자연어 질의 유형판별과 응답 추출을 위한 어휘 의미체계에 관한 연구

윤 성 희
상명대학교 컴퓨터소프트웨어공학전공
shyoon@smu.ac.kr

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

Sung-Hee Yoon
Dept. of Computer Software Engineering, Sangmyung University

요 약

질의응답 시스템이 정보검색 시스템과 다른 중요한 점은 질의 처리 과정이며, 자연어 질의 문장에서 사용자의 질의 의도를 파악하여 질의 유형을 분류하는 것이다. 본 논문에서는 질의 유형을 분류하기 위해 복잡한 분류 규칙이나 대용량의 사전 정보를 이용하지 않고 질의 문장에서 의문사에 해당하는 어휘들을 추출하고 주변에 나타나는 명사들의 의미 정보를 이용하여 세부적인 정답 유형을 결정할 수 있는 질의 유형 분류 방법을 제안한다. 의문사가 생략된 경우의 처리 방법과 동의어 정보와 접미사 정보를 이용하여 질의 유형 분류 성능을 향상시킬 수 있는 방법을 제안한다.

1. 서 론

일반적으로 정보검색 시스템(information retrieval system)은 사용자의 질의에 대해 정보가 포함되어 있을 가능성이 높은 문서들을 찾아주는 시스템이다. 이러한 정보검색 시스템에서는 일반적으로 사용자들이 다시 한번 검색된 문서들로부터 정답을 찾아야 하는 불편함이 있다. 그러나 많은 사용자들은 명확한 의도를 가지고 질문을 하며, 정보 검색 시스템이 대량의 문서를 찾아주기 보다는 정답들을 곧바로 찾아 제시해 주기를 원한다. 이러한 요구를 만족시키기 위한 질의 응답 시스템(question-answering system)은 질의에 대한 결과로 문서를 제시해 주는 것이 아니라 사용자가 원하는 질문에 대해 정답을 추출해서 문장이나 단락으로 제시해 주기 때문에 사용자의 부담을 줄여주는 지능적이고 편리한 시스템이다.

질의응답 시스템이 정보검색 시스템과 다른 중요한 점 중 하나는 질의 처리 과정으로서 질의에서 사용자의 질의 의도를 파악할 수 있는 질의 유형(question type)이나 키워드(keyword) 등의 정보를 출하는 것이다. 특히 질의 유형의 분석 과정은 질의응답 시스템이 문서에서 정답이 될 수 있는 정답 후보(answer candidate)들을 추출하는데 중요한 정보를 제공한다. 사용자의 질의 의도를 정확히 파악하고 정답으로서 구하는 것이 무엇인지 파악하는 과정을 질의 유형 분류라고 한다. 국제적인 정보검색평가대회인 TREC에서는 1999년의 TREC-8에서 질의응답시스템의 평가를 시작하였다. 최근 TREC에서 소개된 질의응답 시스템들은 대부분 질의 유형 분류(question type classification)를 위한 모듈을 포함하고 있다[10].

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 규칙을 작성하지 않

고 질의 유형을 분류하는 방법을 제안한다. 각 질의의 유형마다 질의의 초점을 나타내는 어휘가 존재한다면 어휘 정보만 이용해서 질의 유형을 분류할 수 있도록 설계한다. 명사 의미 정보 사전을 구축하고 동의어 사전, 유의어 사전, 접미사 정보 등을 이용하여 질의 유형 및 정답 유형을 분류할 수 있다.

2. 관련 연구들

질의 유형 분류는 사용자의 질의 의도를 특정한 범주(category)에 할당하는 것으로 질의응답 시스템 연구의 한 분야로 진행되어 왔다[7,8].

규칙에 기반한 질의 유형 분류를 채택하고 있는 시스템들은 일반적으로 어휘-구문 패턴을 구축하고, 이러한 패턴을 유한 상태 오토마타와 매치(match)하여 질의 유형을 분류한다. 규칙을 수정하기 위해서 전문적이 지식을 가진 사람들의 노력이 필요하고, 규칙과 일치되지 않는 질의가 들어 왔을 때는 질의 유형을 분류할 수 없고, 규칙이 많아질수록 튜닝이 더 어려워지게 되며, 시스템이 다른 응용 영역에서 사용될 경우에는 기존의 규칙들을 모두 수정하거나 많은 부분을 다시 작성해야 하는 문제점이 있다.

반면, 통계적인 방법에 기반한 질의 유형 분류는 수동으로 분류된 대량의 학습 데이터로부터 추출한 통계 정보를 이용한다. 대량의 학습 데이터를 이용한 통계 모델을 사용하기 때문에 안정적으로 질의의 유형을 분류할 수 있으며, 자동화된 통계적 방법을 사용하여 시스템 구축을 쉽게 할 수 있다. 그러나 사용자가 질의에서 의도하지 않은 결과를 정답으로 추출하는 경우가 자주 있으며, 대량의 학습 데이터를 이용하여 추출해야 하는 어려움이 있다.

3. 어휘 의미 정보를 이용하는 질의 유형 분류

질의분석 모듈은 주어진 질의의 초점이 무엇인지를 분석하는 모듈이다. 이때 초점은 주로 개체가 대부분인데, 예를 들어 '사람', '조직', '장소', '거리', '시간' 등이다. 질의 분석의 결과와 검색 문서의 단락을 비교하여 적합한 답을 찾는 모듈은 질의 유형에 따라 해당하는 개체가 문서에 나타나고, 질의에 해당하는 단어들이 많은 단락에 높은 가중치를 주어 정답으로 추출한다. 기존 연구에서는 질의 분석 모듈은 대부분 패턴 매칭이나 부분 구문 분석을 통하여 해당 질의 유형을 결정하고, 질의에 해당하는 단락을 찾는 모듈에 대하여 서로 다른 방법론을 제시하는 것이 일반적이었다[4,6].

3.1 질의 유형 어휘

3.1.1 어휘 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. 예를 들어 "대한민국의 수도는 어디인가?"에서 '어디'에 의해 '장소'를 질의 유형으로 정하고, '수도'에 의해 '지명'을 대답 유형으로 결정하는 방법이다. 그림 1은 어휘정보를 이용하는 질의응답 시스템의 구성을 보여준다. 표 1은 질의의 초점을 나타내는 어휘를 중심으로 질의 유형을 분류하는 일부 예를 보여준다. 표의 오른쪽에서 '+의 의미정보'로 표시된 내용에 해당한다.

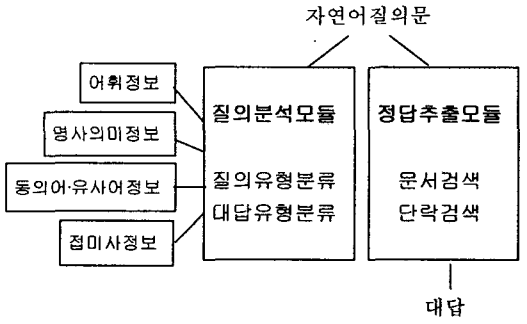

그림 1. 질의응답 시스템의 질의유형 분류

어휘예	질의 유형	질의 문장 예	
누구 누가	Human (사람)	'... 우승자는 누구입니까(누구인가)?' '... 누가 우승하였는가?'	
+ 어디 어느 +	Location (장소)	'... 개최 장소는 어디입니까(어디인가)?' '... 어느 나라에서 개최되었는가?'	+의 의미정보
+ 언제	Time (날짜나시간)	'... 전시회는 언제 열리는가?' '... 개막일은 언제인가"	+의 의미정보
+ 얼마 얼마나 몇 +	Number (수)	'... 강수량은 얼마인가?' '... 몇 시인가?'	+의 의미정보

표 1. 어휘정보와 질의유형의 예

3.1.2 명사 의미 정보 이용

정답의 유형을 찾고 정답 후보를 생성하기 위해서는 질의 유형에 대한 하위 의미 범주를 분류할 필요

가 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 세분화된 하위 의미 범주는 정답 유형을 결정하고 정답 후보를 결정하는데 유용하게 사용될 수 있다. 표 2는 질의 유형 분류의 일부와 어휘들의 하위범주 일부 예를 보여준다.

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

표 2. 질의 유형과 하위 의미 범주

3.1.3 동의어 및 유의어 사전을 활용한 성능 향상

질의 유형에 대해 하위 의미 범주로 분류하기 위해 명사 의미 사전을 이용하였다. 그러나 모든 명사에 대한 의미정보를 사전으로 구축하는 것은 불가능하다. 질의 문장에서 다양한 형태의 출현 가능한 명사들을 분류하기 위해서 대용량의 사전을 구축하지 않고 동의어와 유의어 정보를 사용하는 방법과 접미사 정보를 이용하는 방법을 제안한다.

[예] 동의어 및 유의어
작가 = { 글쓴이, 소설가, 수필가, 집필자, 문예가,
저작가, 제작자, 대문호, 저자, 지은이 ... }
장소 = { 곳, 데, 처소, 지점, 부분, 점, 위치, 지역 ... }
나라 = { 국가, 사직 ... }

3.1.4 접미사 정보를 이용한 성능 향상

접미사는 접사의 하나로 낱말의 끝에 붙어서 의미를 첨가하여 다른 낱말을 이르는 말이다. 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 중의적인 뜻을 갖는 접미사는 질의 유형을 분류하기에 곤란한 경우가 많으므로, 이 문제에 대한 해결 방안을 강구할 필요가 있다. 다음의 예들은 중의적으로 사용되는 접미사의 예이다.

[예] 접미사 "장"의 중의적 사용 예
"운동장" "농구장" "각축장"
"위원장" "위촉장" "초대장" "고추장"

3.2 질의 유형 어휘가 없을 때

질의에 대한 대답 유형을 알 수 있는 실마리가 질의 문장에 나타나지 않을 경우는 한국어에서 나타나는 생략현상으로 인한 경우이다. 따라서 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. 생략형 질의에서 마지막 명사의 의미가 의미 계층구조에서 대답 유형의 하위개념인지에 따라 결정한다. 대답유형으로 정의된 것에 속하지 않을 때는 해당 의미 계층정보를 대답 유형으로 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 "... 미국의 도시는?" "... 사람은?" "... 저자는?" 와 같이 의문의 초점이 생략된 질의의 경우에는 마지막 어휘인 '도시', '사람', '저자' 등의 하위 범주에 의해 질의 유형을 분류할 수 있다. 그림 2는 질의 유형을 분류하는 과정을 보여준다.

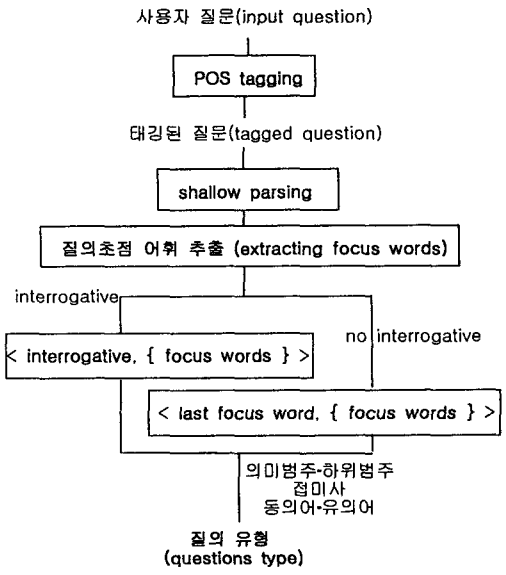

그림 2 질의유형 분류 과정

4. 실험과 평가

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. 약 2,400 문장 자연어 질의 문장을 어휘의 의미정보를 이용하여 분류하는 실험에서 수동 분류 결과와 비교할 때 약 89%의 분류 성공률을 보였다. 의미 정보가 등록되지 않은 경우에 질의 유형이 분류되지 못한 질의 문장이 많았다.

사용자들이 입력한 일부의 질의문들은 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 질문들로서 이질적 정보의 통합을 통한 정답의 제시 등의 지능적 해결을 기대하는 질의들이었다. 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?"과 같은 어떤 정의(definition)에 관련된 질의로서 비전문적 사용자들의 자연어 질의 응답 성능에 대한 기대가 매우 높다는 것을 보여준다.

5. 결론

본 논문에서는 질의응답 시스템의 성능을 향상시키기 위한 필수 조건으로서 사용자의 질의 문장을 분류하는 방법을 제안하였다. 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. 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.

참고 문헌

[1] 김수민, 백대호, 김상범, 임해창, "시소러스 범주정보를 이용한 질의응답 시스템", 한글 및 한국어 정보처리 학술대회 , 2000.

[2] 김학수, 안영훈, 서정연, "한국어 질의응답 시스템을 위한 지지벡터기계 기반의 질의 유형 분류기", 한국정보과학회논문지-소프트웨어 및 응용, 제 30권, 제 5호, 2003

[3] 신승은, 이대연, 서영훈, "구문관계 정보를 이용한 한국어 질의-응답 시스템", 한국콘텐츠학회 논문집, 제 4권, 제 2호, 2004.

[4] 양수정, 서영훈, "질의문의 구문정보를 이용한 키워드 추출", 한국콘텐츠학회 2003 추계 종합 학술대회 논문집, 1권, 2호. 2003.

[5] 이경순, 김재호, 최기선, "질의응답 시스템의 평가를 위한 태스트컬렉션 구축", 한글 및 한국어 정보처리학회, 2000.

[6] 이대연, 서영훈, "구문구조를 이용하여 정답을 추출하는 질의 응답 시스템", 제 15회 한글 및 한국어 정보처리 학술대회, 2003.

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

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

[9] Moldovan Dm Adrian N, "Lexical Chain for Question Answering," Coling 2002.

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