ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2019 September: 25(3): 161~177 http://dx.doi.org/10.13088/jiis.2019.25.3.161

의존 구문 분석을 이용한 질의 기반 정답 추출

이도경
연세대학교 산업공학과
(*galaxy1019@yonsei.ac.kr*)

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

김우주
연세대학교 산업공학과
(*wkim@yonsei.ac.kr*)

질의응답 시스템은 크게 사용자의 질의를 분석하는 방법인 질의 분석과 문서 내에서 적합한 정답을 추출하는 방법인 정답 추출로 이루어지며, 두 방법에 대한 다양한 연구들이 진행되고 있다. 본 연구에서는 문장의 의존 구문 분석 결과를 이용하여 질의응답 시스템 내 정답 추출의 성능 향상을 위한 연구를 진행한다. 정답 추출의 성능을 높이기 위해서는 문장의 문법적인 정보를 정확하게 반영할 필요가 있다. 한국어의 경우 어순 구조가 자유롭고 문장의 구성 성분 생략이 빈번하기 때문에 의존 문법에 기반한 의존 구문 분석이 적합하다. 기존에 의존 구문 분석을 질의응답 시스템에 반영했던 연구들은 구문 관계 정보나 구문 형식의 유사도를 정의하는 메트릭을 사전에 정의해야 한다는 한계점이 있었다. 또 문장의 의존 구문 분석 결과를 트리 형태로 표현한 후 트리 편집 거리를 계산하여 문장의 유사도를 계산한 연구도 있었는데 이는 알고리즘의 연산량이 크다는 한계점이 존재한다. 본 연구에서는 구문 패턴에 대한 정보를 사전에 정의하지 않고 정답 후보 문장을 그래프로 나타낸 후 그래프 정보를 효과적으로 반영할 수 있는 Graph2Vec을 활용하여 입력 자질을 생성하였고, 이를 정답 추출 모델의 입력에 추가하여 정답 추출 성능 개선을 시도하였다. 의존 그래프를 생성하는 단계에서 의존 관계의 방향성 고려 여부와 노드 간 최대 경로의 길이를 다양하게 설정하며 자질을 생성하였고, 각각의 경우에 따른 정답 추출 성능을 비교하였다. 본 연구에서는 정답 후보 문장들의 신뢰성을 위하여 웹 검색 소스를 한국어 위키백과, 네이버 지식백과, 네이버 뉴스로 제한하여 해당 문서에서 기존의 정답 추출 모델보다 성능이 향상함을 입증하였다. 본 연구의 실험을 통하여 의존 구문 분석 결과로 생성한 자질이 정답 추출 시스템 성능 향상에 기여한다는 것을 확인하였고 해당 자질을 정답 추출 시스템뿐만 아니라 감성 분석이나 개체명 인식과 같은 다양한 자연어 처리 분야에 활용 될 수 있을 것으로 기대한다.

주제어 : 질의응답 시스템, 정답 추출, 의존 구문 분석, 그래프 임베딩, Bi-directional LSTM-CRF

논문접수일 : 2019년 6월 4일 논문수정일 : 2019년 6월 27일 게재확정일 : 2019년 7월 3일
원고유형 : 학술대회(급행) 교신저자 : 김우주

1. 서론

정보 통신 기술이 발전하면서 사용자들이 인터넷을 통하여 얻을 수 있는 정보의 양이 폭발적으로 증가해왔다. 이러한 정보의 홍수 속에서 사용자가 원하는 정보와 관련된 문서를 찾기 위한 정보 검색 기술이 연구되어왔다. 그러나 검색 대상 문서의 양이 늘어남에 따라 사용자가 문서 내에서 정보를 찾기 위한 많은 시간과 노력이 필요하였다. 이에 따라 사용자의 질의에 대하여 구체적인 정답을 제공해주는 질의응답 시스템에 대한 요구가 증가했다.

질의응답 시스템은 대용량의 비정형 문서로부터 주어진 질의에 대하여 구체적인 정답을 제공해주는 시스템으로(Kawahara et al., 2002), 크게 사용자의 질의를 분석하는 질의 분석과 문서 내

에서 적합한 정답을 찾는 과정인 정답 추출로 구성된다. 이 중 정답 추출 과정은 시스템의 성능에 큰 영향을 미치기 때문에 정답 추출에 관한 많은 연구가 수행되었다(Abney et al., 2000 ; Ittycheriah et al., 2000). 정답 추출의 성능을 높이기 위해서는 문장의 문법적인 정보를 정확하게 반영해야하기 때문에 정답 추출에 구문 분석의 결과를 반영하려는 방법(Doan-Nguyen et al., 2004; Mendes et al., 2011)들이 시도되었는데 기존의 연구들은 구문 관계 정보나 구문 형식의 유사도를 정의하는 메트릭을 사전에 정의해야한다는 한계점이 있었다. 따라서 본 논문에서는 이러한 한계점을 극복하기 위하여 구문 패턴에 대한 정보를 사전에 정의하지 않고 구문 분석 결과를 활용한 정답 추출 방법에 대하여 연구하였다.

구문 분석은 문장의 구성요소들(어절, 구, 절)이 이루는 문법적 구조를 파악하는 과정으로 구 구조 분석과 의존 구문 분석으로 나누어진다(Shin, 1999). 구 구조 분석은 구의 구조 규칙에 따라서 구를 만드는 방식이고 의존 구문 분석은 문장 안의 의존소와 지배소의 구조적 관계를 파악하는 방식이다. 한국어의 경우 어순 구조가 자유롭고 문장의 구성 성분 생략이 빈번하기 때문에 의존 문법에 기반한 의존 구문 분석이 적합하다(Kwon and Choi, 1992). 따라서 본 연구에서는 한국어의 특성을 고려하여 의존 구문 분석의 결과를 이용한 자질을 추가한 질의 기반의 정답 추출 방법을 제안한다. 구체적으로, 정답 추출의 성능 정답 후보 문장을 그래프로 나타낸 후 그래프 정보를 효과적으로 반영할 수 있는 Graph2Vec을 활용하여 입력 자질을 생성하였고, 이를 정답 추출 모델(Bi-directional LSTM-CRF)의 입력에 추가하여 정답 추출 성능 개선을 시도하였다. 의존 그래프를 생성하는 단계에서 의존 관계의 방향성 고려 여부와 노드 간 최대 경로의 길이를 다양하게 설정하며 자질을 생성하였고, 각각의 경우에 따른 정답 추출 성능을 비교하였다.

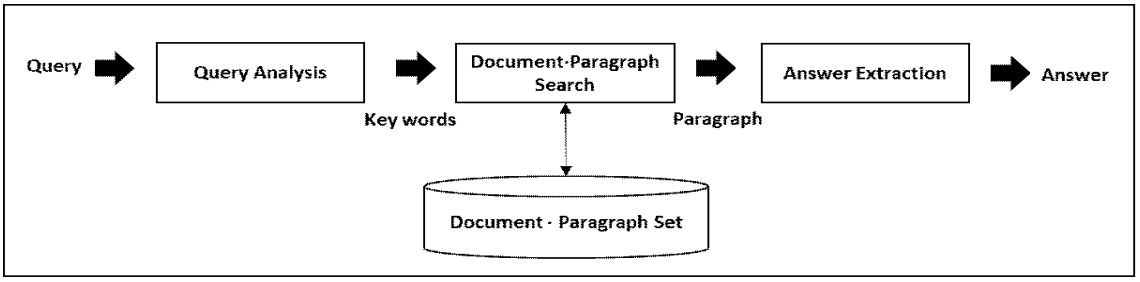
일반적인 질의응답 시스템은 질의 분석, 문장 및 문서 검색, 정답 추출의 과정을 가진다. 그런데 본 연구에서는 질의 분석과 문장 및 문서 검색 과정은 선행되었다고 가정하고 주어진 정답 후보 문장으로부터 정답을 추출할 때 의존 구문 분석의 결과로 생성한 자질을 추가하여 정답 추출의 성능을 측정한 후 비교하였다. 단, 정답 후보 문장들의 신뢰성을 위하여 웹 검색 소스를 한국어 위키백과, 네이버 지식백과, 네이버 뉴스로 제한하였다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련된 질의응답 시스템에서의 정답 추출 연구와 의존 구문 분석에 관한 연구와 구문 분석을 이용한 추출 시스템들에 대해서 소개한다. 3장에서는 정답 추출 모델과 기본 단어 자질 및 의존 구문 분석 결과로 생성한 자질에 대하여 설명한다. 4장에서는 실험 방법과 결과를 설명하고, 5장에서는 본 연구의 결론과 의의를 기술한다.

2. 관련 연구

2.1 질의응답 시스템에서의 정답 추출

<Figure 1>은 일반적인 질의응답 시스템 과정으로 크게 질의 분석과 정답 추출로 나누어진다. 정답 추출의 성능을 향상시키기 위하여 다양한 방식들이 연구되었다. 기존에는 사전이나 정의문을 지니는 코퍼스로부터 언어 규칙 혹은 문법 패턴들을 구축한 후 이를 이용하여 정답을



〈Figure 1〉 General Question Answering System Process

추출하는 패턴 기반의 방식의 연구(Soubbotin et al., 2001; Ravichandran and Hovy, 2002)가 수행되었다. 또 패턴과 통계적인 방법을 결합하여 패턴들의 적절한 가중치를 학습하려는 방법(Ravichandran et al., 2003)도 시도되었다. 그러나 패턴 기반의 방식은 낮은 재현율을 보이고 문장이 길어질 때 먼 거리의 의존성을 고려하지 못한다는 한계점이 존재한다.

이러한 한계점을 극복하기 위하여 최근에는 라벨링이 되어있는 충분한 양의 데이터를 학습시켜 정답을 추출하려는 기계학습 방식의 연구가 활발하게 진행되고 있다. (Yen et al., 2013)은 SVM(Support Vector Machine)을 이용하여 정답 후보들을 질의와 연관 있는 순서대로 랭킹을 정하였다. (Yu et al., 2014)는 정답 문장 선택 시 점수를 계산하기 위하여 CNN(Convolutional Neural Networks) 모델을 이용하였다. Choi et al. (2018)은 시퀀스 태깅에 주로 사용되는 Bi-directional LSTM-CRF모델을 이용하여 질의에 대한 정답을 추출하였다.

2.2 의존 구문 분석

의존 구문 분석은 문장 안의 의존소와 지배소의 구조적 관계를 파악하는 과정으로 2000년대 중반 이후 의존 구문 분석이 영어권을 중심으로 연구의 주요 흐름으로 자리 잡았다. 의존 구문 분석에 대한 연구에는 크게 그래프 기반의 방법과 전이 기반의 방법이 있다. 그래프 기반의 방법(McDonald et al., 2005)은 문장 내의 각 단어를 정점으로, 의존관계를 간선으로 하는 그래프를 만든 후 가장 높은 점수를 갖는 최대 신장 트리를 선택한다. 전이 기반의 방법(Nivre, 2004)은 데이터를 사용하여 전이 행동을 학습한 후 현재의 스택, 버퍼의 정보를 바탕으로 가장 높은 점수를 갖는 다음 행동을 결정한다.

국내에서는 '21세기 세종 계획' 연구 결과물인 세종 구문 분석 말뭉치가 공개되어 이를 기반으로 의존 구문 분석 연구가 진행되었다(Lim et al., 2011; Ahn et al., 2014). <Table 1>은 한국정보통신기술협회(TTA)에서 정의한 의존관계 가이드라인의 기본 원칙이다. Lim et al. (2015) 는 TTA 의존 구문 분석 가이드라인을 참고하여 해외 의존 구문 분석 연구의 결과물을 한국어에서도 활용할 수 있도록 영어권 연구에서 일반적으로 적용되는 단일 지배소 원칙과 투사성 원칙을 적용하여 새로운 의존 관계 가이드라인을 정립하고 이를 기반으로 한 의존 구문 분석 API를 공개하였다.

〈Table 1〉 TTA Standard Dependency Guideline

(1)	자연어처리를 위한 일관성 유지와 효율성 제고에 초점을 주되, 일반 언어학적 관점에서도 크게 벗어나지 않도록 한다.
(2)	문장의 표층 구조를 중시하여 분석한다.
(3)	의존관계 분석의 기본 단위로 어절을 사용한다.
(4)	지배소 후위 원칙에 따라 각 어절의 지배소는 자신보다 뒤에 위치하도록 분석한다.
(5)	각 어절은 1개의 지배소를 가진다.
(6)	각 어절 및 지배소 쌍은 서로 교차하지 않는다.
(7)	보어와 부가어를 구분하되 보어의 범위를 엄격히 제한한다.
(8)	원칙적으로 접속과 내포를 구별하지 않으며, 접속절은 모두 부사절로 분석한다. (다만, 명사구 접속은 인정한다.)
(9)	하나의 주어가 모문과 내포문 모두에 관련되어 있으면 모문과 내포문의 관계에 따라 해당 주어의 지배소를 결정한다.

2.3 구문 분석 결과를 이용한 추출

Shin et al. (2004) 은 질의응답 시스템에서 정답 후보 문장들과 질의 문장의 구문 구조를 비교하여 정답을 추출하였다. (Shen et al., 2006; Shelmanov et al., 2017) 은 두 문장의 구문 형식이 유사한 정도를 계산하는 메트릭을 정의하여 질의 문장과 유사한 구문 형식을 가지는 정답 문장에서 정답을 추출한다. 두 연구들은 문장에 대하여 구문 분석을 수행하거나 의존 관계를 파악하여 그 결과를 질의-응답 시스템에 적용한다는 점에서는 본 연구와 유사하지만 구문 관계 정보나 구문 형식의 유사도를 정의하는 메트릭을 사전에 정의해야한다는 점에서 한계를 갖는다.

(Punyakanok et al., 2004; Yao et al., 2013) 은 질의응답 시스템에서 정답을 추출할 때 질의 문장의 의존 트리와 정답 후보 문장들의 의존 트리로 트리 편집 거리를 계산하였다. 두 연구들은 문장의 의존 구문 분석 결과를 트리 형태로 표현한다는 점에서는 본 연구와 유사하지만 트리 편집 거리를 계산하기 위한 알고리즘의 연산량이 크다는 한계점이 존재한다.

국내에서는 의존 구문 분석 결과를 이용하여 문장 구조의 패턴에 대한 규칙을 정의하고 이를 통하여 두 개체 간의 관계를 표현하는 트리플 구조를 추출하려는 연구가 수행되었다(Kwak et al., 2013; Kim et al., 2015; Hwang et al., 2018). 위의 연구들은 문장에 대하여 의존 구문 분석 결과를 이용하여 문장 안에서 개체를 추출하려는 점에서는 본 연구와 유사하다. 하지만 위의 연구들은 개체 간의 관계를 파악하는 것이 목적인 반면, 본 연구는 질의에 대한 정답을 추출하는 것이 목적이라는 점에서 차이가 있다.

3. 정답 추출 모델과 의존 구문 분석 결과에 대한 자질

본 연구에서는 분석된 질의를 바탕으로 정답 후보 문장을 선택하는 과정은 선행되었다고 가정하고 분석된 질의와 선택된 후보 문장으로부터 질의에 대한 정답을 추출한다. 예를 들어 “하얀 방은 언제 개봉했습니까?”라는 사용자 질의

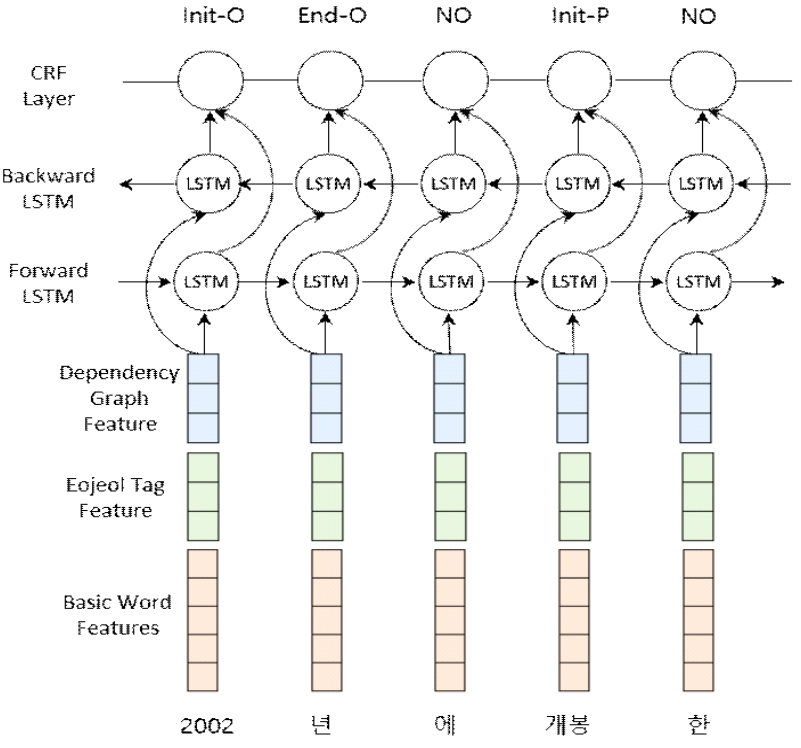
문장을 분석하여 얻은 질의 <하얀 방(주어), 개봉(서술어)>와 분석된 질의로 얻은 후보 문장 “하얀 방은 2002년에 개봉한 대한민국의 스릴러 영화이다”이 주어질 때 문장으로부터 정답 추출 모델을 이용하여 “하얀 방”의 개봉 연도인 “2002년”을 추출한다.

3.1. 장에서는 문장으로부터 정답을 추출하기 위한 모델의 구조와 전체 과정에 대해 설명한다. 3.2. 장에서는 정답 추출 모델의 입력 자질에 대해 설명하는데 3.2.1. 절에서는 베이스라인 실험을 위한 기본 단어 자질, 3.2.2 절에서는 문장 내 어절 태그에 대한 자질, 3.2.3. 절에서는 의존 구문 분석 결과로 생성한 그래프 임베딩 자질에 대한 설명을 한다.

3.1 Bidirectional LSTM-CRF 기반의 정답 추출 모델

본 연구에서는 정답 추출을 위하여 Bidirectional LSTM-CRF(Huang et al., 2015)를 이용하였고 모델의 구조는 <Figure 2>와 같다. Bidirectional LSTM-CRF는 Bidirectional LSTM의 결과와 레이블 인접성 정보를 바탕으로 현재 레이블을 추측하는 방법인 CRF(Conditional Random Fields)가 결합한 네트워크로 시퀀스 태깅에 주로 이용된다.

먼저, 문장을 형태소 분석하여 단어들로 토큰화한 후 모델에 입력하면 임베딩층을 거쳐 단어 임베딩을 구한다. 단어의 임베딩, 단어의 자질,



〈Figure 2〉 Bidirectional LSTM-CRF based Answer Extraction Model Structure

어절 태그 자질, 문장의 의존 그래프 임베딩 자질을 결합한 입력 벡터는 Bidiriectioanl LSTM층과 CRF층을 거쳐 질의 중 주어(S), 질의 중 서술어(P), 정답(O), 나머지(NO) 중 하나의 태그로 결정된다. 이 때 청킹(Chuncking)을 위하여 나머지 태그(NO)를 제외한 각각의 태그 앞에 “Init-”, “Middle-”, “End-”를 추가하였다. 마지막으로 태깅 된 형태소들 중 정답 태그(O)를 포함하는 형태소들을 결합하여 최종 정답을 결정한다.

3.2 자질

본 연구에서는 기본 단어 자질, 어절 태그 자질, 의존 그래프 임베딩 자질을 결합하여 각 형태소의 입력 벡터를 만든다. 자질들 중 단어 임베딩을 포함한 7가지 단어의 정보는 의존 구문 분석 없이도 반영할 수 있는 기본 단어 자질이라고 정의한다. 어절 태그 자질과 의존 그래프 임베딩 자질은 의존 구문 분석을 한 후에 얻을 수 있는 자질이다.

3.2.1 기본 단어 자질

본 연구에서는 정답 추출을 위하여 Choi et al. (2018)을 참고하여 한 형태소에 대해서 7가지의 단어 정보를 기본 단어 자질로 정의하였다. 단어 임베딩, 단어의 형태소 정보, 해당 단어가 문장의 첫 단어 인지에 대한 여부, 해당 단어가 문장의 마지막 단어 인지에 대한 여부, 해당 단어가 숫자 인지에 대한 여부, 서술어에 대한 단위 단어 정보, 질의에 대한 정보이다. 예를 들어 분석된 질의가 <하얀 방(주어), 개봉(서술어)> 일 때 서술어인 “개봉”에 대한 단위 단어는 날짜를 나타내는 “년”, “월”, “일”로 정의하였다. 질의에 대한 정보는 해당 단어가 질의에 해당되는 단어일 경우 주어인지 서술어인지에 대한 정보이다.

3.2.2 어절 태그 자질

본 연구에서는 문장의 의존 구문 분석을 위하여 한국전자통신연구원(ETRI)에서 제공하는 의존 구문 분석 API를 이용하였다. ETRI 의존 구문 분석 결과로 문장 내의 어절들에 대한 태그와 어절들 간의 의존 관계에 대한 정보를 얻을 수 있다. 어절들에 대한 태그는 구문 태그(<Table 2> 참조)와 기능 태그(<Table 3> 참조)가 결합되어 표현된다.

〈Table 2〉 Syntax Tag

NP	체언 (명사, 대명사, 수사)
VP	용언 (동사, 형용사, 보조 용언)
AP	부사구
VNP	긍정 지정사구 (명사 + 이다)
DP	관형사구
IP	감탄사구 (호칭 및 대답 등의 표현)
X	의사 구 (pseudo phrase, 조사 단독 어절 또는 기호 등)
L	부호 (왼쪽 괄호 및 따옴표)
R	부호 (오른쪽 괄호 및 따옴표)

〈Table 3〉 Function Tag

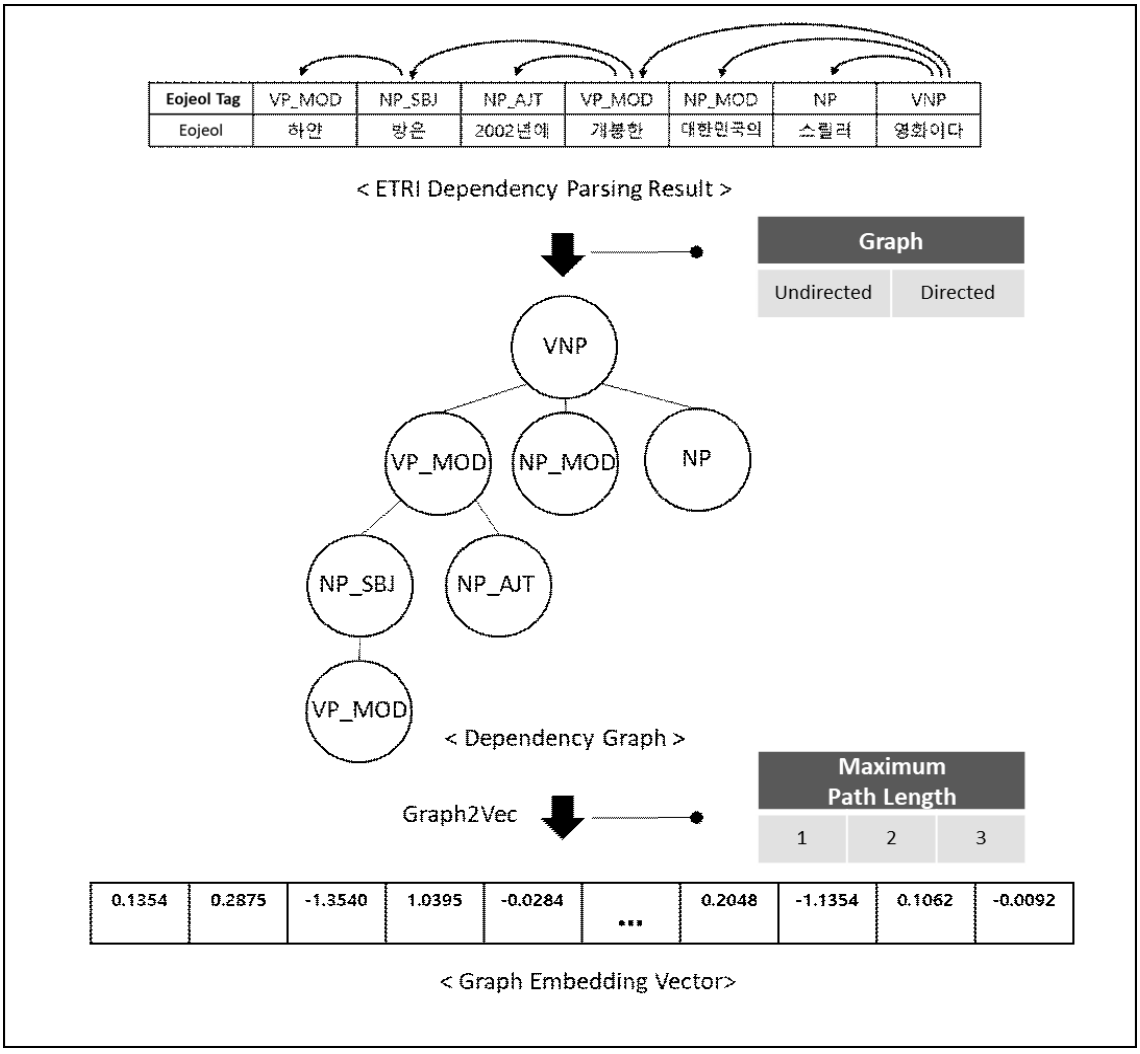
SBJ	주어
OBJ	목적어
MOD	관형어(체언 수식어)
AJT	부사어(용언 수식어)
CMP	보어
CNJ	접속어(~와)

ETRI 의존 구문 분석 API의 결과로 얻은 “하얀 방은 2002년에 개봉한 대한민국의 스릴러 영화이다”라는 문장 내의 어절들의 태그는 “하얀<VP_MOD>”, “방은<NP_SBJ>”, “2002년에<NP_AJT>”, “개봉한<VP_MOD>”, “대한민국의<NP_MOD>”, “스릴러<NP>”, “영화이다<VNP>”이다. 그런데 “2002년에<NP_AJT>”라는 어절은 형태소 분석하면 “2002”, “년”, “에”로 나누어지는데 이러한 경우는 “2002”, “년”, “에”라는 형태소들에 모두 같은 어절 태그 <NP_AJT> 정보를 주었다.

3.2.3 의존 그래프 임베딩 자질

본 연구에서는 의존 구문 분석 결과를 이용하여 의존 그래프 임베딩을 구한다. 의존 그래프 임베딩을 생성하는 과정은 다음과 같다. 먼저, ETRI 의존 구문 분석 결과인 어절들에 대한 태그와 어절들 간의 의존 관계 정보를 이용하여 의존 관계 그래프를 생성한 후 Graph2Vec (Narayanan et al., 2017)을 통하여 그래프의 임베딩을 학습한다. Graph2Vec은 어떤 그래프를 구성하고 있는 부분 그래프들을 추출하여 해당 그래프와 그 그래프를 이루고 있는 부분 그래프들의 쌍으로 학습 데이터를 생성한 후 스킵그램 (Skip-gram) 모델로 학습하여 그래프의 임베딩을 구한다.

<Figure 3>는 예시를 통하여 의존 그래프 임베딩 생성 과정을 보여준다. 먼저, ETRI 의존 구문 분석 결과를 이용하여 의존 관계 그래프를 생성한다. “하얀 방은 2002년에 개봉한 대한민국의 스릴러 영화이다”라는 문장은 7개의 어절로 이루어져 있고 어절 간의 6개의 의존 관계가 존재한다. 각각의 어절을 노드로 하고 어절들 간 의존 관계를 간선으로 하여 그래프를 생성한다. 이때 노드의 라벨은 어절 태그로 한다. “영화이다<VNP>”를 나타내는 노드는 “개봉한<VP_MOD>”, “대한민국의<NP_MOD>”, “스릴러<NP >”와 의존 관계를 가지기 때문에 세 개의 노드와 연결된다. 이렇게 생성된 의존 관계 그래프는 Graph2Vec을 통하여 벡터로 표현된다. 이 때 Graph2Vec에서 그래프에서 추출할 부분 그래프의 최대 경로 길이를 파라미터로 지정할 수 있다. 즉, 최대 경로 길이가 1이면 직접적인 연결 관계만 고려하는 것이고, 최대 경로 길이가 커질수록 간접적인 연결 관계까지 고려한다.



〈Figure 3〉 Dependency Graph Embedding Generation Process

4. 실험

4.1 실험 데이터

본 연구에서는 문서 정보의 신뢰성을 보장하기 위하여 문서의 소스를 위키피디아, 네이버 백과사전, 네이버 뉴스로 한정하였다. 각 웹 문서 소스의 검색 API에 질의를 검색하여 질의와 관련된 문서들을 수집하였다. 위키백과와 네이버 지식백과에서는 질의의 주어만 일치 조건을 주어 검색하였고, 네이버 뉴스에서는 질의의 주어와 서술어를 일치 조건을 주어 검색하였다. 예를 들어, 분석된 질의가 (하얀 방, 개봉)일 때 위키

〈Table 4〉 Examples of Experiment Data

Query		Sentence	Answer
Subject	Predicate		
하얀 방	개봉	하얀 방은 2002년에 개봉한 대한민국의 스릴러 영화이다 (The White Room was a Korean thriller film released in 2002.)	2002년
UCCU 센터	위치	UCCU 센터는 미국 유타 주 오렘에 위치한 실내 경기장이다 (The UCCU Center is an indoor stadium in Orem, Utah, USA.)	미국 유타 주 오렘

백과와 네이버 지식백과에서는 “하얀 방”을 검색하였고, 네이버 뉴스에서는 “하얀 방”, “개봉”을 검색하였다. 그 이유는 백과 문서의 경우 문서가 키워드를 묘사하는 문서이기 때문에 해당 주어만 키워드로 주어도 질의에 대한 문서를 구할 수 있다. 반면 뉴스 문서의 경우 여러 키워드가 혼재된 문서이기 때문에 주어와 서술어로 키워드에 제약을 주어 질의와 보다 더 관련 있는 문서를 가져왔다.

3개 소스로부터 수집된 문서 중 주어, 서술어, 정답을 모두 포함하고 있는 문장 11,877개를 실험 데이터로 사용하였다. 이 중 학습(Train) 데이터는 9,501개, 검증(Validation) 데이터와 테스트(Test) 데이터는 각각 1,188개이다. 실험 데이터의 예시는 <Table 4>와 같다.

4.2 실험 방법

본 연구에서는 의존 구문 분석 결과를 반영할 때 정답 추출의 성능이 향상되는지를 확인하기 위하여 기본 단어 자질만 입력으로 했을 때의 정답 추출의 성능과 기본 단어 자질, 의존 구문 분석 결과를 이용한 자질(어절 태그 자질, 문장의 의존 그래프 임베딩 자질)을 입력으로 했을 때의 정답 추출의 성능을 비교하였다. 의존 구문 분석 결과로 문장의 의존 관계 그래프를 생성할 때 관계의 방향성을 고려하지 않은 무방향 그래프와 관계의 방향성을 고려한 방향 그래프로 나누어 실험하였다. 추가적으로 Graph2Vec 내 부분 그래프 추출하는 과정에서 부분 그래프의 노드 간 최대 경로 길이를 지정할 수 있는데 최대 경로 길이를 1부터 3까지 증가시키면서 실험하였다. 부분 그래프의 노드 간 최대 경로 길이가 1일 때는 문장 내의 어절들의 직접적인 의존 관계만을 고려한 것이고, 2 이상일 때는 간접적인 의존 관계까지 고려한 것이다.

4.3 실험 결과

테스트 데이터에 대하여 입력 자질에 따른 정답 추출의 성능은 <Table 5>와 같다. 기본 단어 자질만 입력으로 했을 때인 정답 추출의 베이스라인 성능은 70.11%이고, 기본 단어 자질에 어절 태그 자질만 추가했을 때의 성능은 70.95%로 0.84% 향상되었다. 기본 자질에 문장의 의존 관계 그래프 벡터 자질만 추가했을 때도 조금씩 성능이 향상되었는데 부분 그래프의 최대 경로 길이가 1일 때 무방향 그래프는 1.52%, 방향 그래프는 1.85% 향상되었다. 기본 단어 자질에 어절 태그 자질과 의존 관계 그래프 임베딩 자질 모두 추가했을 때는 방향성 그래프의 부분 그래프 최대 경로가 1일 때 73.23%로 3.12% 성능이 향상

〈Table 5〉 Accuracy of Answer Extraction on the Test Dataset

	Maximum Path Length	Eojeol Tag Feature (Not Included)	Eojeol Tag Feature (Included)
Basic Word Features		70.11% (833/1,188)	70.95% (843/1,188)
Basic Word Features + Dependency Graph Embedding Feature (Undirected)	**1**	71.63% (851/1,188)	71.04% (844/1,188)
	2	71.21% (846/1,188)	71.29% (847/1,188)
	3	71.04% (844/1,188)	70.87% (842/1,188)
Basic Word Features + Dependency Graph Embedding Feature (Directed)	**1**	71.96% (855/1,188)	73.23% (870/1,188)
	2	71.12% (845/1,188)	71.46% (849/1,188)
	3	70.11% (833/1,188)	71.12% (845/1,188)

되었다.

<Table 6>과 <Table 7>은 의존 구문 분석 결과를 반영하지 않았을 때와 반영하였을 때의 정답 추출 결과를 구체적으로 비교한 결과이다.

<Table 6>은 기본 단어 자질만 반영했을 때는 정답 추출을 올바르게 했지만 의존 구문 분석 결과에 대한 자질을 추가해서 오히려 맞추지 못한 경우이다. 예시 문장들을 확인해보니 모두 잘못된 의존 구문 분석 결과를 보였다. 첫 번째 문장 '이디 아민 다다 오우메는 우간다의 군인 출신 정치인으로 1971년 군사 쿠데타로 대통령에 취임하였다'의 경우는 '이디 아민 다다 오우메'는 사람 이름으로 어절들의 태그가 "이디<NP>", "아민<NP>", "다다<NP>", "오우메는<NPSBJ>" 여야 하는데 "다다<AP>"로 되어 있었다. 두 번째 문장 '영원의 제로오타 출판,2006년 8월 23일,448 페이지'의 경우는 의존 구문 분석은 띄어쓰기 단위인 어절 단위로 분석 하는데 띄어쓰기 오류로 인하여 잘못된 의존 구문 분석 결과를 보였다. 현재 의존 구문 분석기 성능의 한계로 <Table 6>과 같이 문장들의 잘못된 의존 구문 분석 결과들이 오히려 노이즈가 될 수도 있다. 즉, 정답 추출의 성능이 의존 구문 분석기 성능에 의존적이기 때문에 추후 고도화된 의존 구문 분석기가 필요할 것으로 보인다.

<Table 7>은 기본 단어 자질만 반영했을 때는 정답 추출을 제대로 하지 못했지만 의존 구문 분석 결과에 대한 자질을 추가하였을 때 제대로 정답을 맞춘 경우이다. 본 연구의 정답 추출 모델인 Bi-directional LSTM-CRF는 시퀀스의 정보를 반영하는데 유리한 모델로, 질의와 정답이 인접하고 문장의 길이가 짧은 경우에는 정답을 비교적 잘 추출하지만 문장의 길이가 길어지는 경우에는 정답을 추출하기가 어렵다는 한계점이 있

〈Table 6〉 Wrong Result of Answer Extraction 1

Query		Answer	Sentence	Result (Basic Word Features + Dependency Features)
Subject	Predicate			
이디 아민	취임	1971년	이디 아민 다다 오우메는 우간다의 군인 출신 정치인으로 1971년 군사 쿠데타로 대통령에 취임하였다	X
영원의 제로	페이지	448	영원의 제로오타 출판, 2006년 8월 23일, 448 페이지	X

〈Table 7〉 Wrong Result of Answer Extraction 2

Query		Answer	Sentence	Result (Basic Word Features)
Subject	Predicate			
아이맥 G3	개발	Apper	아이맥 G3는 1998년부터 2003년까지 Apper에 의해 개발, 제조, 판매된 개인용 데스크톱 컴퓨터이다	제조
조류인간	감독	신연식	조류인간은 2015년에 개봉한 대한민국의 영화로, 신연식의 감독 전작 러시안 소설의 가상 소설인 조류인간을 영화로 재탄생시킨 작품이다	러시안

다. 올바르게 의존 구문 분석이 된 경우에 의존 구문 분석에 대한 자질을 추가한 결과 <Table 7>과 같이 문장의 길이가 길고 복잡하거나 질의에 포함된 단어와 정답의 거리가 먼 경우에 대체로 더 나은 성능을 보였다.

5. 결론

본 연구에서는 질의응답 시스템 내 정답 추출의 성능을 향상시키기 위하여 문장의 의존 구문 분석 결과를 이용한 다양한 자질을 정답 추출 모델에 반영하여 실험하였다. 정답 추출 모델로 Bidirectional LSTM-CRF를 이용하였고 단어 임베딩을 포함한 8가지 단어의 정보는 의존 구문 분석 없이도 반영할 수 있는 기본 단어 자질로 정의하였다. 의존 구문 분석 결과로는 어절 태그 자질과 의존 그래프 임베딩 자질을 생성하였다. 어절 태그 자질은 ETRI 의존 구문 분석 결과 중 어절들에 대한 태그 정보이다. 의존 그래프 임베딩 자질은 ETRI 의존 구문 분석 결과인 어절들에 대한 태그 정보와 어절들 간의 의존 관계 정보를 바탕으로 의존 그래프를 생성한 후 Graph2Vec을 통하여 구한 그래프 임베딩이다.

본 연구에서는 의존 구문 분석 결과를 반영할 때 정답 추출의 성능이 향상되는지를 확인하기 위하여 기본 단어 자질만 입력으로 했을 때인 정답 추출의 성능과 의존 구문 분석 결과를 이용한

자질(어절 태그 자질, 의존 그래프 임베딩 자질)을 추가로 입력으로 했을 때의 정답 추출 성능을 비교하였다. 의존 그래프 임베딩 자질은 의존 관계의 방향성 고려 여부에 따라 노드 간 최대 경로의 길이를 1부터 3까지 조정하며 실험하였다.

실험 결과 어절 태그와 의존 관계 그래프 임베딩 자질 모두 정답 추출의 성능을 향상시키는 것을 확인하였고 두 자질을 모두 추가하고 방향성을 고려하며 부분 그래프의 노드 간 최대 경로는 1일 때 73.23%로 가장 높은 성능을 보였다. 이는 기본 단어 자질만 입력으로 했을 때인 베이스라인 성능 70.11%보다 3.12% 상승한 수치이다. 이는 정답 추출의 성능 향상을 위해서는 의존 관계의 방향성을 고려하고 어절 간의 직접적인 의존 관계만을 고려하는 것이 더 좋다는 것을 의미한다.

연구가 갖는 의의는 다음과 같다. 첫째, 어순 구조가 자유롭고 문장 구성 성분의 생략이 빈번한 한국어의 특성을 고려하여 의존 구문 분석 결과를 이용한 자질을 추가해서 정답 추출의 성능을 향상시켰다. 의존 구문 분석을 위하여 ETRI에서 제공하고 있는 API를 이용하였는데 향후에 더 나은 성능의 API가 개발된다면 정답 추출의 성능이 더욱 향상될 것을 기대한다. 둘째, 어절 간 의존 관계에 관한 패턴을 사전에 정의하지 않고 학습 기반의 그래프 임베딩 방식으로 의존 구문 분석 결과에 대한 자질을 생성하였다. 본 연구에서 제안하고 있는 학습 기반의 그래프 임베딩 방식은 구문 관계 정보나 구문 형식의 유사도를 정의하는 메트릭을 사전에 정의해야 하는 기존 연구의 한계점을 극복했다는 점에서 의의가 있다.

연구가 갖는 한계점은 다음과 같다. 첫째, 의존 구문 분석기 성능의 한계이다. 본 연구에서는 의존 구문 분석기를 이용하여 의존 구문 분석이 이루어지기 때문에 의존 구문 분석기의 성능에 따라서 정답 추출 성능에 큰 영향을 받는다. 추후 연구에서는 오픈소스가 아닌 고도화된 의존 구문 분석기를 개발하여, 정답 추출의 성능을 더욱 향상시킬 수 있을 것을 기대한다. 둘째, 본 연구에서는 문장 내의 동명이인이나 동음이의어에 대한 구분을 할 수 없어 질의 분석의 결과가 잘못될 가능성이 있다. 때문에 동명이인이나 동음이의어를 처리하는 방법을 모색하여 적용할 필요가 있다.

향후 연구 방향은 다음과 같다. 본 연구에서는 의존 구문 분석 결과로 생성한 자질을 의미를 파악하기 위하여 정답 추출 모델에만 적용시켰다. 하지만 향후에는 감성 분석이나 개체명 인식과 같은 다양한 자연어 처리를 위한 모델에 해당 자질을 적용하여 성능을 확인한다면 자질의 타당성을 보다 정확하게 검증할 수 있을 것이다.

참고문헌(References)

Abney, S., M. Colins, A. Singhal, “Answer Extraction”, *Proceedings of the Sixth Conference on Applied Natural Language Processing*, (2000), 296~301.

Ahn, K. M. and Y. H. Seo, “A Korean Dependency Parsing Algorithm using Sets of Head Candidates”, *Journal of KISS : Software and Applications,* Vol.41, No.1 (2014), 88~95.

Choi, H. S., M. T. Kim, W. J. Kim, D. W. 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.

Doan-Nguyen, H., and L. Kosseim: "Improving the Precision of a Closed-Domain Question-Answering System with Semantic Information", *Coupling approaches, coupling media and coupling languages for information retrieval*, (2004), 850~859.

Huang, Z., X. Wei, and Y. Kai,. "Bidirectional LSTM-CRF models for sequence tagging", *arXiv preprint arXiv: 1508.01991*, (2015).

Hwang, H. S., J. S. Bae and C. K. Lee, "Korean Open Information Extraction using Dependency Parsing and Semantic Role Labeling", *Proceedings of Korean Information Science Society*, No.12(2018), 563-565.

Ittycheriah, A., M. Franz, W. Zhu, and A. Ratnaparkhi, "IBM's Statistical Question Answering System", *In 9th Text Retrieval Conference*, (2000), 229~334.

Kawahara, D., N. Kaji, and S. Kurohashi, "Question and answering system based on predicate-argument matching", *Proceedings of the Third NTCIR*, (2002), 21~24.

Kim, B. S., H. J. Yu and G. B. Lee, "A Syntax-Based Hybrid System for Korean Open Information Extraction", *The 27th Annual Conference on Human & Cognitive Language Technology,* (2015), 41~45.

Kwak, S. J., B. G. Kim and J. S. Lee, "Tiplet Extraction using Korean Dependency Parsing Result", *The 25th Annual Conference on Human & Cognitive Language Technology*, (2013), 86~89.

Kwon, H. and J. Y. Choi, "A Korean Language Parser with a Unification Based Dependency Grammar", *The Journal of Korea Information Science Society,* Vol.19(1992), 467~476.

Lim, J. H., Y. J. Bae, H. K. Kim, Y. J. Kim and K. C. Lee, "Korean Dependency Guidelines for Dependency Parsing and Exo-Brain Language Analysis Corpus", *The 27th Annual Conference on Human & Cognitive Language Technology*, (2015), 234~239.

Lim, S. J., Y. T. Kim and D. Y. Ra, "Korean Dependency Parsing Based on Machine Learning of Feature Weights", *Journal of KIISE: Software and Applications*, Vol.38, No.4(2011), 214~223.

McDonald, R., F. Pereira, K. Ribarov, and J. Hajic, "Non-projective Dependency Parsing using Spanning Tree Algorithms", *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing,* (2005), 523~530.

Mendes, A. C., and L. Coheur, "An approach to answer selection in question-answering based on semantic relations", *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, (2011), 1852~1857.

Narayanan, A., M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal, "graph2vec: Learning distributed representations of graphs", *arXiv preprint arXiv:1707.05005,* (2017).

Nivre, J. "Incrementality in deterministic dependency parsing", *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*, (2004), 50~57.

Punyakanok, V., D. Roth, and W. Yih, "Mapping dependency trees: An application to question answering", *The 8th International Symposium on Artificial Intelligence and Mathematics*,

(2004).

Ravichandran, D and E. Hovy, “Learning surface text patterns for a question answering system”. *Proceedings of the 40th annual meeting on association for computational linguistics*, (2002), 41~47.

Ravichandran, D., I. Abharam, and R. Salim, “Automatic derivation of surface text patterns for a maximum entropy based question answering system”. *Proceedings of the Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics* (2003).

Shelmanov, A., M. Kamenskaya, M. Ananyeva, and I. Smirnov, “Semantic-syntactic analysis for question answering and definition extraction”, *Scientific and Technical Information Processing,* Vol.44, No.6(2017), 412~423.

Shen D., and D. Klakow. “Exploring correlation of dependency relation paths for answer extraction”. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, (2006), 889~896.

Shin, H. P., “Maximally Efficient Syntactic Parsing with Minimal Resources”, *The 11th Annual Conference on Human & Cognitive Language Technology,* (1999), 242~248.

Shin, S. E., D. Y. Yi and Y. H. Seo, “Korean Question-Answering System using Syntactic-Relation Information”, *Journal of the Korea Contents Association,* Vol.4, No.2 (2004), 36~42.

Soubbotin, M. M. and S. M. Soubbotin, “Patterns for potential answer expressions as clues to the right answers”, *Proceedings of the 10th Text REtrieval Conference*, (2001).

Yao, X., B. Van-Durme, C. Callison-Burch, and P. Clark, “Answer extraction as sequence tagging with tree edit distance”, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (2013), 858~867.

Yen, S. J., Y. C. Wu, J. C. Yang, Y. S. Lee, C. J. Lee, and J. J. Liu, “A support vector machine-based context-ranking model for question answering”, *Information Sciences,* Vol.224(2013), 77~87.

Yu, L., K. M. Hermann, P. Blunsom, and S. Pulman. “Deep learning for answer sentence selection”, *arXiv preprint arXiv:1412.1632*, (2014).

Abstract

Query-based Answer Extraction using Korean Dependency Parsing

Dokyoung Lee* · Mintae Kim** · Wooju Kim***

In this paper, we study the performance improvement of the answer extraction in Question-Answering system by using sentence dependency parsing result.

The Question-Answering (QA) system consists of query analysis, which is a method of analyzing the user's query, and answer extraction, which is a method to extract appropriate answers in the document. And various studies have been conducted on two methods. In order to improve the performance of answer extraction, it is necessary to accurately reflect the grammatical information of sentences. In Korean, because word order structure is free and omission of sentence components is frequent, dependency parsing is a good way to analyze Korean syntax. Therefore, in this study, we improved the performance of the answer extraction by adding the features generated by dependency parsing analysis to the inputs of the answer extraction model (Bidirectional LSTM-CRF).

The process of generating the dependency graph embedding consists of the steps of generating the dependency graph from the dependency parsing result and learning the embedding of the graph. In this study, we compared the performance of the answer extraction model when inputting basic word features generated without the dependency parsing and the performance of the model when inputting the addition of the Eojeol tag feature and dependency graph embedding feature.

Since dependency parsing is performed on a basic unit of an Eojeol, which is a component of sentences separated by a space, the tag information of the Eojeol can be obtained as a result of the dependency parsing. The Eojeol tag feature means the tag information of the Eojeol.

The process of generating the dependency graph embedding consists of the steps of generating the dependency graph from the dependency parsing result and learning the embedding of the graph. From the

* Department of Industrial Engineering, Yonsei University
** 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 YONSEI UNIVERSITY College of Engineering D901
Tel: +82-2-2123-7754, E-mail: wkim@yonsei.ac.kr

dependency parsing result, a graph is generated from the Eojeol to the node, the dependency between the Eojeol to the edge, and the Eojeol tag to the node label. In this process, an undirected graph is generated or a directed graph is generated according to whether or not the dependency relation direction is considered. To obtain the embedding of the graph, we used Graph2Vec, which is a method of finding the embedding of the graph by the subgraphs constituting a graph. We can specify the maximum path length between nodes in the process of finding subgraphs of a graph. If the maximum path length between nodes is 1, graph embedding is generated only by direct dependency between Eojeol, and graph embedding is generated including indirect dependencies as the maximum path length between nodes becomes larger.

In the experiment, the maximum path length between nodes is adjusted differently from 1 to 3 depending on whether direction of dependency is considered or not, and the performance of answer extraction is measured. Experimental results show that both Eojeol tag feature and dependency graph embedding feature improve the performance of answer extraction. In particular, considering the direction of the dependency relation and extracting the dependency graph generated with the maximum path length of 1 in the subgraph extraction process in Graph2Vec as the input of the model, the highest answer extraction performance was shown. As a result of these experiments, we concluded that it is better to take into account the direction of dependence and to consider only the direct connection rather than the indirect dependence between the words.

The significance of this study is as follows. First, we improved the performance of answer extraction by adding features using dependency parsing results, taking into account the characteristics of Korean, which is free of word order structure and omission of sentence components. Second, we generated feature of dependency parsing result by learning - based graph embedding method without defining the pattern of dependency between Eojeol. Future research directions are as follows. In this study, the features generated as a result of the dependency parsing are applied only to the answer extraction model in order to grasp the meaning. However, in the future, if the performance is confirmed by applying the features to various natural language processing models such as sentiment analysis or name entity recognition, the validity of the features can be verified more accurately.

Key Words : Question Answering System**,** Answer Extraction, Dependency Parsing, Graph Embedding, Bi-directional LSTM-CRF

Received : June 4, 2019 Revised : June 27, 2019 Accepted : July 3, 2019
Publication Type : Conference(Fast-track) Corresponding Author : Wooju Kim

저 자 소 개

이 도 경
연세대학교 산업공학에서 석사과정 재학 중이다. 주요 관심 분야는 머신러닝, 딥러닝을 활용한 자연어 처리이다. 지능정보시스템학회에서 발표한 바 있다.

김 민 태
연세대학교 산업공학에서 통합과정 재학 중이다. 주요 관심 분야는 머신러닝, 딥러닝을 활용한 자연어 처리, 추천 시스템 등이다. 한국정보과학회, 지능정보과학회, 지능정보시스템학회 등 국내, 국외 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.