 트리플의 주어와 속성에 따라 템플릿을 재정렬한 뒤 결합한다. 마지막으로 각 템플릿의 조사와 동사 정보를 문법에 맞게 실체화하는 과정을 거쳐 문서를 생성한다. 그림 1은 본 연구에서 제안하는 시스템의 전체 흐름도를 나타낸다.

2.1 트리플 선택 및 문장형태 결정

도메인 온톨로지의 모든 트리플을 문장화하면 온톨로지의 방대한 크기 때문에 생성된 결과 문서의 규모 또한 거대해져 주제어와 연관성이 떨어지는 정보까지 문장으로 변환되는 문제가 발생한다. 따라서 본 논문에서는 효율적인 정보 전달을 위해 주제어를 중심으로 거리가 2이내인 트리플을 문장 생성을 위한 트리플로 추출한다.

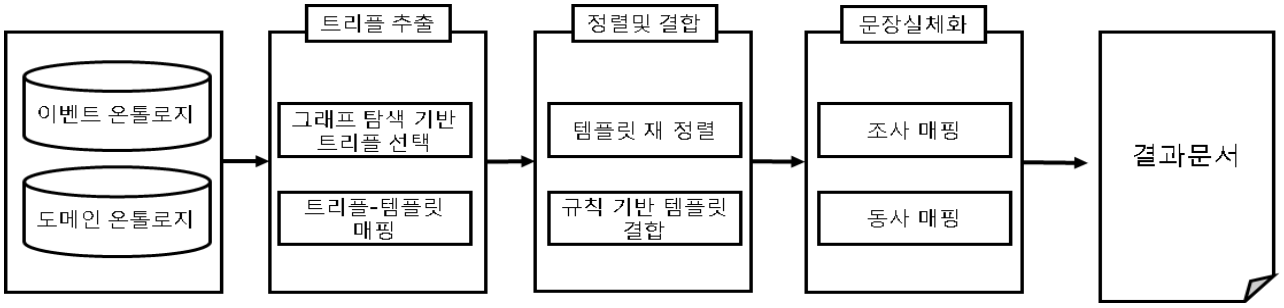

그림 1 시스템 전체 흐름도

추출한 트리플들은 주제어와 인스턴스 간의 속성에 따라 문장 템플릿을 생성함으로써 문장의 뼈대가 될 수 있도록 한다. 문장 템플릿은 문장 순서에 따라 슬롯의 집합 형태로 형성된다. 슬롯은 트리플의 단어가 문장에서 어떤 형태로 표현되는지를 의미하며 단어 별로 각각 표현된다. 슬롯에 들어 있는 정보는 주어, 목적어, 서술어, 보어 등의 문장성분과 문장 성분에 따라 단어 뒤에 따라올 수 있는 조사나 어미 정보가 들어 있다[3]. 이때 서술어의 경우 서술어의 시제 정보도 추가로 들어 있다. 표 1은 트리플의 속성 별로 생성되는 템플릿의 예를 보여준다.

표 1 트리플의 속성별 생성되는 템플릿의 예

트리플	생성되는 템플릿
<3D프린터, 이용, 3D도면>	[3D프린터, 주어, 보조사] [3D도면, 목적어, 목적격조사] [이용, 서술어, 종결어미, 현재형]
<3D프린터, 이다, 프린터>	[3D프린터, 주어, 보조사] [프린터, 보어, 조사없음] [이다, 서술어, 종결어미, 현재형]
<3D빌더, 한 종류, 3D프린터>	[3D빌더, 주어, 보조사] [3D프린터, 관형어, 관형격조사] [한 종류, 보어, 조사없음] [이다, 서술어, 종결어미, 현재형]
<스트라티, 최대속도, 64km>	[스트라티, 관형어, 보조사] [최대속도, 주어, 주격조사] [64km, 보어, 조사없음] [이다, 서술어, 종결어미, 현재형]

이벤트 온톨로지의 경우 기초가 되는 문장이 육하원칙에 맞춰 세부적으로 구조화 되어진 온톨로지이며, 각각의 요소가 이벤트명(주제어+이벤트 번호)을 중심으로 연결되어 있다. 따라서 연결된 트리플을 추출하여 '언제 -> 어디서 -> 누가 -> 무엇을 -> 왜 -> 어떻게' 순서로 템플릿을 만든다. 이렇게 생성된 이벤트 온톨로지의 템플릿은 이후 문장이 생성되면 실제 문서에서 문서의 마지막에 리스트 형식으로 나열하기 때문에 정렬 및 결합 과정에 포함되지 않는다. 표 2는 이벤트 온톨로지의 속성별 슬롯 정보이며 표 3은 이를 바탕으로 실제 템플릿이 생성된 예이다.

표 2 이벤트 온톨로지의 속성별 슬롯의 정보

속성	정보
언제	부사어, 부사격조사(에)
어디서	부사어, 부사격조사(에서)
누가	주어, 주격조사
무엇을	목적어, 목적격 조사
왜	부사어, 부사격조사(에)
어떻게	서술어, 종결어미, 과거형

표 3 이벤트별 생성되는 템플릿의 예

트리플	템플릿
<3D프린터이벤트1, 누가, 쉐이프웨이즈>	[쉐이프 웨이즈, 주어, 주격조사]
<3D프린터이벤트1, 무엇을, 3D프린팅 개인마켓>	[3D프린팅 개인마켓, 목적어, 목적격 조사]
<3D프린터이벤트1, 어떻게, 설립>	[설립, 서술어, 종결어미, 과거형]
<3D프린터이벤트2, 언제, 1989년>	[1989년, 부사어, 부사격조사]
<3D프린터이벤트2, 누가, 스콧크럼프>	[스콧크럼프, 주어, 주격조사]
<3D프린터이벤트2, 무엇을, FDM>	[FDM, 목적어, 목적격조사]
<3D프린터이벤트2, 어떻게, 특허>	[특허, 서술어, 종결어미, 과거형]

2.2 템플릿 재 정렬 및 템플릿 결합

템플릿 재 정렬은 문서가 만들어졌을 때 문장이 자연스럽게 연결되도록 하고 템플릿 결합은 생성된 문장을 복문으로 형성하여 풍부한 표현이 가능하도록 하는 중요한 단계이다. 본 논문에서의 템플릿 재 정렬은 먼저 모든 템플릿을 주어를 기준으로 정렬한다. 그 후 템플릿을 차례대로 돌며 새로운 목적어가 나왔을 때 해당 템플릿의 목적어가 주어인 템플릿을 찾아 해당 템플릿의 뒤에 삽입한다. 위 방식으로 템플릿을 정렬하여 문장을 생성할 경우 문장에서 새로운 단어가 나타났을 때 해당 단어에 대한 개념을 설명하는 문장이 연달아 등장할 수 있어

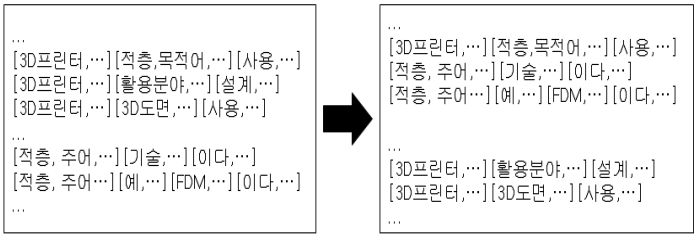

그림 2 템플릿 재 정렬의 예

사용자가 개념을 이해하기 쉽도록 문서가 작성될 수 있다는 장점이 있다. 그림 2는 템플릿 재 정렬 과정을 거친 뒤 단어 '적층' 을 주어로 가지는 템플릿이 단어 '적층' 을 목적어로 가지는 템플릿의 다음으로 순서가 조정이 된 예를 나타낸다.

템플릿 결합의 경우 정렬된 템플릿을 차례대로 순회하면서 기준이 되는 템플릿과 다음에 오는 템플릿이 주어가 같다면 두 개의 템플릿을 결합 후보로 간주한다. 본 논문에서는 결합 후보에 대하여 결합 유무를 결정하는 결합 규칙을 3가지로 정의하였으며 각 규칙과 우선순위는 아래와 같다.

1. 서술어가 같다.
2. 하나의 템플릿의 서술어가 '이다' 이다.
3. 두 템플릿의 서술어가 다르다.

표 4는 해당 규칙별 템플릿이 결합되는 예이다. 규칙 1의 예는 두 개의 트리플이 주어는 'FDM' 으로, 서술어는 '이용' 으로 동일하기 때문에 1의 규칙으로 결합된다. 결합후의 문장 순서를 결정하기 위해 두 템플릿의 문장순서를 이용한다. 예의 경우 두 템플릿 모두 '주어 - 목적어 - 서술어' 의 형식을 갖고 있다. 따라서 두 템플릿이 결합후의 문장순서는 '주어 - 목적어1 - 목적어 2 - 서술어' 의 순이 된다. 결합 후 목적어 사이의 연결이 매끄럽게 될 수 있도록 첫 번째 목적어와 두 번째 목적어를 이어 주기 위해 첫 번째 목적어의 조사를 접속조사로 바꾸어 준다.

표 5 템플릿별 생성되는 문장의 예

템플릿	생성 문장
[FDM, 주어, 보조사] [ABS, 목적어, 접속조사] [PLA, 목적어, 목적격조사] [이용, 서술어, 종결어미, 현재형]	FDM은 ABS와 PLA를 이용한다.
[3D프린터, 주어, 보조사] [3차원물체, 목적어, 목적격조사 [출력, 서술어, 연결어미, 현재형] [프린터, 보어, 조사없음] [이다, 서술어, 종결어미, 현재형]	3D프린터는 3차원물체를 출력하는 프린터이다.
[3D프린터, 주어, 보조사] [건축, 부사어, 부사격조사] [활용, 서술어, 연결어미, 현재형] [스트라티, 목적어, 목적격조사] [제작, 서술어, 종결어미, 현재형]	3D프린터는 건축에 활용되고 스트라티를 제작한다.

2.3 문장 실체화

문장 실체화는 이전의 단계에서 생성된 템플릿을 바탕으로 실제 자연어 문장을 생성하는 과정이다. 실체화 단계에서는 두 가지 사상 방법을 적용하는데 하나는 조사 매핑이고 나머지는 서술어 매핑이다. 조사 매핑은 템플릿 내에 조사정보가 포함된 슬롯을 대상으로 슬롯 정보에 따라 은/는, 이/가 등의 조사형태를 결정하는 방법이다. 서술어 매핑은 서술어 사전을 이용하여 슬롯 정보에 따라 사전에서 찾아 생성되도록 했다. 서술어 사전은 서술어가 어미와 시제에 따라 변하는 형태를 모두 저장하고 있다. 표 5는 문장 실체화 단계를 적용한 예이다.

표 4 템플릿 결합 규칙별 생성 예

규칙번호	템플릿1	템플릿2	결합 결과
1	[FDM, 주어, 보조사] [ABS, 목적어, 목적격조사] [이용, 서술어, 종결어미, 현재형]	[FDM, 주어, 보조사] [PLA, 목적어, 목적격조사] [이용, 서술어, 종결어미, 현재형]	[FDM, 주어, 보조사] [ABS, 목적어, 접속조사] [PLA, 목적어, 목적격조사] [이용, 서술어, 종결어미, 현재형]
2	[3D프린터, 주어, 보조사] [프린터, 보어, 조사없음] [이다, 서술어, 종결어미, 현재형]	[3D프린터, 주어, 보조사] [3차원물체, 목적어, 목적격조사] [출력, 서술어, 종결어미, 현재형]	[3D프린터, 주어, 보조사] [3차원물체, 목적어, 목적격조사 [출력, 서술어, 연결어미, 현재형] [프린터, 보어, 조사없음] [이다, 서술어, 종결어미, 현재형]
3	[3D프린터, 주어, 보조사] [건축, 목적어, 목적격조사] [활용, 서술어, 종결어미, 현재형]	[3d프린터, 주어, 보조사] [스트라티, 목적어, 목적격조사] [제작, 서술어, 종결어미, 현재형]	[3D프린터, 주어, 보조사] [건축, 부사어, 부사격조사] [활용, 서술어, 연결어미, 현재형] [스트라티, 목적어, 목적격조사] [제작, 서술어, 종결어미, 현재형]

3. 결론 및 향후 연구

본 논문은 온톨로지에서 추출한 트리플을 한국어 문법에 기반 하여 문서를 생성하는 시스템을 제안하였다. 트리플 선택, 문장형태 결정, 템플릿 재 정렬/결합, 문장 실체화 단계를 통해 주제어에 대한 개념과 관련된 사건의 템플릿으로부터 문장을 생성할 수 있다. 또한 트리플-템플릿 매핑 방법 규칙과 서술어 사전의 확장을 통해 손쉽게 다른 온톨로지에도 적용이 가능하다.

제안한 방법은 범용성이 좋다는 장점을 지니고 있지만 여러 가지 문제점과 보완점도 내포하고 있다. 같은 주어를 가진 문장이 연속되거나 반전, 인과 관계 등과 같은 인접한 문장의 관계를 나타내는 표현을 사용할 수 없다는 단점이 있다. 이는 추후 담화 정보를 분석하는 단계를 추가하여 문장 사이에 접속사, 부사 등을, 추가하여 문장 간의 연결을 자연스럽게 할 수 있을 것이다. 간혹 주제어와 연결된 다른 목적어에 관련된 내용이 많이 나오는 경우가 생긴다. 이 경우 본 연구에서 제안하는 문장 정렬 방법을 이용하면 주제어에 대한 정보가 분산되어 오히려 주제에 관한 내용을 이해하는데 방해가 될 수 있다. 이러한 문제는 추후 문장 정렬화 과정에서 상단으로 올라오는 템플릿이 많은 경우 일부만 올리거나 전체를 올리지 않고 이를 소주제로 하여 새로운 단락을 쓰는 것을 고려해볼 수 있다.

사사

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구이며 (No.R7117-16-0209, 어떤 주제에 대한 빅데이터를 스마트 보고서로 요약하는 기술 개발), 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음 [B0126-16-1002, 개방형 미디어 생태계 구축을 위한 시맨틱 클러스터 기반 시청상황 적응형 스마트방송 기술 개발]

참고문헌

[1] I. Androutsopoulos, G. Lampouras, D. Galanis “Natural Language Descriptions from OWL Ontologies: the NaturalOWL System”, Journal of Artificial Intelligence Research, No. 48, pp. 671-715, 2013.

[2] Y.-J. Han, S.-Y. Park, S.-B. Park, Y.-H. Lee, K.-Y. Kim "Reconstruction of people information based on an event ontology", In *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering*, pp.446-451 2007

[3] 이강천, 서정연 “의미 중심어에 기반한 한국어 문장 생성 시스템”, 정보과학회논문지, 제4권, 제5호, pp.718-727, 1998