지식베이스 확장을 위한 자동 관계 추출

울산과학기술원 ▌**임성우·한지연·이교운·최재식***

1. 서 론

지식 베이스(Knowledge Base)는 데이터의 비정형적인 관계를 유연하게 표현할 수 있는 효과적인 수단이다. YAGO[1] 및 위키데이터(Wikidata)[1] 같은 지식 베이스는 자동 질의 시스템 등 다양한 인공지능 기술을 현실화하는데 매우 중요하게 사용되고 있다. 그럼에도 불구하고 이런 방대한 지식 베이스를 수동으로 구축하고 유지하는 것은 현실적으로 매우 어렵다. 따라서, 컴퓨터 소프트웨어가 책과 인터넷 등 자료를 스스로 읽어, 중요한 관계를 자동으로 추출할 수 있는 기술(Relation Extraction)은 지식 베이스를 자동으로 확장(Knowledge Base Population)하는데 매우 중요한 기술이다.

관계를 자동으로 추출하는 다양한 기술이 존재하지만[3, 9, 13], 그 중에서 슬롯 채우기(Slot-Filling) 문제는 관계(예, 본사의 위치)와 키워드(예, 구글)가 주어졌을 때, 주어진 문서에서 키워드와 관계를 갖는 답(예, 캘리포니아)을 찾는 문제로 정의된다. 예를 들어, "*구글은 지난 5월부터 캘리포니아 본사 주변에서 차량 공유 서비스를 시범 운영 …*"이라는 문장을 읽고, 구글의 본사가 캘리포니아에 있다는 관계를 찾아, 지식 베이스에 추가하는 작업이다.

슬롯 채우기 문제를 해결하면 기존의 지식 베이스에 저장된 관계가 없는 경우에도, 자연어 문서 만으로 새로운 관계를 추출하여 지식 베이스를 확장할 수 있으며, 추론 알고리즘을 통하여 새로운 지식(관계)을 도출할 수도 있다. 이렇게 확장된 지식 베이스는 개인화된 자동 질의 서비스의 질을 향상 시키고, 의사나 변호사가 환자나 판례에 관련한 방대한 자료를 찾는데 걸리는 시간을 현저하게 줄일 수 있다[4].

관련하여 미국 국립표준기술연구소(NIST)는 매년 문서 분석 학술 대회(TAC, Text Analysis Conference)를 개최하고 있으며, 2009년부터는 지식 베이스 확장 분과(KBP, Knowledge Base Population Track)를 집중적으로 장려하고 있다[2]. 슬롯 채우기는 KBP 분과에서 개최하고 스탠포드대[5], 뉴욕대[6], 매사추세스대[7], 카네기 멜론대[8]등 주요한 연구 기관이 참가하고 있는 주요한 대회 부문이다. 슬롯 채우기는 주어진 관계에 대하여, 주어진 키워드에 대한 답을 방대한 자료에서 찾는 것을 평가한다.

올해 슬롯 채우기는 65개 관계에 대해서, 1,350개의 키워드에 대한 답을 30,000개 문서(약 922,663개 문장)에서 찾도록 구성되었다. 본고에서는 본 연구팀이 슬롯 채우기에 참가하기 위해 개발한, 학습 데이터(문장)에서 주요한 특징 특징을 찾는 기계 학습 모델을 소개하고, 주어진 문장에서 관계를 효과적으로 검출하기 위해 구현된 분산 시스템에 대해서 소개한다.

2. 문서 분석 학술 대회 지식 베이스 확장 분과 2016(TAC KBP 2016)

본고에서는 TAC KBP 2016중에서 슬롯 채우기(Cold Start Slot Filling)를 소개한다[3]. 여기서 Cold Start 문제(희박성 문제)란, 선언적으로 정의된 관계가 희박한 상황에서 대량의 문서들이 주어졌을 때 처음부터 지식 베이스를 구축하는 것을 의미한다. 이는 새로운 관계를 문서에서 학습하기 위한 핵심적인 문제로, 올해 슬롯 채우기는 부모, 출생지 등 사람에 관련된 관계와 주요 임직원, 본사 위치 등 기업, 기관에 관련된 관계를 포함하여 총 65개 관계에 대해서 1350개의 질의에 대한 답을 30,000개 문서(약 922,663개 문장)에서 찾도록 구성되었다.

* 종신회원

† 본 연구는미래창조과학부의 재원으로 한국연구재단의 기초연구사업(NRF-2014R1A1A1002662)과 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업(IITP-2016-R2720-16-0007)의 지원으로 수행되었음.

1) https://www.wikidata.org/

2) http://tac.nist.gov/tracks/index.html

3) http://tac.nist.gov/2016/KBP/ColdStart/index.html

표 1. 슬롯 채우기 대회의 슬롯 예제: 형식은 정답의 종류를 의미한다.

슬롯	형식	설명
per:title	String	사람(키워드)의 직함/직업 관계
per:spouse	Name	사람(키워드)의 부부 관계
per:employee_or_member_of	Name	사람(키워드)가 종사하거나 몸담는 단체
per:parents	Name	사람(키워드)의 부모 관계
per:date_of_birth	Value	사람(키워드)이 태어난 날 관계
org:top_members_employees	Name	기관(키워드)의 고위 임직원 관계
org:subsidiaries	Name	기관(키워드)의 계열사 관계
org:city_of_headquaters	Name	기관(키워드)의 본사가 있는 도시 관계
org:shareholders	Name	기관(키워드)의 주주 관계

대회 참가자는 질의(query)들과 자연어로 이루어진 문서 데이터를 받는다. 각 질의는 <키워드>와 한 개(혹은 두 개)의 <관계>가 주어지는데, 키워드와 첫 번째 관계에 있는 답을 내는 것이 1단계(round 1), 1단계에서 낸 답을 다시 키워드로 두 번째 관계에 대해 답을 내는 것이 2 단계(round 2)이다.

예를 들어, "버락 오바마"를 키워드로 배우자(첫 번째 관계)와 그 배우자의 출생지(두 번째 관계)를 질의하고 싶다면, <키워드-"버락 오바마", 관계1-"배우자", 관계2-"출생지">의 질의를 할 수 있다. 참가자가 질의를 받아 1단계에서 버락 오바마와 배우자 관계에 있는 사람으로 미셸 오바마를 답으로 냈다면, 2단계에서는 미셸 오바마의 출생지를 찾아 답으로 낸다.

올해 대회에는 1350개의 1단계 문제와 697개의 2단계 문제가 출제되었다. 즉, 653개의 질의는 1단계 문제만 포함하고 있다. 점수는 세가지 방식으로 평가되는데, 첫 번째는 참가자가 제출한 정답 중 얼마나 맞았는지 계산한 정밀도(precision)[4]이고, 두 번째는 주최 측에서 제시한 정답 중 참가자가 맞춘 것이 얼마나 되는지 계산한 재현율(recall)[5]이며, 마지막은 앞의 두 점수의 조화 평균인 F1 점수(F1 score)[6]이다.

3. 슬롯 채우기를 위한 기계 학습 기술

이 장에서는 슬롯 채우기를 위한 원거리 감독법을 이용한 관계 특징 추출 알고리즘과 인공 신경망을 이용한 문장 분석을 설명한다.

3.1 원거리 감독법(distant supervision)

원거리 감독법 기반 특징으로 학습된 시스템들은 현재까지 슬롯 채우기 문제를 성공적으로 해결한 경우가 많다[9]. 원거리 감독법은 어떤 관계(예, "수도")를 가진 키워드 쌍(예, "대한민국", "서울특별시")이 포함되어 있는 문장이 있다면, 해당 문장은 두 키워드의 관계(예, "배우자")를 표현할 가능성이 높다는 가정에 기초하여 특징 추출 생성에 드는 비용을 줄이는 방법이다. 예를 들어 "대한민국"과 "서울특별시"의 관계가 "수도"라는 것을 알고 있을 때, "서울특별시는 대한민국의 수도이자 최대 도시이다."라는 문장은 "수도" 관계를 표현한다고 가정한다.

원거리 감독법을 이용한 관계 특징 추출 알고리즘[9]은 특정 관계를 가지는 키워드 쌍을 추출하기 위해 해당 관계를 표현하는 문장의 특징을 이용한다. 이전 "수도"의 예에서, "B는 A의 수도이자 최대 도시이다"라는 특징을 저장한 다음 B와 A 자리에 다른 키워드 쌍이

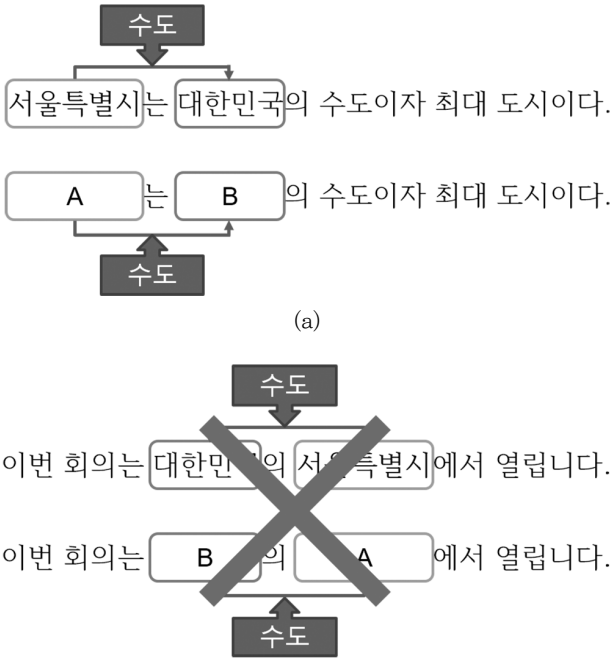


그림 1 원거리 감독법을 이용한 특징 추출의 예시

4) precision = 맞은 정답 수/제출한 정답 수
5) recall = 맞은 정답 수/실제 정답 수
6) F1 score = 2*precision*recall/(precision+recall)

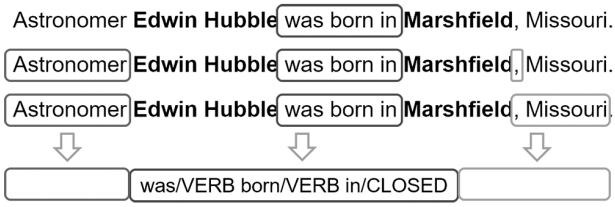


그림 2 한 문장에서 나올 수 있는 여러 가지 특징들의 예시[9]

들어간 문장을 찾게 되면 A와 B의 관계가 "수도"라고 추출한다(그림 1-(a)). 다양한 종류의 특징을 정의 할 수 있는데, 기본적으로 널리 쓰이는 특징은 두 키워드 사이의 단어 구문, 키워드 앞과 뒤에 나온 단어 구문 그리고 문장의 의존 관계 트리(dependency tree)에서 두 키워드 사이의 경로 정보 등이 있다.

하지만 원거리 감독법으로 추출한 특징이 관계를 온전히 표현한다고 판단하기는 어렵다. 키워드가 다른 관계를 표현하는 경우가 있을 수 있고, 더 나아가서는 두 키워드가 한 문장에 우연하게 한 문장에 표현되어 있을 수도 있다. "대한민국"과 "서울특별시"는 "수도"의 관계를 가지고 있지만, "속한 도시"의 관계 또한 가지고 있다. 예를 들어, "이번 회의는 대한민국의 서울특별시에서 열립니다."라는 문장에서 원거리 감독법은 "이번 회의는 A의 B에서 열립니다"라는 특징이 "수도" 관계를 표현한다고 잘못된 가정을 할 수 있다(그림 1-(b)). 또 다른 예로, "대한민국 사람인 철수는 2002년 서울을 방문하였습니다." 라는 문장은 대한민국과 서울의 관계를 정확하게 표현하지 않을 수 있다.

이러한 오류를 해결하기 위한 방법으로 능동 학습(Active Learning)을 이용하여 노이즈를 줄여 해결할 수 있고[10][11], 두 키워드가 포함된 여러 관계를 동시에 추출 및 학습할 수 있으며[12], 원거리 감독법으로 반례(negative example)를 생성하여 반례 및 정례의 데이터 불균형을 해소함[14]으로 해결할 수 있다.

스탠포드 대학에서 개발한 DeepDive시스템[14]은 특징점 추출의 오류 및 불확실성을 제거하는 근본적인 방법으로 확률 관계형 모델(Statistical Relational Model) 기법을 사용하였다. 관계 추출에 필요한 사람들의 상식을 모사하기 위해서 마코프 논리 네트워크(MLN: Markov Logic Network)으로 지식을 확률적으로 표현(factor graph)하고, 가능한 답을 추론하여 답으로 결정된 확률은 높인다. 예를 들어, 부부관계를 찾는 질의의 경우에, 평상적으로 부인은 한 명이 존재한다는 제약조건을 마코프 논리 네트워크에 설정해 주면, 문장에서 서로 다른 두 사람을 동시에 부인으로 답할 확률을 줄 일 수 있다.

3.2 인공 신경망(Neural Network)을 이용한 문장 분석

특징 기반 관계 추출 알고리즘의 경우 특징을 추출할 때 전적으로 키워드 쌍이 동시에 들어간 문장에 의존한다. 하지만 키워드가 대명사와 같은 동일 지시어로 바뀐 문장에서는 특징과 답을 찾지 못한다는 단점을 가지고 있다. 가령 관계대명사를 사이에 둔 긴 문장의 경우 대명사를 키워드라 인식하지 못하기 때문에 이러한 문제를 푸는데 어려움이 있다. 또한 추출된 특징을 바탕으로 관계에 해당하는 문장을 따로 분류 한다 하더라도 정답 후보군을 결정할 수 없는 문제의 경우 답을 찾을 수 없다. 사람/장소/기관의 경우 개체명 인식(Named Entity Recognition)이 가능하여 후보군을 줄일 수 있지만 나이, 직업 같은 경우 적용하기 어려워 정확한 정답을 찾기 어려운 경우도 존재한다. 예를 들어 태어난 도시 관계를 나타내는 "A는 B에서 태어나 C에서 자랐다"라는 문장이 있을 때, 이 문장 안에 도시를 나타내는 단어들(B, C)이 정답 후보군이 되어 분류기를 통해 정답을 추출 할 수 있지만, 직업관계를 나타내는 "A는 B로서 7년간 일했다" 라는 문장에서는 직업인 단어들이 개체명 인식이 안되기 때문에 정답 후보군을 추출하기 어렵다.

이러한 경우 인공 신경망을 이용하여 슬롯 채우기 문제를 해결하려는 시도로 최근 양방향 RNN(Bi-directional

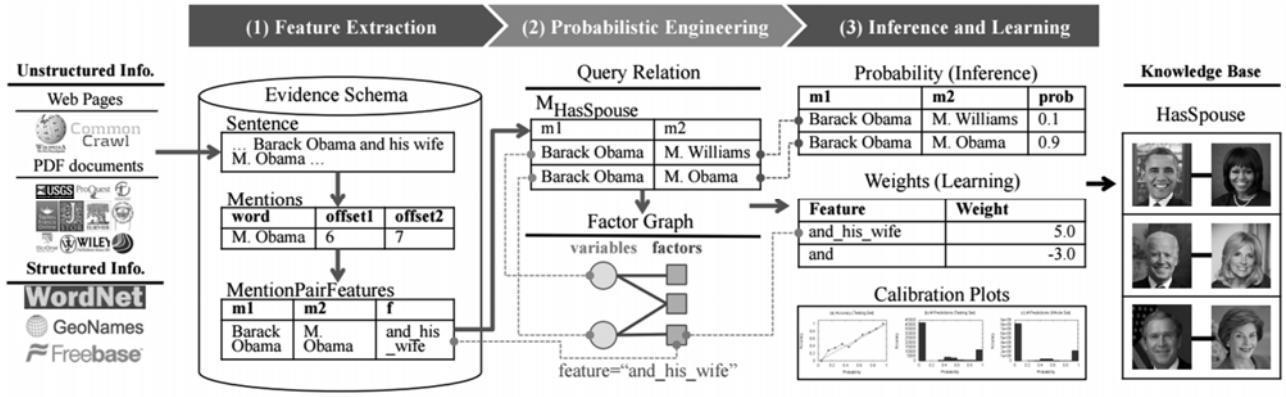


그림 3 스탠포드 대학의 DeepDive 시스템 구조[9]

RNN) 모델[2]이 제안되었다. RNN은 종단간(end-to-end) 학습을 통해서 이전 또는 이후의 정보를 자동으로 근사하는 등 학습이 용이한 점이 장점이다.

4. 시스템 구현

이 장에서는 본 연구팀이 TAC KBP 2016 대회 중 Cold Start 슬롯 채우기 문제를 풀기 위해 구축한 시스템에 대해서 설명한다. 우선 문장에서 개체명 인식을 통해 정답 후보군을 추출 할 수 있는 경우, 원거리 감독법에 기반한 시스템을 이용하여 문제를 해결한다. 만약 문장 내에서 정답 후보군을 추출 할 수 없는 경우, 인공 신경망에 기반한 시스템을 이용하여 문제를 해결한다.

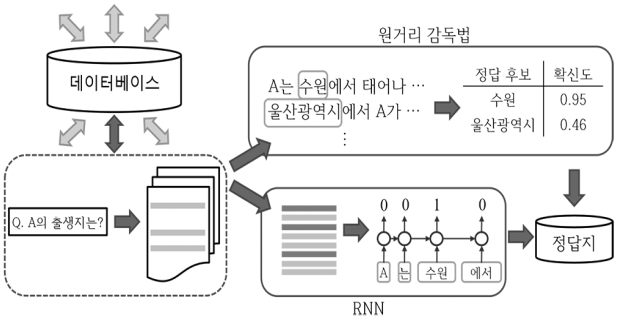


그림 4 시스템 전체 구조도

4.1 특징 기반 문제 해결

원거리 감독법을 이용한 특징 기반 관계 추출 알고리즘이 좋은 성능을 보이기 위한 조건 중 하나는 해당 관계를 표현하는 특징들을 추출하기 위한 충분한 학습 데이터가 존재해야 한다는 것이다. 하지만 학습 데이터가 많으면 많을수록 저장되는 특징(해당 관계를 표현하는 특징뿐만 아니라 표현하지 않는 특징도 함께)의 수가 늘어나고 그로 인해 후에 학습 과정과 실제 적용 과정에서의 계산량 또한 증가하게 되어 속도와 성능을 저하시킬 수 있다. 이러한 문제를 해결하기 위해 일반적으로는 특징 추출 과정 후 추출 과정에서 k번 이하로 나온 특징을 없애는 방법을 사용한다. 또한 해당 알고리즘은 정확하게 일치하는 특징에 대해서만 좋은 성능을 보이기 때문에 같은 의미라도 다양하게 표현 할 수 있는 자연어의 특성상 재현율을 높이기가 힘들다.

본 연구팀은 위와 같은 문제를 해결하기 위하여 복합 특징을 사용하지 않고 개별적인 특징을 추출하였다. 그림 2를 보면 기존의 특징 추출 과정은 키워드 쌍 앞/사이/뒤 특징과 함께 품사 태그까지 모두 하나의 특징으로 저장하였다. 하지만 본고에서는 키워드 쌍 앞/사이/뒤 특징을 모두 분리시키고, 품사 태그와 개체명 인식, 의존 관계 트리에서 두 키워드 사이의 경로들을 모두 분리시켰다(그림 5).

이 방법은 다음과 같은 세 가지 기대효과를 얻을 수 있다. 첫 번째는 더 적은 특징으로 더 많은 특징을 표현 가능하게 된다. 이는 학습 데이터의 필요량을 줄여 줄 수 있다. 두 번째는 데이터가 충분히 많을 경우 중복되는 특징이 다르게 처리되는 경우를 줄인다. 마지막으로

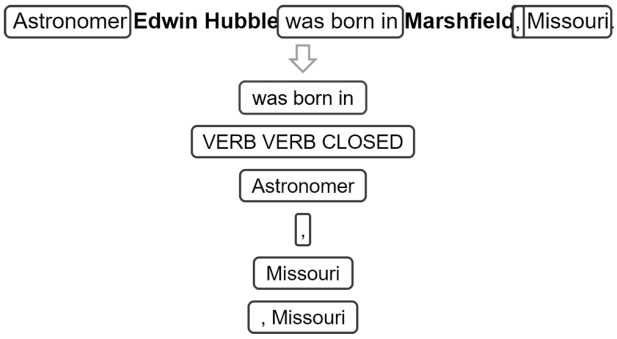


그림 5 본 연구팀에서 사용한 특징 추출의 예

표 2 본고에 사용된 특징의 종류

분류	특징
특징의 위치 기반	키워드 쌍 사이 단어 구문
	키워드 쌍 이전 단어 구문(최대 3 단어)
	키워드 쌍 이후 단어 구문(최대 3 단어)
	의존 관계 트리의 경로(사이): 문장의 의존 관계 트리[7]
	의존 관계 트리의 경로(앞): 문장을 의존 관계 트리에서 먼저 나온 키워드와 연결된 단어(최대 1단어)
	의존 관계 트리의 경로(뒤): 문장을 의존 관계 트리에서 뒤에 나온 키워드와 연결된 단어(최대 1단어)
특징 속성 기반	단어의 원형: 단어의 원형[8]
	품사 태그: 단어의 품사[9]
	개체명 태그: 단어의 개체명[10]
기타	키워드 쌍 사이의 단어 개수: 키워드 쌍 사이의 단어 개수
	관계의 방향: 방향성이 중요한 관계의 경우, 키워드의 순서. (예, "태어난 장소"를 찾는 문제에서 장소가 먼저 나온 문장과 사람이 먼저 나온 문장의 특징은 다르다)

에서 두 키워드를 연결하는 경로에 있는 단어 구문(사람/기관/장소/시간 등)

총 특징의 종류는 (6 (특징 위치) × 3 (특징 속성)) × 2 (관계의 방향) + 1 (단어 개수) = 37 종류가 있다.

7) Natural Language ToolKit(NLTK)와 Stanford dependencies parser를 사용하였다.

8) NLTK와 WordNet을 사용하였다.

9) NLTK와 Stanford Part-Of-Speech (POS) tagger를 사용하였다.

10) NLTK와 Stanford Named Entity Recognizer(NER)를 사용하였다.

특정 특징들은 함께 있을 때 의미를 가지게 되는데 이를 학습 모델을 통해 자동으로 학습 할 수 있다.

또한, 특정 특징의 경우 문맥의 흐름 상 해당 관계를 표현하기도 하고 표현하지 않기도 하는데(예시:"철수(24)/서울특별시"에서 철수가 태어난 장소를 찾는 경우), 본고에서는 해당하는 특징을 직접 정하고 해당 특징이 들어간 학습 데이터를 모두 반례(negative)로 처리를 한 후 학습하였다. 본고에 사용된 특징의 종류는 위와 같다.

학습 모델은 64개의 노드를 갖는 완전 연결 층(Fully-Connected layer)를 2층 쌓은 모델을 사용하였다. 모델의 입력 값은 추출된 특징의 개수만큼의 크기를 가지는 벡터로, 각각의 값은 0 혹은 1로 표현되며 학습하고자 하는 문장에 해당 특징이 포함 된 경우 1 아니면 0 값을 주었다. 출력 값은 2개의 노드로 표현하여 각각 해당 후보가 정답이 아닐 확률과 정답이 맞을 확률을 표현한다.

4.2 인공 신경망 기반 문제 해결

3.2에서 설명한 바와 같이 원거리 감독법을 이용한 특징 기반 관계 추출 알고리즘을 사용하기 위해서는 문장 내에서 주어진 키워드와 정답 후보군을 추려낸 후 특징이 관계를 표현하는지 구분한다. 하지만 특정 관계(예시: 특정 사람이 저지른 범죄의 종류, 특정 사람의 직업, 특정 회사의 홈페이지 도메인)에 대해서는 문장 안에 정답이 있더라도 개체명 인식을 통해 정답 후보군을 추려낼 수 없으므로 특징 기반 관계 추출을 하기가 어려운 문제가 있다. 이를 해결하기 위해 본 연구팀은 두 종류의 인공 신경망을 사용하였다. 첫 번째는 문장을 읽고 해당 문장이 특정 관계를 표현하는지를 판단하고, 두 번째는 문장을 읽고 어느 위치에 정답이 위치하는지를 판단한다.

인공 신경망의 입력으로는 단어의 Word2Vec 벡터에 다음의 벡터들을 순서대로 연결해서 사용하였다. 해당 단어의 품사 태그를 one-hot encoding으로 표현한 것 / 개체명 태그를 one-hot encoding으로 표현한 것 / 해당 단어가 Word2Vec에 포함되어 있는지(0으로 표현) 혹은 포함되어 있지 않은지(1로 표현) / 단어가 없는지 (1로 표현) 혹은 있는지(0으로 표현).

첫 번째 인공 신경망에서는 주어진 키워드의 앞 10단어의 단어 벡터와 뒤 10단어의 단어 벡터를 순서대로 사용하여 양방향 Long-Short Term Networks (LSTM)의 입력 값으로 사용하였다. 양방향 LSTM 층 위에는 1개의 완전 연결 층(Fully-Connected layer)을 추가하였고 최종 출력은 2개의 노드로 각각 "관계를 표현한다/하지 않는다"를 표현하게 하였다.

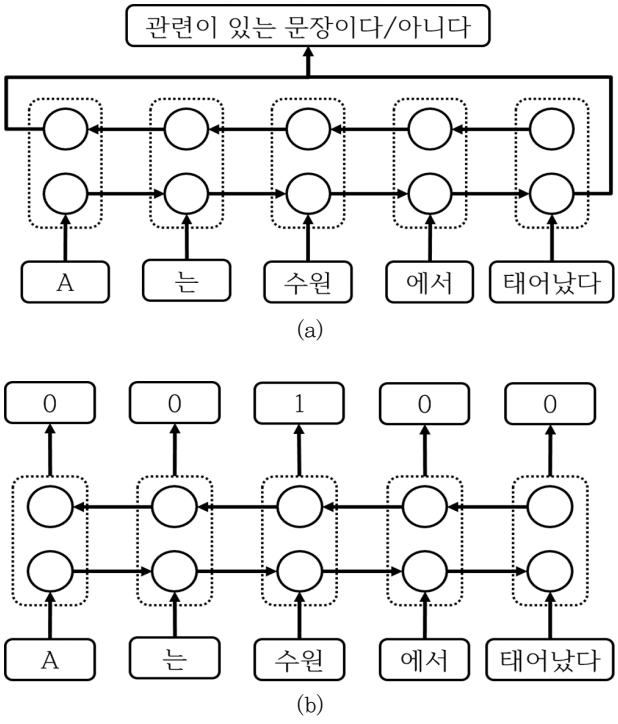


그림 6 (a)관련 있는 문장을 판별하는 LSTM, (b)정답인 단어의 위치를 찾는 LSTM

두 번째 인공 신경망에서는 앞과 같이 입력 값은 주어진 키워드의 앞 10단어의 단어 벡터와 뒤 10단어의 단어 벡터를 순서대로 사용하여 양방향 LSTM의 입력 값으로 사용하였고, 양방향 LSTM 층 위에 1개의 완전 연결 층(Fully-Connected layer)를 추가한 후 최종 출력을 20개의 노드로 설정하였다. 20개의 노드는 각각 순서대로 키워드 앞 10단어와 키워드 뒤 10단어를 표현하고 노드의 값이 해당 단어가 정답일 확률을 표현하게 하였다.

4.3 대용량 문장 처리를 위한 분산 시스템

특징에 기반한 방법과 인공 신경망을 사용한 방법 둘 다 문장을 전처리(품사 태그/ 개체명 태그 등) 한 후 사용한다는 공통점이 있다. 실험 환경에서 한 개의 문장을 전처리 할 때 평균적으로 10초가 걸리므로 TAC KBP 2016 대회에 있는 데이터를 모두 처리하기 위해 걸리는 예상 시간은 10 (초) × 922,663 (TAC KBP 2016 데이터에 있는 문장의 총 개수) / 3,600 (초) = 2,562 (시간)이다. 하지만 실제 슬롯 채우기를 위해서는 키워드가 들어간 문장만 처리해야 하므로 예상되는 시간은 10 (초) × 79,265 (처리해야 할 문장의 예상 개수)[11] / 3,600 (초) = 220 (시간) 이 된다.

11) (TAC KBP 2015 데이터에 키워드가 들어간 문장 수) × (TAC KBP 2016 데이터 개수) / (TAC KBP 2015 데이터 개수)로 계산하였다.

표 3. 실험결과[12)]

모델 \ 역치	0.5			0.7			0.9		
	P	R	F1	P	R	F1	P	R	F1
원거리 감독법	0.13	0.34	0.19	0.20	0.25	0.22	0.48	0.10	0.17
원거리 감독법+규칙	0.18	0.28	**0.22**	0.35	0.21	**0.26**	0.61	0.10	**0.17**
정례 기반 원거리 감독법	0.15	0.23	0.18	0.20	0.21	0.20	0.25	0.12	0.16
정례 기반 원거리 감독법+규칙	0.20	0.18	0.19	0.35	0.17	0.23	0.60	0.08	0.15

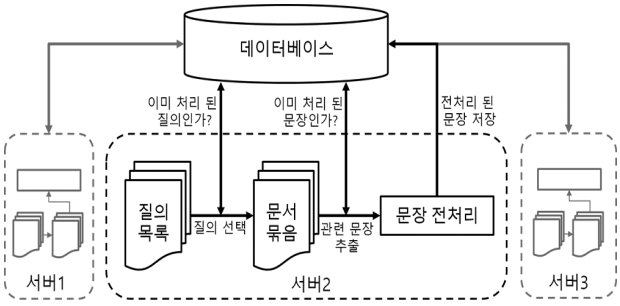


그림 7 분산 시스템 구조도

본고에서는 문장을 처리하는데 걸리는 시간을 줄이기 위해 8개의 서버에 분산 시스템을 구축하였다.

서버 간의 연결을 위하여 MongoDB를 사용하였고, 호스트 서버의 MongoDB에 다른 서버들이 연결되어 계산을 분산하였다. 분산 시스템을 이용한 처리 과정은 다음과 같다(그림 7). 서버에서 질의를 읽고 해당 질의가 이미 처리되었거나 처리되는 중인지 데이터베이스에서 확인한다. 만약 질의가 이미 처리되었거나 처리되는 중이면 다음 질의를 읽어온다. 이 과정을 반복 후 아직 처리 되지 않았거나 처리되는 중이 아닌 질의를 찾으면 해당 질의를 푼다. 질의를 풀기 위해 키워드가 들어간 문장을 찾게 되는데 혹시 해당 키워드가 이전에 한번 처리된 키워드인지 데이터베이스에서 확인한다. 만약 이전에 처리되었던 정보가 있으면 정보를 받아오고 아니면 서버에서 직접 전처리를 수행한 후 데이터베이스에 저장한다.

본 연구팀은 해당 분산처리를 사용하여 TAC KBP 2016 데이터에 대하여 1시간에 약 6679개의 문장을 처리하는 속도를 구현하였다. 이는 질의를 분산한 효과와 질의들 중 중복된 키워드에 대해 처리하는 시간을 줄인 효과다.

5. 실험 결과

5.1 슬롯 채우기 데이터

실험을 위한 학습 데이터에는 TAC KBP 2012, 2013, 2014, 2015 데이터를 사용하였다 (이후 지난 TAC KBP 데이터로 표현). 지난 TAC KBP 슬롯 채우기 평가지에는 정답과 정답을 증명하는 문장이 모두 포함되어 있으므로 해당 문장들을 사용하였다. 원격 지도법을 사용하기 위해 키워드와 정답이 모두 정확히 답지와 문제지와 일치하게 표현된 문장만 사용하니 해당 도시에 본사를 둔 기관(gpe:headquarters_in_city)은 128개의 문장, 해당 도시에 살고 있는 사람(gpe:residents_of_city)은 216개의 문장, 해당 기관에서 공부한 학생(org:students)은 102개의 문장 등 실제 사용 가능한 데이터의 개수가 부족하였다. 본 연구팀은 부족한 데이터의 양을 늘리기 위해 위키 데이터에서 특정 관계를 갖는 키워드 쌍을 추출했다. 추출된 키워드 쌍을 구글에 검색하여 두 키워드가 모두 들어간 문장을 찾은 후 문장이 관계를 표현하는지 직접 표기했다. 그 결과 사람의 부모(per:parents)는 800개, 기관의 본사의 위치(org:city_of_headquarters)는 750개, 기관의 임원(org:top_members_employees)은 500개, 사람이 저지른 범죄(per:charges)와 사람이 죽은 도시(per:city_of_death),기관의 설립자(org:founded_by)는 400개, 사람의 직업(per:title)은 300개, 기관의 자회사(org:subsidiaries)는 200개의 데이터를 추가했다.

반례 데이터를 만들기 위해 지난 TAC KBP 데이터에서 키워드는 들어가지만 정답이 들어가지 않는 문장을 키워드 당 최대 10 문장을 무작위로 추출하여 반례 데이터로 사용하였다.

5.2 결과

TAC KBP 2016 슬롯 채우기 대회의 결과는 아직 나오지 않았으므로 본고에서는 TAC KBP 2012, 2013 데이터를 학습 데이터로, TAC KBP 2014 데이터를 시험 데이터(test data)로 사용하였다. 본고에서는 해당 지역에 본사를 둔 기관(gpe:heaquarters_in_city)에 대해 아래와 같은 변형을 주어 비교한다.

1) 원거리 감독법: 원거리 감독법으로 특징을 추출하고 2번 이하 나온 특징 제거 후 학습
2) 규칙: 특정 특징을 직접 반례로 처리한 후 학습

12) P: 정확도, R: 재현율, F1: F1 점수

3) 정례 기반 원거리 감독법: 특징이 반례에서 나온 횟수를 이용해 횟수가 3번 이상일 경우 특징을 지운 후 학습

또한 각각의 경우에 역치값(threshold)을 변경 하였을 경우의 결과도 같이 첨부하였다(표 3).

결과를 보면 "원거리 감독법 + 규칙"이 가장 좋은 성능을 보여준다. "정례 기반 원거리 감독법"의 경우에는 성능이 오히려 떨어진 것을 볼 수 있다. 본 연구팀이 사용한 특징 추출 방법 상 많은 특징들이 정례와 반례 모두에 공통적으로 나타나고 이를 제거하는 과정을 거치면서 많은 특징들이 없어진다. 그러나 본 연구팀은 문장에서 일치하는 특징이 없을 경우에 대해서 처리를 하지 않아 학습 과정 중 주요한 특징 없이 의미 없는 특징으로 학습하게 되어 F1 점수가 하락했다. 이와 비슷한 문제가 "원거리 감독법"의 경우에도 나타난다. "철수의 출생 정보는 2000.9.10/서울이다."라는 문장에서 "의 출생정보는 2000.9.10/"와 "이다."라는 특징으로 학습을 하는데 "의 출생정보는 2000.9.10/"라는 특징이 나오는 빈도가 적을 경우 학습 전에 해당 특징이 지워지게 되어 "이다."라는 특징만으로 학습이 진행된다. 그 결과, "길동이가 2005년부터 시작한 음식점의 위치는 경기도이다."라는 문장에서 “가 2005년부터 시작한 음식점의 위치는”이라는 특징이 없을 경우 “이다.”라는 특징만으로 정답을 판단하여 길동이의 출생지에 대해 “경기도”라고 판단을 내린다. 또한 정확도에 비해 재현율이 높게 상승하지 않는다. 본 연구팀의 시스템이 문장 내에 키워드와 정답이 모두 있는 경우에 대해서만 문제를 풀 수 있기 때문에 대명사를 사용하거나 약어, 별칭을 사용한 경우 문제를 풀지 못하기 때문이다.

6. 결 론

본고에서는 자연어 처리 분야의 중요한 문제 중 하나인 슬롯 채우기의 설명과 슬롯 채우기를 풀기 위한 방법들, 본 연구팀이 TAC KBP 2016에 참가하며 구현한 시스템에 대해서 설명하였다. 슬롯 채우기 문제를 풀기 위해서는 품사 태그와 개체명 인식, 의존 관계 트리 등 많은 자연어 처리 기술이 사용된다. 그만큼 슬롯 채우기 문제가 복잡하고 어려우며 문제를 풀기 위해 많은 정보와 처리 기술이 필요하다. 원거리 감독법을 이용한 특징 추출 기술도 한 관계를 다양하게 표현하고 여러 관계를 한 문장에 표현 할 수 있는 자연어의 특성상 많은 어려움이 있다. 슬롯 채우기를 사람의 수준으로 풀기 위해서는 키워드와 관련된 문서들을 찾아 읽고 이해하며 지식을 쌓아 정답을 추론하는 기술이 필요할 것으로 보인다.

참고문헌

[1] F. Mahdisoltani, J. Biega and F. Suchanek, "Yago3: A knowledge base from multilingual wikipedias," in 7th Biennial Conference on Innovative Data Systems Research, 2014, .

[2] N. T. Vu, P. Gupta, H. Adel and H. Sch, "Bi-directional recurrent neural network with ranking loss for spoken language understanding," in *2016 IEEE ICASSP*, pp. 6060-6064, 2016.

[3] J. Fan, A. Kalyanpur, D. C. Gondek and D. A. Ferrucci, "Automatic knowledge extraction from documents," *IBM Journal of Research and Development*, vol. 56, pp. 5: 1-5: 10, 2012.

[4] T. F. Gordon, "A use case analysis of legal knowledge-based systems," *Legal Knowledge and Information Systems (JURIX 2003), Frontiers in Artificial Intelligence and Applications*, pp. 81-90, 2003.

[5] G. Angeli, V. Zhong, D. Chen, A. Chaganty, J. Bolton, M. J. Premkumar, P. Pasupat, S. Gupta and C. D. Manning, "Bootstrapped self training for knowledge base population," in *Proc. TAC* 2015, 2015.

[6] Y. He and R. Grishman, "The NYU Cold Start System for TAC 2015," .

[7] B. Roth, N. Monath, D. Belanger, E. Strubell, P. Verga and A. McCallum, "Building Knowledge Bases with Universal Schema: Cold Start and Slot-Filling Approaches," *TAC* 2015, 2015.

[8] B. Kisiel, B. McDowell, M. Gardner, N. Nakashole, E. A. Platanios, A. Saparov, S. Srivastava, D. Wijaya and T. Mitchell, "CMUML System for KBP 2015 Cold Start Slot Filling," .

[9] M. Mintz, S. Bills, R. Snow and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proc. of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 1003-1011, 2009.

[10] G. Angeli, J. Tibshirani, J. Wu and C. D. Manning, "Combining distant and partial supervision for relation extraction." in *EMNLP*, pp. 1556-1567, 2014.

[11] L. Sterckx, T. Demeester, J. Deleu and C. Develder, "Using active learning and semantic clustering for noise

reduction in distant supervision," in *4e Workshop on Automated Base Construction at NIPS2014 (AKBC-2014)*, pp. 1-6, 2014.

[12] M. Surdeanu, J. Tibshirani, R. Nallapati and C. D. Manning, "Multi-instance multi-label learning for relation extraction," in *Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 455-465, 2012.

[13] D. Zelenko, C. Aone and A. Richardella, "Kernel methods for relation extraction," *Journal of Machine Learning Research*, vol. 3, pp. 1083-1106, 2003.

[14] C. Zhang, *DeepDive: A Data Management System for Automatic Knowledge Base Construction*, 2015.

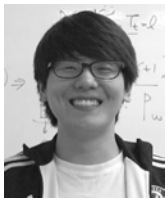
[15] 정석원, 최맹식, 김학수. "위키백과로부터 기계학습 기반한국어 지식베이스 구축." 한국정보과학회 논문지, 재42권 제8호 pp.1065-1070, 2015

[16] 윤희근, 박성배. "지식 베이스 확장을 위한 트리플 추출 " 한국정보과학회 논문지, 재34권 제8호 pp.17-24, 2016

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735-1780, 1997.

[18] M. Liwicki, A. Graves, H. Bunke and J. Schmidhuber, "A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks," in *Proc. 9th Int. Conf. on Document Analysis and Recognition*, pp. 367-371, 2007.

약 력



임 성 우

2015 울산과학기술원 졸업(학사)
2015～현재 울산과학기술원 전기전자컴퓨터공학과 석사과정
관심분야: 인공지능, 기계학습, 자연어처리
Email: seongwoo@unist.ac.kr

한 지 연

2016 울산과학기술원 졸업(학사)
2016～현재 울산과학기술원 전기전자컴퓨터공학과 석박사통합과정
관심분야: 인공지능, 기계학습, 자연어처리
Email: jiyeon@unist.ac.kr

이 교 운

2016 울산과학기술원 졸업(학사)
2016～현재 울산과학기술원 전기전자컴퓨터공학과 석박사통합과정
관심분야: 인공지능, 기계학습, 자연어처리, 자동통계학자
Email: leekwoon@unist.ac.kr

최 재 식

2004 서울대학교 컴퓨터공학과 졸업(학사)
2012 University of Illinois at Urbana-Champaign 전산학과 박사(졸업)
2013 Lawrence Berkeley National Laboratory 박사후연구원
2013～현재 울산과학기술원 전기전자컴퓨터공학과 조교수
관심분야: 인공지능, 기계학습, 컴퓨터비전, 로보틱스
Email: jaesik@unist.ac.kr