뉴로 심볼릭 기반 규칙 유도 및 추론 엔진을 활용한 지식 완성 시스템

Knowledge Completion System using Neuro-Symbolic-based Rule Induction and Inference Engine

저자 (Authors)
신원철, 박현규, 박영택
Won-Chul Shin, Hyun-Kyu Park, Young-Tack Park

출처 (Source)
정보과학회논문지 48(11), 2021.11, 1202-1210 (9 pages)
Journal of KIISE 48(11), 2021.11, 1202-1210 (9 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10662634

APA Style
신원철, 박현규, 박영택 (2021). 뉴로 심볼릭 기반 규칙 유도 및 추론 엔진을 활용한 지식 완성 시스템. 정보과학회논문지, 48(11), 1202-1210.

ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 48, No. 11, pp. 1202-1210, 2021. 11
https://doi.org/10.5626/JOK.2021.48.11.1202

뉴로 심볼릭 기반 규칙 유도 및 추론 엔진을 활용한 지식 완성 시스템

(Knowledge Completion System using Neuro-Symbolic-based Rule Induction and Inference Engine)

신 원 철 † 박 현 규 †† 박 영 택 †††
(Won-Chul Shin) (Hyun-Kyu Park) (Young-Tack Park)

요 약 최근 지식 그래프의 불완전성 문제를 해결하기 위한 다양한 지식 완성 연구중 딥러닝 학습 방법과 로직 시스템의 장점을 결합한 NTP(Neural Theorem Prover)와 같은 연구가 기존 연구들에 비해 좋은 성능을 내고 있다. 하지만 NTP는 하나의 입력에 대한 예측 결과를 얻기 위해 지식 그래프의 모든 트리플이 연산에 관여하게 되므로 대용량 지식 그래프 처리에 한계가 있다. 본 논문에서는 NTP의 계산 복잡도 문제를 개선한 모델로부터 심볼의 벡터 표현을 학습하여 규칙을 유도하고, 추론 엔진을 사용하여 유도된 규칙으로부터 지식 추론을 수행할 수 있는 딥러닝 학습 방식과 로직 추론 방식의 통합시스템을 제안한다. 본 논문에서 사용한 규칙 생성모델의 규칙유도 성능 검증을 위해 NTP와 Nations, Kinship, UMLS 데이터 셋을 대상으로 유도된 규칙을 활용한 테스트 데이터 추론가능 여부를 비교하였으며, 대규모 지식그래프인 Kdata와 WiseKB를 사용한 실험에서는 추론 엔진을 통한 지식 추론 결과 실험에 사용된 지식 그래프에 비해 각각 Kdata는 30%, WiseKB는 95%증가된 지식 그래프를 얻을 수 있었다.

키워드: 지식 그래프, 딥러닝, 로직 시스템, 추론 엔진, 지식 추론

Abstract Recently, there have been several studies on knowledge completion methods aimed to solve the incomplete knowledge graphs problem. Methods such as Neural Theorem Prover (NTP), which combines the advantages of deep learning methods and logic systems, have performed well over existing methods. However, NTP faces challenges in processing large-scale knowledge graphs because all the triples of the knowledge graph are involved in the computation to obtain prediction results for one input. In this paper, we propose an integrated system of deep learning and logic inference methods that can learn vector representations of symbols from improved models of computational complexity of NTP to rule induction, and perform knowledge inference from induced rules using inference engines. In this paper, for rule-induction performance verification of the rule generation model, we compared test data inference ability with NTP using induced rules on Nations, Kinship, and UMLS data set. Experiments with Kdata and WiseKB knowledge inference through inference engines resulted in a 30% increase in Kdata and a 95% increase in WiseKB compared to the knowledge graphs used in experiments.

Keywords: knowledge graphs, deep learning, logic system, inference engine, knowledge inference

· 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.No.2019-0-00067,대용량 지식그래프 자동완성을 위한 시맨틱 분석 추론기술 개발)

† 학생회원 : 숭실대학교 컴퓨터학과 학생
ocshin1201@naver.com
†† 학생회원 : 숭실대학교 컴퓨터학과 박사
phkchr09@gmail.com
††† 종신회원 : 숭실대학교 컴퓨터학과 교수(Soongsil Univ.)
park@ssu.ac.kr
(Corresponding author임)

논문접수 : 2021년 4월 12일
(Received 12 April 2021)
논문수정 : 2021년 10월 25일
(Revised 25 October 2021)
심사완료 : 2021년 10월 26일
(Accepted 26 October 2021)

정보과학회논문지 제48권 제11호(2021. 11)

1. 서 론

지식 그래프는 다양한 소스 로부터 수집된 사실 또는 규칙을 개체와 개체들 사이의 의미적 관계로 표현한 네트워크로써 최근 많은 인공지능 연구나 지능형 시스템에서 지식 그래프가 사용되고 있다. 지식그래프에서 지식 표현 방법은 <Subject, Predicate, Object>의 트리플 구조로 Subject와 Object는 표현하고자 하는 개체 또는 엔티티, Predicate은 두 개체간의 관계(Relation)를 나타내도록 한다. 대표 적으로 많이 사용되는 지식 그래프로는 Freebase[1], DBpedia[2], YAGO[3]를 예로 들 수 있다. 하지만 지식 그래프의 대부분이 웹으로부터 자동적으로 데이터를 수집하여 생성되기 때문에 개체나 개체 간의 관계가 누락되는 불완전성을 갖는다. 지식 그래프의 불완전성은 지식 그래프를 활용하는 연구나 시스템에 부정적인 영향을 줄 수 있다. 이러한 지식 그래프의 불완전성을 해결하기 위해 지식 그래프 자동완성에 대한 연구가 다양하게 진행 되고 있다. 지식 그래프 자동완성의 방법으로는 크게 지식 그래프 임베딩을 통한 Neural link prediction방법이 과 사용자 정의 규칙을 기반으로 하는 지식 추론 방법이 있다. Neural link prediction의 임베딩 벡터 학습방법 특성상 모델을 학습하는데 많은 양의 데이터가 필요하고, 모델의 예측 결과에 대해서 해석 할 수 없는 한계점이 있다. 사용자 정의 규칙을 기반으로 하는 지식 추론 방법인 SWRL (Semantic Web Rule Language)[4]추론 방법 또한 효율적인 추론을 위한 추론 엔진개선 연구가 진행되어 지식 그래프 완성에 사용되고 있다. 지식 그래프 완성의 최근 연구 중 신경망 학습과 로직 기반 추론 시스템의 장점을 결합한 NTP(Neural Theorem Prover)[5]가 기존 연구에 비해 좋은 성능을 내고 있다. NTP는 end-to-end Learning을 제공하며, 학습이 끝난 뒤에 해석 가능한 규칙을 유도할 수 있다. NTP의 Link prediction 결과는 우수하지만 NTP는 하나의 입력에 대한 예측 결과를 얻기 위해 지식 그래프의 모든 트리플이 관여하므로 대용량 지식 그래프에 대해서 계산적 한계가 있다. 이를 해결하기 위해 NTP의 계산 그래프를 효과적으로 감소시키고, 학습시간을 단축시킨 연구가 제안되었다. 하지만 이러한 연구는 주어진 지식 그래프의 테스트 데이터에 대한 Link prediction결과만을 제공한다. 본 논문에서는 NTP의 계산 복잡도를 개선한 모델을 활용하여 규칙을 유도하고, 유도된 규칙을 활용하여 추론 엔진을 통해 지식 추론을 수행하여 지식 그래프에 추가함으로써 완성된 지식 그래프를 얻을 수 있는 통합 시스템을 제안한다. 제안하는 시스템은 대용량 지식 그래프를 활용한 지식 완성 실험을 통해 누락된 정보의 상당량이 완성된 지식 그래프를 얻을 수 있었다.

2. 관련 연구 및 배경지식

2.1 Knowledge Graph Embedding

대규모의 지식 그래프를 다루는 것과 지식 그래프의 누락된 정보를 완성하는 것은 여전히 어려운 과제로 남아있다. 지식 그래프 완성에 있어서 확장성과 성능문제를 동시에 해결 하는 방법은 지식 그래프를 벡터 공간상에 임베딩 하는 것이다. 지식 그래프 임베딩은 특정 목적 함수를 만족하도록 지식 그래프에 존재하는 엔티티와 릴레이션에 대한 벡터표현을 학습하는 것이다. 가장 기본적인 지식 그래프 임베딩 방법으로는 TransE[6]가 있다.

TransE의 기본적인 아이디어는 지식 그래프의 엔티티집합 E, 릴레이션 집합R 로부터 $h,t \in E,\ l \in R$인 조건을 만족하는 트리플 $<h,l,t>$에 대해서 $h+l \approx t$를 만족하도록 엔티티와 릴레이션 임베딩 벡터를 학습한다. TransE는 1-to-1 릴레이션 관계에 대해서는 좋은 임베딩 성능을 보이지만 N-to-1, 1-to-N, N-to-N의 관계를 보이는 트리플을 학습하는데 문제가 있다. 문제 해결을 위해 TransR[7]은 엔티티와 릴레이션을 두 개의 서로 다른 임베딩 공간에 표현한다. TransR은 트리플 $<h,l,t>$의 각 릴레이션 l에 대하여 엔티티 임베딩 공간에서 릴리에션 임베딩 공간으로 투영할 수 있는 투영 행렬$M_l \in R^{k \times d}$을 갖는다. 이 때 k는 엔티티 임베딩 벡터의 크기이고, d는 릴레이션 임베딩 벡터의 크기이다. 이러한 맵핑 행렬을 사용하여 투영된 엔티티 벡터h_l, t_l은 $h_l = hM_l,\ t_l = tM_l$과 같다. TransE와 TransR과 같은 얕은 Neural link prediction 모델을 더 큰 지식 그래프로 확장하기 위해 효율적인 파라미터 사용과 GPU에 최적화된 연산으로 빠른 연산속도를 제공하는 Convolution neural network를 사용하는 지식 그래프 임베딩 방법으로 ConvE[8]가 제안되었다. 이 외에도 TransE 모델을 확장하는 많은 연구가 제안되었다.

2.2 Neural Theorem Prover

Neural unification을 활용하는 NTP(Neural Theorem Prover)는 심볼의 벡터 표현을 활용하여 작동하며, 지식 그래프의 질의(Query)에 대해 증명 가능한 end-to-end 학습방식의 신경망을 제안한다. 이러한 신경망은 Prolog[9]의 Backward chaining추론 방식을 사용하여 재귀적으로 구축된다. 특히 NTP에서는 심볼에 대한 벡터 표현을 활용한 Neural unification방식을 사용하여 심볼의 벡터 표현에 대한 유사도 계산함수를 사용함으로써 심볼릭 추론 과정을 심볼에 대한 벡터 표현 학습문제와 결합한다. NTP는 벡터 표현 학습에 있어 명시되지 않은 학습 가능한 릴레이션으로 표현된 규칙 템플릿을 사용하고, 이를 Parameterized rule[10]이라 정의한다.

Parameterized rule(e.g. #1(X, Y) :- #2(X, Z), #3(Z,Y))은 Prolog에서 사용하는 규칙의 기본 형식을 따른다. Parameterized rule은 ":-"를 기준으로 좌변은 결론, 우변은 전제로 구성되며 NTP는 심볼의 벡터표현 학습시 규칙 템플릿의 릴레이션(e.g. #1, #2, #3)이 지식 그래프에 존재하는 트리플의 특정 릴레이션 벡터와 유사해지도록 심볼의 벡터에 대한 학습이 이루어짐으로써 해석 가능한 논리 규칙을 유도할 수 있다.

2.3 뉴로 심볼릭 기반 규칙 생성을 통한 지식완성

지식 그래프의 불완전성 문제를 해결하기 위해 지식 완성 기법 연구가 중요하게 요구되며, 딥러닝 학습 방법을 활용하여 지식 그래프 임베딩을 통한 연구와 규칙 기반 추론을 활용하여 심볼릭 추론을 통한 지식 완성 수행과 같은 다양한 연구들이 진행되었다. 기존에 제안된 연구 중 뉴로 심볼릭 방식을 이용하여 함축적인 규칙을 명시적으로 추출하여 지식 완성에 활용한 연구가 제안되었다. 해당 연구에서 제안된 방법은 NTP의 계산량과 학습시간에 대한 문제점 개선과 지식 그래프로부터 규칙 추출을 위해 Symbolic unification기반의 릴레이션 임베딩 경로를 구하고, 릴레이션 임베딩 학습을 위한 Cost function을 정의하여 자동으로 규칙을 생성한다. NTP의 Neural unification은 심볼의 벡터 표현간 유사도비교를 통해 이루어지기 때문에 지식 그래프의 모든 트리플의 임베딩 벡터를 비교하여야 한다. 하지만 벡터간 유사도 비교 대신 심볼의 동일성 검사를 수행하는 Symbolic unification은 NTP의 계산 복잡도에 대한 문제점을 해결함과 동시에 학습속도와 규칙 유도 성능을 향상시킬 수 있다. 해당 연구는 지식 그래프의 학습 데이터를 사용하여 NTP에 비하여 더 많은 규칙을 유도하고, 유도된 규칙을 통한 테스트 데이터의 추론 가능 여부에 대한 성능을 검증 하였다. 하지만 제안된 방법은 테스트 트리플의 추론 가능 여부에 대한 실험결과만 제시하기 때문에 실제로 누락된 정보를 완성한 지식 그래프를 얻지 못한다. 본 논문에서는 기존에 제안된 규칙 추출을 위한 릴레이션 임베딩 학습방법을 통해 규칙을 유도하고, 추론엔진을 통해 지식 추론을 수행할 수 있는 통합 시스템을 제안한다. 제안하는 시스템은 현존하는 지식그래프의 누락된 정보에 대한 추론을 통해 지식 그래프에 추가함으로써 완성된 지식 그래프를 얻는 것을 목표로 한다.

3. 연구 내용

본 논문에서 제안하는 통합 시스템의 구조도는 그림 1과 같다. 제안하는 시스템은 먼저 뉴로 심볼릭 기반 규칙 생성 모델을 통해 규칙을 유도하고 추론엔진을 통해 유도된 규칙으로부터 지식 추론을 수행함으로써 누락된 지식 추론을 통한 완성된 지식 그래프를 얻는다.

3.1 규칙 유도를 위한 Proof Path 생성 방법

기존 연구인 NTP는 Neural unification을 사용하여 Proof tree를 구성하게 된다. 이러한 방식은 지식 그래프의 크기가 커질수록 계산 그래프의 크기가 기하급수적으로 커지게 된다. 때문에 이는 방대한 양의 지식 그래프를 처리하는데 계산적 한계가 있다. Symbolic unification은 벡터간 유사도 비교 대신 심볼의 동일성 검사를 수행함으로써 계산 그래프가 효과적으로 감소할 수 있고, 규칙 생성이 가능한 경로만을 선택하도록 Proof tree를 구성할 수 있다. 그림 2는 Symbolic unification을 통한 Proof tree와 Proof path생성 과정을 보여준다. Proof path를 생성하는 과정은 NTP에서의 계산 그래프 생성 방법인 Backward chaining과 유사한 방법으로 진행된다. Proof path는 주어진 목적 트리플과 해당 목적 트리플을 사실로서 증명할 수 있는 지식 그래프에 존재하는 트리플 간의 집합을 리스트로 표현한 것이다. 예를 들어 지식 그래프 K의 엔티티 집합 E, 릴레이션 집합 R로부터 $h_i, t_i \in E$, $r_i \in R$인 조건을 만족하는 목적 트리플 $r_i(h_i, t_i) \in K$대해 Proof path는 $[r_i(h_i, t_i), r_j(h_j, t_j), r_k(h_k, t_k)]$의 형태이다. 이때 $r_i(h_i, t_i)$는 목적 트리플에 해당하고, $r_j(h_j, t_j)$와 $r_k(h_k, t_k)$는 목적트리플을 사실로 증명 가능한 근거에 해당하는 트리플이다. 그림 2와 같이 목적 트리플이 [nationality, Tom, France]라고 할 때, Symbolic unification에서는 먼저 목적 트리플과 규칙의 Head인 #1(X,Y)와 Unification 되어 동일성검사를 수행하고 변수의 경우 Substitution 생성하게 된다. 목

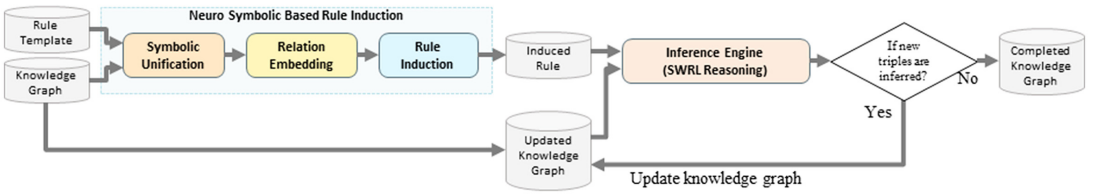

그림 1 시스템 구조도

Fig. 1 System architecture

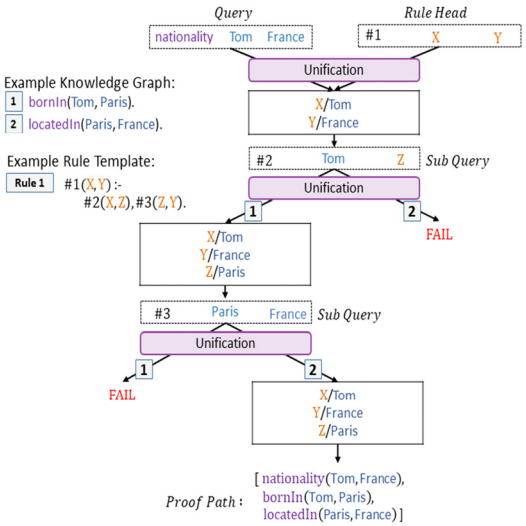

그림 2 Proof path 생성 예시
Fig. 2 Proof path creation example

적 트리플의 릴레이션인 nationality와 규칙 템플릿의 릴레이션인 #1과 동일성검사를 하게 되면 규칙 템플릿의 릴레이션인 #1은 특정 릴레이션으로 명시되지 않은 Parameterized rule의 릴레이션이므로 Unification이 성공하게 된다. 이후 변수에 대한 Substitution생성되어 규칙 변수인 X,Y에 각각 Tom과 France을 바인딩하여, {X/Tom, Y/France}라는 Substitution을 얻는다. 다음으로 규칙 Body #2(X,Z)에 Substitution을 적용하여 규칙의 Body에 해당하는 X는 Tom으로 대체되어 [#2, Tom, Z]와 같은 Sub-Query를 생성할 수 있다. 이후 Sub-Query와 지식 그래프의 모든 트리플과 Unification을 수행하는 기존 NTP와 달리, 생성된 Sub-Query와 Unification될 트리플을 지식 그래프로부터 탐색하게 된다. Sub-Query와 Unification될 트리플의 탐색은 Substitution 값을 참조하여 진행되는데, Substitution에 의해 대체된 Tom은 Sub-Query의 Subject에 해당하는 엔티티 이므로 지식 그래프로부터 Subject가 Tom인 1번 트리플이 Unification된다. 이러한 방식으로 주어진 규칙 템플릿의 모든 Body에 대해 Substitution을 구하고, Unification이 가능한 트리플을 찾게 됨으로써 Proof path를 구할 수 있다.

3.2 Relation Embedding 학습 방법

Proof path는 Symbolic unification과정을 통해 생성될 수 있다. 규칙을 유도하기 위해선 규칙 템플릿의 릴레이션 벡터와 Symbolic unification을 통해 생성된 Proof path의 릴레이션 벡터가 유사해지도록 학습되어야 한다. 이러한 심볼의 벡터 표현간 유사도 학습은 딥러닝 기반의 Neural link prediction 방법을 사용하는 지식 완성 모델에서 활용되는 지식그래프 임베딩 방법과 동일하다. 대표적인 Neural link prediction방법인 TransE나 TransR의 경우 목적함수를 정의하고 지식 그래프에 존재하는 엔티티와 릴레이션에 대한 벡터표현을 학습하고, ConvE는 CNN과 다층 퍼셉트론으로 이루어진 신경망과 지식 그래프의 엔티티와 릴레이션에 대한 벡터표현을 동시에 학습한다. 본 논문에서 제안하는 방법은 목표 트리플로부터 구해진 Proof path의 릴레이션(e.g. nationality, locatedIn, bornIn)과 규칙 템플릿의 릴레이션(e.g. #1, #2, #3)간 유사도 평균이 1이 되도록 손실 함수를 정의하고 릴레이션 벡터를 학습한다. 목표 트리플 $r_i(h_i, t_i)$로부터 도출 가능한 Proof path수가 n개일 때, 목표 트리플로부터 구해진 하나의 Proof path $[r_i(h_i, t_i), r_j(h_j, t_j), r_k(h_k, t_k)]$를 구성하는 각 트리플의 릴레이션 집합인 π에 대하여 π=(r_i, r_j, r_k)의 형태가 된다. 추가로 하나의 목적 트리플 $r_i(h_i, t_i)$로부터 도출 가능한 모든 Proof path가 P일 때 $\pi \in P$를 만족하고, $P = \bigcup_{l=1}^{n} \pi_l$과 같다. 벡터 학습을 위해 Proof path의 릴레이션 집합인 (r_i, r_j, r_k)과 규칙 탬플릿 [#1(X,Y) :- #2(X,Z), #3(Z,Y)]의 릴레이션 집합(#1, #2, #3)간의 그룹핑을 수행한다. 그룹핑된 릴레이션 집합 Π에 대해 Π =$[(\#1, r_i), (\#2, r_j), (\#3, r_k)]$형태가 되며, 그룹핑된 릴레이션은 서로 Unify된 릴레이션이 된다. 이후 그룹핑된 릴레이션의 벡터 표현간 유클리디안 유사도 평균값을 계산한다. 예를 들어 유사도를 구하고자 하는 두 릴레이션 #1, r_i에 대하여 유클리디안 유사도는 다음과 같다.

$$L2sim(\#1, r_i) = \exp(-\| \overrightarrow{\#1} - \overrightarrow{r_i} \|_2)$$

유사도 평균은 Symbolic unification을 통해 구해진 Proof path의 신뢰값에 해당하며 유사도 평균값이 1이 되도록 릴레이션 임베딩 벡터를 학습한다. Symbolic unification을 통해 구해진 Proof path는 모두 참이며 규칙 템플릿의 릴레이션(e.g. #1, #2, #3)은 해당 규칙 템플릿을 따라 Unification결과로 도출된 n개의 Proof path P에 적용하기 위해 augment하고, 하나의 augment a에 대한 Proof path의 신뢰값은 다음과 같다.

$$\begin{aligned} avgL2sim(\Pi_l^a) = avg(&L2sim(\#1^a, r_i \vee r_i'), \\ &L2sim(\#2^a, r_j \vee r_j'), \\ &L2sim(\#3^a, r_k \vee r_k')) \end{aligned}$$

이후 augment된 규칙 템플릿중 가장 유사도가 큰 하나의 Π를 선택하기 위해 augment된 m개의 Proof path중 Max Pooling을 수행한다. 이후 하나의 목적 트리플(r_i, h_i, t_i)로부터 도출된 n개의 참인 Proof path중

릴레이션간 유사도 업데이트가 필요한 하나의 경로를 선택하기 위해 Min pooling을 수행하며 다음과 같다.

$$u\_ntp(P) = \min_n(\max_m(\bigcup_{l=1,a=0}^{n,m} avgL2sim(\Pi_l^a)))$$

위와 같은 과정으로 도출된 하나의 Proof path에 대한 유사도는 1이 되도록 하고, unification으로부터 도출된 Proof path의 릴레이션 집합 $(r_i, r_j, r_k) \in P$로부터 릴레이션을 무작위 샘플링한 $(r_i', r_j', r_k') \notin P$에 대해선 릴레이션간 유사도가 0이 되도록 하기 위해 Negative log-likelihood 손실 함수를 사용한다.

$$Loss_{u\_ntp} = \sum_{(l_i \in L, y)} -y\log(u\_ntp(P)) - (1-y)\log(1-u\_ntp(P'))$$

위와 같은 방법으로 학습된 규칙의 각 릴레이션은 Unify된 지식그래프의 릴레이션 벡터와 유사도가 높고, Unify 되지 않은 릴레이션 벡터는 유사도가 낮아지도록 학습된다. 학습이 완료되면 규칙의 각 릴레이션은 가장 유사한 트리플의 릴레이션으로 디코딩됨으로써 규칙을 유도할 수 있다. 이와 같은 방법으로 유도된 규칙은 주어진 트리플정보를 기반으로 새로운 지식 추론을 가능하게 한다.

3.3 유도 규칙을 활용한 지식 추론 방법

본 논문에서 제안하는 시스템은 릴레이션 임베딩 학습과 지식 추론을 위한 추론엔진간의 통합을 위해 Parameterized rule 형태의 규칙 템플릿을 분석하여 지식 추론에 필요한 참조 테이블을 정의한다. 정의된 참조 테이블에서 규칙 탬플릿의 릴레이션에 해당하는 값은 릴레이션 임베딩 벡터 학습 이후 지식 그래프의 릴레이션중 가장 유사한 릴레이션으로 디코딩된다. 이와 같이 규칙 분석을 통해 생성된 테이블은 추론엔진의 추론과정에서 참조되어 규칙을 만족하는 새로운 트리플을 추론하게 된다. 참조 테이블은 키 테이블, 공통변수 테이블, 결론 테이블 3개를 사용한다. 예를 들어 다음과 같은 두 개의 Parameterized rule을 사용할 경우 정의되는 참조 테이블은 각각 표 1, 2, 3과 같다. Parameterized rule의 각 릴레이션은 릴레이션 번호와 규칙 번호, augment 번호로 표현된다. 예를 들어 #2_1_0은 1번 규칙에서 릴레이션 번호가 2이고, augment번호가 0인 릴레이션을 의미한다.

Rule 1: #1_1_0(X,Y) :- #2_1_0(X,Z), 3_1_0(Z,W), 4_1_0(W,Y)
Rule 2: #1_2_0(X,Y) :- #2_2_0(X,Z), #3_2_0(Z,Y)

키 테이블은 릴레이션별로 사용되는 규칙 번호(Rule#)와 인접한 두 조건간의 공통변수(ComVar) 및 해당 규칙에서 조건의 위치정보(CondIdx)에 대해 저장한다. 결론 테이블은 새로운 트리플을 생성하기 위해 참조되는 테이블로 결론부의 Subject와 Object에 대한 위치 정보를 저장한다. 공통변수 테이블은 규칙의 조건의 개수가 3개 이상인 경우 참조되는 테이블로써 첫 번째와 두 번째 조건에 대한 조인연산이 완료된 후 마지막 세 번째 조건에 대한 조인 수행시 먼저 조인된 트리플 데이터의 키값을 다음 매칭에서 사용될 공통변수로 치환해야 하고, 해당 작업을 위해 참조된다. Parameterized rule은 3.2절에서 설명한 릴레이션 임베딩 학습 방법을 수행하고, 지식 그래프에 존재하는 트리플의 릴레이션간 유사도를 구하여 가장 유사한 릴레이션으로 변환될 수 있다. 예를 들어 Parameterized rule을 트리플의 릴레이션으로 변환하는 디코딩 과정을 거쳐 유도된 규칙이 다음과 같을 때 참조 테이블 또한 추론 엔진에서 사용가능한 형태로 변환될 수 있고 각각 표 4, 5, 6과 같다.

Rule 1: nationality(X,Y) :-
bornIn(X,Z), locatedIn(Z,W), nation(W,Y)
Rule 2: locatedIn(X,Y) :- belongsTo(X,Z), within(Z,Y)

릴레이션 임베딩 학습 후 위와 같이 참조 테이블을 생성하게 되면 추론엔진을 통한 지식 추론을 수행할 준비가 완료된다. 이후 진행 되는 추론과정은 규칙의 조건 해당하는 트리플에 키 테이블을 참조하여 추론과정에 필요한 정보를 맵핑하는 과정과, 규칙 번호와 공통변수를 키로 하여 테이블에 대한 조인연산을 통해 수행된다. 맵핑 과정은 키 테이블을 참조하여 모든 트리플에 대해 Predicate을 기준으로 맵핑 테이블을 정의한다. 예를 들어 아래와 같은 5개의 트리플에 대한 키 테이블 정보가 맵핑된 테이블은 표 7과 같다.

Triples: bornIn(john, chicago), locatedIn(chicago, illinois)
nation(illinois, USA), belongsTo(L.A., california)
within(california, USA)

표 1 키 테이블 예시
Table 1 Key table example

Relation	Rule#	ComVar	CondIdx
#2_1_0	1	object	1
#3_1_0	1	subject	2
#4_1_0	1	subject	3
#2_2_0	2	object	1
#3_2_0	2	subject	2

표 2 결론 테이블 예시
Table 2 Conclusion table example

Rule#	Subject	Relation	Object
1	(#2_1_0, subj)	#1_1_0	(#4_1_0, obj)
2	(#2_2_0, subj)	#1_2_0	(#3_2_0, obj)

표 3 공통변수 테이블 예시
Table 3 Common variable table example

Relation	Rule#	new_ComVar	phase
#3_1_0	1	obj	2

표 4 디코딩된 키 테이블 예시
Table 4 Decoded key table example

Relation	Rule#	ComVar	CondIdx
bornIn	1	object	1
locatedIn	1	subject	2
nation	1	subject	3
belongsTo	2	object	1
within	2	subject	2

표 5 디코딩된 결론 테이블 예시
Table 5 Decoded conclusion table example

Rule#	Subject	Relation	Object
1	(bornIn, subj)	nationality	(nation, obj)
2	(belongsTo, subj)	locatedIn	(within, obj)

표 6 디코딩된 공통변수 테이블 예시
Table 6 Decoded common variable table example

Relation	Rule#	new_ComVar	phase
locatedIn	1	obj	2

표 7 맵핑 테이블 예시
Table 7 Mapping table example

subject	relation	object	Rule#	comVar	condIdx
john	bornIn	chicago	1	chicago	1
chicago	locatedIn	illinois	1	chicago	2
illinois	nation	USA	1	illinois	3
L.A.	belongsTo	california	2	california	1
california	within	USA	2	california	2

표 7과 같이 맵핑 테이블은 트리플의 각 릴레이션에 따라 키 테이블의 규칙 번호, 해당 규칙에서의 조건의 위치정보가 모든 트리플에 맵핑되고, 공통변수의 위치정보를 참조하여 트리플의 실제 엔티티가 맵핑된 형식을 갖는다. 맵핑 테이블 생성 이후 규칙 조건의 위치정보(condIdx)에 따라 맵핑 테이블로부터 조건에 대한 트리플을 추출하여 조인 연산을 수행함으로써 트리플을 추론하게 되는데 n번의 조인연산을 통해 조건의 개수가 n+1개인 규칙에 대한 추론이 가능하다. 트리플 추론을 위한 조인 과정은 먼저 조건의 위치정보(condIdx)가 1인 트리플과 2인 트리플을 맵핑 테이블로부터 추출하여 각각 테이블로 구성하고 규칙 번호와 공통변수를 키로 하여 조인함으로써 표 8과 같은 조인 테이블을 생성하고, 결론 테이블을 참조하여 2개의 조건으로 이루어진 규칙 2에 대해서 locatedIn(L.A., USA)를 결론으로 추론한다.

첫 번째 조인 연산 이후 두 번째 조인 연산을 통해 조건의 개수가 3개 이상인 규칙에 대한 추론을 위해 조인 키값인 공통변수를 공통변수 테이블을 참조하여 변

표 8 첫 번째 조인 테이블 예시
Table 8 First join table example

Grouped Triples	Rule#	comVar
[john, bornIn, chicago, chicago, locatedIn, illinois]	1	chicago
[L.A., belongsTo, california, California, within, USA]	2	california

표 9 두 번째 join 테이블 예시
Table 9 Second join table example

Grouped Triples	Rule#	comVar
[john, bornIn, chicago, chicago, locatedIn, illinois illinois, nation, USA]	1	illinois

경한다. 공통 변수테이블의 phase 2는 조인연산 반복횟수를 의미하며 두 번째 조인 연산에서 규칙 1에 대한 조인 키값은 그룹핑된 트리플 중 locatedIn의 Object라는 정보를 알 수 있으므로 조인 테이블의 규칙 1의 공통변수(comVar)는 chicago에서 illinois로 변경된다. 이후 맵핑 테이블로부터 조건의 위치정보(condIdx)가 3인 트리플을 추출하고 테이블로 구성하여 첫 번째 조인 테이블과 규칙 번호와 공통변수를 키로 하여 조인 연산을 수행함으로써 표 9와 같은 조인 테이블을 얻을 수 있고 해당 테이블은 결론 테이블을 참조함으로써 규칙 1에 대해서 nationality(john, USA)를 결론으로 추론할 수 있다.

4. 실 험

4.1 실험 데이터

제안된 규칙 생성 모델의 규칙 유도에 대한 성능 검증을 위해 기존 NTP모델에서 사용한 Benchmark데이터인 Nations[11], UMLS[12], Kinship[13]을 사용하였고, 지식 추론을 통한 대규모 지식 그래프의 지식 증강도에 대한 검증을 위해 NTP에서 사용한 데이터셋에 비해 규모가 큰 Kdata와 WiseKB를 사용하였다. Nations는 전세계 국가간의 관계에 대한 지식그래프로 14개의 관계, 55개의 엔티티 및 1793개의 트리플로 구성되어있다. UMLS는 유기체, 해부학적 구조, 생물학적 기능 등과 같은 생의학적 관계에 대한 지식 그래프로 49개의 관계, 135개의 엔티티 및 5827개의 트리플로 구성된다. Kinship은 친족관계에 관한 지식 그래프로 26개의 관계, 104개의 엔티티 및 9618개의 트리플을 포함하고 있다. 추론엔진을 사용한 지식 추론 실험에서 사용된 Kdata와 WiseKB는 한글로 구성된 대용량 지식 그래프로 출생지, 장르, 직업, 국적 등 의 트리플이 주를 이룬다. 실험에 사용한 지식그래프의 통계를 표 10에서 나타낸다.

표 10 실험 데이터 요약
Table 10 Summary of the experiment data

Dataset	# Train	# Test	# Entities	# Relations
Nations	1,592	201	55	14
UMLS	5,216	661	135	49
Kinship	8,544	1,074	104	26
Kdata	2,850k		1,871k	15,278
WiseKB	15,581k		4,077k	389

4.2 실험 방법 및 결과

Kdata는 500만개가 넘는 트리플을 포함하고 있고, WiseKB는 2억개가 넘는 트리플을 포함하는 대규모 지식 그래프이다. 실험을 위해 두 데이터는 먼저 정제작업을 통해 규칙을 통해 추론 될 수 없거나 조건으로 사용될 수 없는 트리플을 제외한다. 정제 작업을 통해 얻은 Kdata의 트리플 수는 280만개, WiseKB의 트리플수는 1,500만개를 실험에 사용하게 된다. WiseKB와 Kdata는 학습 데이터와 테스트 데이터가 분리되어 있는 데이터가 아니기 때문에 규칙 유도를 위한 학습 데이터와 규칙에 대한 검증을 수행할 테스트 데이터를 분리하는 전처리 작업이 필요하다. 이를 위해 지식 그래프에서 각 릴레이션별 트리플 개수를 카운트 하여 내림차순 정렬 후 트리플수가 가장 많은 릴레이션부터 상위 8개의 목표 릴레이션을 선정하였다. 이렇게 선정된 8개의 릴레이션은 규칙을 통해 추론될 테스트 트리플의 릴레이션이 된다. 목표 릴레이션 선정후 PRA(Path Ranking Algorithm)[14]의 Random walk Algorithm을 사용하여 지식 그래프로부터 목표 릴레이션에 대한 릴레이션 경로를 추출한다. 예를 들어 목표 릴레이션 nationality로부터 추출된 릴레이션 경로가 bornIn → nation이라고 하면 목표 릴레이션은 규칙의 결론에 해당하고 릴레이션 경로가 규칙의 전제에 해당하는 형태의 규칙 nationality(X,Y) :- bornIn(X,Z), nation(Z,Y)로 변환될 수 있다. 이렇게 생성된 규칙셋을 바탕으로 규칙셋을 실제 트리플로 인스턴스화할 수 있는 트리플을 지식 그래프로부터 추출하여 추출된 트리플중 결론부에 해당하는 트리플들의 10%를 테스트 데이터로 사용하고, 나머지 트리플을 학습 데이터로 사용하였다. 실험을 위해 Kdata로부터 추출된 학습 데이터는 총 100,578개 트리플이 사용되었으며, 3,377개의 트리플을 테스트 데이터로 사용하였다. WiseKB는 총 627,118개의 트리플을 학습 데이터로 사용하고 테스트 데이터로 20,355개의 트리플을 사용하였다.

실험 방법으로는 먼저 기존의 NTP와 본 논문에서 사용한 규칙생성 모델의 규칙 유도 성능 검증을 위하여 비교적 크기가 작은 Nations, UMLS, Kinship 3개의 지식 그래프를 사용하여 선행 실험을 진행 하였다. 선행 실험은 모델 학습을 수행하였을 때 유도된 규칙의 개수와 규칙과 학습 데이터를 활용하여 추론 가능한 테스트 데이터의 트리플 수를 비교하였다. 유도된 규칙과 트리플 추론 결과를 표 11에 기술하였으며, 기존 뉴로 심볼릭 기반 연구인 NTP에 비해 테스트 트리플을 추론 가능한 규칙이 더 많이 유도되었음을 확인하였다. 또한 유도된 규칙은 NTP에 비해 더 많은 트리플을 추론할 수 있음을 확인하였다.

Kdata와 WiseKB에 대해서는 규칙 생성 모델을 통해 학습 데이터 으로부터 유도된 규칙과 학습 데이터를 활용하여 추론 엔진을 사용한 지식 추론을 수행하였다. 추론 엔진을 통해 지식 추론을 수행하게 되면 유도된 규칙의 조건을 만족하는 모든 트리플을 지식 그래프로부터 가져오게 되고, 만족하는 규칙의 조건에 대해서 결론 도출이 가능한 모든 트리플을 추론하게 된다. 표 12와 표 13에 각각 Kdata, WiseKB의 목표 릴레이션별 테스트 데이터의 트리플수와 추론엔진으로부터 추론된 테스트 트리플 개수에 대한 실험 결과를 기술하였다. 그림 3은 Kdata와 WiseKB에서 주어진 학습 데이터와 모델로부터 유도된 규칙을 통해 추론된 전체 트리플 개수를 나타낸다. WiseKB는 Kdata에 비해 트리플 수가 많기 때문에 규칙 인스턴스의 개수도 늘어나므로, Kdata보다 더 많은 인스턴스를 추론 할 수 있었다. 추론 결과 그림 3과 같이 Kdata는 100,578개의 트리플로부터 30,751개의 트리플을 추론하여 실험에 사용된 학습 데이터기준 30%의 지식 증강도를 보였고, WiseKB의 경우 627,118개의 트리플로부터 596,350개의 트리플을 추론하여 95%의 지식 증강도를 보였다.

그림 4는 Kdata와 WiseKB에 대한 유도된 규칙에 대한 검증 결과로 아래와 같은 두가지 규칙 탬플릿에 대한 MAP(Mean Average Precision)과 AP(Average

표 11 모델 별 추론된 규칙과 추론된 트리플 결과
Table 11 Model inference results comparison

	Inferred Rules			Inferred Triples	
Dataset	NTP	Our Model	# Test	NTP	Our Model
Nations	21	**94**	201	52	**186**
UMLS	11	**30**	661	291	**377**
Kinship	12	**89**	1,074	490	**844**

표 12 Kdata 추론 결과
Table 12 Kdata inference results

	Kdata	
Relation	#Test Triples	#Inferred Triples
출생지	1,302	1,298
장르	838	529
국적	484	427
국가	274	211
직업	214	153
제작	128	87
출연	79	73
감독	58	51

표 13 WiseKB 추론 결과
Table 13 WiseKB inference results

	WiseKB	
Relation	#Test Triples	#Inferred Triples
belongsTo	4,110	3,684
locatedIn	7,325	7,181
nationality	2,953	2,762
bornIn	2,791	2,701
nation	1,783	1,775
genre	876	814
director	301	182
work	216	176

Precision)산출 결과이다.

Rule 1: #1(X,Y) :- #2(X,Z), #3(Z,Y)

Rule 2: #1(X,Y) :- #2(X,Z), #3(Z,W), #4(W,Y)

실험에 사용된 규칙 템플릿은 각각 20의 Augmentation을 수행하였고, 각 규칙 탬플릿 마다 20개의 규칙이 유도 될 수 있다. Kdata와 WiseKB의 경우 PRA를 사용하여 생성된 릴레이션 경로로부터 정답 규칙셋을 생성할 수 있기 때문에 유도된 규칙 중 유의미한 규칙이 얼마나 많이 유도되었는지에 대하여 정답 규칙셋과의 비교를 통해 수치화 할 수 있다. 유도된 규칙은 augment된 각 규칙 템플릿의 릴레이션(e.g. #1_0_0, #2_0_0, #3_0_0)이 가장 높은 유사도를 갖는 지식그래프의 릴레이션으로 디코딩 되어 생성되고 각 릴레이션간 유클리디안 유사도의 최소값을 유도된 규칙에 대한 신뢰값으로 사용하였다. 이렇게 구해진 신뢰값을 기준으로 하나의 규칙 템플릿에서 구해진 규칙들을 내림차순 정렬 할 수 있다. 이후 두개의 규칙 템플릿에 대해 각각 AP를 구하고 이에 대한 평균인 MAP을 산출 하였다. 사용된 두개의 규칙 템플릿 각각으로부터 유도된 40개의규칙중 정답 규칙셋에 있는 유의미한 규칙들이 높은 신뢰값을 갖을 수록 AP와 MAP은 높아지게 된다. Kdata의 장르, 출생지 릴레이션에 대한 유도 규칙들은 PRA 결과로부터 생성된 정답 규칙셋에 없는 규칙들이 높은 신뢰값을 갖기때문에 MAP과 AP수치가 낮은 결과를 보인다. 하지만 높은 신뢰값을 갖는 규칙중 PRA에서는 도출 하지 못한 테스트 트리플을 추론할 수 있는 유의미한 규칙이 다수 존재함을 확인하였다.

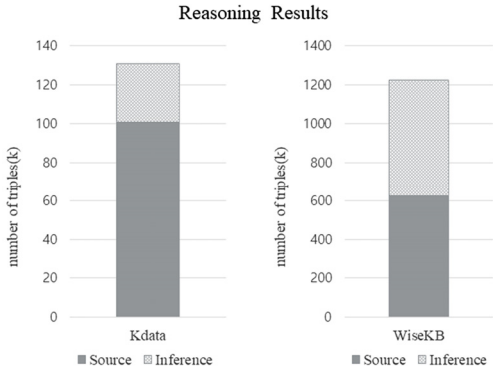

그림 3 추론된 트리플의 개수
Fig. 3 Number of inferred triples

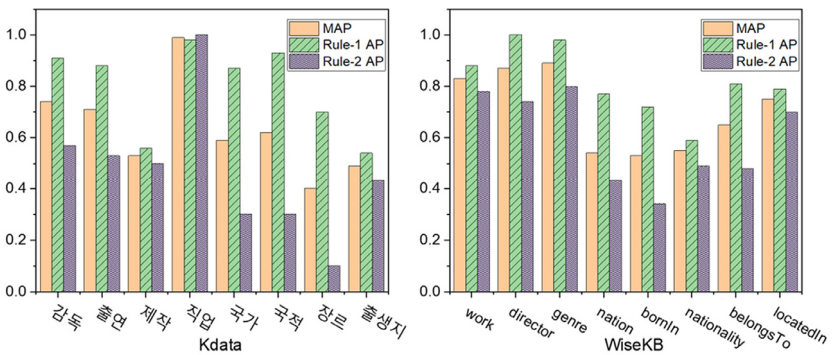

그림 4 Kdata 및 WiseKB 유도 규칙에 대한 MAP 결과
Fig. 4 MAP results for Kdata and WiseKB induced rules

5. 결 론

본 논문에서는 뉴로 심볼릭 기반 모델을 활용하여 지식 그래프의 릴레이션 임베딩 벡터를 학습하고, 유도된 규칙을 사용하여 새로운 지식을 추론할 수 있는 딥러닝 학습 방식과 로직 추론 방식의 통합 시스템을 제안한고, 지식 추론 실험을 통해 제안한 시스템에 대한 성능을 검증하였다. 본 논문에서는 기존 NTP의 계산 복잡도 문제를 개선한 Symbolic unification방법과 릴레이션 임베딩 학습 방법을 사용하여 작은 크기의 지식 그래프뿐만 아니라 대용량 지식 그래프에 대한 벡터 학습과 규칙유도를 가능하게 하고, 기존 NTP의 확장성 문제를 해결함과 동시에 더 많은 규칙을 유도해 낼 수 있었다. 또한 제안하는 모델을 통해 유도된 규칙은 새로운 지식 추론에 유의미하게 사용될 수 있고, 누락된 지식을 효과적으로 완성 할 수 있음을 검증하였다.

References

[1] Bollacker. K, Evans. C, Paritosh. P, Sturge. T, and Taylor, J., "Freebase: A collaboratively created graph database for structuring human knowledge," *Proc. of the 2008 ACM SIGMOD international Conference on Management of data*, pp. 1247-1250, 2008.

[2] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, ... and Bizer, C., "DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web*, Vol. 6, No. 2, pp. 167-195, 2015.

[3] Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum, "Yago: a core of semantic knowledge," *Proc. of the 16th international conference on World Wide Web. ACM*, 2007.

[4] Horrocks, Ian, et al., "SWRL: A semantic web rule language combining OWL and RuleML," *W3C Member submission 21 (2004)*: 79.

[5] T. Rocktäschel and S. Riedel, "End-to-End Differentiable Proving," *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 3791-3803, Dec. 2017.

[6] Antoine. Bordes, Nicolas. Usunier, Alberto. Garcia-Duran, Jason. Weston, Oksana. Yakhnenko, "Translating Embeddings for Modeling Multi-relational Data," *Advances in neural information processing systems*, pp. 2787-2795, 2013.

[7] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, Learning entity and relation embeddings for knowledge graph completion, *Proc. of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2181-2187. Austin, TX. 2015.

[8] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel, Convolutional 2D Knowledge Graph Embeddings, *Proc of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 1811-1818, 2018.

[9] H. Gallaire. Logic and Data Bases, *Symposium on Logic and Data Bases, Centre d'études et de recherches de Toulouse*, Plemum Press, New York, 1978.

[10] R. Socher, D. Chen, C. D. Manning, and A. Y. Ng, "Reasoning with Neural Tensor Networks for Knowledge Base Completion," *Proc. of the 26th International Conference on Neural Information Processing Systems*, pp. 926-934, 2013.

[11] Rummel, R. J., Dimensionality of Nations Project: Attributes of Nations and Behavior of Nation Dyads, 1950-1965, *Inter-university Consortium for Political and Social Research (ICPSR)*, 1984.

[12] McCray, Alexa, "An Upper-Level Ontology for the Biomedical Domain," *Comparative and functional genomics*, Vol. 4, pp. 80-4, 2003.

[13] Denham, W., The Detection of Patterns in Alyawara Nonverbal Behaviour, Doctoral dissertation, Department of Anthropology, University of Washington, 1973.

[14] M. Gardner, T. Mitchell, "Efficient and Expressive Knowledge Base Completion Using Subgraph Feature Extraction," *Proc. of the 2015 Conference on Empirical Methods in Natural Language*.

신 원 철
정보과학회논문지
제 48 권 제 4 호 참조

박 현 규
정보과학회논문지
제 48 권 제 4 호 참조

박 영 택
정보과학회논문지
제 48 권 제 4 호 참조