관계추출 모델 학습을 위한 반자동 패턴 마이닝

최규현○, 남상하, 최기선
한국과학기술원, 기계독습연구실
wiany11@kaist.ac.kr, nam.sangha@kaist.ac.kr, kschoi@kaist.ac.kr

Semiautomatic Pattern Mining for Training a Relation Extraction Model

GyuHyeon Choi○, Sangha nam, Key-Sun Choi
Korean Advanced Institute of Science and Technology, Machine Reading Lab.

요 약

본 논문은 비구조적인 자연어 문장으로부터 두 개체 사이의 관계를 표현하는 구조적인 트리플을 밝히는 관계추출에 관한 연구를 기술한다. 사람이 직접 언어적 분석을 통해 트리플이 표현되는 형식을 입력하여 관계를 추출하는 규칙 기반 접근법에 비해 기계가 데이터로부터 표현 형식을 학습하는 기계학습 기반 접근법은 더 다양한 표현 형식을 확보할 수 있다. 기계학습을 이용하려면 모델을 훈련하기 위한 학습 데이터가 필요한데 학습 데이터가 수집되는 방식에 따라 지도 학습, 원격지도 학습 등으로 구분할 수 있다. 지도 학습은 사람이 학습 데이터를 만들어야하므로 사람의 노력이 많이 필요한 단점이 있지만 양질의 데이터를 사용하는 만큼 고성능의 관계추출 모델을 만들기 용이하다. 원격지도 학습은 사람의 노력을 필요로 하지 않고 학습 데이터를 만들 수 있지만 데이터의 질이 떨어지는 만큼 높은 관계추출 모델의 성능을 기대하기 어렵다. 본 연구는 기계학습을 통해 관계추출 모델을 훈련하는데 있어 지도 학습과 원격지도 학습이 가지는 단점을 서로 보완하여 타협점을 제시하는 학습 방법을 제안한다.

주제어: 관계추출, 자연어처리, 패턴 마이닝, 패턴 인식

1. 서론

관계추출(Relation Extraction)은 비구조적인 자연어 문장으로부터 구조적인 트리플(triple)를 추출하는 작업을 의미한다. 트리플이란 두 개체 간의 관계(relation)를 <주어, 관계, 목적어>와 같이 세 개의 항으로 표현하는 구조이다. 예를 들어 “조선의 군주인 세종대왕은 정안군 이방원의 아들이다.”라는 문장으로부터 <세종대왕, 부모, 이방원>이라는 트리플을 추출할 수 있다. 이렇게 구조적으로 표현된 정보는 자연어로 표현된 정보보다 기계가 해석하기 수월하기 때문에 질의응답을 비롯한 다양한 분야에서 유용하게 사용될 수 있다. 기계가 해석하고 처리할 수 있는 정보는 활용가치가 높기 때문에 도처에 존재하는 방대한 자연어 데이터를 RDF(Resource Description Framework) 형식인 트리플로 표현하여 연결 데이터(Linked Data)를 구축하기 위한 관계추출이 지속적으로 연구되고 있다.

관계추출에 관한 접근법은 크게 규칙 기반과 기계학습 기반으로 나눌 수 있다. 규칙 기반 접근법은 사람이 언어적인 분석을 통하여 트리플이 문장에서 표현되는 형식을 찾아 그 패턴을 규칙으로 사용하여 문장으로부터 트리플을 추출한다. 사람이 직접 정의하는 만큼 패턴의 질은 높을 것으로 기대할 수 있지만 간과하거나 미처 생각하지 못한 패턴이 많이 있을 수 있다. 더 많은 패턴을 정의하기 위해서는 사람 또한 데이터를 분석해야하는데 사람이 분석할 수 있는 데이터의 양은 한정적이기 때문에 충분한 양의 패턴을 얻기 힘든 단점이 있다. 기계학습 기반 접근법은 자연어 문장으로부터 기계가 직접 패턴을 발견하여 학습한다. 기계는 사람과는 비교할 수 없을 정도로 많은 데이터를 분석할 수 있기 때문에 다양한 패턴을 수집할 수 있다는 장점이 있다.

기계학습 기반 접근법은 학습 데이터를 준비하는 과정에 따라 다시 지도 학습(Supervised Learning), 원격지도 학습(Distantly Supervision) 등으로 나뉜다. 지도 학습으로 모델을 훈련하기 위해서는 사람이 직접 학습 데이터를 준비해야한다. 손으로 충분한 양의 학습 데이터를 만드는 것은 굉장히 힘든 작업이지만 학습 데이터의 질은 우수하기 때문에 괜찮은 성능의 관계추출 모델을 기대할 수 있다. 원격지도 학습에서는 기존의 지식 베이스(Knowledge Base)에 있는 트리플을 활용하여 자동으로 학습 데이터를 준비할 수 있다. 특정 관계에 대한 학습 데이터를 준비하기 위해 먼저 그 관계를 포함하는 트리플을 모으고 트리플에 나타난 두 개체를 모두 포함하는 문장을 수집하여 학습 데이터로 사용할 수 있다. 구체적인 예시로, <세종대왕, 부모, 이방원>이라는 트리플을 사용하여 “조선의 군주인 세종대왕은 정안군 이방원의 아들이다.”와 같이 두 개체 세종대왕과과 이방원을 모두 포함하는 문장을 수집하는 것이다. 사람이 수고롭게 학습 데이터를 만들 필요가 없는 것은 원격지도 학습의 장점이지만 “태종 이방원의 노력이 있었기에 세종대왕은 많은 치적을 남길 수 있었다.”라는 문장이 이방원이 세종대왕의 부모라는 사실을 설명하지 않는 것처럼 수집된 자연어 문장이 반드시 올바른 트리플을 표현하지는 않기 때문에 학습 데이터의 질이 떨어지는 단점이 존재한다.

본 논문에서는 기계학습을 통해 관계추출 모델을 학습하는데 있어 지도 학습과 원격지도 학습 사이의 타협점을 찾아 향상된 성능의 모델을 훈련할 수 있는 방법을 제시하고자 한다. 손쉽게 학습 데이터를 수집할 수 있는 원격 지도적 방법에서 시작하되 학습 데이터를 그대로

사용하지 않고 사람의 힘을 빌려 학습에 필요한 패턴을 수집(mining)하게 된다. 사람이 훈련 데이터를 만들기 위해 막대한 시간과 노력을 들일 필요는 없지만 적정 수준의 개입을 통해 학습 데이터의 질을 개선하고 이를 토대로 모델을 훈련하여 관계추출의 정확도를 향상시킬 수 있다.

2. 관련 연구

1998년 일곱 번째 MUC(Message Understanding Conference)에서 관계추출에 대한 이론이 공식적으로 진술된 이후 자연어처리 및 기계학습 분야에서 관련 연구가 많이 수행되어오고 있다.

전통적인 기계학습 접근법인 지도학습을 사용하되 분류기(classifier)의 성능을 높이기 위해 고려되는 자질의 종류와 수에 따라 다양한 방식으로 분류기를 학습하는 관계추출 연구가 영어를 대상으로 수행되었다 [1]. 사용된 자질은 단어, 개체의 종류(named-entity type), 개체의 명사가 표현되는 방식, 구(phrase), 의존구조(dependency tree) 등을 포함한다. 또한 자질로부터 또 다른 자질을 계산하는 함수를 정의하는데, 여러 개의 단어가 포함된 범주(category)까지 고려한 자질과 의존구조를 후처리한 정보(parse tree) 등이 있다. 마지막으로 워드넷(WordNet)을 이용한 의미론적인 자질까지 고려하여 총 8개의 자질을 사용해 서포트 벡터 머신(Support Vector Machine) 분류기를 훈련한다. 실험에 사용된 학습 데이터는 약 300만개의 단어로 구성된 674개의 주석이 달린 문서로 구성된 말뭉치로 9683개의 관계를 표현한다. 이중 1386개의 관계를 표현하는 약 5만 단어로 구성된 97개의 문서가 분류기 학습에 사용되었다. 여러 가지 자질을 포함하여 학습된 분류기는 84.8%의 정밀도와 66.7%의 재현율을 보였다.

영어를 대상으로 원격지도학습을 도입하여 수행된 연구가 있다 [2]. 프리베이스(Freebase), 위키피디아의 구조화된 데이터, NNDB(biographical information), MusicBrainz(music), SEC(financial and corporate data) 등으로부터 수집된 정보를 처리하여 총 7,300종류의 관계에 대해 900만 개의 개체 사이에서 1억 1600만 개의 트리플로 구성된 지식베이스를 만들었다. 말뭉치는 프리베이스와 위키피디아로부터 추출된 약 1800만 개의 문서로 구성되어 있으며 문서 당 문장 수는 평균 14.3개, 총 단어는 약 600만 개다. 이중 800만 개의 문서가 학습 데이터로, 400만 개의 문서가 시험 데이터로 사용되었다. 학습은 다계층 로지스틱 회귀(multi-class logistic regression) 분류기를 훈련하였고 실험에 대한 결과는 교차 검증(cross-validation)으로 평가하였다. 실험에 대한 결과는 재현율이 5% 정도일 때는 정밀도가 75%~80% 사이로 꽤 높았지만 재현율이 40%~45%로 올라감에 따라 정밀도가 20% 이하로 급격히 하락하였다.

한국어를 대상으로 원격지도학습을 사용한 서술어 연결 연구도 수행된 바 있다 [3]. 해당 연구에서는 한국어 디비피디아(DBpedia)와 한국어 위키피디아(Wikipedia)를 사용하였다. 한국어 디비피디아에 있는 215만여 개의 관계 중 약 95%를 차지하는 1500개의 관계 중 영어나 숫자가 포함된 관계를 제외하고 약 1300개의 씨앗 관계를 대상으로 실험을 진행하였다. 각 씨앗 관계에 해당하는 트리플에서 주어와 목적어를 수집하여, 주어와 목적어를 모두 포함하는 문장을 약 270만 개의 한국어 위키피디아 문장으로부터 수집해 원격지도적으로 학습 데이터를 만들었다. 이후 해당 연구에서 정의한 삼항관계화라는 과정을 통해 문장으로부터 트리플의 표현방식을 추출하고 나이브 베이즈(Naive Bayes) 분류기를 훈련하였다. 실험 검증은 405개의 관계에 대한 서술어 연결 결과를 세 명의 평가자가 수동으로 평가한 다음 다수의 판단 결과를 정답으로 인정하는 방식으로 진행되었다. 서술어 연결에 대한 재현율은 정확한 측정이 어려운 관계로 약 63.70%로 추정하였고 연결된 서술어 854개 중 444개가 타당한 것으로 판단되어 약 51.99% 정밀도를 보였다.

3. 방법

본 연구의 주된 방향은 지도 학습과 원격지도 학습 사이에서 서로의 단점을 보완하는 타협점을 찾는 것이다. 지도학습의 단점은 사람이 직접 양질의 학습 데이터를 준비하는 과정에 필요한 노력이 매우 크다는 점이고 비지도학습의 단점은 자동으로 수집되는 학습 데이터가 상당량의 잡음(noise)을 포함하고 있어 질이 떨어진다는 것이다. 따라서 사람이 적정수준으로 개입하여 양질의 학습 데이터를 얻고 관계추출 모델을 훈련할 수 있는 방법을 찾고자 한다.

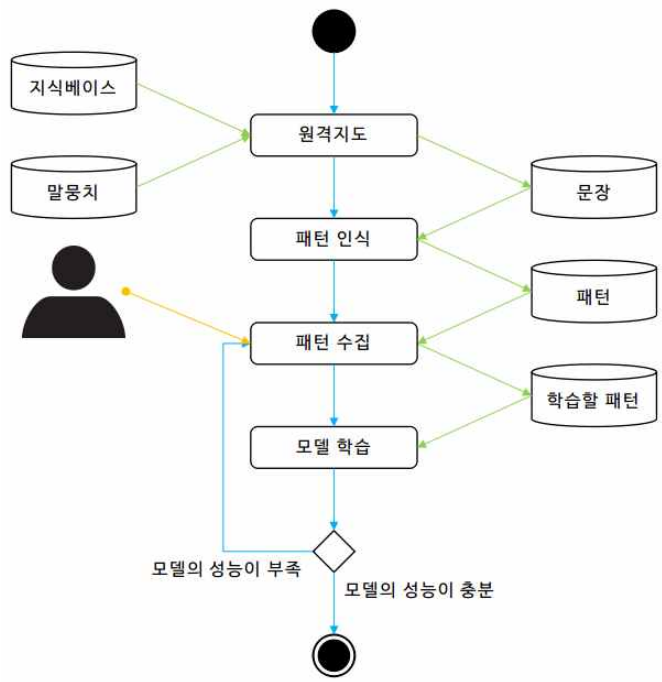

그림 1 관계추출 모델이 학습되는 과정

관계추출 모델을 훈련하는 과정은 그림 1과 같다. 그림 1에서 볼 수 있듯이 모델의 학습 과정은 원격지도적으로 문장을 수집하는 것에서 시작한다. 본 연구에서는

패턴 인식(Pattern Recognition)이라는 기계학습 개념을 사용하여 관계추출 모델을 학습한다. 패턴이 학습 데이터가 되기 때문에 먼저 문장이 트리플을 표현하는 형식을 주요 단어와 상대적인 위치 정보로 해석하여 패턴을 만드는 방법을 정의할 필요성이 있다. 그 다음 정의한 방법에 따라 패턴을 만들어 모델 학습 및 시험에 사용할 수 있다. 추출한 패턴을 모두 긍정적인 학습 데이터로 사용하여 순수한 원격지도학습으로 모델을 훈련하는 대신 사람이 수동으로 한 번 더 패턴을 걸러내도록 한다. 이렇게 얻어진 패턴을 학습 데이터로 모델을 훈련하고 시험 데이터에 대해 관계추출을 수행한다. 표본 평가(sample evaluation) 결과를 참고하여 모델의 성능이 충분해질 때까지 모델 학습 과정을 반복할 수 있다. 각 중요 과정의 자세한 설명은 하위 단원에서 설명한다.

3.1. 패턴의 정의

패턴은 문장의 의존구조로부터 만들어진다. 패턴의 속성값(attribute value)은 의존구조 상에 있는 각 어절에서의 어근 혹은 트리플을 구성하는 주어와 목적어의 조사이다. 의존구조 상에서 어절의 상대적인 위치 또한 패턴의 속성값을 결정짓는 데 중요한 역할을 하게 된다.

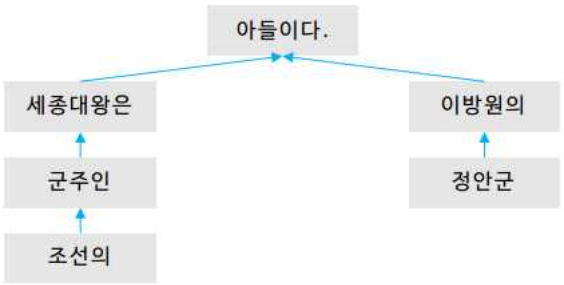

그림 2 문장의 의존구조

그림 2는 “조선의 군주인 세종대왕은 정안군 이방원의 아들이다.”라는 문장의 의존구조를 나타내고 있다. 의존구조는 문장을 구성하는 각 어절이 노드가 되므로 각각의 어절이 서로 어떻게 의존하고 있는지 알 수 있다.

예시 문장이 트리플 <세종대왕, 부모, 이방원>을 표현할 수 있으므로 부모라는 관계를 표현하는 패턴을 추출할 수 있다. 패턴을 추출하는데 중심 역할을 하는 세 개의 노드는 주어와 목적어에 해당하는 두 개의 노드와 두 노드의 첫 번째 공통 조상 노드(common ancestor node)이다. 공통 조상 노드는 서술어 자리에 위치하기 때문에 편의를 위해 이후부터 서술어 노드로 부르도록 한다.

그림 3은 예시문장으로부터 추출된 예시 패턴이다. 먼저 문장의 의존구조에 빈 노드가 7개 추가된 것을 볼 수 있는데, 이는 패턴을 추출하는 과정에서 주어, 목적어, 서술어, 세 개의 노드뿐만 아니라 그 주변 노드까지 고려하기 때문이다. 문장의 의존구조 상에서 어느 노드 주변에 위한 노드들은 주로 그 노드를 꾸며주는 역할을 하기 때문에 관계를 표현하는데 있어 굉장히 중요한 정보를 제공할 수도 있고 반대로 문장의 의미를 완전히 바꿔버릴 수도 있다. 따라서 세 개의 중심노드로부터 주변에 위치한 노드 또한 반드시 고려되어야 한다. 그림 3의 예시 패턴에서는 주변 노드를 세 개의 중심 노드로부터 깊이 2까지의 들어오거나 나가는 노드로 정의하였다. 서술어 노드에 대해서는 들어오는 노드가 많아 어느 노드가 들어오는 노드인지 정의하기 다소 애매한 부분이 있어 나가는 노드만 고려한다. 주변 노드가 없거나 기대하는 개수보다 적을 경우 빈 노드를 추가하여 위치 정보를 보존한다.

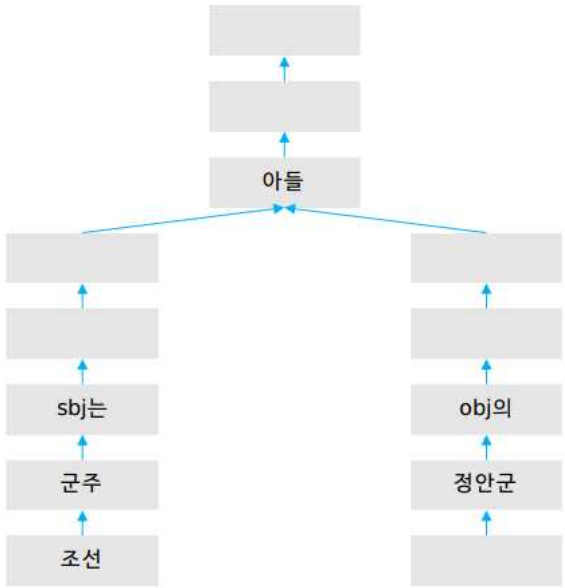

그림 3 문장에서 추출한 패턴

다음 각 노드에 대한 후처리 작업을 통해 패턴의 다양성(diversity)를 줄인다. 우선 주어와 목적어 노드는 조사 정보만을 남긴다. 이 때 조사의 다양성 또한 낮추기 위해 ‘-은’을 ‘-는’으로 바꾼 것처럼 대표조사를 사용한다. 서술어 노드를 비롯한 나머지 노드에 대해서는 어근 정보만을 남기게 된다. 어근만 있어도 의존구조 자체 내의 위치에 따라 어근이 하는 역할을 어느 정도 한정할 수 있기 때문에 이러한 일반화를 통해 패턴의 의미는 간직하면서 개수가 불필요하게 많아지는 것을 방지하게 된다.

패턴을 문자열로 표현하더라도 패턴의 구조적인 정보는 반드시 유지되어야 한다. 표 1은 문자열로 표현된 예시 패턴을 보여준다. 언더바(‘_’) 기호는 위치정보를 유지하기 위한 비어 있는 노드를 의미한다. 주어 노드는 세 번째, 목적어 노드는 여덟 번째, 서술어 노드는 열한 번째에 위치하며, 주어, 목적어 노드의 왼쪽 자리 두 개는 들어오는 노드, 주어, 목적어, 서술어 노드 모두에 대해 오른쪽 자리 두 개는 나가는 노드를 의미한다.

표 1 문자열로 표현된 패턴

패턴
obj 아들 sbj가 _ _ _ _ obj의 아들 _ 편집 _ _
_ _ sbj는 _ _ _ 아버지 obj의 계구 공부 계승 _ _
_ _ sbj는 _ _ _ 어머니 obj와 함께 _ 살 _ _
_ _ sbj는 _ _ _ _ obj와 함평이 사이 태어나 _ _
_ _ sbj를 _ _ _ _ obj는 혼인 _ 낳 되 _

3.2. 패턴의 수집

원격지도적 방법을 사용하여 자동으로 모아진 패턴을 그대로 학습 데이터로 사용하면 데이터의 질이 나쁘기 때문에 관계추출 모델의 성능을 기대하기 어렵다. 따라서 학습 데이터로 사용할 패턴을 수동적인 방법으로 사람이 다시 수집하는 과정이 필요하다.

본 연구의 목표가 사람이 학습 데이터를 만드는데 필요한 상당량의 노력을 적정 수준으로 줄이며 기계의 도움을 받아 비교적 양질의 학습 데이터를 얻는데 있는 만큼, 패턴을 수집하는 작업은 최대한 간결해져야 한다. 이를 위해 패턴을 일반화하는 작업이 선행되어야 한다.

패턴은 특정 관계를 설명하는데 별로 필요치 않은 어휘 정보까지 포함하고 있어 대개 필요 이상으로 세부적이다. 그림 3의 패턴에서 속성값 '조선,' '군주,' '정안군' 등은 사실 아버지라는 관계를 설명하는데 있어서 중요하지 않은 정보다. 이런 불필요한 속성값은 관계추출 모델을 학습하는데 과적합(overfitting)의 문제를 야기할 뿐만 아니라 패턴의 다양성을 증가시켜 패턴을 수집하는데 필요한 작업량이 많아지게 한다. 이와 같은 이유로 패턴을 구성하는 속성값의 집합을 수동으로 정의하여 불필요한 속성값은 배제하도록 한다. 패턴을 수집하는 과정에서 각 속성값이 얼마나 빈번하게 등장하였는지 참조하면 속성값의 집합을 정의하는데 많은 도움을 받을 수 있다.

표 2 부모 관계에서 각 속성값이 차지하는 비율

속성값	비율
%SBJ%의	5.37 %
아들	3.54 %
%OBJ%의	3.24 %
…	
%OBJ%를	1.81 %
하-	1.74 %
…	
있-	1.55 %
%OBJ%와	1.34 %
되-	1.21 %
죽-	1.05 %
…	

속성값	비율
아버지	0.90 %
사이	0.88 %
딸	0.88 %
태어나-	0.84 %
것	0.74 %
…	
%SBJ%를	0.70 %
낳-	0.68 %
…	
어머니	0.34 %
…	
동생	0.29 %

표 2는 부모 관계에 대해 원격지도적으로 수집된 패턴으로부터 나타나는 속성값에 대한 통계자료로, 비중이 큰 상위 40개의 속성값 중 일부를 제외한 18개의 속성값이 전체에서 차지하는 비율을 나타내고 있다.

먼저 이 속성값들이 어떤 역할을 할 수 있는지 언어적 지식을 바탕으로 살펴볼 수 있다. '아들,' '아버지,' '딸,' '어머니' 와 같은 속성값은 완벽하게 부모라는 관계를 설명할 수 있다. 반면 '동생' 은 부모라는 관계를 설명하기에는 부족하다. '낳-' 과 같은 속성값도 '누가 누구를 낳았다' 라는 표현으로 충분히 부모라는 관계를 설명할 수 있다. 하지만 '태어나-' 와 같은 경우는 '누구와 누구 사이에서 태어났다' 라는 표현을 사용하면 부모라는 관계를 나타낼 수 있어도 '~에서 태어났다' 처럼 사용될 수도 있기 때문에 혼자서는 부모라는 관계에 대한 충분한 정보를 제공한다고 보기는 힘들다. 이 외에 상당히 큰 비중을 차지하는 '하-,' '있-,' '되-,' '죽-,' '것' 과 같은 속성값은 독립적으로 부모라는 관계를 표현할 수 있다고 보기는 어렵다. 이런 관찰을 통해 패턴을 구성하는데 필요한 속성값 집합을 정의한다.

속성값 집합이 정해지면 패턴을 일반화할 수 있다. 속성값 집합에 속하지 않은 패턴의 구성요소는 모두 빈 노드로 대체하는 것이다. 아래 표 3은 표 1의 패턴이 일반화 과정을 거친 이후의 패턴을 나타내고 있다.

표 3 일반화된 패턴

패턴
obj 아들 sbj가 _ _ _ _ obj의 아들 _ _ _ _
_ _ sbj는 _ _ _ 아버지 obj의 _ _ _ _ _
_ _ sbj는 _ _ _ 어머니 obj와 함_ _ _ _ _
_ _ sbj는 _ _ _ _ obj와 _ 사이 태어나 _ _
_ _ sbj를 _ _ _ _ obj는 _ _ 낳 _ _

일반화 과정을 거치면 패턴의 다양성이 줄어든다. 이러한 사실은 자연어의 특성과도 연관이 있는 부분이다. 자연어에서 특정 관계를 표현하는 표현 방식과 표현에 사용되는 어휘를 비교해보면 표현 방식이 훨씬 다양하게 나타난다. 왜냐하면 같은 어휘라도 어휘가 나타나는 위치나 어휘를 꾸며주는 수식어들에 의해 다양하게 표현될 수 있기 때문이다.

일반화과정을 거친 후 주요 속성값을 하나도 포함하고 있지 않은 패턴은 수집 대상에서 미리 제거한다. 이후 패턴의 다양성이 얼마나 줄어들었는지 살펴보면 아래에 있는 표 3과 같다.

표 3 일반화 과정을 통한 패턴의 다양성 변화

관계	일반화 전	일반화 후
occupation	36,884 개	410 개
birthPlace	15,931 개	264 개
region	13,956 개	356 개
parent	727 개	272 개
award	402 개	50 개

표 3의 결과는 상당히 긍정적이다. 패턴의 다양성을

줄어들자 고려해야할 패턴이 적게는 63%에서 많게는 98% 가까이 줄어들었다. 패턴마다 자연어로 패턴을 표현할 수 있는 방법의 수가 제각기 다르므로 일반화된 이후 패턴의 종류도 서로 다르지만 일반화되기 전과 비교해 보면 그 편차가 굉장히 작다는 것을 알 수 있다. 즉, 관계를 표현하는데 반드시 필요한 핵심 패턴은 얼마 되지 않는 것이다. 확인해야할 패턴의 수가 줄어들기 때문에 패턴을 수집하는 작업이 굉장히 수월해진다. 수천, 수만 개의 학습 데이터를 만들거나 혹은 검증할 필요 없이 수백 개 정도의 패턴만 확인하면 모델을 학습하는데 필요한 패턴을 수집할 수 있게 된 것이다.

패턴을 만드는데 꼭 필요한 속성값을 정의하고 일반화된 패턴으로부터 학습에 필요한 패턴을 골라내는 작업이 현재 단순히 기계만으로는 해결할 수 없다고 판단하여 사람의 수작업을 동원하였다. 여기서 어차피 마지막에 패턴을 수집하는 게 사람이라면 처음부터 사람이 패턴을 정의하는 편이 더 나을 지도 모른다는 의문이 생기는데, 아무것도 없는 상태에서 손으로 패턴을 만드는 것과 있는 패턴으로부터 패턴을 수집하는 것에는 상당한 차이가 있다. 실제로 사람이 생각할 수 있는 표현 방식은 굉장히 제한적이다. 어느 표현 방식을 보고 올바른 표현인지 아닌지 구분할 수 있어도 다양한 표현 방식을 모두 기억하기란 상대적으로 굉장히 어렵기 마련이다. 따라서 기계의 힘을 빌려 반자동으로 학습 데이터로 사용할 수 있는 패턴을 수집하는 것은 효율적이다. 이러한 상호작용적인 과정을 반복하며 표본 평가를 통해 모델의 관계추출 정확도가 수렴하거나 일정 수준 이상이 될 때까지 반자동으로 학습 데이터의 질을 개선하며 모델을 훈련할 수 있다.

4. 실험

한국어 디비피디아와 한국어 위키피디아를 가공하여 실험에 사용된 지식베이스와 말뭉치를 만들었다. 한국어 디비피디아에서는 다양한 서술어를 그대로 관계로 사용하는 경우가 많아 2만여 개의 관계가 존재한다. 때문에 2만여 개의 관계를 영어 디비피디아에서 정의하는 240개의 개체형 관계(Object-type relation)로 변환하는 작업을 통해 각 관계에 속하는 트리플의 개수를 늘려 학습이 용이하도록 하였다. 말뭉치는 534,768 한국어 위키피디아 문서로부터 길이가 너무 짧거나 긴 문장을 제외하고 문장의 구문 분석을 방해하는 이상한 기호를 제거하는 등의 전처리 과정을 거쳐 만들어졌다.

학습 단계에서는 원격지도적으로 학습 데이터로 사용될 패턴을 수집하였고 네 명의 사람이 서로 다른 관계에 대해 반자동으로 패턴을 주석 처리하는 작업을 거쳐 총 200개의 관계에 대해 관계추출 모델을 훈련하였다. 훈련된 모델로 관계추출을 수행하였고 157개의 모델이 1개 이상의 트리플을 추출하였고, 그중 78개의 모델이 100개 이상의 트리플을 추출하였다. 나머지 43개의 모델은 트리플을 추출하지 못하였다.

추출된 트리플은 표본 검사를 통해 정확도를 추정하였고 추출된 트리플의 수에 따른 관계별 표본 정확도는 아래 표 5와 같다.

표 5 트리플이 가장 많이 추출된 10개의 관계에 대한 모델의 표본 정확도

관계	삼항관계 추출 수	표본 정확도
occupation	50,447	0.91
region	21,095	0.98
birthPlace	15,577	0.84
locationCountry	10,462	0.77
genre	10,243	0.73
team	9,985	0.8
residence	9,782	0.88
location	8,884	0.78
education	5,949	0.92
nationality	5,215	0.82

1000개 이상의 트리플이 추출된 관계 중 정확도가 높은 상위 10개의 관계와 해당하는 표본 정확도는 아래 표 6과 같다.

표 6 1000개 이상의 트리플이 추출된 관계 중 정확도가 높은 상위 10개의 관계에 대한 결과

관계	트리플 추출 수	표본 정확도
region	21,095	0.98
country	3,045	0.98
citizenship	1,417	0.97
youthClub	4,327	0.97
almaMater	5,067	0.96
parentOrganisation	2,590	0.95
education	5,949	0.92
industry	1,419	0.92
award	50,447	0.91
battle	3,968	0.9

모든 관계에 대해 도합 217,909개의 트리플을 추출하였으며 관계 구분 없이 측정한 표본 정확도는 0.85 정도로 전반적으로 우수한 관계추출 성능을 보여준다. 학습 데이터로 사용된 반자동으로 수집한 패턴의 질이 우수함을 알 수 있다.

5. 결론

본 논문은 적정 수준에서 사람이 개입함으로써 양질의 패턴으로 구성된 학습 데이터를 준비하는 방법에 대해 설명하였다. 사람이 직접 학습 데이터를 만드는데 필요한 어마어마한 노력에 비해 필요한 패턴을 구성하는 속성값의 집합을 줄임으로써 패턴을 일반화한 다양성이 줄어든 패턴을 다시 한 번 걸러주는 정도의 노력을 통해 반자동으로 양질의 패턴을 상당수 수집하고 이를 학습 데이터로 사용할 수 있음을 실험을 통해 검증하였다.

한편 반자동으로 얻은 학습 데이터를 사용하여 훈련한 관계추출 모델이 정확도 측면에서 훌륭한 성능을 보였는

데, 이러한 결과는 패턴을 구성하는 속성값이 전혀 많을 필요가 없다는 점을 보여준다. 이는 언어적 직관과도 상통하는 부분인데, 자연어에서 특정 관계를 표현하는 방법은 비록 다양할지라도 다양한 표현에 사용되는 어휘의 종류는 비교적 제한적이기 때문이다. 특정 관계를 표현하는 패턴을 구성하는데 필요한 속성값이 사실 크지 않은 집합이라면 기계가 속성값의 집합을 구하도록 하는 것 또한 가능할 수 있다. 이와 관련된 후속 연구를 통해 본 연구에서 제안한 반자동인 관계추출 모델 훈련 방법을 상당부분 자동화할 수 있을 것으로 기대할 수 있다.

사사

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임.
이 논문은 2016년도 미래창조과학부의 재원으로 한국연구재단 바이오의료기술개발사업의 지원을 받아 수행된 연구임.

참고문헌

[1] G. Zhou, J. Su, J. Zhang, and M. Zhang. “Exploring Various Knowledge in Relation Extraction,” ACL 2005, 427-434, 2005.
[2] M. Mike, B. Steven, S. Rion, and J. Dan. “Distant supervision for relation extraction without labeled data,” ACL 2009, 1003-1011, 2009.
[3] 원유성, 우종성, 김지성, 함영균, 최기선, “한국어 서술어와 지식베이스 프로퍼티 연결,” 정보과학회 논문지, 제42권, 제12호, pp. 1568-1574, 2015.