

질의 유형 분류기를 활용한 지식 베이스 기반의 복합 질의 응답 시스템

홍동균, 심홍매, 최동근, 김광민, 정용일, 이반 베를로체
(주)솔트룩스

e-mail: dghong@saltlux.com

Knowledge based Complex Question Answering System using CNN Based Question Type Classifier

Dong-Gyun Hong, Hong-Mei Shen, Dong-Geun Choi, Kwang-Min Kim,
Yong-Il Jeong, Ivan Berlocher Kim
Saltlux Inc.

요 약

최근 지식 베이스의 발전과 함께 지식 베이스 기반의 질의 응답에 관한 연구가 많은 관심을 받고 있다. 특히 지식 베이스상의 여러 개의 사실이 필요한 복합 질의에 대한 처리의 중요성이 높아지고 있다. 그러나 기존 연구에서는 일반적인 지식을 묻는 질의 처리에만 집중하여, 그 외의 다른 유형을 갖는 복합 질의에 대한 처리의 연구는 시작 단계에 머물러 있다. 이에 본 논문은 질의 유형 분류기를 활용한 지식 베이스 기반의 복합 질의 응답 시스템을 제안한다. 복합 질의 응답 시스템은 단순 질의를 포함하여 다양한 유형(일반형, 판정형, 비교형)을 갖는 복합 질의를 처리한다. 우리는 실험을 통해서 질의 유형 분류기가 복합 질의 응답 시스템의 정답률을 높임을 보였다.

1. 서론

최근 지식 베이스 기반 질의 응답(KBQA)은 자연어 질의에 대하여 지식 베이스에 존재하는 사실을 기반으로 답을 도출하는 것이다. 대표적인 지식 베이스로는 Google Knowledge Graph¹, DBpedia² 등이 있으며, 이러한 지식 베이스에 대한 KBQA 연구도 활발히 진행되고 있다[1,2,3].

지식 베이스는 지식의 집합으로 지식은 개체 간의 관계(속성) 혹은 개체와 그 개체가 가진 특정 값 사이의 관계를 사실로 표현한 것이다. 그리고 하나의 사실은 <주어, 술어, 목적어> 트리플 형태를 가진다. 예를 들면 “대한민국 수도는 서울”이라는 하나의 사실에서 주어는 대한민국이고 술어는 수도이며 목적어는 서울이다.

KBQA는 사용자의 자연어 질의로부터 의도를 파악하고 그 의도에 맞는 답을 도출하기 위하여 지식베이스 내에 존재하는 사실들을 이용한다. 이때 자연어 질의를 적절한 SPARQL 질의 언어로 변환하여 사용한다.

하지만 국내의 KBQA 관련 연구에서는 한 개의 Fact로 답변이 가능한 단순 질의에 대한 응답을 처리하는 게 대부분이다[1,2]. 여러 개의 Fact를 사용해야 하는 복합 질의 응답에 대한 연구는 시작 단계에 머물

러 있기 때문에 어려움이 있다[3,4].

본 논문에서는 복합 질의 처리를 위해 하나의 통합 시스템을 제안한다. 이 시스템은 질의유형 분류기를 활용하여 복합 질의의 유형을 파악하여 처리하는 시스템으로 구성되어 있다. 제안하는 QA 시스템에서 처리하는 복합 질의를 아래와 같이 일반형, 판정형 그리고 비교형 세가지 형태로 나눴다. 그 중 일반형에는 단순 질의도 포함되어 있다.

<표 1> 복합 질의 예시

질의유형	질의 예시	패턴
단순 질의	“대한민국(Entity)의 수도(Property)는?”	EP
일반형 복합 질의	“서울(Entity) 중구(Entity)의 인구(Property)는?”	EEP
판정형 복합 질의	“서울(Entity)은 대한민국(Entity)의 수도(Property)야?”	EEP
비교형 복합 질의	“중구(Entity)랑 강남구(Entity) 중에 인구(Property)가 더 많은 곳은?”	EEP

위와 같이 복합 질의의 경우 개체 2개와 속성 1개인 동일한 패턴이지만, 서로 다른 의도를 내포하고 있기에 의도마다 다른 처리 방식을 요구한다. 따라서 논문에서 제안하는 통합시스템에서는 이를 해결하기 위해 복합 질의의 유형을 판단할 수 있는 분류기를 사용하였고, 분류된 결과에 적합한 QA 모듈을 사용하여 자연어 질의를 처리하였다.

본 논문의 구성은 2장에서 질의유형을 분류하는 CNN 기반 질의유형 분류기 및 복합 질의응답 모듈인

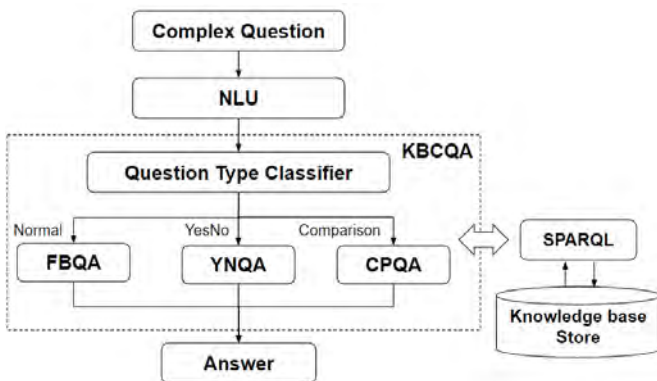
¹ <https://developers.google.com/knowledge-graph/>

² <https://wiki.dbpedia.org/>

일반형 질의 처리 모듈, 판정형 질의 처리 모듈, 비교형 질의 처리 모듈을 소개한다. 3 장에서는 실험 및 결과에 대해 토의하며, 마지막으로 4 장에서는 본 논문에 대한 결론 및 향후 연구방향에 대하여 제안한다.

2. 지식베이스 기반 복합 질의 응답 시스템 (KBCQA)

2.1 복합 질의 응답 프로세스



(그림 1) 지식베이스 기반 복합 질의 응답 처리 과정

본 논문에서 제안하는 지식베이스 기반의 복합 질의 응답 시스템의 구조는 (그림 1)과 같다. 질의 응답 처리 과정은 아래 세 단계로 진행된다.

1 단계) 자연어 이해(NLU: Natural Language Understanding)

자연어 질의에 대하여 형태소 분석, 개체명 인식, 구문 분석 등을 수행하고, 이 결과를 이용하여 자연어 질의로부터 개체, 속성, 클래스와 같은 시맨틱 정보를 인식한다.

2 단계) 질의유형 분류

1 단계의 NLU 결과를 바탕으로 질의유형 분류기 (Question Type Classifier)를 통해 해당 질의유형이 ‘일반형’ 인지 ‘판정형’ 인지 ‘비교형’ 인지를 판단한다. 판단된 질의는 유형 별로 해당 질의 처리 모듈로 보낸다.

3 단계) 지식베이스 기반 복합 질의 응답 처리

KBCQA 에는 총 3 개의 질의 처리 모듈로 구성되어 있다.

먼저 FBQA 는 일반형 질의 응답 모듈로써, 단순 질의와 일반형 복합 질의를 처리한다. 그리고 YNQA 모듈과 CPQA 모듈은 각각 판정형 복합 질의와 비교형 복합 질의를 처리하는 모듈이다.

각 KBCQA 모듈의 질의 처리 과정에서는 공통적으로 시맨틱 정보를 조합하여 시맨틱 패턴을 얻는 과정이 포함되어 있다. 예를 들어 “대한민국의 수도는?” 이라는 질문의 시맨틱 정보는 “대한민국” (개체), “수도” (속성)이고, 이를 조합하여 “EP” 라는 시맨틱 패턴을 얻는다. 이후 지식베이스에 시맨틱 패턴에 대응하는 SPARQL 질의를 실행하여 결과를 얻는다.

2.2 합성곱 신경망을 이용한 질의유형 분류

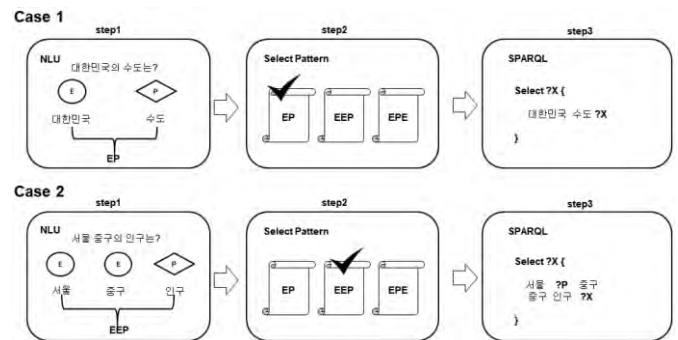
통합 시스템에서 질의유형 분류는 자연어 질의의 유형이 ‘일반형’ 인지 ‘판정형’ 인지 ‘비교형’ 인지를 판단하여 각각의 처리 모듈로 넘겨주는 역할을 한다. 이 때 사용하는 질의유형 분류기는 합성곱 신경망(Convolutional neural network)을 이용했다. 합성곱 신경망은 문장 내 단어의 위치 정보를 보존해주는 특징이 있어 질의유형에 따라 패턴을 잘 분류해주는 장점이 있다. 본 논문에서는 KIM[5]의 합성곱 신경망 네트워크를 이용하여 질의유형 분류기를 구축하였다. 이 네트워크는 NLU 의 결과를 입력으로 사용하여 질의유형을 판단한다. 신경망의 학습을 위한 데이터는 아래와 같이 2 가지 방법으로 구축하였다.

- Case 1: “대한민국 수도”
- Case 2: “대한민국/NNP 수도/NNG”

Case1 은 자연어 질의에 대하여 형태소 분석 결과 중 불필요한 조사를 제외한 데이터이고, Case2 는 case1 의 결과에 특징을 늘려주기 위해 형태소 Tag 를 붙인 데이터이다. 신경망에 입력으로 들어가는 벡터는 Word2Vec 과 같은 방법을 사용하지 않고 학습에서 사용된 단어들로 Lookup 테이블을 만들어 문장 분류를 잘할 수 있는 방식으로 업데이트 되도록 정하였다.

2.3 지식베이스 기반 복합 질의 응답 처리(KBCQA)

2.3.1 사실 기반 질의 응답 (FBQA)



(그림 2) 사실기반 질의응답(FBQA) 처리 예시

지식베이스 기반 복합 질의 응답 처리 시스템의 일반형 질의 처리는 (그림 2)와 같이 진행된다.

- Case 1: “대한민국의 수도는?”
- Case 2: “서울 중구의 인구는?”

사실 기반 질의 처리 모듈에서는 단순질의와 일반형 복합 질의를 처리한다. 첫번째 단계에서는 NLU 분석 결과인 시맨틱 정보를 조합하여 시맨틱 패턴을 생성한다. case 1 은 대한민국이란 하나의 개체와 인구라는 하나의 속성이 인식되어 EP 패턴임을 알 수 있고, case 2 는 서울과 중구와 같은 두개의 개체와 인구라는 하나의 속성을 인식되어 EEP 패턴이다. 그 후

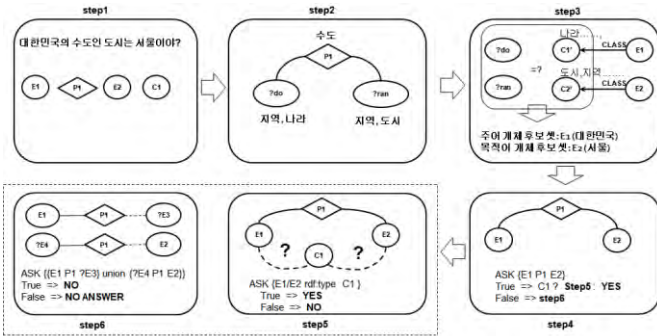
인식된 패턴을 바탕으로 최종 단계에서 SPARQL 질의를 생성한다.

2.3.2 판정형 질의 응답

판정형 질의 응답은 자연어 질의의 내포된 사실 관계(정보)를 지식 베이스의 사실 기반으로 검증하여 맞는지 (예), 틀렸는지 (아니오) 를 응답하는 것을 말한다.

하지만 지식베이스 기반의 판정형 질의 응답 처리는 지식베이스 내의 정보 부재로 인하여 답변을 하지 못하는 경우도 있다. 본 시스템에서 이런 경우에는 최종 답변을 “예/아니오” 가 아닌 “답변불가” 로 처리하였다.

판정형 질의 응답 처리 과정은 아래 (그림 3)과 같이 크게 2 개의 부분으로 나눌 수 있다. 첫번째는 질의 내 내포된 정보를 정확히 식별하는 부분으로 그림에서는 step1 부터 step3 절차이다. 두번째는 정확히 식별된 정보에 대해 지식 베이스의 사실 기반으로 검증하여 답변을 하는 부분으로서 그림에서는 step4 부터 step6 절차이다.



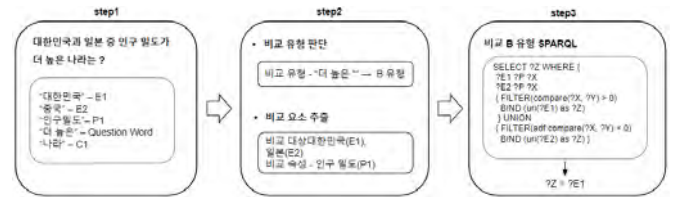
(그림 3) 판정형 질의응답(YNQA) 처리 예시

각 단계별 상세한 프로세스는 다음과 같다. (그림 3)의 step 1)은 사용자 질의에 대하여 NLU 분석을 통해 얻은 개체, 속성, 클래스 등을 포함한 시맨틱 정보이다. step 2)에서는 속성의 주어와 목적어의 클래스인 Domain, Range 를 인식하고 step 3)에서는 질문 내의 개체들의 클래스를 인식해서 step 2)에서 인식한 Domain, Range 에 속하는 후보주어인 개체 (E1)와 후보목적어인 개체(E2)를 구한다. 이를 통해 후보 리스트를 줄이고 성능을 향상시킬 수 있다. step 4)에서는 후보 리스트의 조합을 통해 SPARQL ASK 질의를 하여 결과 값이 True 일 경우 질문 내의 클래스정보(C1)가 없다면 최종 답변을 [YES]로 하고 클래스 정보가 있으면 step 5)에서 클래스 정보와 질문 내의 개체들(E1, E2)과의 연결성을 체크하여 연결이 있을 경우 최종 답변을 [YES]로, 연결이 없을 시에는 [NO]로 한다. 그리고 step4)의 결과가 False 일 경우에는 step 6)를 수행한다. step 6)는 후보리스트 조합에 대하여 SPARQL ASK 질의를 통해 사용자 질의 내 포함된 정보가 지식베이스 내에 존재하는지를 체크하여 최종 답변을 [NO] 혹은 [NO ANSWER]로 할지 판단한다.

2.3.3 비교형 질의 응답

본 논문에서 정의하는 비교형 질의 응답이란, 지식 베이스 상의 두 개체의 속성을 비교하는 질의에 대하여 응답을 하는 것을 의미한다.

비교형 질의 응답의 처리는 비교 마이닝 시스템의 일환으로 볼 수 있다. 기존 연구[6]에서 비교 마이닝은 1) 비교 문장인지 아닌지를 식별, 2) 비교 유형 분류, 3) 비교 요소 추출 등의 단계를 거치게 된다. 본 논문에서도 이러한 절차를 통해 비교 질의 응답 처리를 수행하고자 하였다. 아래는 비교형 질의 응답 처리 과정을 간략하게 보여준다.



(그림 4) 비교형 질의 응답(CPQA) 처리 예시

비교 문장의 식별은 앞서 설명한 질의 유형 분류기에서 수행한다. 비교형 질의의 경우 보통 (그림 4)의 step1 의 예시 질의처럼 ‘중’ 이나 ‘더’ 와 같은 비교를 암시하는 단어가 주로 등장한다. 이러한 특징이 질의 유형 분류기가 비교형 질의를 다른 질의와 구분하도록 한다.

다음으로 비교형 질의에는 두 가지 유형이 존재한다. 하나는 ‘더 낮은’, ‘더 작은’ 등 속성의 값이 더 작은 개체를 답으로 요구하는 질의가 있으며, 반대로 더 큰 속성 값을 갖는 개체를 요구하는 질의 유형이 존재한다. 편의상 전자를 비교 A 유형, 후자를 비교 B 유형이라고 하겠다. Step2 에서 예시 질의는 ‘더 높은’ 이라는 단어가 등장하기 때문에 비교 B 유형의 질의로 판단하였다. 그리고 Step2 의 비교 요소 추출 결과를 보면 ‘대한민국’ 과 ‘일본’ 이라는 두 개체에 대하여 ‘인구 밀도’ 라는 속성을 비교하는 질의임을 파악할 수 있다.

최종 단계인 step3 에서는 비교 요소들에 대하여 B 유형의 비교를 수행하는 SPARQL 을 생성하고, 이를 지식 베이스에 질의하여 원하는 결과를 얻는다.

3. 실험 및 결과

3.1 실험 환경

복합 질의 응답 시스템을 평가하기 위하여 지식베이스의 사실을 기반으로 300 건의 질의를 생성하여 평가 셋으로 사용하였다. 실험에서 사용한 지식베이스³ 는 위키피디아 데이터를 기반으로 구축하였으며, 평가 셋은 일반형과 판정형 그리고 비교형 질의를 각 100 개씩 생성하였다. 이때 일반형 질의 100 개중 50 개는 단순 질의이고 50 개는 복합

³ 실험에서 사용한 지식 베이스는 adams.ai 에서 접근할 수 있다.

질의 이다. 또 판정형 질의 100 개중 65 개는 답이 ‘예’ 이고 35 개는 ‘아니요’ 인 질의로 구성하였다.

질의유형 분류기 모형은 위에서 설명한 평가 질의를 제외한 일반형, 판정형, 비교형을 포함하여 총 10,000 건의 데이터를 이용해서 학습했다.

합성곱 신경망에서 Convolution 필터가 의미하는 것은 사용자 질의에 대해 각 사이즈 별로 n-gram 을 고려하여 특징맵을 추출한다는 의미이다. 그리고 pooling 은 Max pooling 을 사용하였다. convolution layer 와 pooling layer 는 각각 2 개씩 구성하였고, 과적합을 피하기 위해 drop-out 의 비율은 0.5 로 설정하여 학습모형을 만들었다.

복합 질의 응답 시스템에 질의유형 분류기를 적용하는 것의 유용함을 확인하기 위해 분류기 적용 전후의 정답률을 비교하는 실험을 진행하였다. 분류기를 적용하지 않은 경우에는 질의 처리를 임의의 순서로 수행하도록 하였다(예: FBQA → YNQA → CPQA). 분류기가 적용된 경우는 분류기의 분류 점수가 높은 순서대로 질의 처리를 진행한다. 각 질의 처리 모듈에서 답변이 가능하면 KBCQA 시스템은 답변을 사용자에게 반환하게 되며, 만약 지정된 Timeout 을(실험에서는 3 초) 초과하거나 모든 질의 처리 모듈에서 답변하지 못할 경우 시스템은 응답하지 않은 것으로 간주하였다.

3.2 실험 결과

<표 1> 질의유형 분류기 실험 결과

	학습 정확도	테스트 정확도
Case 1	0.968	0.960
Case 2	0.984	0.980

먼저 본 실험에서 300 건의 질의에 대한 질의유형 분류기의 분류 정확도는 <표 1>과 같다. 이 때 입력 방식이 case2 이고, 필터사이즈는 (2,3,4) 였을 때 정확도가 가장 높았다. 통합 테스트에서는 해당 모형을 사용했다.

<표 2> 복합 질의 응답 시스템 정답률

	Recall	Precision	F1-Score
KBCQA	0.76	0.78	0.76
CNN-KBCQA	0.81	0.86	0.83

분류기 적용 전 시스템(KBCQA)과 분류기 적용 후 시스템(CNN-KBCQA)의 정답률은 <표 2>와 같다. 분류기를 적용한 후 재현율과 정확률이 각각 6.5%, 10.2%만큼 상승한 것을 확인하였다. 분류기 적용 전후의 시스템 답변을 비교해본 결과, 분류기가 적용되기 전 시스템의 경우 질의유형을 고려하지 않아 다른 모듈로 답변하여 오답을 반환하는 경우가 많은 것을 볼 수 있었다.

일례로 판정형 질의인 “ 김치전의 칼로리는

153kcal 이야?” 와 같은 질문은 일반형 질의인 “ 김치전의 칼로리는?” 과 동일하게 하나의 개체와 하나의 속성을 갖는 질의였다. 그러나 질의유형을 파악하지 못했기 때문에 일반형 질의 처리에서 김치의 칼로리를 답변하여 오답을 반환하는 결과를 보였다.

4. 결론 및 향후 연구 방향

본 논문에서는 합성곱 신경망 기반의 질의유형 분류기를 활용한 복합 질의 응답 시스템을 제안하였다. 복합 질의는 동일한 시맨틱 패턴이더라도 일반형, 판정형, 비교형 등 다양한 질의유형을 갖는다. 우리는 실험을 통해서 질의유형 분류기의 유용함을 보였다. 또한 복합 질의 응답 실험 결과에서 시맨틱 패턴을 활용하는 각 모듈이 높은 정답률을 보임을 확인하였다.

그러나 분류기 적용 후에도 해결하기 어려운 질의가 존재했다. 이들은 개체 인식, 속성 탐지와 같은 자연어 이해 단계의 한계로 인한 것으로 학계에서도 활발히 연구하고 있는 분야이다[7,8]. 우리도 이 부분을 집중적으로 연구하여 개선할 예정이며, 나아가 부정형 질의, 다중 조건 질의 등 더 다양한 유형의 복합 질의에 대해서도 연구를 진행할 계획이다.

Acknowledgement

본 연구는 산업통상자원부와 한국산업기술진흥원의 “국제공동기술개발사업”의 지원을 받아 수행된 연구결과임.

참고문헌

- [1] Berant, Jonathan, et al. "Semantic parsing on freebase from question-answer pairs." Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013.
- [2] Yin, Wenpeng, et al. "Simple question answering by attentive convolutional neural network." arXiv preprint arXiv:1606.03391(2016).
- [3]. Khot, Tushar, Ashish Sabharwal, and Peter Clark. "Answering complex questions using open information extraction." arXiv preprint arXiv:1704.05572 (2017).
- [4]. Saha, Amrita, et al. "Complex Sequential Question Answering: Towards Learning to Converse Over Linked Question Answer Pairs with a Knowledge Graph." arXiv preprint arXiv:1801.10314(2018).
- [5] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).
- [6] 양선, and 고영중. "한국어 비교 마이닝을 위한 비교 요소 자동 추출." 정보과학회논문지: 소프트웨어 및 응용 38.12 (2011): 689-696.
- [7] Dubey, Mohnish, et al. "EARL: Joint Entity and Relation Linking for Question Answering over Knowledge Graphs." arXiv preprint arXiv:1801.03825 (2018).
- [8]. Yu, Mo, et al. "Improved neural relation detection for knowledge base question answering." arXiv preprint arXiv:1704.06194(2017).