Journal of Korean Institute of Intelligent Systems
Vol. 25, No. 1, February 2015, pp. 063-069
http://dx.doi.org/10.5391/JKIIS.2015.25.1.063

ISSN(Print) 1976-9172
ISSN(Online) 2288-2324

한국어 자가 지식 학습을 위한 패턴 및 인스턴스 생성

Pattern and Instance Generation for Self-knowledge Learning in Korean

윤희근* · 박성배*†
Hee-Geun Yoon, and Seong-Bae Park†

*경북대학교 IT대학
† College of IT Engineering, Kyungpook National University

요 약

웹의 비구조 문서로부터 자동으로 인스턴스를 생성하기 위한 다양한 연구가 제안되었다. 영어권의 기존 연구들에서는 간단한 규칙과 정규식 기반의 패턴을 활용하였다. 영어에서는 단순한 정규식 기반의 패턴만으로도 충분히 높은 정확도를 보여주었지만, 한국어는 영어와 다른 언어적인 특성으로 인하여 기존의 정규식 형태의 패턴으로는 적합한 패턴을 생성할 수 없다. 이에 본 논문에서는 한국어에 적합한 패턴 및 인스턴스 생성 방법을 제안한다. 제안한 방법은 대상 문장의 의존 관계를 고려함으로써 높은 정확도를 가지는 패턴 집합을 생성한다. 또한 인스턴스의 주어(subject)와 목적어(object) 판별을 위하여 조사 정보를 함께 활용함으로써 한국어의 자유로운 어순으로부터 오는 제약을 해결한다. 실험 결과에 따르면 본 논문에서 제안한 패턴 생성 방법이 단순 어순만을 고려하여 생성된 패턴들에 비하여 더 높은 정확률을 보여주어, 한국어 대상 자동 인스턴스 생성에 적합함을 확인하였다.

키워드 : 패턴 생성, 자가 지식 학습, 지식 추출, 지식 베이스 확장, 트리플 인스턴스 생성

Abstract

There are various researches which proposed an automatic instance generation from freetext on the web. Existing researches that focused on English, adopts pattern representation which is generated by simple rules and regular expression. These simple patterns achieves high performance, but it is not suitable in Korean due to differences of characteristics between Korean and English. Thus, this paper proposes a novel method for generating patterns and instances which focuses on Korean. A proposed method generates high quality patterns by taking advantages of dependency relations in a target sentences. In addition, a proposed method overcome restrictions from high degree of freedom of word order in Korean by utilizing postposition and it identifies a subject and an object more reliably. In experiment results, a proposed method shows higher precision than baseline and it is implies that proposed approache is suitable for self-knowledge learning system.

Key Words : Pattern generation, Self-knowledge learning, Knowledge retrieval, Knowledge base expansion. Triple instance generation

1. 서 론

시맨틱 웹의 등장으로 기존에 인터넷 자연어로 존재하는 수많은 정보들에 메타정보를 부착하여, 컴퓨터가 이해할 수 있고 계산 가능한 정보를 생성하기 위한 요구가 크게 증대하였다. 대표적으로 DBPedia, Freebase와 같은 지식 베이스들은 Wikipedia와 같은 웹에 존재하는 수많은 정보들을 각자의 정의된 구조에 맞도록 정형화하여 제공하고 있다.

구조적인 지식을 구축하는 일은 노동집약적인 일로 매우 많은 비용이 소모된다. 그래서 기존의 지식 베이스들은 이 문제를 해결하기 위하여 크라우드소싱 방법을 채택하였다. 인터넷을 사용하는 사용자라면 누구든 정보를 생성, 수정할 수 있도록 함으로써 대규모의 인력을 동원하였다. 하지만 크라우드소싱 방법은 참여하는 사용자의 수에 의존적이기 때문에, 영어, 독일어와 같은 몇몇 주요 언어를 제외하고는 실질적으로 큰 효과를 보지 못하고 있다. 이런 이유로 웹상에 존재하고 있는 대량의 정보를 자동으로 지식 베이스에

Received: Jan. 18, 2014
Revised : Jan. 21, 2015
Accepted: Feb. 12, 2015
† Corresponding author(sbpark@sejong.knu.ac.kr)
본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [10044494, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발].

부착하기 위한 연구가 꾸준히 이루어지고 있다.

일반적으로 지식 베이스는 특정 개념을 대표하는 컨셉과, 이 컨셉들 사이의 관계를 정의하는 릴레이션으로 구성되어 있다. 이 관계들은 흔히 주어(subject)와 릴레이션, 그리고 목적어(object)로 구성된 트리플로 표현이 되는데, 지식 생성 연구들에서는 이 트리플 관계를 가지는 실제 인스턴스를 생성하는 것을 목표로 한다. 이를 위하여 각 트리플 관계를 표현하는 패턴들을 정의하고, 이 패턴들을 이용하여 지식을 생성하는 과정을 거치게 된다.

기존의 지식 생성 연구는 크게 두 개의 카테고리로 분류된다. 하나는 상향식 방법으로, 특정 지식 베이스를 고려하지 않고 지식을 추출하기 위한 대상 문서만을 고려하는 방법이다. 이 방법에서는 주어진 대상 문서들을 다양한 자연어처리 기술을 통하여 분석한다. 그리고 사용자가 분석된 결과들을 바탕으로 특정 릴레이션 관계를 나타내는 패턴을 정의하고 이 패턴을 이용하여 새로운 인스턴스들을 생성한다. 이 방법은 특정 지식 베이스에 의존하지 않기 때문에 릴레이션에 대한 제한 없이 다양한 지식을 생성할 수 있다는 장점이 있다. 하지만 이를 위하여 사용자가 각 릴레이션에 대한 패턴을 직접 정의해야하기 때문에 각 릴레이션에서 나타날 수 있는 다양한 패턴들을 모두 구축하기가 힘들다는 단점이 존재한다. 또한 사용자가 선택한 릴레이션들은 지식 베이스에 존재하는 릴레이션들과 아무런 관계없이 선정되었기 때문에, 최종적으로 생성된 트리플들을 지식 베이스에 저장하기 위해서는 사용자가 정의한 릴레이션과 지식 베이스 릴레이션들 사이의 관계를 식별해야 하는 과정이 필요하다.

지식 생성을 위한 다른 방법은 하향식 방법으로, 주어진 지식 베이스에 정의되어 있는 릴레이션만을 대상으로 새로운 인스턴스를 생성하는 방법이다. 이는 NELL[7]과 같은 자가 지식 학습 프레임워크에서 채택된 방법으로, 주어진 지식 베이스에 이미 구축되어 있는 인스턴스 정보들을 활용한다. 이들을 이용하여 릴레이션 관계를 나타내는 패턴들을 자동으로 추출한다. 그리고 이 패턴들을 이용하여 새로운 인스턴스들을 자동으로 생성한다. 이 방법은 대상 릴레이션 자체를 대상 지식 베이스에 존재하는 것들만 고려하기 때문에 상향식 방법에 비하여 생성할 수 있는 릴레이션의 수가 제약이 된다는 단점이 존재한다. 하지만 지식을 생성하는 단계에서 이미 지식 베이스의 대상 릴레이션이 정해져 있기 때문에 인스턴스를 지식 베이스에 부착할 때, 릴레이션들의 관계를 식별해야 하는 작업이 요구되지 않는다. 또한 본 방법은 지식 추출 과정을 통해 생성한 지식을 재활용하여 추가적인 패턴을 생성하는 것이 가능하기 때문에 패턴 생성을 위한 큰 비용이 요구되지 않는다는 장점이 있다. 이를 통해 하향식 방법으로는 사용자의 개입 없이 시스템이 스스로 끊임없이 새로운 지식을 학습할 수 있는 자가 지식 학습이 가능하다.

본 논문에서는 한국어를 대상으로 자연어 문장으로부터 온톨로지의 인스턴스를 자동으로 생성하여 확장할 수 있는 자가 지식 학습 프레임워크를 제안한다. 구체적으로 한국어 대상 패턴 생성 및 지식 생성 모델을 제안한다. 생성된 패턴이 너무 구체적이면 매칭되는 사례가 거의 존재하지 않아 다른 인스턴스 생성에 활용할 수 없다. 반대로 너무 일반적인 패턴을 생성하면 이는 수많은 오류를 발생시키게 된다. 이에 본 논문에서는 한국어를 위한 의존 관계 트리 구조와 조사 정보에 기반한 패턴 생성 방법을 제안한다. 이를 통해 매우 구체적이지는 않지만 높은 정확도를 보이는 패턴 및 인스턴스 생성 방법을 제안한다. 실험 결과에 따르면, 단순히 어순만을 고려하여 생성한 패턴보다 높은 정확도를 가지는 패턴을 생성하고, 이를 바탕으로 생성한 인스턴스도 더 높은 정확률을 보임을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서 기존의 인스턴스 자동 생성을 위한 연구들에 대해서 살펴본다. 3장에서는 본 논문에서 제안하는 패턴 생성 방법에 관해서 기술한다. 4장에서는 제안한 방법의 성능을 보이기 위한 실험 설정과 결과에 대해서 설명한다. 그리고 5장에서 결론을 맺는다.

2. 관련 연구

다양한 도메인에서 비구조 문서로부터 자동으로 인스턴스를 생성하여 온톨로지를 확장하기 위한 많은 연구가 제안되었다. Juana et al[1]은 여행 도메인의 온톨로지를 확장할 수 있는 시스템을 제안하였다. 해당 시스템에서는 주어진 비구조 문서에서 개체명을 인식한 다음 해당 개체명을 온톨로지의 컨셉에 부착하는 과정으로 온톨로지를 확장하였다. 개체명의 모호성 해결과 개체명들 사이의 관계 분석등을 위하여 형태소 분석, 구문 구조등의 정보를 활용하였다.

Ontosophie[2]는 비구조 문서로부터 온톨로지의 각 컨셉의 인스턴스가 될 수 있는 후보들을 생성하고 이들을 온톨로지에 부착하는 메소드를 제안하였다. 비구조 문서에서 인스턴스 후보를 생성하기 위하여 Crystal[3]을 이용하였다. Crystal은 주어진 학습 코퍼스로부터 각 컨셉에 해당하는 개체들을 추출할 수 있는 룰을 생성해준다. 이 룰을 이용하여 비구조 문서로부터 컨셉의 인스턴스 후보들을 생성하고, 간단한 룰에 기반하여 충돌을 해결하여 온톨로지 확장을 수행한다.

Clara et al[4]는 도메인에 독립적인 온톨로지 확장 시스템을 제안하였다. 본 시스템에서는 비구조 문서의 어휘, 개체명, 동일지시어 분석 등을 통해 인스턴스의 후보들을 생성하고, 이들을 다양한 기계학습 방법을 적용하여, 해당 인스턴스 후보들의 온톨로지 컨셉을 분류하는 방법으로 접근하였다.

강문수[8]는 웹문서의 구조화된 정보를 바탕으로 자동으로 온톨로지 인스턴스의 속성 정보를 추출하는 방법을 제안하였다. 웹문서에서 구조화된 정보를 추출하기 위하여 테이블 구조를 추출한다. 그리고 이 테이블에서 속성의 의미를 나타내는 헤더와, 속성값을 나타내는 셀들의 결합 형태로 데이터를 가공하여 온톨로지에 부착하기 위한 형태로 생성하였다.

자연어로부터 자동으로 트리플 형태의 인스턴스를 생성하기 위한 많은 연구도 많이 제안되었다. Hearst patten[5]는 상하위어 관계를 가지는 인스턴스들을 생성하기 위한 패턴 생성 방법이다. 이 방법에서는 기본적으로 관계를 찾고자 하는 릴레이션에서 주어와 목적어에 해당하는 단어가 존재하는 문장을 찾고, 이 문장에서 두 객체 사이에 존재하는 단어들과 품사 정보를 이용한다. 그리고 이 패턴으로 정의되는 간단한 룰들을 활용하여 상하위어 관계를 생성할 수 있는 패턴을 정의하였다. 이 접근법은 매우 단순하지만 영어권에서 매우 높은 정확도를 보임으로써, 다른 릴레이션들을 찾기 위한 연구들에서도 많이 적용되었다[6]. CPL[7]은 대표적인 자가 지식 학습 프레임워크인 NELL의 모듈 중 하나로, 비구조 문서로부터 패턴 및 지식 생성을 담당하는

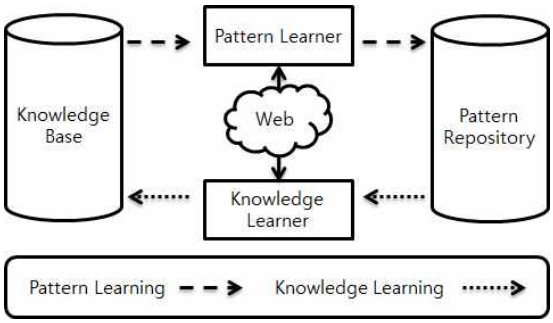

그림 1. 자가 지식 학습 프레임워크
Fig. 1. Framework of self-knowledge learning

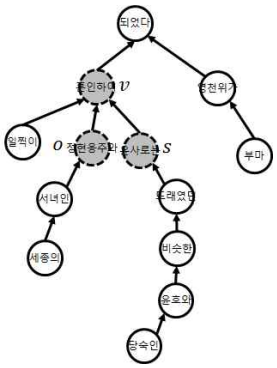

그림 2. 패턴 생성을 위한 대상 노드 선택
Fig. 2. Target nodes election for pattern generation

모듈이다. 이 모듈은 hearst pattern과 매우 유사하게 주어진 패턴 생성을 위해 주어진 인스턴스에서 주어와 목적어 단어 사이에 포함되어 있는 단어를 그대로 활용하여 패턴을 정의한다. 이 패턴들을 정규화 형태로 정의하여, 해당 패턴이 매칭되는 문장으로부터 새로운 인스턴스를 생성한다.

3. 한국어 자가 지식 학습을 위한 자연어 패턴 및 지식 생성

3.1 자가 지식 학습 프레임워크

자가 지식 학습이란 대상으로 지식 베이스에 정의된 릴레이션 관계를 가지고 있는 인스턴스들을 스스로 생성하는 방법이다. 이 과정은 한번으로 끝나지 않고, 계속 반복되어 수행된다. 자가 지식 학습은 패턴 생성 과정과 지식 생성 과정 구성되는데, 이 두 과정이 반복적으로 수행됨으로써 인스턴스를 생성한다. 그림 1은 자가 지식 학습 프레임워크의 개념도이다.

자가 지식 학습 프레임워크에는 크게 2가지의 학습 방향이 존재한다. 한 가지는 패턴 학습 과정이다. 이는 주어진 지식 베이스의 특정 릴레이션 관계를 표현하는 자연어 표현을 수집하는 과정이다. 이를 통해 각 릴레이션을 표현하는 패턴들을 수집하고, 새로운 인스턴스를 생성하는데 활용할 수 있게 한다. 또 다른 학습 과정은 지식 학습 과정으로 각 릴레이션들의 패턴이 주어지면 이 패턴으로부터 동일한 릴레이션을 가지는 새로운 인스턴스들을 생성하는 과정이다. 이렇게 생성된 인스턴스들은 다시 지식 베이스에 저장되어 새로운 패턴을 학습하기 위한 지식으로 활용된다.

자가 지식 학습 프레임워크에서 가장 중요한 부분은 너무 구체적이지 않으면서도 릴레이션의 관계를 정확하게 추출할 수 있는 패턴을 생성하는 것이다. 패턴에 오류가 존재하게 되면 이는 잘못된 지식을 생성하게 된다. 이렇게 발생된 오류는 계속 누적되어 완전히 잘못된 방향으로 학습이 수행되게 된다. 그렇기 때문에 각 학습 단계에서 오류를 최대한 발생시키지 않을 정확한 패턴을 생성하는 것은 매우 중요한 문제이다.

3.2 한국어 자가 지식 학습을 위한 패턴 생성

한국어는 영어와 달리 어순이 매우 자유로우며 또한 어휘들의 활용형이 많아 영어권에서 많이 적용된 단순 구문 정보만을 이용한 패턴 생성 방법은 적합하지 않다. 이에 본 논문에서 한국어에 적합한 패턴을 생성하는 것을 목표로 한다. 본 논문에서 제안하는 패턴은 릴레이션을 표현하는 술어 부분과 주어, 목적어를 판별하기 위한 조건d을 기술하는 3개의 튜플로 구성된다.

패턴을 생성하기 위한 과정은 주어, 목적어 판별 조건 선택 과정과 술어 표현 선정 단계로 구성된다. 패턴 생성의 첫 번째 과정은 주어와 목적어 판별 조건을 추출하는 것이다. 영어의 경우는 어순이 고정되어 있기 때문에 정규식과 같은 형태로 표현된 패턴 매칭으로도 주어와 목적어를 정확하게 판별할 수 있다. 하지만 한국어는 어순이 자유롭기 때문에 주어와 목적어에 해당하는 대상들이 다양한 어순을 가지고 나타날 수 있다. 한국어에서는 이러한 정보를 조사를 확인함으로써 판별이 가능하다. 이 정보를 활용하기 위하여 문장에서 주어와 목적어에 해당하는 단어가 포함된 어절을 선택한 후, 각 어절의 조사 정보를 추출하여 패턴의 각 튜플을 구축한다.

술어 표현 선정 단계는 릴레이션을 표현하는 적합한 술어를 선택하는 단계이다. 일반적으로 지식 베이스의 릴레이션 관계는 자연어 문장에서 술어로 표현되는데, 한국어는 영어와 달리 두 대상 사이의 관계를 설명하는 술어의 위치가 정형화되어 있지 않다. 만약 문장에 하나 이상의 술어가 존재한다면, 두 대상의 관계를 표현하는 술어를 선택해야 한다.

이 문제를 해결하기 위하여 본 논문에서는 문장의 의존 관계 정보를 이용한다. 의존 관계 정보는 파스트리에서 구단위 정보와 같은 상세 구조를 생략하고, 단어들 사이의 의존 관계 정보만을 이용하여 구성함으로써, 본 논문에서 필요로 하는 단어 사이의 관계 정보만을 파악하기에 적합한 형태이다. 이 의존 관계 트리구조에서, 인스턴스에 포함된 두 대상의 관계를 표현하는 술어는 두 대상에 해당하는 노드와 가장 가까이 위치하게 된다. 아래 수식을 통해 이 특성이 반영된 술어 v^*를 선택할 수 있다.

$$v^* = \min_{v \in V} len(n_s, n_v) + len(n_o, n_v)$$

위의 수식에서 s, o, v는 주어, 목적어, 술어를 의미하며, n_s, n_o, n_v는 의존 관계 트리에서 각각 주어, 목적어, 술어의 노드를 의미한다. $len()$는 주어진 두 노드 사이의 거리를 측정하는 함수이다.

다음의 예를 통해 패턴 생성 과정을 살펴보자. ‘isSpouseOf’ 라는 릴레이션의 패턴을 생성하기 위하여 해

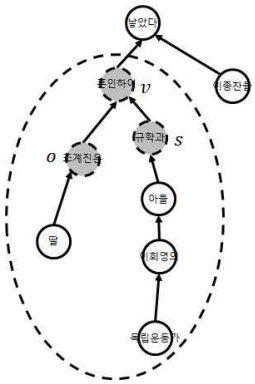

그림 3. 인스턴스 생성을 위하여 선택된 대상 노드
Fig. 3. Selected nodes for instance generation

당 릴레이션 관계를 가지고 있는 인스턴스 (윤사로, isSpouseOf, 정현옹주)가 주어진다. 이 릴레이션을 포함하고 있는 문장을 찾기 위하여 웹으로부터 해당 인스턴스의 주어와 목적어에 해당하는 (윤사로, 정현옹주)를 포함한 문장, '당숙인 윤호와 비슷한 또래였던 윤사로는 일찍이 세종의 서녀인 정현옹주와 혼인하여 부마 영천위가 되었다.' 문장을 추출한다. 의존 관계 분석을 통하면 이 문장으로부터 그림 2와 같은 트리 구조를 생성할 수 있다.

해당 트리로부터 주어와 목적어 판별을 위한 조건 추출을 위하여 주어 단어가 포함된 노드 '윤사로는'와 목적어가 포함된 노드 '정현옹주와'를 추출한다. 이를 바탕으로 (는, 와, ?)의 패턴이 생성된다. 다음 단계로 해당 패턴의 술어를 선택하기 위하여, 앞에서 선택된 두 노드로부터 가장 가까운 술어인 '혼인하여'를 선택한다. 이를 바탕으로 본 문장에서 최종적으로 (는, 와, 혼인하여)의 패턴을 생성하게 된다.

3.3 한국어 자가 지식 학습을 위한 인스턴스 생성

인스턴스 생성 단계는 패턴 생성 단계에서 생성된 패턴을 이용하여 새로운 인스턴스를 생성하는 과정이다. 본 논문에서 제안하는 패턴을 이용하여 새로운 인스턴스를 생성하는 과정은 패턴 생성 단계와 반대의 방향으로 이루어진다.

본 논문에서 제안하는 패턴은 문장에 포함되어 있는 릴레이션을 술어로 표현한다. 그러므로 패턴이 주어지면, 해당 패턴에서 술어에 해당하는 단어가 포함되어 있는 문장을 찾아냄으로써 해당 릴레이션에 해당하는 인스턴스 후보가 있는 문장을 찾는다. 두 번째 단계에서는 해당 릴레이션의 주어와 목적어에 해당하는 대상을 추출하는 것이다. 이들은 추출하기 위하여 패턴 생성에서 술어 선정의 원리와 동일한 방법을 적용한다. 의존 관계 트리에서 특정 술어의 관계를 가지는 주어와 목적어는 해당 술어 노드의 자손 노드로 구성된다. 그렇기 때문에 주어, 목적어 노드의 후보를 제약하기 위하여 의존 관계 트리에서 매칭된 술어를 루트로 하는 부분 트리를 추출한다. 그리고 최종적으로 주어와 목적어를 추출하기 위하여 패턴의 각 조건에 맞는 노드들을 선택한다. 만약 주어와 목적어의 조건에 일치하는 노드가 다수 존재할 경우, 술어로부터 더 가까운 노드들을 선택한다. 그림 3는 앞에서 패턴을 학습한 'isSpouseOf' 릴레이션의 새로운 인스턴스 생성 예이다.

앞서 학습된 (는, 혼인하여, 와) 패턴의 술어에 해당하는 '혼인하여'가 포함된 문장 '딸 조계진은 독립운동가 이회영의 아들 규학과 혼인하여, 이종찬을 낳았다.'을 추출한다. 해당 문장에서 매칭된 술어 '혼인하여'가 루트가 되는 부분 트리를 추출한다. 다음 주어와 목적어의 조건에 부합하는 어절 '규학과', '조계진은'을 추출함으로써, (규학, isSpouseOf, 조계진)이라는 새로운 인스턴스를 생성한다.

표 1. 실험을 위한 기초 지식 통계
Table 1. Simple statistics of pre-knowledge base

Relation	# of Instances
born	12
make	7
isChildOf	23
isSpouseOf	5
isAppointed	4
mountThrone	2

표 2. 릴레이션들의 기초 인스턴스 예
Table 2. Initial instance examples of relations

Subject	Relation	Object
이방원	born	1367년
최무선	make	신기전
이도	isChildOf	태종
신숙주	isSpouseOf	무송윤
이도	isAppointed	충년군
충녕내군	mountThrone	세종

4. 실험

본 논문에서 제안한 패턴 생성 및 지식 생성 방법론의 우수성을 보이기 위하여 실험을 수행하였다. 첫 단계의 패턴 생성을 위한 초기 지식은 인물과 관련있는 총 6개의 릴레이션에 대해서 수행되었다. 각 릴레이션의 인스턴스들은 임의로 생성된 트리플들이 사용되었다. 표 1은 실험을 위하여 선택된 릴레이션 및 각 릴레이션 별로 제공된 인스턴스의 통계를 보여준다. 그리고 표 2는 릴레이션 별 입력으로 사용된 기초 인스턴스의 예를 보여준다.

패턴 생성 및 새로운 트리플 생성을 위한 대상 비구조 문서는 위키피디아에 존재하는 문서를 대상으로 하였다. 한국어 위키피디아 문서 중, 총 25,000개의 문서를 임의로 선택하였다. 이 문서들을 ETRI의 자연어 처리 분석툴을 통하여 전처리를 수행하였다. 분석 결과 중, 형태소, 품사, 개체명 및 의존 관계 정보를 사용하였다. 위키 문서 중 인물에 관한 문서에서는 본문에서 대상 인물의 이름 대신에 대명사로

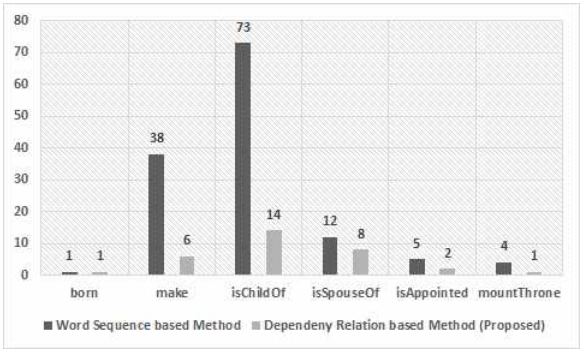

그림 4. 패턴 생성 개수
Fig. 4. Number of generated patterns

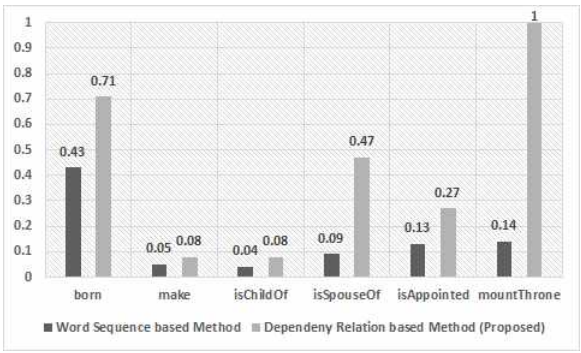

그림 5. 생성된 패턴의 정확률
Fig. 5. Accuracy of generated patterns

나타나는 경우가 많아, 추출된 인스턴스에 고유명사 대신에 대명사로 나타나는 경우가 존재한다. 이런 인스턴스들은 정보로써 가치를 가질 수 없기 때문에 적절한 고유명사로 복원을 해 주어야 한다. 하지만 대명사 복원은 매우 어렵고 본 논문의 범위를 벗어나는 문제이기 때문에, 본 논문에서는 단순하게 그, 그녀와 같은 인물을 나타내는 대명사를 해당 위키문서의 제목으로 대치하는 방법을 적용하였다.

본 실험에서는 성능 비교를 위하여 파스트리 정보를 활용하지 않고, 어순에만 기반하여 술어을 선정하는 메소드와 비교 실험을 수행하였다. 비교 메소드는 문장에서 주어진 인스턴스의 주어와 목적어에 해당하는 명사구를 찾은 다음, 두 단어 이후에 가장 가깝게 나타나는 동사를 선택하여 패턴을 생성하였다. 주어와 목적어 판별을 위한 조사 정보는 두 방법 모두 동일하게 적용하였다. 그림 4는 두 방법을 통해 생성된 릴레이션별 패턴의 수를 보여준다.

실험 결과에 따르면 릴레이션 'born'을 제외한 5개의 릴레이션에 대해서 비교 메소드가 제안한 방법에 비해 더 많은 수의 패턴을 생성하였다. 이는 비교 방법이 의미적인 관계는 전혀 고려하지 않고, 단지 표면적인 정보에만 기인하여 패턴을 생성하기 때문에 잘못된 패턴들이 대량으로 생성된 결과로 해석된다. 이를 확인하기 위하여 릴레이션 별로 생성된 패턴들에 대한 정확률을 측정해보았다. 그림 5는 생성된 패턴들의 정확률을 보여준다.

평가 결과에 따르면 제안한 패턴 생성 방법이 적은 수의 패턴을 생성하지만 훨씬 높은 정확률을 보임을 확인할 수 있다. 즉 비교 메소드가 추가적으로 생성한 대부분의 패턴들이 오류이다. 하지만 제안한 방법에서는 의존 관계 분석 결과를 통해 문장의 의미적인 관계까지 활용하기 때문에 노

표 3. 생성된 패턴들의 예
Table 3. Examples of generated patterns

Relation	Subject	Object	Verb
born	은, 는	의	딸이
make	은, 는	을, 를	내놓
isChildOf	은, 는	와, 과	혼인하
isSpouseOf	은, 는	에	태어나
isAppointed	이, 이가	로, 으로	즉위하
mountThrone	은, 는	에	봉하

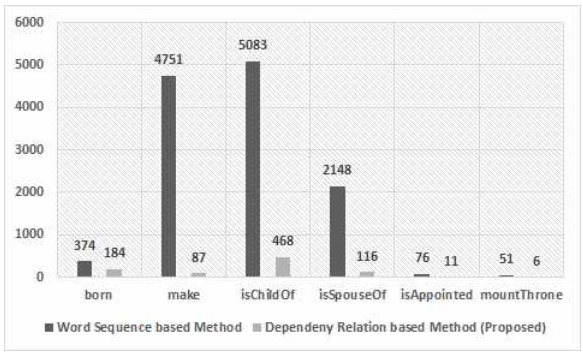

그림 6. 인스턴스 생성 개수
Fig. 6. Number of generated instances

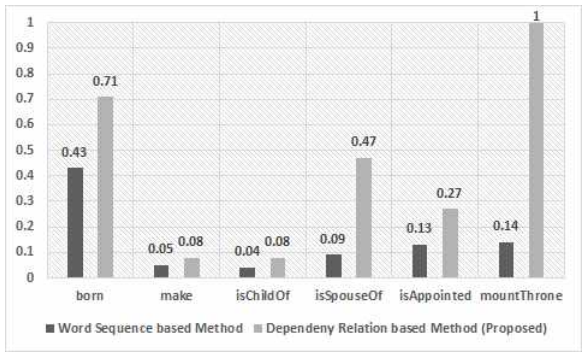

그림 7. 생성된 인스턴스의 정확률
Fig. 7. Accuracy of generated instances

이즈 패턴들이 제외되었다. 표 3은 제안한 메소드에서 생성된 패턴들의 예제이다.

다음은 생성된 패턴들을 이용하여 새로운 인스턴스를 생성하는 과정을 수행해 보았다. 비교모델은 패턴 생성과 유사하게, 패턴의 술어가 포함된 문장을 추출한 후, 매칭된 술어에서 주어와 목적어 조건에 부합하는 가장 가까운 어절들을 선정하여 추출하였다. 그림 6은 비교 방법과 제안한 방법을 통해 생성된 인스턴스의 수를 보여준다.

패턴 생성 결과와 유사하게 모든 릴레이션에 대해서 비교 메소드가 제안한 방법에 비해 월등히 많은 수의 인스턴스를 생성하였다. 이는 비교 방법이 제안한 방법에 비하여 월등히 많은 수의 패턴을 생성하였기 때문이다. 하지만 지식도 패턴과 마찬가지로 생성된 수보다 정확률이 중요한 요

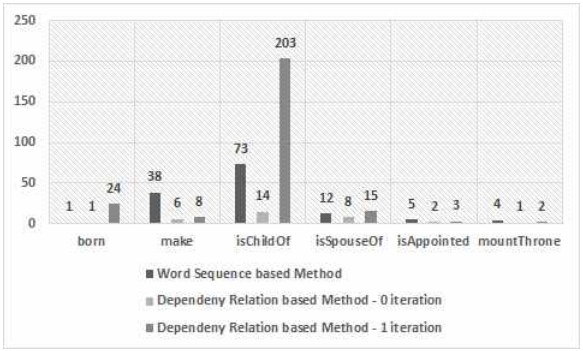

그림 8. 반복 수행 후, 패턴 생성 수
Fig. 8. Number of generated pattern after iteration

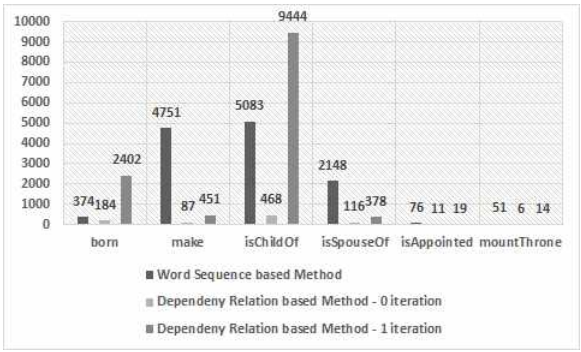

그림 9. 반복 수행 후, 인스턴스 생성 수
Fig. 9. Number of extracted instances after iteration

소이다. 정확률을 체크하기 위하여 생성된 인스턴스들에 대하여 평가를 수행하였다. 비교 방법의 경우는 생성된 지식의 수가 너무 많아 모든 결과를 평가하기가 쉽지 않기 때문에 'make', 'isChildOf', 'isSpouseOf'에 대해서는 각 200개씩의 인스턴스를 임의 표본 추출하여 정확률을 측정하였다.

평가 결과에 따르면 제안한 방법이 비교 방법에 비하여 훨씬 높은 정확률을 보임을 확인할 수 있었다. 비교 메소드의 경우 많은 수의 인스턴스를 생성하였으나 대부분이 틀린 인스턴스였다. 특히 본 결과를 통해 생성된 패턴의 정확률이 낮았던 릴레이션들의 경우 생성된 지식의 정확률도 매우 낮음을 확인할 수 있었다. 이 실험 결과를 통해 정확률이 높은 패턴 생성의 필요성을 확인할 수 있다.

제안한 방법이 비교 방법에 비하여 더 높은 정확률을 가지는 패턴 및 지식을 생성하는 것을 확인하였으나 지식 생성에서 재현율 또한 매우 중요한 요소이다. 아무리 정확한 지식을 생성하더라도 그 수가 적다면 이는 활용될 수 없다. 이에 본 논문에서 제안한 방법이 실제로 자가 지식 학습 프레임워크의 반복 학습 과정을 거치면 낮은 재현율이 극복될 수 있음을 확인하기 위한 실험을 수행하였다. 이를 위하여 제안한 방법을 통해 생성된 지식들을 재이용하여 패턴 생성과 지식 생성 과정을 수행하였다. 그림 8, 9은 반복 과정을 통해 생성된 패턴과 지식 수를 이전 결과들과 비교하여 보여준다.

실험 결과에 따르면, 제안한 방법이 각 단계에서 생성하는 패턴 및 지식의 수는 적지만 반복 과정을 통해 충분히 낮은 재현율을 극복할 수 있음을 볼 수 있다. 이를 통해 본 논문에서 제안한 패턴 및 지식 학습 방법이 자가 지식 학습 시스템 적합함을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 한국어를 대상으로 자가 지식 학습을 위한 패턴 생성 및 지식 추출 방법에 대하여 제안하였다. 최근 많은 지식 베이스에서 크라우드소싱에 기반한 인스턴스 생성을 수행하고 있다. 하지만 몇몇 주요 언어를 제외하고는 큰 실효를 얻지 못하고 있다. 이를 해결하기 위해서 웹에 존재하는 대량의 문서로부터 자동으로 지식을 생성하기 위한 연구가 제안되고 있다.

기존의 지식 추출에 관한 연구는 대부분 영어권을 대상으로 수행되었다. 영어권의 연구에서는 단순한 규칙에 기반한 정규식 형태의 패턴만으로도 높은 정확률을 가지는 패턴과 지식을 생성하였다. 하지만 한국어는 영어에 비해 어순과 활용이 자유로워 이런 단순한 형태로는 정확한 패턴을 생성할 수 없다. 이에 본 논문에서는 파스 트리와 조사 정보를 정보를 활용한 패턴 생성 및 지식 생성 방법을 제안하였다. 간단한 규칙 기반의 패턴 생성 방법과 비교해본 결과 본 논문에서 제안한 방법이 훨씬 높은 정확률을 보임을 확인할 수 있었다. 또한 낮은 재현율도 반복 학습 과정을 거침으로써 극복할 수 있음을 확인하였다. 이를 바탕으로 본 논문에서 제안하는 패턴 생성 및 인스턴스 생성 방법이 한국어 자가 지식 학습을 위한 시스템 구축에 충분히 적용될 수 있음을 확인하였다.

향후 연구에서는 좀 더 다양한 정보를 활용한 패턴 생성에 관한 연구가 필요할 것으로 생각된다. 실험에서 비교 방법에 비하여 높은 정확률을 보였으나, 아직까지 작지 않은 오류율이 존재한다. 비록 각 단계에서 오류율이 낮다 할지라도, 반복 과정을 거침으로 인해 지식 베이스가 완전히 잘못된 방향으로 학습이 수행될 수 있다. 이를 해결하기 차후에는 더 다양한 특징 정보를 활용하여 패턴 및 인스턴스 생성 결과의 정확도를 향상시키는 연구를 수행할 예정이다. 또한 현재는 생성된 패턴 및 인스턴스에 대한 신뢰도를 측정에 관한 연구를 수행하지 않았다. 만약 패턴과 인스턴스의 신뢰도를 측정하여 신뢰도가 높은 정보들만을 이용한다면 더 안정적인 시스템을 구축할 수 있다. 이를 위하여 차후에는 패턴과 인스턴스의 신뢰도를 측정할 수 있는 연구를 함께 수행할 것이다.

References

[1] Juana Mar´ıa Ruiz-Mart´ınez, Jose Antonio Min˜arro-Gime´nez Dagoberto Castellanos-Nieves, Francisco Garc´ıa-Sa´nchez and Rafael Valencia-Garc´ıa, "ONTOLOGY POPULATION: AN APPLICATION FOR THE E-TOURISM DOMAIN," *International Journal of Innovative Computing, Information and Control*, Vol. 7, No. 11, pp. 6115-6133, 2011.

[2] David Celjuska, and Dr. Maria Vargas-Vera, "Ontosophie: A Semi-Automatic System for Ontology Population from Text," *Proceedings of WOP2009 collocated with ISWC2009*, Vol. 516, 2009.

[3] Stephen Soderland, David Fisher, Jonathan Aseltine, and Wendy G. Lehnert, "CRYSTAL: Inducing a Conceptual Dictionary," *Journal of CoRR,* Vol. cmp-lg/9505020, 1995.
[4] Carla Fariaa, Ivo Serrab, Rosario Girardib, "A domain-independent process for automatic ontology population from text," *Journal of Science of Computer Programming,* 2013.
[5] Marti A. Hearst, "Automatic Acquisition of Hyponyms Large Text Corpora," *Proceedings of Conference on Computational Linguistics*, 1992.
[6] Patrick Pantel, and Marco Pennacchiotti, "Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations," *Proceedings of Conference on Computational Linguistics*, pp. 113-120, 2006.
[7] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr.2, and Tom M. Mitchell, "Toward an Architecture for Never-Ending Language Learning," *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2010.
[8] Moon-Soo Chang, and Sun-Mee Kang, "An Extraction of Property of Ontology Instance Using Stratification of Domain Knowledge," *Journal of Korean Institute of Intelligent Systems,* Vol. 17, No. 3, pp. 291-296, 2007.

저 자 소 개

윤희근(Hee-Geun Yoon)
2007년 : 경북대학교 컴퓨터공학과 졸업 (학사)
2009년 : 경북대학교 대학원 전자전기컴퓨터학부 졸업 (석사)
2009년～현재 : 경북대학교 대학원 컴퓨터학부 박사과정

관심분야 : 기계학습, 자연어처리
Phone : +82-53-940-8692
E-mail : hkyoon@sejong.knu.ac.kr

박성배(Seong-Bae Park)
1994년 : 한국과학기술원 컴퓨터과학과 졸업 (학사)
1996년 : 서울대학교 대학원 컴퓨터공학과 졸업 (석사)
2002년 : 서울대학교 대학원 컴퓨터공학과 졸업 (박사)
2004년～현재 : 경북대학교 IT대학 컴퓨터학부 교수

관심분야 : 기계학습, 자연어처리, 텍스트마이닝, 정보추출, 생명정보학
E-mail : sbpark@sejong.knu.ac.kr