ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2019 March: 25(1): 43~61 http://dx.doi.org/10.13088/jiis.2019.25.1.043

지식베이스 구축을 위한 한국어 위키피디아의 학습 기반 지식추출 방법론 및 플랫폼 연구*

김재헌
기술연구소, 리스트
(*wogjs217@li-st.com*)

이명진
기술연구소, 리스트
(*mjlee@li-st.com*)

최근 4차 산업혁명과 함께 인공지능 기술에 대한 연구가 활발히 진행되고 있으며, 이전의 그 어느 때보다도 기술의 발전이 빠르게 진행되고 있는 추세이다. 이러한 인공지능 환경에서 양질의 지식베이스는 인공지능 기술의 향상 및 사용자 경험을 높이기 위한 기반 기술로써 중요한 역할을 하고 있다. 특히 최근에는 인공지능 스피커를 통한 질의응답과 같은 서비스의 기반 지식으로 활용되고 있다. 하지만 지식베이스를 구축하는 것은 사람의 많은 노력을 요하며, 이로 인해 지식을 구축하는데 많은 시간과 비용이 소모된다. 이러한 문제를 해결하기 위해 본 연구에서는 기계학습을 이용하여 지식베이스의 구조에 따라 학습을 수행하고, 이를 통해 자연어 문서로부터 지식을 추출하여 지식화하는 방법에 대해 제안하고자 한다. 이러한 방법의 적절성을 보이기 위해 DBpedia 온톨로지의 구조를 기반으로 학습을 수행하여 지식을 구축할 것이다. 즉, DBpedia의 온톨로지 구조에 따라 위키피디아 문서에 기술되어 있는 인포박스를 이용하여 학습을 수행하고 이를 바탕으로 자연어 텍스트로부터 지식을 추출하여 온톨로지화하기 위한 방법론을 제안하고자 한다. 학습을 바탕으로 지식을 추출하기 위한 과정은 문서 분류, 적합 문장 분류, 그리고 지식 추출 및 지식베이스 변환의 과정으로 이루어진다. 이와 같은 방법론에 따라 실제 지식 추출을 위한 플랫폼을 구축하였으며, 실험을 통해 본 연구에서 제안하고자 하는 방법론이 지식을 확장하는데 있어 유용하게 활용될 수 있음을 증명하였다. 이러한 방법을 통해 구축된 지식은 향후 지식베이스를 기반으로 한 인공지능을 위해 활용될 수 있을 것으로 판단된다.

주제어 : 딥러닝, 온톨로지, 인공지능, 지식베이스, 지식추출

논문접수일 : 2018년 12월 18일 논문수정일 : 2019년 3월 5일 게재확정일 : 2019년 3월 11일
원고유형 : 학술대회(급행) 교신저자 : 이명진

1. 개요

최근 4차 산업혁명과 함께 인공지능에 대한 관심이 점차 높아지고 있으며, 이에 따라 자율주행자동차, 영상, 로봇 등 다양한 분야에서 활발히 연구가 이루어지고 있는 실정이다. 이러한 인공지능 기술에 대한 연구는 1950년대부터 시작해서 최근까지 인간의 지능과 관련되어 있는 학습, 문제 해결 등과 같은 인지 문제를 해결하는데 많은 노력을 기울여 왔다(Russell et. al., 2009). 인공지능 분야는 최근 기술에 대한 높은 관심과 다양한 알고리즘에 대한 연구로 인해 그 어느 때보다도 많은 기술적 발전을 이루어가고 있다. 지식베이스를 기반으로 한 시스템은 인공

* 본 연구는 2017년도 중소벤처기업부의 창업성장기술개발사업 지원에 의한 연구임 [S2534527]

지능의 하위 영역으로서(Engelmore, 1984), 다양한 분야의 복잡하고 비정형화된 전문 지식과 규칙 등을 저장 및 운용하여 인공지능 에이전트가 의사결정을 하는데 활용할 수 있도록 하는 것을 목적으로 한다(Krishna, 1992). 이러한 지식베이스는 최근 기계학습을 수행하는데 있어 특징 선택(feature selection)을 위해 활용되거나 설명 가능한 기계학습 알고리즘(Explainable AI)에 적용되는 등 통계 기반의 인공지능과 융합되어 그 활용성이 점차 높아지고 있다(Bergman, 2014).

최근의 지식베이스는 인터넷이 보편화되고 확산됨에 따라 웹에 존재하는 자원들을 연결하고 의미를 부여함으로써 지식을 표현하고 공유하는 형태로 활용되고 있다. 또한 인공지능 스피커를 통해 이루어지는 질의응답 시스템 등 다양한 영역에서 지능화된 처리를 위한 기반 지식으로 활용되고 있다. 하지만 유용한 지식베이스를 구축하는 것은 여전히 전문가의 많은 시간과 노력을 요구한다(Forsythe, 1993). 최근에는 지식을 활용하는 많은 인공지능 기술들이 위키피디아(Wikipedia)를 사용하고 있으며(Higashinaka et. al., 2007; Kaisser, 2008), 위키피디아에 존재하는 각각의 위키페이지(wikipage)로부터 다양한 정보를 추출하여 이를 주어(subject), 동사(predicate), 목적어(object)로 구성된 트리플(tiple) 형태의 지식으로 변환하여 제공하는 DBpedia(Bizer, 2009)를 활용해 다양한 지식베이스 기반의 인공지능 서비스를 연구 및 개발하고 있다. DBpedia가 위키피디아로부터 제목, 분류, 이미지 등 다양한 정보를 추출하여 트리플 형태의 지식을 생성하고 있지만 가장 유용하게 활용될 수 있는 지식은 위키페이지에 존재하는 인포박스(infobox)로부터 사전에 정의된 맵핑 규칙 혹은 속성의 이름에 따라 지식을 생성하여 제공하는 것이다. 이처럼 DBpedia는 사용자가 작성한 반 정형화된 데이터로부터 지식을 생성하는 방법을 이용함으로써 지식의 정확도 측면에서는 높은 신뢰도를 기대할 수 있지만 이 역시 사람에 의해 작성되는 인포박스에 의존함으로써 결국 사람의 시간과 노력을 필요로 한다는 사실에는 변함이 없다. 또한 한국어 위키피디아의 경우 인포박스를 포함하고 있는 위키페이지가 약 50% 정도밖에 되지 않기 때문에 위키피디아에 작성되어 있는 인포박스로부터 데이터를 추출하여 지식을 추출하는 것은 지식의 확장성 측면에서 한계를 가지고 있다.

이러한 문제점을 해결하기 위해 본 논문에서는 온톨로지의 지식을 구축하기 위해 기계학습을 이용하여 지식베이스의 구조에 따라 학습을 수행하고, 이를 통해 자연어 문서로부터 지식을 추출하여 지식화하는 방법에 대해 제안하고자 한다. 이를 위한 다양한 선행 연구가 수행되었지만(Wu and Weld, 2007; Lange et. al., 2010; Brandão et. al., 2010), 해당 연구들은 지식베이스의 구조를 고려하지 않고 인포박스를 통해 학습한 단순 문자열을 그 값으로 추출하는데 초점이 맞추어져 있다. 하지만 본 연구에서는 지식베이스의 구조를 바탕으로 학습을 수행하고, 속성의 타입 및 속성이 가져야 하는 값의 형식인 XML 스키마 데이터타입에 따라 적절한 유형의 값을 추출하여 지식베이스를 구축한다. 이러한 방법의 적절성을 보이기 위해 DBpedia의 온톨로지 구조에 맞추어 자동화된 지식베이스 확장을 위한 기계학습 기반의 방법론을 제안하고자 한다. 이를 위해 DBpedia의 온톨로지 구조에 따라 위키피디아의 인포박스에 기술되어 있는 정보를 이용하여 학습 데이터를 만들고 학습을 수행하여, 지식을 추출할 수 있는 모델을 구축한다. 위

키피디아의 인포박스는 분류체계와 해당 분류가 가질 수 있는 속성이 템플릿(template)으로 정의되어 있으며, DBpedia는 이러한 템플릿을 바탕으로 온톨로지 구조가 이루어져 있다. 따라서 지식을 추출하는 모델은 입력된 문서에 대해 적합한 분류를 결정하고, 결정된 분류에 따라 추출할 지식을 식별한 후 문장 단위로 지식을 추출하기에 적합한 문장인지를 식별한다. 적합한 문장으로 식별되면 해당 문장으로부터 지식을 추출하고 이를 트리플 형태로 변환하여 지식베이스에 저장한다. 이와 같은 과정을 통해 자연어 문서로부터 DBpedia의 구조에 따라 지식을 추출하고 확장함으로써 구조화된 지식을 활용할 수 있게 된다. 이렇게 구축된 지식은 향후 지식베이스 기반의 인공지능을 위한 기초 지식으로 활용될 수 있다.

2장에서는 본 연구와 관련된 연구를 소개하며, 3장에서는 본 논문에서 제안하고자 하는 모델에 대해 자세히 살펴보고자 한다. 4장에서는 실험을 통해 본 논문에서 제안하는 방법이 실제로 유용함을 보이고, 마지막으로 결론과 향후 연구로 마치고자 한다.

2. 관련 연구

위키페이지의 인포박스는 전체 글에 대한 표 형태의 핵심적인 요약정보로써 위키 문법에 따라 일부 정형화된 형태를 가지고 있으며, DBpedia와 같이 구조화 된 메타 데이터를 생성하는데 큰 역할을 하고 있다. 초기에는 생물학과 관련된 위키페이지에서 분류체계 정보를 표현하기 위해 만들어 졌지만 현재는 모든 문서에 다양한 정보를 포함하는 형태로 작성된다. 하지만 이와 같은 유용성에도 불구하고 인포박스가 사용자에 의해 작성되어야 하기 때문에 인포박스를 생성하는데 많은 시간과 노력이 필요하다. 실제로 인포박스를 작성하기 위해서는 작성할 인포박스의 분류를 결정한 후 결정된 분류에 맞추어 속성을 작성하여야 한다. 물론 위키피디아가 사용자에 의해 만들어지는 사전이기 때문에 정의되어 있지 않은 분류를 만들거나 속성을 만드는 등 많은 자율성이 주어지지만, 분류에 따라 사용할 수 있는 속성이 템플릿(template) 형태로 정의되어 있기 때문에 분류 및 속성에 대한 결정 과정이 필요하다. 이로 인해 2010년에 발표된 논문에 따르면 인포박스를 포함하고 있는 위키페이지의 비율이 약 33% 정도에 불과하였으며(Lange et. al., 2010), 2018년 4월 기준 한국어 위키피디아의 경우 인포박스를 포함하고 있는 위키페이지의 비율이 약 41만 개의 전체 한국어 위키페이지 중 약 20만 개인 50%에 불과하다.

인포박스의 작성을 효과적으로 지원하거나 자동화된 인포박스의 생성을 위해 위키피디아의 구조화되지 않은 텍스트에서 자동화된 정보 추출을 이용하여 인포박스와 같은 구조화 된 데이터를 생성하는 다양한 연구가 수행되었다. 이를 위한 대표적인 연구로서 Wu and Weld(2007)는 위키피디아의 문서 및 문장을 분류하고 CRF (Conditional Random Field) 기법을 이용하여 인포박스를 생성하는 연구를 수행하였다. Lange et. al. (2010)이 제시한 iPopulator는 Wu and Weld(2007)의 연구와 동일한 절차와 학습 알고리즘을 사용하지만 속성에 대한 값을 추출하는데 있어 값의 구조적인 정보를 이용하여 값을 추출하는 방법을 적용하였으며, 대부분의 위키피디아 인포박스 템플릿을 대상으로 실험을 수행하였다. Brandão et. al.(2010)의 연구 역시 Wu and Weld

(2007)가 제시한 KYLIN과 동일한 절차와 학습 알고리즘을 사용하고 있지만 해당 시스템을 통해 나온 결과를 다시 학습하도록 자기 지도 학습(self-supervised learning)을 이용하여 성능 향상을 꾀하고 있다. Choi et. al.(2018)은 지식베이스 확장을 위해 위키피디아를 포함하여 웹에 존재하는 멀티소스 비정형 문서로부터 질의에 대한 정보를 추출하기 위한 시스템의 개발 방법론을 제안하였다. 이러한 논문들이 비정형 텍스트로부터 인포박스를 생성하는데 있어 훌륭한 방법론을 제안하고 있지만, 이는 단지 텍스트 형태의 값을 추출하는데 초점을 맞추고 있을 뿐 지식베이스의 구조에 대한 고려를 하고 있지 않다. DBpedia의 경우 인포박스 속성의 유형에 따라 그 값이 URI 형태가 될 수도 있으며, 일반적인 텍스트 형태의 값이 될 수도 있다. 따라서 인포박스를 생성하는데 있어 단순히 텍스트 형태의 값을 생성하는 것이 아니라 온톨로지 구조에 따라 링크를 포함하거나 적절한 데이터 타입을 가진 값을 추출할 수 있어야 한다.

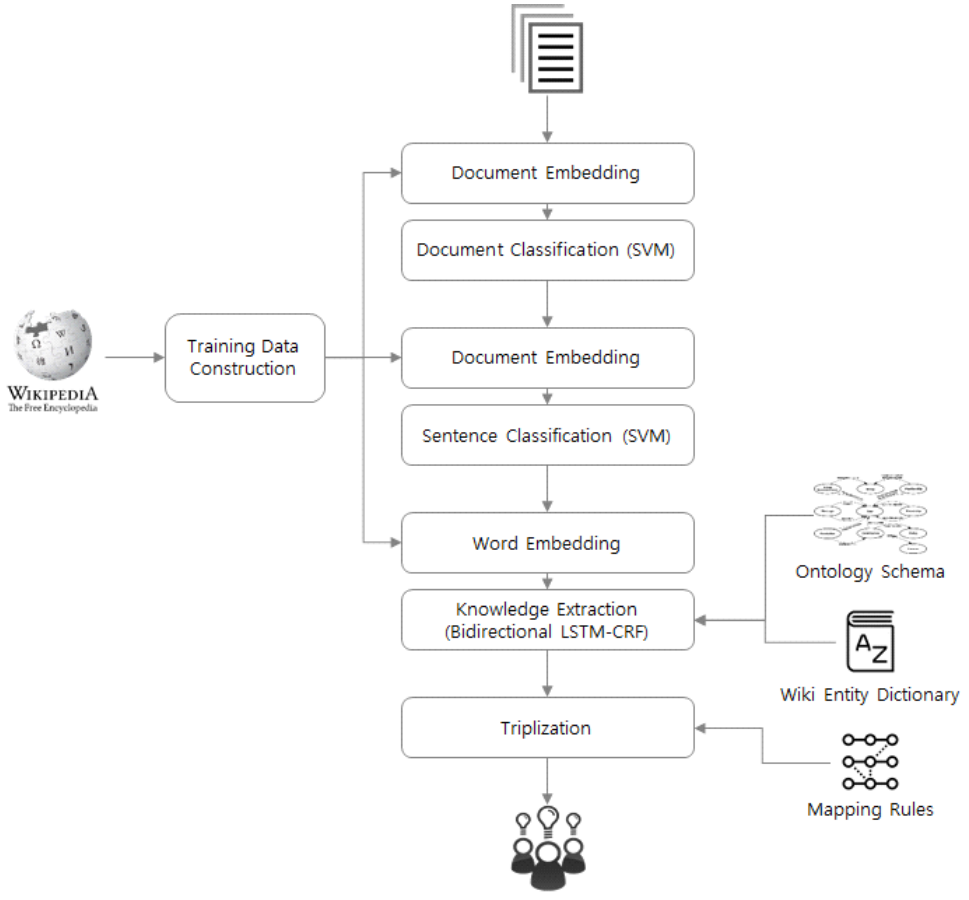
인포박스를 생성하는 또 다른 연구로 Bhuiyan et. al.(2017)은 인포박스를 생성하기에 충분한 내용을 포함하고 있지 않은 위키페이지에서 인포박스를 생성하기 위해 영화 도메인을 대상으로 IMDB[1] 및 프리베이스(Freebase)[2] 등으로부터 관련 정보를 수집하여 유사성 기반의 방법으로 인포박스를 생성하는 방법을 제안하고 있다. 많은 위키페이지들이 실제로 값을 추출할 만큼의 충분한 문장을 포함하고 있지 않기 때문에 충분히 고려되어야 할 문제점이기는 하나 도메인이 한정되어 있고 테이블 구조의 데이터 간 유사도를 기반으로 하기 때문에 제약사항이 존재한다.

위키피디아로부터 한국어 지식베이스를 구축하기 위한 연구로서 Jeong et. al.(2015)은 위키피디아로부터 YAGO(Suchanek, 2007) 형태의 지식베이스를 생성하기 위해 YAGO의 트리플을 한국어 DBpedia와 단순 비교하여 초기 지식을 생성한 후 지식의 확장을 위해 문장의 자질을 선정하고 ME(Maximum Entropy)(Berger, 1996)로 확장하는 방법을 통해 자동화된 지식베이스를 구축하였다. 하지만 휴리스틱한 방법을 이용했다 하더라도 주어와 목적어를 모두 포함하는 문장만을 선별하여 학습하거나 한 문장만을 자질 추출에 사용하는 것은 성능을 낮추는 요인이 될 수 있다.

3. 한국어 위키피디아의 학습 기반 지식추출

이번 장에서는 텍스트 형태의 문서로부터 지식을 추출하여 지식베이스를 구축하는 방법에 대해 설명하고자 한다. <Figure 1>은 이에 대한 전체 프로세스를 보여주고 있다. 지식추출을 위한 학습은 위키피디아의 인포박스를 이용하여 학습을 수행하며, 지식추출의 프로세스는 문서 분류, 문장 분류, 지식 추출, 마지막으로 트리플 형태의 지식으로 변환하는 절차로 이루어진다. 이러한 절차의 목적은 특정 자연어 문서에 나타날 수 있는 모든 지식을 추출하기 보다는 사전에 온톨로지의 구조에 따라 정의된 유형의 지식을 추출하기 위함이다. 따라서 문서를 분류하는 모델은 입력된 문서가 속할 클래스(class)를 정의하

1) https://www.imdb.com/
2) https://developers.google.com/freebase/



〈Figure 1〉 Knowledge Extraction Process

는 것과 동일하며, 문장 분류와 지식 추출 과정을 통해 온톨로지의 정의를 따르는 속성과 값을 추출하는 과정이라 볼 수 있다.

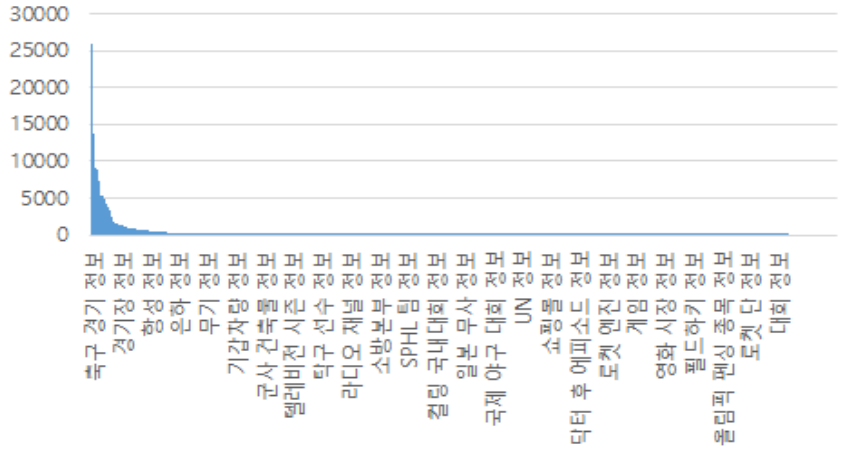
3.1 지식추출을 위한 대상 및 범위 선정

본 논문의 목적은 위키피디아의 인포박스를 학습하여 지식을 추출함으로써 지식베이스를 구축할 수 있는 방법을 제안하는 것이다. 이를 위해 이번 절에서는 우선 위키피디아의 인포박스 중 학습을 위한 범위를 결정한다.

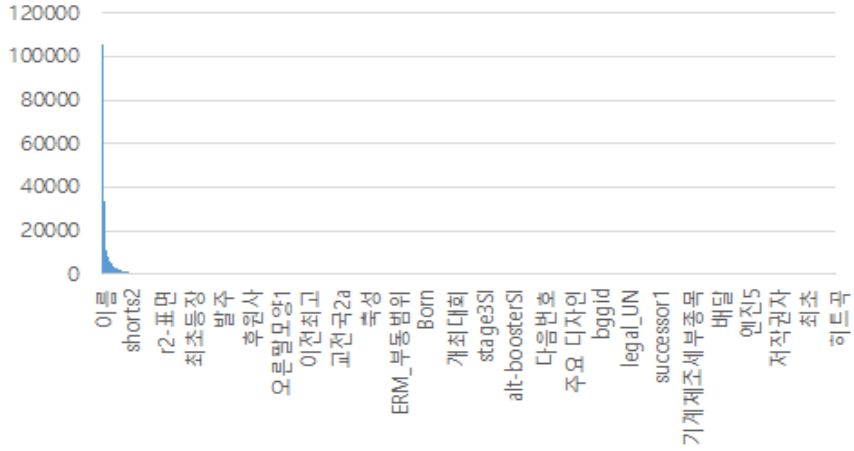
위키피디아의 위키페이지는 인포박스를 포함할 수 있으며, 인포박스는 해당 위키페이지의 분류 정보와 분류에 따른 속성과 값 정보를 담고 있다. 위키피디아는 이러한 분류를 사전에 정의해 놓고 해당 분류가 가질 수 있는 속성들을 템플릿이라는 이름으로 정의해 놓았다. 따라서 사용자는 인포박스 작성 시 작성할 텍스트의 분류를 결정하고 해당 분류의 템플릿을 찾아 정의된 속성에 적합한 값을 채우는 방식으로 작성한다.

하지만 영어 위키피디아의 경우 2,300개 이상의 템플릿이 정의[3]되어 있어 적합한 템플릿을 찾는 것이 쉽지 않으며, 한국어 위키피디아의 경우에는 템플릿에 대한 정의가 많지 않아 작성자가 임의로 작성한 인포박스가 많은 상황이다. <Figure 2>와 <Figure 3>은 2018년 4월 기준 한국어 위키피디아에 작성되어 있는 인포박스의 통계 그래프이다.

전체 인포박스 분류는 635개 였으며, 가장 많이 사용된 인포박스 분류는 축구 경기 정보로써 약 26,000건 정도 사용되었다. 속성 별로 살펴보면 만 개 이상의 속성이 인포박스에 사용되었으



〈Figure 2〉 Statistics of Infobox Categories



〈Figure 3〉 Statistics of Infobox Attributes

3) https://en.wikipedia.org/wiki/Category:Infobox_templates

며, 가장 많이 사용된 속성은 이름 속성으로써 약 100,000번 정도 사용되었다. 하지만 <Figure 2>와 <Figure 3>에서 볼 수 있는 것처럼 많은 인포박스의 분류와 속성이 한 번씩 선언되어 사용된 것들이 대부분을 차지한다. 이러한 분류와 속성의 경우 학습 데이터로서의 활용이 불가능하기 때문에 적절한 개수 이상 사용된 분류와 속성으로 범위를 제한하여야 한다. 일부 샘플을 이용하여 SVM 분류에 대한 선행 시험을 수행해 본 결과 최소 50개 이상에 대해 신뢰할 만한 성능을 보였으며, 이에 따라 본 논문에서는 분류 및 속성 모두 50번 이상 사용된 것들을 대상으로 학습을 수행하였다. 결국 약 200개의 분류와 약 2,500개의 속성을 학습하였다.

3.2 문서의 분류를 결정하기 위한 학습 모델

문서에 대한 분류 모델은 입력된 문서가 다루고 있는 주제를 분류하기 위한 모델로써, 지식베이스의 관점에서 본다면 해당 인스턴스가 속할 클래스를 결정하는 것이다. 온톨로지에서 특정 클래스가 가질 수 있는 속성이 정의되어 있는 것과 같이 문서의 분류를 결정함으로써 해당 문서에서 추출할 속성 또한 결정된다. 예를 들어, 인물에 대한 정보를 담고 있는 문서에서는 출생일에 대한 지식을 추출하기에 적합하지만 대학에 대한 정보를 담고 있는 문서에는 적합하지 않기 때문에 문서의 분류 과정을 통해 클래스에 대한 결정 및 그에 따른 추출 대상 속성을 결정한다.

위키피디아에서 하나의 위키페이지는 인포박스를 포함하여 본문, 각주, 외부링크 등 다양한 정보를 포함하고 있다. 문서 분류를 위한 학습 데이터는 모든 위키 문법을 제거한 본문 만을 사용하였으며, 본문의 명사 단위로 문서 임베딩(document embedding)(Dai et. al., 2014)을 이용한 벡터 값을 SVM(Support Vector Machine)(Hearst et. al., 1998)을 이용해 학습을 수행하였다.

3.3 적합 문장을 결정하기 위한 학습 모델

문서 분류 모델을 통해 문서의 주제 정보가 분류되면 다음 프로세스는 해당 주제에 속하는 각각의 속성(위키피디아의 템플릿에 의해 사전에 정의된 속성)에 대해 지식을 추출하기에 적합한 문장인지 아닌지를 결정하는 것이다. 이를 수행하기 위한 적합 문장 분류 모델은 각 분류의 속성 별로 존재하게 되며, 따라서 주제 분류의 속성 단위로 학습 데이터를 생성하고 학습을 수행한다.

적합 문장 분류 모델을 위한 학습 데이터를 생성하기 위해 각각의 위키페이지 본문을 문장 단위로 분리한 후 문장 단위로 문서 임베딩을 통해 학습용 벡터를 생성한다. 다음으로 각각의 문장이 인포박스의 특정 속성에 대해 값을 포함하고 있을 경우 “Good”으로 태깅하고, 그렇지 않을 경우에는 “Bad”로 태깅한다. 예를 들어, 국적 속성에 대한 적합 문장 분류 모델을 학습하기 위해 문재인 위키 페이지로부터 추출된 “문재인(文在寅, 1953년 1월 24일 ~)은 대한민국의 제19대 대통령이다”라는 문장은 문재인 대통령의 국적인 “대한민국”이라는 값을 포함하고 있으므로 “Good”으로 태깅한다. 다음의 “본관은 남평(南平)이다”라는 문장은 국적 속성의 값인 “대한민국”을 포함하지 않으므로 국적 속성에 대해 “Bad”로 태깅한다. 이와 같은 방법을 이용하여 각 속성 단위로 학습 데이터를 만들고 문서 분류 모델과 마찬가지로 SVM을 이용해 학습을 수행하였다.

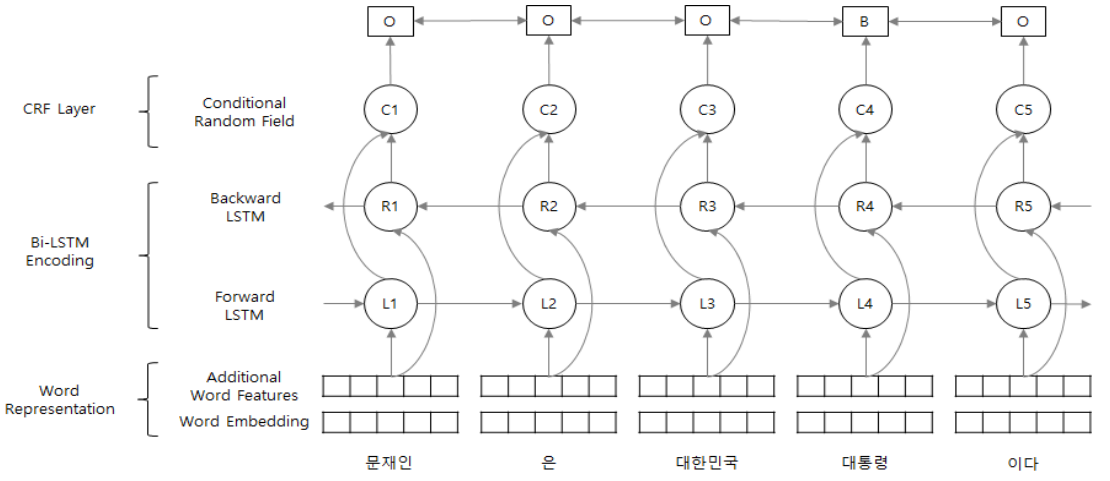
3.4 지식 추출을 위한 학습 모델

특정 속성에 대해 적합문장으로 분류된 문장은 지식을 추출하기 위해 지식 추출 모델의 입력으로 사용된다. 지식 추출 모델은 적합 문장 분류 모델과 마찬가지로 속성 단위로 학습을 수행하여 모델을 생성하게 된다. 이를 위한 학습 데이터는 인포박스에 있는 속성을 포함하는 문장에 대해 BIO(Inside-outside-beginning) 태깅을 사용하였다(Ramshaw and Marcus, 1995).

지식 추출을 위한 학습 방법은 비교 평가를 수행하기 위해 CRF와 Bidirectional LSTM-CRF 두 가지를 적용하였다. 우선 CRF는 순차적으로 입력되는 데이터에 대해 주변의 문맥(context)를 고려하는 학습 모델을 만들어 내는 통계 모델링 방법의 한 종류(Lafferty, 2001)로써, 본 연구에서는 적합 문장으로부터 특정 속성에 대한 값을 추출하기 위한 목적으로 사용된다. CRF의 입력은 적합 문장에 대해 형태소 분석을 수행한 후 출현빈도와 역출현빈도(TF/IDF)를 이용하여 특성 값을 부여한 벡터를 활용하였다.

그리고 기존 연구들에서 제안된 방법과의 비교평가를 위해 Bi-LSTM-CRF(Bidirectional Long Short-Term Memory-Conditional Random Field)를 이용하였다. Bi-LSTM-CRF은 RNN(Recurrent Neural Network)의 장기의존성 문제를 극복하기 위해 RNN의 셀(cell)에 LSTM을 적용하고 이를 양방향 쌍으로 엮은 후 CRF의 x와 y의 관계 함수를 Bidirectional LSTM으로 정의한 것을 의미한다(Huang et. al., 2015; Chiu and Nichols, 2016; Jin et. al., 2018). 최근의 딥러닝에 대한 실험을 통해 많은 문제에서 CRF보다 Bi-LSTM-CRF가 더 좋은 성능을 보인다고 증명된 바 있다(Wu et. al., 2017; Ljubešić, 2018). 이를 위해 본 논문에서는 <Figure 4>와 같이 Bi-LSTM을 통해 도출된 양방향의 연속된 특징을 CRF에 입력으로 사용하여 최적의 태그 열 결과값을 도출하였다.

Bi-LSTM-CRF의 입력은 형태소 분석을 통해 출력된 형태소 단위의 단어를 워드 임베딩을 수행한 후 그 벡터 값을 순서대로 연결해 입력으로 사용하였다. 양방향 LSTM의 입력 값으로 사용



〈Figure 4〉 Bi-LSTM-CRF for Knowledge Extraction

된 단어 벡터는 30으로 지정하였다. 이는 앞서 언급한 CRF 방법을 적용하여 실험을 수행한 후 실험을 통해 도출된 정답 문장의 길이들로부터 통계적 변곡점이 두드러지는 지점의 값을 단어 벡터의 길이로 지정한 것이다. 학습 최적화를 위해 Adam Optimizer(Kingma and Ba, 2015)를 이용하였으며, Viterbi 알고리즘(Viterbi, 1967)을 사용하여 최적의 태그열 결과값을 도출하였다.

도출된 값은 온톨로지 구조를 반영하여 최종적인 결과값의 대상이 된다. 온톨로지는 그 속성의 타입에 따라 목적어 값의 유형이 결정된다. DatatypeProperty일 경우에는 문자열, 숫자 값 등 일반적인 텍스트 값이 위치하게 되며, ObjectProperty일 경우에는 URI 값을 목적어로 갖는다. 따라서 값을 추출하기 위해 사용된 모델이 어떤 속성을 위한 것인지를 온톨로지를 통해 확인하고 온톨로지에 선언된 유형에 따라 값의 타입을 결정하게 된다. 이 때 URI 값으로 생성하기 위해서는 이에 대응되는 적절한 자원(resource)이 존재하여야 하기 때문에 위키피디아의 각 페이지를 이용하여 개체 사전을 만들고 추출된 값에 대한 개체 사전의 존재유무를 검사한다. 이를 이용하여 속성이 ObjectProperty이고 추출된 값이 개체 사전에 존재할 경우 URI 형태의 개체 값을 출력을 제공한다.

또한 인포박스의 속성에 따라 출생일과 같이 하나의 값을 갖는 속성과 직업과 같이 두 개 이상의 값을 갖는 속성이 존재한다. 따라서 전체 속성에 대해 값의 빈도를 조사한 후 그 유형에 따라 최종적으로 제시할 후보의 개수를 결정한다. 최종적으로 제시되는 결과값에 대한 신뢰값은 적합 문장 분류 모델의 점수와 지식 추출 모델의 점수를 곱한 값을 사용한다. 이는 지식 추출 모델뿐만 아니라 적합 문장 분류 모델이 정답을 예측하는데 있어 중요한 영향을 미치기 때문이다. 이와 같은 과정을 통해 특정 속성에 대해 주어진 적합 문장으로부터 해당 속성의 값을 추출하게 된다.

3.5 지식베이스 구축

위키피디아 학습을 이용한 지식베이스 구축의 마지막 프로세스는 앞서의 과정을 통해 추출된 지식을 온톨로지화 시키는 것이다. 다시 말해, 문서 분류, 적합 문장 분류, 지식 추출의 예측 과정을 통해 추출된 지식을 RDF(Resource Description Framework) 트리플의 형태로 변환하는 것이다. 이때 입력된 문서의 제목 혹은 URL이 트리플의 주어가 되고, 속성은 동사, 그리고 값은 목적어로 변환된다.

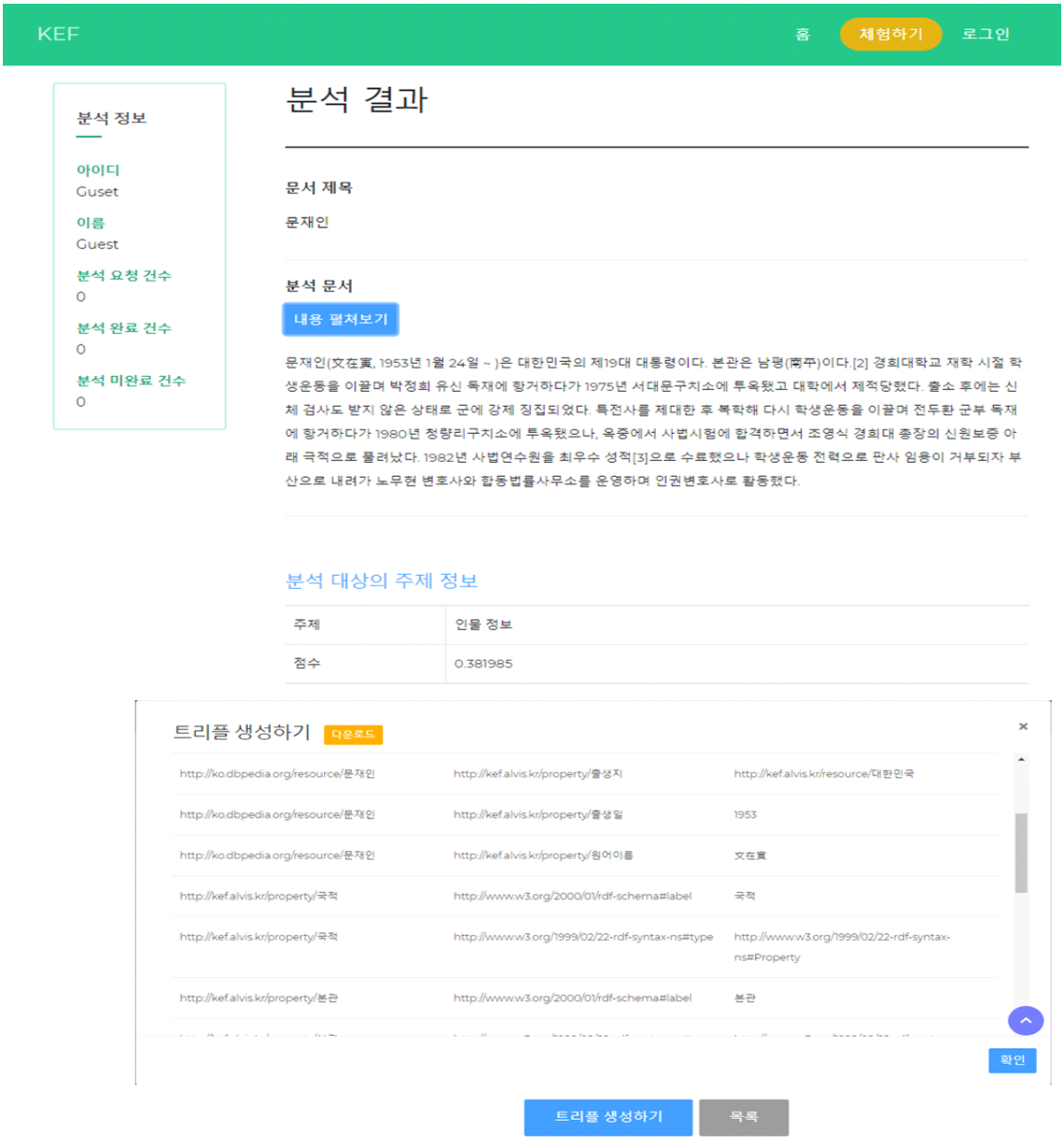
이러한 예측 프로세스를 통해 추출된 지식을 온톨로지 구조에 따라 적합한 형태로 변환하기 위한 방법으로 맵핑 규칙을 이용한다. DBpedia는 DBpedia Extraction Framework를 이용하여 인포박스에 존재하는 데이터를 트리플 형태로 변환하며(Lehmann et. al., 2015), 본 논문에서도 이와 동일하게 맵핑 규칙을 이용하여 트리플로의 변환을 수행하였다. 다만 맵핑 규칙에 정의된 것들보다 본 논문을 위해 학습한 분류와 속성이 많아 맵핑 규칙을 추가적으로 확장하여 적용하였다.

3.6 학습 기반 지식 추출 플랫폼

앞서 제시한 방법론에 따라 분석 기능은 파이썬(Python)으로 구현하였으며, 웹 인터페이스는 자바(Java)를 이용하여 웹 기반의 플랫폼 형태로 구현하였다. 사용자가 분석 대상의 제목과 텍스트를 입력하고 분석 요청 버튼을 누르면 해당 텍

스트에 대한 주제 분류를 수행한 후 해당 주제에 속한 적합 문장 분류 모델과 지식 추출 모델을 수행하여 트리플 형태의 지식을 추출하게 된다.

이러한 방법론에 따라 구현된 결과물은 <Figure 5>와 같다. <Figure 5>는 문재인 대통령에 대한 텍스트를 입력으로 수행한 결과이며, 해당 텍스트



〈Figure 5〉 Knowledge Extraction Web Service

의 주제를 인물 정보로 분류하였다. 또한 해당 텍스트로부터 인물 정보에 해당하는 지식을 추출하여 트리플 형태로 출력하고 있다.

4. 실험 결과

이번 장에서는 앞서 제안한 방법론에 따라 지식 추출에 대한 평가를 수행하고자 한다. 평가를 수행하기 위해 관련 연구(Wu and Weld, 2007; Lange et. al., 2010; Brandão et. al., 2010)와 동일한 인포박스 분류를 대상으로 실험을 진행하였다. 다만 국가, 항공사, 배우, 대학 분류 중 국가 정보의 경우 실험할 만큼의 충분한 학습 데이터가 한국어 위키피디아에 존재하지 않아 항공사, 배우, 대학 세 개의 분류에 대해 실험을 수행하였다. <Table 1>과 같이 각각의 분류에 대해 5개씩의 속성을 대상으로 전체 데이터 중 80%를 학습 데이터로, 20%를 테스트 데이터로 하여 시험을 수행하였다.

지식 추출을 위해 CRF와 Bi-LSTM-CRF를 적용한 비교 실험 결과는 <Table 2>와 같다. 실험 결과를 살펴보면 몇몇 개를 제외하고는 CRF에 비해 Bi-LSTM-CRF의 적용 결과가 더 나은 성능을 보인 것을 알 수 있다. 제시된 결과 중 대학 정보 분류의 설립 속성의 값이 다른 속성들에 비해 유독 낮은 것을 확인할 수 있다. 위키피디아에서 대학 정보에 존재하는 설립 속성의 경우 해

〈Table 1〉 Number of Data for Training and Testing

Category	Attribute	Number of Train Data	Number of Test Data
항공사 정보	항공사	731	183
	창립일	601	151
	허브공항	384	96
	본사	457	115
	모기업	255	64
대학 정보	종류	744	186
	설립	909	228
	학교법인	155	39
	국가	959	240
	위치	956	239
영화인 정보	본명	2,202	551
	출생일	13,666	3,417
	출생지	6,724	1,681
	국적	10,222	2,556
	직업	13,237	3,310

〈Table 2〉 Experimental Results of CRF and Bi-LSTM-CRF

Category	Attribute	CRF			Bi-LSTM-CRF		
		Precision	Recall	F1	Precision	Recall	F1
항공사 정보	항공사	0.7980	0.9454	**0.8429**	0.7574	0.9672	0.8147
	창립일	0.8918	0.9868	0.9227	0.9029	0.9931	**0.9326**
	허브공항	0.9375	0.9479	0.9410	0.9743	1.0000	**0.9809**
	본사	0.5102	0.9652	0.6364	0.6477	0.9545	**0.7379**
	모기업	0.5938	0.6094	0.5990	0.6797	0.7031	**0.6875**
대학 정보	종류	0.8235	0.9194	**0.8548**	0.7773	0.8280	0.7934
	설립	0.7405	0.7939	**0.7572**	0.3039	0.3256	0.3109
	학교법인	0.7073	0.8205	0.7410	0.8846	0.9744	**0.9137**
	국가	0.8874	0.9375	0.9015	0.8812	0.9583	**0.9036**
	위치	0.3754	0.7350	0.4640	0.6716	0.8987	**0.7378**
영화인 정보	본명	0.6543	0.6724	0.6594	0.6679	0.7536	**0.6962**
	출생일	0.5036	0.9871	0.6636	0.9463	1.0000	**0.9655**
	출생지	0.7063	0.8252	0.7087	0.7660	0.9412	**0.8215**
	국적	0.8048	0.9789	0.8523	0.8741	0.9956	**0.9083**
	직업	0.8935	0.8856	**0.8760**	0.8455	0.9050	0.8614

당 대학이 현재까지 설립되어 온 연혁에 대한 정보를 <Figure 6>와 같이 담고 있다. 이는 본 논문에서 제안하는 방법론이 텍스트로부터 값을 추출하기 때문에 연혁과 같은 정보를 정확히 추출하기가 어려운 구조를 가지고 있으며, 또한 평가 시 문자열의 일치 여부를 이용하여 평가를 수행하였기 때문에 글자 단위의 차이가 낮은 성능의 원인이 되었을 것으로 판단된다. 향후 이러한 유형의 정보를 추출하기 위한 방법과 평가 방법의 개선이 필요할 것으로 보인다.

<Table 3>은 학습한 모든 속성을 대상으로 속성의 값 유형에 따른 평가 결과를 보여준다. 단순한 숫자 유형의 값에 대해 가장 좋은 성능을

〈Figure 6〉 Establish Attribute of University Category

〈Table 3〉 Experimental Results by Value Type

Value Type	Korean	Number	English	Chinese	Date	Address
Precision	0.80	0.94	0.70	0.86	0.74	0.69

보이고 있으며, 추출하여야 될 값이 길거나 복잡한 경우에 낮은 성능을 보임을 알 수 있다. 이는 앞서 언급한 것과 같이 학습 및 평가 방법 두 가지 측면의 요인이 기인한 것으로 판단된다.

5. 결론 및 향후 연구

인공지능 기술의 발전과 함께 최근 지식베이스에 대한 필요성과 중요성이 점차 높아지고 있지만 지식베이스를 구축하는 것은 사람의 많은 노력과 시간을 필요로 하는 작업이다. 이러한 문제를 해결하기 위해 본 논문에서는 기계학습을 이용해 자연어 텍스트로부터 지식을 추출하여 지식베이스를 구축 및 확장해 나가는 방법을 제안하였다. 이를 위해 위키피디아의 인포박스를 이용하여 학습 데이터를 만들고 추출될 지식의 유형을 결정하기 위한 문서 분류, 지식 추출에 적합한 문장을 고르기 위한 적합 문장 분류, 적합한 문장으로 분류된 문장을 대상으로 실제 지식을 추출하는 과정 및 지식베이스 구조에 따른 검증 과정으로 지식 추출 방법을 제안하였다. 추출된 지식은 마지막으로 RDF 형태의 트리플 구조로 변환하는 작업을 거쳐 최종적인 지식베이스 형태로 만들어진다. 이러한 과정은 학습된 모델을 이용하여 단순히 문서로부터 적합한 값을 추출하는 것이 아니라, 지식베이스의 구조를 고려하여 그 구조에 따른 값을 추출하는 방법으로 구성된다. 이를 통해 자연어 문서로부터 지식을 추출해 낼 수 있으며, 실험을 통해 본 논문에서 제안하는 방법이 효과적으로 지식을 추출할 수 있음을 증명하였다. 또한 실제의 서비스를 통해 충분히 유용하게 활용될 수 있음 보였다. 이러한 방법을 이용함으로써 지식베이스의 구조에 따라 인스턴스를 확장해 나가는데 필요한 사람의 노력을 현저히 줄일 수 있으며, 보다 빠른 지식베이스의 구축이 가능할 것으로 판단된다. 또한 구축된 지식베이스는 최근 인공지능 스피커 등 다양한 분야에서 사용자와의 질의응답(Question Answering)을 위한 기반 지식으로써 사용자의 질의에 대해 적합한 응답을 찾고 제시하기 위한 목적으로 활용될 수 있다. 통상적으로 질의응답에 활용될 수 있는 지식을 구축하기 위해 웹에서 데이터를 수집하여 구조적으로 변환 및 저장한 후 사용자의 자연어 질의가 입력되면 이를 지식베이스에 따라 해석하고 답을 찾아 제시하게 된다. 이러한 과정에 있어서 제안된 방법을 활용할 경우 주어진 지식베이스의 구조에 따라 자연어 문서로부터 지식을 생성할 수 있으므로 보다 효과적인 질의응답 시스템의 구축이 가능할 것으로 보인다.

본 논문에서 제안한 방법론을 보다 발전시키기 위해서는 몇 가지 지속적인 향후 연구가 필요하다. 첫 번째는 일반적인 문서를 대상으로 학습을 확장해 나가는 것이다. 현재 학습을 수행한 대상이 위키피디아 문서이기 때문에 위키피디아

나 네이버 백과사전, 나무위키 등과 같이 무엇인가에 대한 정의를 담고 있는 문서에서는 만족할 만한 성능을 보이나 뉴스나 블로그와 같은 일상적으로 작성될 수 있는 문서에서는 좋은 성능을 보이지 않는다. 이러한 문제를 해결하기 위해서는 위키피디아의 인포박스를 이용하여 뉴스나 블로그와 같은 글도 학습할 수 있도록 학습 데이터를 만들어야 한다. 다만 이 과정 또한 사람의 시간과 노력이 필요하기 때문에 이를 자동화 할 수 있도록 단순 문자열 매칭 뿐만 아니라 본 논문에서 제안한 방법론을 이용하여 재검토를 수행하는 등 학습 데이터 생성의 자동화에 대한 연구를 수행하고 있다. 두 번째는 지식 추출의 속성을 확장하는 것이다. 앞서 언급한 것과 같이 유효한 학습 데이터를 확보하기 위해 출현 빈도가 50번 이상인 것만을 대상으로 학습을 수행하였는데 이를 보다 확장하여 추출될 수 있는 지식의 유형을 늘려나가는 것이다. 마지막 향후 연구 주제로써 학습 및 예측 프로세스를 보다 단순화하는 것이다. 본 논문에서 제안한 절차는 문서 분류, 적합 문장 분류, 지식 추출 및 검증 세 단계의 학습 및 예측 프로세스로 구성되어 있다. 또한 지식 추출의 정확도를 높이기 위해 분류 별 속성 단위로 모델이 존재하는 형태를 가지고 있다. 이러한 프로세스나 모델을 통합하거나 제거하는 등 다양한 방법을 통해 단순화할 수 있는 연구를 진행해 나갈 계획이다.

참고문헌(References)

Berger, A. L., V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, Vol.22, No.1(1996), 39~71.

Bergman, M., *Knowledge-based Artificial Intelligence*, AI3, 2014. Available at http://www.mkbergman.com/1816/knowledge-based-artificial-intelligence/ (Accessed 13 November, 2018).

Bhuiyan, H., K. J. Oh, M. D. Hong, and G. S. Jo, "An effective approach to generate Wikipedia infobox of movie domain using semi-structured data," Journal of Internet Computing and Services, Vol.18, No.3(2017), 49~61.

Bizer, C., T. Heath, K. Idehen, and T. Berners-Lee, "Linked Data on the Web (LDOW2008)," *Workshop at the 17th International World Wide Web Conference*, (2008).

Bizer, C., J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia - A Crystallization Point for the Web of Data," *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 7, No. 3(2009), 154~165.

Brandão, W. C., E. S. Moura, A. S. Silva, and N. Ziviani, "A Self-Supervised Approach for Extraction of Attribute-Value Pairs from Wikipedia Articles," *Proceedings of the 17th international conference on String processing and information retrieval*, (2010), 279~289.

Chiu, J. and E. Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs," *Transactions of the Association for Computational Linguistics*, Vol. 4, No. 1(2016), 357~370.

Choi, H., M. Kim, W. Kim, D. Shin, and Y. H. Lee, “Development of Information Extraction System from Multi Source Unstructured Documents for Knowledge Base Expansion,” Journal of Intelligence and Information Systems, Vol. 24, No. 4(2018), 111~136.

Dai, A. M., C. Olah, and Q. V. Le, "Document Embedding with Paragraph Vectors," *NIPS Deep Learning Workshop*, (2014).

Engelmore, R. S., "Artificial Intelligence and Knowledge Based Systems: Origins, Methods and Opportunities for NDE," *Review of Progress in Quantitative Nondestructive Evaluation*, Springer Science, New York, 1987.

Forsythe, D. E., "Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence," Social Studies of Science, Vol.23, No.3(1993), 445~477.

Hearst, M. A., S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, Vol.13, No.4(1998), 18~28.

Higashinaka, R., K. Dohsaka, and H. Isozaki, "Learning to rank definitions to generate quizzes for interactive information presentation," *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, (2007), 117~120.

Huang, Z., W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *arXiv.org preprint*, 2015. Available at https://arxiv.org/pdf/1508.01991.pdf (Downloaded 15 November, 2018).

Jeong, S., M. Choi, and H. Kim, "Construction of Korean Knowledge Base Based on Machine Learning from Wikipedia," *Journal of KIISE*, Vol. 42, No. 8(2015), 1065-1070.

Jin, S., H. Jang, and W. Kim, "Improving Bidirectional LSTM-CRF model Of Sequence Tagging by using Ontology knowledge based feature," Journal of intelligence and information systems, Vol.24, No.1(2018), 253~266.

Kaisser, M., "The qualim question answering demo: Supplementing answers with paragraphs drawn from wikipedia," *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, (2008), 32~35.

Kingma, D. and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of the 3rd International Conference for Learning Representations*, (2015).

Krishna, S, *Introduction to Database and Knowledge-base Systems*, World Scientific Publishing, Singapore, 1992.

Lafferty, J., A. McCallum, and F. C.N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *Proceedings of the Eighteenth International Conference on Machine Learning*, (2001), 282~289.

Lange, D., C. Böhm, and F. Naumann, "Extracting structured information from Wikipedia articles to populate infoboxes," *Proceedings of the 19th ACM international conference on Information and knowledge management*, (2010), 1661-1664.

Lehmann, J. R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer, "DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web*, Vol.6, No.2(2015), 167~195.

Ljubešić, N., "Comparing CRF and LSTM performance on the task of morphosyntactic tagging of non-standard varieties of South Slavic languages," *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects*, (2018), 156–163.

Ramshaw, L. A. and M. P. Marcus, "Text

Chunking using Transformation-Based Learning," *ACL Third Workshop on Very Large Corpora*, (1995), 82~94.

Russell, S. J., and P. Norvig, *Artificial Intelligence : A Modern Approach*, Prentice Hall, 2009.

Suchanek, F. M., G. Kasneci, and G. Weikum, "Yago:a core of semantic knowledge," *Proceedings of the 16th international conference on World Wide Web*, (2007), 697~706.

Sun, R., Artificial intelligence: Connectionist and symbolic approaches, In: N. J. Smelser and P. B. Baltes (eds.), *International Encyclopedia of the Social and Behavioral Sciences*, Pergamon/Elsevier, Oxford, 2001.

Viterbi, A. J., "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," IEEE Transactions on Information Theory, Vol.13, No.2(1967), 260~269.

Wu, F. and D.S. Weld, “Autonomously semantifying Wikipedia,” *Proceedings of the sixteenth ACM conference on Conference on Information and knowledge management*, (2007), 41-50.

Wu, J., X. Hu, R. Zhao, F. Ren, and M. Hu, "Clinical Named Entity Recognition via Bi-directional LSTM-CRF Model," *Proceedings of the Evaluation Task at the China Conference on Knowledge Graph and Semantic Computing*, (2017), 31~36.

Abstract

Knowledge Extraction Methodology and Framework from Wikipedia Articles for Construction of Knowledge-Base

JaeHun Kim* · Myungjin Lee**

Development of technologies in artificial intelligence has been rapidly increasing with the Fourth Industrial Revolution, and researches related to AI have been actively conducted in a variety of fields such as autonomous vehicles, natural language processing, and robotics. These researches have been focused on solving cognitive problems such as learning and problem solving related to human intelligence from the 1950s. The field of artificial intelligence has achieved more technological advance than ever, due to recent interest in technology and research on various algorithms. The knowledge-based system is a sub-domain of artificial intelligence, and it aims to enable artificial intelligence agents to make decisions by using machine-readable and processible knowledge constructed from complex and informal human knowledge and rules in various fields. A knowledge base is used to optimize information collection, organization, and retrieval, and recently it is used with statistical artificial intelligence such as machine learning. Recently, the purpose of the knowledge base is to express, publish, and share knowledge on the web by describing and connecting web resources such as pages and data. These knowledge bases are used for intelligent processing in various fields of artificial intelligence such as question answering system of the smart speaker. However, building a useful knowledge base is a time-consuming task and still requires a lot of effort of the experts. In recent years, many kinds of research and technologies of knowledge based artificial intelligence use DBpedia that is one of the biggest knowledge base aiming to extract structured content from the various information of Wikipedia. DBpedia contains various information extracted from Wikipedia such as a title, categories, and links, but the most useful knowledge is from infobox of Wikipedia that presents a summary of some unifying aspect created by users. These knowledge are created by the mapping rule between infobox structures and DBpedia ontology schema defined in DBpedia Extraction Framework.

* Research Laboratory, LiST
** Corresponding Author: Myungjin Lee
Research Laboratory, LiST
3, Beodeunaru-ro 19-gil, Yeongdeungpo-gu, Seoul 07229, Republic of Korea
Tel: +82-2-2632-5133, Fax: +82-2-2632-5134, E-mail: mjlee@li-st.com

In this way, DBpedia can expect high reliability in terms of accuracy of knowledge by using the method of generating knowledge from semi-structured infobox data created by users. However, since only about 50% of all wiki pages contain infobox in Korean Wikipedia, DBpedia has limitations in term of knowledge scalability. This paper proposes a method to extract knowledge from text documents according to the ontology schema using machine learning. In order to demonstrate the appropriateness of this method, we explain a knowledge extraction model according to the DBpedia ontology schema by learning Wikipedia infoboxes. Our knowledge extraction model consists of three steps, document classification as ontology classes, proper sentence classification to extract triples, and value selection and transformation into RDF triple structure. The structure of Wikipedia infobox are defined as infobox templates that provide standardized information across related articles, and DBpedia ontology schema can be mapped these infobox templates. Based on these mapping relations, we classify the input document according to infobox categories which means ontology classes. After determining the classification of the input document, we classify the appropriate sentence according to attributes belonging to the classification. Finally, we extract knowledge from sentences that are classified as appropriate, and we convert knowledge into a form of triples. In order to train models, we generated training data set from Wikipedia dump using a method to add BIO tags to sentences, so we trained about 200 classes and about 2,500 relations for extracting knowledge. Furthermore, we evaluated comparative experiments of CRF and Bi-LSTM-CRF for the knowledge extraction process. Through this proposed process, it is possible to utilize structured knowledge by extracting knowledge according to the ontology schema from text documents. In addition, this methodology can significantly reduce the effort of the experts to construct instances according to the ontology schema.

Key Words : Deep learning, Artificial Intelligence, Ontology, Knowledge base, Knowledge extraction

Received : December 18, 2018 Revised : March 5, 2019 Accepted : March 11, 2019
Publication Type : Conference(Fast-track) Corresponding Author : Myungjin Lee

저 자 소 개

김 재 헌
가톨릭대학교 컴퓨터공학과에서 학사 학위를 취득한 후 현재 주식회사 리스트의 기술연구소에서 연구원으로 근무 중에 있다. 딥러닝과 관련된 다양한 연구에 관심을 가지고 있으며, 관련 기술 및 서비스를 연구 개발하고 있다.

이 명 진
연세대학교 정보산업공학과에서 박사 학위를 취득한 후 현재 주식회사 리스트에서 CTO로 근문 중에 있다. 최근 지식베이스 기반의 질의응답 시스템과 같이 지식베이스를 활용한 인공지능과 딥러닝에 대한 연구 개발을 수행하고 있으며, 이와 관련된 50편 이상 다수의 서적과 논문을 출판하였다.