한국산학기술학회논문지
Vol. 5, No. 6, pp. 539-545, 2004

자연어 질의유형 판별과 응답 추출을 위한 어휘 의미 체계에 관한 연구

윤 성 희[1*]

A Study on Word Semantic Categories for Natural Language Question Type Classification and Answer Extraction

Sung-Hee Yoon[1*]

요 약 자연어 질의를 입력하고 문서로부터 질의에 대한 정답을 추출하여 제공하는 질의응답 시스템에서는 사용자의 질의 의도를 파악하여 질의 유형을 분류하는 과정이 매우 중요하다. 본 논문에서는 질의 유형을 분류하기 위해 복잡한 분류 규칙이나 대용량의 사전 정보를 이용하지 않고 질의의 의도를 나타내는 어휘들을 추출하고 인접 명사들의 의미 정보를 이용하여 질의 및 정답 유형을 결정할 수 있는 방법을 제안한다. 또 동의어 정보와 접미사 정보를 이용하고, 의문사가 생략된 경우 어휘 의미 정보를 이용하여 질의 유형 분류기의 성능을 향상시킬 수 있음을 보인다.

Abstract For question answering system that extracts an answer and output to user's natural language question, a process of question type classification from user's natural language query is very important. This paper proposes a question and answer type classifier using the interrogatives and word semantic categories instead of complicated classifying rules and huge dictionaries. Synonyms and postfix information are also used for question type classification. Experiments show that the semantic categories are helpful for question type classifying without interrogatives.

Key Words : Question Answering System, question type classification, answer extraction

1. 서 론

일반적인 정의로서 정보검색 시스템(information retrieval system)은 사용자의 질의에 대해 정보가 포함되어 있을 가능성이 높은 문서들의 집합을 찾아주는 시스템이다. 대량의 문서를 검색하고 순위화하여 사용자의 질문에 대한 정답 문서 집합으로 응답한다[1,2]. 검색 과정은 수집된 정보 또는 자료의 내용을 분석한 뒤 적절히 가공하여 축적해 놓은 데이터베이스로부터 사용자의 요구에 적합한 정보를 탐색하여 찾아내는 일련의 과정을 의미한다. 이와 같은 정보검색 시스템에서는 사용자들이 정답 문서 집합으로부터 다시 정답을 찾기 위한 수고를 해야 한다. 반면에 많은 사용자들은 명확한 의도를 가지고 질문을 하며, 정보 검색 시스템이 대량의 문서를 찾아주기 보다는 정답들을 곧바로 찾아 제시해 주기를 원하는 진보된 요구를 만족시키기 위하여 질의 응답(question-answering)이라는 개념이 출현하였고, 많은 연구들이 AAAI와 TREC(Text REtrieval Contest)를 중심으로 수행되어 왔다[3-5]. 질의에 대한 결과로 문서 집합을 제시해 주는 것이 아니라 사용자가 원하는 질문에 대해 문서들로부터 정답을 추출해서 문장이나 단락으로 제시해 주기 때문에 사용자의 부담을 줄여주는 더 지능적이고 편리한 시스템이다. 질의응답 기술은 사용자의 자연어 질의와 검색 대상 문서의 의미를 파악하기 위한 고정밀 자연어 처리 기술과 대상 문서로부터 답을 추출하기 위한 정보추출 기술을 필요로 하며 문서를 걸러주는 역할을 위해 기존의 문서검색 기술도 적용된다[6-10].

자연언어로 된 질의어를 처리할 수 있는 질의응답 시스템에서 중요한 점은 사용자의 질문을 이해하고 질문의 의도를 정확하게 파악하는 것이다. 이 과정은 질의응답 시스템이 정보검색 시스템과 다른 중요한 점은 질의 처리 과정으로서 자연어 질의로부터 사용자의 질의 의도를 파악할 수 있는 질의 유형(question type)이나 키워드(keyword) 등의 정보를 추출하는 것이다. 사용자

[1]상명대학교 컴퓨터소프트웨어공학전공
*교신저자: 윤성희(shyoon@smu.ac.kr)

의 질의 의도를 정확히 파악하고 정답으로서 구하는 것이 무엇인지 결정하는 과정을 질의 유형 분류라고 한다. 특히 질의 유형의 분류 과정은 질의응답 시스템이 문서에서 정답이 될 수 있는 정답 후보(answer candidate)들을 추출하는데 중요한 정보를 제공하게 되므로 질의 유형 분류 과정이 정확하게 이루어진다면 질의응답 시스템의 질적 성능을 향상시킬 수 있다. 국제적인 정보검색평가대회인 TREC에서는 1999년의 TREC-8에서 질의응답 시스템의 평가를 시작하였다[3,5]. 최근 TREC에서 소개된 질의응답 시스템들은 대부분 질의 유형 분류(question type classification)를 위한 모듈을 포함하고 있다.

본 연구에서는 영어권의 언어들에 대한 대규모 언어 지식베이스 등의 풍부한 자원에 비해 상대적으로 부족한 언어 자원의 문제를 해결하기 위해 대량의 코퍼스(corpus)를 이용하거나 복잡한 질의 유형 분류 규칙을 작성하지 않고 어휘 의미 정보를 이용하여 질의 유형을 분류하는 방법을 제안한다. 사용자 질의마다 질의의 초점을 나타내는 어휘가 존재한다면 어휘 정보만을 이용해서 질의 유형을 분류할 수 있으며, 인접 명사의 의미 정보를 활용하기 위해 명사 의미 정보 사전을 구축하고 동의어 사전, 유의어 사전, 접미사 정보 등을 이용하여 질의 유형을 분류하고 정답 유형을 결정할 수 있음을 보여준다.

2. 연구 배경

정보검색의 분야에서는 새로운 방법으로 지식검색이라는 개념이 도입되어 질의에 대한 답을 제공하지만, 사용자의 질의에 대해 다른 사용자가 입력한 응답을 정답으로 제공한다는 점에서 근본적으로 질의응답 시스템의 개념과 크게 다르다. 질문에 대한 정답을 사람이 제시하는 지식검색과 달리 질의응답 기술에서는 사용자의 질문에 대해 질의응답시스템이 문서 등으로부터 정답을 추출하여 사용자에게 제시한다. 단순히 키워드에 의한 검색과 후보 문서 제공의 수준을 넘어서 정답을 포함하는 단락을 검색하고 정답을 추출하는 과정을 포함한다. 질의응답 기술은 현재 정형 데이터베이스에 담겨 있는 상품 정보나 기업 정보를 찾아주는 데이터베이스 질의응답과 게시판 혹은 전자메일 정보를 찾아주는 FAQ(Frequently Asked Question) 질의응답에 부분적으로 활용되고 있다[11-13]. 데이터베이스에서 사용자가 원하는 자료를 검색하는 데이터베이스 자연어 질의처리 기술은 자연어 질의문을 SQL 등의 질의처리 언어를 대신하여 입력한 다음 질문을 의미적 수준에서 파싱(parsing)하고 이를 직접 데이터베이스 정형 질의어(formal database query)로 변환하는 과정을 거친다. 그러나 고정밀 자연어 처리, 정교한 정보추출 등의 텍스트 마이닝(text mining) 기술이 요구되는 백과사전 지식정보나 전자 매뉴얼 정보에 대한 질의응답 기술은 현재 국내외적으로 연구가 진행 중이다[7,8,9,14,15].

질의응답 시스템과 정보검색 시스템의 큰 차이는 자연어를 질의로 입력하고 문서를 검색하는 것이 아니라 정답을 찾아 제공하는 것이다[1,16]. 이를 위해 질의문의 처리과정에서 사용자의 질의 의도를 파악하기 위해 질의유형을 분류하고 키워드 등의 정보를 질의로부터 추출한다. 한편으로는 일반적인 정보검색과 비슷한 방법으로 정답을 포함할 후보 문서를 선정하고, 문서에서 다시 정답을 포함할 가능성이 있는 단락을 추출한 후, 단락에서 질의 유형과 일치하는 개체를 찾아서 사용자에게 정답으로 제공한다. 일반적인 질의응답 시스템의 구조는 그림 1과 같다.

질의 유형 분류는 사용자의 질의 의도를 특정한 범주(category)에 할당하는 것으로 질의응답 시스템 연구의 한 분야로 진행되어 왔다. 기존의 질의 유형 분류 기법은 규칙에 기반한 방법(rule-based method)과 통계에 기반한 방법(statistical method)으로 나뉘어 진다.

규칙에 기반한 질의 유형 분류를 채택하고 있는 시스템들은 일반적으로 어휘-구문 패턴(lexico-syntactic pattern)을 구축하고, 이러한 패턴을 유한 상태 오토마타(finite state automata)와 매치(match)하여 질의 유형을 분류한다. 일반적으로 규칙에 기반한 접근 방법을 채택한 질의응답 시스템들은 질의 유형 분류 과정이 유한 상태 오토마타로 구현되므로 사용자의 질의에 대해서 즉각적으로 질의 유형을 분류해 낼 수 있다. 또한 수동으로 기술된 규칙에 따라 질의 유형을 분류하므로 질의 유형을 잘못 분류해서 전혀 관계없는 대답을 하는 경우를 방지할 수 있다. 그리고 응용 영역이 정해져 있을 경우 간단한 튜닝(tuning)으로 성능을 향상시킬 수 있다. 그러나 규칙을 수정하기 위해서 전문적이 지식을 가진 사람들의 노력이 필요하고, 규칙과 일치되지 않는 질의가 들어 왔을 때는 질의 유형을 분류할 수 없는 문제점을 가지고 있다. 또한 규칙이 많아질수록 좋은 성

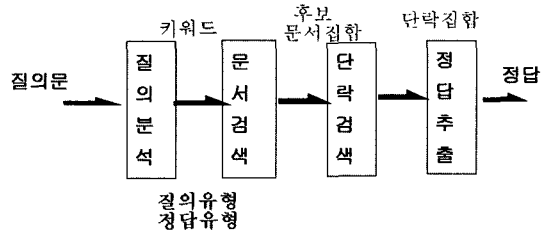

그림 1. 질의응답 시스템의 구조

능을 내기 위한 튜닝이 더 어려워지게 되며, 시스템이 다른 응용 영역에서 사용될 경우에는 기존의 규칙들을 모두 수정하거나 많은 부분을 다시 작성해야 하는 문제점이 있다. 규칙이 많아질수록 좋은 성능을 얻기 위한 튜닝이 점점 더 어려워지게 되며, 시스템이 다른 응용 영역에서 사용될 경우에는 기존의 규칙들을 모두 수정하거나 재작성해야 하는 문제점이 있다.

반면, 통계적인 방법에 기반한 질의 유형 분류는 수동으로 분류된 대량의 학습 데이터로부터 추출한 통계 정보를 이용한다. 질의 유형 분류를 위해 최대 엔트로피 모델(maximum entropy model)을 사용하거나 결정적 LR 파서인 Contex를 확장한 예가 있다. 통계에 기반한 질의 유형 분류 방법은 대량의 학습 데이터를 이용한 통계 모델을 사용하기 때문에 안정적으로 질의의 유형을 분류할 수 있으며, 자동화된 통계적 방법을 사용함으로써 시스템 구축을 쉽게 할 수 있다. 그러나 사용자가 질의에서 의도하지 않은 결과를 정답으로 추출하는 경우가 자주 있으며, 이러한 경우 보완이 쉽지 않고 대량의 학습 데이터를 이용하여 추출해야 하는 어려움이 있다[6,17].

본 연구에서는 질의 유형을 분류하기 위해 복잡한 분류 규칙이나 대용량의 사전 정보를 이용하지 않고 사용자의 자연어 질의문에서 질의의 초점이 되는 의문사에 해당하는 어휘들을 추출하고 주변에 나타나는 명사들의 의미 정보를 이용하여 질의에 대한 정답 유형을 결정할 수 있는 질의 유형 분류 방법을 소개한다. 또한 의문사가 생략된 경우에 주변에 출현하는 어휘들의 의미 정보를 이용하는 질의 유형 분류 방법과 동의어 및 유의어 정보와 접미사 정보를 이용하여 질의 유형 분류 성능을 향상시킬 수 있는 방법을 소개한다.

3. 질의 분석과 유형 분류

질의응답 시스템의 구축은 그림 2에서 보는 바와 같이 크게 질의분석 모듈(module)과 정답추출 모듈로 구성된다. 질의분석 모듈은 주어진 질의의 초점이 무엇인지를 분석하는 모듈로서 질의가 무엇을 초점으로 하는가에 따라 질의의 답을 찾는데 필요로 하는 질의의 특성을 분석한다. 정답추출 모듈은 문서검색과 단락검색 후 정답추출 과정을 거치는데 본 연구는 질의 분석 모듈의 질의 유형 분류에 초점을 맞춘 연구이다. 질의의 유형 분류하고 정답의 유형으로 결정되는 개체는 예를 들어 '사람', '조직', '장소', '거리', '시간' 등에 해당하는 어휘들이다. 질의 분석의 결과와 검색 문서의 단락을 비교하여 적합한 답을 찾는 모듈은 질의 유형에 따라 해당하는 개체가 문서에 나타나고 질의에 해당하는

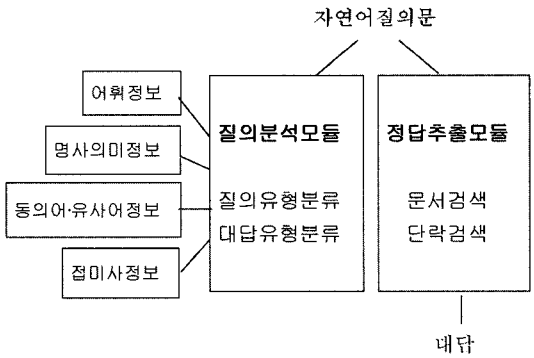

그림 2. 어휘정보를 이용하는 질의유형분류

단어들이 많은 단락에 높은 가중치를 주어 정답으로 추출한다. 기존 연구에서는 질의 분석 모듈은 대부분 패턴 매칭(pattern matching)이나 부분 구문 분석(syntatic analysis)을 통하여 해당 질의 유형을 결정하고, 질의에 해당하는 단락을 찾는 모듈에 대하여 서로 다른 방법론을 제시하는 것이 일반적이었다[7,8,17,18].

3.1 질의 어휘에 의한 유형 분류

3.1.1 의문사 정보 이용

한국어의 의문문 형태로 나타나는 질의 문장의 경우 대부분은 문장의 마지막에 의문의 초점을 나타내는 중요한 정보를 가지고 있다. 각각의 질의 유형마다 질의의 초점을 나타내는 어휘로서 의문사가 존재한다면 어휘 정보만 이용해서 질의 유형을 분류할 수 있다. 질의 문장에 질의의 초점을 나타내는 어휘를 이용해서 질의 유형을 결정한다. '언제'가 나타났을 때는 기본적으로 '시간'을 대답 유형으로 결정하고 앞에 나타나는 어휘에 따라 세부적인 대답 유형을 결정하도록 한다. 예를 들어 "대한민국의 수도는 어디인가?"에서 '어디'에 의해 '장소'를 질의 유형으로 정하고, '수도'에 의해 '지명'을 대답 유형으로 결정하는 방법이다. 예에서 '조직'과 '장소'는 질의의 초점을 나타내는 어휘가 모두 '어디'로 나타나기도 하는데, '어디'의 앞에 나타나는 어절의 단어를 살펴서 질의 유형을 판단할 수 있다.

질의유형에 따라 정답의 유형을 찾고 정답 후보를 생성하기 위해서는 질의 유형에 대한 하위 의미 범주를 분류할 필요가 있다. TREC-10 및 TREC-11에 참가한 질의응답 시스템들을 참고하여 구축한 의미 정보 사전의 예가 다음의 표에 나타나있다. 표의 질의 유형 분류기는 자연어로 된 질의 문장에 대하여 질의 유형을 분류하고 각각의 질의 유형에 대해 다시 세분화된 하위 의미 범주로 분류한다. 각각의 질의 유형에 대해 세분

화된 하위 의미 범주는 정답 유형을 결정하고 정답 후보를 결정하는데 유용하게 사용될 수 있다. 표 1은 질의의 초점을 나타내는 어휘를 중심으로 질의 유형을 분류하는 예를 보여주며, 표 2는 질의 유형에 관련되는 일부 어휘들의 하위범주 예를 보여준다.

3.1.2 동의어 및 유의어

질의 유형에 대해 하위 의미 범주로 분류하기 위해 명사 의미 사전을 이용하였다. 그러나 모든 명사에 대한 의미정보를 사전으로 구축하는 것은 불가능하므로 질의 문장에서 다양한 형태의 출현 가능한 명사들을 분류하기 위해서 대용량의 사전을 구축하지 않고 동의어와 유의어 정보를 사용하는 방법과 접미사 정보를 이용하는 방법을 채택한다. 예를 들어, '작가'는 '글쓴이, 소설가, 문필가, 집필자, 문예가, 저작가, 제작자, 대문호, 저자, 지은이 ...' 등과 같은 의미로, 또 '장소'는 '곳, 데, 처소, 지점, 부분, 점, 위치, 지역 ...' 등과 같은 의미로 사용되는 단어들로 등록된다.

3.1.3 접미사 정보를 이용한 성능 향상

접미사는 접사의 하나로 낱말의 끝에 붙어서 의미를 첨가하여 다른 낱말을 이르는 말이다. 단독으로는 사용

표 1. 어휘정보와 질의유형의 예

어휘	질의 유형	질의 문장 예	
누구 누가	Human (사람)	'... 우승자는 누구입니까(누구인가)?' '... 누가 우승하였는가?' '... 누구의 발명품인가?'	
+ 어디 어느 +	Location (장소)	'... 개최 장소는 어디입니까(어디인가)?' '... 어디에서 개최되었는가?' '... 어느 나라에서 개최되었는가?' '... 어느 곳에서 ...?'	+의 의미정보
+ 언제	Time (날짜나 시간)	'... 전시회는 언제 열리는가?' '... 언제 전시회가 열리는가?' '... 개막일은 언제인가'	+의 의미정보
+ 얼마 얼마나 몇 +	Number (수)	'... 인구는 얼마나 많은가?' '... 강수량은 얼마인가?' '... 금메달은 몇 개인가?' '... 몇 시인가?'	+의 의미정보
어떤 + 어느 + + 어디	Organization (조직)	'... 어느 학교인가?' '... 어떤 물질인가?' '... 우승팀은 어디인가?' '... 가장 비가 많이 오는 것은 어느 계절인가?'	+의 의미정보
무엇 무슨 + 어떤 + 어느 +	Object (물체나 사물)	'... 무슨 곡인가?' '... 어떤 행성인가?' '... 어떤 방법인가?' '... 물질은 무엇인가?' '... 이유는 무엇인가?' '... 어느 기관인가?'	+의 의미정보
왜	Why (이유)	'왜 ... 참석하였는가?' '... 왜 해체되었는가?'	
어떻게	How (방법)	'... 어떻게 발생하였는가?' '어떻게 ... 분리하였는가?'	
+		'... 사람은?' '... 장소는?' '... 년도는?' '... 수는?' '... 우승자는?' '... 영화는?'	+의 의미정보
	Unknown		

표 2. 질의유형과 의미 하위범주 예

어휘	질의 유형	하위 의미 범주 예
누구 / 누가	Human (사람)	Artist, Politician, Economics
+ 어디 / 어느 +	Location (장소)	Place, Continent, State, Capital
+ 언제	Time (날짜나 시간)	Year, Month, Day, Season, Period
+ 얼마 / 얼마나 / 몇 +	Number (수)	Count, Price, Percent, Weight, Height
어떤 + / 어느 + / + 어디	Organization (조직)	School, Company, Government
무엇 / 무슨 + / 어떤 + / 어느 + /	Object (물체나 사물)	Planet, War, Religion, Reason, Organ
	Unknown	

될 수 없고 항상 다른 단어의 어근 뒤에 결합되어 여러 가지 의미를 첨가해 주는 역할을 한다. 질의 문장이 여러 가지 종류의 접미사가 붙어 다양한 형태로 나타나기 때문에 접미사 처리를 하지 않는다면 시스템의 성능을 저하시키는 원인이 된다. 예를 들어 "... 개최국은 어디인가?" "... 참가국들은?"에서 '어디'는 '국가'를 대답 유형을 결정해야 한다.

반면에 중의적(ambiguous)인 뜻을 갖는 접미사는 질의 유형을 분류하기에 곤란한 경우가 많으므로, 이 문제에 대한 해결 방안을 강구할 필요가 있다. 중의적으로 사용되는 접미사의 예로서 접미사 "장"은 "운동장", "농구장", "각축장", "위원장", "위촉장", "초대장", "고추장", "청국장" 등과 같이, 또 접미사 "국"은 "개최국", "생산국", "참가국", "동맹국", "미역국", "북어국", "된장국", "토란국" 등과 같이 중의적 의미로 사용될 수 있다.

3.2 인접 명사에 의한 질의 유형 분류

앞의 절에서 서술한 바와 같이 질의문에 대한 질의 유형을 분석하고 결정하는 데는 의문사가 중요한 역할을 한다. 그러나 질의에 대한 대답 유형을 알 수 있는 실마리가 질의 문장에 나타나지 않을 경우는 한국어에서 매우 흔하게 나타나는 생략 현상으로 인한 경우이다. 한국어는 생략이 빈번하게 발생하는 언어이기 때문에 질의의 초점을 나타내는 어휘가 존재하지 않는 경우가 많기 때문에 의문사 정보만을 가지고 질의 유형을 분석하는 데는 한계가 있다. 이러한 경우는 질의문의 마지막 어절에 위치하는 명사의 의미 정보를 분석하여 질의 유형과 대답 유형을 결정할 수 있다. 이와 같은 생략형 질의에서 마지막 명사의 의미 계층 구조에서 대답 유형의 하위 개념인지에 따라 결정한다. 대답 유형으로 정의된 것에 속하지 않을 때는 해당 의미 계층 정보를 대답 유형으로 결정한다. "삼국지의 저자는?"에서 '저자'의 상위 개념을 거슬러 올라갔을 때, '사람'의 개념 분류와 일치하므로 대답 유형을 '사람'으로 결정한다. 질의 문장에 나타나는 어휘들의 의미와 이들의 출현 규칙을 살피고 추출한 어휘들의 정보를 담고 있는 사전을 이용하여 질의 유형을 분류할 수 있다. 또한 질의 유형에 대한 자세한 분류를 하여 정답 후보들의 적합성을 판단하고 정답을 추출하는데 중요한 역할을 한다. 예를 들어 "... 영국의 도시는?" "... 인물은?" "... 선수는?"와 같이 의문의 초점이 생략된 질의의 경우에는 마지막 어휘인 '도시', '인물', '선수' 등의 하위 범주에 의해서 '장소', '사람', '사람' 등으로 질의 유형을 분류할 수 있다. '어느 ...'와 '어떤 ...'의 경우에도 인접하여 나타나는 명사의 의미 정보에 따라 대답 유형을 결정할 수 있다. 대답 유형으로 두 개 이상이 나타나는 경우 각각의 경우에 대해서 대답 유형을 찾는다.

4. 실험 결과 및 평가

질의응답 시스템을 평가하는 대표적인 TREC의 QA Track에서는 질의응답 시스템을 평가하기 위해 대량의 평가셋을 구축하고 성능평가를 위한 평가 척도를 개발하고 있다. TREC-8에서는 약 53만 문서의 코퍼스로부터 수작업으로 작성한 200여 질문에 대해서 50 바이트 및 200 바이트 크기의 정답을 포함한 단락을 제시하도록 하였다[18,20,22]. 표 3은 TREC-8에서 제시된 질문의 유형을 분류한 분포이다.

본 연구에서는 제안한 방법을 이용하여 사용자의 자연어 질의유형과 대답 유형의 분류가 얼마나 정확하게 이루어졌는지를 실험 평가하기 위하여 실험을 위하여 키워드 검색에 매우 익숙한 대학생들을 대상으로 자연어 질의 문장을 수집하였다. 정보검색 시스템과 질의응답 시스템의 차이를 간단히 설명하고, 정답 문장이나 정답 단락을 얻기 위한 목적의 가상의 질의응답 시스템에 대한 질의 문장을 자연어로 입력하도록 하였다.

표 3. TREC-8의 질의유형 분포

유형	질의 수	유형	질의 수
what	64	which	10
who	47	why	2
how	31	whom	1
where	22		
when	19	기타	

실험 질의문들을 수동으로 분류한 결과와 본 논문에서 제안한 어휘 의미정보를 이용하는 분류결과를 비교하기 위하여 검색과 정답 추출의 대상이 되는 문서들을 수집하고 처리과정을 거쳤다. 검색 대상의 문서들은 최근의 뉴스, 개인 및 단체 홈페이지, 연예 및 스포츠 기사들을 중심으로 선정하였으며, 문서의 내용을 태깅(tagging)하여 얻은 명사들을 중심으로 의미 범주를 할당하였으며, 수집된 어휘들에 대한 실험용 동의어 및 유의어 사전을 구축하였다. 사전의 규모는 품사별, 기능어별로 구분할 수 있으며, 또한 형태소 분석을 위한 기능어 사전이나 구문분석을 위한 품사 사전들로 구분될 수 있다. 본 실험에 사용된 기본 동사 사전의 엔트리 수는 약 1,200개이며 질의 유형 분류를 위한 의미 표지를 담은 명사 사전의 엔트리 수가 약 27,400이다. 사전의 구축 방식과 실험 집합에 따라서 엔트리 수는 쉽게 증가될 수 있으므로 그 절대 엔트리 수는 본 연구에서는 크게 의미가 있는 것은 아니다. 예를 들면, '하다'와 결합된 다수의 동사들은 기본 엔트리로 사전에 등록되는 방법과 형태소 분석에 의해서 명사와 분리되어 해석될 수 있다. 본 연구에서 보다 중요한 점은 정확한 응답 생성을 위해 질의문의 의도를 유형별로 분류하기 위해 어휘의 의미정보를 활용한다는 점이다. 또한 본 논문에서 언급하고 있는 대용량의 언어자원이란 사전의 엔트리 수를 의미하기 보다는 대량의 코퍼스로부터 가공되고 추출되어야 하는 고급 통계적 지식자원을 의미하고 있다. 사전 어휘 항목이 증가되는 것은 쉬운 일이지만 각 항목이 담고 있는 유용한 고급 지식 자원을 축적하는 것이나 이러한 사정항목의 고급 언어자원을 대신할 대량의 복잡한 규칙집합을 구축하는 것이 매우 어려운 작업이므로 이를 대신하여 어휘의 의미정보를 이용하여 질의유형을 분류하고자 하는 것이 본 연구의 의도이다. 따라서 사전의 항목수와 크게 관계없이 의미정보들이 얼마나 체계적으로 구축되어 있는가에 따라서 시스템의 성능이 크게 좌우될 수 있다.

본 연구를 위하여 약 2,400 문장의 자연어 질의문이 수집되었으며, 질의 유형은 위 TREC-8의 분류와 유사한 분포를 보였다. 어휘 의미정보를 이용하여 질의 유형을 분류하는 실험에서 수동 분류 결과와 비교할 때 약 88.6%의 분류 성공률을 보여서 질의 유형의 자동 분류에 대한 높은 가능성을 나타냈다. 또한 앞에서 소개한 생략형 질의가 38%를 차지하였으나, 그 중에서 인접 명사의 의미 정보를 이용하여 질의 유형 분류가 성공적으로 이루어진 질의는 87.3%가량이며 실패 경우의 대부분은 명사 의미 범주가 등록되지 않은 경우에 해당되었다.

실패 유형을 중심으로 살펴보면 등록되지 않은 명사에 대한 의미 정보를 참조하는 경우에 질의 유형이 분류되지 못한 질의 문장이 많아서 보다 정교한 사전의 구축될 경우 극복할 수 있는 문제로 보인다. 특이한 점은 사용자들이 입력한 상당 수 질의 문장들은 대답으로서 단순히 검색된 문서와 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 등의 지능적 해결을 기대하는 질의들이었다는 점이다. 예를 들면 "대북관계는 앞으로 어떻게 진행될까?"와 같은 질문들이다. 이와 같은 질의들의 경우는 어휘 의미에 의존하여 문서나 단락 검색에 의해 정답을 추출할 수 없는 경우들이기 때문에 수동 분류의 경우에도 객관적 유형 분류가 어려운 질의들이다. 또한, 대답 추출의 난이도 측면에서 살펴볼 때, TREC 2001에서 보인 바와 비슷하게 수집된 질의 문장의 약 21% 정도가 "세마포어 연산이란 무엇입니까?"나 "6-시그마란 무엇입니까?" 등과 같은 어떤 정의(definition)에 관련된 질문 문장들이었다[21,22]. 다양한 자연어 질의문의 갖는 이와 같은 특이한 현상들은 일반 사용자들의 자연어 질의 처리와 정답 추출 성능에 대한 기술적 기대가 매우 높다는 것을 보여준다.

5. 결 론

질의응답 시스템의 성능을 향상시키기 위해서 질의 분석 과정에서 질의 및 정답 유형의 분류가 정확하게 이루어져야 한다. 본 논문에서는 영어권의 언어들에 대한 대규모 언어 지식베이스 등의 풍부한 자원에 비해 상대적으로 부족한 언어 자원의 문제를 해결하기 위해 대량의 코퍼스(corpus)를 이용하거나 복잡한 분류 규칙을 작성하지 않고 질의 유형을 분류하는 방법을 제안하였다. 사용자의 질의 의도를 파악하기 위해서 사용자 질의 문장에서 어휘 정보와 명사 의미 정보를 중심으로

이용하여 질의 유형과 정답 유형을 결정할 수 있도록 하였다. 의문문의 형태로 나타나는 한국어 질의 문장에서 대부분은 문장의 마지막에 질의의 초점을 나타내는 중요한 정보를 가지고 있어서 질의마다 질의의 의도을 나타내는 어휘가 존재한다면 어휘 정보만 이용해서 질의 유형을 분류할 수 있다는 것을 보였다. 질의의 초점을 나타내는 어휘가 의문사로서 질의 문장에 존재하는 경우, 주변에 출현하는 명사들을 추출하여 명사 의미 정보 사전을 이용하여 질의 문장을 세부 단계까지 분류하여 질의응답 시스템에서 정답 후보 생성 시 효과적으로 사용할 수 있다. 생략에 의해 의문의 초점을 나타내는 어휘가 생략되는 경우, 질의문의 마지막 어절의 명사를 추출하여 명사 의미 정보 사전을 이용하여 질의 유형의 해당 범주를 분류하는 방법을 사용한다. 동의어, 유의어, 접미사 정보를 이용하여 질의 유형 분류의 성능을 향상시킬 수 있다. 복잡한 구문 규칙이나 언어 자원, 대용량의 사전 정보, 코퍼스, 통계 정보 등을 이용하지 않고도 충분히 만족할 만한 질의 유형 분류를 할 수 있음을 실험을 통하여 확인하였다. 동의어 사전, 유의어 사전, 접미사 사전 정보 등을 이용하여 질의 유형 분류 성능을 향상시킬 수 있고, 질의의 초점 어휘가 생략된 경우에도 정교하게 구축된 명사 의미 정보 사전과 정답 유형을 결정할 수 있음을 실험을 통해 보여주었다. 질의 유형의 하위 의미 분류를 보다 다양하고 폭넓게 적용하여 사용자 질의 문장에서 정답 유형을 더 구체적으로 제시할 수 있도록 하는 연구가 계속될 필요가 있다.

참고문헌

[1] 장명길, 김현지, 장문수 외, "의미기반 정보검색", 정보과학회지 한글정보처리 특집, 제19권 제10호, pp. 7-18, 2001. 10.

[2] 정규철, 서영훈. "어휘정보와 명사 의미 정보를 이용한 사용자 질의 문장 분석", 한국콘텐츠학회 추계종합학술대회 논문집- IT 기반기술 분야, 2003.

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

[4] Jimmy L., Boris K., "Question Answering Techniques for the World Wide Web" 10th Conference of the European Chapter of the Association for Computational Linguistics(EACL-2003), 2003.

[5] TREC(Text Retrieval Conference) Overview, http://trec.nist/overview.html.

[6] 강승식, 이하규, 손소현, 문병부, 홍기채, "자연언어 질의 문장의 용어 가중치 부여 기법", 한글 및 한국어 정보처리학술대회 발표논문집, pp. 223-227, 2003. 10.

[7] 김학수, 안영훈, 서정연, "한국어 질의응답 시스템을 위한 지지벡터기계 기반의 질의 유형 분류기", 한국정보과학회논문지, 제 30권, 제 5호, pp. 466-475, 2003. 6.

[8] 신승은, 이대연, 서영훈, "구문관계 정보를 이용한 한국어 질의-응답 시스템", 한국콘텐츠학회논문집, 제4권, 제2호, pp. 36-42, 2004.

[9] 이경순, 김재호, 최기선, "한국어 질의응답 시스템에서 개체인식에 기반한 대답 추출", 한글 및 한국어 정보처리학술대회 발표논문집, pp. 184-189, 2000. 10.

[10] 이경순, 김재호, 최기선, "질의응답 시스템의 평가를 위한 테스트컬렉션 구축", 한글 및 한국어 정보처리학술대회 발표논문집, pp. 190-197, 2000.

[11] 김현돈, 조성배, "한메일넷 질의 자동응답을 위한 이단계 자기구성 지도", 정보과학회 춘계학술발표논문집, pp. 481-483, 2000. 4.

[12] 안현준, 김현돈, 조성배, "한메일 FAQ의 개념적 검색을 위한 계층적 브라우징 시스템", 한국정보처리학회 춘계학술발표논문집, 제7권, 제1호, 2000. 4.

[13] 원정임, 윤지희, 이건배, "유사객체 검색에 의한 협력질의 응답", 한국정보처리학회 추계 학술발표논문집, pp. 481-486, 1997. 10.

[14] 김학수, 안영훈, 서정연, "하이브리드 방법의 사용자 질의 의도 분류", 정보과학회 논문지, 제30권, 제1호, pp. 51-57, 2003. 2.

[15] 양수정, 서영훈, "질의문의 구문정보를 이용한 키워드 추출", 한국콘텐츠학회 2003 추계 종합학술대회 논문집, 제1권, 제2호. pp. 190-194, 2003.

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

[17] 이대연, 서영훈, "구문구조를 이용하여 정답을 추출하는 질의 응답 시스템", 제15회 한글 및 한국어 정보처리학술대회 발표논문집, pp. 89-94, 2003. 10.

[18] 이재홍, 최호섭, 옥철영, "개념어의 습득을 위한 지식기반 질의응답 시스템", 제15회 한글 및 한국어 정보처리학술대회 발표논문집, pp. 95-100, 2003. 10.

[19] Baeza-Yates Ricardo and Reberio-Neto Berthier, "Modern Information Retrieval", 1999.

[20] Edward H, Hermjakov U, Lin CY, Ravichandran D, "Using Knowldege to Facilitate Factoid Answer Pinpointing", COLING, 2002.

[21] Jimmy L, "The Web as Resource for Question Answering", LREC, 2002.

[22] J. Burer, C. Cardoe., "Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)", NIST DUC Vision and Roadmap Documents, http://www-nlpir.nist.gov/projects/duc/roadmapping.html, 2001.