트리플 필터링을 통한 한국어 자가 지식 학습 정확률 향상

이지수○, 김경훈, 최수정, 박성배, 박세영
경북대학교
{jslee, khkim, sjchoi, sbpark, sypark}@sejong.knu.ac.kr

Accuracy Improvement of Self-knowledge Learning by Filtering Triple

Jisu Lee○, Kyounghun Kim, Su Jeong Choi, Seong-Bae Park, Se-Young Park
School of Computer Science and Engineering, Kyungpook National University

요 약

자가 지식 학습 프레임워크는 자연어 텍스트에서 지식 트리플을 생성하기 위한 방법 중 하나로, 문장의 의존 관계 트리 상에서 주어 개체와 목적어 개체 사이의 관계를 패턴으로 학습해 이 패턴을 바탕으로 새로운 지식 트리플을 생성한다. 그러나 이 방법은 의존 관계 트리를 생성하는 도구의 성능에 영향을 받을 뿐만 아니라 생성된 지식 트리플을 반복적으로 사용하는 자가 지식 학습의 특성상 오류가 누적될 가능성이 있다. 이러한 문제점을 해결하기 위해서 본 논문에서는 자가 지식 학습 프레임워크에서 생성된 지식 트리플을 TransR 신뢰도 함수를 사용해 신뢰도 값을 측정하여 그 값에 따라 지식 트리플을 필터링하는 방법을 제안한다. 실험 결과에 따르면 필터링 된 지식 트리플들이 그렇지 않은 지식 트리플들에 비하여 더 높은 정확률을 보여주어, 제안한 방법이 자가 지식 학습 프레임워크의 정확률 향상에 효과적임을 증명하였다.

주제어: Knowledge embedding, knowledge graph, self-knowledge learning

1. 서론

인터넷의 발달에 따라 수많은 지식 정보들이 웹상에 등장했고, 이러한 정보들을 활용하기 위한 노력이 계속 되고 있다. 일반적으로 웹상의 지식 정보들은 자연어 텍스트로 존재하고 있기 때문에 이를 컴퓨터가 활용하기 위해서는 자연어 텍스트를 컴퓨터가 이해 할 수 있는 데이터로 정형화할 필요가 있다.

Freebase, 디비피디아 등은 지식 데이터들을 모아 놓은 대표적인 지식 베이스로, 웹상의 많은 정보들을 각자의 정의된 구조에 맞춰 정형화된 데이터로 제공하고 있다. 지식 베이스의 지식은 정형화된 구조인 지식 트리플로 표현될 수 있다. 지식 트리플은 흔히 주어와 목적어, 그리고 주어와 목적어의 관계를 서술하는 술어로 이루어진다. 이러한 구조를 가진 지식 트리플의 집합이 바로 지식 베이스가 되는 것이다.

최근에는 웹상의 정보를 자동으로 추출해 지식 베이스를 구축하기 위한 연구가 계속 이루어지고 있다. Rusu et al.[1]은 텍스트를 파싱한 뒤 몇 가지 규칙을 이용하여 트리플을 생성하는 방법을 제안하였다. 정창후 외[2]는 주어와 목적어 사이에 존재하는 관계-논항 구조를 분석하고 이를 정규화한 어휘 패턴 문자열을 생성한 후 스트링 커널을 사용하여 지식 트리플 관계를 추출하였다.

Yankai et al.[3]이 제안한 TransR은 주어 개체와 관계가 주어졌을 때 가장 적합한 목적어 개체를 찾아 지식 트리플을 구성한다. 이를 위해 개체 공간상에 존재하는 개체를 각각의 관계 후보가 속한 관계 공간에 투영하여 개체와 관계를 나타내는 지식 트리플을 예측한다. 윤희근 외[4]는 한국어를 대상으로 자가 지식 학습을 위한 패턴 생성 및 지식 트리플 생성 방법을 제안하였다. 주어진 문장을 의존 관계 트리로 변환 후 주어 개체와 목적어 개체에 존재하는 단어들과 품사 정보를 이용하여 패턴을 생성한다. 그리고 생성된 패턴을 이용하여 문장으로부터 새로운 지식 트리플을 만든다. 그러나 이 방법은 단순히 패턴의 개체들과 일치하는 단어를 찾는 것이기 때문에 의존 관계 트리 도구의 성능에 따라서 오류가 있는 지식 트리플들을 생성할 가능성이 높다. 자가 지식 학습은 이렇게 생성된 오류가 있는 지식 트리플들을 반복하여 사용하므로 오류가 누적되어 생성된 지식 트리플들의 전체적인 정확률이 낮아진다. 그러므로 지식 트리플들의 정확률을 높이기 위하여 생성된 지식 트리플들을 필터링 할 필요가 있다.

본 논문에서는 자가 지식 학습 프레임워크를 통해 생성된 지식 트리플을 필터링하여 지식 트리플들의 정확률을 향상시키는 방법을 제안한다. 자가 지식 학습 프레임워크를 통해 코퍼스에서 지식 트리플을 생성한 뒤 TransR을 신뢰도 함수를 활용하여 지식 트리플마다 신뢰도 값을 부여한 후 상위의 지식 트리플을 필터링한다. TransR 신뢰도 함수로 필터링 한 지식 트리플들의 정확률은 필터링 하기 전 지식 트리플들의 정확률과 비교해 약 1.8배 정도 높았다.

2. 지식 트리플 생성 및 필터링

그림 1은 본 논문에서 제안한 방법의 흐름도이다. 먼저 디비피디아의 지식 트리플과 위키피디아 문서를 이용하여 지식 트리플의 주어와 목적어 사이의 관계를 패턴으로 학습한다. 그 다음 자가 학습된 패턴을 사용하여 새로운 지식 트리플들을 생성한다. 본 논문에서는 자가 지식 학습의 정확률을 높이기 위하여, 생성된 지식 트리플들에 대해 TransR의 신뢰도 함수를 사용해 신뢰도 값을 측정하고 그 값을 기준으로 트리플을 필터링한다.

2.1. 지식 트리플 생성

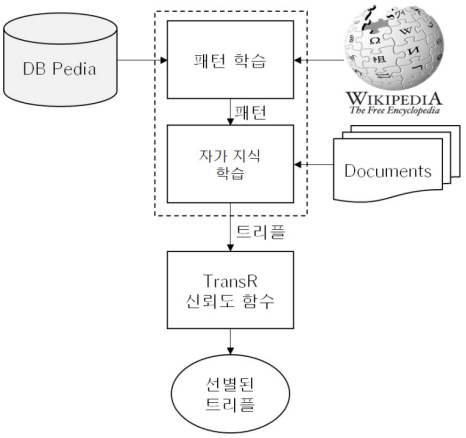

그림 1. 제안한 방법의 흐름도

우리는 지식 트리플을 생성하기 위해서 윤희근 외[4]가 제안한 자가 지식 학습 프레임워크를 사용한다. 자가 지식 학습 프레임워크는 지식 트리플들이 이미 정의되어 있는 기존의 지식 베이스를 이용하여 새로운 지식 트리플들을 반복적으로 스스로 생성하는 방법으로, 학습 과정은 크게 패턴 학습 과정과 지식 트리플 생성 과정으로 나뉜다.

먼저 패턴 학습 과정은 기존의 지식 베이스에 있는 지식 트리플들을 시드 트리플로 사용하여, 이 시드 트리플들의 주어와 목적어 사이의 특정 관계를 패턴으로 학습하는 것이다. 패턴은 주어, 목적어를 판별하기 위한 조건을 기술하는 부분과 관계를 표현하는 술어 부분으로 이루어진 튜플로 구성된다. 우리는 패턴을 생성하기 위하여 위키피디아 문서 중에서 시드 트리플의 주어와 목적어를 포함한 문장들을 추출한다. 추출한 문장을 의존 관계 트리로 변환하여 패턴을 구성하는 3개의 튜플을 찾아 정의하는 방법으로 패턴을 학습한다.

학습된 패턴은 새로운 지식 트리플을 생성하는데 사용된다. 웹상의 문서의 문장들에서 학습된 패턴과 일치하는 주어, 목적어 그리고 술어를 지식 트리플로 생성한다. 생성된 지식 트리플은 다시 패턴 학습을 위한 시드 트리플로 사용된다. 이렇게 지식 트리플을 생성하고 생성된 지식 트리플을 다시 자가 학습 과정에 사용하는 반복 과정을 통해 많은 지식 트리플들을 얻을 수 있다.

2.2. 지식 트리플 필터링

자가 지식 학습 프레임워크를 사용하여 우리는 지식 트리플을 생성하였고, 이렇게 생성된 지식 트리플들의 정확률은 비교적 높지 않았다. 왜냐하면 자가 지식 학습 프레임워크는 기본적으로 의존 관계 트리를 사용하여 패턴을 학습하기 때문에 의존 관계 트리를 생성하는 도구의 성능에 많은 영향을 받을 수 있기 때문이다. 또한 생성된 지식 트리플을 반복적으로 사용해 패턴을 학습하는 자가 지식 학습의 특성으로 인해 잘못 생성된 패턴의 오류가 누적되기도 한다. 또 다른 오류로는 지식 트리플을 생성할 때, 패턴에 있는 주어나 목적어 등이 동음이의어로 포함된 문장에서 트리플을 생성하는 오류가 있다.

이러한 오류들로 인해 지식 트리플들의 낮은 정확률을 개선하기 위해서 본 논문에서는 자가 지식 학습 프레임워크를 통해 생성된 지식 트리플들을 Yankai et al.[3]이 제안한 TransR 신뢰도 함수를 사용해 신뢰도 값을 측정하여 필터링한다. TransR의 신뢰도 함수는 특정 주어 개체와 관계가 주어졌을 때 주어진 주어 개체, 관계와 함께 올바른 지식 트리플을 형성할 가능성이 높은 목적어 개체들을 찾을 수 있게 한다. 이 때 신뢰도 값 순위대로 목적어 개체를 순서대로 나열 할 수 있고 이 순위를 기준으로 지식 트리플을 필터링한다. 제안한 방법을 통해 지식 트리플을 필터링 한다면 필터링하기 전보다 높은 정확률이 나올 것이다.

TransR에서는 지식 트리플(h,r,t)을 구성하는 주어 개체 h와 목적어 개체 t, 관계 r이 서로 독립된 공간에 존재한다고 가정되었다.

$$h_r = hM_r \ , \ t_r = tM_r \tag{1}$$

위 식(1)에서 h와 t는 각각 주어 개체와 목적어 개체를 나타내며 M_r은 개체 공간에서 관계 공간으로의 투영 행렬을 의미한다. 주어 개체와 목적어 개체는 투영 행렬을 통해 관계 공간으로 투영할 수 있다. 이렇게 얻어진 h_r와 t_r을 바탕으로 정의된 신뢰도 함수는 다음과 같다.

$$f_r(h,t) = \| h_r + r - t_r \|_2^2 \tag{2}$$

여기서 h_r과 t_r은 각각 관계 공간에 투영된 주어 개체와 목적어 개체를 나타낸다. 이 때 신뢰도 함수 식(2)를 지식 트리플에 적용하여 지식 트리플마다 신뢰도 값을 계산할 수 있는데, 신뢰도 값이 낮을수록 지식 트리플(h,r,t)이 올바른 지식을 표현하는 트리플일 가능성이 높아진다.

또한 지식 트리플들의 주어 개체와 목적어 개체의 개체 공간상에서의 위치를 학습하기 위해, 다음 식(3)과 같은 마진 기반의 함수를 사용한다.

$$L = \sum_{(h,r,t)\in S} \sum_{(h',r,t')\in S'} \max(0, f_r(h,t) + \gamma - f_r(h',t')) \tag{3}$$

위의 식에서 S는 시드 트리플을 나타내고, γ는 마진을 나타낸다. 개체의 공간상에서의 위치를 학습하기 위해 확률적 경사 하강법(Stochastic Gradient Descent)을 사용해 위의 마진 기반 함수의 L값이 최소가 되는 지식 공간 모델을 만든다.

학습된 지식 공간 모델을 바탕으로 우리는 식(2)를 통해 지식 트리플들의 신뢰도 값을 측정하여 앞서 자가 지식 학습에서 생성된 지식 트리플들을 필터링한다.

3. 실험

디비피디아의 지식 트리플을 시드 트리플로 사용하여 패턴을 학습하였다. 새로운 지식 트리플을 생성하는 과정에서는 위키피디아 문서를 사용하였다. TransR 학습은 학습 비율 0.001, 마진 1, 개체 차원 100, 학습 횟수 1,000번으로 진행하였다. 표 1은 자가 지식 학습을 통해

생성된 지식 트리플의 수를 나타내고, 5개의 관계에서 총 27,501개의 지식 트리플을 생성하였다.

표 1. 생성된 지식 트리플의 수

관 계	지식 트리플 수
Birthplace	7,541
Country	6,741
Genre	1,370
Location	843
Occupation	11,006

본 논문에서 제안한 필터링의 효과성을 보이기 위해, 먼저 생성된 지식 트리플을 각 관계별로 100개씩 랜덤 샘플링(A)하여 정확률을 평가하였다. 그 다음 TransR의 신뢰도 값을 측정해, 상위 100개의 지식 트리플(B)들을 평가하였다. 표 2는 평가 결과를 보여준다.

표 2. 지식 트리플의 정확률 평가

관 계	(A)	**(B)**
Birthplace	0.24	**0.84**
Country	0.59	**0.79**
Genre	0.38	**0.85**
Location	0.47	**0.84**
Occupation	0.54	**0.70**
평 균	0.44	**0.80**

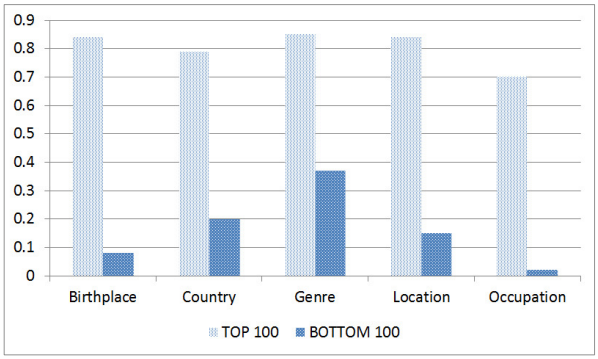

그림 2. 상위 100개와 하위 100개의 지식 트리플에 대한 정확률

실험 결과, 100개의 랜덤 샘플링 된 트리플(A)의 평균 정확률은 0.44이고 신뢰도 값 상위 100개의 트리플(B)에 대한 평균 정확률은 0.8로 약 1.8배 상승하였다. 각 5가지 관계 모두에 대해서 랜덤 샘플링 된 트리플보다 높은 정확률을 보여주었다. 또한 우리는 신뢰도 값 상위 100개와 하위 100개의 트리플에 대해서도 정확률을 평가하였다. 그림 2는 그 결과를 나타낸 그래프이다. 5가지의 관계 모두에서 상위 100개 트리플과 하위 100개 트리플의 정확률은 현저한 차이를 보였다.

이를 통해 TransR이 지식 공간 모델을 효과적으로 학습함을 알 수 있었다. 또한 학습된 지식 공간 모델을 신뢰도 값 측정에 활용하여 지식 트리플 선별에 적용하는 것이 매우 효과적임을 증명하였다.

4. 결론 및 향후연구

자가 지식 학습 프레임워크는 지식 트리플 생성 과정 중 올바르지 않은 지식 트리플이 생성될 경우 그 지식 트리플을 반복해서 사용하기 때문에 그로 인한 오류가 누적될 가능성이 크다. 그리고 의존 관계 트리를 생성하는 도구의 성능에 영향을 받기 때문에 의존 관계 트리를 사용해 패턴을 학습하는 과정에서 오류가 발생할 가능성도 있다. 이에 본 논문에서는 자가 지식 학습을 통해 생성된 지식 트리플들의 TransR 신뢰도 값을 측정하여 그 값에 따라 지식 트리플을 필터링하는 방법을 제안하였다. 5가지 관계의 필터링 한 지식 트리플 100개에 대한 정확률이 필터링 되기 전 임의로 샘플링 한 지식 트리플 100개의 정확률보다 약 1.8배 높았다. 이를 통해 본 논문에서 제안한 방법이 자가 지식 학습 프레임워크의 정확률 향상에 효과적임을 증명하였다.

또한 실험을 통해 신뢰도 값 상위 지식 트리플과 하위 지식 트리플의 정확률이 현저하게 차이나는 것을 확인할 수 있었다. 그러나 이는 필터링의 기준이 되는 신뢰도 값 없이 단지 지식 트리플의 개수를 상위, 하위 각 100개로 한정해 평가한 것이다. 그렇기 때문에 평균적으로 높은 정확률을 얻기 위해 필요한 지식 트리플의 수를 알 수 없다. 그러므로 향후에는 자가 지식 학습으로 생성된 지식 트리플들을 필터링 할 때 평균적인 정확률이 높은 지식 트리플들을 얻기 위한 임계값을 구하는 것에 대한 연구를 수행 할 것이다.

감사의 글

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원(No.R0101-15-0054, WiseKB : 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)과 2015년도 교육부 및 한국연구재단의 BK21 플러스 사업으로 지원을 받아 수행된 연구임 (No. 21A20131600005, 경북대학교 컴퓨터학부 Smart Life실현을 위한 SW인력양성사업단)

참고문헌

[1] D. Rusu, L. Dali, B. Fortuna, M. Grobelnik, and D. Mladenic, "Triplet Extraction from sentences", 10th International Multiconference, Information Society-IS, pp.8-12, 2007.

[2] 정창후, 전홍우, 송사광, 홍순찬, 정한민, 최성필, "술어-논항 구조의 어휘 패턴을 이용한 스트링 커널 기반 관계 추출", 정보과학회논문지: 소프트웨어 및 응용, 제39권, 제12호, pp.927-934, 2012.

[3] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning Entity and Relation Embeddings for Knowledge Graph Completion", Proceedings of AAAI, pp.2181-2187, 2015.

[4] 윤희근, 박성배, "한국어 자가 지식 학습을 위한 패턴 및 인스턴스 생성", 한국지능시스템학회 논문지, 제25권, 제1호, pp.63-69, 2015.