질의 응답 시스템을 위한 질의문 심층 분석

Deep Analysis of Question for Question Answering System

신승은, 서영훈
충북대학교 전기전자컴퓨터공학부

Seung-Eun Shin(seshin@nlp.chungbuk.ac.kr), Young-Hoon Seo(yhseo@chungbuk.co.kr)

요약

본 논문에서는 질의 응답 시스템의 성능 향상을 위한 질의문 심층 분석을 제안한다. 일반적인 질의응답 시스템들은 사용자의 자연언어 질의의 의미를 분석하지 않기 때문에 정확한 정답을 제공하는 것이 어렵다. 질의문 심층 분석은 의미자질 추출 문법과 자연언어 질의 특성을 이용하여 사용자의 질의를 의미적으로 분석하고, 의미자질들을 추출한다. 의미자질 추출 문법과 자연언어 질의 특성은 사용자 질의의 의미와 구문 구조를 반영하기 위해 의미자질과 형식형태소로 표현된다. 웹에서 추출한 세부 정답 유형이 '인물'인 100개의 질의에 대한 실험을 통해, 비교적 짧지만 사용자의 질의 의도를 충분히 표현하고 있는 자연언어 질의에 대해 질의문 심층 분석을 수행함으로써 사용자의 질의 의도를 분석하고, 의미자질들을 추출할 수 있음을 보였다.

■ 중심어 : | 질의 응답 시스템 | 질의문 분석 | 의미자질 | 정답 유형 |

Abstract

In this paper, we describe a deep analysis of question for question answering system. It is difficult to offer the correct answer because general question answering systems do not analyze the semantic of user's natural language question. We analyze user's question semantically and extract semantic features using the semantic feature extraction grammar and characteristics of natural language question. They are represented as semantic features and grammatical morphemes that consider semantic and syntactic structure of user's questions. We evaluated our approach using 100 questions whose answer type is a person in the web. We showed that a deep analysis of questions which are comparatively short but enough to mean can analysis the user's intention and extract semantic features.

■ keyword : | Question Answering System | Question Analysis | Semantic Feature | Answer Type |

I. 서 론

정보검색 기술은 인터넷의 발전과 더불어 상업적 응용이 확대되면서 급속히 발전하고 있다. 최근에는 웹 문서의 양이 급격히 증가하면서 세계적으로는 7억 페이지 이상을 색인하는 대용량 문서 색인 기술과 함께 수만에서 수십만의 검색 결과 중에서 사용자가 원하는 의도에

* 본 논문은 2004년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었습니다.
(This work was supported by Chungbuk National University Grant in 2004)

접수번호 : #051230-001
접수일자 : 2005년 12월 30일
심사완료일 : 2006년 02월 22일
교신저자 : 서영훈, e-mail : yhseo@chungbuk.ac.kr

맞는 정보를 정확하게 찾아주는 효과적인 검색 랭킹 기술이 요구되고 있다. 특히 웹과 같은 영역에서의 정보검색은 다양한 분야의 정보들이 서로 연결되어 있는 상황에서 빠르고 정확하게 찾아주는 점에 초점을 맞추어 기술 개발이 집중적으로 이루어지고 있다. 웹 검색 서비스에서 두각을 나타내고 있는 구글(www.google.com)의 경우, 이러한 웹의 특성을 십분 반영하여 Page Rank라는 랭킹 시스템을 도입하여 사용자 입장에서 높은 검색 효과를 얻는 서비스를 제공하고 있다[1][2].

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 정답 또는 정답을 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 정보검색평가대회인 TREC(Text REtrieval Conference)에서는 1999년 TREC-8에서 질의응답시스템의 평가[3]를 시작하였으며, TREC 2001의 경우, 약 25% 정도가 정의(definition)에 관련된 질문이었고, TREC 2002에서는 정답을 포함한 문자열을 제시하는 것이 아니라 실제 정답을 제시하도록 요구하였다. TREC QA 로드맵에 따르면 정답이 단순히 문장의 일부분을 제시하는 것이 아니라 문장들 사이의 추론이 필요한 질문, 새로운 문장을 생성하여 이를 정답으로 제시하거나, 주어진 정답에 대한 배경 설명, 정답의 정당성 검증, 정답의 모호성 해결, 전문가 수준의 의견 제시가 필요한 질문, 이질적 정보의 통합을 통한 정답의 제시 등 점점 질의응답 시스템의 난이도를 높여 갈 계획이다.

보다 더 지능적이고 난이도 높은 질의응답 시스템을 위해 사용자의 질의 의도를 정확하게 분석하여 정답추출에 활용하는 것이 필요하다. 이를 위해 본 논문에서는 한국어 질의응답시스템을 위한 질의문 심층 분석을 제안한다. 사용자의 질의 의도를 파악하기 위해 의미자질 추출 문법과 자연언어 질의 특성을 이용하여 자연언어 질의로부터 세부 정답 유형을 결정하고, 의미자질들을 추출한다. 결정된 세부 정답 유형과 추출한 의미자질들은 질의응답 시스템의 성능 향상을 위해 활용될 수 있다.

II. 관련 연구

질의응답 기술은 사용자의 자연언어 질문과 검색 대상 문서의 의미를 파악하기 위한 자연언어 처리 기술과 대상 문서로부터 정답을 추출하기 위한 정보추출 기술을 필요로 하며, 많은 후보 문서들로부터 답을 포함하는 문서를 걸러주는 역할을 위해 기존의 문서 검색 기술도 활용한다. 질의응답 시스템과 정보검색 시스템의 큰 차이점 중 하나는 자연언어를 입력하여 문서를 검색하는 것이 아니라 정답을 찾는 것에 있다. 이를 위해 질의의 처리과정에서 사용자가 원하는 정답이 무엇인지 질의 의도를 파악할 수 있는 정답유형이나 색인어 등의 정보를 질의로부터 추출한다. 또한, 기존의 정보검색 방법에 의해 질의와 유사한 문서를 추출하고, 문서에서 다시 정답을 포함할 가능성이 있는 단락을 추출한 후, 단락에서 정답 유형과 동일한 개체를 찾아내어 사용자에게 정답으로 제시한다[4].

질의분석은 주어진 질의의 초점이 무엇인지를 분석하는 모듈로서, 질의가 무엇을 초점으로 하는가에 따라 질의의 정답을 찾는데 필요로 하는 질의의 특성을 분석한다. 질의와 단락을 비교하여 적합한 답을 찾는 모듈은 정답 유형에 따라 해당하는 개체가 문서에 나타나고, 질의의 키워드에 해당하는 단어들이 많은 단락에 높은 가중치를 주어 정답으로 추출한다[5-7].

기존 연구에서는 질의분석 모듈은 대부분 패턴 매칭이나 부분 구문 분석을 통하여 해당 정답 유형을 결정하고, 정답에 해당하는 단락을 찾는 모듈에 대하여 서로 다른 방법론을 제시하는 것이 일반적이었다. 이들 연구에는 단순히 질의에 나타나는 키워드만 매칭하여 정답을 찾는 방법, 개체 인식과 사건 인식을 통하여 정답을 찾는 방법, 키워드와 개체를 이용하여 정답을 찾는 방법, 키워드와 의미관계, 개체 등을 이용한 방법 등이 있다[8].

기존의 질의분석은 형태소 분석이나 n-gram 방식으

로 단순히 명사 단어를 키워드로 추출하고, 대분류 수준의 정답 유형을 결정하기 때문에 정답추출 단계에서 사용자의 질의 의도에 적합한 정답을 찾는 것을 어렵게 한다. 따라서 질의응답 시스템의 성능 향상을 위해 사용자의 질의 의도를 정확하게 파악하는 것이 필요하다.

자연언어처리 기술을 이용하여 단어의 의미를 정보검색에 활용하고자 하는 연구들도 있었다. 기존의 의미기반 정보검색에서는 기본적으로 자연언어 텍스트로부터 적절히 색인어를 추출하여 이를 통계적인 검색 모델에 반영하고 있다. 이를 위해서는 의미있는 색인어를 추출하는 것이 무엇보다 중요하다[1][9]. 기존의 의미기반 정보검색은 의미있는 색인어를 추출하기 위해 단어의 의미를 분석하여 색인에 반영하고자 하는 연구들이었다. 따라서 단어 자체의 의미만을 고려하기 때문에 사용자의 질의 의도를 충분히 정보검색에 반영하지 못한다. 따라서 효율적인 정보검색과 질의응답을 위해 단어 자체의 의미뿐만 아니라 사용자가 질의에서 단어를 어떤 의미로 사용하였는지 그 사용 의도를 정확하게 분석해야 한다.

III. 질의문 심층 분석

질의문 심층 분석은 사용자의 질의를 대상으로 자연언어 처리 기법을 적용하고 분석하여 질의의 의도와 내용을 파악하는 것으로, 질의응답 시스템의 성능 향상을 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, 대분류 수준의 정답 유형을 결정하는 역할만을 했으며, 키워드 추출에 의미를 반영하더라도 단어 자체의 의미만을 고려하였다[1][10-12]. 따라서 질의응답 시스템의 성능 향상을 위해 사용자 질의로부터 정답의 세부 유형 결정과 단어가 문장에서 어떤 의도로 사용되었는지를 나타내는 의미자질을 추출하는 질의문 심층 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, 정답 유형을 '인물'로 결정할 것이다. 또한, 문서검색과 정답추출 과정에서 키워드와 정답 유형을 이용한 통계적인 방법에 의해 정답을 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 정답추출을 수행할 수 있다. 예를 들어, "베니스 영화제에서 감독상을 받은 이창동감독은……", "오아시스의 감독은 이창동 씨이며, ……."와 같은 문장을 포함하는 문서를 키워드에 대한 통계 정보와 상관없이 정답 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 정답추출에 반영하지 못한다. 따라서 효율적인 질의응답 시스템을 위해 단어 자체의 의미뿐만 아니라 사용자가 질의에서 단어를 어떤 의미로 사용하였는지 그 사용 의도를 정확하게 분석하는 질의문 심층 분석이 필요하다.

1. 의미자질

의미자질들은 자연언어 질의에서 사용자의 의도를 나타내기 위해 사용되는 단어나 연속된 단어들에 부여하는 의미이다. 자연언어 질의의 형태소 분석 결과로부터 의미있는 형태소들을 함께 결합하는 과정을 거친 후, 결합된 형태소들에게 의미자질을 부여한다.

다음은 세부 정답 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 정답 유형이 '저자'인 위의 자연언어 질의를 살펴보면 '저자'에 대한 질의에 공통적으로 사용되는 의미자질들이 있음을 알 수 있다. '저자'에 대한 질의에 사용된 공통적인 의미자질들은 '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 의미자질들을 이용하여 구성되고 있다.

세부 정답 유형과 의미자질 정의를 위해 TREC의 Test Collection과 Web으로부터 추출된 643개의 인물 관련 자연언어 질의 말뭉치를 분석하여 정답의 세부 유형에 따라 분류하고, 각각의 세부 정답 유형에 따라 질의 구성에 공통적으로 사용된 의미자질들을 정의하였다. 인물 관련 질의들은 '기타'를 포함한 전체 24개의 세부 정답 유형으로 분류되었고, 세부 정답 유형에 대한 전체 의미자질들은 125개의 의미자질들로 정의되었다.

표 1. 세부 정답 유형과 의미자질

세부 정답 유형	의미자질 (Semantic Features)
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 1]은 세부 정답 유형과 의미자질의 일부를 보여주는 예이다. 세부 유형 중 '공통'은 모든 세부 유형에서 공통적으로 사용되는 의미자질들을 말한다.

인물 관련 자연언어 질의 말뭉치로부터 정의된 의미자질에 따라 실질형태소를 추출하고, 동의어/유의어 사전을 이용한 확장을 통해 의미자질 사전을 구축하였다.

의미자질 사전은 2,039개의 실질형태소로 구성되었으며, 278개의 용언들을 포함한다. 의미자질 사전은 의미자질, 의미태그, 어휘목록으로 구성된다.

표 2. 의미자질 사전의 예

의미자질	의미태그	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 의미자질 사전의 예이다. 의미자질에 따른 의미태그는 '자질속성'+'의미'로 붙여진다. 자질속성은 '%', '@', '#' 3가지로 나뉜다. '%' 속성은 문장에서의 사용 위치가 자유로운 의미자질들이며, 질의분석 과정에서 사용자의 자연언어 질의를 의미 태깅한 후 별도로 추출하며 의미자질 추출 문법의 구성과 적용에서 제외하였다. '%' 속성의 의미자질들은 규칙과 의미자질 사전을 이용하여 의미 태깅을 한다. '@' 속성은 문법을 구성하는 기본적인 의미자질로 의미자질 사전을 검색하여 의미 태깅을 수행한다. '#' 속성은 사용자 질의에서 자신을 제외한 모든 의미자질들이 의미자질 추출 문법을 만족하였을 때 질의로부터 추출할 의미자질을 나타낸다.

2. 의미자질 추출 문법

본 논문에서 제시하는 의미자질 추출 문법은 643개의 인물 관련 자연언어 질의 말뭉치를 기반으로 하여 작성되었다. 인물 관련 질의의 의미자질은 실질형태소를 중심으로 부여되며, 의미자질 추출 문법은 세부 정답유형에 따라 각각 의미자질들의 어절 정보들로 구성된다. 문법의 구성에서 문장에서의 위치가 자유로운 의미자질들('%' 속성을 갖는 의미자질들)은 제외하였다. 이러한 의미자질들은 질의분석 과정에서 사전이나 규칙을 통해 먼저 추출되고, 남은 질의에 대해 문법을 이용한 질의분석

을 수행한다. 이것은 자유로운 어순을 갖는 한국어의 특성을 반영하고 질의분석 문법의 수를 줄임으로써 질의분석의 효율을 높이기 위함이다.

[그림 1]은 의미자질 추출 문법의 BNF(Backus- Naur Form) 표현이다.

어절 정보는 실질형태소의 의미자질로만 구성되는 것과 실질형태소의 의미자질과 형식형태소로 구성되는 것이 있으며, 의미자질 추출 문법은 어절 정보들의 리스트로 구성된다.

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

그림 1. 의미자질 추출 문법의 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 의미자질 추출 문법의 예

세부 정답유형 '저자'에 대한 의미자질 추출 문법은 [그림 2]와 같다. [그림 2]에서 실질형태소의 의미자질로만 구성된 어절 정보는 '(@About)'과 '(@Who)'이며 실질형태소의 의미자질과 형식형태소로 구성된 어절 정보는 '(@Author_V etm)', '(#Book jc)', '(#Book co+etm)' 등이 있다. 또한, 문법에서 사용된 기호 '?'는 생략 가능한 형식형태소와 어절 정보를 나타낸다.

3. 자연언어 질의 특성

세부 정답 유형에 대한 의미자질이 정의되지 않은 질의들에 대해서는 자연언어 질의의 특성을 이용하여 의미자질을 추출한다. 이러한 경우 세부 정답유형이 결정되지 않고, 다양한 의미자질을 추출할 수는 없지만 의미자질 추출 문법을 적용한 질의들과 유사한 성능 향상을 보일 수 있다.

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 용언-질의, 질의 7은 명사-질의로 모두 23개의 세부 정답유형에 포함되지 않는 질의이며, 의미자질이 정의되지 않은 질의들이다. 이러한 질의들의 분석은 자연언어 질의 특성을 이용하여 의미자질들을 추출한다.

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

용언-질의와 명사-질의의 자연언어 특성에 따른 어절 정보 리스트는 [그림 3]과 같으며, '#Event_V'는 사용자 질의로부터 추출할 용언을 의미하고 '@' 속성의 의미자질은 의미자질 사전을 이용하여 의미자질을 부여한다.

[용언-질의 어절 정보 리스트]
1. (NP jc) (#Event_V etm) (@Person|@Property_N) (@Who)?
2. (NP jx) (@Who) (#Event_V ef)
3. (@Who) (NP jc) (#Event_V ef)

[명사-질의 어절 정보 리스트]
1. (NP jm) (@Property_N) (@Who)?
2. (NP jc) (@Property_N) (@Who)?
3. (@Property_N) (@Who)?

그림 3. 용언-질의와 명사-질의의 어절 정보 리스트

4. 질의문 심층 분석 결과

질의문 심층 분석 과정은 먼저 질의문의 형태소분석 결과와 의미자질 사전을 이용하여 어절 정보 리스트를 구성하고, 의미자질 추출 문법을 적용한다. 적합한 의미자질 추출 문법이 없는 경우에는 자연언어 질의 특성에 따른 어절 정보 리스트를 적용한다.

(질의 5)와 (질의 8)에 대한 질의문 심층 분석 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)
- 적용 문법
  - (#Book co+etm) (@Genre jm) (@Author_N jx?) (@Who)?
- 질의문 심층 분석 결과
  - 정답 유형 : 인물
  - 세부 정답 유형 : 저자
  - 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)
- 적용된 어절 정보 리스트 : 용언-질의 1
  - (NP jc) (#Event_V etm) (@Person|@Property_N) (@Who)?
- 질의문 심층 분석 결과
  - 정답 유형 : 인물
  - 장소(Location) : 우리나라
  - 단서부사(Adverb) : 최초
  - 이벤트용언(Event_V) : 성공하
  - 특성명사(Property_N) : 의사
  - 명사구(NP) : 간이식 수술

IV. 실험 및 평가

본 논문에서는 실험을 위해 웹에서 100개의 인물 관련 자연언어 질의를 추출하여 실험 질의 말뭉치를 구축하였다. 실험 질의 말뭉치를 대상으로 질의문 심층 분석을 수행하고, 그 결과에 대한 평가로써 정답유형 및 의미자질 추출에 대하여 실험하였다.

정답유형 및 의미자질 추출 실험에서 의미자질 추출 문법과 자연언어 질의 특성을 이용하여 실험 질의 말뭉치의 질의들에 대해 질의문 심층 분석을 수행하고, 분류된 정답유형과 추출된 의미자질의 정확률을 평가하였다. 의미자질은 사용자의 질의 의도를 파악하기 위해 질의로부터 꼭 추출해야 하는 요소로 의미자질 추출 문법을 적용한 경우에는 세부 정답유형에 따라 추출할 의미자질이 다르며, 자연언어 질의 특성을 이용한 경우에는 이벤트 용언, 특성명사, 명사구가 있다.

표 3. 정답 유형 및 의미자질 추출 정확률

	의미자질 추출 문법을 적용한 질의분석	자연언어 질의 특성을 이용한 질의분석
적용된 질의 수	69	31
정답유형 분류 정확률	1.000	0.9032
세부 정답유형 분류 정확률	1.000	*
의미자질 추출 정확률	0.9179	0.8449

[표 3]은 정답유형 및 의미자질 추출 실험의 결과를 나타낸다.

실험 질의 말뭉치 중 69개의 질의가 의미자질 추출 문법이 적용되어 분석되었으며, 31개의 질의가 자연언어 질의 특성에 따라 분석되었다.

4개의 질의가 세부 정답유형이 정의되었으나 의미자질 추출 문법에 의해 분석되지 않고 자연언어 질의 특성에 따라 분석되었다. 그 중 2개는 의미자질 추출 문법이 구성되지 않았고, 나머지 2개는 의미자질 사전에 포함되지 않은 어휘를 포함하고 있었다.

자연언어 질의 특성을 이용한 질의분석의 정답유형 분류 정확률이 다소 낮은 원인은 웹에서 추출한 질의에서 일반적으로 특성명사로 구성된 짧은 질의를 많이 사용하였기 때문이며, 이를 보완하기 위해 의미자질 사전의 확장이 필요하다.

의미자질 추출에서는 명사구 인식 문제가 정확률을 떨어뜨리는 주된 원인이었다.

웹에서 추출한 비교적 짧은 질의에 대해 높은 정답유형 분류 정확률과 의미자질 추출 정확률을 보였으나, 의미기반 질의분석에서 정답유형 분류와 의미자질 추출 정확률을 높이기 위해 의미자질 추출 문법과 의미자질 사전의 보완이 필요하며, 특히 명사구 인식에 대한 추가적인 연구가 필요하다.

V. 결론 및 향후 연구

본 논문에서는 질의응답시스템의 성능 향상을 위한 질의문 심층 분석을 제안하였다. 질의문 심층 분석은 자연언어 질의를 의미적으로 분석하여 사용자의 질의 의도를 파악하고 문서검색 및 정답추출에 필요한 세부 정답 유형과 의미자질들을 추출한다.

자연언어 질의 말뭉치로부터 세부 정답 유형에 따른 의미자질들을 정의하고, 이를 이용하여 의미자질 추출 문법을 구성하였다. 의미자질 추출 문법을 통해 사용자의 자연언어 질의를 의미적으로 분석하여 정답의 세부 유형을 결정하고 의미자질을 추출하였다. 의미자질이 정의되지 않은 경우에는 자연언어 질의 특성을 이용하여 질의문 심층 분석을 수행하였다.

실험에서 세부 정답 유형은 0.69의 재현율과 1.00의 정확률을 보였으며, 의미자질 추출은 0.89의 정확률을 보였다. 실험을 통해 비교적 짧지만 사용자의 질의 의도를 충분히 표현하고 있는 자연언어 질의에 대해 질의문 심층 분석을 수행함으로써 사용자의 질의 의도를 분석하고, 의미자질들을 추출할 수 있음을 보였다.

본 논문에서 제안한 질의문 심층 분석은 정답유형이 인물인 질의를 대상으로 실험하였기 때문에 다른 정답유형에 대한 실험이 필요하며, 질의문 심층 분석 결과를 이용한 다양한 질의응답 시스템의 성능 향상 방안을 고안해야 할 것이다.

참 고 문 헌

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

[2] M. Henzinger, "Google Tutorial: Web Information Retrieval," Tutorial on Web Information Retrieval at ICDE'2000 (16th International Conference on Data Engineering), 2000.

[3] E. M. Voorhees and D. Tice, "The TREC-8 Question Answering Track Evaluation," In Proceedings of the TREC-8, 1999.

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

[5] 김학수, 안영훈, 서정연, "한국어 질의응답시스템을 위한 지지벡터기계 기반의 질의유형분류기", 정보과학회 논문지, 제30권, 제5호, pp.466-475, 2003.
[6] 이승우, 이근배, "유한패턴매칭을 이용한 자연어 질의응답 시스템", 정보과학회지, 제22권, 제4호, pp.21-27, 2004.
[7] 황이규, 윤보현, "HMM에 기반한 한국어 개체명 인식", 정보처리학회 논문지, 제10권, 제2호, pp.229-236, 2003.
[8] 이경순, 김재호, 최기선, "한국어 질의응답 시스템에서 개체인식에 기반한 대답추출", 제12회 한글 및 한국어 정보처리 학술대회, pp.184-189, 2000.
[9] 맹성현, "정보검색 기술의 현황과 발전방향", 정보과학회지, 제22권, 제4호, pp.6-14, 2004.
[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.

저 자 소 개

신 승 은(Seung-Eun Shin) 정회원

- 1999년 : 충북대학교 컴퓨터공학과(공학사)
- 2001년 : 충북대학교 컴퓨터공학과(공학석사)
- 현재 : 충북대학교 컴퓨터공학과(공학박사)

<관심분야> : 정보검색, 자연언어처리

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