지식기반 질의응답을 위한 질문분석 방법

허정, 황이규, 최미란, 장명길
한국전자통신연구원, 미래기술연구본부, 음성언어정보연구부, 지식마이닝연구팀
{jeonghur, yghwang, miranc, mgjang}@etri.re.kr

Question Analysis for Knowledge based Question/Answering

Jeong Heo, Yi-Gyu Hwang, Mi-Ran Choi, Myung-Gil Jang
Knowledge Mining Research Team, ETRI

요 약

AnyQuestion 1.0은 ㈜두산의 '두산세계대백과 엔싸이버'의 인물분야만을 대상으로 한 질의응답형 정보검색 시스템이다. 본 시스템에서는 지식기반 질의응답, Logical Form 기반 질의응답, 단락 기반 질의응답을 통합한 3단계 정답 추출 방법을 제안하고 있다.

지식기반 질의응답은 본문의 구조화된 정보와 비구조화된 정보로부터 정보추출 기술을 이용하여 구축한 지식베이스에 대한 질의응답을 목적으로 한다. "사용자의 질문에 대한 정답을 지식베이스에서 제시할 수 있는가?"와 "지식베이스에서 어떤 정보를 정답으로 제시해야 하는가?"는 3단계 정답 추출 방법에서는 상당히 중요하다. 이를 위해서 질문 분석에서는 수동으로 구축한 지식베이스 속성 자질 정보와 다양한 규칙을 기반으로 질문 분석을 수행하였고, 이를 이용하여 지식기반 질의응답을 하였다. 실험결과, 지식기반 질의응답 할당 재현율은 65.4%, 지식기반 질의응답의 정확률은 81.25%였다.

백과사전 인물분야에 대한 지식기반 질의응답은 기존의 데이터베이스 분야에서 연구되어온 자연어 DB 인터페이스를 활용한 질의응답으로 속도가 빠르며, 상대적으로 높은 정확률을 보였다.

1. 서 론

디지털 정보들 중 사용자가 원하는 정보를 찾기 위해 다양한 검색기술이 연구되어 왔다. 그러나, 인터넷 기술의 발전과 더불어 범람하는 정보들로 인해, 일반적인 정보검색 기술을 이용한 검색 결과가 정확한 정보를 빠르게 얻고자 하는 사용자의 요구를 충족시키지 못하게 되었다. 일반적인 정보검색 기술들은 방대한 양의 결과를 사용자에게 제시한다. 사용자는 원하는 정보를 찾기 위해 검색결과를 검토하는 노력을 해야 한다. 사용자들은 이러한 부차적인 노력을 회피할 수 있는 새로운 정보 검색 기술을 기대하고 있다.

질의응답(Question Answering : QA) 기술은 사용자의 질문을 분석하여 정확한 의도를 파악한 후, 사용자가 원하는 정확한 정보를 다양한 문서로부터 추출하여 답으로 제시하는 기술이다. 따라서, 검색결과를 다시 검토하는 수고를 회피할 수 있다. 그러나, 질의응답 기술은 정밀한 언어처리 기술, 문서검색 기술, 정보추출 기술 및 추론 기술 등이 요구된다.

질의응답 시스템은 기술적인 측면을 고려할 때, 크게 세가지로 나눠 볼 수 있다.

첫째, 최근 검색 포털 업체들이 제시한 새로운 검색 서비스 모델인 '지식검색[1]'이 있다. 사용자가 자연어 질문을 통해 검색을 수행하면, 기존의 유사한 질문들이 제시된다. 이 중 가장 유사한 질문을 선택하여 그 질문의 정답을 보고 사용자가 정보를 찾는다. 사용자가 기존 지식검색에서 정보를 찾지 못하면, 게시판에 질문을 올리고, 다른 사용자가 질문에 대한 정답을 제시한다. 제시된 정답에 대한 신뢰도 및 유용성에 대한 평가는 사용자들이 한다[1].

둘째, 데이터베이스로 구축되어 있는 다양한 정보를 자연어 질의로 검색하는 자연어 DB 인터페이스(Natural Language DataBase Interface)가 있다. 정보추출 기술을 이용하여 문서들로부터 지식을 추출하여 지식 DB에 저장하고, 사용자의 질문을 분석하여 지식

1) 네이버(www.naver.com)의 '지식iN',
엠파스(www.empas.com)의 '지식거래소',
세이클럽(www.sayclub.com)의 '세이테마',
야후(www.yahoo.co.kr)의 '야후!지식검색'

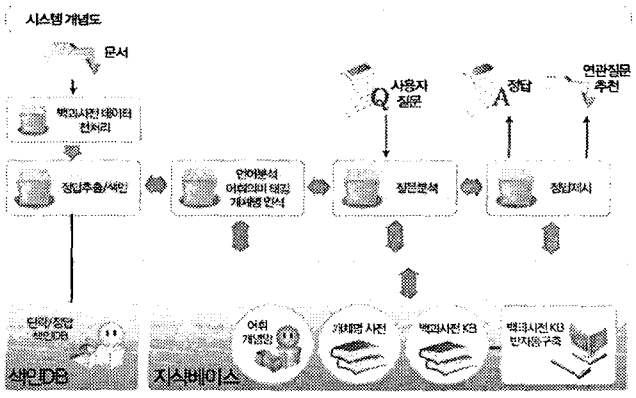
DB에 대한 구조화된 질의 언어(Structured Query Language)로 변환하여 정답을 제시한다[2,3,4].

셋째, 일반적인 형태의 질의응답로써, 사용자의 질문을 분석하여 질문의 의도를 파악하고, 질문에 사용된 키워드를 기반으로 문서검색을 수행한다. 검색된 문서를 대상으로 사용자가 원하는 정보를 정보추출 기술을 이용하여 제시한다[5,6].

본 논문에서 언급하고 있는 AnyQuestion 1.0은 자연어 DB 인터페이스를 이용한 질의응답과 일반적인 질의응답의 하이브리드된 형태로 3단계 정답추출 방법이 있다. 1단계는 정보추출 기술을 이용해 구축된 지식베이스(Knowledge Base)를 대상으로 정답을 추출하는 지식기반 정답 추출이다. 2단계는 언어 분석된 결과를 논리적인 형태(Logical Form)로 색인하고, 질문의 LF형식을 색인과 매칭하여 정답을 추출하는 LF 기반 정답 추출이다. 마지막 3단계는 본문을 단락단위로 나눠 주제를 색인하고, 질문의 주제와 동일한 단락을 검색하는 단락 기반 정보 추출이다. 본 논문에서는 지식기반 질의응답(Knowledge Based Question Answering : KBQA)을 위한 질문 분석 방법을 제안한다. 논문의 구성은 다음과 같다. 2절에서 AnyQuestion 1.0에 대해 간략히 소개하고, 3절에서는 지식베이스의 반자동 구축에 대해서 기술한다. 4절에서는 지식기반 질의응답을 위한 질문 분석 방법을 제안하고, 5절에서는 실험 및 결과를, 6절에서는 결론 및 향후 연구 계획을 기술한다.

2. AnyQuestion 1.0

AnyQuestion 1.0은 ㈜ 두산의 '두산세계대백과 엔싸이버'의 인물분야를 대상으로 질의응답를 수행하는 시스템이다. 본 시스템은 언어처리, 정답색인, 지식베이스 반자동 구축, 질문분석, 정답제시 모듈로 나뉜다[7]. [그림 1]은 AnyQuestion 1.0의 구성도이다.



[그림 1] AnyQuestion 1.0의 구성도

언어처리는 본문과 질문의 자연어 분석을 담당하며 형태소 분석, 어휘의미 분석, 개체명 인식, 문장구조 분석을 수행한다. 개체명 인식은 본문에서 정답이 될 가능성이 있는 어휘들에 대해서 개체명을 인식하고 태그를 부착하는 것으로써 계층적으로 구성된 63개의 개체명 태그가 이용된다[9]. 문장구조 분석은 격틀을 기반으로 문장을 분석하여 논리적인 형태로 제공한다[10].

[표 1] 상위 19개의 개체명

개체명 태그	의미
PERSON	인명, 호, 별칭
ANIMAL	동물
PLANT	식물
DISEASE	병명
OCCUPATION	직업명
POSITION	직위
LOCATION	장소명
QUANTITY	수량
DATE	날짜, 시대
TIME	시간
CHEMICALMAT	화학 성분명
MUSICALINSTRUMENT	악기명
CULTURALASSET	문화재명
TRANSPORT	교통장비
PRIZE	수상명, 상명
WORKS	작품, 저서
THEORY	이론, 정책, 경향
EVENT	사건명

정답색인에서는 본문에 대한 언어분석 결과를 대상으로 정보추출 기술을 이용하여, 정답일 가능성이 높은 정보들을 색인한다.

지식베이스 반자동 구축에서는 본문에 기술되어 있는 다양한 인물 정보들 중, 인물범주 별로 공통되는 정보들을 선정하여, 다양한 규칙와 통계적인 방법으로 인물정보에 대한 지식베이스를 반자동으로 구축한다.

질문분석은 사용자가 입력한 자연어 질문을 분석하여, 사용자가 요구하는 정보의 유형 분석, 지식 검색을 위한 질문 분석과 문서 검색을 위한 키워드 확장 등을 수행한다.

정답제시에서는 질문 분석된 결과를 기반으로 지식베이스와 색인 DB로부터 정답을 검색하고 순위화하여 사용자에게 제시한다. 정답을 제시할 때, 사용자의 질문과 연관성이 많은 질문과 정답을 함께 제시하는 연관질문

추천도 있다.

3. 지식베이스 반자동 구축

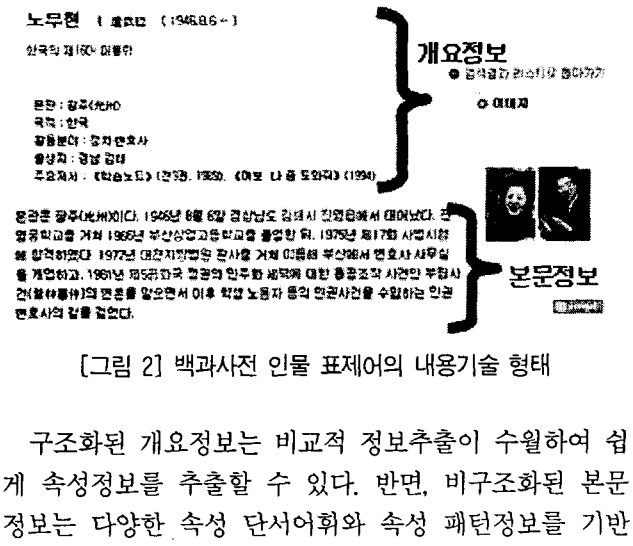
인물에 대해 기술한 문서들에는 출생정보, 사망정보, 업적정보등 다양한 정보들이 개별 인물에 관계없이 공통적으로 출현한다. 이처럼 인물과 관련된 문서에 공통적으로 출현하는 정보들을 인물의 속성이라고 정의하고, 관련된 속성들을 묶어 템플릿이라 정의한다.

백과사전의 인물분야는 인물의 특징에 따라 범주가 나뉘어져 있다. 범주는 계층적인 구조로 최상위 범주는 총 24개로 구성되어 있다. 동일범주에 속한 인물들은 공통된 속성들을 가진다. 과학범주의 인물정보 문서에서는 발견물이나 개발품 등의 속성이 공통적으로 나타난다. 또한, 범주에 관계없이 모든 인물에 공통적으로 표현되는 인물속성이 있다. 출생과 사망에 관련된 속성이 이에 해당한다. 이처럼 범주에 관계없이 모든 인물에서 기술되는 공통속성 21개를 선정하였고, 범주 별로 개별 기술된 속성 31개를 선정하였다. [표 2]은 지식베이스를 구성하는 공통속성의 예를 보여주고 있다.

[표 2] 인물 공통 속성의 예

템플릿	속성
출생	출생장소, 출생일, 국적, 본관
사망	사망장소, 사망일, 사망원인
명칭	별칭
학력	졸업학교, 졸업일

백과사전에서 인물 표제어에 대한 내용기술은 크게 두 부분으로 나뉜다. 구조화된 개요정보와 비구조화된 본문정보로 나뉜다.

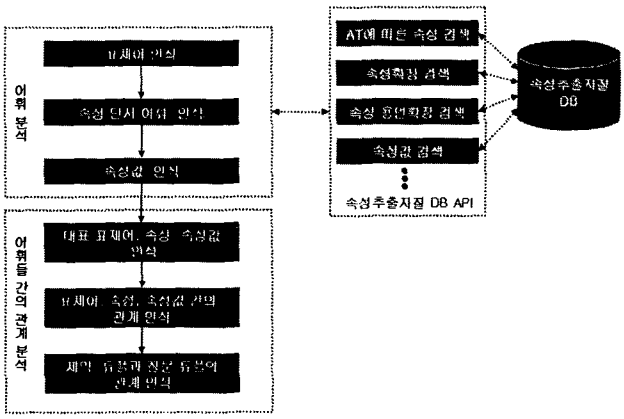


[그림 2] 백과사전 인물 표제어의 내용기술 형태

구조화된 개요정보는 비교적 정보추출이 수월하여 쉽게 속성정보를 추출할 수 있다. 반면, 비구조화된 본문정보는 다양한 속성 단서어휘와 속성 패턴정보를 기반으로 정보추출 기술을 이용하여야 속성정보를 추출할 수 있다[8]

4. 지식기반 질의응답을 위한 질문분석

질문분석에서는 질문의 정답유형 인식, 키워드 추출 및 확장, 질문주제 인식, 지식베이스 속성 인식 등의 작업을 수행한다. 기 구축된 지식베이스로부터 정확한 정답을 제시하기 위해서는 자연어 질문을 분석하여 사용자가 요구하는 정보가 지식베이스에서 제시할 수 있는 인물 속성인지를 파악하여야 한다. 이를 위해서는 질문에 대한 언어 분석과 속성인식을 위한 질문 분석이 수행되어야 한다.



[그림 3] 지식베이스 속성 인식 흐름도

[그림 3]은 지식베이스 속성 인식 흐름도인데, 크게 어휘분석과 어휘들 간의 관계 분석으로 나뉜다. 어휘분석에서는 질문을 구성하는 개별 어휘들이 지니는 정보를 분석하고, 어휘들 간의 관계 분석에서는 개별 어휘들이 가지는 정보를 기반으로 어휘들 간의 정보관계를 분석한다. 이처럼 어휘 정보 및 어휘 정보 관계 분석을 위해서는 어휘들의 속성을 정의한 속성 추출 자질 DB가 이용된다.

4.1. 속성 단서 어휘, 개체명 태그, 질문 정답유형의 관계와 속성 추출 자질 DB

"질문에 대한 정답을 지식베이스에서 제시할 수 있는가?"와 "질문에 대한 정답을 지식베이스의 어느 속성에서 찾을 것인가?"는 지식베이스 속성 인식에서 처리해야 할 가장 중요한 내용이다. 이를 판단하기 위해 이용할 수 있는 정보는 질문의 정답유형, 질문에 출현한 어휘들과 개체명 태그들이 있다.

질문의 정답유형은 질문의 정답에 해당할 가능성이 있는 속성들의 수를 제약시켜준다. 정답유형이 'LOCATION'이라면, 52개의 속성 중, 장소와 관련되는 '출생장소', '사망장소'등과 같은 속성을 대상으로 정답을 찾을 수 있다.

질문에 출현하는 어휘들은 개체명 태그가 붙은 어휘와 그렇지 않은 어휘로 구분할 수 있다. 개체명이 붙은 어휘들은 지식베이스의 각 속성에 대한 값에 해당하고, 그렇지 않은 어휘들은 속성을 제약하는 속성 단서 어휘들이다.

1) 김구는 어디에서 태어났는가?
정답유형 : LOCATION
2) 백범일지를 쓴 사람은?
정답유형 : PERSON

1)에서는 정답유형 'LOCATION'으로 정답을 검색할 대상이 되는 속성을 '출생장소', '사망장소' 등으로 제약한다. 그리고, '태어났는가'는 속성 단서 어휘로 '출생'과 관련된 속성으로 속성값(정답이 있을 속성)을 제약한다. 그러므로, 1)에서는 '김구'의 '출생장소' 속성에 정답이 있다는 것을 알 수 있다.

2)에서는 표제어(사람)를 찾는 질문으로써, '백범일지'가 개체명 인식에 의해서 'WORKS'로 태그가 부착된다. 따라서, '백범일지'는 'WORKS'로 제약된 속성들 중, 하나의 속성에 값으로 저장되어 있다는 것을 알 수 있다. 'WORKS'로 제약되는 속성은 '주요저서', '주요작품' 등이 있다. 그러나, '쓴'은 속성 단서 어휘로써, 'WORKS'에 의해 제약된 속성들 중, '주요저서' 속성으로 제약을 한다. 그러므로, 2)에서는 '주요저서' 속성의 값으로 '백범일지'가 저장되어 있는 표제어를 정답으로 제시할 수 있게 된다.

이처럼 속성을 제약할 수 있는 단서어휘, 개체명 태그, 질문의 정답유형 등을 속성별로 정리한 지식정보가 속성 추출 자질 DB이다. 속성 추출 자질 DB는 지식베이스 설계와 지식베이스 반자동 구축 시 고려된 여러 정보를 기반으로 반자동으로 구축하였다. 또한, 다양한 질문 코퍼스를 분석하여, 질문 중에 나타나는 어휘들의 속성을 분석하여 반영하였다.

4.2. 질문튜플과 제약튜플

지식베이스 속성을 인식하여, 구조화된 질의 언어(SQL)로 변환하기 위해서는 표제어, 속성, 속성값 정보가 반드시 있어야 한다. 표제어는 누구에 대한 질문인지를 제시하는 정보이고, 속성은 표제어의 어떤 정보를 요구하는지를 제시하는 정보이다. 속성값은 해당 속성에 들어있는 값을 의미하는데, 질문의 유형에 따라 질문의 정답유형이나 질문 내에서 개체명 태그가 붙은 어휘가 된다.

지식기반 질의응답의 질문은 크게 두 종류로 구분할 수 있다.

A. 특정 인물의 속성에 해당하는 값을 요구하는 질문.
→ "박정희의 고향은 어디인가?"
#정답유형 : LOCATION
B. 속성으로 특정 값을 가지는 인물을 요구하는 질문.
→ "로미오와 줄리엣의 저자는?"
#정답유형 : PERSON

A유형의 질문에서 표제어는 질문 내에서 PERSON으로 개체명 태깅이 된 어휘가 된다. 그리고, 속성값은 질문의 정답유형이 되고, 이 속성값과 속성 단어 어휘를 분석하여 속성을 결정하게 된다. 반면, B유형에서는 표제어는 질문의 정답유형이 되고, 속성값은 질문 내에서 'WORKS'로 개체명 태깅이 된 '로미오와 줄리엣'이 된다. 속성은 속성값의 개체명 태그와 속성 단어 어휘를 분석하여 결정된다.

앞서 언급된 바와 같이, 질문을 분석하여 구조화된 질의 언어로 변환하기 전에 요구되는 표제어, 속성, 속성값을 묶어서 튜플이라고 정의한다. 튜플의 형태는 크게 5가지로 나뉜다.

I. 〈표제어, 속성, ?정답유형〉
II. 〈?PERSON(정답유형), 속성, 속성값〉
III. 〈표제어, 속성, 속성값〉
IV. 〈V_PERSON, 속성, 속성값〉
V. 〈V_PERSON, 속성, ?정답유형〉

Ⅰ과Ⅱ는 A와 B유형의 질문에 대한 튜플에 해당한다. A유형 질문의 예에 대한 튜플은 (박정희,출생장소,?LOCATION)이고, B유형의 예에 대한 튜플은 (?PERSON,주요저서,로미오와 줄리엣)이다. Ⅲ유형의 튜플은 표제어를 제약하는 튜플로써, Ⅰ유형의 튜플과 함께 출현한다. Ⅳ과 Ⅴ유형은 함께 생성되어야 하고, 표제어가 없고, 정답유형이 PERSON이 아닌 질문에서 생성된다.

3) 노무현 대통령이 쓴 책은?
정답유형 : WORKS
튜플 : 〈노무현,직위,대통령〉 → 〈노무현, 주요저서, ?WORKS〉
4) 중일전쟁이 발생한 해는?
정답유형 : DATE
튜플 : 〈V_PERSON, 사건명, 중일전쟁〉

→ 〈V_PERSON, 사건일, ?DATE〉

3)의 예문에서는 〈노무현, 직위, 대통령〉 튜플이 〈노무현, 주요저서, ?WORK〉튜플을 제약한다. 백과사전의 인물 중에는 동명이인이 많은데, 위 질문에서는 '노무현' 이라는 사람들 중에 '대통령'의 직위를 가진 '노무현'의 '주요저서'를 정답으로 요구하는 것이다. 4)의 예문에서는 〈V_PERSON, 사건명, 중일전쟁〉 튜플이 〈V_PERSON, 사건일, ?DATE〉튜플을 제약하는데, 템플릿 내의 속성들은 서로 연관성을 가진다. '사건' 템플릿에 속한 '사건명'와 '사건일' 속성은 서로 연관되어 있으므로, 위의 제약관계에 의해 정답을 제시할 수 있다. 이처럼 튜플은 제약을 하는 튜플과 제약을 받으면서 질문의 정답을 찾는 튜플로 나뉠 수 있다. 제약을 하는 튜플을 제약튜플이라 정의하고, 질문에 대한 정답을 찾는 튜플을 질문튜플이라 정의한다. 제약튜플과 질문튜플의 관계는 '제약튜플질문튜플'로 표현한다.

4.3. 속성 중의성 문제

질문 내에 포함되어 있는 속성단서 어휘와 개체명 태그나 정답유형을 이용하여 속성값에 대한 속성을 인식하는 것은 질문에 대한 정답을 지식베이스의 어느 속성에서 찾을 것인가를 결정하는 중요한 과정이다. 그러나, 속성 단서 어휘가 제약하는 속성과 개체명 태그가 제약하는 속성이 언제나 유일한 것은 아니다. '작품'이라는단서어휘는 '주요작품', '건축물', '주요저서' 등에 대한 단서이고, 개체명 태그 'WORKS'는 '주요작품', '주요저서' 등 다양한 속성들과 관련된다. 이처럼, 여러 속성과 관련되는 속성 단서 어휘와 개체명 태그간에 일치하는 속성은 반드시 하나일 수 없는 것이다. 이럴 경우 속성 중의성이 발생했다고 한다.

5) 레오나르도 다빈치의 작품은 무엇이 있나?
정답유형 : WORKS

5)의 예문에서는 단서어휘 '작품'에 의해 '주요작품', '건축물'과 '주요저서' 속성으로 속성값이 제약되고, 속성값에 해당하는 정답유형 'WORKS'에 의해 '주요작품', '주요저서' 속성으로 제약된다. 속성 단서와 정답유형에 의해 공통적을 제약된 속성은 '주요작품'과 '주요저서'이다. 따라서 5)예문에서는 두 개의 질문튜플이 생성된다. 그러므로, 생성된 두 튜플, 〈레오나르도 다빈치, 주요작품, ?WORKS〉과 〈레오나르도 다빈치, 주요저서, ?WORKS〉의 관계는 OR의 관계를 가지게 된다. 각각이 튜플에서 검색된 결과를 OR연산을 해서 사용자에게 제시하여야 한다.

4.4. 복문형태의 질문처리

연결어미를 이용한 복문형태의 질문도 있다. 이럴 경우, 언어분석에서 복문을 단문형태로 분할한다. 이 정보를 활용하여 단문 별로 튜플을 생성할 수 있다. 생성된 각 단문의 튜플들은 연결어미의 관계을 보고, 불리언 연산인 OR나 AND로 연결한다.

6) 1998년에 대통령에 당선되었고, 2000년에 노벨평화상을 받은 사람은?
정답유형 : PERSON

6)의 질문은 연결어미로 두 단문이 연결된 복문형태의 질문이다. 연결어미를 기준으로 단문 분할한다.

6-1) 1998년에 대통령에 당선된 사람은?
6-2) 2000년에 노벨평화상을 받은 사람은?

분할된 두 단문형태의 질문에 대해서, 튜플을 생성하면 다음과 같다.

6-1') (?PERSON, 당선직위, 대통령) & (?PERSON, 당선일, 1998년)
6-2') (?PERSON, 상, 노벨평화상) & (?PERSON, 수상일, 2000년)

생성된 각 단문들의 튜플을 통해 검색된 정답은 연결어미, '-고-'에 의해 AND연산을 수행하여 정답을 제시한다.

4.5. 표제어 선택 및 복원

표제어는 일반적으로 문장내의 각 단문에서 생략되는 경우가 많다. 그러므로, 표제어를 복원하는 작업도 요구된다. 표제어의 선택에서 발생할 수 있는 경우는 다음의 네 가지로 나눌 수 있다.

I. 단문 내에 표제어가 있고, 정답유형이 PERSON이 아닌 경우.
   예) 박정희의 사망일은?
II. 단문 내에 표제어가 없고, 정답유형이 PERSON이 아닌 경우.

예) 임시정부 주석을 역임하고, 백범일기의 주인공인 김구가 암살된 해는?
→ 주변 단문에서 '김구'를 표제어로 복원하여 (김구,

직위, 임시정부주석)이라는 튜플을 생성할 수 있다.

예) 발명왕이라고 불리우는 사람이 가진 특허 수는?
→ 임의의 V_PERSON을 표제어로 생성

III. 단문 내에 표제어가 있고, 정답유형이 PERSON 인 경우.
예) 백범 김구를 죽인 사람은?
IV. 단문 내에 표제어가 없고, 정답유형이 PERSON 인 경우.
예) 노벨상은 받은 한국 대통령은?

Ⅰ와 Ⅲ 의 경우에는 단문 내의 표제어를 튜플의 표제어로 선택한다. Ⅱ의 경우에는 해당 단문 좌우의 단문에 존재하는 표제어를 이용하여 표제어를 복원한다. 만약 좌우에도 표제어가 존재하지 않으면, 임의의 V_PERSON으로 표제어를 생성한다. Ⅳ의 경우에는 정답유형의 PERSON을 표제어로 선택한다.

표제어 선정의 우선 순위는 "단문 내의 표제어 〉 정답유형의 PERSON 〉 인접 단문의 표제어 〉 임의의 V_PERSON" 이다.

6. 실험 및 결과

실험은 질의응답 평가셋의 200 질문을 기준으로 하였다. 실험은 크게 지식기반 질의응답 할당 정확률과 지식기반 질의응답 정확률로 나누어 진행하였다. 지식기반 질의응답 할당 정확률은 3단계 정답추출 방법 중 지식기반 질의응답으로 할당된 질문이 올바르게 할당되었는지 여부를 판단하는 근거이다. 지식기반 질의응답 정확률은 할당된 질문이 지식베이스 검색을 하여 정답을 찾았는지 여부를 판단하는 근거이다.

200개의 평가셋 질문에서 지식기반 질의응답으로 할당되어야 하는 질문의 수는 78개이다. 그리고, 지식베이스 속성인식을 통해 지식기반 질의응답으로 할당된 질문 수는 61개이나, 올바르게 할당된 수는 51개이다. 그러므로, 지식기반 질의응답 할당 정확률은 83.6%이고, 지식기반 질의응답 할당 재현률은 65.4%이다. 지식기반 질의응답으로 할당되어야 하는데 할당되지 못한 27개의 질문을 분석한 결과, 오류의 원인은 네 가지로 요약할 수 있다. 오류의 대부분은 인식오류로서 개체명, 작품명(문학작품, 음악작품 등), 별칭(호, 본명, 별명 등) 등을 인식하지 못하여 발생하는 오류로 20개의 질문이 해당되었다. 그 외에는 알고리즘적인 문제가 5개, 개체명 모호성 문제가 1개, 질문에 대한 정답유형 오류가 1개였다.

지식기반 질의응답 정확률은 올바르게 할당된 51개의 질문에서 지식베이스에 그 값이 없는 19개를 제외하면, 32개의 질문이 대상이 된다. 32개의 질문 중 정답을 찾은 질문 수는 26개이다. 따라서, 지식기반 질의응답 정확률은 81.25%이다. 오류는 인식오류, 정답유형 오류와 알고리즘적인 오류로 나뉜다.

지식기반 질의응답으로 할당되지 않았지만, 생성된 일부 튜플을 이용하여 지식베이스에서 정답을 찾을 수 있는 질문이 5개 있었다. 또한, 지식기반 질의응답의 질문이 아닌데 지식베이스에서 정답을 찾을 수 있는 질문이1개가 있었다. "주원장이 세운 나라는?"이라는 질문에서 요구하는 정답에 해당하는 내용은 지식베이스의 속성으로 없다. 그러나, 이 질문을 분석하여 나온 튜플이 ("주원장, 국적, ?Country)"이다. 이 튜플을 기반으로 찾은 답이 정답이었다. 이처럼 질문의 정확한 의미와 분석된 결과가 차이는 있으나, 분석된 결과로 정확한 정답을 찾는 경우도 있었다.

6. 결론 및 향후 연구 계획

본 논문에서는 AnyQuestion 1.0의 3단계 정답추출 방법 중, 지식기반 질의응답을 위한 질문 분석에 대해서 기술하였다. 백과사전의 인물분야라는 도메인의 특수성을 최대한 고려하여 자연어 DB 인터페이스 기술을 이용하였다. 실시간으로 다양한 언어처리를 해야 하는 일반적인 질의응답에 비해, 속성 추출 자질 정보와 규칙만을 이용한 지식기반 질의응답은 상당히 빠른 속도로 사용자의 질문에 정확한 답을 제시할 수 있다. 또한, 인물 도메인에 대한 질문분석의 방법을 최대한 활용하여, 다양한 인물정보 DB와 연결하여 인물정보를 제공할 수 있을 것이다.

지식기반 질의응답에 대한 실험은 200개의 평가셋 질문을 대상으로 하였으며, 질문의 지식기반 질의응답 할당 정확률과 재현률, 지식기반 질의응답의 정확률을 실험하였다. 지식기반 질의응답 할당 정확률은 83.6%이고, 할당 재현률은 65.4%였다. 지식기반 질의응답의 정확률은 81.25%였다.

할당 재현율이 낮은 것은 개체명 태깅도 되지 않으면서, 속성 단서의 역할도 하지 않는 어휘들에 의해서 지식기반 질의응답으로 할당되지 못했기 때문으로 분석된다. 이처럼, 의미없는 어휘들이 특정 속성에서는 속성단서의 역할을 하는 경우도 있다. 따라서, 각 속성별 불용어휘에 해당하는 어휘들을 선정하여, 제거하는 방법에

대한 연구를 진행해야 할 것이다.

3종류의 정답추출의 결과를 통합하여야 정답을 추출할 수 있는 질문들이 많다. 이처럼 다양한 정답추출 방법을 통합하여 정답을 제시하는 방법에 대해서도 연구가 진행될 것이다.

그리고, 백과사전 전분야에 대한 지식기반 질의응답을 위해서 어떤 기술들이 필요하며, 속성의 정의는 어떻게 되어야 할 것인가에 대한 연구를 진행 중에 있다.

참고문헌

[1] 황이규, 김현진, 장명길, "질의응답 기술 개발", 한국정보처리학회지, VOL. 11 NO. 02, 2004,03

[2] Bert Green, Alice Wolf, Carol Chomsky, and Kenneth Laughery, "BASEBALL : An Automatic Question Answerer", In Proceedings of the Western Joint Computer Conference, 1961.

[3] Gary G. Hendrix, "Human Engineering for Applied Natural Language Processing", In Proceedings of the 5thInternational Joint Conference on AI(IJCAI-1977), 1977.

[4] Hoo-Jung Chung 외 6명, "A Practical QA System in Restricted Domains", In Proceedings of the ACL 2004 Workshop(Question Answering in Restricted Domains), 2004.

[5] Marius A.Pasca, Sanda M. Harabagiu, "High Performance Question/Answering", SIGIR 2001.

[6] Ellen M. Voorhees, "The TREC-8 Question Answering Track Report", http : //trec.nist.gov/pubs/trec8/papers/qa_report.pdf

[7] 김현진 외 4명, "3단계 정답 추출 방법론을 이용한 백과사전 인물분야 질의응답 시스템 구현", 제 16 회 한글.언어.인지 학술대회(제출), 2004.

[8] 왕지현 외 2명, "인물 백과사전 지식베이스 구축을 위한 속성패턴기반 정보추출", 제 31 회 정보과학회 추계 학술 발표회(제출), 2004.

[9] Euisok Chung 외 3명, "Hybrid Named Entity Recognition for Question-Answering System", 8th International Conference on Spoken Language Processing(accepted), 2004.

[10] 임수종 외 2명, "백과사전 질의응답 시스템을 위한 격틀 기반 의존관계 분석", 제 16 회 한글.언어.인지 학술대회(제출), 2004.