문장패턴을 이용한 자연어 질의 시스템에 대한 연구

청주대학교 컴퓨터정보공학과
우근신 · 송재관 · 홍성웅 · 연제용 · 박찬곤

A Study on the Natural Language Query System Using Sentence-Pattern

Keunsin Woo, Jaegwan Song, Sungwoong Hong, Cheyong Yon, Changun Park
Department of Computer and Information Engineering Chongju University, Chongju, Korea

요 약

질의응답 시스템은 인터넷과 같은 실용적 환경에서 사용될 경우, 실제 사용자의 질의는 다양한 유형으로 나타나게 된다. 따라서 실용적인 시스템에서 사용되는 질의는 문장의 형태나 단어의 쓰임에 관계없이 같은 의도를 가진 질의를 같은 유형으로 분류할 수 있는 의문형 문장패턴을 태깅하여 다양한 형태의 자연어로 기술된 문서에서 원하는 응답으로 처리할 수 있는 질의 응답 시스템은 정보 검색 시스템으로서의 가능성을 보여준다.

서 론

자연어 처리는 인간이 습득하여 사용하는 언어적 지식을 컴퓨터에 이식하여 인간이 언어를 사용함으로써 얻는 많은 이점을 컴퓨터에서 얻으려는 연구이다. 자연어처리에 있어서 가장 핵심적인 문제는 입력된 문장을 분석하는 것으로 이 분석된 문장을 중심으로 질의-응답 시스템에서는 질의에 관련된 응답을 생성하는 것이다. 기존의 정보검색(information retrieval, IR)은 사용자의 질문에 대한 응답으로 대량의 문서를 검색하고 순위화하는데 초점을 맞추어 왔다. 그러나, 많은 사용자들은 명확한 의도를 가지고 질문을 하며, 정답을 곧바로 찾아 제시해 주기를 바란다.

이러한 요구를 만족시키기 위하여 질의응답(question answering, QA)이라는 개념이 출현했으며, 많은 연구들이 AAAI와 TREC[1]을 중심으로 수행되어 왔다.

질의응답 시스템이 정보 검색 시스템과 다른 점 중 하나는 질의 처리 과정(question answering, QA)에 있다.

질의 처리 과정은 질의에서 사용자의 질의 의도를 파악 할 수 있는 질의 유형(question type)이나 키워드(keyword) 등의 정보를 질의로부터 추출하는 것이다. 특히 질의 유형은 질의응답 시스템이 문서에서 정답이 될 수 있는 정답 후보(answer candidate)들을 추출하는데 중요한 정보를 제공한다. 최근 TREC(Text REtrieval Contest)에서 소개된 질의응답 시스템들은 대부분 질의 유형 분류(question type classification)를 위한 모듈을 포함하고 있다.[1)]

질의응답 시스템이 인터넷과 같은 실용적 환경에서 사용될 경우, 실제 사용자의 질의는 다양한 유형으로 나타나게 된다. 따라서 실용적인 시스템에서 사용되는 질의는 문장의 형태나 단어의 쓰임에 관계없이 같은 의도를 가진 질의를 같은 유형으로 분류해 낼 수 있어야 한다. 예를 들어 "올해의 프로골프 우승자는?"과 "누가 프로골프에서 승리했죠?"는 같은 의도를 가진 질의다. 이와 같은 질의에서 만족한 응답을 얻기 위해 의문형 문장패턴은 다양한 형태의 텍스트에서 원하는 응답으로 처리할 수 있고 응용 영역의 변화에 유연하게 대처할 수 있다.

관련연구

이 장에서는 자연어 질의 응답 시스템에 대한 관련 연구들을 통하여 질의 응답 시스템의 유형을 나누어 보고자한다. HQL[5)]은 구현 방법으로 parse tree를 구성하기까지의 상부 구조의 설계와 기존 데이터베이스 관리 시스템인 INGRES의 하부 모듈을 이용하는 방법을 택하였고 한글 질의어 시스템에서 처리하는 한글 데이터의 내부표현 상태를 통일화함으로써 데이터 처리의 효율을 기하였다. HCQ[6)]는 문장 단위의 질의 실행 방법과 대화식 질의 방법이 갖는 단점을

E-mail : jmwoo@chongju.ac.kr

제거하고자 메뉴 선택에 의한 질의로 개발 환경의 하드웨어 제약점들을 극복할 수 있고 시스템과 사용자의 대화 회수를 최소한으로 할 수 있도록 HCQ를 설계하였다.

NHI[7]는 Colmerauer에 의해 1차 프레디키드 논리(first-order predicate logic)의 clause로 문법을 나타내는 DCG (Definitive Clause Graar)를 이용하여 한글 질의어의 문법을 정의하고 관계 데이터 모델에 기초한 데이터베이스에 대한 자연 한글 질의문을 정형 질의문으로 변환시키는 자연 어 인터페이스를 설계 및 구현하였다. QUIK[8]은 지식에 기초한 시스템으로, 데이터베이스의 구조 의미를 표현하는 방법으로서 aggregation을 사용하였으며 자연어의 의미 구조를 표현하기 위하여 용언 중심의 격체계를 사용하였다. 이 격 체계와 aggregation을 연결짓기 위해 '관계 그래프'를 제시하였다. FQAS[9]는 fuzzy정보를 처리하는 자연어 질의 응답 시스템을 구현하고자 하여 블랙보드 개념을 적용하였다. FQAS는 자연어로 표현된 지식을 처리하고자 서술적인 지식을 표현하기에 적합한 프로덕션 시스템을 사용하였다. K-NLQ[10]는 NHI가 새로운 분석기를 활용할 수 없다는 단점을 보완한 것으로 다양한 질의를 처리하기 위해 입력 질의문이 유형을 분류하였으며, 부정(negation)이나 등위 접속사(and), 집단 함수(aggregation function)가 포함된 문장에 대한 처리를 가능하게 하였다. KID[11]는 한국어 질의를 객체 지향 데이터 모델에서 사용되는 질의 그래프로 변환해주는 객체 지향 데이터베이스를 위한 자연어 인터페이스의 설계에 관해 기술하고, 객체 지향 질의 모델에서 나타나는 경로식의 자연어 표현을 처리하기 위해 채택한 프레임 기반 기법을 기술하였다. 채진석[11]은 한국어 질의를 OQL 명령문으로 변환하는 한국어 질의 처리기와 OQL 명령문으로 변환하는 한국어 질의 처리기와 OQL 처리기, 저장 관리기, 스키마 관리기 등으로 이루어진 객체 지향 데이터베이스를 위한 한국어 질의 시스템의 설계 및 구현을 하였다. 한국어 질의로부터 OQL 명령문으로의 변환에 필요한 중간 단계의 지식을 표현하기 위해 프레임을 사용하였다.

위의 연구들을 통하여 아래와 같이 질의 응답 시스템을 분류하여 볼 수 있다.

- 데이터 모델에 따라 관계 데이터 모델 기반 시스템, 객체 지향 데이터 모델 기반 시스템
- 질의 방법에 따라 문장 단위의 질의 기반 시스템, 대화식 질의 기반 시스템, 메뉴 선택 질의 기반 시스템
- 질의어 분석 단계에 적용하는 분석기에 따라 자연어 처리에서의 분석기 기반 시스템, 응용에 적합한 지식베이스를 이용한 분석기 기반 시스템 그 밖에 키워드에 template matching 기법을 이용한 방법, 제한된 영역에 관한 지식을 시스템 내의 procedure로 표시하는 방법, 지식 기반 시스템 등이 있다.

질의어 문장패턴 추출

1. 문장패턴

일상생활 가운데서 서로의 생각을 교환하기 위해 문장이라는 기본 단위를 이용한다. 실제 언어 행위에서 사용되는 문장은 다양하다. 그러나 어떤 문장이든지 모두 일정한 문법적 규칙에 따라 이루어지므로 구조적 형식에 있어서 모든 문장에 공통되는 일정한 틀을 갖고 제한된 수의 유형으로 나뉠 수 있다. 이러한 제한된 수의 유형들은 개개의 구체적인 문장을 대표한다.

한국어 문장은 서술어에 따라 문장의 성분이 결정되어지는 특징이 있다. 말뭉치 구축을 통하여 한국어 문장에서 나타나는 같은 종류의 문장들을 대표할 수 있는 틀을 '문장패턴'이라 한다.[4]

한국에서의 문장패턴 연구는 미국이나 일본에서의 문장패턴 연구보다 뒤늦게 시작되었으며 또 뒤떨어져있다. 미국이나 일본의 경우에서 보면 미국에서는 1920년대 초부터, 일본에서는 1940년대부터 문장패턴에 대한 연구가 본격적으로 진행되기 시작하여 이미 수십 편의 논문과 저서들을 통하여 그 연구성과들이 발표되고 있다. 그런데 한국어의 경우에서는 문장패턴이 문장론의 한 독자적인 분과로 등장되어 본격적으로 연구되기 시작한 것은 1960년대 후반기부터이다.

한국어에서 '문장패턴'이란 용어는 다른 용어로 사용되고 있었다. 최현배는 고등말본(1959)에서 문장패턴을 '월골'이라고 불렀으며, 또 김민수, 이기문도 인문계고등학교 표준문법(1958)에서 문장이 기본이 되는 틀을 '문형'이라 하였다.[13] 이렇게 한국어에서도 벌써 60년대 이전 시기에 문장패턴에 대한 논의가 있었던 것만은 사실이나 모두가 문법서의 서술에서 문장성분을 분류하고 기술하는 한 방편으로 제시한 것에 지나지 않았다. 한국어에서의 문장패턴 연구는 다른 언어의 문장패턴 연구에 비해 뒤늦게 시작되었으며 연구성과도 크지 못하다. 문장패턴의 설정에 대한 논문이나 저서는 많지 않으며, 그 연구가 주로 국어학적인 입장에서 기술된 것이어서 전산학적인 측면에서의 재조명이 필요하다.

2. 질의어 문장패턴 추출을 위한 문장 패턴분석

한국어의 문장패턴은 3형, 4형, 5형, 6형, 7형, 12형, 41

Table 1. 조사 분류표

대표조사	코 드	조사그룹
가	1	이, 가, 께서, 에서, 서
을	2	르, 을, 를
에	3	에, 에게, 한테, 께, 더러, 보고
와	4	과, 하고
로	5	로, 에게로, 으로
보다	6	과/와, 처럼, 만큼, 보다, 하고
에서	7	으로부터, 로부터, 서
를 위해	8	
에 의해	9	
라고	10	
에 대해	11	

Table 2. 명사 의미소성 분류표

소성	코 드	의 미
Abs	NA	추상적인 명사
Act	NB	행 위
Ani	NI	동 물
Con	NC	구 상 체
Div	ND	종 류
Hum	NH	인 간
Loc	NL	장 소
Num	NN	수 량
Mat	NM	물 질
Temp	NT	시 간

형 등 설정된 수가 학자에 따라서 각각 다르게 나타나고 있다.[2] 한국어의 문장패턴 설정 기준을 살펴보면 첫째, 서로 다른 문장성분의 부동한 배합방식을 전면적으로 고려해야한다는 견해와, 둘째, 근간성분만 고려하되 그 중에서도 서술어를 중심으로 고려해야 한다는 견해로 나누어 볼 수 있다. 본 장에서는 말뭉치로부터 추출된 32형의 문장패턴에 명사의 의미소성을 부여하여 추출함으로써 의문질의어 문장패턴의 수가 213형으로 늘어났다.

추출된 문장패턴은 편의상 서술어에 따라 동사형, 형용사형, 지정사형으로 분류하였다.

한국어말뭉치로부터 문장패턴을 추출하기 위하여 말뭉치 태깅에 사용된 기호는 Table 1, 2와 같다.

Table 1로부터 추출된 문장패턴의 유형은 Table 3과 같다.

Table 2로부터 추출된 문장패턴의 유형은 Table 4와 같다.

이 논문에서 사용된 기본 문장패턴 예문은 아래와 같다

누가 우니? N1+V

누가 서울에 도착했나? N1+N3+V

Table 3. 추출된 문장패턴

종 류	유 형
동사형	19형 문장패턴
형용사형	8형 문장패턴
지정사형	5형 문장패턴
합 계	32형 문장패턴

Table 4. 의문질의어 문장패턴

종 류	유 형
동사형	153형 문장패턴
형용사형	41형 문장패턴
지정사형	19형 문장패턴
합 계	213형 문장패턴

누가 회장으로 선출되었나? N1+N5+V

누가 영희와 짝꿍이냐?N1+N4+V, N1+N4+NA

3. 의문질의어 문장패턴 추출

의문질의어 패턴을 추출하기 위하여 말뭉치로부터 추출된 32형의 문장패턴에 명사의 의미소성을 부여하여 자연어 질의어를 태깅함으로써 Table 5-7과 같이 의문질의어 문장패턴을 추출하였다.

<문장패턴> NH1+NC1+V
<예문> (철이는 숙제공부에 열중했다.)
<질의어패턴> WNH1+NC1+V
<예문> (누가 숙제공부에 열중했느냐)
NH1+WNC1+V
(철이는 무엇에 열중했느냐)
<문장패턴> NH1+NH4+NA3+V
<예문> (영희는 철수와 같은반이 되었다.)
<질의어패턴> WNH1+NH4+NA3+V
<예문> (누가 철수와 같은반이 되었느냐)
NH1+WNH4+NA3+V
(영희는 누구와 같은반이 되었느냐)
NH1+NH4+WNA3+V
(영희는 철수와 어떤반이 되었느냐)
<문장패턴> NH1+NC2+NC3+V
<예문> (철수가 밭을 논으로 만들었다.)
<질의어패턴> WNH1+NC2+NC3+V
<예문> (누가 밭을 논으로 만들었느냐)
NH1+WNC2+NC3+V
(철수가 무었을 논으로 만들었느냐)
NH1+NC2+WNC3+V

(철수가 밭을 무엇으로 만들었느냐)

<문장패턴> NH1+NA9+NC2+V
<예문> (철수가 은행에서 돈을 찾았다.)
<질의어패턴> WNH1+NA9+NC2+V
<예문> (누가 은행에서 돈을 찾았느냐)
NH1+WNA9+NC2+V
(철수가 어디에서 돈을 찾았느냐)
NH1+NA9+WNC2+V
(철수가 은행에서 무엇을 찾았느냐)

<문장패턴> NH1+NH4+NA+NB2+V
<예문> (철수가 교수님과 언어의 기원에 대해 견해를 나누었다.)
<질의어패턴> WNH1+NH4+NA+NB2+V
<예문> (누가 교수님과 언어의 기원에 대해 견해를 나누었느냐)
NH1+WNH4+NA+NB2+V
(철수가 누구와 언어의 기원에 대해 견해를 나누었느냐.)
NH1+NH4+WNA+NB2+V
(철수가 교수님과 무엇에 대하여 견해를 나누었느냐)
NH1+NH4+NA+WNB2+V
(철수가 교수님과 언어의 기원에 대해 무엇을 했느냐)

질의응답 시스템 구성

인간의 생각을 표현하는 요소 중 가장 구체적인 방식이 자연언어이다. 이러한 자연어의 구조를 밝히는 연구는 인간의 사고방식과 그 과정에 대한 연구이다. 계산기의 등장으로 인해 자연언어는 공학적인 입장에서 처리되기 시작하여 대화형 시스템에 연결되어 개발되고 데이터베이스는 이러한 연구들의 대상으로 적합한 시스템이며, 질의응답 시스템은 질의문을 이용해 얻고 싶은 데이터를 얻고, 이를 활용할 수 있는 시스템으로서 인간이 자신의 생각을 표현하고 자신의 지식을 전달하는 시스템과 흡사한 시스템이다.

자연어 질의 시스템의 분석 과정은 크게 자연어 질의문 분석과 정형 질의문으로의 변환과정으로 나눌 수 있다. 자연어 질의문 분석 과정은 자연어처리에서의 분석과정과 같은 방식으로 이루어지게 되며, 그 과정으로 제 1 단계인 형태소 분석은 주어진 문을 형태소 · 말의 나열로 분해하고, 각각의 형태소 · 말의 품사 등을 결정한다. 형태소나 단어의 정의는 언어의 종류에 따라 달라지므로 이 논문에서의 형태소 분석기는 질의어문장 패턴사전과 어휘사전을 사용하여 형태소 분석기에서 질의 문장패턴을 분석하게된다.

응답후보생성기는 일반구문이나 문서로부터 얻은 정답이 될 수 있는 정답후보 문장패턴을 문장패턴 적용기를 이

Table 5. 동사형 질의어패턴

기본 구성	패턴 구성	수
N1+V	WNI1+V WNH1+V :	14
N1+N3+V	WNH1+NH3+V NH1+WNL3+V :	72
N1+N5+V	WNH1+NL5+V NC1+WNC5+V :	15

Table 6. 지정사형 질의어패턴

기본 구성	패턴 구성	수
N1+N	WNI1+NI NL1+WNT :	16
N1+N3+N	WNH1+NA3+NA NH1+WNH3+NB :	9
N1+N4+N	WNH1+NH4+NA NH1+WNH4+NB	6

Table 7. 형용사형 질의어패턴

기본 구성	패턴 구성	수
N1+A	WNA1+A NH1+WA :	14
N1+N3+A	WNA1+NH3+WA NH1+WNH3+A :	42
N1+N4+A	WNA1+NA4+A NH1+WNH4+A :	12

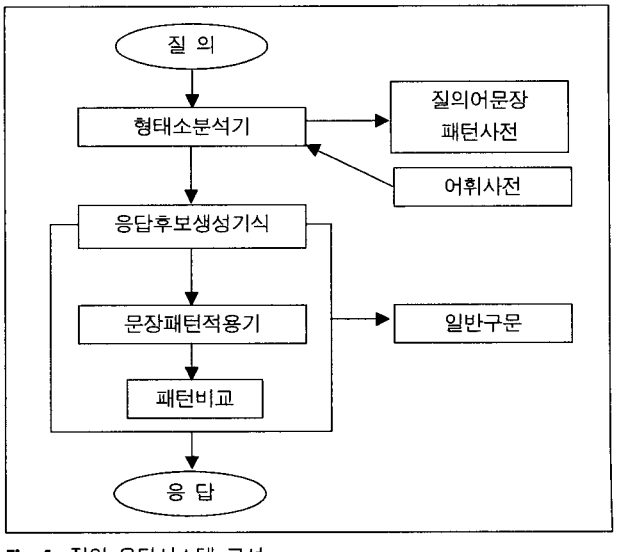

Fig. 1. 질의 응답시스템 구성.

용하여 태깅한 후 의문질의어 패턴과 비교하여 같은 형의 패턴을 응답후보로 결정하게 된다.

이와 같이 이 논문에서의 질의응답시스템에서는 정보검색시스템과는 달리 질의 처리과정이 사용자의 질의 의도를 파악할 수 있는 질의 유형이나 키워드 등의 정보를 질의로부터 다양한 형태의 패턴으로 태깅되며, 일반 문서에서 정답이 될 수 있는 정답후보(Answer Candidate)들을 추출하는데 중요하다.

또한 인터넷과 같은 실용적 환경에서 사용되면 사용자의 질의는 다양한 문장의 형태나 같은 의도를 가진 질의에 대한 정답후보를 추출해 낼 수 있을 것이다.

결론 및 향후 과제

이 논문에서는 기존의 자연어 질의 시스템들의 연구 고찰을 통하여 앞으로의 자연어 질의 시스템의 구성과 의문형 질의패턴에 대하여 살펴보았다.

본 논문의 연구 결과 일반 문장에서 도출된 의문문의 문장패턴과 도출대상의 일반 문장의 문장패턴이 일치하는 것을 확인하였고, 의문문의 의문요소와 일치되는 문장패턴의 요소를 추출함으로써 응답후보의 생성이 가능하였다. 향후 과제는 질의응답시스템에 맞는 분석기와 응답후보로부터 좀더 정확한 정답유도를 위한 다양한 패턴연구와 다양한 형태의 질의를 처리할 수 있는 자연어처리 기법을 개선해야 할 것이다.

REFERENCES

1) Mann GS(2001) : *"A Statistical Method for Short Answer Extraction", In Proceedings of the ACL Workshop Open-Domain Question Answering, pp13-30*
2) 강은국(1993) : 조선어 문형 연구, 서광학술자료사
3) 송재관(1998) : "한영 기계번역을 위한 한국어 품사 분류". *'98* 추계 정보과학회 논문집
4) 김진한(1987) : "한국어 결합가 패턴에 의한 기계번역에 관한 연구". 청주대학교 석사학위논문
5) 이석호, 홍봉희 : "한글 질의어 시스템의 설계 및 구현". 한국정보과학회 논문지, *Vol.11, No. 1, 84. 5*
6) 윤성희(1996) : "한국어 자연 언어 질의 문장분석에서 스키마 도메인 정보를 이용하는 중의성 해결 기법",상명여대 산업과학 연구
7) 이석호, 김성기 : "자연 한글 질의어 처리를 위한 인터페이스의 설계 및 구현" , 한국정보과학회 논문지, *Vol.12, No. 1, 85.2*
8) 이신영 : "정형질의어로부터 한국어의 생성". 서울대학교 석사학위 논문, *87. 1*
9) 이현주, 오경환*(1990)* : "지식 기반형 *fuzzy* 질의 응답 시스템". 인지과학 *2 : 2*
10) 채진석(1992) : "자연 한글 질의어 시스템의 설계 및 구현" , 서울대학교 석사학위 논문, *92. 2*
11) 채진석, 이석호(1995) : "객체 지향 데이터베이스를 위한 한국어 질의 인터페이스에서의 경로식 처리". 한국정보과학회 논문지 *22 (8) : 1194*
12) 김진한(1987) : "한국어 결합가 패턴에 의한 기계번역에 관한 연구". 청주대학교 석사학위논문
13) 최현배(1955) : 고등말본, *pp16-17*