정확한 해답 추출을 위한 개념 기반의 질의 분석

Concept-based Question Analysis for Accurate Answer Extraction

신승은*, 강유환**, 안영민**, 박희근**, 서영훈**
충북대학교 BK21 충북정보기술사업단*, 충북대학교 전기전자컴퓨터공학부**

Seung-Eun Shin(seshin@nlp.chungbuk.ac.kr)*, Yu-Hwan Kang(eric@nlp.chungbuk.ac.kr)**,
Young-Min Ahn(nlpmania@paran.com)**, Hee-Guen Park(pinetree@nlp.chungbuk.ac.kr)**,
Young-Hoon Seo(yhseo@chungbuk.ac.kr)**

요약

본 논문에서는 정확한 해답 추출을 위해 키워드보다 중요한 역할을 하는 개념을 분석하는 개념 기반 질의 분석에 대해 기술한다. 해답 유형이 같은 질의들에서 나타나는 개념은 유사하기 때문에 이러한 개념들을 잘 정의하여 이용할 경우, 해답을 포함하는 다양한 형태의 구문으로부터 보다 정확한 해답을 추출할 수 있다는 것이 본 논문의 주요 아이디어이다. 즉, 해답을 포함하는 문서와 그 문서 내에 있는 해답을 좀 더 정확하게 추출하기 위해 질문에 있는 각 단어나 구절들의 구문 및 의미 역할을 파악하고자 하는 것이다. 이를 위해, 정답 유형별로 그 유형의 질문에서 공통으로 나타나는 주요 개념들로 구성된 개념 프레임을 정의하고, 사용자 질의를 분석하여 개념 프레임을 채우는 과정으로 질의 분석을 수행한다. 실험 결과 본 논문에서 제안한 개념 기반 방식이 기존의 질의분석 기법에 비해 높은 정답 추출 성능을 보여주었다. 본 논문에서 제안한 개념 기반 접근 방법은 언어에 관계없이 적용 가능한 모델이며, 또한 기존 방식과 함께 사용할 수 있는 장점도 있다.

■ 중심어 : | 질의응답 시스템 | 질의 분석 | 개념 | 해답 추출 |

Abstract

This paper describes a concept-based question analysis to analyze concept which is more important than keyword for the accurate answer extraction. Our idea is that we can extract correct answers from various paragraphs with different structures when we use well-defined concepts because concepts occurred in questions of same answer type are similar. That is, we will analyze the syntactic and semantic role of each word or phrase in a question in order to extract more relevant documents and more accurate answer in them. For each answer type, we define a concept frame which is composed of concepts commonly occurred in that type of questions and analyze user's question by filling a concept frame with a word or phrase. Empirical results show that our concept-based question analysis can extract more accurate answer than any other conventional approach. Also, concept-based approach has additional merits that it is language universal model, and can be combined with arbitrary conventional approaches.

■ keyword : | Question Answering System | Question Analysis | Concept | Answer Extraction |

* 본 논문은 2005년도 충북대학교 학술연구 지원사업의 연구비지원에 의하여 연구되었음.
(This work was supported by the research grant of the Chungbuk National University in 2005)
접수번호 : #061113-004
접수일자 : 2006년 11월 13일
심사완료일 : 2007년 01월 03일
교신저자 : 서영훈, e-mail : yhseo@chungbuk.ac.kr

I. 서 론

정보검색 기술은 인터넷의 발전과 더불어 크게 발전하고 있다. 특히 웹 문서의 양이 급격히 증가하면서 수만에서 수십만의 검색 결과 중에서 사용자가 원하는 정보를 정확하게 찾아주는 효과적인 검색 랭킹 기술이 요구되고 있다[1].

최근 대부분의 정보 검색 시스템들은 사용자의 질의에 대해 관련 있는 문서들을 결과로 제시한다. 그러나 사용자의 질의가 구체적인 대답을 요구할 경우에 사용자들은 찾고자 하는 정답을 정보 검색 시스템의 결과 문서들로부터 찾아야 하는 불편함이 있다. 반면 질의응답 시스템은 사용자들에게 질의에 대한 응답으로 해답 또는 해답을 포함하는 어절들이나 문장들을 제공하기 때문에 더 지능적이고 편리한 시스템이라 할 수 있다. 따라서 질의응답 시스템에 대한 요구가 점점 증가하고 있으며, 보다 더 지능적인 질의응답 시스템이 요구되고 있다.

국제적인 정보검색 평가대회인 TREC(Text REtrieval Conference)에서는 1999년 TREC-8에서 사실에 기반한 단답형 질의에 대한 질의응답시스템의 평가를 시작하였으며, TREC 2003 QA Track에서는 단답형 질의에 여러 개의 답을 요구하는 리스트 질의와 정의에 관련된 질의들이 추가적으로 포함되었다[2]. TREC QA 로드맵에 따르면 해답이 단순히 문장의 일부분을 제시하는 것이 아니라 문장들 사이의 추론이 필요한 질문, 새로운 문장을 생성하여 이를 해답으로 제시하거나, 주어진 해답에 대한 배경 설명, 해답의 정당성 검증, 해답의 모호성 해결, 전문가 수준의 의견 제시가 필요한 질문, 이질적 정보의 통합을 통한 해답의 제시 등 점점 질의응답 시스템의 난이도를 높여 갈 계획이다.

보다 더 지능적이고 난이도 높은 질의응답 시스템을 위해 사용자의 질의 의도를 정확하게 분석하는 것이 필요하다. 이를 위해 본 논문에서는 정확한 해답 추출을 위해 키워드보다 중요한 역할을 하는 개념을 분석하는 개념 기반 질의 분석에 대해 기술한다. 개념 기반 질의 분석은 해답을 포함하는 문서와 그 문서 내에 있는 해답을 좀 더 정확하게 추출하기 위해 질문에 있는 각 단어나 구절들의 구문 및 의미 역할을 파악하고자 하는 것이다. 이를 위해, 해답 유형별로 그 유형의 질문에서 공통으로 나타나는 주요 개념들로 구성된 개념 프레임을 정의하고, 사용자 질의를 분석하여 개념 프레임을 채우는 과정으로 질의 분석을 수행한다.

II. 관련 연구

질의응답 기술은 사용자의 자연언어 질문과 검색 대상 문서의 의미를 파악하기 위한 자연언어 처리 기술과 대상 문서로부터 해답을 추출하기 위한 정보추출 기술을 필요로 하며, 많은 후보 문서들로부터 답을 포함하는 문서를 걸러주는 역할을 위해 기존의 문서 검색 기술도 활용한다. 일반적인 질의응답 시스템은 해답을 찾기 위해 질의 분석 과정에서 사용자가 원하는 해답이 무엇인지 질의 의도를 파악할 수 있는 해답 유형이나 키워드 등의 정보를 질의로부터 추출한다. 또한, 기존의 정보검색 방법에 의해 질의와 유사한 문서를 추출하고, 문서에서 다시 해답을 포함할 가능성이 있는 단락을 추출한 후, 단락에서 해답 유형과 동일한 개체를 찾아내어 사용자에게 해답으로 제시한다[3].

기존의 질의 분석은 질의 유형을 결정하고, 질의로부터 키워드를 추출하는 단계로써, 질의 유형 분류 기법은 규칙에 기반한 방법(rule-based method)[4-6]과 통계에 기반한 방법(statistical method)[7-9]으로 나뉘어진다.

규칙에 기반한 질의 유형 분류를 채택하고 있는 시스템들은 일반적으로 lexico-syntactic 패턴을 구축하고, 이러한 패턴을 유한 상태 오토마타와 매치하여 질의 유형을 분류한다. MURAX[6]는 규칙에 기반한 질의 유형 분류를 사용하는 대표적인 질의응답 시스템으로, 대부분의 규칙에 기반한 시스템은 MURAX와 유사한 질의 처리 방법을 사용한다. 규칙에 기반한 질의 유형 분류는 사용자의 질의에 대해 즉시 질의 유형을 분류할 수 있고, 응용 영역이 정해져 있을 경우 간단한 규칙 수정을 통해 성능을 향상시킬 수 있다는 장점이 있다. 그러나 규칙 수정을 위해 전문적인 지식을 가진 사람들의 노력이 필요하고, 규칙이 없는 질의에 대한 처리와 규칙의 수가 증가함에 따라 전체 시스템의 성능 향상을 위한 수정이 어려

워진다는 단점을 갖는다.

통계적 방법에 기반한 질의 유형 분류는 수동으로 분류된 대량의 학습 데이터로부터 추출한 통계 정보를 이용한다. Ittycheriah[7]는 질의 유형 분류를 위해 최대 엔트로피 모델(maximum entropy model)을 이용하였다. Mann[8]은 가공되지 않은 데이터를 이용해서 질의 유형을 분류하는 방법을 제안하였다. 이 방법은 상호 정보 척도(mutual information)에 기반하여 질의 데이터를 학습한다. 통계에 기반한 질의 유형 분류는 안정적으로 질의 유형을 분류할 수 있고, 응용 영역의 변화에 크게 영향을 받지 않으며, 자동화된 통계적 방법을 사용함으로써 시스템 구축을 쉽게 할 수 있다는 장점이 있다. 그러나 질의 유형이 다르지만 구조적으로 매우 유사한 질의들에 대해 질의 유형을 제대로 구분하기 어렵다. 규칙에 기반한 시스템들은 이러한 문제를 보완할 수 있는 규칙을 쉽게 수정하거나 추가할 수 있지만, 통계적 방법의 경우에는 보완이 쉽지 않다. 이를 위해 지지 벡터 기계 기반의 질의 유형 분류에 대한 연구도 있었다[9].

이러한 기존의 질의분석은 단순히 명사 단어를 키워드로 추출하고, 해답 유형을 결정하기 때문에 해답추출 단계에서 사용자의 질의 의도에 적합한 해답을 찾는 것을 어렵게 한다. 또한 기존의 질의분석은 사용자 질의의 의미를 고려할지라도 질의에서의 사용 의미를 고려하지 않고 단어 자체의 의미만을 활용하였기 때문에 사용자의 정확한 질의 의도를 분석할 수 없었다. 따라서 보다 지능적인 질의응답 시스템을 위해 정확한 해답 추출이 필요하며, 사용자의 질의 의도를 정확하게 파악하는 것이 필요하다. 이를 위해 단어 자체의 의미뿐만 아니라 사용자가 질의에서 단어를 어떤 의미로 사용하였는지 그 사용 의도를 정확하게 분석해야 한다.

III. 개념 기반 질의 분석

개념 기반 질의 분석은 사용자의 질의를 대상으로 자연언어 처리 기법을 적용하고 분석하여 질의의 의도와 내용을 파악하는 것으로, 보다 지능적인 질의응답 시스템을 위해 단어 자체의 의미뿐만 아니라 사용자가 질의에서 단어를 어떤 의미로 사용하였는지 사용 의도를 정확하게 분석하여 문서검색과 정답추출에 활용하고자 하는 것이다.

기존의 질의분석은 사용자의 질의로부터 키워드를 추출하고, 정답 유형을 결정하는 역할만을 했으며, 키워드 추출에 의미를 반영하더라도 단어 자체의 의미만을 고려하였다[1][10-12]. 따라서 질의응답 시스템에서 정확한 해답 추출을 위해 사용자 질의로부터 해답의 세부 유형 결정과 단어가 문장에서 어떤 의도로 사용되었는지를 나타내는 개념을 추출하는 개념 기반 질의 분석이 필요하다.

다음의 (질의 1)을 살펴보자.

(질의 1) '베니스 영화제'에서 <오아시스>로 감독상을 수상한 사람은?

기존의 질의분석에서는 (질의 1)에서 '베니스 영화제', '오아시스', '감독상', '수상', '사람'을 키워드로 추출하고, 해답 유형을 '인물'로 결정할 것이다. 또한, 문서검색과 해답추출 과정에서 키워드와 해답 유형을 이용한 통계적인 방법에 의해 해답을 추출한다. 그러나 단어의 사용 의도를 분석하면 '베니스 영화제'는 '수상식'을, '오아시스'는 '작품명', '감독상'은 '수상명'을 나타내기 위해 사용되었음을 알 수 있다. 이러한 정보를 이용한다면 통계적인 방법보다 정확한 해답추출을 수행할 수 있다. 예를 들어, "베니스 영화제에서 감독상을 받은 이창동감독은…….", "오아시스의 감독은 이창동 씨이며, ……."와 같은 문장을 포함하는 문서를 키워드에 대한 통계 정보와 상관없이 해답 문서로 선택할 수 있다. 특히, 사용자 질의에 포함된 고유명사는 질의 의도를 표현하기 위해 중요한 역할을 수행한다. 그러나 기존의 질의분석에서는 단어 자체의 의미만을 고려하기 때문에 사용자의 질의 의도를 충분히 문서검색과 해답추출에 반영하지 못한다.

1. 개념

개념들은 자연언어 질의에서 사용자의 의도를 나타내기 위해 사용되는 단어나 연속된 단어들에 부여하는 의미이다. 자연언어 질의의 형태소 분석 결과로부터 의미

있는 형태소들을 함께 결합하는 과정을 거친 후, 결합된 형태소들에게 개념을 부여한다.

다음은 세부 해답 유형이 '저자'인 자연언어 질의들의 예이다.

(질의 2) 동의보감을 저술한 사람은 누구인가?
(질의 3) 햄릿의 저자는?
(질의 4) '로미오와 줄리엣'을 쓴 영국의 대문호는 누구인가?
(질의 5) '철의 여인 마가렛 대처'란 책의 저자는 누구입니까?

세부 해답 유형이 '저자'인 위의 자연언어 질의를 살펴보면 '저자'에 대한 질의에 공통적으로 사용되는 개념들이 있음을 알 수 있다. '저자'에 대한 질의에 사용된 공통적인 개념들은 '저서명'-(동의보감, 햄릿, 로미오와 줄리엣, 철의 여인 마가렛 대처), '저자 관련 명사'-(저자, 대문호), '저자 관련 용언'-(저술하, 쓰), '장르'-(책), '인물명사'-(사람), '국가'-(영국), '인물의문사'-(누구) 등이며, 질의는 이러한 개념들을 이용하여 구성되고 있다.

세부 해답 유형과 개념 정의를 위해 TREC의 Test Collection과 Web으로부터 추출된 643개의 인물 관련 자연언어 질의 말뭉치를 분석하여 해답의 세부 유형에 따라 분류하고, 각각의 세부 해답 유형에 따라 질의 구성에 공통적으로 사용된 개념들을 정의하였다. 인물 관련 질의들은 '기타'를 포함한 전체 24개의 세부 해답 유형으로 분류되었고, 세부 해답 유형에 대한 전체 개념들은 125개의 개념들로 정의되었다. [표 1]은 세부 해답 유형과 개념의 일부를 보여주는 예이다. 세부 유형 중 '공통'은 모든 세부 유형에서 공통적으로 사용되는 개념들을 말한다.

인물 관련 자연언어 질의 말뭉치로부터 정의된 개념에 따라 어휘형태소를 추출하고, 동의어/유의어 사전을 이용한 확장을 통해 개념 사전을 구축하였다. 개념 사전은 2,039개의 어휘형태소로 구성되었으며, 278개의 용언들을 포함한다. 개념 사전은 개념, 태그, 어휘목록으로 구성된다.

표 1. 세부 해답 유형과 개념

세부 해답 유형	개념 (Concept)
공통	장소, 시간, 성, 인물명사, 단서부사, …
저자	저서명, 필명, 저자관련명사, 저자관련용언, …
가족	기준인물, 가족관계, 인물정보, 관계정보, …
수상자	수상명, 수상식, 수상관련명사, 수상관련용언, …
정치가	지위, 사건, 단체, 선출관련명사, 선출관련용언, …

표 2. 개념 사전의 예

개념	태그	어휘 목록
단서부사	%Adverb	최초, 처음, 마지막, 가장
시간	%Time	
인물의문사	@Who	누구
인물명사	@Person	사람, 인물, 분
저자명사	@Author_N	저자, 지은이, 글쓴이, 편찬자
저자용언	@Author_V	저술하, 쓰, 짓, 편찬하
장르	@Genre	책, 소설, 시, 수필
가족관계	@Family	아들, 부인, 아버지, 아빠
기준인물	#Standard_Person	
저서명	#Book	

[표 2]는 개념 사전의 예이다. 개념에 따른 태그는 '속성'+'개념'으로 붙여진다. 속성은 '%', '@', '#' 3가지로 나뉜다. '%' 속성은 문장에서의 사용 위치가 자유로운 개념들이며, 질의 분석 과정에서 사용자의 자연언어 질의를 개념 태깅한 후 별도로 추출하며 개념 프레임의 구성에서 제외하였다. '%' 속성의 개념들은 규칙과 개념 사전을 이용하여 개념 태깅을 한다. '@' 속성은 문법을 구성하는 기본적인 개념들로 개념 사전을 검색하여 개념 태깅을 수행한다. '#' 속성은 사용자 질의에서 자신을 제외한 모든 개념들이 개념 프레임을 구성하였을 때 질의로부터 추출할 개념들을 나타낸다.

2. 개념 프레임

본 논문에서 제시하는 개념 프레임은 643개의 인물 관련 자연언어 질의 말뭉치를 기반으로 하여 작성되었다. 인물 관련 질의의 개념은 어휘형태소를 중심으로 부여되며, 개념 프레임은 세부 해답 유형에 따라 각각 개념들의 어절 정보들로 구성된다. 프레임의 구성에서 문장에서의 위치가 자유로운 개념들('%' 속성을 갖는 개념들)은 제

외하였다. 이러한 개념들은 질의 분석 과정에서 사전이나 규칙을 통해 먼저 추출되고, 남은 질의에 대해 개념 프레임을 이용한 질의 분석을 수행한다. 이것은 자유로운 어순을 갖는 한국어의 특성을 반영하고 개념 프레임의 수를 줄임으로써 질의 분석의 효율을 높이기 위함이다.

[그림 1]은 개념 프레임의 BNF(Backus-Naur Form) 표현이다. 어절 정보는 어휘형태소의 개념으로만 구성되는 것과 어휘형태소의 개념과 문법형태소로 구성되는 것이 있으며, 개념 프레임은 어절 정보들의 리스트로 구성된다.

```
<개념 프레임> ::= <어절 리스트>
<어절 리스트> ::= <어절 정보> | <어절 리스트><어절 정보>
<어절 정보> ::= '(' <개념> ')' | '(' <개념> <문법형태소> ')'
<개념> ::= '@Who' | '@Person' | '@Author_N' | '@Author_V'
          | '@Genre' | '@Family' | '#Standard_Person' | '#Book' |...
<문법형태소> ::= 'jc' | 'jx' | 'jm' | 'etm' | 'co' | 'ef' | 'oj' | 'co+etm'
```

그림 1. 개념 프레임의 BNF 표현

```
1. (#Book_Info jc) (@About) (@Genre jc) (@Author_V etm)
                    (@Person|@Author_N jx?) (@Who)?
2. (#Book co+etm) (@Genre jc) (@Author_V etm)
                    (@Person|@Author_N jx?) (@Who)?
3. (#Book co+etm) (@Genre jm )(@Author_N jx?) (@Who)?
4. (#Book jc) (@Author_V etm)
                    (@Person|@Author_N jx?) (@Who)?
5. (#Book jm) (@Author_N jx?) (@Who)?
6. (@Who) (#Book jc) (@Author_V ef)
7. (#Book jc) (@Who) (@Author_V ef)
```

그림 2. '저자' 에 대한 개념 프레임의 예

세부 해답 유형 '저자'에 대한 개념 프레임은 [그림 2]와 같다. [그림 2]에서 어휘형태소의 개념으로만 구성된 어절 정보는 '(@About)'과 '(@Who)'이며 어휘형태소의 개념과 문법형태소로 구성된 어절 정보는 '(@Author_V etm)', '(#Book jc)', '(#Book co+etm)' 등이 있다. 또한, 개념 프레임에서 사용된 기호 '?'는 생략 가능한 문법형태소와 어절 정보를 나타낸다.

세부 해답 유형에 대한 개념이 정의되지 않은 질의들에 대해서는 자연언어 질의의 특성을 이용하여 개념 프레임을 정의한다.

아래의 질의들을 살펴보자.

(질의 6) 남극에 도착한 최초의 사람은?

(질의 7) 세계에서 가장 부자는 누구입니까?

(질의 8) 우리나라에서 간이식 수술을 최초로 성공한 의사는?

(질의 6)과 (질의 8)은 용언을 포함하는 질의, (질의 7)은 용언을 포함하지 않는 질의로 모두 23개의 세부 해답 유형에 포함되지 않는 질의이며, 개념이 정의되지 않은 질의들이다. 이러한 질의들의 분석은 자연언어 질의 특성을 이용한 개념 프레임을 구성하여 수행한다.

용언을 포함하는 질의에 대한 특성은 이벤트 관련 용언(Event_V)과 인물명사(Person) 혹은 특성명사(Property), 이벤트 관련 용언과 인물 의문사(Who)로 구성되며, 용언을 포함하지 않는 질의는 특성명사와 특성명사를 한정하는 명사구로 구성된다. 따라서 기타 질의에 대한 개념 프레임은 이벤트용언, 특성명사, 인물명사, 관련 명사구의 구문구조에 따라 구성된다. 질의분석에서 명사구를 정확하게 분석하는 것 자체가 어려운 연구이나 질의의 어절 수가 적은 자연언어 질의에 대해서는 단순 명사나 명사 나열 등 간단한 명사구 형태를 적용하여 분석하는 것이 가능하며, 전체적으로 비교적 높은 정확률을 보인다.

용언을 포함하는 질의와 용언을 포함하지 않는 질의의 자연언어 특성에 따른 개념 프레임은 [그림 3]과 같으며, '#Event_V'는 사용자 질의로부터 추출할 용언을 의미하고 '@' 속성의 개념은 개념 사전을 이용하여 개념을 부여한다.

```
[용언을 포함하는 질의의 개념 프레임]
1. (NP jc) (#Event_V etm) (@Person|@Property_N) (@Who)?
2. (NP jx) (@Who) (#Event_V ef)
3. (@Who) (NP jc) (#Event_V ef)

[용언을 포함하지 않는 질의의 개념 프레임]
1. (NP jm) (@Property_N) (@Who)?
2. (NP jc) (@Property_N) (@Who)?
3. (@Property_N) (@Who)?
```

그림 3. 용언을 포함하는 질의와 용언을 포함하지 않는 질의의 개념 프레임

3. 개념 기반 질의 분석 결과

개념 기반 질의 분석 과정은 먼저 질의의 형태소분석 결과와 개념 사전을 이용하여 어절 정보 리스트를 구성하고, 개념 프레임을 적용한다. 적합한 개념 프레임이 없는 경우에는 자연언어 질의 특성에 따른 개념 프레임을 적용한다.

(질의 5)와 (질의 8)에 대한 개념 기반 질의 분석의 과정과 결과는 다음과 같다.

(질의 5) '철의 여인 마가렛 대처'란 책의 저자는 누구입니까?

- 형태소분석 결과
  - 철의 여인 마가렛 대처/nc+이/co+라고/ec+하/pv+ㄴ/etm 책/nc+의/jm 저자/nc+는/jx 누구/np+이/co+ㅂ니까/ef
- 어절 정보 리스트
  - (철의 여인 마가렛 대처 co+etm) (@Genre jm) (@Author_N jx) (@Who co+ef)
- 개념 프레임
  - (#Book co+etm) (@Genre jm) (@Author_N jx?) (@Who)?
- 개념 기반 질의 분석 결과
  - 해답 유형 : 인물
  - 세부 해답 유형 : 저자
  - 추출된 개념
    - 저서명(Book) : 철의 여인 마가렛 대처

(질의 8) 우리나라에서 간이식 수술을 최초로 성공한 의사는?

- 형태소분석 결과
  - 우리나라/nc+에서/jc 간/nc+이식/nc 수술/nc+을/jc 최초/nc+로/jc 성공/nc+하/xsv+ㄴ/etm 의사/nc+는/jx
- 전처리 결과
  - (%Location) <간이식 수술>/nc+을/jc (%Adverb) 성공하/pv+ㄴ/etm 의사/nc+는/jx
- 어절 정보 리스트
  - (간이식 수술 jc) (성공하 etm) (@Property_N)
- 개념 프레임
  - (NP jc) (#Event_V etm) (@Person|@Property_N) (@Who)?
- 개념 기반 질의 분석 결과
  - 해답 유형 : 인물
  - 추출된 개념
    - 장소(Location) : 우리나라
    - 단서부사(Adverb) : 최초
    - 이벤트용언(Event_V) : 성공하
    - 특성명사(Property_N) : 의사
    - 명사구(NP) : 간이식 수술

IV. 개념을 이용한 해답 추출

본 논문에서는 질의응답 시스템의 해답 추출에서 개념 기반 질의 분석 결과의 활용 방안을 제시하고, 실험을 통해 평가한다. 질의응답 시스템에서 정확한 해답을 추출하기 위해 세부 해답 유형에 따라 개념과 문법형태소를 이용하여 개념 기반 해답 추출 규칙을 구성하고, 이를 적용하여 해답을 추출하였다.

[그림 4]는 세부 해답 유형 '저자'에 대한 개념 기반 해답 추출 규칙의 예이다. 해답 추출 규칙은 개념과 문법형태소를 이용한 어절 정보 리스트로 구성된다. 문장에 대한 형태소 분석과 전처리 과정을 수행하고, 개념 태깅을 하여 어절 정보 리스트를 구성한다. 구성된 어절 정보 리스트에 [그림 4]와 같은 개념 기반 해답 추출 규칙을 적용하여 문장으로부터 해답을 추출한다. 해답 추출 규칙을 구성할 때, 단서부사와 문장에서 쓰임이 자유로운 개념들은 제외하며, 규칙을 적용할 때에는 해답 문장에 반드시 포함해야 하는 개념들에 대해 별도로 추출하여 규칙 적용 전에 검사하고, 조건을 만족한 경우에 규칙을 적용한다.

```
1. ($Book oj) (@Author_V etm) (Answer)
2. (Answer j_) ($Book oj) (@Author_V)
3. ($Book oj) (@Author_V etm)
                (@Author_N|@Person j_) (Answer co+*)
4. (Answer j_) ($Book oj) (@Author_V) (@Author_N|@Person co+*)
5. ($Book oj) (@Author_V etm)
                (@Author_N|@Person co+etm) (Answer)
6. ($Book jm) (@Author_N j_) (Answer co+*)
7. ($Book jm) (@Author_N co+etm) (Answer)
8. (Answer j_) ($Book jm) (@Author_N co+*)
```

그림 4. '저자' 에 대한 개념 기반 해답 추출 규칙의 예

개념 기반 해답 추출 규칙을 이용한 해답 추출 과정을 살펴보면 다음과 같다.

(질의 2) 동의보감을 저술한 사람은 누구인가?

- 개념 기반 질의 분석 결과
  - 세부 해답 유형 : 저자
  - 추출된 개념 :
    - 저서명 : 동의보감
- 해답 포함 문장
  - 구암 허준은 1610년 동의보감을 편찬하였다.
- 형태소분석 결과
  - 구암/nc 허준/nh+은/jx 1610/nn+년/nb 동의보감/nc+을/jc 편찬/nc+하/xsv+었/ep+다/ef
- 전처리 결과
  - 구암/nc 허준/nh+은/jx (%Time) 동의보감/nc+을/oj 편찬하/pv+었/ep+다/ef
- 어절 정보 리스트
  - 구암/nc 허준/nh+은/jx ($Book oj) (@Author_V)
- 적용 규칙 : 개념 기반 해답 추출 규칙 2
  - (Answer j_) ($Book oj) (@Author_V)
- Answer : 허준

기타 질의에 대한 해답 추출은 자연언어 질의의 구문 구조 특성을 이용한 해답 추출 규칙을 이용한다. [그림 5]는 각각 용언을 포함하는 질의와 용언을 포함하지 않는 질의에 대한 해답 추출 규칙이다.

```
[용언을 포함하는 질의 해답 추출 규칙]
1. (!NP jc) (!Event_V etm) (Answer)
2. (Answer j_) (!NP jc) (!Event_V)
3. (!NP jc) (!Event_V etm) (@Person|!Property_N j_) (Answer co+*)
4. (Answer j_) (!NP jc) (!Event_V etm) (@Person|!Property_N co+*)

[용언을 포함하지 않는 질의 해답 추출 규칙]
5. (!NP jm) (!Property_N co+etm) (Answer)
6. (Answer j_) (!NP jm) (!Property_N co+*)
7. (!NP jm) (!Property_N j_) (Answer co+*)
8. (!NP jc) (!Property_N co+etm) (Answer)
```

그림 5. 기타 질의에 대한 해답 추출 규칙

해답 추출 규칙은 해답이 이벤트용언이나 특성명사의 수식을 받는 경우(규칙 1, 규칙 5, 규칙 8)와 해답이 문장의 주어로 사용되는 경우(규칙 2, 규칙 4, 규칙 6), 해답이 문장의 술어로 쓰인 경우(규칙 3, 규칙 7)로 나뉜다.

다음은 용언을 포함하는 질의에 대한 해답 추출 규칙을 사용하여 해답을 추출하는 과정이다.

(질의 6) 남극에 도착한 최초의 사람은?

- 질의 분석 결과
  - 단서부사(Adverb) : 최초
  - 이벤트용언(Event_V) : 도착하
  - 명사구(NP) : 남극
- 해답 포함 문장
  - 아문센은 처음으로 남극에 이르러 환호하였다.
- 형태소분석 결과
  - 아문센/nh+은/jx 처음/nc+으로/jc 남극/nc+에/jc 이르/pv+어/ec 환호/nc+하/xsv+었/ep+다/ef
- 전처리 결과
  - 아문센/nh+은/jx (%Adverb) 남극/nc+에/jc 이르/pv+어/ec 환호하/pv+었/ep+다/ef
- '%' 속성을 갖는 개념
  - 사용자 질의 단서부사(Adverb) :

최초(처음, 시초, 초번, ...)

- 해답 문장 단서부사(Adverb) : 처음
- 해답문장 단서부사 ∈ 사용자 질의 단서부사

■ 어절 정보 리스트

- 아문센/nh+은/jx (!NP jc) (!Event_V)
  환호하/pv+었/ep+다/ef

■ 적용 규칙

- (Answer j_) (!NP jc) (!Event_V)

■ Answer : 아문센

용언을 포함하지 않는 질의에 대한 해답 추출 규칙을 사용하여 해답을 추출하는 과정은 다음과 같다.

(질의 7) 세계에서 가장 부자는 누구입니까?

■ 질의 분석 결과

- 장소(Location) : 세계
- 단서부사(Adverb) : 가장
- 특성명사(Property_N) : 부자

■ 해답 포함 문장

- 빌게이츠는 세계에서 최고의 갑부이며
  자선사업가다.

■ 형태소분석 결과

- 빌게이츠/nh+는/jx 세계/nc+에서/jc
  최고/nc+의/jm 갑부/nc+이/co+며/ec
  자선/nc+사업/nc+가/xsn+이/co+다/ef

■ 전처리 결과

- 빌게이츠/nh+는/jx (%Location) (%Adverb)
  갑부/nc+이/co+며/ec 자선사업가/nc+이/co+다/ef

■ '%' 속성을 갖는 개념

- 사용자 질의 단서부사(Adverb) :
  가장(제일, 최고, 매우 ...)
- 정답문장 단서부사(Adverb) : 최고
- 정답문장 단서부사 ∈ 사용자 질의 단서부사
- 사용자 질의 장소(Location) :
  세계(세상, 천하, 지구, ...)
- 정답문장 장소(Location) : 세계
- 정답문장 장소 ∈ 사용자 질의 장소

■ 어절 정보 리스트

- 빌게이츠/nh+는/jx (!Property_N co+*)
  자선사업가/nc+이/co+다/ef

■ 적용 규칙

- (Answer j_) (!Property_N co+*)

■ Answer : 빌게이츠

위와 같이 개념과 문법형태소를 이용하여 해답 추출 규칙을 만들고 이를 적용하여 해답을 포함하는 문장으로부터 정확한 해답을 추출할 수 있다.

V. 실험 및 평가

본 논문에서는 실험을 위해 웹에서 100개의 인물 관련 자연언어 질의를 추출하여 실험 질의 말뭉치를 구축하였다. 실험 질의 말뭉치를 대상으로 개념 기반 질의 분석을 수행하고, 그 결과에 대한 평가로써 해답 유형 및 개념 추출과 해답 추출의 정확률에 대하여 실험하였다.

해답 유형 및 개념 추출 실험에서 개념 프레임과 자연언어 질의 특성을 이용하여 실험 질의 말뭉치의 질의들에 대해 개념 기반 질의 분석을 수행하고, 해답 유형과 추출된 개념의 정확률을 평가하였다. 개념은 사용자의 질의 의도를 파악하기 위해 꼭 추출해야 하는 요소로 세부 해답 유형이 정의된 경우에는 세부 해답 유형에 따라 추출할 개념이 다르며, 정의되지 않은 경우에는 이벤트 용언, 특성명사, 명사구가 있다.

표 3. 해답 유형 및 개념 추출 정확률

	세부 해답 유형이 정의된 질의	세부 해답 유형이 정의되지 않은 질의
질의 수	69	31
해답 유형 분류 정확률	1.000	0.9032
세부 해답 유형 분류 정확률	1.000	세부 해답 유형 없음
개념 추출 정확률	0.9179	0.8449

[표 3]은 해답 유형 및 개념 추출 실험의 결과를 나타낸다. 유형 분류 정확률은 (정확하게 유형이 분류된 질의 수)/(전체 질의 수)이며, 개념 추출 정확률은 (정확하게 개념이 추출된 수)/(추출되어야 할 전체 개념 수)이다. 실험 질의 말뭉치 중 69개의 질의가 세부 해답 유형이 정의된 질의였으며, 31개의 질의가 세부 해답 유형이 정의되지 않은 질의였다. 4개의 질의가 세부 해답 유형이 정의되었으나 개념 프레임에 의해 분석되지 않았다. 그 중 2개는 개념 프레임이 구성되지 않았고, 나머지 2개는 개념 사전에 포함되지 않은 어휘를 포함하고 있었다.

세부 해답 유형이 정의되지 않은 질의의 해답 유형 분류 정확률이 다소 낮은 원인은 웹에서 추출한 질의에서 일반적으로 특성명사로 구성된 짧은 질의를 많이 사용하였기 때문이며, 이를 보완하기 위해 개념 사전의 확장이 필요하다. 개념 추출에서는 명사구 인식 문제가 정확률을 떨어뜨리는 주된 원인이었다. 웹에서 추출한 비교적 짧은 질의에 대해 높은 해답 유형 분류 정확률과 개념 추출 정확률을 보였으나, 개념 기반 질의 분석에서 해답 유형 분류와 개념 추출 정확률을 높이기 위해 개념 프레임과 개념 사전의 보완이 필요하며, 특히 명사구 인식에 대한 추가적인 연구가 필요하다.

해답 추출 실험은 개념 기반 질의 분석 결과를 이용하여 Google과 Yahoo의 검색 결과로부터 정답을 포함하는 문서를 선택하고, 해당 문서를 대상으로 수행하였다.

표 4. 해답 추출 정확률

	Google	Yahoo	평 균
해답 추출 정확률	0.7407	0.7586	0.7497

[표 4]는 해답 추출 정확률을 나타낸다. 해답 추출 실험에서 개념 기반 해답 추출 규칙을 이용하였을 때, 해답 추출의 정확률이 평균 0.7497이었다. 일반적인 질의응답 시스템의 해답 추출 정확률(TREC 2005 Best 0.713)보다 높은 정확률을 보였다[13]. 다양한 문서 형태를 보이는 웹 문서의 특성으로 인해 해답이 추출되지 않는 문서들도 있었다. 예를 들어 백과사전 문서에서는 해답 추출 규칙에 맞는 문장에서 주어가 생략되어 해답을 추출하지 못하는 경우가 많았다. 해답 추출에서 이러한 웹의 특성을 반영한다면 해답 추출 정확률은 더 많은 향상이 있을 것으로 기대된다. 또한 여러 문장에 걸쳐서 표현된 해답 등과 같이 다양한 형태의 해답 추출을 위해 개념 기반 해답 추출 규칙이 확장된다면 보다 정확한 해답 추출을 기대할 수 있을 것이다. 또한 제한된 영역에서의 실험으로 일반적인 질의응답 시스템과의 직접적인 비교는 힘들지만 개념 기반 방법의 장점을 반영하고, 단점을 기존의 방법론으로 보완하여 질의응답 시스템의 성능을 향상시킬 수 있을 것으로 기대한다.

VI. 결론 및 향후 연구

본 논문에서는 정확한 해답 추출을 위한 개념 기반의 질의 분석을 제안하였다. 개념 기반 질의 분석은 자연언어 질의를 의미적으로 분석하여 사용자의 질의 의도를 파악하고 문서 검색 및 해답 추출에 필요한 세부 해답 유형과 개념들을 추출하였고, 질의 분석 결과를 이용하여 정확하게 해답을 추출하는 방법을 기술하였다.

실험에서 세부 해답 유형 분류는 0.69의 재현율과 1.00의 정확률을 보였으며, 개념 추출은 0.89의 정확률을 보였다. 실험을 통해 비교적 짧지만 사용자의 질의 의도를 충분히 표현하고 있는 자연언어 질의에 대해 개념 기반 질의 분석을 수행함으로써 사용자의 질의 의도를 분석하고, 개념들을 추출할 수 있음을 보였다. 해답 추출 실험에서는 제한된 문서에 대한 실험이었으나, 일반적인 질의응답 시스템의 해답 추출 정확률보다 높은 0.7497의 정확률을 보임으로써 개념 기반 질의 분석의 결과가 정확한 해답 추출을 위해 활용될 수 있음을 보였다. 또한 본 논문에서 제안한 개념 기반 접근 방법은 언어에 관계없이 적용 가능한 모델이며, 또한 기존 방식과 함께 사용할 수 있는 장점도 있다.

향후 실용적인 개념 기반 질의응답 시스템을 위해 개념 프레임과 개념 기반 해답 추출 규칙의 확장이 필요하며, 검색대상이 되는 문서들에 대한 효율적인 개념분석 과정과 개념 색인에 대한 연구도 필요하다.

참 고 문 헌

[1] 장명길, 김현진, 장문수, 최재훈, 오효정, 이충희, 허정, "의미기반 정보검색", 정보과학회지, 제19권, 제10호, pp.7-18, 2001.

[2] E. M. Voorhees and H. T. Dang, "Overview of the TREC 2005 Question Answering Track," In Proceedings of the TREC 2005, 2005.

[3] 황이규, 김현진, 장명길, "질의응답 기술 개발", 정보처리학회지, 제11권, 제2호, pp.48-56, 2004.

[4] D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju, and V. Rus, "LASSO: A Tool for Surfing the Answer Net," In the 8th Text REtrieval Conference (TREC-8), 1999.

[5] J. Prager, D. Radev, E. Brown, and A. Coden, "The Use of Predictive Annotation for Question Answering in TREC8," In the 8th Text REtrieval Conference (TREC-8), 1999.

[6] J. Kupiec, "MURAX: A Robust Linguistic Approach For Question Answering Using An On-Line Encyclopedia," In Proceedings 16'th ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.181-190, 1993.

[7] A. Ittycheriah, M. Franz, W. J. Zhu, and A. Ratnaparkhi, "Question Answering Using Maximum Entropy Components," In Proceeding of NAACL, 2001.

[8] G. S. Mann, "A Statistical Method for Short Answer Extraction," In Proceedings of the ACL Workshop Open-Domain Question Answering, pp.13-30, 2001.

[9] 김학수, 안영훈, 서정연, "한국어 질의응답시스템을 위한 지지벡터기계 기반의 질의유형분류기", 정보과학회 논문지, 제30권, 제5호, pp.466-475, 2003.

[10] E. Voorhees, "Query Expansion using Lexical Semantic Relation," In Proceedings of the 17th ACM-SIGIR Conference, pp.61-69, 1994.

[11] B. V. Dobrow, N. V. Loukachevitch, and T. N. Yudina, "Conceptual Indexing Using thematic Representation of Texts," TREC-6, 1997.

[12] 강승식, "한글 문서의 색인어와 색인 기법," 정보과학회지, 제22권, 제4호, pp.72-77, 2004.

[13] E. M. Voorhees and H. T. Dang, "Overview of the TREC 2005 Question Answering Track," TREC 2005, 2005.

저 자 소 개

신 승 은(Seung-Eun Shin) 정회원

- 1999년 : 충북대학교 컴퓨터공학과(공학사)
- 2001년 : 충북대학교 컴퓨터공학과(공학석사)
- 2006년 : 충북대학교 컴퓨터공학과(공학박사)
- 2006년 ~ 현재 : 충북대학교 BK21 충북정보기술사업단 연수연구원

<관심분야> : 정보검색, 자연언어처리

강 유 환(Yu-Hwan Kang) 정회원

- 1998년 : 충북대학교 컴퓨터공학과(공학사)
- 2000년 : 충북대학교 컴퓨터공학과(공학석사)
- 2000년 ~ 2001년 : (주)L&H Lorea 근무
- 2001년 ~ 현재 : 충북대학교 컴퓨터공학과 박사과정

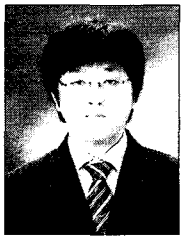
<관심분야> : 기계번역, 구문분석, 질의응답시스템

안 영 민(Young-Min Ahn) 정회원

- 2000년 : 충북대학교 컴퓨터공학과(공학사)
- 2002년 : 충북대학교 컴퓨터공학과(공학석사)
- 2002년 ~ 2004년 : (주)제이너 시스템 근무
- 2003년 ~ 현재 : 충북대학교 컴퓨터공학과 박사과정

<관심분야> : 정보검색, 자동분류, 구문분석

박 희 근(Hee-Guen Park) 준회원



- 2006년 : 충북대학교 컴퓨터공학과(공학사)
- 2006년 ~ 현재 : 충북대학교 컴퓨터공학과 석사과정

<관심분야> : 자연언어처리, 형태소분석, 구문분석, 질의응답시스템

서 영 훈(Young-Hoon Seo) 종신회원

- 1983년 : 서울대학교 컴퓨터공학과(공학사)
- 1985년 : 서울대학교 컴퓨터공학과(공학석사)
- 1991년 : 서울대학교 컴퓨터공학과(공학박사)
- 1994년 ~ 1995년 : 미국 Carnegie-Mellon 대학 기계번역센터 객원교수
- 1988년 ~ 현재 : 충북대학교 전기전자컴퓨터공학부 교수

<관심분야> : 정보검색, 자연언어처리, 기계번역