ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2018 December: 24(4): 111~136 http://dx.doi.org/10.13088/jiis.2018.24.4.111

지식베이스 확장을 위한 멀티소스 비정형 문서에서의 정보 추출 시스템의 개발*

최현승
연세대학교 산업공학과
(*gustmd820@ajou.ac.kr*)

김민태
연세대학교 산업공학과
(*iammt@yonsei.ac.kr*)

김우주
연세대학교 산업공학과
(*wkim@yonsei.ac.kr*)

신동욱
SK텔레콤 지식기술 Cell
(*dwshin@sk.com*)

이용훈
SK텔레콤 지식기술 Cell
(*Yhlee@sk.com*)

지식베이스를 구축하는 작업은 도메인 전문가가 온톨로지 스키마를 이해한 뒤, 직접 지식을 정제하는 수작업이 요구되는 만큼 비용이 많이 드는 활동이다. 이에, 도메인 전문가 없이 다양한 웹 환경으로부터 질의에 대한 답변 정보를 추출하기 위한 자동화된 시스템의 연구개발의 필요성이 제기되고 있다. 기존의 정보 추출 관련 연구들은 웹에 존재하는 다양한 형태의 문서 중 학습데이터와 상이한 형태의 문서에서는 정보를 효과적으로 추출하기 어렵다는 한계점이 존재한다. 또한, 기계 독해와 관련된 연구들은 문서에 정답이 있는 경우를 가정하고 질의에 대한 답변정보를 추출하는 경우로서, 문서의 정답포함 여부를 보장할 수 없는 실제 웹의 비정형 문서로부터의 정보추출에서는 낮은 성능을 보인다는 한계점이 존재한다. 본 연구에서는 지식베이스 확장을 위하여 웹에 존재하는 멀티소스 비정형 문서로부터 질의에 대한 정보를 추출하기 위한 시스템의 개발 방법론을 제안하고자 한다. 본 연구에서 제안한 방법론은 "주어(Subject)-서술어(Predicate)"로 구분된 질의에 대하여 위키피디아, 네이버 백과사전, 네이버 뉴스 3개 웹 소스로부터 수집된 비정형 문서로부터 관련 정보를 추출하며, 제안된 방법론을 적용한 시스템의 성능평가를 위하여, Wu and Weld(2007)의 모델을 베이스라인 모델로 선정하여 성능을 비교분석 하였다. 연구결과 제안된 모델이 베이스라인 모델에 비해, 위키피디아, 네이버 백과사전, 네이버 뉴스 등 다양한 형태의 문서에서 정보를 효과적으로 추출하는 강건한 모델임을 입증하였다. 본 연구의 결과는 현업 지식베이스 관리자에게 지식베이스 확장을 위한 웹에서 질의에 대한 답변정보를 추출하기 위한 시스템 개발의 지침서로서 실무적인 시사점을 제공함과 동시에, 추후 다양한 형태의 질의응답 시스템 및 정보추출 연구로의 확장에 기여할 수 있을 것으로 기대한다.

주제어 : 정보추출, 질의응답 시스템, 기계독해, Bi-directional LSTM-CRF, 지식베이스

논문접수일 : 2018년 10월 29일 논문수정일 : 2018년 12월 17일 게재확정일 : 2018년 12월 18일
원고유형 : 학술대회(급행) 교신저자 : 김우주

1. 서론

지식베이스는 질의응답 시스템에 사용되는 요소로서, 사용자의 질의에 대한 답변지식을 저장, 탐색하는 기능을 수행하기 위한 기술로 인공지능 분야의 중요한 연구과제로 여겨지고 있다. 이

* 본 연구는 SK Telecom에서 지원을 받아 수행한 연구임.

러한 지식베이스를 구축하는 작업은 특정 도메인 전문가가 직접 지식을 정제하는 수작업이 요구되므로 비용이 많이 드는 활동이며(Khot et al., 2017), 도메인에 독립적인 지식베이스를 구축하는 작업 또한 검색엔진에 질의를 통해 얻어진 문서로부터 사람이 후보군을 수작업으로 탐색을 하고, 직접 추출 결과의 신뢰성을 고려해야 하는 노력이 요구된다(Etzioni, 2004). 따라서 사람의 수작업 노력 없이 위키피디아와 같은 웹의 다양한 비정형문서로부터 정보를 추출(Information Extraction; IE)하기 위한 자동화된 시스템 연구 개발의 필요성이 제기되고 있다.

기존의 정보 추출 관련 선행 연구들은 정보 추출규칙을 생성하여 규칙에 맞는 패턴이 발견되면 지식을 추출하거나, 훈련 문서와 라벨데이터를 생성한 뒤 모델을 학습하여, 학습된 모델 기반의 정보를 추출하는 방법을 주로 연구하였다. 이러한 연구들은 특정 형태의 패턴만을 정보추출의 대상으로 하여 새로운 형태의 정보 추출에는 적합하지 않다는 한계점 존재하거나(규칙 기반 정보추출 연구), 훈련 데이터와 상이한 형태의 문서에서는 정보를 효과적으로 추출하기 어렵다(학습 모델 기반 정보추출 연구)는 한계점이 존재한다. 그러나 지식베이스 확장을 위한 정보추출 문제는 실제 웹에 존재하는 멀티 소스로 부터 수집된 다양한 형태의 문서를 정보 추출의 대상으로 하기 때문에, 훈련 데이터와 이질적인 문서 형태에 대해서도 효과적으로 정보를 추출할 필요성이 존재한다.

또한 기계 독해(Machine Reading Comprehension; MRC) 관련 연구들은 문서에 정답이 있는 경우를 가정하고 질의에 대한 답변 정보를 추출하는 경우로서, 문서의 정답 포함 여부를 보장 할 수 없는 실제 웹이 비정형 문서에서의 정보추출에서는 낮은 성능을 보인다는 한계점이 존재한다. 지식베이스 확장을 위한 정보추출 문제는 실제 웹에 존재하는 문서를 정보 추출의 대상으로 하기 때문에, 문서의 정답 포함 여부를 보장할 수 없으며 문서, 문장, 지식 추출기의 종합적인 신뢰성이 함께 고려되어야 한다.

본 연구에서는 지식베이스 확장을 위하여 웹에 존재하는 멀티소스 비정형 문서로부터 질의에 대한 정보를 추출하기 위한 시스템의 개발 방법론을 제안한다. 본 연구에서 제안한 방법론은 “주어(Subject)-서술어(Predicate)”로 구분된 질의에 대하여, 위키피디아, 네이버 백과사전, 네이버 뉴스 3개 소스로부터 수집된 다양한 형태의 비정형 문서에서 관련 정보를 추출하며, Wu and Weld(2007)의 모델을 베이스라인 모델로 선정하여 성능을 비교분석 하였다.

본 연구는 다음과 같이 구성한다. 2장에서는 본 연구와 관련된 정보추출 및 기계독해 관련 연구들에 대해서 소개한다. 3장에서는 본 논문에서 제시하는 연구 방법론을 자세히 기술하고, 4장에서는 제안 모델의 성능을 평가하고 그 결과에 대해 논의한다. 마지막으로 5장은 본 논문의 결론과 연구의 한계점, 향후 연구 방향에 대해 논의한다.

2. 관련연구

2.1 정보추출

정보추출은 자연어 처리의 세부 영역으로 주어진 사용자 질의에 대하여 대량의 정형 혹은 비정형 문서로부터 구조화된 정보를 자동적으로 추출하는 기법을 의미한다(Gaizauska and Wilk,

1998; Line Eikvil, 1999; Qiu et al., 2018). 예컨대, (“피사의 사탑”, “높이”)라는 사용자 질의에 대해 위키피디아 피사의 사탑 문서의 “이탈리아의 천재 건축가 보나노 피사노가 설계한 피사의 사탑은 높이 55.8m, 지름 16m의 종탑입니다.”라는 문장으로부터 (“피사의 사탑”, “높이”, “55.8m”)라는 트리플을 추출 할 수 있다. 사용자 질의에 대하여 추출된 정보는 추후 질의응답에 활용될 지식베이스에 저장되거나, 사용자에게 답변으로 직접 제공될 수 있다.

정보추출은 정보추출의 방법론 및 문서 형태에 따라 구분 할 수 있다(Line Eikvil, 1999). 정보추출 방법론에 따른 구분: 1) 지식 공학적 접근방법(Knowledge engineering approach), 2) 자동화된 훈련 접근방법(Automatic training approach). 지식공학적 접근방법은 도메인 지식에 기반한 문법적 규칙으로서 정보 추출 패턴을 정의하고, 패턴에 맞는 문장이 발견되는 정보를 추출하는 방식을 의미하며, 자동화된 훈련 접근방법은 훈련 문서와 라벨 데이터를 생성한 뒤, 학습된 모델을 활용해 정보를 추출하는 방식을 의미한다. 지식공학적 접근 방법론은 기존에 알려진 패턴의 정보를 추출하기에는 적합하지만, 알려지지 않거나 새로운 형태의 정보를 추출하기에는 어렵다는 한계점 및 도메인 전문가(지식 공학자)의 노력이 요구되기 때문에, 주로 자동화된 훈련 접근방법론 기반의 정보추출 연구가 주로 이루어져 왔다(Banko et al., 2007).

문서 형태에 따른 구분: 1) 비정형 문서(Free text), 2) 정형 문서(Structured text), 3) 반정형문서(Semi-structured text). 비정형 문서는 정보 추출 분야의 주요 대상으로서, 자연어 처리기술 및 규칙기반의 시스템을 설계하여 정보를 추출한다. 정형문서의 경우는 미리 정의되어 있는 구조화된 포맷의 문서 혹은 데이터베이스를 의미하며, 상대적으로 간단한 기술을 통해 정보가 추출된다. 반정형 문서는 HTML 이나 테이블 등 고정된 포맷을 가지지 않은 형태의 문서를 의미하며, 각 상황에 맞는 토큰이나 구분자 등의 패턴을 통해 주로 정보를 추출한다. 웹은 주로 텍스트로 구성되었기 때문에, 정보 추출 연구는 웹에서 지식을 발견하기 위한 주요 기술로서 다양하게 활용되어 왔다(Line Eikvil, 1999).

웹에 존재하는 문서로부터 정보를 추출한 선행연구들은 미리 정해진 규칙에 맞는 패턴을 찾아 정보를 추출하거나, 훈련 문서와 라벨데이터를 생성한 뒤 모델을 학습하여, 학습된 모델 기반의 정보를 추출하는 방법을 주로 연구하였다. Etzioni et al.(2004)은 온톨로지 지식베이스 및 규칙 템플릿을 활용하여, 온톨로지 클래스와 관계에 대한 정보 추출 규칙을 생성 한 뒤, Naïve Bayes 분류기 기반의 추출결과의 신뢰성을 판단하는 모듈을 설계 함으로서, 웹 규모의 도메인 독립적 정보추출 연구방법론 KNOWITALL을 제안하였다. Banko et al.(2007)은 지식베이스 없이, 의존 구문분석을 통한 Self-supervised 데이터를 구축한 뒤, 분류기를 통해 웹문서로부터 신뢰성 있는 관계 트리플을 추출하는 TEXTRUNNER를 개발하였다. Wu and Weld(2007)은 위키피디아 문서와 infobox의 value 맵핑을 통한 훈련 데이터셋을 구축함으로서, 위키피디아 문서로부터 infobox를 생성하기 위한 정보추출 시스템 KYLIN을 제안하였다.

정보추출 선행연구와 관련하여, Etzioni et al.(2004), Banko et al.(2007)의 연구는 미리 정해진 규칙에 맞는 패턴을 찾아 정보를 추출하는 연구로서, 규칙에 맞지 않은 새로운 패턴의 데이터에는 정보추출 적용이 어렵다는 한계점이 존재한

다. 또한 Wu and Weld(2007)의 연구는 위키피디아를 활용한 자기지도학습 모델을 설계하여, 위키피디아 문서 형태를 가정하고 정보를 추출하기 위한 시스템을 개발하였기 때문에, 훈련데이터(위키피디아)와 유사한 문서 형태의 테스트 데이터에서는 좋은 성능을 보이는 반면, 학습데이터와 이질적인 형태의 테스트 데이터에서는 낮은 성능을 보인다는 한계점이 존재한다. KYLIN은 문서의 위키피디아 카테고리(클래스)를 분류한 뒤, 카테고리에 존재하는 속성에 대하여 '카테고리-속성'별로 존재하는 다수의 전용 모델 선택을 통해 정보를 추출하게 된다. 따라서, 훈련데이터(위키피디아)와 다른 형태의 문서(예: 네이버 뉴스)에서는 카테고리 분류를 잘못하게 되거나, 분류된 카테고리-속성에 해당하는 전용 모델이 없는 경우 정보 추출을 시도하지 못해, 훈련데이터와 이질적인 문서에 대한 정보 추출 환경에서는 낮은 수준의 성능을 보인다는 한계점이 존재한다. 지식베이스 확장을 위한 정보 추출 문제는 실제 웹에 존재하는 비정형 문서를 정보추출의 대상으로 하고 있기 때문에, 도메인 별 문서의 형태가 이질적인 특성이 존재하며, 다양한 형태의 문서에서 효과적으로 관련 정보를 추출하기 위한 연구방법론의 필요성이 제기되고 있는 상황이다.

본 연구에서 제안한 정보추출 방법론은 웹에 존재하는 다양한 형태의 멀티소스 비정형 문서로부터, 주어진 질의에 대해 관련 정보를 추출할 수 있는 강건한 모델을 개발 하였다는 점에 있어서 의미가 있다. 구체적으로 위키피디아 문서 카테고리 정보에 따른 모델 선택 없이, 질의의 서술어 자질에 기반한 Bi-directional LSTM-CRF 알고리즘 기반의 정보추출 모델을 활용하여, 멀티소스로부터 수집된 다양한 형태의 비정형 문서에서도 재현율 성능을 유지할 수 있다.

2.2 기계독해

기계독해는 주어진 문맥을 기계가 이해하고 관련된 질의에 대해 답을 하는 질의응답 모델을 의미한다. 전통적인 정보추출 관련 연구방법론은 질의 분류, 문서 분류, 문장 분류, 언어학적 분석, 특성 추출 등 여러 분석 단계를 통해 질의에 대한 답변 정보를 추출 하는 프로세스로 구성 되어있다. 여러 분석 단계를 거치는 전통적인 정보추출 연구와 달리, 기계독해 연구는 end-to-end 인공신경망 구조를 질의응답에 적용하여, 질의와 문서가 주어져 있을 때 정답을 문서 안에서 찾아내는 것을 목적으로 하며, 인코딩, 상호 집중, 응답 추출 3단계 프로세스로 구성된다. 기계독해 방법론은 attention mechanism 적용 연구를 통해 많은 연구 성과(Herman et al., 2015; Wang and Jiang, 2016; Seo et al., 2016)를 보여주며, 주어진 질문에 대한 정답의 시작과 끝 경계를 지문 내에서 발견하는 SQuAD(Rajpurkar et at., 2016) 데이터 셋을 중심으로 연구가 진행되어 왔다. Herman et al.(2015)는 CNN/Daily Mail 데이터 셋을 활용하여 "문서-질의-답변"기계독해 트리플 생성 방법을 제안 하였으며, Recurrent Neural Network, Attention 기반의 인공신경망 모델의 기계독해 적용을 연구하였다. Wang and Jiang.(2016)은 문서와 질의를 매칭하여 질의에 attention 가중치를 반영하는 기계독해 모델을 제안하였다. 또한 Seo et al.(2016)은 문서와 질의를 양방향 attention 기반의 매칭을 통해 향상된 성능의 기계독해 모델 BIDAF를 제안하였다.

이러한 기계독해 관련 선행연구들은 주로 문서 안에 정답이 존재함을 가정하고, 질의에 대한

문서 안의 정답 위치를 찾아내는 end-to-end 인공신경망 모델에 대한 연구로서, 문서 및 문장 안의 정답 포함 여부를 판단하는 절차가 부재하다. 따라서, 웹에 존재하는 비정경 문서를 대상으로 기계독해 선행연구를 적용할 경우, 정답이 존재하지 않는 문서에서도 정보를 추출하고자 하는 시도가 빈번하게 발생하기 때문에 낮은 수준의 성능을 보인다는 한계점이 존재한다. 따라서, 지식베이스 확장을 위하여 실제 웹에 존재하는 비정형 문서로부터 관련 정보의 포함여부에 대한 종합적인 신뢰성을 고려한 정보 추출 시스템의 연구개발이 필요한 상황이다.

또한 기계독해는 질의응답의 데이터 단위가 문서이기 때문에, 학습 문서와 이질적인 형태의 문서에서는 낮은 질의응답 성능을 보이게 된다. 하지만, 지식베이스 확장을 위한 정보 추출 문제는 실제 웹에 존재하는 다양한 비정형 문서를 정보 추출의 대상으로 하고 있기 때문에, 도메인별 문서의 형태가 이질적인 특성이 존재한다. 따라서, 다양한 형태의 문서에서 효과적으로 관련 정보를 추출하기 위한 연구방법론의 필요성이 제기되고 있는 상황이다.

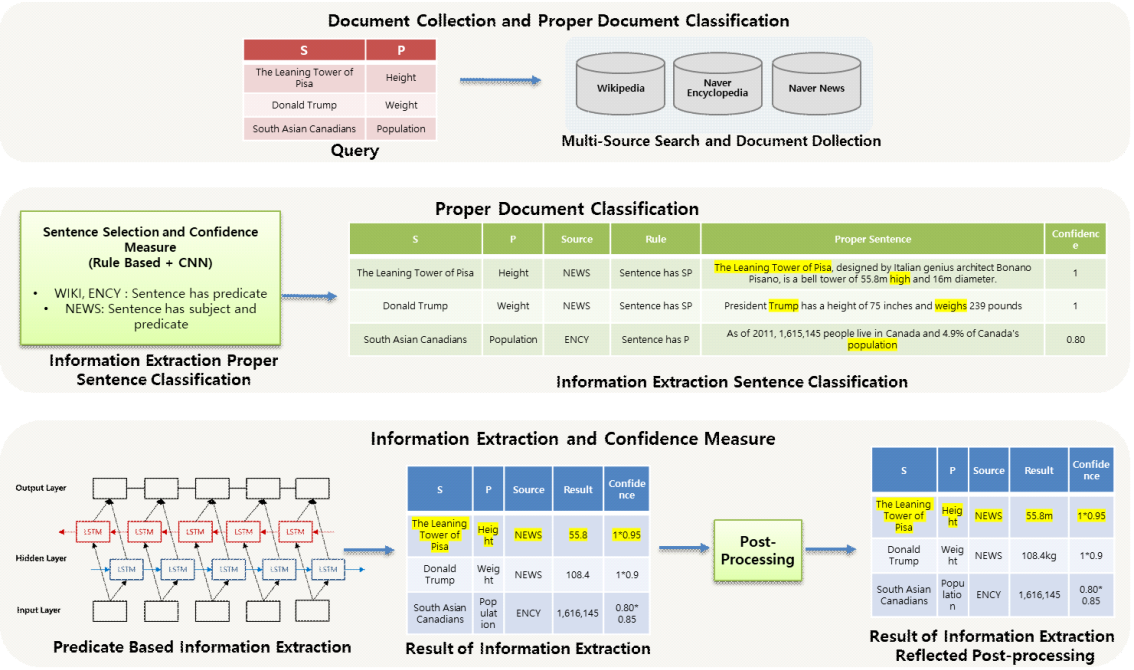
본 연구에서는 웹에 존재하는 멀티소스 비정형 문서로부터 주어진 질의에 대하여, 문장 및 문장내 토큰들의 정보 추출 적합성을 종합적으로 고려한 정보추출 방법론을 제안한다. 이를 통해, 정답이 존재하지 않은 데이터에서 불필요한 정보추출 시도를 방지함으로서, 실제 웹 환경에서도 성능 유지 할 수 있는 모델이라는 점에 있어 의미가 있다. 또한, 본 연구에서 제안하는 모델은 문서가 아닌 문장 단위의 정보추출을 통해, 학습 문서와 이질적인 형태의 문서에서도 정보추출의 성능을 유지할 수 있다는 점에 의미가 있다.

3. 모델

본 연구의 목적은 “주어(Subject)-서술어(Predicate)”로 구분된 질의에 대하여, 위키피디아, 네이버 백과사전, 네이버 뉴스 3개 소스로부터 수집된 비정형 문서에서 관련 정보를 추출하는 시스템을 개발하는 것이다. <Figure 1>은 본 연구에서 제안하는 정보추출 방법론은 개념적으로 설명하고 있다. 제안하는 방법론은 질의 문서 수집 및 적합문서 분류, 적합문장 분류, 정보 추출 및 신뢰도 측정으로 구분된다. 문서 수집 및 적합문서 분류 단계에서는 주어-서술어로 구분된 사용자 질의에 대하여 위키피디아, 네이버 백과사전, 네이버 뉴스의 3개 소스로부터 관련 문서를 수집하고, 정보추출이 적합한 문서 여부를 분류한다. 적합문장 분류 단계에서는 적합 문서라고 판정된 문서안의 문장 중, 질의에 대한 답변 정보를 포함하고 있을 개연성이 있는 문장들을 선별한다. 마지막으로, 정보 추출 및 신뢰도 측정 단계에서는 적합 문장으로 판단된 문장들에 대하여, 서술어 자질벡터를 활용한 정보추출 모델로 문장안의 관련 정보를 추출하고, 정보 추출 결과의 신뢰도를 측정하게 된다.

3.1 문서수집 및 적합문서 분류

“주어(Subject)-서술어(Predicate)로 구분된 질의로그에 대하여, 검색 키워드를 생성하여 위키피디아, 네이버 백과사전, 네이버 뉴스 3개 소스로부터 관련 문서를 수집하고, 정보 추출이 가능한 적합문서를 분류한다. <Table 1>은 위키피디아, 네이버 백과사전, 네이버 뉴스 3개 소스 별 검색 키워드, 문서 수집 방법, 적합문서 분류 방법을 요약하고 있다.



〈Figure 1〉 Proposed Information Extraction Methodology

〈Table 1〉 Document Collection & Proper Document Classification

Source	Search Keyword	Document Collection	Proper Document Classification
Wikipedia	"Subject"	Collect 1 Document Based on Wiki API	Include "Subject" in Title or 1st Paragraph of Document
Naver Encyclopedia	"Subject"	Collect 3 Documents Based on Naver API	
Naver News	"Subject", "Predicate"	Collect 3 Documents Based on Crawler	Distance between Sentence including "Subject" and Sentence including "Predicate" <= 1

위키피디아와 네이버 백과사전으로부터 수집된 문서는 질의의 주어를 활용하여 검색 키워드를 생성하며, 네이버 뉴스는 주어-서술어 질의를 모두 활용하여 검색 키워드를 생성한다. 위키피디아와 네이버 백과사전은 특정 주어를 중점으로 기술된 문서 형태로서 주어 키워드로 검색 하였을 때 적절한 문장이 수집될 개연성이 있는 반면, 네이버 뉴스는 다양한 주어-서술어에 대한 지식이 혼재 되어있는 형태의 문서이기 때문에 주어-서술어 정보를 함께 활용한 키워드를 통해 검색하는 것이 효과적이기 때문이다.

또한 위키피디아를 소스로 하는 문서의 경우

Wikipedia API 기반의 크롤러를 활용하여 1건을 수집하였으며, 네이버 백과사전과 네이버 뉴스의 문서의 경우 각각 네이버 API, 자체개발 크롤러(관련도 순 문서수집)를 활용하여 3건을 수집하였다. 위키피디아의 경우 특정 개체(Entity)에 대한 1건의 문서가 존재하는 형태이기 때문에 1건의 문서를 수집하였고, 네이버 백과사전, 네이버 뉴스의 경우에는 같은 개체 이더라도 다수의 문서가 존재하는 형태이기 때문에 3건의 문서 수집하였다.

위키피디아와 네이버 백과사전의 경우, 특정 주어를 중점으로 기술된 문서 형태로서, 문서 제목이나 본문의 첫 문단에 주어의 내용이 포함되었는지를 바탕으로 적합한 문서를 수집하였는지 여부를 판단 하였다. 네이버 뉴스의 경우 다양한 주어-서술어에 대한 지식이 혼재된 형태이기 때문에, 주어를 포함하는 문장과 서술어를 포함한 문장사이의 거리가 짧을수록 해당 주어-서술어의 정보를 포함할 개연성이 있다. 따라서, 네이버 뉴스에서 수집된 문서의 경우 주어가 포함된 문장과 서술어가 포함된 문장 사이의 거리가 1 이하인 경우 적합문서로 분류하였다. 적합문서 분류과정에서 적합 문서로 판정된 문서의 경우에만 적합 문장 분류 단계를 수행하여 정보추출을 시도하게 된다.

3.2 적합문장 분류

3.2.1 적합문장 분류 및 신뢰도 측정방법

문서 수집 및 적합문서 분류단계에서 적합 문서로 판정된 수집 문서들에 대하여, 주어-서술어로 구분된 질의로그에 대한 답변정보를 포함하고 있을 개연성이 높은 문장들을 규칙 기반으로 선별하며, 각 적합문장의 신뢰도를 계산하기 위하여, CNN기반의 문장 분류 모델 및 규칙기반으로 구성하였다.

본 연구에서는 Yoon Kim이 제안한 Convolution Neural Network(CNN) 기반의 문장 분류 모델(Yoon Kim, 2014)을 응용하여 적합문장의 신뢰도를 판정하기 위한 모델을 개발하였다. 훈련 데이터를 생성하기 위하여, 위키피디아 infobox의 속성-값(Key-Value)와 위키백과 본문의 데이터를 함께 활용하였다. 위키피디아 infobox의 속성이 포함된 문장들 중, 값을 포함하고 있는 문장(적합문장)함하고 있지 않은 문장(부적합 문장)을 구분하여, 문장의 값 포함 여부를 태깅 하였다.

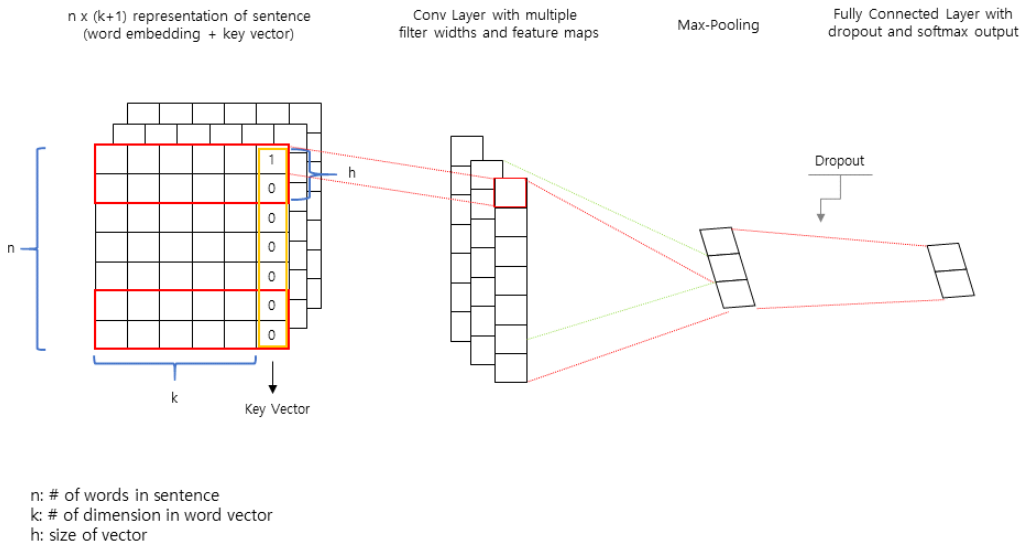
<Figure 2>는 Yoon Kim이 제안한 CNN기반 문장 분류 연구(Yoon Kim, 2014)를 변형하여 본 연구에서 개발한 CNN 기반 적합문장 신뢰도 계산 모델의 구조도를 설명하고 있다. 모델의 구조는 문장 표상 단계, 컨볼루션 단계, max-pooling 단계, Fully-connected layer 단계로 구분된다.

x_i가 문장안의 i번째 어휘에 대한 k차원의 어휘 벡터, $\oplus$는 concatenation 연산을 의미, p가 각 단어들이 질의의 서술어에 해당하는지를 나타내는 n차원의 이진 벡터(n=문장 길이), $x_{1:n}$는 문장으로부터 연결된 어휘 벡터, x'가 문장 표상이라고 하자. 문장 표상 단계에서 모델은 문장을 단어들의 연결된(concatenated) 어휘 벡터($x_{1:n}$)와 질의의 서술어에 해당 하는지를 나타내는 key 벡터(p)의 연결로 표상($x' = x_{1:n} \oplus p$)한다.

$$x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n \quad (1)$$

$$x' = x_{1:n} \oplus p \quad (2)$$

h, w, b가 각각 필터크기, $h*k$차원의 필터, 편향을 의미, f가 비선형 함수, c_i, c가 i+h-1번째 어휘에서 생성된 i번째 feature, feature map을 각각



〈Figure 2〉 Architecture of Confidence Calculation Model Based on Convolutional Neural Network

의미한다고 하자. 컨볼루션 단계에서, 컨볼루션 연산은 필터 w를 $x'^{i:i+h-1}$ h 윈도우 크기의 단어(각 단어가 key인지 여부가 포함)에 적용하여 새로운 feature c_i을 생성하며, 이 필터는 문장표상의 각 단어들의 윈도우에 적용되어 feature map c를 생성한다(b는 편향을 의미).

$$c_i = f(w \cdot x'^{i:i+h-1} + b) \quad (3)$$

$$\mathrm{c} = [c_1, c_2, \ldots, c_{n-h+1}] \quad (4)$$

$\hat{c}$, m은 각각 max-pooled feature map, 필터 개수를 의미하고, $\hat{c_j}$가 j번째 필터의 feature map, z는 flatten layer라고 하자. max-pooling 단계에서는 각 필터의 feature map c에 존재하는 벡터 중 가장 큰 값을 도출하며($\hat{c} = \max\{c\}$), 개의 필터의 max-pooling 결과인 m개 노드는 <Figure 2>와 같이 펼쳐진(flatten) 계층 $\mathrm{z} = [\hat{c_1}, \hat{c_2}, \ldots, \hat{c_m}]$으로 이루어진다.

$$\hat{c} = \max\{c\} \quad (5)$$

$$\mathrm{z} = [\hat{c_1}, \hat{c_2}, \ldots, \hat{c_m}] \quad (6)$$

fully-connected 단계에서는 펼쳐진(flatten) 계층 에 대하여 소프트맥스와 드롭아웃을 적용한 fully-connected layer 인공신경망을 통해, 각 문장의 Value 포함(적합문장) 여부를 예측하게 된다.

<Table 2>는 소스별 적합문장을 선별하는 방법을 요약하고 있다. 위키피디아와 네이버 백과사전으로부터 수집된 문서에서는 질의의 서술어가 포함된 문장을 정보 추출 적합문장으로 분류하였으며, CNN 기반의 적합문장 분류 모델의 점수를 적합문장의 신뢰도로 도출하였다. 위키피디아와 네이버 백과사전에 대하여 서술어가 포함된 문장을 적합문장으로 분류한 이유는 위키피디아와 네이버 백과사전이 특정 주어를 중점으로 기술한 문서로서, 해당 문서에서 질의의 서술어가 포함된 문장이 주어-서술어에 대한 관련

〈Table 2〉 Proper Sentence Classification

Source	Proper Sentence Classification Method
Wikipedia	Include “Predicate” in Sentence
Naver Encyclopedia	
Naver News	Include “Subject” and “Predicate” in Sentence

정보를 포함할 개연성이 있기 때문이다. 또한 네이버 뉴스로부터 수집된 문서에서는 질의의 주어-서술어가 모두 포함된 문장을 정보 추출 적합 문장으로 선정하였으며, 적합문장으로 판정된 문장들의 신뢰도를 1로 도출하였다. 네이버 뉴스의 적합문장 판정을 주어-서술어에 기반한 이유는 다음과 같다. 네이버 뉴스에서 수집된 문서는 특정 주어를 중점으로 기술한 형태가 아니라, 다양한 주어-서술어에 대한 지식이 혼재되어 있어 주어와 서술어가 함께 포함되어 있는 문장이 질의의 주어-서술어와 관련된 답변정보를 포함할 개연성이 있기 때문이다.

3.2.2 적합문장 신뢰도 측정모델 훈련결과

<Table 3>은 적합문장 신뢰도 측정모델의 데이터 셋을 설명하고 있다. 수집된 적합 문장과 부적합 문장은 각각 89,780건, 76,003건 수집되었으며, 본 연구에서는 데이터 셋의 20%를 테스트 데이터 셋을 구축하는데 활용하였으며, 남은 80% 중 72%는 훈련 데이터 셋, 8%는 검증용 데이터셋을 구축하는데 활용하였다.

<Table 4>는 적합문장 신뢰도 측정모델의 훈련 파라미터를 설명하고 있다. 본 연구에서는 128차원의 단어 차원을 활용하였고, 3, 4, 5의 크기를 가진 필터를 각 128개를, $1 * e^{-4}$의 학습속도, 128의 배치크기, 0.5의 드롭아웃 비중을 파라미터로 선정하여 학습에 활용하였다. 본 연구에서는 검증용 데이터 셋의 정확도 성능을 기준으로 학습중도 종료(early stopping)접근 방법을 통해 최적의 학습시간을 선택하였다.

<Table 5>, <Table 6>, <Table 7>은 적합문장 신뢰도 측정모델의 훈련, 검증, 테스트 성능을

〈Table 3〉 Data set of Proper Sentence Confidence Measure Model

Sentences	Train	Validation	Test	Total
Good Sentences	64,709	7,115	17,956	89,780
Bad Sentences	54,655	6,147	15,201	76,003
Ratio	72%	8%	20%	100%

〈Table 4〉 Training Parameters Proper Sentence Confidence Measure Model

Embedding Size	Filter Size	# of Filters	Learning Rate	Batch Size	Drop Out Rate
128	3,4,5	128	$1 * e^{-4}$	128	0.5

〈Table 5〉 Train Performance of Proper Sentence Confidence Measure Model

Label	Precision	Recall	F1-Score	Support	Accuracy
Bad	0.8764	0.8763	0.8763	54,655	0.8867
Good	0.8956	0.8956	0.8956	64,709	

〈Table 6〉 Validation Performance of Proper Sentence Confidence Measure Model

Label	Precision	Recall	F1-Score	Support	Accuracy
Bad	0.7571	0.7651	0.7611	6,147	0.7773
Good	0.7952	0.7879	0.7915	7,115	

〈Table 7〉 Test Performance of Proper Sentence Confidence Measure Model

Label	Precision	Recall	F1-Score	Support	Accuracy
Bad	0.7566	0.7860	0.7710	15,201	0.7859
Good	0.8127	0.7859	0.7991	17,956	

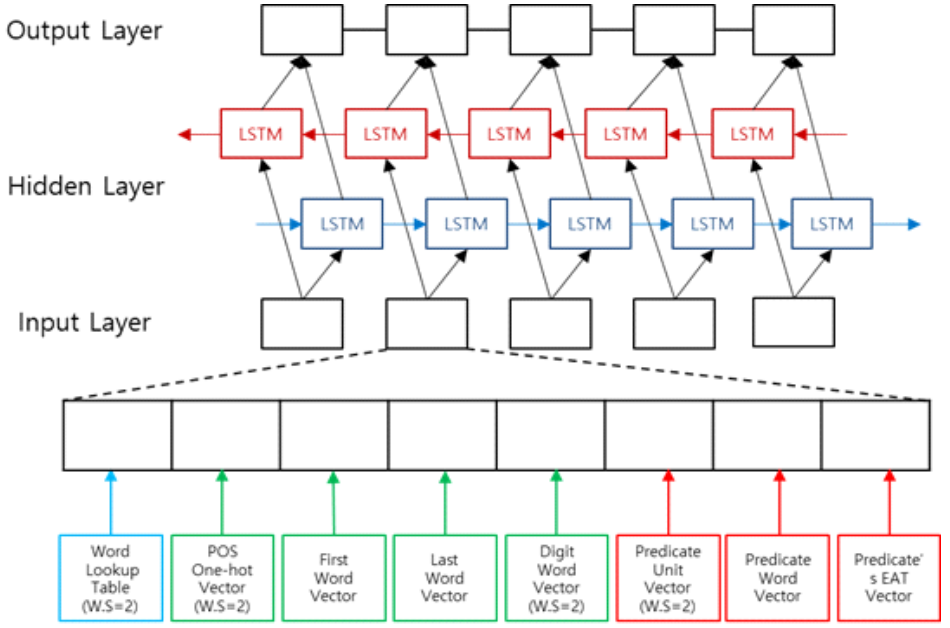
각각 나타내고 있다. 적합문장 신뢰도 측정모델은 훈련용 데이터 셋에서 89.56%, 89.56%. 88.67%의 정밀도와 재현율, 정확도 성능을 보여주었고, 검증용 데이터 셋에서 79.52%, 78.79%, 77.73%의 정밀도, 재현율, 정확도 성능을 보여주었다. 또한 테스트 데이터 셋에서 81.27%, 78.59%, 78.59%의 정밀도, 재현율, 정확도 성능을 각각 보여주었다.

3.3 정보추출 및 신뢰도 측정

3.3.1 정보추출 및 신뢰도 측정방법

본 연구에서는 속성-값(Key-Value)이 포함된 문장에서, 속성의 위치정보에 기반한 값을 추출하기 위한 Bi-directional LSTM-CRF 알고리즘 기반의 시퀀스 태깅 모델을 개발하였다. 정보 추출 모델의 훈련 데이터를 생성하기 위하여, 위키피디아 infobox의 속성-값(Key-Value)와 위키백과 본문의 데이터를 함께 활용하였다. 위키피디아 infobox의 속성-값이 모두 등장한 문장을 본문에서 탐색하고, 각 문장을 형태소 단위로 분리하여, 각 형태소의 속성-값 여부 라벨을 태깅 하였다. 이때, 형태소가 속성의 시작부분일 경우 <init:Key>, 속성의 중간부분일 경우 <middle:Key>, 마지막 부분일 경우 <end:Key>, 값의 시작부분일 경우 <init:Value>, 값의 중간부분일 경우 <middle:Value>, 값의 마지막 부분일 경우 <end:Value>, 속성과 값이 아닌 형태소인 경우 <No> 라벨을 태깅 한다. 생성된 훈련 데이터는 추후, Bi-directional LSTM-CRF 알고리즘 기반의 시퀀스 태깅(Sequence Tagging) 모델을 훈련하는데 활용된다.

<Figure 3>는 정보추출 적합 문장에 대하여,



〈Figure 3〉 Architecture of the Query's Predicate Based Information Extraction Model

〈Table 8〉 Feature Description of Bi-directional LSTM CRF Model

Feature Type	Feature	Description
Word Vector Feature	Word embedding (window size = 2)	Embeddings of the word
Syntactic Feature	Word Morpheme (window size = 2)	Morpheme of the word
	Last Word	Whether it is last word of the sentence
	First Word	Whether it is first word of the sentence
	Digit Word (window size = 2)	Whether it is digit word
Query's Predicate Feature	Unit Word of Query's Predicate (window size = 2)	Whether it is unit word of the query's predicate
	Query's Predicate Word	Whether it is word of the query's predicate
	Expected Answer Type of Query's Predicate	Expected answer type feature of the query's predicate

질의의 서술어(Predicate)정보 기반의 정보 추출 모델의 설계도를 나타내고 있으며, <Figure 3>는 정보 추출 모델을 구현하는데 활용된 Bi-directional LSTM-CRF 모델의 자질들의 명세서를 설명한

다. 본 연구에서는 주어진 문장과 질의의 서술어 정보를 활용하여, 문장안에 존재하는 주어-서술어 질의에 대한 속성-값(Key-Value)를 추출하는 Bi-directional LSTM-CRF기반 시퀀스 태깅 모델로서 정보추출 모델을 구현하였다. 정보 추출 모델은 입력층(Input Layer)과 은닉층(Hidden Layer), 출력층(Output Layer)로 구분되는데, 입력층에서는 어휘 벡터 자질(Word Vector Feature; <Figure 3>의 푸른색으로 표현), 문법 자질(Syntactic Feature; <Figure 3>의 녹색으로 표현), 질의의 서술어 자질(Query's Predicate Feature; <Figure 3>의 붉은색으로 표현)정보를 기반으로 각 단어를 표상한다. 이때, 어휘 벡터 자질은 어휘벡터(window size=2)를 의미하여, 문법 자질은 어휘의 형태소 정보(window size=2), 각 단어들이 문장의 마지막 단어인지 여부, 첫 번째 단어인지 여부를 나타내는 벡터, 각 단어들이 숫자인지 여부를 나타내는 벡터(window size=2)를 나타낸다. 마지막으로, 질의의 서술어 자질은 각 단어가 질의 서술어에 대한 단위정보에 해당하는지의 여부(window size=2), 질의의 서술어에 해당하는지에 대한 여부, 질의의 서술어에 대한 기대되는 답변 데이터 타입 벡터를 의미한다. 입력층의 표상된 문장의 각 단어는 bi-directional LSTM 은닉층을 통과하게 되며, 마지막으로 출력층의 CRF는 각 단어에 대한 속성-값 여부를 태깅하여(문장의 속성은 질의의 서술어를 의미하며 값은 질의에 대한 정답정보를 의미), 값으로 태깅된 단어를 질의에 대한 정답 정보로 출력하게 된다. 또한 값으로 태깅된 단어의 Score 평균값을 바탕으로 정보 추출 신뢰도로 계산한다.

정보추출 모델에서 출력된 질의에 대한 정보는 후처리 모듈을 통해, 단위 정보가 누락되었을 경우 완전한 형태로 결과를 반환 된다. 본 연구에서는 질의의 서술어 카테고리 10종(길이, 무게, 속도, 크기 등)에 대하여, 서술어에 대한 답변 단위정보를 미리 정의 함(<Table 9>)으로서 추출된 정답 주변 에 단위정보 존재 여부를 탐색하여, 불완전하게 추출된 정보를 보정하는 후처리 모듈을 설계하였다.

〈Table 9〉 Post Processing Dictionary

Index	Query's Predicate	Unit Dictionary
1	Release Date, Opening Day, Birth Day, …	year, month, day, …
2	Length, Height, Diameter, Radius, …	*cm, m, km, …*
3	Weight	*kg, g, ton, …*
4	Speed, Maximum Speed	*km/h, m/h, …*
5	Population	people
6	Age, Life Span	age, year, month, day, …
7	Size	$cm, m, km, cm^2, m^2, km^2, \cdots$
8	Area	$cm^2, m^2, km^2, \cdots$
9	Price	Won, Dollar
10	Blood Type	A, B, AB, O

정보추출 모델에서 출력된 질의에 대한 정보는 후처리 모듈을 통해, 단위 정보가 누락되었을 경우 완전한 형태로 결과를 반환 된다. 본 연구에서는 질의의 서술어 카테고리 10종(길이, 무게, 속도, 크기 등)에 대하여, 서술어에 대한 답변 단위정보를 미리 정의 함(<Table 9>)으로서 추출된 정답 주변 에 단위정보 존재 여부를 탐색하여, 불완전하게 추출된 정보를 보정하는 후처리 모듈을 설계하였다.

후처리 과정을 거친 이후 추출된 지식은 적합 문장 스코어와 정보추출 스코어를 곱한 값을 최종 정보 추출 스코어로 도출되며, 최종 정보 추출 스코어를 통해 각 추출결과의 신뢰도를 평가할 수 있다.

3.3.2 정보추출 모델 훈련결과

<Table 10>은 정보추출 모델을 훈련하기 위해 수집된 데이터 셋을 설명하고 있다. 총 82,227건의 데이터가 수집되었으며, 이중 20%는 테스트 데이터 셋을 구축하는데 활용되었고, 남은 64%, 16%의 데이터는 각각 훈련용 데이터 셋, 검증용 데이터 셋을 구축하는데 활용되었다.

<Table 11>는 정보추출 모델의 훈련 파라미터를 설명하고 있다. 본 연구에서는 100차원의 단어 차원을 활용하였고, $5 * e^{-4}$의 학습속도, 256의 배치크기, 0.5의 드롭아웃 비중을 파라미터로 선정하여 학습에 활용하였다. 본 연구에서는 검증용 데이터 셋의 정확도(올바로 태깅된 문장의 비율)성능을 기준으로 학습중도 종료(early stopping)접근 방법을 통해 최적의 학습시간을 결정하였다.

<Table 12>, <Table 13>, <Table 14>는 정보추출 모델의 훈련, 검증, 테스트 성능을 각각 나타내고 있다. 정보추출 모델은 훈련용 데이터 셋에서 평균 99.22%, 99.11%의 정밀도와 재현율 성능을 보여주었고 98.64%의 정확도(올바로 태깅된 문장의 비율 성능)을 보여주었다. 또한 검증용 데이터 셋에서 92.45%, 92.11%, 82.75%의 평균 정밀도, 평균 재현율, 정확도 성능을 보여주었으며, 테스트 셋에서 92.57%, 91.93%, 82.81%의 평균 정밀도, 평균 재현율, 정확도 성능을 보여주었다.

〈Table 10〉 Data Set Description for Information Extraction Model

Sentences	Train	Validation	Test	Total
Frequency	52,587	13,146	16,494	82,227
Ratio	64%	16%	20%	100%

〈Table 11〉 Training Parameters Proper Information Extraction Model

Embedding Size	Learning Rate	Batch Size	Drop Out Rate
128	$5 * e^{-4}$	128	0.5

〈Table 12〉 Train Performance of Information Extraction Model

Label	Precision	Recall	F1-Score	Support
middle:Key	1.0000	1.0000	1.0000	100
middle:Value	0.9767	0.9979	0.9872	26,727
end:Key	1.0000	1.0000	1.0000	2,090
end:Value	0.9783	0.9979	0.9880	28,583
init:Key	1.0000	1.0000	1.0000	52,587
init:Value	0.9994	0.9993	0.9993	52,587
avg/total	0.9992	0.9991	0.9956	162,671
Accuracy(Ratio of correctly tagged sentence): 98.64%				

〈Table 13〉 Train Performance of Information Extraction Model

Label	Precision	Recall	F1-Score	Support
middle:Key	1.0000	1.0000	1.0000	29
middle:Value	0.8629	0.8647	0.8638	6,622
end:Key	1.0000	1.0000	1.0000	504
end:Value	0.8629	0.8743	0.8680	7,086
init:Key	1.0000	1.0000	1.0000	13,146
init:Value	0.9111	0.8927	0.9018	13,146
avg/total	0.9245	0.9211	0.9228	40,533
Accuracy(Ratio of correctly tagged sentence): 82.75%				

〈Table 14〉 Test Performance of Information Extraction Model

Label	Precision	Recall	F1-Score	Support
middle:Key	1.0000	1.0000	1.0000	31
middle:Value	0.8590	0.8587	0.8588	8,321
end:Key	1.0000	1.0000	1.0000	639
end:Value	0.8686	0.8739	0.8712	8,983
init:Key	1.0000	0.9999	1.0000	16,434
init:Value	0.9132	0.8910	0.9020	16,434
avg/total	0.9257	0.9193	0.9225	50,842
Accuracy(Ratio of correctly tagged sentence): 82.81%				

4. 실험

본 연구에서 제안하는 정보추출 모델의 성능을 평가 하기 위하여, SK 텔레콤의 대화형 인공지능 스피커 사용자 질의 로그 데이터 400건을 활용하여, 위키피디아, 네이버 백과사전, 네이버 뉴스 3개 웹소스로부터 수집된 비정형 문서로부터 질의에 대한 올바른 정보를 추출해 내는지 성능을 측정하였다. 제안 모델의 성능 평가를 위해, 질의별 각 소스에서 추출된 정보의 신뢰도가 가장 높은 결과 1건을 선정하였으며, 정보추출 관련 선행연구 Wu and Weld(2007)의 모델을 베이스라인으로 선정하여 베이스라인, 제안된 모델, Wu and Weld(2007)과 제안된 모델을 함께 적용한 앙상블 모델의 성능을 비교분석 하였다.

4.1 성능평가 질의 데이터

정보 추출 모델의 성능을 평가하기 위해 SK 텔레콤 대화형 인공지능 스피커 NUGU에서 수집된 사용자 질의 로그 중 400건을 선정하였다. <Table 15>는 성능 평가 질의 데이터의 질의 타입에 따른 빈도수와 각 비중을 나타낸다. 사용자 질의 로그는 질의의 논리성 측면과 질문의 답변 가능성의 기준을 바탕으로 구분될 수 있다. 질의의 논리성 측면에 따른 구분: 1) 적합질의(Proper Query), 2) 부적합 질의(Improper Query). 적합 질의는 상식적으로 대답이 가능하다고 판단되는 질의를 의미하며, 부적합 질의는 남성의 출산일과 같은 비논리적인 질의이거나, 비상식적 질의를 의미한다. 평가 질의 데이터셋 400건에는 적합질의와 부적합질의가 각 389건, 11건으로 이루어져 있으며 각 비중이 97.25%, 2.75%로서, 논리적이거나 비논리적인 질의가 혼재되어 있다. 질의의 답변 가능성에 따른 구분: 1) 응답 가능 질의(Answerable Query), 2) 검색 불가 질의(Unfindable Query), 3) 응답 불가 질의(Unanswerable Query). 응답 가능 질의는 해당 질의를 통해 본 연구에서 구축한 검색엔진에서 답을 도출할 수 있는 질의를 의미하고, 검색 불가 질의는 상식적으로는 가능하지만 본 연구의 검색엔진이 수집한 문서에서는 답을 찾을 수 없는 질의를 의미한다. 마지막으로 응답 불가 질의는 비논리적인 질의이거나 비상식적인 질의로서 답을 도출하면 안되는 질의 형태를 의미한다. 평가 질의 데이터셋 400건에는 응답가능 질의, 검색 불가 질의, 응답 불가 질의 각각 274건, 104건, 22건으로 이루어져 있으며, 각 비중은 68.50%, 26.00%, 5.50%로서, 혼재되어 있는 상황이다.

〈Table 15〉 Query Set Description

	Answerable Query	Unfindable Query	Unanswerable Query	Total	Ratio
Proper Query	274	104	11	389	97.25%
Improper Query	0	0	11	11	2.75%
Total	274	104	22	400	-
Ratio	68.50%	26.00%	5.50%	-	

4.2 성능 평가 방법

<Table 16>는 제안 모델의 정보추출 성능을 측정하기 위한 혼동행렬을 설명하고 있다. 질의에 대한 정답이 실제 문서에 존재하는 경우(Actual: True)와 그렇지 않은 경우(Actual: False)로 구분할 수 있으며, 정답이 문서에 존재하지 않는 경우는 그 질의가 검색 불가 질의(Unfindable Query)인 경우와 응답 불가 질의(Unanswerable Query)인 경우로 다시 구분할 수 있다. 또한 정보추출 시스템이 예측한 추출 결과가 존재하는 경우(Predicted: True)와 존재 하지 않는 경우(Predicted: False)로 크게 구분되며, 예측한 추출 결과가 존재하는 경우는 다시 옳은 정보를 추출한 경우(Correct)와 틀린 정보를 추출하는 경우(Incorrect)로 구분된다.

정답이 존재하는 문서에서 올바른 정보를 추출할 경우: True Positive Correct Answer(*TPCA*), 틀린 정보를 추출할 경우: True Positive Incorrect Answer(*TPIA*), 정보를 추출하지 못할 경우: False Positive Answer(*FPA*)로 구분된다. 또한, 정답이 존재하지 않는 문서에서 검색 불가 질의에 대해 정보를 추출한 경우: False Positive Answer & Unfindable(FPA_{UF}), 정보를 추출하지 못한 경우: True Positive Answer & Unfindable(TPA_{UF})로 구분된다. 마지막으로 정답이 존재하지 않는 문서에서 응답 불가 질의에 대해 정보를 추출한 경우: False Positive Answer & Unanswerable(FPA_{UA}), 정보를 추출하지 못한 경우: True Positive Answer & Unanswerable(TPA_{UA})로 구분된다.

이렇게 구해진 혼동행렬의 각 셀(<Table 16> 참조)은 <Table 17>과 같이 정보추출 성능을 평가하기 위한 지표를 계산하는데 활용된다. 본 연구에서는 정답이 문서에 존재하는 경우와 그렇지 않은 경우의 예측 정밀도(Precision), 재현율(Recall), F1-Score를 각각 성능평가 지표로 활용하며, 모델의 전반적인 성능을 확인할 수 있는 정확도(Accuracy)를 도출하여 성능을 평가하였다.

문서에 정답이 존재하는 경우의 예측 정밀도 ($Precision_{True} = TPCA/(TPCA + TPIA + FPA_{UF} + FPA_{UA})$) 성능은 모델의 정보 추출 결과가 존재할 경우의 옳은 정보를 도출한 비중을 의미한다. 문서에 정답이 존재하는 경우의 재현율 ($Recall_{True} = TPCA/(TPCA + TPIA + FNA)$) 성능은 문서안의 정답이 존재하는 경우 중 모델이

〈Table 16〉 Confusion Matrix for Proposed Information Extraction

		Actual: True	Actual: False	
			Unfindable Query	Unanswerable Query
Predicted: True	Correct	True Positive Correct Answer *(TPCA)*	False Positive Answer & Unfindable *(FPA_{UF})*	False Positive Answer & Unanswerable *(FPA_{UA})*
	Incorrect	True Positive Incorrect Answer *(TPIA)*		
Predicted: False		False Positive Answer *(FPA)*	True Positive Answer & Unfindable *(TPA_{UF})*	True Positive Answer & Unanswerable *(TPA_{UA})*

〈Table 17〉 Evaluation Metric for Proposed Information Extraction

Metric	Formula
$Precision_{True}$	$\frac{TPCA}{TPCA + TPIA + FPA_{UF} + FPA_{UA}}$
$Recall_{True}$	$\frac{TPCA}{TPCA + TPIA + FNA}$
$F1 - Score_{True}$	$\frac{2 * Precision_{True} * Recall_{True}}{Precision_{True} + Recall_{True}}$
$Precision_{False}$	$\frac{TPA_{UF} + TPA_{UA}}{FNA + TPA_{UF} + TPA_{UA}}$
$Recall_{False}$	$\frac{TPA_{UF} + TPA_{UA}}{FPA_{UF} + TPA_{UF} + FPA_{UA} + TPA_{UA}}$
$F1 - Score_{False}$	$\frac{2 * Precision_{False} * Recall_{False}}{Precision_{False} + Recall_{False}}$
Accuracy	$\frac{TPCA + TPA_{UF} + TPA_{UA}}{TPCA + TPIA + FNA + FPA_{UF} + TPA_{UF} + FPA_{UA} + TPA_{UA}}$

옳은 정보를 도출한 비중을 의미한다.

문서에 정답이 존재하지 않은 경우의 예측 정밀도($Precision_{False} = (TPA_{UF} + TPA_{UA})/(FNA + TPA_{UF} + TPA_{UA})$)성능은 모델이 정보 추출 결과가 없는 경우 중 실제로 정답이 문서에 존재하지 않은 비중을 의미한다. 문서에 정답이 존재하지 않는 경우의 재현율($Recall_{False} = (TPA_{UF} + TPA_{UA})/(FPA_{UF} + TPA_{UF} + FPA_{UA} + TPA_{UA})$) 성능은 실제 문서에 정답이 존재하지 않는 경우 중 모델에서 정보를 추출하지 않은 비중을 의미한다.

문서의 정답이 존재하는 경우의 F1-Score ($F1 - Score_{True} = (2 * Precision_{True} * Recall_{True})/(Precision_{True} + Recall_{True})$), 정답이 존재하지 않은 경우의 F1-Score 성능($F1 - Score_{False} = (2 * Precision_{False} * Recall_{False})/(Precision_{False} + Recall_{False})$은 각각 정답이 존재하는 경우와 존재하지 않은 경우의 예측 정밀도와 재현율 성능의 조화평균을 의미한다. 마지막으로, 모델의 정확도(Accuracy = $(TPCA + TPA_{UF} + TPA_{UA})$ / $(TPCA + TPIA + FNA + FPA_{UF} + TPA_{UF} + FPA_{UA} + TPA_{UA})$ 성능은 정보추출 혼동행렬의 경우의 수 중 정답이 존재할 경우 옳은 정답을 추출한 경우와 정답이 존재하지 않을 경우 정보를 추출하지 않은 비중을 의미한다.

4.3 성능 평가 결과

<Table 18>, <Table 19>, <Table 20>는 각각 위키피디아, 네이버 백과사전, 네이버 뉴스로부터 수집된 문서를 활용하여 400건의 질의문에 대해 정보 추출 연구의 베이스 라인 모델 KYLIN(Wu and Weld(2007)), 본 연구에서 제안하는 모델, KYLIN과 제안된 모델을 함께 활용한 정보 추출 결과(KYLIN에서 정보추출 결과가 없을 경우, 제안된 모델을 적용)의 성능평가 결과를 요약하고 있다.

〈Table 18〉 Performance Measure on Wikipedia

	Actual: True			Actual: False		
	KYLIN	Proposed Model	KYLIN + Proposed Model	KYLIN	Proposed Model	KYLIN+ Proposed Model
Precision	87.23%	60.00%	70.08%	59.49%	61.87%	69.25%
Recall	21.92%	25.66%	43.85%	98.59%	92.95%	92.01%
F1-Score	35.04%	35.95% (+0.91%)	53.94% (+18.90%)	74.20%	74.29% (+0.90%)	79.09% (+4.89%)
Accuracy	KYLIN: 62.75%					
	Proposed Model: 61.50%(-1.25%)					
	KYLIN + Proposed Model: 69.50%(+6.75%)					

〈Table 19〉 Performance Measure on Naver Encyclopedia

	Actual: True			Actual: False		
	KYLIN	Proposed Model	KYLIN + Proposed Model	KYLIN	Proposed Model	KYLIN+ Proposed Model
Precision	90.19%	64.91%	70.54%	52.14%	58.39%	65.74%
Recall	21.10%	33.94%	47.24%	100.00%	91.75%	91.75%
F1-Score	34.20%	44.57% (+10.37%)	56.59% (+22.40%)	68.54%	71.36% (+2.82%)	76.60% (+8.06%)
Accuracy	KYLIN: 57.00%					
	Proposed Model: 60.25%(+3.25%)					
	KYLIN + Proposed Model: 67.50%(+10.50%)					

〈Table 20〉 Performance Measure on Naver News

	Actual: True			Actual: False		
	KYLIN	Proposed Model	KYLIN + Proposed Model	KYLIN	Proposed Model	KYLIN+ Proposed Model
Precision	100.00%	48.57%	49.29%	78.64%	89.69%	89.96%
Recall	2.29%	39.08%	40.22%	100.00%	94.56%	94.56%
F1-Score	4.49%	43.34% (+38.82%)	44.30% (+38.81%)	88.04%	92.06% (+4.02%)	92.21% (+4.17%)
Accuracy	KYLIN: 78.75%					
	Proposed Model: 82.50%(+3.75%)					
	KYLIN + Proposed Model: 82.75%(+4.30%)					

〈Table 21〉 Performance Measure on Multi Source

	Actual: True			Actual: False		
	KYLIN	Proposed Model	KYLIN + Proposed Model	KYLIN	Proposed Model	KYLIN+ Proposed Model
Precision	89.47%	64.49%	69.80%	36.73%	47.61%	55.55%
Recall	18.61%	39.78%	51.45%	100.00%	87.30%	87.30%
F1-Score	30.81%	49.20% (+18.49%)	59.24% (+28.43%)	53.73%	61.62% (+7.90%)	67.90% (+14.17%)
Accuracy	KYLIN: 44.25%					
	Proposed Model: 54.75%(+10.50%)					
	KYLIN + Proposed Model: 62.75%(+18.50%)					

위키피디아 소스로부터의 정보추출 결과 (<Table 18>)와 관련하여, KYLIN과 비교하여 제안된 모델이 $F1-Score_{True}$성능은 0.91% 향상되었으며, $F1-Score_{False}$, Accuracy, 성능은 각각 0.90% 향상, 1.25% 감소 되었다. 그러나, 베이스라인 모델과 제안된 모델을 함께 활용한 경우, $F1-Score_{True}$, $F1-Score_{False}$, Accuracy 각각 18.90%, 4.89%, 6.75% 성능이 향상되었다.

네이버 백과사전 소스로부터 정보추출 결과 (<Table 19>), KYLIN 모델과 비교하여, 제안된 모델의 $F1-Score_{True}$, $F1-Score_{False}$, Accuracy 성능이 각각 10.37%, 2.82%, 3.25% 향상되었으며, KYLIN 모델과 제안된 모델을 함께 활용한 경우 성능 향상 정도가 더 두드려 졌다 ($F1-Score_{True}$, $F1-Score_{False}$, Accuracy 각각 22.40%, 8.06%, 10.50% 향상).

네이버 뉴스 소스로부터 정보추출 결과 (<Table 20>), KYLIN에 비하여, 제안된 모델의 $F1-Score_{True}$, $F1-Score_{False}$, Accuracy 성능이 각각 38.82%, 4.02%, 3.75% 향상되었으며, KYLIN과 제안된 모델을 함께 활용한 경우 성능 향상 정도가 더 극대화 되었다($F1-Score_{True}$, $F1-Score_{False}$, Accuracy 각각 38.81%, 4.17%, 4.30% 향상).

<Table 21>은 위키피디아, 네이버 백과사전, 네이버 뉴스 멀티소스를 모두 활용한 경우의 KYLIN, 제안 모델, KYLIN과 제안 모델을 함께 활용한 정보추출의 결과를 요약하고 있다. 각 모델은 위키피디아-네이버 백과사전-네이버뉴스 순으로 정보 추출 결과가 존재하지 않으면 다음 소스에서 정보를 추출하는 형태로 시스템을 구현하였으며, KYLIN과 제안된 모델을 함께 활용한 경우, KYLIN 위키피디아-KYLIN 네이버 백과사전-KYLIN 네이버 뉴스-제안된 모델의 위키피디아–제안된 모델의 네이버 백과사전-제안된 모델의 네이버 뉴스의 우선순위별로 이전 소스별 모델의 정보 추출결과가 없을 경우 다음 우선순위의 소스별 모델로 정보를 추출하는 형태로 시스템을 구현하였다. KYLIN 보다 제안된 모델의 $F1-Score_{True}$, $F1-Score_{False}$, Accuracy 성능이 18.49%, 7.90%, 10.50% 향상되었으며, KYLIN과 제안된 모델을 함께 활용한 경우, 성능 향상 정도가 극대화 되었다($F1-Score_{True}$, $F1-Score_{False}$, Accuracy 각각 28.43%, 14.17%,

18.50% 향상).

위키피디아 소스의 정보추출 결과가, 제안 모델이 KYLIN 보다 일부 낮은 성능지표를 보여준 이유는, KYLIN은 훈련데이터로 활용된 위키피디아 문서형태에 최적화된 반면, 제안된 모델이 웹에 존재하는 다양한 형태의 문서에 범용적으로 적용 될 수 있는 전략을 추구했기 때문으로 판단된다. 이는 위키피디아가 아닌 네이버 지식백과, 네이버 뉴스에서의 $F1-Score_{True}$, $F1-Score_{False}$, Accuracy 성능이 제안된 모델이 KYLIN 보다 더 높은 수준을 보여주었다는 점을 통해서 확인할 수 있다. 특히, 네이버 뉴스의 경우 KYLIN 모델의 $Recall_{True}$ 성능이 2.29%로서 낮은 정보추출 성능을 보여주었으나, 제안 모델의 경우 성능이 크게 증가되어 $F1-Score_{True}$가 향상되었는데, 이는 KYLIN 모델이 훈련데이터의 문서형태(위키피디아)와 이질적인 기사 형태의 문서에서는 정보 추출을 효과적으로 하지 못하는 반면, 제안 모델은 훈련데이터와 이질적인 기사 형태의 문서에서 정보를 추출하는 경우에도 높은 수준의 재현율을 유지 할 수 있는 강건한 모델임을 의미한다. 또한, KYLIN은 제안된 모델에 비해 높은 $Precision_{True}$ 성능이 측정된 반면, $Recall_{True}$은 제안 모델이 더 우수한 성능이 측정되었다는 점, 두 모델을 함께 활용하였을 경우, 더 우월한 $F1-Score_{True}$성능이 측정되었다는 점을 통해, KYLIN 모델과 제안 모델이 보완적 관계에 있음을 알 수 있다. 따라서, 상대적으로 정보추출 커버리지가 낮지만, 신뢰성 높은 정보를 추출하는 KYLIN으로 먼저 정보를 추출한 뒤 추출결과가 존재 하지 않았을 경우, 상대적으로 정보추출 커버리지가 우월한 제안된 모델을 활용하였을 때, 상호 보완을 통해 효과적인 정보추출이 이루어진 것이라 판단된다.

5. 결론

본 연구에서는 지식베이스 확장을 위하여 실제 웹에 존재하는 멀티소스에서 수집된 다양한 형태의 비정형 문서로부터, 질의에 대해 답변 정보를 추출하기 위한 방법론을 제안하였다. 이를 위해, 주어-서술어로 구분된 질의에 대하여, 위키피디아, 네이버 백과사전, 네이버 뉴스 3개 소스로부터 관련 문서를 수집하고 문서, 문장의 정보추출 적합여부를 판단한 뒤, 서술어 질의의 위치 정보를 기반으로 문장안의 관련 정보를 추출하여 정보 추출결과의 종합적인 신뢰도를 도출하는 시스템을 개발하였다. 정보추출 모델의 성능 평가를 위하여 SK텔레콤의 대화형 인공지능 스피커 사용자 질의 400건을 선정하여, 정보추출 연구의 베이스라인 모델 Wu and Weld(2007)의 성능을 비교 함으로서, 기존의 모델보다 더 높은 성능 지표를 보임을 확인하였다.

본 연구의 학술적 기여점은 질의의 서술어 자질을 활용한 Bi-directional LSTM-CRF 기반의 시퀀스 태깅 모델을 개발하여, 멀티 소스로부터 수집된 다양한 형태의 비정형 문서에서도 재현율 성능을 유지 할 수 있는 강건한 모델을 제시 하였다는 점이다. 지식베이스 확장을 위한 정보 추출 문제는 실제 웹에 존재하는 비정형 문서를 정보 추출의 대상으로 하고 있기에 소스별 문서의 형태가 이질적인 특성이 존재하며, 제안 하는 방법론은 베이스라인 모델에 비하여 다양한 형태의 비정형문서에서도 효과적으로 정보를 추출함을 입증하였다. 이는 훈련 데이터의 문서 형태와 이질적인 문서에 대한 정보를 추출하는 경우 낮은 재현율 성능을 보이는 기존 정보추출 관련 선행연구와 차별화 되는 점이다.

또한 본 연구는 정보 추출 단계 이전에 문서,

문장의 정보 추출 적합성 여부를 예측하는 규칙을 통해, 정답이 존재하지 않는 데이터에서 불필요한 정보추출 시도를 방지 함으로서, 실제 웹 환경에서도 예측 정밀도 성능을 유지 할 수 있는 정보 추출 적합성 여부 판단 정책을 제공하였다는 점에 있어 의미가 있다. 지식베이스 확장을 위한 정보 추출 문제는 벤치마크 데이터 셋이 아닌 실제 웹에 존재하는 비정형 문서를 대상으로 하고 있기에, 문서의 정답 포함여부를 보장할 수 없다는 특성이 존재한다. 주로 문서안에 정답이 존재하는 벤치마크 데이터셋을 활용하여 질의에 대한 정답위치를 찾아내는 기계 독해 관련 연구 방법론은 웹에 존재하는 비정형 문서를 대상으로 질의응답을 수행할 경우 정답이 존재하지 않는 문서에서도 답변추출을 빈번히 시도하기 때문에 낮은 수준의 예측 정밀도 성능을 보인다는 한계점이 존재한다. 본 연구에서 제안한 문서와 문장의 정보추출 적합성을 예측하는 정책은 실제 웹 환경에서도 예측 정밀도 높은 정보를 추출하는데 기여한다는 점에 의미가 있다.

본 연구의 실무적 기여점은 다음과 같다. 본 연구는 “주어-서술어”로 구분된 사용자 질의에 대하여 답변정보를 추출하기 위한 시스템의 개발 방법론을 제안 함으로서, 국내 대화형 인공지능 스피커의 개발자로 하여금 구체적인 멀티소스 지식검색 기능을 구현하는데 도움을 줄 수 있을 것이다. 대화형 인공지능 스피커의 지식베이스를 구축 하는 작업은 사람이 검색엔진에 질의를 통해 얻어진 문서로부터 직접 후보군을 탐색하고 신뢰성을 함께 고려해야 하는 수작업을 거쳐야 하는 작업이므로 비용이 많이 드는 활동이다. 본 연구에서 제안 하는 방법론은 사용자 질의에 대하여 답변 정보를 추출하고 추출 결과의 종합적인 신뢰도를 제공하는 자동화된 시스템으로서, 사람의 수작업이 요구되는 지식베이스 트리플 추출 작업의 비용을 감소시키는데 기여할 수 있을 것으로 기대한다. 또한 본 연구에서 제안하는 방법론을 통해, 대화형 인공지능 스피커가 대답하지 못하였던 사용자 질의에 대하여 선택적으로 답변 정보를 추출 함으로서, 지식검색의 커버리지 개선에도 기여할 것으로 판단된다.

연구의 한계점은 다음과 같다. 첫째, 데이터 전처리와 관련된 문제이다. 본 연구에서는 오픈소스인 Konlpy python 패키지(Park and Cho, 2014) 기반의 형태소 분석을 통해 지식 추출의 단위를 구분 하였는데, 형태소 분석을 제대로 수행하지 못해 정보 추출결과가 부적절하게 수행될 수 있다. 추후 연구에서는 오픈소스가 아닌 고도화된 형태소 분석기를 개발하여, 정보 추출 결과의 성능을 고도화 할 필요가 있다. 둘째, 개체 모호성의 문제이다. 본 연구의 정보 추출 시스템은 동명이인을 구분할 수 없어, 같은 이름을 가진 여러 사람이 뉴스에 등장할 경우 질의에서 의도한 주어-서술어에 대한 정보를 추출하지 못할 가능성이 있다. 이에, 향후 연구에서는 동명이인의 인물을 특정하기 위한 조치가 필요할 것으로 판단된다. 셋째, 평가 질의 데이터의 문제이다. 본 연구에서는 SK텔레콤의 대화형 인공지능 스피커로부터 수집된 사용자 질의 중 400건을 선정하여 정보추출 시스템의 성능을 평가하였다. 본 연구에서는 총 2,800건의 문서(400건 질의 * 각 질의 당 7개 문서(위키피디아 1건, 네이버 백과사전 3건, 네이버 뉴스3건))의 정답 포함여부 등을 판단하여 평가 지표를 계산하였다. 연구의 외적 타당성을 보장하기 위해서는 보다 많은 질의를 사용하여 시스템의 성능을 판단하는 것이 바람직하지만, 이러한 작업은 사람의 수작업을 거쳐야 하는 비용이 많이 드는 활동이기에

본 연구에서는 400건의 평가 질의 데이터셋 및 라벨링을 통하여 평가환경을 구축하였다. 향후 연구에서는 보다 많은 질의에 대한 시스템 평가를 통해 연구의 타당성을 입증하기 위한 작업이 필요할 것이다. 또한 멀티소스 웹문서로부터 질의에 대한 정보추출 시스템의 한국어 벤치마크 데이터셋 개발을 통해, 더욱 객관적으로 연구결과를 평가할 수 있는 환경을 구축하기 위한 연구도 필요할 것으로 판단된다.

또한 본 연구가 가지는 향후 연구 방향은 다음과 같다. 첫째, 정보추출 모델의 앙상블 전략을 통한 검색 소스별 최적화에 대한 내용이다. 본 연구의 기여점은 훈련데이터(위키피디아)와 다른 형태의 문서(예: 네이버 뉴스)에서 정보를 효과적으로 추출할 수 있는 강건한 시스템을 개발한 점이다. 제안한 모델이 훈련데이터와 다른 네이버 백과사전과 네이버 뉴스로부터 수집된 문서에서는 KYLIN보다 높은 성능을 보여주었으나, 위키피디아 검색소스에서는 KYLIN보다 비슷하거나 약간 낮은 수준의 성능을 보여주었다. 이는 훈련데이터인 위키피디아와 유사한 문서에서는 기존의 KYLIN의 모델이 최적화되어 있으며, 훈련데이터와 다른 형태의 소스에서는 제안 모델이 효과적으로 정보를 추출한 것을 의미한다. 따라서, 후속 연구에서는 KYLIN과 제안 모델을 함께 적용하여 정보추출 결과를 종합하기 위한 앙상블 모델이 필요할 것으로 판단된다. 이를 통해 검색 소스별 정보추출 모델의 가중치를 고려함으로서, 정보추출의 고도화에 기여할 수 있을 것으로 기대된다. 둘째, 검색 소스 확장에 대한 문제이다. 본 연구에서는 위키피디아, 네이버 백과, 네이버 뉴스의 검색소스에 한정하여 정보추출 방법론을 설계하였다. 향후 연구에서는 소비자 후기 및 네이버 지식인과 같은 비정형 특성이 강하게 나타나는 검색소스에서도 정보를 추출하기 위한 확장 연구가 필요할 것으로 판단된다.

참고문헌(References)

Ahn, S., “Deep Learning Architectures and Applications”, *Journal of Intelligence and Information Systems*, Vol.22, No.2(2018), 127~142.

Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni, "Open information extraction from the web." *IJCAI*. Vol. 7, (2007), 2670~2676.

Eikvil, L, "Information extraction from world wide web-a survey.", *Technical Report 945*, Norwegian Computing Center, 1999.

Etzioni, O., M. Cafarella, and D. Downey, "Web-scale information extraction in knowitall:(preliminary results)." *Proceedings of the 13th international conference on World Wide Web*. ACM, (2004), 100~110.

Gaizauskas, R., and Y. Wilks, "Information extraction: Beyond document retrieval.", *Journal of documentation*, Vol 54, No.1 (1998), 70~105.

Hermann, K., et al. "Teaching machines to read and comprehend." *Advances in Neural Information Processing Systems*, (2015), 1693~1701.

Huang, Z., W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging." *arXiv preprint*, (2015).

JIN, S., H. Jang, 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), 253~267).

Khot, T., A. Sabharwal, and P. Clark, "Answering complex questions using open information extraction." *arXiv preprint*, (2017).

Kim, Y. "Convolutional neural networks for sentence classification." *arXiv preprint*, (2014).

Park, E.L., and Cho. S, "KoNLPy: Korean natural language processing in Python." *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, (2014). 133~136.

Park, H., M. Song, K. Shin, " Sentiment Analysis of Korean Reviews Using CNN: Focusing on Morpheme Embedding", *Journal of Intelligence and Information Systems*, Vol.24, No.2(2018), 59~83.

Qiu , Lin., H. Zhou, Y. Q, W. Zhang, S. Li, S. Rong, D. Ru, L. Qian, W. Tu, Y. Yu, "QA4IE: A Question Answering based Framework for Information Extraction." *arXiv preprint*, (2018).

Rajpurkar, P., J. Zhang, K. Lopyrev, P. Liang, "Squad: 100,000+ questions for machine comprehension of text." *arXiv preprint*, (2016).

Seo, M., A. Kembhavi, A. Farhadi, H. Hajishirzi, "Bidirectional attention flow for machine comprehension." *arXiv preprint*, (2016).

Wang, W., N. Yang, F. Wei, B. Chang, M. Zhou, "Gated Self-matching networks for reading comprehension and question answering." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, (2017). 189~198.

Wang, S., and J. Jiang, "Machine comprehension using match-lstm and answer pointer." *arXiv preprint*, (2016).

Wu, F., and D. S. Weld, "Autonomously semantifying wikipedia." *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, ACM, (2007), 41~50.

Abstract

Development of Information Extraction System from Multi Source Unstructured Documents for Knowledge Base Expansion

Hyunseung Choi* · Mintae Kim* · Wooju Kim** · Dongwook Shin*** · Yong Hun Lee***

In this paper, we propose a methodology to extract answer information about queries from various types of unstructured documents collected from multi-sources existing on web in order to expand knowledge base. The proposed methodology is divided into the following steps. 1) Collect relevant documents from Wikipedia, Naver encyclopedia, and Naver news sources for "subject-predicate" separated queries and classify the proper documents. 2) Determine whether the sentence is suitable for extracting information and derive the confidence. 3) Based on the predicate feature, extract the information in the proper sentence and derive the overall confidence of the information extraction result.

In order to evaluate the performance of the information extraction system, we selected 400 queries from the artificial intelligence speaker of SK-Telecom. Compared with the baseline model, it is confirmed that it shows higher performance index than the existing model.

The contribution of this study is that we develop a sequence tagging model based on bi-directional LSTM-CRF using the predicate feature of the query, with this we developed a robust model that can maintain high recall performance even in various types of unstructured documents collected from multiple sources. The problem of information extraction for knowledge base extension should take into account heterogeneous characteristics of source-specific document types. The proposed methodology proved to extract information effectively from various types of unstructured documents compared to the baseline model. There is a limitation in previous research that the performance is poor when extracting information about the document type that is different from the training data.

In addition, this study can prevent unnecessary information extraction attempts from the documents

* Department of Industrial Engineering, Yonsei University
** Corresponding Author: Wooju Kim
Department of Industrial Engineering, Yonsei University
50, Yonsei-ro, Seodaemun-gu, Seoul, Republic of Korea
Tel: +82-2-2123-5716, Fax: +82-2-364-7807, E-mail: wkim@yonsei.ac.kr
*** Knowledge Technology Cell, AI Technology Unit, AI Center, SK Telecom

that do not include the answer information through the process for predicting the suitability of information extraction of documents and sentences before the information extraction step. It is meaningful that we provided a method that precision performance can be maintained even in actual web environment. The information extraction problem for the knowledge base expansion has the characteristic that it can not guarantee whether the document includes the correct answer because it is aimed at the unstructured document existing in the real web. When the question answering is performed on a real web, previous machine reading comprehension studies has a limitation that it shows a low level of precision because it frequently attempts to extract an answer even in a document in which there is no correct answer. The policy that predicts the suitability of document and sentence information extraction is meaningful in that it contributes to maintaining the performance of information extraction even in real web environment.

The limitations of this study and future research directions are as follows. First, it is a problem related to data preprocessing. In this study, the unit of knowledge extraction is classified through the morphological analysis based on the open source Konlpy python package, and the information extraction result can be improperly performed because morphological analysis is not performed properly. To enhance the performance of information extraction results, it is necessary to develop an advanced morpheme analyzer.

Second, it is a problem of entity ambiguity. The information extraction system of this study can not distinguish the same name that has different intention. If several people with the same name appear in the news, the system may not extract information about the intended query. In future research, it is necessary to take measures to identify the person with the same name.

Third, it is a problem of evaluation query data. In this study, we selected 400 of user queries collected from SK Telecom 's interactive artificial intelligent speaker to evaluate the performance of the information extraction system. n this study, we developed evaluation data set using 800 documents (400 questions * 7 articles per question (1 Wikipedia, 3 Naver encyclopedia, 3 Naver news) by judging whether a correct answer is included or not. To ensure the external validity of the study, it is desirable to use more queries to determine the performance of the system. This is a costly activity that must be done manually. Future research needs to evaluate the system for more queries. It is also necessary to develop a Korean benchmark data set of information extraction system for queries from multi-source web documents to build an environment that can evaluate the results more objectively.

Key Words : Information Extraction, Question Answering System, Machine Reading Comprehension, Bi-directional LSTM-CRF, Knowledge Base

Received : October 29, 2018 Revised : December 17, 2018 Accepted : December 18, 2018
Publication Type : Conference(Fast-track) Corresponding Author : Wooju Kim

저 자 소 개

최 현 승

연세대학교 산업공학과에서 석사과정 재학 중이다. 주요 관심 분야는 자연어 처리, 이상 탐지, 머신러닝/딥러닝 활용 등이다. 지능정보연구, 한국빅데이터학회지 등 국내 저널에 논문을 게재한 바 있다.

김 민 태

연세대학교 산업공학에서 통합과정 재학 중이다. 주요 관심 분야는 머신러닝, 딥러닝을 활용한 자연어 처리, 추천시스템 등이다. 한국정보과학회, 지능정보시스템학회 등 국내 저널에 논문 게재 및 발표한 바 있다.

김 우 주

1987년 연세대학교 BBA 과정 학사 학위를 취득하고, 1994년 KAIST 경영과학 박사를 취득하였으며, 현재 연세대학교 정보산업공학과 교수로 재직 중이다. 관심분야는 시맨틱 웹, 시맨틱 웹 환경의 의사결정지원 시스템, 시맨틱 웹 마이닝, 지식관리 및 인공지능 웹 서비스이다.

신 동 욱

2014년 한양대학교 컴퓨터공학 박사를 취득하고, 2014년부터 2017년까지 Naver Search에서 재직하였으며, 현재 SK텔레콤의 AI센터 지식기술 Cell에 재직 중이다. 인공지능플랫폼 NUGU의 자연어처리 응용 기술을 개발하고 있다. 관심분야는 정보추출, 질의응답 서비스, 텍스트 마이닝, 자연어처리 응용, 머신러닝 등이다.

이 용 훈

2011년 포항공과대학교 컴퓨터공학 박사를 취득하고, 현재 SK텔레콤의 AI센터 지식기술 Cell의 리더로 재직하며, 인공지능플랫폼 NUGU의 지식베이스 관리와 NLP Core 엔진을 개발하고 있다. 관심분야는 구문분석, 정보추출, 질의응답 서비스, 머신러닝 등이다.