Journal of KIIT. Vol. 18, No. 3, pp. 39-48, Mar. 31, 2019. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2020.18.3.39

지식베이스 질의응답을 위한 자연언어 질문의 시맨틱 프레임 구조 표현 방법 연구

류 법 모*

A Study on the Semantic Frame Representation of Natural Language Question for Knowledge Base Question and Answering

Pum-Mo Ryu*

논문은 2019년도 부산외국어대학교 학술연구조성비에 의해 연구되었음.

요 약

본 논문은 지식베이스 질의응답 시스템(KBQA)를 위한 자연언어 질문분석에서 필수적으로 고려되어야 하는 제약 조건을 제시하고 이를 시맨틱 프레임으로 표현하는 방법을 제안한다. 단답형 자연언어 질문을 대상으로 서술어-논항구조 및 그 의미역할에 대한 언어 분석을 수행하고, 질문 문장의 특성을 반영하여 질문초점, 정답 제약서술어, 질문분할, 시간제약, 공간제약, 부정표현 등을 추가한다. 기존의 정답유형 또는 키워드 중심의 질문분석 방법과 비교하여 문장의 구문 및 의미구조를 분석하기 때문에 지식베이스에 저장된 구조화된 콘텐츠와 정확하게 비교하여 정답을 추론할 수 있는 장점이 있다.

Abstract

This paper proposes constraints that must be considered in natural language question analysis for the knowledge base question and answering system (KBQA), and presents a way to express these constrains in the form of semantic frame. The proposed method analyzes the predicate-arguments structure and semantic roles for factoid natural language questions, and represents additional characteristics of question sentences such as question focus, answer-constraint predicates, question segmentations, temporal restriction, spatial restriction and negation expressions. Compared with the existing question analysis methods which focus on answer type and keyword, our method has the advantage that it can precisely compare with the structured contents stored in the knowledge base in that it analyzes the syntax and semantic structure of question sentences.

Keywords

question-answering system, question analysis, knowledge base, frame semantics, question focus

* 부산외국어대학교 사이버경찰전공 교수
- ORCID: https://orcid.org/0000-0002-9777-9211

· Received: Feb. 05, 2020, Revised: Mar. 20, 2020, Accepted: Mar. 23, 2020
· Corresponding Author: Pum-Mo Ryu
Dept. of Cyber Security & Police, Busan University of Foreign Studies, Busan, Korea,
Tel.: +82-51-509-6782, Email: pmryu@bufs.ac.kr

Ⅰ. 서 론

웹의 폭발적인 성장과 모바일 환경의 도래로 인해 정보검색 분야에서 질의응답기술의 중요성이 높아지고 있다. 사용자가 원하는 정보는 불필요한 정보, 왜곡 정보들과 함께 웹상에 혼재되어 있으므로, 사용자가 요구하는 정보를 기존의 문서검색 방식으로 찾기가 점점 어려워지고 있다. 또한, 비교적 화면크기가 작은 모바일 기기에서 기존의 웹 검색 방식과 유사한 사용자 인터페이스를 제공하면 검색결과의 활용성 측면에서 단점으로 작용한다. 이와 같은 환경 변화로 인해 사용자의 구체적인 질문에 정확한 정답과 근거를 제시하는 자연언어 질의응답기술에 대한 수요가 급증하게 되었다[1].

최근 Freebase[2], DBpedia[3], YAGO2[4]과 같은 대규모 지식베이스의 구축과 공개가 활발해 지면서, 지식베이스 기반 질의응답 시스템(KBQA)에 대한 관심이 높아지고 있다. 지식베이스들은 기계가 읽을 수 있는 구조화된 데이터를 저장하고 있으며 대표적으로 <s, p, o>의 트리플 형태로 구축되어 있으며, 지식베이스에 대하여 접근하기 위해서는 SPARQL과 같은 기계가 이해할 수 있는 질의문을 사용하여야 한다. 그러나 SPARQL은 일반 사용자가 사용하기에는 어렵고 복잡하다는 측면이 있어서, 사용자가 직관적으로 이용할 수 있는 자연언어 기반 인터페이스에 대한 관심이 증대하고 있다. 이러한 문제의식으로 QALD[5] 또는 OKBQA[6]와 같은 워크샵 등을 통해, 자연언어 질문을 분석하여 지식베이스에 적합한 질의로 변환해주는 자연언어 질문 이해 연구가 국제적으로 활발하게 이루어지고 있다[7].

KBQA에서 자연언어 질문을 지식베이스의 구조화된 지식과 비교하여 위해서는 자연언어 질문에서 정답유형 및 다양한 근거와 제약조건을 지식베이스 수준의 시맨틱 프레임 형식으로 표현하는 것이 중요하다. 본 연구에서는 지식베이스 대상 질의응답 시스템에서 자연언어 질문으로부터 추출해야 하는 의미적 제약조건을 프레임 시맨틱스 기반으로 표현하는 방법을 제안한다. 기존의 자연언어 질문분석과 관련된 연구는 정답유형 인식에 제한된 연구[8] 또는 일반적인 언어분석 결과를 이용한 시맨틱 프레임 생성 연구 등이 진행되고 있지만[9][10], 실제 정답후보 생성 및 순위화에 필요한 시간제약, 공간제약, 부정(Negation) 제약 등을 포함하는 통합적인 질문 시맨틱 프레임 표현 방법과 관련된 연구는 활발하게 진행되지 않고 있다. 따라서 본 연구에서는 KBQA를 위한 질문분석에서 필수적으로 고려되어야 하는 제약조건을 제시하고 이를 시맨틱 프레임 형식으로 표현하는 방법을 제안한다.

2장에서는 자연언어 질의응답에서 질문분석과 관련된 기존의 연구를 살펴보고, 3장에서 자연언어 질문을 시맨틱 프레임으로 변환하는 절차를 설명하고, 4장에서 자연언어질문에 대한 의미적 제약조건을 시맨틱 프레임으로 표현하는 방법을 제안한다. 마지막으로 5장에서 결론을 맺는다.

Ⅱ. 관련 연구

자연언어 질의응답 기술의 발전에서는 1999년부터 시작된 TREC(Text Retrieval Conference)의 질의응답 경연대회가 큰 역할을 하였다. TREC에서는 사실기반 질의응답(Factoid QA)부터 시작하여 리스트형 질의응답(list QA), 정의형 질의응답(Definition QA), 대화형 질의응답(Interactive QA)을 거쳐 실시간 질의응답(live QA)으로 기술적 난이도를 높이고 있다[11].

LASSO 시스템의 질문분석모듈은 크게 3단계로 구분된다. 첫째, 질문의 유형을 결정한다. 질문유형은 5W1H에 기반을 둔다. 둘째, 정답유형을 결정한다. 정답유형은 정답의 의미범주로 정의한다. 이때, Who형 질문의 경우 정답유형이 PERSON으로 명확하지만, What형 질문의 경우 모호성 문제가 발생한다. 이 문제를 해소하기 위해서, 질문초점의 개념을 정의하고 있다. 질문초점은 질문에서 찾고자 하는 것을 지칭하여 질문의 모호성을 해소하는 질문 내의 단어나 구로 정의된다[12].

IBM은 1997년 ‘Deep Blue’ 이후 인공지능의 새로운 도전 분야로 질의응답을 선택하였고 2011년에 DeepQA인 왓슨(Watson)을 야심차게 발표하였다. 왓슨은 ‘Jeopardy! 퀴즈쇼’에서 인간챔피언을 이기면서 크게 주목을 받았다. 이는 인공지능 연구 부흥에 큰 역할을 하였다[13].

왓슨에서는 정답유형으로 기 정의된 의미범주를

사용하지 않는다. LASSO 시스템과 패턴을 이용하여 질문초점을 인식한다. 인식된 질문초점의 중심어를 어휘정답유형(Lexical answer type)으로 인식한다[14].

자연언어 질문 분석에서 정답유형 및 질문초점 인식과 관련된 연구는 활발히 진행되고 있지만 질문에 표현된 다양한 제약조건을 인식하고 구조화하는 자연언어 질문에 특화된 심층의미분석 기술 개발과 관련한 연구는 아직 초기단계이다. Heo의 연구에서는 자연언어 질분 분석의 여러 가지 요소 중 정답유형 인식을 제한적으로 다루고 있다[8].

Ryu와 Hahm의 연구는 심층 언어분석 단계와 질문중간표현 생성 단계를 거치면서 자연언어 질문을 프레임 시맨틱스 기반의 의미표현으로 변환하는 방법을 특허와 논문으로 제안하였다. 각 질문을 대상으로 의문대명사를 포함한 질의, 의문대명사를 포함하지 않는 질의, 평서문 형태의 질의로 구분하고 각 구분의 특징에 맞게 주어진 질문에서 <?x, p, o> 수준의 트리플 구조를 생성한다. 이 과정에서 동사가 없는 질의를 처리하는 방법, 발견되지 않는 주요 논항을 처리하는 방법 등 세부적인 문제에 대한 해결방안을 설명한다. 그러나 형태소분석, 개체명 인식, 구문분석, 의미역 인식 등 일반적인 언어분석 방법론만 적용하고, 질문분석에서 중요한 요소인 시간, 공간, 부정(Negation) 정보 등에 대한 고려가 부족하다[9][10].

III. 자연언어 질문 시맨틱 프레임 생성 절차

본 절에서는 자연언어 질문을 대상으로 시맨틱 프레임을 생성하는 절차를 설명한다. 그림 1과 같이 자연언어 질문을 입력받아서 언어분석 과정을 수행한 후, 심층 질문분석을 통해서 질문 시맨틱 프레임을 생성한다.

언어분석 단계에서는 형태소 분석, 의존구문분석, 개체명 인식, 의미역 분석을 순차적으로 수행한다. 그림 2는 자연언어 문장에 대한 언어분석 결과 예제이다. 형태소 분석에서는 문장을 형태소 단위로 분절하고 각 형태소에 품사정보를 부착한다. 개체명 인식 단계에서는 문장에서 나타나는 인명, 지명, 조직명, 시간표현 등을 인식하고 각각의 유형을 부착한다. 예시에서는 지명 '그리스'와 시간표현 'BC 5~4세기'를 인식한다. 의존구문분석은 서술어를 중심으로 각 논항들의 구문역할(주어, 목적어, 부사어 등)을 표현하거나, 명사구 내부의 수식 관계를 표현한다. 예를 들어 서술어 '말한다'는 '철학자들을'을 목적어로 가지고, '뜻으로'를 부사어로 가진다.

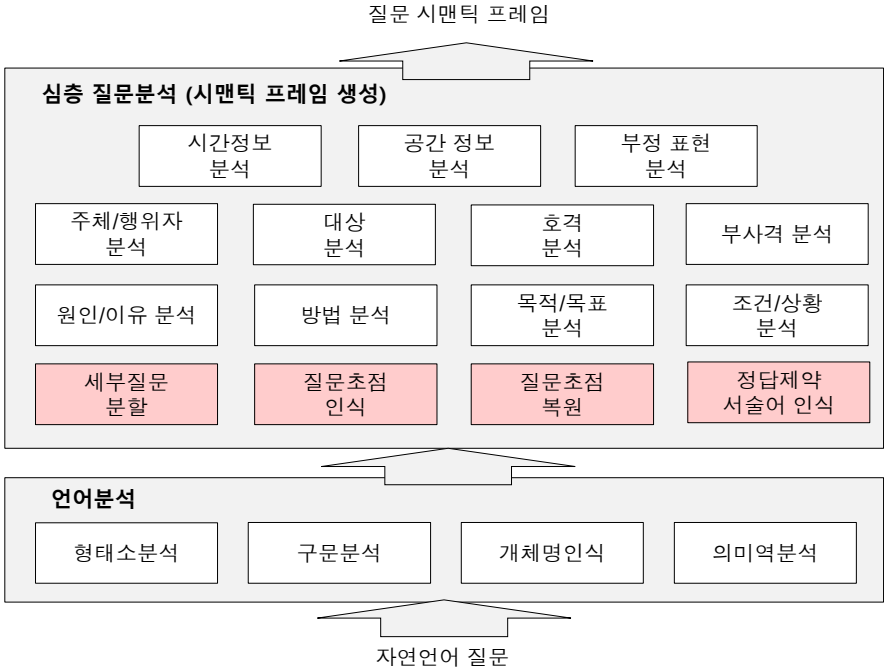

그림 1. 자연언어 질문 시맨틱 프레임 생성 단계

Fig. 1. Steps of generating question semantic frame from natural language question

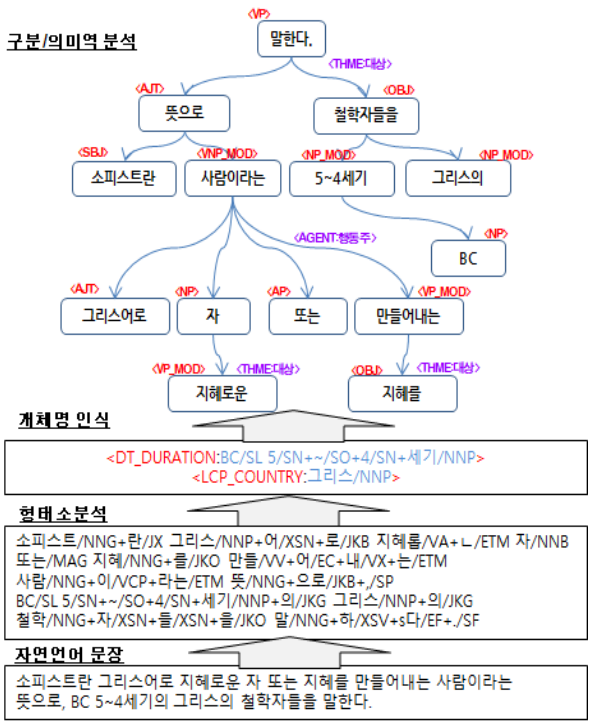

그림 2. 단계별 언어분석 결과 예시
Fig. 2. Examples of language analysis steps

또한 '5~4세기'는 'BC'를 명사수식어로 취한다. 의미역 분석에서는 서술어의 각 논항의 의미적 역할(Agent/행동주, Theme/대상 등)을 부여한다. 예를 들어 서술어 '말하다'는 '철학자들을'을 대상으로 지정하고, 서술어 '만들어내다'는 '사람이라는'을 행동주로, '지혜를'을 대상으로 지정한다.

Ⅳ. 자연언어 질문 의미적 제약조건

본 절에서는 단답형 자연언어 질문을 대상으로 추출할 수 있는 의미적 제약조건을 정의하고, 심층 질문분석 즉 시맨틱 프레임 생성 단계의 개념들을 쉽게 이해할 수 있도록 제약조건을 태깅(tagging)하는 과정을 통해서 설명한다. 자연언어 질문에서 질문초점(QF: Question Focus)을 먼저 표시하고, 이를 중심으로 정답제약 서술어를 태깅하고, 정답제약 서술어 중심의 논항들의 의미역할 태깅, 질문 분할 태깅, 생략된 질문초점 복원 등 정답제약 정보를 태깅한다. 태깅 단위를 대괄호 '[', ']'으로 묶고, 태깅 단위 뒤에 '//' 기호와 정보유형을 차례로 부착한다. 한 개의 태깅 단위에 여러 개의 정보를 태깅할 수 있다. 아래 예문에서는 뒤 절에서 설명되는 태그를 미리 부착한 경우도 있기 때문에 자세한 정의는 뒤 절의 설명을 참조하기 바란다.

4.1 질문초점

질문초점은 자연언어 질문에서 정답제약의 중심이 되는 어휘를 가리키며 [QF] 태그를 이용하여 표시한다. 질문초점은 다음과 같은 특징이 있다.

- 정답의 유형을 표현한다. 예문 (1-1)에서 질문초점 '화산'은 정답의 유형을 표현한다.
- 정답을 제약하는 다른 어휘의 수식을 받는다. 예문 (1-1)에서 정답제약 조건 '세계 최대의'은 질문초점 '화산'을 수식한다.
- 정답을 제약하는 다른 논항들과 공통의 정답제약 서술어를 공유한다. 예문 (1-2)에서 질문초점 '이것은'은 정답제약 서술어 '만들었다'의 여러 논항 중 하나이다. 그림 3은 예문 (1-2)를 서술어 '만들다'를 중심으로 프레임 형식으로 표현한 것이다.
- 질문초점을 정답후보로 대체하면 정답가설(Question hypothesis)이 생성된다. 예문 (1-3)에서 '이것'을 정답후보 '자격루'로 대체하면 정답가설이 생성되고, 정답후보 순위화 단계에서는 정답가설을 다양한 근거를 기반으로 검증하는 절차를 수행한다.

(1-1) *세계 최대의 [화산]//[QF]*
(1-2) *[이것은]//[QF][A1] [세종 때]//[ATS] [장영실, 이천 등이]//[A0] [왕명을 받아]//[AC] [만들었다.]//[PV]*
(1-3) *정답가설: '이것'을 '자격루'로 대체함*
[자격루는]//[QF][A1] [세종 때]//[ATS] [장영실, 이천 등이]//[A0] [왕명을 받아]//[AC] [만들었다.]//[PV]

[만들었다.]//[PV] (정답제약서술어)
- *[이것은]//[QF][A1]* (질문초점, 대상)
- *[세종 때]//[ATS]* (시간제약)
- *[장영실, 이천 등이]//[A0]* (주체)
- *[왕명을 받아]//[AC]* (원인)

그림 3. 예문 (1-2)에 해당하는 시맨틱 프레임 구조
Fig. 3. Semantic frames corresponding to example 1-2

4.2 정답제약 서술어

정답제약 서술어는 질문초점을 논항으로 가지는 서술어를 의미한다. 예문 (2-1)~(2-3)은 태그 [PV]를 이용하여 정답제약 서술어를 태깅한 예제이다. 정답초점 '화가', '말', '현악기'는 각각 서술어 '그리다', '가리키다', '매다'의 주체 또는 대상 논항이다.

(2-1) *진경산수화의 하나인 '인왕제색도'를 [그린]///[PV] [화가]///[QF]*
(2-2) *아들이 많은 집의 외동딸을 [가리키는]///[PV] [말]///[QF]*
(2-3) *12줄의 명주실을 [맨]///[PV] 우리나라 대표 [현악기]///[QF]*

서술어가 일부 명사 뒤에 붙어 동사를 만드는 접미사 '-하다', '-시키다', '-당하다', '-받다' 등의 형태인 경우에는 목적어까지 포함하여 서술어로 태깅한다. 마찬가지로 접미사 '-되다' 의 경우에도 보격 명사까지 포함하여 서술어로 태깅한다. 의미수준의 질문분석에서는 단순히 어절 단위 분석이 아니고 다어절로 구성된 의미단위 수준의 분석이 이루어져야 한다. 다만 이 때 정답제약 서술어에 포함되는 목적어와 보어는 단일명사만 가능하고, 복합명사나 명사구, 관형절의 수식을 받는 경우에는 불가하다. 예문 (2-4)~(2-7)에서 복합어 '재료가 되다', '효과가 있다', '관련이 있다', '뜻을 가지다'를 정답제약 서술어로 태깅한다.

(2-4) *박지원이 쓴 <허생전>의 주인공 허생은 [이것의]///[QF] [재료가 되는]///[PV] 말총을 매점매석해 큰 이익을 취한다.*
(2-5) *생체친화성이 있어 화상과 피부질환에 큰 [효과가 있는]///[PV] [이것은]///[QF] 무엇일까?*
(2-6) *사람의 신체와 [관련이 있는]///[PV] [말이]///[QF] 아닌 것은 무엇일까?*
(2-7) *흰 사슴이 물을 마시던 연못이라는 [뜻을 가진]///[PV] 한라산 정상의 [화구호]///[QF] 이름*

지위나 신분 또는 자격을 나타내는 격조사 '-로/으로', '-로서/으로서', 서술격조사 '이다(be)' 또는 부정형용사 '아니다(not be)'는 의미적으로 서술어 역할을 담당하기 때문에 서술어로 볼 수 있다. 따라서 이와 같은 지정사형 표현을 [PC] 태그를 이용하여 서술어로 태깅한다. 예문 (2-8)에서 의미적으로 볼 때, '-으로' 서술어를 중심으로 '위성'이 주어가 될 수 있다. 이 경우 '위성으로'에 [QF]와 [PC] 태그를 모두 부착한다. 그림 4는 '-으로'를 의미적으로 '이다'를 표현하는 서술어로 표기하고, 이를 중심으로 '위성'이 논항으로 표현된 프레임 구조이다. 이 질문은 '-으로' 프레임과 '고안했다' 프레임의 2개의 세부 프레임으로 구성된다. 세부 질문분할과 관련된 자세한 내용은 4.3절에서 설명한다.

(2-8) *가로, 세로 10cm 초소형 [위성으로]///[QF][PC][A0] [SN] [미국에서]///[ASS] [대학생들의 위성개발 실습을 위해]///[AP] [X]///[QF][A1] [고안했다.]///[PV]*

[으로]///[PC] (정답제약서술어)
- *[위성]///[QF][A0]* (질문초점, 주체)

[고안했다.]///[PV] (정답제약서술어)
- *[X]///[QF][A1]* (질문초점, 대상)
- *[미국에서]///[ASS]* (장소)
- *[대학생들의 위성개발 실습을 위해]///[AP]* (목적)

그림 4. 예문 (2-8)에 해당하는 시맨틱 프레임 구조
Fig. 4. Semantic Frames corresponding to example 2-8

4.3 세부 질문 분할 및 질문초점 생략 복원

자연언어 질문이 여러 절의 담화 연결(Discourse relation) 또는 대등 연결 형식으로 구성된 경우, 세부 질문 단위로 분할하고, 각 분할 단위로 질문초점과 정답제약 서술어를 태깅한다. 각 분할 단위에 질문초점이 생략된 경우, 질문초점을 복원하여 태깅한다. 질문 분할은 크게 용언-용언 간 연결, 체언-용언 간 연결로 구분한다.

(1) 용언 - 용언 간 대등절 연결 [SV]

질문에서 두 개의 용언이 대등연결어미로 연결된 경우 [SV]를 이용하여 분할 위치를 태깅한다. [SV]를 중심으로 두 개의 세부 질문으로 분할할 수 있다. 예문 (3-1)에서 '통칭하며'에 [SV]를 태깅하여 분할 위치를 표시하고, 이를 중심으로 분할 된 두 개의 세부질문을 표기한다.

(3-1) *[이것은]///[QF] 숲의 식물들이 만들어내는 살균성을 가진 모든 물질을 [통칭하며]///[PV][SV] 1937년 러시아의 생화학자 토킨이 [제안한///[PV] [용어이다.]///[QF] 이것은 무엇일까?*

(세부질문1) *[이것은]///[QF] 숲의 식물들이 만들어내는 살균성을 가진 모든 물질을 [통칭하며]///[PV]*

(세부질문2) *1937년 러시아의 생화학자 토킨이 [제안한///[PV] [용어이다.]///[QF] 이것은 무엇일까?*

⑵ 체언 - 용언 간 대등절 연결 [SN]

형태상 질문문장에 용언이 없지만 격조사 '-로/으로' 나 '-로서/으로서', 또는 서술격조사 '-이다' 가 서술어 역할을 하여([PC]인 경우) 의미상 대등절로 볼 수 있을 경우에 [SN]을 이용하여 분할 위치를 태깅한다. 예문 (3-2)에서 '소설가로' 에 [SN]을 태깅하여 분할위치를 표시한다.

(3-2) *러시아 [소설가로]///[PC][SN][QF] 카라마 조프가의 형제들을 [쓴]///[PV] [작가는]///[QF] 누구일까?*

(세부질문1) *러시아 [소설가로]///[PC][QF]*

(세부질문2) *카라마 조프가의 형제들을 [쓴]///[PV] [작가는]///[QF] 누구일까?*

⑶ 질문초점 복원

한 개의 질문이 여러 개의 세부 질문으로 분할된 경우, 각 세부 질문에는 질문초점이 생략된 경우가 있다. 이 경우 생략된 질문초점을 정답제약 서술어 앞에 복원하여 [X]로 표기한다. 예문 (3-3)를 2개의 세부 질문으로 분리하고, 세부질문 1에 질문초점이 생략되어 있기 때문에 질문초점을 복원하여 표시한다.

(3-3) *아테네에서 수사와 변론을 [X]///[QF] [가르치며]///[PV][SV] 그리스어로 '지혜'를 [뜻하는]///[PV] [사람]///[QF]*

(세부질문1) *아테네에서 수사와 변론을 [X]///[QF] [가르치며]///[PV]*

(세부질문2) *그리스어로 '지혜'를 [뜻하는]///[PV] [사람]///[QF]*

예문 (3-4)는 4개의 세부 질문으로 분리되고, 세부질문 1과 세부질문 3에는 질문초점을 복원하여 표기한다. 각 세부 질문 뒤에 [SV], [SN] 기호를 태깅한다.

(3-4) *치매를 [X]/[QF] [일으키는]///[PV][SV] 가장 흔한 퇴행성 [뇌질환으로]///[QF][PC][SN] 독일의 정신과 의사에 의해 최초로 [X]///[QF] [보고됐다.]///[PV][SV] 미국의 전 대통령 로널드 레이건이 [이 질병에]///[QF] [걸려]///[PV] 더욱 알려졌다.*

(세부질문1) *치매를 [X]/[QF] [일으키는]///[PV]*

(세부질문2) *가장 흔한 퇴행성 [뇌질환으로]///[QF][PC]*

(세부질문3) *독일의 정신과 의사에 의해 최초로 [X]///[QF] [보고됐다.]///[PV]*

(세부질문4) *미국의 전 대통령 로널드 레이건이 [이 질병에]///[QF] [걸려]///[PV] 더욱 알려졌다.*

4.4 정답제약 서술어 중심의 논항 의미역할

정답제약 서술어를 중심으로 각 논항의 의미역할을 태깅한다. 본 연구에서 사용하는 의미역할은 전체적으로 의미역 태깅(Semantic role labeling) 관련 연구 [15]에서 정의한 의미역할과 유사하나, 자연언어 질문분석에 특화된 의미역을 추가한다. 질문초점도 해당하는 의미역할을 태깅한다. 이를 위해서 ExoBrain과제에서 제공한 QA 데이터셋 [16]에 포함된 단답형 질문 138개를 대상으로 수작업 시맨틱 프레임 태깅을 진행하였고, 이를 기반으로 중요도가 높은 제약조건을 선정하였다. 표 1은 데이터셋에서 발견한 정답 제약조건의 빈도수를 제시하고 있다.

표 1. 평가용 ExoBrain 질의응답 데이터에서 나타난 제약조건별 빈도수

Table 1. Frequencies of restrictions in ExoBrain QA dataset

Restriction type	Freq.	Restriction type	Freq.
Causality	21	Space/Location	26
Method	34	Vocative	47
Instrument	12	Negation	9
Purpose	13	Agent	223
General condition	26	Theme	38
Time	57	Adverbial	13

(1) 원인/이유 [AC: Argument of Causality]

정답제약서술어에 해당하는 사건이 발생한 이유 또는 원인을 표현하는 논항에 의미역할 [AC]를 태깅한다. 예문 (4-1)에서 '베수비오 화산의 폭발로'를 정답제약서술어 '사라지다'의 원인으로 표기한다. 예문 (4-2)에서 '인구의 증가, 자동차 통행의 증가, 각종 인공 시설물의 증가 등으로'를 정답제약서술어 '나타나다'의 원인으로 태깅한다.

(4-1) *서기 79년 [베수비오 화산의 폭발로]///[AC] 한 순간에 [사라졌던]///[PV] 로마의 [도시]///[QF]*

(4-2) *[인구의 증가, 자동차 통행의 증가, 각종 인공 시설물의 증가 등으로]///[AC] 도시 중심부의 기온이 주변 지역보다 현저히 높게 [나타나는]///[PV] [현상은?]///[QF]*

(2) 방법 [AM: Argument of Method]

정답제약서술어에 해당하는 사건을 수행하는 방법 또는 수단에 해당하는 논항의 의미역할을 태깅한다. 예문 (4-3)에서 '지도를 보여줘'를 정답제약서술어 '도와주다'를 수행하는 방법으로 태깅한다. 예문 (4-4)에서 '청자에 백토를 칠하고 투명한 유약을 입혀'를 정답제약서술어 '구워내다'의 방법으로 태깅한다.

(4-3) *[지도를 보여줘]///[AM] 자동차 운전을 [도와주는]///[PV] 길 안내 [장치]///[QF]*

(4-4) *[청자에 백토를 칠하고 투명한 유약을 입혀]///[AM] 다시 [구워낸]///[PV] [자기는]///[QF] 무엇일까?*

(3) 도구 [AI: Argument of Instrument]

정답제약 서술어에 해당하는 사건을 행할 때 사용하는 도구에 해당하는 논항의 의미역할을 태깅한다. '도구'와 '방법'의 판단 기준이 애매한 경우가 있는데, '방망이로', '택시로', '금으로' 등 구체명사에 '-로/으로'가 붙은 경우 '도구'로 분석하고, 그 외 구체명사가 아닌 경우는 방법으로 분석한다. 예문 (4-5)에서 '그리다' 행위를 수행할 때 '먹으로'는 도구의 역할, '짙고 엷음을 이용하여'는 방법의 역할을 수행한다.

(4-5) *[먹으로]///[AI] [짙고 엷음을 이용하여]///[AM] [그린]///[PV] 전통 [회화]///[QF]*

(4) 목적/목표 [AP: Argument of Purpose]

정답제약 서술어에 해당하는 사건을 행하는 목적 또는 목표를 나타내는 논항의 의미역할을 태깅한다. 예문 (4-6)에서 '줄이다' 행위의 목적은 '실제 거리를 지도에 표현하기 위해'이다. 예문 (4-7)에서 '인용하다' 사건의 목적은 '원작자에 대한 존경의 의미로'이다.

(4-6) *[실제 거리를 지도에 표현하기 위해]///[AP] 일정하게 [줄인]///[PV] [비율][QF]*

(4-7) *예술작품에서 [원작자에 대한 존경의 의미로]///[AP] 원작의 요소를 [인용하는]///[PV] [것은]///[QF] 무엇일까?*

(5) 일반적인 조건이나 상황 [ACD: Argument of Condition]

정답제약 서술어에 해당하는 사건이 발생하는 조건이나 상황을 표현하는 논항을 태깅한다. 시간 및 공간과 관련된 조건은 별도의 의미태그를 이용하기 때문에 제외한다. 예문 (4-8)~(4-10)은 각각의 사건 '생겨나다', '확보하다', '생산하다'가 발생할 수 있는 조건 또는 상황을 [ACD]로 표기하고 있다.

(4-8) *매우 급하게 서두르는 모양을 [나타내는]///[PV] [부사로]///[QF][PC] [불이 나서 급히 내달리는 상황에서]///[ACD] [생겨난]///[PV] [단어는]///[QF] [A0] 무엇일까?*

(4-9) *[채무자가 자신의 재산을 함부로 처분할 가능성이 큰 경우]///[ACD] 채무자의 재산을 임시로 [확보하는]///[PV] [이 조치가]///[QF] 필요하다.*

(4-10) *[해안에 방조제를 설치한 후]///[ACD] [밀물과 썰물의 조차를 이용해]///[AM] 전기를 [생산하는]///[PV] [발전 방식]///[QF]*

(6) 특정 시간 정보 [ATS: Argument of Time Specific]

시간표현, 연대표현 또는 고유명사에 시간단서어휘가 추가로 결합한 경우 전체를 하나의 구체적

인 시간으로 인식하여 태깅한다. ([시간표현, 연대표현, 고유명사] + [초창기, 초기, 말기, 중반, 즈음, 당시, 이후, 시기 등]). 예문 (4-11)에서 사건 '구제하다'의 구체적인 시간 제약으로 '고려 광종 시기'를 태깅한다.

(4-11) *[고려 광종 시기]///[ATS] [농민 생활의 안정을 위해]///[AP] [기금을 모아]///[AM] [그 이자로]///[AI] [빈민을]///[A1] [구제했던]///[PV] [기구는]///[QF][A0] 무엇일까?*

(7) 특정한 공간 정보 [ASS: Argument of Space Specific]

지명 고유명사나 지명 고유명사에 장소를 표현하는 어휘가 붙어있는 경우에는 전체를 특정한 공간으로 인식한다. ([지명 고유명사] + [유역, 인근, 근해, 해안, 연안, 근처, 상류, 지역, 전역, 동부, 서부 등]). 예문 (4-12)은 사건 '펼쳐지다'의 구체적인 공간 제약 정보로 '아마존 강 유역에'를 표기한다.

(4-12) *[아마존 강 유역에]///[ASS] [펼쳐진]///[PV] 세계 최대의 열대우림 [지역을]///[QF] 일컫는 말은 무엇일까?*

(8) 호격 [AV: Argument of Vocative]

정답제약 서술어 '불리다', '이르다' 등과 호응하여 해당 행위의 대상이 되는 경우 특별히 호격 기호 [AV]를 이용하여 태깅한다. 호격 태그를 이용하면 정답의 별칭 정보를 추론할 수 있다.

예문 (4-13)에서 '불리다' 사건의 호격 논항은 '바타스라고' 이고, 이 정보를 질문초점 '신'의 별칭으로 인식할 수 있다. 또한 예문 (4-14)에서 '불리다' 사건의 호격 논항으로 '폴라비우스 투기장'을 태깅하고, 이 정보는 질문초점 '경기장'의 별칭으로 사용할 수 있다.

(4-13) *[그리스 로마 신화에서]///[A2] [바카스라고도]///[AV] [불리는]///[PV] 술의 [신]///[QF][A0]*

(4-14) *['플라비우스 투기장'으로]///[AV] [불렸던]/// [PV] [이탈리아 로마의]///[ASS] 거대한 원형 [경기장]///[QF][A0]*

(9) 사건의 부정 [NEG: Negation]

정답제약서술어를 부정하는 표현을 [NEG] 태그를 이용하여 구체적으로 표기하여, 정답후보 순위화 모듈에서 활용할 수 있도록 한다. 예문 (4-15)에서 '해당하지 않다' 사건은 부정적인 표현임을 명시적으로 표기한다. 자연언어에서 '부정' 다양하게 표현될 수 있다. 예문 (4-16)은 부정 표현 '아니다'가 서술어 '유래되다' 와 인접하고 있지 않다.

(4-15) *[용비어천가에서]///[ACD] [조선을 세운 해동육룡에]///[A2] [해당하지 않는]///[PV][NEG] [왕은]///[QF][A0] 누구일까?*

(4-16) *[위인의 이름이나 호에서]///[A2] [유래된]///[PV] [지명이]///[QF][A0] [아닌]///[NEG] 것은 무엇일까?*

(10) 주체/행위자 [A0: Agent], 대상 [A1: Theme]

정답제약 서술어와 호응하여 행위 또는 관계의 '주체/행위자' 역할을 하는 경우 [A0]를 태깅한다. 또한 정답제약 서술어와 호응하여 행위 또는 관계의 '대상' 역할을 하는 경우 [A1]을 태깅한다. [A0], [A1]은 각각 의미역 태깅의 ARG0, ARG1과 동일하다.

예문 (4-17)에서 '실시하다' 사건의 대상은 '유학 교육 등을'이고, 주체는 '교육기관'이다. '교육기관'은 질문초점이면서 동시에 의미적으로 사건의 '주체' 역할을 한다. 예문 (4-18)에서 '가리키다' 사건의 대상은 '인질들이 그들을 풀어주려는 경찰보다 인질범에게 동조하는 비이성적 현상을'이고, 주체는 '용어다.'이다.

(4-17) *[조선 왕조 시기]///[ATS] [인재 양성을 위해]///[AP] [유학 교육 등을]///[A1] [실시한]///[PV] 최고 [교육기관]///[QF][A0]*

(4-18) *[인질들이 그들을 풀어주려는 경찰보다 인질범에게 동조하는 비이성적 현상을]///[A1] [가리키는]///[PV] 범죄 심리학 [용어다]///[QF][A0] 이것은?*

(11) 부사격 의미역 [A2]

[A0], [A1], [ATS], [ATG], [ASS], [ASG], [AV]를 제외하고 기타 격조사가 붙은 부사어는 [A2] 태그를 부착한다. 예문 (4-19) '뜻하다' 사건의 부사격 의미역으로 '그리스어로'를 태깅한다.

(4-19) *[아테네에서]///[ASS] [수사와 변론을]///[A1] [X]///[QF][A0] [가르치며]///[PV][SV] [그리스어로]///[A2] ['지혜'를]///[A1] [뜻하는]///[PV] [사람]///[QF][A0]*

그림 5는 예문 (4-19)에 태깅된 정보를 시맨틱 프레임 형식으로 표현한 것이다. 두 개의 세부 질문으로 분할하고, 세부질문 단위로 서술어-논항 구조를 표현한다. 각 프레임의 질문초점(QF)과 각 논항의 의미적 역할도 표기하고 있다.

그림 6은 예문 (4-11)에 대한 시맨틱 프레임 구조를 자연언어 텍스트를 기반으로 생성한 지식베이스의 구조화된 지식과 비교하는 과정을 설명한다. 지식베이스에 '제위보'를 주체(A0)로 가지는 2개의 프레임이 있다. 정답 추론 단계에서 질문분석의 '고려 광종 시기'와 지식베이스의 '963년(광종 14년)'과 연결될 수 있고, 질문 시맨틱 프레임과 지식베이스 두 번째 프레임이 서술어 수준에서 연결되어, 점선으로 연결된 논항들이 노드 의미와 프레임 구조에서 유사성을 가진다. 예를 들면, '구제하다'는 '구호하다'와 '이자'는 '변리'와 그리고 '빈민'은 '가난한 사람'과 노드의 의미가 유사하고, 프레임 구조 내에서 의미역할이 유사하다. 이와 같은 시맨틱 비교 과정을 통해서 질문초점(QF)의 위치에 지식베이스의 '제위보'가 연결될 수 있음을 추론할 수 있다.

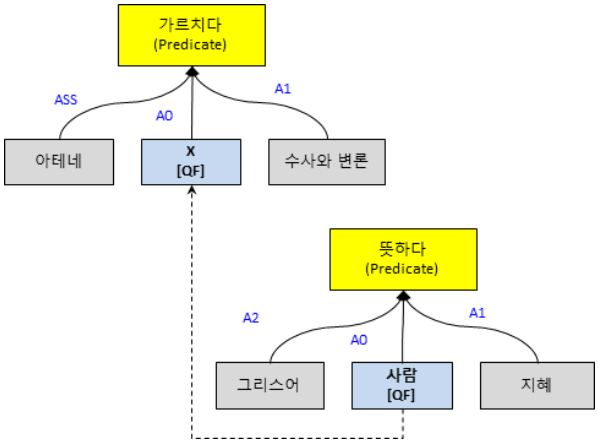

그림 5. 예문 (4-19)에 표현된 시맨틱 프레임 구조
Fig. 5. Semantic frames corresponding to example 4-19

V. 결론 및 향후 과제

본 연구에서는 지식베이스 기반 질의응답시스템에서 필수적인 자연언어 질문의 시맨틱 분석 방안을 제안하였다. 향후 제안한 방법론을 적용하여 질문분석 모듈을 개발하고, 이를 트리플 형식으로 구축된 지식베이스에 적용하여, 기존의 질문분석 모듈보다 우수함을 보일 계획이다. 또한 제안한 질문분석 표현 방법의 상호운용성을 높이기 위하여 XML 또는 JSON 형식의 표준화된 표현 방법을 제안할 예정이다.

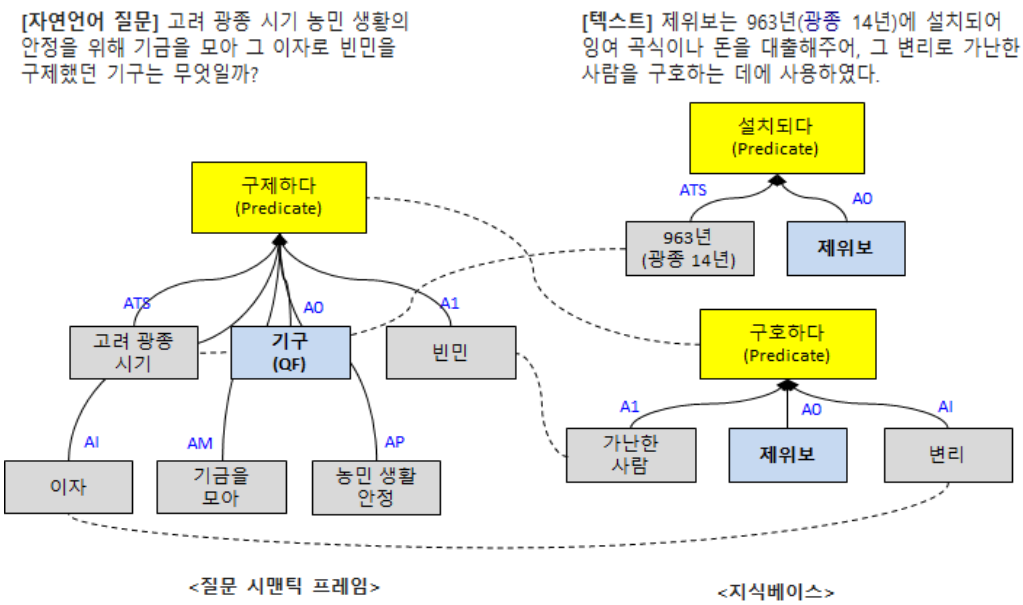

그림 6. 예문 (4-11)에 표현된 시맨틱 프레임 구조와 지식베이스의 비교
Fig. 6. Comparison of knowledge fractions in KB and semantic frame of example 4-11

References

[1] J. Heo, "Constructing answer type taxonomy and understanding questions for deep question and answering", Doctoral dissertation of Ulsan University, Korea, Feb. 2017.

[2] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge", ACM SIGMOD Conference, Vancouver Canada, pp. 1247-1249, Jun. 2008.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of opendata", The semantic web, Springer Berlin Heidelberg, Busan, Korea, pp. 722-735, Nov. 2007.

[4] J. Hoffarta, F. M.Suchanek, K. Berbericha, and G. Weikuma, "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia", Artificial Intelligence, Vol. 194, pp. 28-61, Jan. 2013.

[5] Question Answering over Linked Data (QALD), http://qald.aksw.org/ [accessed: Jan. 01. 2020]

[6] Open Knowledge Base and Question-Answering (OKBQA), http://www.okbqa.org/ [accessed: Jan. 01. 2020]

[7] H. Wang, M. Bansal, K. Gimpel, and D. McAllester, "Machine comprehension with syntax, frames, and semantics", The 53rd Annual Meeting of the ACL 2015, Beijing, China, pp. 700-706, Jun. 2015.

[8] J. Heo, P. Ryu, H. Kim, and C. Ock, "Recognition of answer type for WiseQA", KIPS Tr. Software and Data Eng., Vol. 4, No. 7, pp. 283-290, Jun. 2015.

[9] P. Ryu, H. Kim, Y. Bae, H. Oh, C. Lee, S. Lim, J. Lim, M. Jang, M. Choi, and J. Heo, "Question answering system and method for structured knowledge-base using deep natural language question analysis", Korean Patent, No. 102033395, Oct. 2019.

[10] Y. Hahm, S. Nam, and K. Choi, "Semantic parsing of questions based on the frame semantics for Korean question answering system", The 28th Annual Conference on HCLT, pp. 122-127, Oct. 2016.

[11] J. Burger, C. Cardie, and V. Chaudhri, "Issues, tasks and program structures to roadmap research in question & answering (Q&A)", Document Understanding Conferences Roadmapping Documents, Oct. 2001.

[12] D. I. Moldovan, S. M. Harabagiu, M. Pasca, and R. Mihalcea., "The structure and performance of an open-domain question answering system", The 38th Annual Meeting on Association for Computational Linguistics, pp. 563-570, Oct. 2000.

[13] D. A. FERRUCCI, "Introduction to 'this is watson", IBM Journal of Research and Development, Vol. 56, No. 3.4, pp. 1:1-1:15, May 2012.

[14] A. Lally, J. M. Prager, M. C. McCord, B. K. Boguraev, S. Patwardhan, J. Fan, P. Fodor, and J. Chu-Carroll, "Question analysis: How watson reads a clue", IBM Journal of Research and Development, Vol. 56, No. 3.4, pp. 2:1-2:14, May 2012.

[15] S. Lim, Y. Bae, H. Kim, and D. Ra., "Korean semantic role labeling using domain adaptation technique", Journal of KIISE, Vol. 42, No. 4, pp. 475-482, Apr. 2015.

[16] http://aiopen.etri.re.kr [accessed: Jan. 01. 2020]

저자소개

류 법 모 (Pum-Mo Ryu)

1995년 2월 : 경북대학교 컴퓨터공학과(공학사)
1997년 2월 : POSTECH 컴퓨터공학과(공학석사)
2009년 2월 : KAIST 전산학과 (공학박사)
2009년 ~ 2015년 : 한국전자통신연구원 지식마이닝연구실 책임연구원
2015년 ~ 현재 : 부산외국어대학교 사이버경찰전공 부교수
관심분야 : 정보검색, 자연어처리, 온톨로지, 질의응답