한국어 질의응답 시스템을 위한 프레임 시멘틱스 기반 질의 의미 분석

함영균○, 남상하, 최기선
한국과학기술원
hahmyg@kaist.ac.kr, nam.sangha@kaist.ac.kr, kschoi@kaist.ac.kr

Semantic Parsing of Questions based on the Frame Semantics for Korean Question Answering System

Younggyun Hahm○, Sangha Nam, Key-Sun Choi
KAIST

요 약

본 논문에서서는 질의응답 시스템을 위한 자연언어 질의 이해를 위하여 프레임 시멘틱스 기반 의미 분석 방식을 제안한다. 지식베이스에 의존적인 질의 이해는 지식베이스의 불완전성에 의해 충분한 정보를 분석하지 못한다는 점에 착안하여, 질의의 술부-논항구조 및 그 의미에 대한 분석을 수행하여 자연언어 질의에서 나타난 정보들을 충분히 파악하고자 하였다. 본 시스템은 자연언어 질의를 입력으로 받아 이를 프레임 시멘틱스의 구조에 기반하여 기계가 읽을 수 있는 임의의 RDF 표현방식의 모형 쿼리를 생성한다.

주제어: 질의응답시스템, 의미 분석, 프레임넷, 프레임 시멘틱스

1. 서론

현재 Freebase[1], DBpedia[2], YAGO2[3]과 같은 지식베이스의 발달로 인해, 이러한 지식베이스에 대하여 원하는 지식을 얻고자 하는 지식베이스 기반 질의응답 시스템(KBQA)에 대한 관심이 높아지고 있다. 지식베이스들은 기계가 읽을 수 있는 구조화된 데이터로 구축되어 있으며 대표적으로 <s, p, o> 의 트리플 형태의 RDF(Resource Description Framework) 데이터로 구축되어 있으며, 이 지식베이스에 대하여 접근하기 위해서는 SPARQL과 같은 기계가 읽을 수 있는 쿼리를 사용하여야 한다. 그러나 이러한 쿼리는 일반 사용자가 사용하기에 어렵고 복잡하다는 측면이 있어 최종 사용자가 사용하기 위한 좀 더 직관적이고 사용하기 쉬운 인터페이스에 대한 관심이 증대하고 있다. 이러한 문제의식으로 QALD[1] 나 OKBQA[2]와 같은 워크샵 등을 통해, 자연언어 질의를 분석하여 지식베이스에 적합한 쿼리로 변환해주는 자연언어 질의 이해 연구가 국제적으로 활발히 이루어지고 있다.

전통적으로, 자연언어 질의를 기계가 읽을 수 있는 쿼리로 변경하는 방법은 크게 두 가지 접근법이 있다. 하나는 정보추출(Information Extraction) 방식이고, 하나는 의미 분석(Semantic Parsing) 방식이다. 정보추출 방식은 지식베이스의 스키마와 질의의 구문구조, 지식베이스 온톨로지 어휘와 자연언어의 어휘간의 유의미한 관계 등을 패턴화 하여 학습하는 방식이다[4]. 예를 들어, 질문 “남극에 최초로 도착한 사람은 누구인가?”라는 질문을 SPARQL 쿼리로 변경하기 위하여, 기존의 정보추출 방식에 의해 학습된 패턴에 의해 <?x, p, o> 와 같은 트리플 패턴 형태의 쿼리를 생성한다. 이때, 지식베이스의 하나인 한국어 디비피디아[3]에서는, 질의의 정답을 나타내는 지식이 <dbr:로알_아문센, prop-ko:knownFor, dbr:남극> 의 형태의 트리플로 존재하기 때문에, 위의 질의에 대해서는 다음과 같은 SPARQL 쿼리를 생성할 수 있다.

```
SELECT ?x where {
?x prop-ko:knownFor dbr:남극 }
```

이러한 정보추출 방식의 질의 이해는 목표 지식베이스에 대하여 초점을 맞추어 학습된 방식으로, 특정 도메인과 특정 지식베이스에 대하여서 학습이 쉽고 높은 정확률을 보이는 장점이 있다. 그러나 이러한 정보추출 방법은 지식베이스 스키마의 표현력 부족과 지식 자체의 부족으로 인해 학습할 수 있는 룰의 커버리지 역시도 한정적일 수밖에 없다[5]. 특히 디비피디아 스키마의 경우 위키피디아 인포박스에 기반을 둔 것이기 때문에, 백과사전의 목적에 따라 단순사실을 기술하기에 적합하다(예: 이름, 직업, 인구수, 높이, 출생지 등). 위 예시와 같이 ‘도착하다’와 같은 어휘의 경우, 지식베이스 온톨로지 어휘와의 불규칙한 매칭이 이루어지기 때문에, 그 의미를 온전히 해석하는 데에는 한계가 있다[6].

위와 같은 이유로, 본 논문에서는 한국어 질의 이해를 위하여 의미 분석 방식의 접근법을 적용하였다. 의미 분석 방식은 정보추출 방식과 달리, 지식베이스를 고려하

1) http://qald.sebastianwalter.org/
2) http://www.okbqa.org/
3) http://ko.dbpedia.org

지 않고 질의에서 포함하고 있는 의도와 정보의 의미를 충분히 분석하는 것을 목표로 한다. 의미 분석 기반 질의 이해 방법은 도메인에 의존적이지 않기 때문에 정보추출 방식에 비해 오픈도메인 질의응답 시스템에서 효과적이다. 또한 디비피디아와 같이 제한된 온톨로지 어휘를 사용하는 경우 유용하게 적용될 수 있다[7].

특히, 의미 분석 접근법은 정보추출 접근법에 비해 자연언어에서 나타난 의미를 보다 풍부하게 해석할 수 있다는 장점이 있다. 위의 예시 질문에서 나타난 "최초로", "도착한" 과 같은 어휘에 대해서 정보추출 방식의 접근에서는 지식베이스의 스키마, 즉 프로퍼티로서 그 의미의 모호성을 해소하여 주는데, 한국어 디비피디아의 경우에는 "최초"의 개념과 같은 일종의 한정사에 대해 표현할 수 있는 적절한 온톨로지 어휘를 갖지 못하고 있다. 또한 위의 예시에서 보았듯, "도착한"과 같은 행위에 대하여서도 의미를 일관성 있고 정확하게 표현하지 못한다. 본 논문에서는 이러한 언어 수준의 의미를 충분하게 표현하기 위하여 프레임넷[8]의 프레임 시멘틱스(이하 프레임) 개념을 적용하였다. 프레임넷은 propBank와 유사한 구조, 즉 술부-논항의 관계 구조를 갖고 있다. 또한 술부-논항 구조를 갖도록 만드는 어휘에 대해서 그 의미의 모호성을 프레임으로 해소하여 준다. 프레임넷에서는 이러한 어휘를 "target"으로 정의하여 사용한다. 프레임은 디비피디아 온톨로지처럼 단순 사실 정보를 표현하는 어휘들을 포함할 뿐만 아니라, 원인과 결과, 감정, 의견, 행동 등의 다양한 의미들에 대해서도 표현 가능하다는 장점이 있다. 예를 들어, "최초의"는 frame:First_experience, "도착한"에 대하여서는 frame:Arriving 과 같은 프레임이 태깅된다. 이러한 프레임은 종종 지식베이스의 스키마로서 사용되기도 한다[9]. 본 논문에서는, 한국어 질의에 대한 의미 이해 방법으로서, 프레임 구조에 기반한 모형 쿼리(pseudo query)를 생성하는 것을 목표로 하며, 현재의 연구 범위에서는 단문 질의와 단답형 질의를 대상으로 수행하였다. 아래 표 1에서 위에서 상술된 정보추출(IE) 방식과, 전통적인 의미 분석(SP) 방식, 그리고 본 논문과의 연구 위치를 비교하였다.

	IE 방식	SP 방식	본 논문
초점	지식베이스	언어적 의미	언어적 의미
도메인	특정 도메인	오픈 도메인	오픈 도메인
KB	의존적	독립적	독립적
스키마	KB	단어	프레임

표 1 본 논문의 연구 위치 비교표

2장에서는, 한국어 질의 이해의 조건과 그 프레임 기반의 의미 분석에 대해 상술하고, 3장에서는 그 방법론, 4장에서는 평가 및 논의를 기술하고 결론을 5장에 기술한다.

2. 질의응답 시스템을 위한 한국어 질의 이해

2.1 자연언어 질의 이해 태스크의 요구사항

자연언어 질의 이해 태스크의 요구사항이란, KBQA 관점 하에서, 질의를 기계가 읽을 수 있는 형태의 쿼리로 변경할 때 최소한으로 분석되어야 할 구성요소를 의미한다. 예를 들어 다음 질의:

"이순신 장군이 1597년에 명량해협에서 지휘한 해전은 무엇인가?"

이때, 이 질의에 대하여 한국어 디비피디아를 대상으로 한 SPARQL 쿼리는 다음이 적절해 보인다.

```
SELECT ?x WHERE {
    ?x rdf:type dbo:MilitaryConflic .
    ?x dbo:commander dbr:이순신
    ?x dbo:place dbr:명량해협 .
    ?x dbo:date "1597" . }
```

전통적으로 지식베이스 질의응답 시스템에서는 아래의 세 요소를 질의 이해의 주요 요소로 여긴다[4].

(1) 정답의 유형 (dbo:MilitaryConflict)
(2) 의문사 (무엇)
(3) 정답의 근거 (dbo:commander/place/date 등)

지식베이스에서 각각의 개체들은 온톨로지 클래스에 의해서 그 의미가 정의되고(예: 사람, 장소, 사건 등) 이러한 정보는 지식베이스에서 정답 후보들을 선택하는데 있어서 검색공간을 줄이는 효과 및 보다 모호성이 없는 정답을 선택하는데 도움이 된다(예: 영화 "명량"과 사건 "명량해전"을 구분). 따라서 자연언어 질의를 이해할 때에 정답의 유형을 질의에서 발견하여 분석(1)하는 것은 주요 태스크가 된다. 또한 의문대명사(2)의 경우 질의에서 묻고자 하는 화자의 의도를 파악할 수 있다. 예를 들어 "가장 높은 산은?", "몇 개?", "누구인가?" 등으로부터 질문에서 얻고자 하는 정답을 얻는 방식이 달라지고, 따라서 쿼리의 형태도 달라지기 때문이다[10]. 그리고 질의에서 제공하는 정답의 근거들(3)은 SPARQL 쿼리에서 <?answer, p, o> 와 같은 트리플 패턴으로 작성되어 수많은 지식베이스의 개체들 중에서 가장 정답에 가까운, 즉 트리플 패턴의 조건들에 부합하는 개체들을 정답으로 내어주게 된다. 본 논문에서는, 한국어 질의를 충분히 의미 분석하는 것을 목표로 하고 있어 위의 세 가지 요소를 모두 발견하는 시스템을 개발하고자 한다.

2.2 QAF: 프레임 기반 모형 쿼리

본 시스템을 개발하기 전에 한국어 질의에 대한 프레임 분석의 유용성을 검토하기 위하여 [11]에서 사용된 질의 데이터인 NLQ400을 사용하였다. NLQ400은 384개의 한국어 질의로 구성되어 있고, 역사, 과학, 예술 등 다양한 도메인으로 이루어 져 있다. 본 논문은 단답형 질의를 연구 범위로 하여, 위키피디아 1개 문장으로 답변 가능한 질의 95개를 선별하였고, 이 중 객관식이나 O/X 문제 등을 제외한 단답형 문제 72개에 대한 수작업 프레

임 어노테이션 작업을 수행하였다. 위의 예로든 질의에 대한 어노테이션 결과는 다음과 같다.

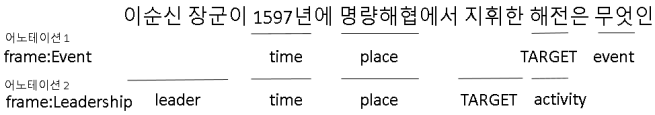

그림 1 질의 “이순신 장군이 1597년에 명량해협에서 지휘한 해전은 무엇인가?”에 대한 프레임 어노테이션

그림 1에서, “해전”은 frame:Event을 만드는 target 어휘이고, 이 어휘에 대한 논항들은 frame:Event에 대한 time, place, event 로서 그 역할이 정의된다. 이 어노테이션에서 의문사인 “무엇”을 event 로 어노테이션 함으로서, frame:Event 의 실제 사건으로 고려하였다. 이는, time:1597년, place:명량해협 에 해당하는 사건의 이름을 ?x로 생성하고자 하는 의도이다. 또한 동사 “지휘한”의 경우, frame:Leadership을 만들고, 그 각각에 대한 논항으로서 leader:이순신 장군, time:1597년, place:명량해협, activity:해전 등을 갖는다. 그러나 어노테이션 2 에서는 의문사 “무엇”에 대하여서는 논항으로 갖지 않는다.

이러한 주석을 통하여 본 연구팀은 정답의 유형을 의미하는 어휘 “해전”에 대한 어노테이션 1, 즉 의문사를 포함하는 어노테이션은 (1)정답의 유형 및 (2)의문사를 논항으로 갖고 있으며, (3)정답의 근거에 대한 정보는 어노테이션 2에서 나타난다는 것을 확인하였다. 특히 72개 질의의 경우 모두 어노테이션 1의 유형을 모두 포함하고 있었다. 또한 어노테이션 1과 어노테이션 2의 경우 대명사 “해전”을 공통의 논항으로 갖는 연결노드가 있었다. 이러한 어노테이션으로부터 질의는 아래와 같은 프레임 그래프로 표현될 수 있다.

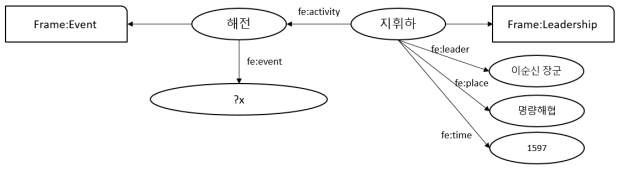

그림 2 질의 “이순신 장군이 1597년에 명량해협에서 지휘한 해전은 무엇인가?”의 프레임 그래프

이때, 두 어노테이션을 연결하는 노드인 “해전”에 대해서는 편의상 Q-Frame으로, 의문사 “무엇”에 대하여서는 Q-FE으로 부르고, 정답에 대한 근거를 표현하는 프레임인 “지휘하” 및 그 논항들(이순신 장군, 명량해협, 1597)은 Sub-frame으로 부른다.

본 논문에서는 위와 같은 그래프를 QAF(Question Answering with Frame Semantics)로 부르며, 질문에서 분석되어야 할 필수요소들이 충분히 분석되는 구조로서 보았다. 이때 질의에서 나타난 조사는 의미 분석에서 주요한 특질로서 고려되지만, QAF에서는 실제 논항이 아니라고 간주하여 제외하도록 하였다. 본 논문에서는 입력으로 한국어 질의를 받고, 출력으로 QAF를 내어주는 시스템 개발을 목표로 하였다.

3. 한국어 질의 프레임 의미 분석

본 시스템을 개발하기 위하여 다음의 목표들을 산정하였다.

(1) 적은 학습데이터의 사용

영어 프레임넷은 양질의 데이터가 오랫동안 구축되어 왔고, 이를 통해 154,607의 예문 및 3,256개의 학습데이터 문장을 사용하여 프레임 의미 분석기를 개발한 바 있다[12]. 한국어의 경우, [13]을 통해 구축된 한국어 프레임넷 코퍼스가 존재하지만, 전체 문장이 5,507개로서 학습데이터로서 부족하고, 영어의 경우 19,582개의 target 어휘를 확보하고 있지만 한국어의 경우 6,802개의 target 어휘가 있어 커버리지 측면에서도 상대적으로 부족하다. 또한 질의에 대한 프레임넷 코퍼스는 한국어는 물론 영어에 대해서도 진행된 바 없는 것으로 알려져 있다. 따라서 본 연구는 한국어 의미 분석을 위해 최소한의 학습데이터를 사용하기 위해 기존에 공개된 자연언어처리 도구를 활용하는 방향으로 진행되었다.

(2) 자연언어 질의에 대한 커버리지

본 논문에서는 다양한 한국어 질의의 형태에 대하여, 질의로부터 이해되어야 하는 요소들을 모두 분석하는 것을 목표로 삼았다. 이에 대응하기 위하여 정보추출 방식이 아닌 의미 분석 방식을 채택하였고, 전통적 의미 분석 방식의 한계인 어휘의 의미모호성 문제는 프레임 의미 모호성의 문제로 국한시켰다.

(3) 표준화된 형식 사용

본 논문에서 개발된 시스템은, 자연언어 질의에 대해 프레임 구조로 밝혀진 술부-논항 구조로부터 기계가 읽을 수 있는 쿼리를 만드는 데에 사용될 계획이다. 이때, 기존의 지식베이스의 형태에 맞춰진 정보추출 방식이 아니기 때문에, 가상의 지식베이스가 있다고 가정하고 프레임을 통해 표현할 수 있는 모든 언어적 의미를 표현하는 데에 초점을 맞추었다. 이러한 결과는 사용의 편의를 위하여 JSON 포맷과 RDF 포맷의 두 가지 형태로 출력된다. 여기에는 추후 디비피비아와 같은 실제 지식베이스에 적합한 SPARQL 쿼리로 변경하는 모듈을 개발하는데 있어 상호운용성을 확보하고자 하는 의도가 있다.

3.1 Q-Frame 및 Q-FE 발견

우리는 72개의 한국어 질의에 대해, 크게 다음의 세 가지 유형이 있음을 발견하였다.

유형 1	의문대명사를 포함한 질의
유형 2	의문대명사를 포함하지 않는 질의
유형 3	평서문 형태의 질의

표 2 본 논문에서 고려한 한국어 질의의 유형

유형 1은 전형적인 단답형 질의로, 예를 들어 "...한 해전은 무엇인가?"와 같은 유형이다. 이러한 유형에 대해서는 우리가 얻고자 하는 정답에 대한 대명사가 의문사의 형태로 "무엇"으로 표현되어 있다. 유형 2는 의문대명사가 생략된 표현으로, 예를 들어 "...한 해전은?"과 같다. 유형 3은 물음표 기호나 의문사가 없는 평서문 표현으로 "...한 해전을 말해보시오"와 같은 표현이다. 본 시스템에서는 입력 질의에 대해 규칙을 적용하여 위의 세 유형으로 분류하는 질의유형분류 모듈을 개발하였다. 해당 모듈에 따라 밝혀진 질의의 유형에 따라 질의의 의존구문구조 속에서 Q-Frame과 Q-Fe를 발견하는 아래 표 4와 같은 룰을 적용하였다. 예컨대 "...지휘한 해전은 무엇?"의 질의의 경우, 의존구조에서 뿌리노드의 태그를 확인 후(NP), 뿌리노드의 자식노드(dist=1)의 태그를 확인(NP_SBJ), 이후 유형을 분류한 뒤 Q-Frame의 target을 "해전"으로 발견한다.

질의	유형	의존구조 정보	뿌리노드
..무엇인가?	1	NP_SBJ, dist=1	VNP
..해전은 무엇?	1	NP_SBJ, dist=1	NP
..한 해전은?	2	NP_SBJ, dist=0	NP_SBJ
..한 해전?	2	NP, dist=0	NP
..말해보시오	3	NP_OBJ, dist=0	VP

표 3 질의의 유형에 따른 Q-Frame 발견 규칙

이후, 각 유형에 따라 Q-FE를 발견하는데, 유형 2 및 3의 경우에는 의문대명사가 생략된 것으로, 가상의 노드를 만들어 Q-FE로 간주하고 유형 1의 경우에는 뿌리노드의 의문사를 Q-FE로 발견한다(그림 3 참고).

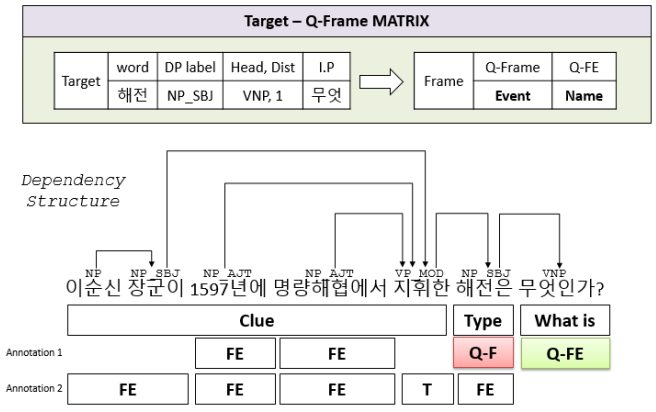

그림 3 의존구조 정보를 사용한 Q-Frame, Q-FE 발견 및 어휘-프레임 매칭

그림 3에서 볼 수 있듯, 어휘 "해전" 및 의문사 "무엇"이 특정 조건하에서 발견되었을 때, 해당 조건에 따라 Q-Frame을 "해전", Q-FE를 "무엇"으로 발견한 뒤, 해당 어휘에 대하여 사전에 구축된 문자열-프레임 매핑 테이블을 사용하였다. 이러한 Q-Frame은 정답의 유형의 모호성 해소를 위해 사용되는데, "해전"의 경우 frame:Event로 매핑되어, 정답의 유형은 Event라고 볼 수 있다.

3.2 Sub-Frame 발견

Sub-Frame의 목적은 질의에서 나타난 정답의 근거를 쿼리에 포함하는 것이다. 예를 들어 SPARQL 쿼리에서는 <?x, p, o> 와 같은 트리플 패턴을 의미한다. 예컨대 한국어 질의 "이순신 장군이 1597년에 명량해협에서 지휘한 해전은 무엇인가?"에서 "이순신 장군", "1597년", 그리고 "명량해협"과 같은 정보들이다. 이러한 트리플 패턴을 생성하기 위해 본 논문에서는 의미 분석으로서 프레임의 술부-논항 구조를 적용하였다. 프레임은 ProbBank 스타일의 술부-논항 구조를 가지며, 술부 및 논항의 역할에 대한 보다 자세한 의미를 부여한다는 점에서만 차이가 있다. 본 논문에서는 이러한 술부-논항 구조를 밝혀내기 위한 목적으로 한국어 의미역 분석(SRL)도구를 사용하였다[14].

그림 3에서와 같이, 한국어 SRL 도구의 술부를 Sub-Frame의 target 어휘를 발견하는데 사용하였고, 각각의 논항을 프레임 구성요소로서 사용하였다. 위와 같은 처리를 통해 발견된 target 어휘에 대하여 프레임을 매핑하기 위하여 한국어 프레임넷[4]의 6,820개의 LU(Lexical Unit), 즉 각 프레임에 해당하는 어휘들과의 문자열 매칭 방법으로 이루어졌다.

그러나 몇몇 질의의 경우 SRL 도구에서 술부-논항 구조가 밝혀지지 않는 경우가 존재하고, 혹은 발견되어야 할 논항을 충분히 발견하지 못하는 등의 문제가 발생하였다. 특히, QAF 표현에서는 Q-Frame 논항이 프레임 술부-논항 그래프들을 연결시키는 중간노드라는 점에서, 이러한 논항을 발견하지 못하는 경우에는 문제가 된다.

본 논문에서는 이러한 문제들을 해결하기 위해 다음과 같은 후처리 작업을 수행하였다.

(1) 동사가 없는 질의

예를 들어, 질의 "신간회의 구성원은 누구인가?"와 같은 경우, 동사가 없어 ProbBank 스타일의 의미역분석 도구에서는 술부-논항 구조를 발견하지 못한다. 그러나 이 경우, 어휘 "구성원"은 정답의 유형이 '사람'임을 의미하고, "누구"의 경우에는 그 사람의 '이름'을 알고자 하는 의도가 포함되어 있다. 그리고 일종의 재귀적 표현으로서, "신간회의 구성원"이라는 어휘 자체에 정답에 대한 근거가 포함되어 있다. 따라서 이러한 질의의 경우 술부-논항 구조가 발견되지 못한다고 하더라도 쿼리의 형태로 변환해 주어야 한다. 본 시스템에서는 위 질의에 대해 다음과 같은 정보를 정답에 대한 근거 트리플 패턴으로서 내어준다:

<구성원, description, 신간회의 구성원>.

(2) 발견되지 않은 주요 논항에 대한 처리

질의 "이순신 장군이 1597년에 명량해협에서 지휘한 해전은 무엇인가?"에 대한 그래프 표현인 그림 2에서 볼 수 있듯, 어휘 "해전"은 두 개의 프레임 그래프를 연결시켜주는 연결노드로서 역할을 한다. 그러나, 술부-논항 구조에서 두 그래프 중 하나라도 해당 어휘를 논항

4) http://framenet.kaist.ac.kr

으로 발견하지 못하는 경우에는 두 그래프를 하나의 그래프로 합치는 처리를 하지 못한다. 이를 위해 SRL에서 Q-Frame의 논항을 발견하지 못하는 경우 이를 해당 술부에 대한 논항으로 포함하도록 후처리 하였다.

(3) 각 술부-논항 그래프의 연결

SRL 결과에서 발견된 복수개의 술부-논항 그래프의 경우 독립적인 어노테이션 층으로 존재하여, 두 그래프를 연결해주는 모듈의 개발이 필요하다. 본 시스템에서는 각 논항들의 문장 내 위치 정보를 사용하여 어노테이션들을 연결하는 모듈을 개발하여 포함하였다.

(4) 논항에 대한 구 묶음

현재 존재하는 한국어 SRL 도구는 여러 어절로 이루어진 명사 구 중, 가장 마지막 어절에 대하여서만 논항으로 인식하여준다. 본 논문에서는 의존구조에서 연속적으로 연결된 명사구에 대해 하나의 논항으로 인식하여 주는 후처리 구 묶음 처리를 수행하여 주었다. 이때 접속조사로 연결되었을 경우, 술부와 동일한 의미역을 갖는 두 개의 논항으로 분류하여 주었다. 또한 실제 논항이 아닌 조사 등에 대하여서는 논항에서 제외하는 후처리를 수행하였다.

예를 들어 "이순신 장군과 안위 장군이 1597년에 지휘한 해전은 무엇?"이라는 질의의 경우, SRL 도구에서는 **지휘하(장군이, 1597년에, 명량해협에서)**의 술부-논항 정보를 파악하지만, 본 시스템에서는 다음과 같은 결과를 내어준다: **지휘하(이순신 장군, 안위 장군, 1597년, 명량해협)**

그림 4과 그림 5는 위의 SRL 결과에 대한 후처리와 프레임 정보가 매핑된 결과를 보여주는 예시이다. 그림 4에서 볼 수 있듯 SRL 도구는 "설립되어", "합병되"의 정보를 술부-논항 구조로 파악하고, 본 시스템에서 Q-Frame의 target으로서 "회사"를 발견한다. 위의 후처리 과정을 통하여 그림 5와 같은 결과로서 수정되어, 서로 연결되지 않았던 술부-논항 그래프가 연결되고, 또한 각각의 술부에 대하여 프레임 정보가 부착되어 의미 모호성이 해소되는 것을 볼 수 있다.

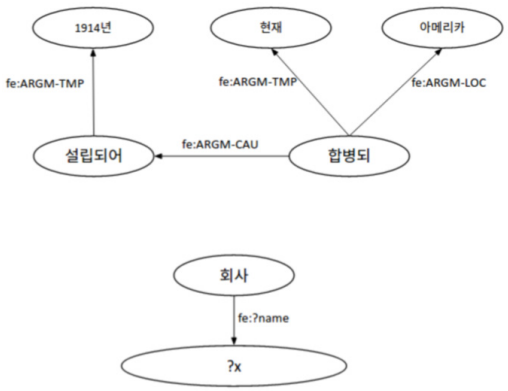

그림 4 질의 "1914년 설립되어 현재 뱅크 오브 아메리카에 합병된 회사는 무엇인가?"에 대한 SRL 결과

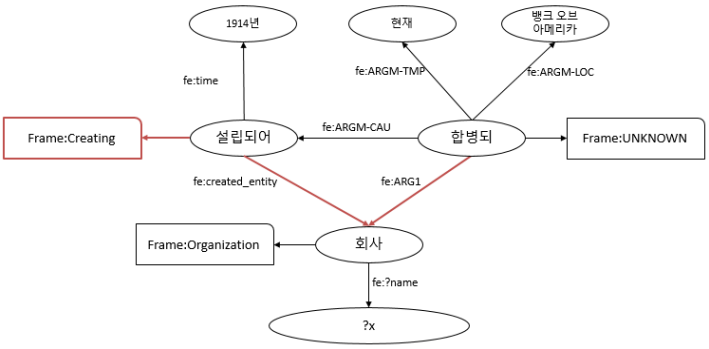

그림 5 질의 "1914년 설립되어 현재 뱅크 오브 아메리카에 합병된 회사는 무엇인가?"에 대한 본 시스템의 결과

3.3 QAF 결과

위의 처리 등을 통하여, 본 시스템은 자연언어 질의에 대한 RDF표현에 기반한 모형 쿼리인 QAF를 생성해 준다. 본 논문에서 계속해서 사용된 예시인 질의 "이순신 장군이 1597년에 명량해협에서 지휘한 해전은 무엇인가?"에 대한 본 시스템의 출력은 아래와 같다:

```
frdf-event:해전#1
        rdf:type frame:Event ;
        fe:event ?answer .

frdf-event:지휘하#1
        rdf:type frame:Leadership ;
        fe:leader "이순신 장군" ;
        fe:time "1597년" ;
        fe:place "명량해협" ;
        fe:activity frdf-event:해전#1 .
```

각 target 어휘인 "해전" 및 "지휘하"의 경우, 고유의 URI값(예: frdf-event)을 부여받아, RDF 트리플에서 하나의 SUBJECT로서 역할을 수행하고, 각각의 논항 및 논항의 역할은 OBJECT와 PREDICATE으로 표현된다. 이러한 RDF 표현은 한국어 디비피디아와 같은 개체-개체간의 정보를 표현하는 이항관계(binary relation)가 아닌, 하나의 사건(위의 예시에서는 "해전" 및 "지휘하")에 대한 각각의 논항들을 RDF로 표현한다는 점에서 차이가 있다. 위와 같은 이벤트 중심의 지식표현은 개체-개체 관계에서의 관계에 대한 부가적인 속성을 기술하는데 적절하다. 이러한 모형 쿼리인 QAF는 추가 연구를 통해 DBpedia나 Freebase와 같은 지식베이스에 적합한 SPARQL로 변경될 수 있다[9].

4. 평가 및 논의사항

4.1 성능평가

평가는 OKBQA에서 사용된 NLQ50 평가 데이터셋을 사용하여 수행되었다. 이때 50개 질의 중, O/X 질의나 서술형 질의를 제외한 45개 질의에 대해 평가하였다.

본 시스템은 45개 질의 모두에 대하여 각각 1개씩의 Q-Frame을 발견하였고, 추가적으로 Sub-Frame을 생성하는 51개의 target 어휘를 발견하였다(평균 2.13개). 전체 96개 target 어휘에 대하여 58개의 프레임을 부착하였고, 이에 대한 수작업 검증에서 모두 옳은 프레임으로 평가되었다. 그리고 45개에 질의에 대하여 질문에 대한 근거에 해당하는 36개의 논항 정보들을 생성할 수 있었다. 이러한 Sub-Frame들은 수작업 검증을 통하여 0.90의 정밀도(precision)와 0.73의 재현율(recall), 0.8137의 F-1 수치로 평가되었다.

4.2 논의사항

(1) 프레임 매핑: 본 시스템은 target 어휘에 대해 프레임을 부착하기 위하여 한국어 프레임넷의 LU 데이터를 사용하였다. 그러나 해당 데이터셋의 커버리지는 위의 평가에서 60%가량으로 나타나 이에 대한 성능 향상이 필요하다. 본 연구팀에서는 이를 해결하기 위하여 유의어 사전 및 워드임베딩 방법을 적용한 프레임 발견[15] 방법을 적용할 계획이다.

(2) 다양한 형태의 질의 처리: 본 시스템은 단문 단답형 질의에 대하여 처리를 제공하고 있으나, 실제 질의응답시스템을 위해 적용하기 위해서는 복문의 질의나 괄호문제, O/X문제 등에 대한 다양한 처리를 제공할 필요가 있다. 특히 복문 질의에 대한 처리가 차후 연구 목표로 남아있다.

(3) 온톨로지 매핑: 본 시스템은 3장에서 논의된 바와 같이 가상의 지식베이스를 가정한 모형 쿼리 생성을 위하여 질의에서 분석되어야 하는 최대한의 정보를 의미분석하는 데에 목표로 하였다. 실제 지식베이스를 대상으로 한 질의응답 시스템의 개발을 위해서는, 디비피디아와 같은 현존하는 지식베이스의 스키마에 대하여 매핑된 SPARQL 쿼리를 생성하는 온톨로지 매핑이 추가 연구로 남아있다.

5. 결론

본 논문에서는 자연언어 질의를 분석하기 위하여 프레임 구조를 적용한 의미 분석으로서 QAF 모형 쿼리를 생성하는 시스템을 개발하였다. 지식베이스에 의존적인 쿼리 생성은 지식베이스의 불완전성으로 질의에서 나타나는 정보를 충분히 분석하지 못한다는 점에 착안하여, 의미 분석으로서 프레임 구조를 분석하는 것을 목표로 하였다. 추후 연구에서는 복문 질의에 대한 처리 및 실제 지식베이스에의 적용 등을 수행할 예정이다. 본 시스템은 오픈소스로 공개되어 있다:
(https://github.com/machinereading/FRDF)

사사

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임. (No. R0101-16-0054, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

이 논문은 2016년도 미래창조과학부의 재원으로 한국연구재단 바이오의료기술개발사업의 지원을 받아 수행된 연구임.(2015M3A9A7029735)

참고문헌

[1] Kurt Bollacker, Coling Evans, Praveen Paritosh, Tim Sturge, Jamie Taylor, “Freebase: a collaboratively created graph database for structuring human knowledge”, In Proceedings of ACM, 2008.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, “Dbpedia: A nucleus for a web of open data”, In The semantic web. Springer Berlin Heidelberg, 722-735. 2007.

[3] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, Gerhard Weikum, “YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia”, Artificial Intelligence, 194, 28-61. 2013.

[4] Xuchen Yao and Benjamin Van Durme, “Information extraction over structured data: Question answering with freebase”, In proceedings of ACL, 2014.

[5] Jonathan Berant, Percy Liang. “Semantic Parsing via Paraphrasing”, In ACL (1) (pp. 1415-1425). 2014.

[6] Younggyun Hahm, Youngsik Kim, Yousung Won, Jongsung Woo, Jiwoo Seo, Jiseong Kim, Seongbae Park, Dosam Hwang, Key-Sun Choi. “Toward Matching The Relation Instantiation From DBpedia Ontology To Wikipedia Text: Fusing FrameNet To Korean”, In proceedings of iSemantics. 2014.

[7] Xuchen Yao, Jonathan Berant, Benjamin Van Durme, “Freebase QA: Information Extraction or Semantic Parsing?”, In ACL 2014, 82. (a), 2014.

[8] Collin F. Baker, Charles J. Fillmore, John B. Lowe, “The berkeley framenet project”, In Proceedings of ACL, 1998.

[9] Jacobo Rouces, Gerard de Melo, Katja Hose, “FrameBase: representing n-ary relations using semantic frames”, In European Semantic Web Conference(pp. 505-521), 2015.

[10] Christina Unger, Corina Forascu, Vanessa Lopez, Axel-Cyrille Ngonga Ngomo, Elena Cabrio, Philipp Cimiano, Sebastian Walter, “Question answering over linked data (QALD-4)”. In Working Notes for CLEF 2014 Conference.. 2014.

[11] Sangha Nam, Younggyun Hahm, Sejin Nam, Key-Sun Choi, “Design and Implementation of an Evaluator for Building a Good Knowledge Base in Question Answering”, ISWC 2015 Workshop on Natural Language Interfaces for Web of Data, NLIWoD, 2015.

[12] Dipanjan Das, Nathan Schneider, Desai Chen, Noah A. Smith, “SEMAFOR 1.0: A probabilistic frame-semantic parser”. Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2010.

[13] Jungyeul Park, Sejin Nam, Youngsik Kim, Younggyun Hahm, Dosam Hwang, Key-Sun Choi, “Frame Semantic Web: a Case Study for Korean”, ISWC, 2014.

[14] Soojong Lim, Changki Lee, Pum-Mo Ryu, Hyunki Kim, Sangkyu Park, Dongyul Ra, “A Domain Adaptation Technique for Semantic Role Labeling with Structural Learning”, In ETRI Journal vol. 36, no. 3, June. 2014, pp429-438, 2014.

[15] Karl Moritz Hermann, Dipanjan Das, Jason Weston, Kuzman Ganchev, “Semantic frame identification with distributed word representations”, In ACL. 2014.