단어 의미 정보를 활용하는 이용자 자연어 질의 유형의 효율적 분류*

Efficient Classification of User's Natural Language Question Types using Word Semantic Information

윤 성 희(Sung-Hee Yoon)**
백 선 욱(Seon-Uck Paek)***

초 록

질의응답 시스템에서의 질의 분석 과정은 이용자의 자연어 질의 문장에서 질의 의도를 파악하여 그 유형을 분류하고 정답 추출을 위한 정보를 구하는 것이다. 본 연구에서는 복잡한 분류 규칙 집합이나 대용량의 언어 지식 자원 대신 이용자 질의 문장에서 질의 초점 어휘를 추출하고 구문 구조적으로 관련된 단어들의 의미 정보에 기반하여 효율적으로 질의 유형을 분류하는 방법을 제안한다. 질의 초점 어휘가 생략된 경우의 처리와 동의어와 접미사 정보를 이용하여 질의 유형 분류 성능을 향상시킬 수 있는 방법도 제안한다.

ABSTRACT

For question-answering system, question analysis module finds the question points from user's natural language questions, classifies the question types, and extracts some useful information for answer. This paper proposes a question type classifying technique based on focus words extracted from questions and word semantic information, instead of complicated rules or huge knowledge resources. It also shows how to find the question type without focus words, and how useful the synonym or postfix information to enhance the performance of classifying module.

키워드: 질의응답 시스템, 질의 유형 분류, 정답 추출, 단어 의미 정보, 자연어 질의
question-answering system, question type classification, answer extraction, word semantic information, natural language question

---

* 이 논문은 상명대학교 2004년 교내 연구비 지원의 결과임.
** 상명대학교 컴퓨터소프트웨어공학전공 부교수(shyoon@smu.ac.kr)
*** 상명대학교 컴퓨터소프트웨어공학전공 부교수(paeksu@smu.ac.kr)
■ 논문접수일자 : 2004년 11월 19일
■ 게재확정일자 : 2004년 12월 6일

1. 서 론

이용자의 질의에 대해 대량의 문서들을 검색하고 정답이 포함되어 있을 가능성이 높은 문서들을 선별하여 순위화된 정답 후보 문서 집합으로 제시하는 소프트웨어 도구인 정보 검색 시스템(information retrieval system)에서는 일반적으로 이용자들이 검색된 결과 문서들로부터 실제로 원하는 정답을 다시 찾아야 하는 불편함이 있다. 그러나 많은 이용자들은 명확한 의도를 가지고 질문을 하며, 정보 검색 시스템이 대량의 후보 문서들을 찾아주기 보다는 더 나아가 정답들을 곧바로 찾아 제시해 주기를 원한다. 예를 들면, "한글은 누가 만들었는가?" 또는 "누가 한글을 만들었나?"라는 질의에 대해 "세종대왕"이라는 정답을 제공하는 것이다. 이러한 이용자의 요구를 만족시키기 위하여 질의응답(question answering)이라는 개념이 출현하였고, 많은 연구들이 AAAI와 TREC (Text REtrieval Contest)를 중심으로 수행되어 왔다(AAAI Fall Symposium on Question Answering; Burger J. and Cardie C. 2001; Ellen M. Voorhees 1999; TREC). 이용자의 질의에 대한 응답으로 후보 문서 집합을 제시해 주는 것이 아니라 구체적인 정답을 문서로부터 추출해서 문장이나 단락으로 제시해 주기 때문에 이용자의 부담을 줄여 주는 더 지능적이고 편리한 시스템이다.

질의응답 기술은 궁극적으로 이용자의 자연어 질의와 검색 대상 문서의 의미를 파악하기 위한 고정밀 자연어 처리 기술과 기존의 문서 검색 기술을 적용하며, 대상 문서로부터 정답을 추출하기 위한 정보 추출(information extraction) 기술도 필요로 한다(황이규, 김현진, 장명길 2004; Baeza Yates R. and Reberio Neto B. 1999).

질의응답 시스템이 정보검색 시스템과 다른 중요한 점 중의 하나는 질의 처리 과정으로서 이용자의 질의를 이해하고 질의의 의도를 정확하게 파악하고자 하는 과정이다. 이용자의 자연어 질의에서 질의 의도를 판별할 수 있는 질의 유형(question type) 결정 어휘나 키워드(keyword) 등의 정보를 추출하는 것이다. 이용자가 정답으로서 구하는 것이 무엇인지 질의 의도(question point)를 파악하는 과정을 질의 유형 분류(question type classification)라고 하며 질의응답 시스템이 대량의 문서 집합으로부터 정답이 될 수 있는 정답 후보(answer candidates)들을 추출하는데 매우 중요한 정보를 제공한다. 질의 유형 분류가 정확하게 이루어진다면 정답 추출 과정을 비롯한 질의응답 시스템의 성능을 크게 향상시킬 수 있다.

질의응답 시스템을 평가하는 대표적인 것은 미국 NIST(National Instituge of Standatds and Technology)에서 주관하는 질의응답 평가대회 TREC의 QA Track이다. 국제적인 정보검색 평가로서 1999년의 TREC-8에서부터 질의응답 시스템의 평가를 위해 대량의 평가 셋(set)을 구축하고 성능 평가를 위한 평가 척도를 개발하였다(Burger J. and Cardie C. 2001; Ellen M. Voorhees 1999). 최근 TREC에서 소개된 질의응답 시스템들은 대부분 질의 유형 분류(question type classification)를 위한 별도의 모듈을 포함하고 있다.

본 논문에서는 영어권의 언어들에 대한 대규모 언어 지식베이스 등의 풍부한 자원에 비

해 상대적으로 부족한 한국어 언어 자원의 문제를 해결하기 위해 대량의 코퍼스(corpus)를 이용하거나 복잡한 규칙을 작성하지 않고 단어 의미 정보를 중심으로 질의 유형을 분류하는 방법을 제안한다. 이용자의 자연어 질의 문장에서 질의의 초점을 나타내는 단어를 추출하고 구조적으로 인접한 단어들의 의미 체계 정보(semantic category)를 이용해서 질의 유형을 분류할 수 있음을 보이고자 한다. 단어 의미 체계 정보 사전을 구축하고, 한편으로는 동의어 사전, 유의어 사전, 접미사 정보 등을 이용하여 효과적으로 질의 유형을 분류하고 정답 유형을 결정하는 방법을 제안하고 있다.

2. 관련 연구

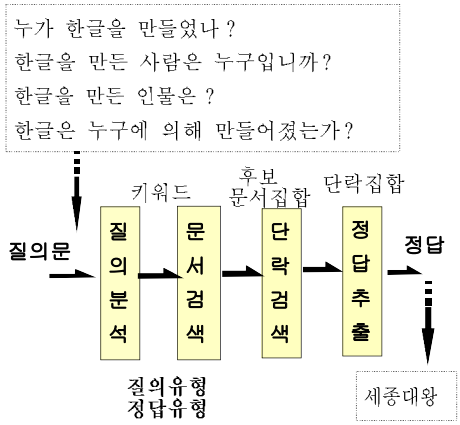
2.1 질의응답 기본 기술

새로운 정보검색의 방법으로서 '지식검색'이라는 개념이 도입되어 질의에 대한 답을 제공하지만 이용자의 질의에 대해 다른 이용자가 입력한 응답을 정답으로 제공한다는 점에서 기본적인 질의응답 시스템의 개념과 크게 다르다. 질문에 대한 정답을 사람이 제시하는 지식검색과 달리 질의응답 기술에서는 시스템이 이용자의 자연어 질의에 대해 많은 문서들로부터 자동으로 정확한 정답을 추출하여 제시하고자 한다. 단순히 키워드에 의한 검색과 후보 문서 집합 제공의 수준을 넘어서 문서들로부터 정답을 포함하는 단락을 검색하고 정답을 추출하는 과정을 포함한다.

질의응답 기술은 현재 정형 데이터베이스(formal database)에 담겨 있는 상품 정보나 기업 정보를 찾아주는 데이터베이스 질의응답과 게시판 혹은 전자메일 정보를 찾아주는 FAQ (Frequently Asked Question) 질의응답에 부분적으로 활용되고 있다(김현돈, 조성배 2000). 데이터베이스에서 이용자가 원하는 자료를 검색하는 데이터베이스 자연어 질의 처리 기술은 SQL 등의 정형 질의어(formal query)를 대신하여 자연어 질의문을 입력하고 질의 문장을 의미적 수준에서 파싱(parsing)하여 이를 직접 데이터베이스 정형 질의어로 변환하는 과정을 거친다(원정임, 윤지희, 이건배 1997). 그러나 고정밀 자연어 처리(deep natural language processiong)와 정교한 정보 추출(information extraction) 등의 텍스트 마이닝(text mining) 기술이 요구되는 백과사전 지식정보나 전자 매뉴얼(electronic manual) 정보에 대한 질의응답 기술은 현재 국내외적으로 연구가 진행 중이다(신승은, 이대연, 서영훈 2004; 이경순, 김재호, 최기선 2000; 황이규, 김현진, 장명길 2004; Lee G. et al. 2001).

질의응답 시스템이 정보검색 시스템과 크게 다른 점은 후보 문서를 검색하고 제공하는 것이 아니라 이용자가 궁극적으로 원하는 정답을 찾아 제공하는 것이다. 질의응답 시스템의 일반적인 처리 과정은 다음의 〈그림 1〉과 같이 구분해 볼 수 있다.

질의 분석 단계는 이용자 질문이 요구하는 정답의 종류를 구분하기 위한 질의 유형(question type)을 결정하고, 검색 엔진 가동을 위한 키워드를 추출한다. 문서 검색 단계는 정답을 포함하는 후보 문서를 선별하기 위해



〈그림 1〉 질의응답의 일반적 처리과정

일반적인 정보검색 과정과 비슷한 방법으로 키워드를 이용하는 검색 엔진을 구동한다. 단락 검색은 정답을 추출하기 위해 검색된 후보 문서에서 키워드와 관련이 있는 단락을 추출한다. 마지막 정답 추출 단계에서는 분류된 질의 유형과 대답 유형에 맞추어 일치되는 개체(entity)를 정답으로 추출하여 이용자에게 제시한다. 이와 같은 처리 과정 중에서 본 연구는 질의 분석 단계의 질의 유형 분류 과정에 초점을 두고 있다.

2.2 질의 유형 분류 기법

질의 유형 분류는 이용자의 질의 의도를 특정한 범주(category)에 할당하는 것으로서 질의응답 시스템 연구의 중요한 한 분야로 진행되어 왔다. 기존의 질의 유형 분류 기법들은 규칙에 기반한 방법(rule-based method)과 통계에 기반한 방법(statistical method)으로 크게 나누어 볼 수 있다.

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(Lee G. et al. 2001). 일반적으로 규칙에 기반한 접근 방법을 채택한 질의응답 시스템들은 유한 상태 오토마타로 구현된 과정을 통해 이용자의 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 있으며, 응용 영역이 정해져 있을 경우 간단한 튜닝(tuning) 과정으로 성능을 향상시킬 수 있다. 그러나 규칙을 수정하기 위해서 전문적인 지식을 가진 사람들의 노력이 필요하고, 규칙과 일치되지 않는 질의가 들어 왔을 때는 질의 유형을 분류할 수 없는 문제점을 가지고 있다. 또한 규칙이 많아질수록 좋은 성능을 내기 위한 튜닝이 더 어려워지게 되며, 시스템이 다른 응용 영역에서 사용될 경우 기존의 규칙들을 모두 수정하거나 많은 부분을 다시 작성해야 하는 문제점이 있다.

반면, 통계에 기반한 질의 유형 분류 방법은 수동으로 분류된 대량의 학습 데이터로부터 추출한 통계 정보를 이용한다. 질의 유형 분류를 위해 최대 엔트로피 모델(maximum entropy model)을 사용하거나 결정적 LR 파서를 확장한 예가 있다(Ittycheriah A. et al. 2000). 통계에 기반한 질의 유형 분류 방법은 대량의 학습 데이터를 이용한 통계 모델을 기반으로 하기 때문에 안정적으로 질의의 유형을 분류할 수 있으며, 자동화된 통계적 방법을 사용함으로써 시스템 구축을 쉽게 할 수 있다. 그러나

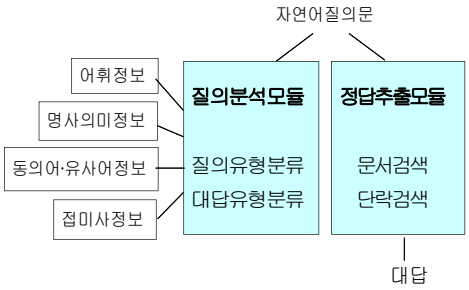
이용자가 질의에서 의도하지 않은 결과를 정답으로 추출하는 경우가 자주 있으며, 이러한 경우 보완이 쉽지 않고 대량의 학습 데이터를 이용하여 통계 정보를 추출해야 하는 어려움이 있다. 또는 이 두 가지 방법을 모두 적용한 하이브리드(hybrid) 방식의 연구 사례도 있다(김학수, 안영훈, 서정연 2003).

본 논문에서는 질의응답 시스템에서 이용자 질의의 유형을 분류하기 위해 복잡한 분류 규칙이나 대용량의 언어 자원을 이용하지 않고 질의문에 나타나는 단어의 의미 정보를 이용하는 방법을 제안하고 있다. 이용자의 자연어 질의 문장에서 의문사(interrogatives)와 같은 질의 초점 어휘들을 추출하고 구문 구조적으로 인접하여 나타나는 단어들의 의미 정보를 이용하여 세부적인 정답 유형을 결정할 수 있는 질의 유형 분류 방법이다. 또한 질의 초점 어휘로서 의문사가 생략된 경우에는 인접 출현하는 어휘들의 의미 정보를 이용하는 분류 처리 방법과 동의어 및 유의어 정보와 접미사 정보를 이용하여 질의 유형 분류 과정의 성능을 향상시킬 수 방법을 제안한다.

3. 단어 의미 정보 기반의 질의 유형 분류

질의응답 시스템은 다음의 〈그림 2〉에서 보이는 바와 같이 주어진 질의 유형을 분류하고 검색 키워드를 추출하는 '질의 분석 모듈'과 질의 키워드와 문서내의 단락을 비교하여 정답을 찾는 '정답 추출 모듈'로 구축되는 것이 일반적이다.

본 연구의 주제가 되는 질의 분석 모듈에서



〈그림 2〉 단어 정보와 질의 유형 분류

이용자 입력 질의의 초점이 무엇인지를 분석하는 과정은 이용자의 질의가 어떤 종류의 개체를 정답으로 요구하는가에 따라 정답을 찾는데 필요로 하는 질의의 특성을 분석한다. 이때 질의의 초점은 특정 범주의 개체로서, 예를 들어 '사람', '장소', '시간', '조직', '거리' 등의 범주가 된다.

질의 초점 어휘 역할을 하는 의문사와 함께 인접 단어들의 의미 정보를 기반으로 질의 유형을 결정하기 위해 구문 구조적으로 관련된 단어들을 판별해야 하며, 이를 위해 이용자의 자연어 질의 문장에 대한 얕은 수준의 개략적 구문 분석(shallow parsing)을 하고 그 결과를 이용한다.

질의 분석의 결과로 구한 질의 유형과 검색 키워드를 검색 문서의 단락과 비교하여 정답을 찾는 정답 추출 모듈은 질의 유형에 따라 해당하는 개체가 문서에 나타나고 키워드에 일치하는 단어들이 많은 단락에 높은 가중치를 주어 정답으로 추출한다.

3.1 질의 초점 어휘와 질의 유형 분류

3.1.1 질의 초점 어휘와 단어 의미 정보

질의 응답 시스템을 평가하는 TREC의

QA Track에서는 평가 셋(evaluation set)을 중심으로 의문사를 기준으로 한 질의의 분포를 제시한 바 있으며, TREC-9부터는 실제 질의 로그(query log)를 이용하여 질문을 수집하였다. 다음의 〈표 1〉은 TREC-8의 평가 셋에서 나타난 질의 의문사의 분포이다(Ellen M. Voorhees. 1999; Ellen M. Voorhees. Tice D. 2000; TREC).

〈표 1〉 TREC 평가 셋의 질의 분포

의문사	질의수	의문사	질의수	의문사	질의수	의문사	질의수
what	64	who	47	how	31	where	22
when	19	which	10	why	2	whom	1

TREC은 또한 질의응답 시스템에서 질의 유형을 분류하고 정답을 추출하기 위해 이용자 질의 로그로부터 계층적으로 세분된 의미 범주를 제시한 바 있다.

본 연구는 TREC에서 질의응답 시스템의 성능을 평가하기 위해 제시하는 이와 같은 척도들을 참조하여 한국어 자연어 질의 분석을 위해 질의 초점 어휘인 의문사의 종류와 그에 따른 질의 유형의 분류를 체계화하고자 하였다.

한국어의 의문문 형태로 나타나는 질의 문장의 많은 경우는 문장의 요소로서 질문의 초점을 나타내는 의문사(interrogative)를 가지고 있다. 각 질의마다 질의의 초점을 나타내는

〈표 2〉 질의 초점 어휘와 질의 유형의 예

어휘	질의 유형	질의 문장 예
누구 누가	Human (사람)	'... 우승자는 누구입니까(누구인가)?' '... 누가 우승하였는가?' '... 누구의 발명품인가?'
N 어디 어느 *N*	Location (장소)	'... 개최 장소는 어디입니까(어디인가)?' '... 어디에서 개최되었는가?' '... 어느 나라에서 개최되었는가?'
N 언제	Time (날짜나 시간)	'... 전시회는 언제 열리는가?' '... 언제 전시회가 열리는가?' '... 개막일은 언제인가'
N 얼마 얼마나 몇 *N*	Number (수)	'... 인구는 얼마나 많은가?' '... 강수량은 얼마인가?' '... 금메달은 몇 개인가?' '... 몇 시인가?'
어떤 *N* 어느 *N* *N* 어디	Organization (조직)	'... 어느 학교인가?' '... 우승팀은 어디인가?' '... 가장 많이 우승한 것은 어떤 팀인가?'
무엇 무슨 *N* 어떤 *N* 어느 *N*	Object (물체나 사물)	'... 물질은 무엇인가?' '... 무슨 곡인가?' '... 어떤 행성인가?' '... 어느 신문인가?'
왜	Why (이유)	'왜 ... 참석하였는가?' '... 왜 해체되었는가?'
어떻게	How (방법)	'... 어떻게 발생하였는가?' '어떻게 ... 분리되었는가?'
N (생략형)	(*N*에 따라 분류)	'... 사람은?' '... 장소는?' '... 년도는' '... 수는?' '... 우승자는?' '... 영화는'
	Unknown	

N : 구문적으로 인접한 단어의 의미정보를 이용함.

〈표 3〉 질의 유형과 하위 의미 범주

질의 유형	하위 의미 범주
Human (사람)	Artist, Politician, Economics, Sports ...
Location (장소)	Place, City, Continent, State, Capital ...
Time (날짜나 시간)	Year, Month, Day, Season, Hour...
Number (수)	Count, Price, Percent, Weight, Height ...
Organization (조직)	School, Company, Government, Team ...
Object (물체나 사물)	Planet, War, Religion, Organ ...
Unknown	

어휘로서 일차적으로 의문사로부터 질의 유형을 분류할 수 있다. 예를 들면 '언제'가 나타났을 때는 기본적으로 '시간'을 대답 유형으로 결정하고 인접해 나타나는 단어에 따라 하위분류로 대답 유형을 결정하도록 한다. 예를 들어 "대한민국의 수도는 어디인가?"에서 '어디'에 의해 '장소'를 질의 유형으로 정하고, '수도'에 의해 '도시'를 대답 유형으로 결정하는 방법이다. 〈표 2〉는 질의의 초점을 나타내는 의문사를 중심으로 질의 유형을 분류하는 예를 보여준다. 많은 경우에 의문사와 함께 문법적으로 직접 관련된 단어의 의미를 정보를 이용하여 질의 유형을 분류하게 된다. 예를 들어 '조직'과 '장소'를 정답 유형으로 구하는 질의가 동일한 초점 어휘인 '어디'와 함께 나타나는 경우도 있으며, 이러한 경우에는 의문사 '어디'의 앞에 나타나는 어절의 단어를 살펴서 질의 유형을 판단할 수 있다. 이와 비슷하게 이용자 질의에서 매우 빈번하게 사용되는 의문사 '무슨', '어느', '어떤' 등은 그 단어 자체로 정답 개체 종류를 판별할 수 없고 구문 관계를 맺고 있는 인접 단어의 의미 정보로부터 질의 유형 정보를 구해야 한다. 〈표 2〉의 분류에서 의문사 '무슨' 이나 '어떤'과 함께 표기된 *N*이 이러한 의미를 나타낸다.

따라서 질의 유형을 분류하고 정답 후보를 생성하기 위해서는 질의 유형에 따라 정답 개체가 될 단어들에 대한 의미 범주를 체계적으로 분류할 필요가 있다. TREC-10 및 TREC-11에 참가한 질의응답 시스템들을 참고하여 구축한 의미 범주 계층의 예를 〈표 3〉에서 보이고 있다. 질의 문장에 대하여 질의 유형을 분류하고, 각 질의 유형에 대해 다시 세분화된 하위 의미 범주를 분류하였다. 단어 의미 체계로서 세분화된 하위 의미 범주는 각 질의 유형에 대해 정답 유형을 결정하고 정답 후보를 선택하는데 결정적으로 사용된다.

3. 1. 2 동의어와 유의어 정보를 활용한 성능 향상

질의 유형에 대해 단어들의 하위 의미 범주를 분류하기 위해 단어 의미 사전을 구축하여 사용한다. 그러나 모든 단어에 대한 의미 정보를 사전의 독립된 엔트리로 구축하는 것은 매우 비효율적이므로 대용량의 사전을 구축하지 않고 동의어와 유의어 정보를 사용하는 방법과 접미사 정보를 이용하는 방법을 적용할 수 있다. 예를 들면, {작가, 글쓴이, 소설가, 문필가, 집필자, 문예가, 저작가, 제작자, 대문호, 저자, 지은이 ...}나 {장소, 곳, 데, 처소, 지점, 부분, 점, 위치, 지역 .. } 등은 이용자의 자연어 질의

에서 동일하거나 유사한 의미로 사용되는 단어들이다. 동의어와 유의어 사전은 정보검색 시스템을 위해서도 구축되어 검색 대상 어휘인 키워드에 대한 동의어와 유의어를 이용해서 검색어를 확장하는 다중 검색 방법에 효율적으로 활용된 예가 있다(윤성희, 장혜진 2004). 뿐만 아니라 질의응답 시스템의 정답 추출 과정을 비롯하여 자연어 처리 응용 시스템 전반에서 활용되는 주요 언어자원이다.

3. 1. 3 접미사 정보를 이용한 성능 향상

접미사는 접사의 하나로 낱말의 끝에 붙어서 의미를 첨가하여 다른 낱말을 만드는 역할을 한다. 단독으로는 사용될 수 없고 항상 다른 단어의 어근 뒤에 결합되어 여러 가지 의미를 첨가해 주는 역할을 한다. 자연어 질의로서 한국어 문장에서는 여러 가지 종류의 접미사가 붙어 다양한 형태의 새로운 단어로 나타나기 때문에 접미사 처리를 하지 않는다면 많은 접미사와 조합되는 많은 단어들이 미등록어로 처리되어 시스템의 성능을 저하시키는 원인이 된다. 예를 들어 "2004 년 올림픽은 어느 나라에서 개최되었는가?"라는 질의 문장에서는 '어느'와 '나라'에 의해 국가명을 정답 개체로 요구하는 '장소' 유형의 질의임을 판별할 수 있다. 반면에 "2004년 올림픽의 개최국은 어디인가?" "2004년 올림픽의 개최국은?"에서 '어디'와 함께 '개최국'이라는 질의 초점 의미 정보를 중심으로 대답 유형을 결정해야 하므로 접미사 '국'이 '국가'의 의미를 갖는다는 정보가 사용되어야 한다. '개최국' '생산국' '참가국' '동맹국' 등의 단어들을 모두 사전 등록하는 대신 접미사 '나라'의 의미를 갖는 '국'으로 분류된 정보를 이용하여 향상된 성능을 얻을 수 있다.

반면에 중의적인 뜻을 갖는 접미사는 질의 유형을 분류하기에 곤란한 경우가 많으므로, 이 문제에 대한 해결 방안으로서 중심 단어의 하위 정보를 이용하는 방법을 찾을 필요가 있다. 다음의 예들은 중의적(ambiguous)으로 사용되는 접미사의 예이다.

[예] 접미사 '장'의 중의적 사용 예
'운동장' '농구장' '각축장'
'위원장' '부회장'
'위촉장' '초대장' '임명장'
'고추장' '청국장'

[예] 접미사 "국"의 중의적 사용 예
'개최국' '생산국' '참가국' '동맹국'
'미역국' '북어국' '된장국' '토란국'

위와 같이 동의어와 유의어 정보, 또 접미사 정보를 질의 유형 분류에 이용하는 방법은 단어 의미 사전이 지나치게 많은 수의 엔트리를 포함하는 방대한 규모가 되지 않도록 체계화한다. 또한 기존의 등록된 단어들로부터 파생적으로 끊임없이 생성될 수 있는 단어들을 쉽게 등록하고 관리할 수 있어서 유연하고 능동적인 시스템이 될 수 있다.

3. 2 질의 초점 어휘가 생략된 경우

이용자 자연어 질의문의 유형을 분석하고 결정하는 데는 질의 초점 어휘로서 의문사가 중요한 역할을 함을 보였다. 그러나 한국어에서 문장의 일부 요소가 생략되는 현상이 빈번하게 나타나며, 질의에서도 대답 유형을 분류

할 수 있는 실마리인 의문사가 생략되는 경우도 매우 많다. 따라서 의문사 정보만을 가지고 질의 유형을 분석하는 데는 한계가 있다. 이러한 경우는 질의문의 마지막 어절에 위치하는 명사의 의미 정보를 취하여 질의 유형과 대답 유형을 결정할 수 있다. 생략형 질의에서 마지막 명사의 의미가 의미 계층구조에서 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 질의 유형을 '사람'으로 결정한다. 질의 문장에 나타나는 단어들의 의미와 이들의 출현 규칙을 살피고 추출한 단어들의 의미 정보를 담고 있는 사전을 이용하여 질의 유형을 분류할 수 있다. 또한 질의 유형에 대한 자세한 분류를 통해 정답 후보들의 적합성을 판단하고 정답을 추출하는데 중요한 역할을 한다. 예를 들어 "... 미국의 도시는?" "... 사람은?" "... 저자는?" 와 같이 의문의 초점이 생략된 질의의 경우에는 마지막 단어인 '도시', '사람', '저자' 등 단어의 상위 의미 범주에 의해서 '장소', '사람', '사람' 등으로 질의 유형을 분류할 수 있다. 질의 유형으로 두 개 이상이 나타나는 경우 각각의 경우에 대해서 대답 유형을 찾는다.

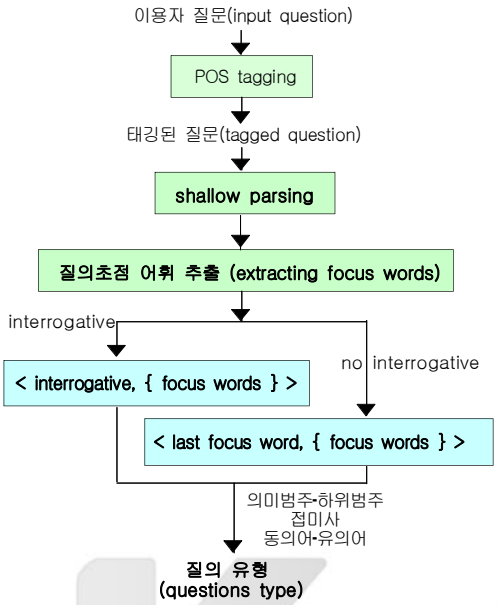
4. 실험 및 평가

4.1 실험 방법

질의응답 시스템을 평가하는 TREC의 QA Track 중 TREC-8에서는 약53 만 문서의 코퍼스로부터 수작업으로 작성한 200여 질문에 대해서 질의 유형을 분류하고 정답 추출 성능을 실험한 바가 있다. TREC-8에서는 질의에 대한 정답 추출 방법으로서 질의 유형에 따라 정답 유형에 일치하는 특정 개체를 추출하는 방법(50 바이트 focus-based answer) 와 정답을 포함한 단락을 제시하는 방법(200 바이트 bag-of-words approach)으로 구분하여 실험하였다. TREC-8에서 제시된 실제 질문의 예는 다음과 같으며, 각각 'who', 'when', 'what'에 의해 'person or organization' 'time' 'place-city' 등으로 질의 유형이 분류된다.

"Who is the prime minister of Japan?"
"When did the Jurassic Period end?"
"What is the capital of Kosovo?"
"Name the first private citizen to fly in space?"

TREC-8은 추출된 정답의 정확성을 중심으로 질의응답 시스템의 성능을 측정하지만, 본 연구의 실험에서는 단어 의미 정보에 기반한 질의 유형의 분류 방법에 의해 이용자의 자연어 질의가 얼마나 정확하게 분류되었는지에 대한 질의 유형 분류의 성능을 평가하고자 하였다. <그림 3>은 본 연구에서 단어 의미 정보를 중심으로 이용자 질의의 유형을 판별하는 과정을 흐름도(flow chart)로 나타낸 것이다. 실험을 위하여 기존의 정보 검색 시스템에서 키워드 검색에 매우 익숙한 대학생들을 대상으로 다량의 검색 대상 문서들을 제공하고 자연어 질의 문장을 수집하였다. 질의응답의 검색



〈그림 3〉 단어 의미정보 기반의 질의유형 분류 과정

대상 문서는 시사 뉴스, 스포츠 뉴스, 학술 내용 등을 담은 웹 문서들과 개인 학과 홈페이지 등이다. 〈표 3〉과 같은 방법으로 단어 의미 범주 및 하위 범주 체계를 적용하기 위해 태깅된(tagged) 문서의 단어들을 수집하여 실험용 의미 정보 사전으로 구축하였다.

실험 대상자들에게 일반 정보검색 시스템과 질의응답 시스템의 차이를 충분히 설명하였으며, 대상 문서들의 내용을 이해한 다음 정답 개체명이나 정답을 포함하는 문장 또는 정답 단락을 제공하는 가상의 질의응답 시스템에 대한 질의 문장을 자연어로 입력하도록 하였다. 약 2,200 자연어 질의 문장을 수집하여 우선적으로 질의 유형을 수동 분류하였다.

본 연구에서 제안하는 자동 분류의 성능을 실험하기 위해 사전 튜닝 과정을 통해서 이용자 질의에 나타나는 단어들 중 실험 사전에 등록되지 않은 단어들과 접미사, 동의어, 유의어에 해당하는 단어들을 등을 구분하여 등록하였다.

4. 2 결과 및 이용자 분석

이용자 질의 유형의 수동 분류 결과를 100%로 보았을 때 본 연구에서 소개한 단어 의미 정보를 이용한 질의 유형의 자동 분류 결과는 89.2%의 성공률 나타냈다. 〈표 4〉는 실험 질의에서 나타난 의문사 출현 빈도를 나타낸다.

〈표 4〉 실험에서 의문사 출현 빈도

의문사	출현(%)	의문사	출현(%)
누가/누구(의)	14.3	무엇	8.4
언제	9.7	얼마(나)	3.6
어디	8.5	몇	5.3
어느	6.3	어떻게	4.4
어떤	4.8	왜	4.1
무슨	4.1	(의문사생략	26.5

이용자 질의 중에서 의문사 생략된 형태의 질의가 26.5%를 차지해서 실험 대상자들이 기존 키워드 검색에 매우 익숙한 경향을 간접적으로 나타냈다. 의문사를 갖는 질의들도 많은 경우에 '무슨' '어떤' '어느' 등이 사용되어서 구문 구적으로 인접한 단어들의 의미 정보를 이용하여 질의 유형이 분류되었다. 구문적으로 인접한 단어의 의미 정보가 질의 유형 분류의 기반이 되므로 질의 문장의 구문 구조적 정보를 얕은 수준의 파싱(shallow parsing)을 통해 얻고 시스템의 계산 부담을 줄일 수 있도록 한다. 〈표 5〉는 단어 의미 정보를 기반으로 구축된 실험 시스템에서 성공적으로 자동 분류된

〈표 5〉 실험 결과의 질의유형 분포

질의유형	비율(%)	질의유형	비율(%)
사람	23.6	조직	6.8
장소	17.4	사물	12.2
시간	18.2	이유	7.3
수	8.4	방법	6.1

질의 유형의 분포를 보이고 있다.

TREC 2001 의 실험 결과로 보고된 바와 비슷하게 본 연구의 실험을 위해 수집된 이용자 질의의 21% 정도가 어떤 정의에 관련된 질의 문장들이었다. "What is Head Start?" "세마포어 연산이란 무엇입니까?"나 "6-시그마란 무엇인가?" 등과 같은 질의 문장으로서 질의문 "올해 노벨 문학상을 수상한 작가는 누구인가?" 등과는 달리 정답 추출의 난이도가 매우 높을 뿐만 아니라 질의 유형의 분류도 어렵다.

이용자들은 또한 정답으로서 단순히 검색된 문서와 문장의 일부분이나 특정 유형의 개체를 요구하는 것이 아니라 문서들의 내용 사이에서 매우 지능적이고 종합적인 추론이 필요한 질문을 던지는 경우가 많이 나타났다. 문서 내용으로부터 새로운 문장을 생성하여 이를 정답으로 제시하거나, 주어진 정답에 대한 배경 설명, 정답의 정당성 검증, 정답의 모호성 해결, 전문가 수준의 의견 제시, 이질적 정보의 통합을 통한 정답의 제시 등 매우 지능적 해결을 기대하는 질의들로서 "시맨틱 웹이 실현가능하려면 몇 년이나 걸릴까요?", "대북 관계는 앞으로 어떻게 진행될까?", "스크린쿼터에 대한 여론의 향방은?"과 같은 예들이다.

이와 같은 이용자들의 질의 경향은 비전문적 이용자들의 자연어 질의응답 시스템의 지능적 성능에 대한 기대가 매우 높다는 것을 보여준다.

5. 결 론

질의 유형 분류 과정이 시스템에 의해 정확하게 이루어진다면 질의응답 시스템의 정답 추출 성능을 크게 향상시킬 수 있다. 본 연구에서는 영어권의 언어들에 대한 대규모 언어 지식베이스 등의 풍부한 자원에 비해 상대적으로 부족한 언어 자원의 문제를 해결하기 위해 대량의 코퍼스를 이용하거나 복잡한 규칙을 작성하지 않고 단어들의 의미 정보 체계를 이용하여 질의 유형을 분류하는 방법을 제안하였다. 이용자의 질의 의도를 파악하기 위해서 질의 문장에서 질의 초점 어휘와 구문적으로 인접한 단어의 의미 정보를 기반으로 질의 유형과 정답 유형을 결정할 수 있는 방법이다. 질의에서 구조적으로 인접한 단어들의 의미 정보 체계를 이용하여 질의 유형을 하위 단계까지 분류하여 시스템에서 정답 후보 생성과 정답 추출을 위해 효과적으로 사용할 수 있다.

본 연구에 대한 실험을 통하여 복잡한 규칙이나 방대한 언어 자원, 또는 대량의 통계 정보 등을 이용하지 않고, 질의 분석 모듈의 처리 부담을 크게 줄이면서 만족할 만한 분류 결과를 얻을 수 있음을 확인하였다. 단어 의미 범주를 체계화한 의미 사전을 구축하고 동의어 사전, 유의어 사전, 접미사 정보 등을 실용적 수준으로 생성하는 과정이 뒤따라야 할 것이다. 또한 질의 유형의 하위 의미 분류를 보다 다양하고 폭넓게 적용하여 이용자 질의 문장에서 정답 유형 및 정답 개체의 종류를 더 구체적으로 제시할 수 있도록 하기 위한 연구가 계속될 필요가 있다.

참 고 문 헌

김수민, 백대호, 김상범, 임해창. 2000. 시소러스 범주정보를 이용한 질의응답 시스템. 『한글 및 한국어 정보처리 학술대회』.

김영택. 2001. 자연언어처리. 『생능출판사』.

김학수, 안영훈, 서정연. 2003. 하이브리드 방법의 사용자 질의 의도 분류. 『한국정보과학회 논문지-소프트웨어 및 응용』, 30(1): 51-57

김현돈, 조성배. 2000. 한메일넷 질의 자동응답을 위한 이단계 자기구성 지도. 『정보과학회 춘계학술대회』.

박세영, 강현규. 1998. 한글공학: 정보검색. 『한국정보처리학회지』, 특집, 5(5).

박소연, 이준호. 2002. 로그 분석을 통한 이용자의 웹 문서 검색 행태에 관한 연구. 『정보관리학회지』, 19(3): 111-122.

신승은, 이대연, 서영훈. 2004. 구문관계 정보를 이용한 한국어 질의응답 시스템. 『한국콘텐츠학회 논문집』, 4(2).

양수정, 서영훈. 2003. 질의문의 구문정보를 이용한 키워드 추출. 『한국콘텐츠학회 추계 종합 학술대회』, 1(2).

윤성희, 장혜진. 2004. 자연어 질의 분석과 검색어 확장에 기반한 웹 정보 검색. 『정보관리학회지』, 21(2): 235-248.

원정임, 윤지희, 이건배. 1997. 유사객체 검색에 의한 협력 질의 응답. 『한국정보처리학회 추계 학술대회』.

이경순, 김재호, 최기선 2000. KorQuA: 질의응답에서 자료 유형을 고려한 대답 검색과 대답 해석. 한글 및 한국어 정보처리 학술대회.

이재홍, 최호섭, 옥철영 2003. 개념어의 습득을 위한 지식기반 질의응답시스템. 『제15회 한글 및 한국어 정보처리 학술대회』.

황이규, 김현진, 장명길. 2004. 질의응답 기술 개발. 『정보처리학회지』, 11(2): 48-56.

AAAI Fall Symposium on Question Answering. 〈http://www.aaai.org/Press/Reports/Symposia/Fall〉

Baeza Yates Ricardo. and Reberio Neto Berthier. 1999. *Modern Information Retrieval*. Addison Wesley.

Burger J. and Cardie C. 2001. Issues, Tracks and Program Structures to Roadmap Research in Question & Answering(Q&A). 〈http://www-nlpir.nist.gov/projects/duc/pubs.html〉

Edward H, Hermjakov U, Lin CY, Ravichandran D. 2002. Using Knowledge to Facilitate Factoid Answer Pinpointing. Coling 2002.

Ellen M. Voorhees. 1999. The TREC-8 Question Answering Track Report. 〈http://trec.nist.gov/pubs/trec__8/papers/qa__report.pdf〉

Ellen M. Voorhees. Tice D. 2000. Building a Question Answering Test Collecion. Proceedings of SIGIR 2000. 200-207.

Ittycheriah A. Franz M. etc. 2000. IBM's Statistical Question Answering System. TREC-9. 229-234.

Jimmy L. 2002. The Web as Resource for Question Answering. LREC 2002.

Jimmy L. and Boris Katz. 2003. Question Answering Techniques for the World Wide Web. 10th Conference of the European Chapter of the Association for Computational Linguistics(EACL-2003).

Lee G. Lee S. etc. 2001. Site/Q: Engineering high performance QA system using Lexico-semantic pattern matching and shallow NLP. TREC-10. 437-446.

Moldovan D. Adrian N. 2002. Lexical Chain for Question Answering. Coling 2002.

TREC(Text Retrieval Conference). 〈http://trec.nist.gov/pubs.html〉