기계학습 기반 K-YAGO 구축

정석원○, 최맹식, 김학수
강원대학교, IT대학, 컴퓨터정보통신공학전공
{nlpsw, nlpmschoi, nlpdrkim}@kangwon.ac.kr

Construction of K-YAGO Based on Machine Learning

Seokwon Jeong○, Maengsik Choi, Harksoo Kim
Program of Computer and Communication Engineering,
College of Information Technology, Kangwon National University

요 약

자연어 처리를 이용한 다양한 응용 시스템에서 지식베이스는 중요한 요소이다. 지식베이스의 대표적인 예로 YAGO와 디비피디아 등이 있다. YAGO는 고성능의 지식베이스지만 한국어를 지원하지 않는다는 문제점이 있다. 그리고 디비피디아는 한국어를 지원하지만 트리플의 속성이 세분화되어 있어서 사용이 어렵다. 본 논문에서는 YAGO와 디비피디아의 트리플을 매칭하여 디비피디아 트리플의 속성을 YAGO에서 사용하는 관계명으로 변환하고 MEM을 이용해 매칭되지 않은 트리플의 속성을 자동으로 분류하는 시스템을 제안한다. 제안한 방식으로 실험한 결과 F1-Measure 79.04%의 성능을 보였다.

주제어: 기계학습, 언어간 매칭, K-YAGO

1. 서론

자연어 처리를 이용하는 다양한 응용 시스템에서 지식베이스는 중요한 요소이다. 이러한 지식베이스의 대표적인 예로 YAGO[1], 디비피디아(DBPedia)[2] 등이 있다. YAGO는 IBM 왓슨(Watson)의 질의응답 시스템 등에 이용된 고성능의 지식베이스이다[3]. 그러나 YAGO는 한국어를 지원하지 않는다는 문제점이 있다. 그리고 디비피디아는 한국어를 지원하지만 추출된 지식들의 속성이 세분화되어 있어서 직접적인 활용이 어렵다.

본 논문에서는 디비피디아 인포박스 트리플의 속성(Property)들을 YAGO 관계명으로 변환하고, 변환되지 않은 디비피디아 인포박스 트리플의 속성을 MEM[4]을 이용하여 YAGO 형태로 자동 분류하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 기술하고 3장에서는 본 논문에서 제안하는 시스템을 설명한다. 4장에서는 실험방법 및 실험의 성능을 살펴보고 5장에서 결론을 맺는다.

2. 관련 연구

지식베이스를 구축하는 다양한 프로젝트들이 진행되어 왔다. YAGO는 위키백과, 워드넷 등을 통해 추출된 지식베이스로 1,000만 개 이상의 개체명과 그 개체명에 대한 1억 2천만 개 이상의 정보(facts)를 포함하고 있다[1]. 디비피디아는 위키백과로부터 구조화된 정보를 추출하고, 이 정보를 웹에 이용 가능하도록 만드는 커뮤니티로서 온톨로지를 개발, 유지하고 있다. 디비피디아 온톨로지는 위키백과 인포박스로부터 규칙 기반 방식을 통해 반자동으로 구축된다[5].

지식베이스를 확장하기 위한 연구도 진행되었다. [6]은 사전 기반의 번역 및 패턴을 이용하여 영어 지식베이스의 트리플을 한국어로 변환함으로써 한국어 지식베이스를 확장하였다.

3. 제안 시스템

YAGO에서는 다양한 형태로 지식베이스를 제공하고 있는데, 본 논문에서는 yagoFacts만을 사용한다. yagoFacts는 트리플의 형태로 정보를 표현한다. 트리플은 정보를 주어-서술어-목적어의 형태로 표현한 것이다.

본 논문에서 제안하는 시스템은 yagoFacts의 개체명을 한글로 변환한 한국어 yagoFacts와 디비피디아 인포박스 트리플의 양 개체명을 비교하여 모두 같은 경우 매칭 트리플을 생성한다. 생성된 매칭 트리플을 MEM을 이용해 학습하여 관계 분류 모델을 생성하고 매칭되지 않은 디비피디아 인포박스 트리플(이하 비매칭 트리플이라 함)을 관계 분류 모델에 적용하여 관계명을 분류한다. 전체 시스템 구조도는 [그림 1]과 같다.

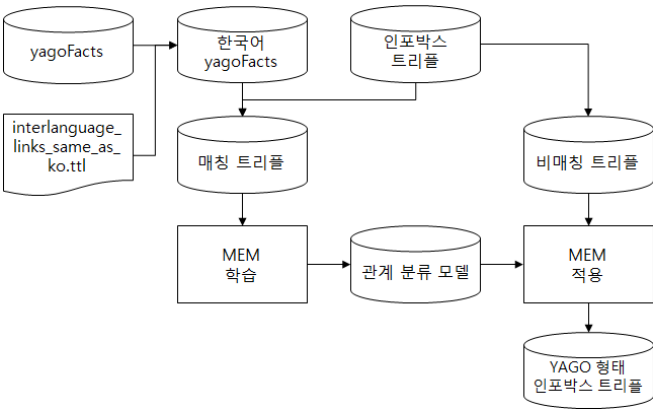

그림 1. 시스템 구조도

3.1 데이터 생성

디비피디아의 interlanguage_links_same_as_ko.ttl에는 한국어 개체명과 그 개체명에 해당하는 영어 개체명

이 저장되어 있다. 저장된 정보를 이용하여 yagoFacts 트리플의 개체명들을 한글로 변환한다. 변환은 [그림 2]와 같이 진행된다.

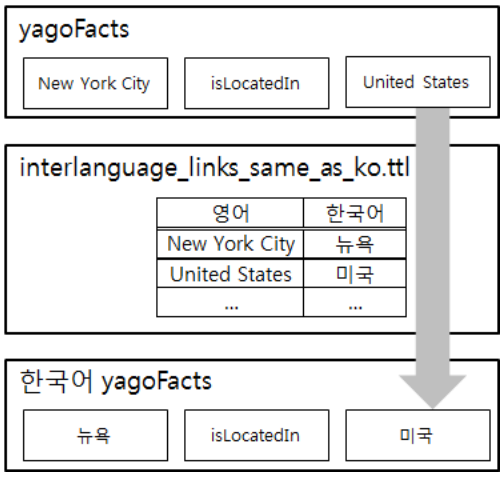

그림 2. 개체명 변환 예

[그림 3]과 같이 한국어 yagoFacts와 디비피디아 인포박스 트리플의 양쪽 개체명을 비교하여 양쪽 개체명이 모두 일치하는 경우 yagoFacts의 관계명으로 분류된 매칭 트리플을 생성한다. 개체명 매칭의 결과로 "뉴욕 - 나라 - 미국" 트리플은 isLocatedIn 클래스의 트리플로 분류된다. 양쪽 개체명이 하나만 일치하거나 모두 일치하지 않는 경우엔 비매칭 트리플로 분류된다.

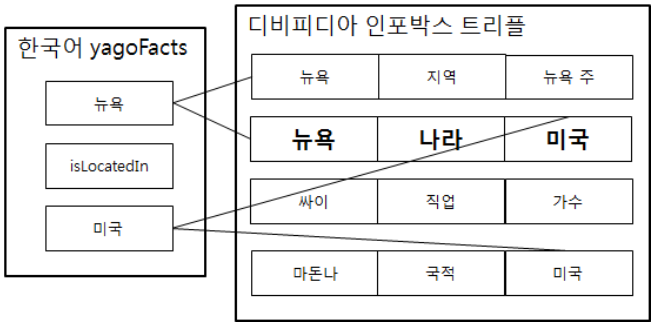

그림 3. 개체명 매칭 예

3.2 관계명 분류

비매칭 트리플을 분류하기 위해 생성된 매칭 트리플로부터 자질을 추출한 뒤 MEM을 이용하여 학습한다. 학습에는 [표 1]과 같은 자질을 사용하였고 MEM은 R언어 패키지인 maxent[7]를 이용하였다.

자질은 디비피디아 인포박스 트리플의 주어 개체명과 목적어 개체명을 그대로 사용하였다. 속성(Property)의 경우 비슷한 개체를 설명하는 여러 인포박스에서 특정 속성들이 같은 순서로 나오는 경향을 보였기 때문에 트리플의 속성을 하나만 사용하지 않고 연속된 3개의 속성을 자질로 사용하였다. 위키백과 카테고리는 해당 개체명으로 검색된 위키백과 문서가 어떤 범주인지를 나타내는 정보이다. 카테고리를 통해 개체명에 대한 많은 정보를 얻을 수 있을 것으로 보였다. 그러나 카테고리는 비슷한 의미를 가지는 개체명끼리도 '1924_태어남', '1932_태어남'과 같이 다양하게 표현되어 있기 때문에 카테고리를 그대로 사용하는 것으로는 관계명을 분류하기 어렵다. 그러므로 본 논문에서는 '태어남'과 같이 카테고리의 마지막 어절만 추출하여 자질로 사용하였다. 타입은 개체명들이 온톨로지 내에서 가지는 정보들을 나타낸다. 주어와 목적어의 온톨로지 타입에 따라 가질 수 있는 관계명이 다르므로 관계명을 분류하는데 많은 영향을 줄 것으로 예상되어 자질로 사용하였다.

표 1. 사용 자질 예

사용 자질	예
주어 개체명	하노 코플러
이전 속성	이름
속성	출생지
다음 속성	직업
목적어 개체명	베를린
주어 위키백과 카테고리	연주자, 배우, 태어남, 사람
목적어 위키백과 카테고리	주도, 수도, 주, 도시, 베를린
주어 타입	Actor, Artist, Person, Agent
목적어 타입	Settlement, PopulatedPlace, Place

4. 실험 및 평가

실험에서는 YAGO2S의 yagoFacts 트리플 데이터 4,431,523개와 디비피디아 3.9의 인포박스 트리플 데이터 1,686,742개를 사용하였다. yagoFacts 트리플 중 양쪽 개체명 모두 한글로 변환된 124,277개의 트리플만을 실험에 사용하였고, 양쪽 개체명 모두 변환된 yagoFacts 트리플과 디비피디아 인포박스를 매칭하여 28,409개의 매칭 트리플을 생성하였다.

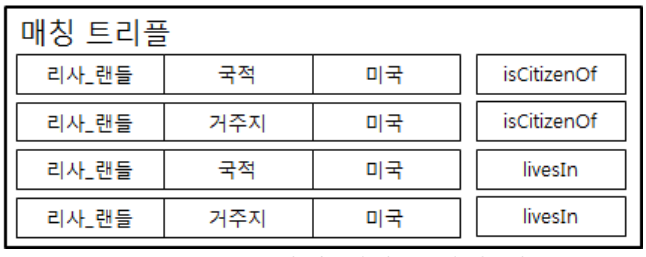

그림 4. 중복된 매칭 트리플 예

한국어 yagoFacts와 디비피디아 인포박스에서 양쪽 개체명이 같고 관계명, 속성만 다른 트리플들이 존재할 경우에 [그림 4]와 같은 매칭 트리플 결과가 생기게 된다. 이런 경우 트리플이 특정 클래스로 분류되지 못하는 애매성이 있으므로 매칭 트리플 데이터에서 제거하였다[5].

매칭 트리플에서 빈도수 100 이상인 12개의 관계명으로 실험을 진행하였다. 매칭 트리플 중 일부(12,997개)를 학습 데이터로 사용하였고, 나머지(1,439개)를 실험 데이터로 사용하였다. 비매칭 데이터는 무작위로 샘플링

한 비매칭 트리플에 수동으로 yagoFacts 관계명을 부착하여 생성하였다. 선택된 12개의 관계명과 데이터별 빈도수는 [표 2]와 같다.

표 2. 관계명의 데이터별 빈도

관계명	학습 데이터	실험 데이터	비매칭 데이터
isLocatedIn	4,598	525	85
wasBornIn	4,069	442	52
hasChild	716	85	6
isMarriedTo	730	83	6
isAffiliatedTo	675	74	52
diedIn	654	72	10
happenedIn	383	45	2
influences	316	35	0
graduatedFrom	294	29	9
hasCapital	195	21	1
hasWonPrize	244	20	3
created	123	8	3
계	12,997	1,439	229

실험 결과는 [표 3]과 같다. [표 3]에서 실험 데이터 모델은 학습 데이터로 학습하고 실험 데이터로 측정한 결과이다. 비매칭 데이터 모델은 학습 데이터로 학습하고 비매칭 데이터로 측정한 결과이다.

성능 측정을 통해 실험 데이터 모델이 96.42%의 F1-Measure를 보이는 것을 확인하였고 비매칭 데이터 모델의 경우 그보다 낮은 79.04%의 F1-Measure를 보이는 것을 확인하였다.

실험 결과 비매칭 데이터 모델에서 'isAffiliatedTo'로 분류되어야 할 트리플들이 'wasBornIn'으로 분류되었다. 잘못 분류된 20개의 트리플 중 18개는 목적어 카테고리와 목적어 타입이 존재하지 않고, 주어 위키백과 카테고리에 '태어남'이 포함되어 있는 경우 해당 트리플을 'wasBornIn'으로 분류하였다. 또한 비매칭 데이터에 'influences'가 나타나지 않아서 전체 성능이 낮아졌다.

표 3. 시스템 성능

	Accuracy	Macro Precision	Macro Recall	F1 Measure
실험 데이터 모델	0.9708	0.9660	0.9623	0.9642
비매칭 데이터 모델	0.8734	0.7830	0.7980	0.7904

5. 결론 및 향후 연구

본 논문에서는 yagoFacts와 디비피디아 인포박스의 트리플을 매칭하여 디비피디아 트리플의 세분화된 속성들을 YAGO 관계명으로 분류하는 방법과 매칭된 트리플을 MEM을 통해 학습하여 비매칭 트리플을 YAGO 관계명으로 자동 분류하는 방법을 제안하였다. 제안한 방법으로 실험한 결과 79.04%의 F1-Measure로 비매칭 트리플의 관계명을 분류할 수 있었다.

향후 연구로 실험에 사용한 12개의 관계명들 이외에 다른 관계명들로 트리플을 매칭할 수 있는 방법을 연구하여 다양한 관계명이 실험에 적용될 수 있도록 하고 다양한 자질들을 사용하여 시스템의 성능을 향상시킬 수 있도록 할 예정이다.

감사의 글

본 연구는 산업통상자원부/미래창조과학부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10041678, 다중영역 정보서비스를 위한 대화형 개인 비서 소프트웨어 원천 기술 개발]. 또한 이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2013R1A1A4A01005074).

참고문헌

[1] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago
[2] http://dbpedia.org/About
[3] http://en.wikipedia.org/wiki/Watson_(computer)
[4] Berger, Adam L., Vincent J. Della Pietra, and Stephen A. Della Pietra. "A maximum entropy approach to natural language processing." Computational linguistics 22.1, pp.39-71, 1996.
[5] Aprosio, Alessio Palmero, Claudio Giuliano, and Alberto Lavelli. "Extending the Coverage of DBpedia Properties using Distant Supervision over Wikipedia." NLP-DBPEDIA@ ISWC. 2013.
[6] Kim, Eun-kyung, Matthias Weidl, and Key-Sun Choi. "Metadata Synchronization between Bilingual Resources: Case Study in Wikipedia." MSW. 2010.
[7] http://cran.r-project.org/web/packages/maxent/maxent.pdf