한국어 질의응답시스템에서 구문정보에 기반한 질의분석

신승은[O] 서영훈
충북대학교 컴퓨터공학과
seshin@dcenlp.chungbuk.ac.kr[O], yhseo@cbu.ac.kr

Question Analysis based Syntactic Information in Korean Question Answering System

Seung-Eun Shin[O] Young-Hoon Seo
Dept. of Computer Engineering, Chungbuk National University

요 약

본 논문에서는 한국어 질의응답시스템에서 정확한 정답추출을 위한 구문 정보에 기반한 질의분석을 제안한다. 질의분석은 세부 정답 유형 결정, 세분화된 키워드 추출을 통해 정확한 정답추출을 목적으로 한다. 술어 유형 정보를 이용하여 대분류 수준의 정답 유형으로 질의분석을 수행하고, 구문 구조 정보를 이용하여 중요 키워드와 일반 키워드를 추출한다. 마지막으로 정답 유형 자질 명사를 이용하여 세부 정답 유형을 결정한다. 실험을 통해 세부 정답 유형 결정에서 정확률 59%, 세분화된 키워드 추출에서 정확률 66%를 보였다.

1. 서론

최근 대부분의 정보 검색 시스템들은 사용자의 질의에 대해 관련 있는 문서들을 결과로 제시한다. 사용자의 질의가 구체적인 대답을 요구하는 것일 경우에는 사용자들은 찾고지 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 있다.

이러한 요구의 증가에 따라 최근 국내외 연구도 활발히 진행되고 있다. 국제적인 정보검색평가대회인 TREC (Text REtrieval Conference)에서는 1999년 TREC-8에서 질의응답시스템의 평가[1]를 시작하였으며, 국내외 많은 연구소와 대학에서도 관련된 연구를 수행하고 있다.

질의응답 시스템은 크게 질의분석과 정답추출이 시스템의 핵심을 이루고 있다. 기존 연구 중 질의분석에 관하여서는 의문사 등의 키워드 추출 및 가중치 부여와 개체인식 기법에 관한 연구가 이루어져왔으며, 정답추출 부분에서는 고유명사에 대한 인식과 추정명사 등에 대한 의미 속성 결정 등을 위한 대규모 지식베이스 구축 및 활용방법에 관한 활발한 연구가 이루어지고 있다[2][3][4][5][6].

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. 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[3].

기존의 질의분석은 일반적인 키워드 추출과 대분류 수준의 정답 유형을 결정하기 때문에 정답추출 단계에서 적합한 정답을 찾는 것을 어렵게 한다. 따라서 본 논문에서는 한국어 질의응답시스템에서 구문 정보에 기반한 질의분석을 제안한다. 구문 정보는 질의응답시스템에서 질의의 술어 유형 정보와 구문 구조 정보를 의미하며, 술어 유형 정보를 이용하여 대분류 수준의 질의분석을 수행하고, 구문 구조 정보를 이용한 세분화된 키워드 추출을 한다. 추출한 키워드와 정답 유형 자질 명사를 이용하여 소분류 수준의 질의분석을 수행한다. 정확한 정답추출을 목적으로 한 세분화된 키워드 추출과 세부 정답 유형을 결정하는

질의분석은 질의응답시스템에 효과적으로 사용할 수 있다.

2. 술어 유형 정보를 이용한 질의분석

질의응답시스템에서 질의의 통상적인 구문은 술어로 끝나는 경우와 술어가 생략된 경우로 구분된다. 술어로 끝나는 질의 구문 유형은 그림 1과 같다. 그림 2는 술어가 생략된 질의의 예이다.

--는 누구일까?/누구입니까?
--에 대하여(대해) 알고 싶어요/싶습니다/싶다
무엇이 다르죠?
무엇이 있나요/있을까요
무슨 뜻이에요(뜻인가요)

그림 1. 술어로 끝나는 질의 유형

~는 누구일까?/누구입니까?
--에 대하여(대해) 알고 싶어요/싶습니다/싶다
무엇이 다르죠?
무엇이 있나요/있을까요
무슨 뜻이에요(뜻인가요)

그림2. 술어가 생략된 질의 유형

그림 1의 술어로 끝나는 질의 유형은 질의분석에 필요한 많은 정보를 가지고 있으며, 이를 이용하여 대분류 수준의 질의분석을 수행할 수 있다. 술어로 끝나는 질의 유형을 분석하여 술어 유형 정보를 구축하고, 술어 유형 정보를 이용하여 대분류 수준의 질의분석을 수행한다. 질의분석은 술어로 끝나는 질의를 대상으로 하며, 술어 유형 정보에 따라 정답 유형(대분류)을 결정하고 핵심어 정보를 이용하여 질의의 핵심어를 추출한다. 술어 유형 정보는 그림 3과 같고, 질의분석의 결과는 그림 4와 같다. 정답 유형 1은 대분류 정답 유형, 정답 유형 2는 세부 정답 유형을 의미한다. 중요 키워드는 질의의 초점을 나타내는 중요도가 높은 키워드를 의미한다.

질의 구문 유형	정답 유형	중요 키워드	
		문장 성분	예외
누구입니까	인물	주어	
알고 싶어요 알고 싶습니다		목적어	~에 대하여 ~이(가)
무엇이 다르죠	일반/비교	주어	
무엇이 있나요		주어	
무슨 뜻이에요	일반/정의	주어	
어디인가요	장소	주어	
몇 년인가요	수	주어	

그림 3. 술어 유형 정보

질의 문장	"동의보감"의 저자는 누구입니까?		
정답 유형1	정답 유형2	중요 키워드	일반 키워드
인물		저자	
질의 문장	EQ와 IQ는 무엇이 다르죠?		
정답 유형1	정답 유형2	중요 키워드	일반 키워드
일반	비교	EQ, IQ	
질의 문장	HTML은 무슨 뜻이에요?		
정답 유형1	정답 유형2	중요 키워드	일반 키워드
일반	정의	HTML	

그림 4. 술어 유형 정보를 이용한 질의 분석 결과

3. 구문 구조 정보를 이용한 질의분석

술어 유형 정보를 이용한 질의 분석 결과와 술어가 생략된 질의를 대상으로 구문 구조 정보를 이용하여 질의분석을 한다. 분석 과정은 구문 구조 정보를 이용한 중요 키워드 및 일반 키워드 추출 부분과 정답 유형 자질 명사를 이용한 세부 정답 유형 결정 부분으로 구분된다.

키워드 추출은 구문 구조 정보를 이용하고, 중요 키워드는 그림 4의 중요 키워드와 비교하여 결정한다. 그림 5는 구문 구조 정보로써, 4가지 구문에 대한 구문 구조와 중요 키워드를 나타낸다. 중요 키워드는 구문 구조에서 중요 키워드에 해당하는 키워드를 의미한다. 또한 구문 구조는 수식어를 제외한 명사들로 제한한다.

구문1:<N1>	$N_1\ N_2 \cdots N_n$
중요 키워드	$N_n >> \cdots >> N_2 >> N_1$
구문2:<N2>	$<N1>_1$+<접속 조사> $<N1>_2$
중요 키워드	$<N1>_1$중요키워드, $<N1>_2$중요키워드
구문3:<N3>	[<N1>|<N2>]$_1$ <접속 부사> [<N1>|<N2>]$_2$
중요 키워드	[<N1>|<N2>]$_1$중요키워드, [<N1>|<N2>]$_2$중요키워드
구문4:<N4>	[<N1>|<N2>|<N3>]$_1$+<속격 조사> [<N1>|<N2>|<N3>]$_2$
중요 키워드	[<N1>|<N2>|<N3>]$_2$중요키워드 >> [<N1>|<N2>|<N3>]$_1$중요키워드

>> : 키워드의 중요도 표시

[<N1>|<N2>]: <N1>또는<N2>가 나타날 수 있음을 의미

그림 5. 구문 구조 정보

정답 유형 결정은 정답 유형 자질 명사를 이용한다. 중요 키워드를 정답 유형 자질 명사와 비교하여 세부 정답 유형을 결정한다. 중요 키워드가 정답 유형을 결정하지 못할 경우에는 중요도가 높은 키워드 순으로 정답 유형 자질 명사와 비교하여 세부 정답 유형을 결정한다. 또한 중요 키워드가 정답 유형을 결정하는 경우에는 해당 중요 키워드를 일반 키워드로 바꾸고, 중요도가 가장 높은 일반 키워드를 중요 키워드로 바꾼다. 그림 6은 정답 유형 자질 명사의 예이며, 그림 7은 1차 질의분석 결과에 대한 2차 질의분석의 결과이다.

정답 유형1	정답 유형 2	자질 명사
인물	가족	남편, 부인, 아버지, 어머니, ……
인물	연예인	가수, 주연, 조연, 개그맨, ……
인물	학자	과학자, 박사, 교수, 수학자, ……
인물	저자	저자, 작가, 글쓴이, 지은이, ……

그림 6. 정답 유형 자질 명사

질의 문장	"동의보감"의 저자는 누구입니까?		
정답 유형1	정답 유형2	중요 키워드	일반 키워드
인물	저자	동의보감	저자
질의 문장	EQ와 IQ는 무엇이 다르죠?		
정답 유형1	정답 유형2	중요 키워드	일반 키워드
일반	비교	EQ, IQ	
질의 문장	고정IP와 유동IP의 차이점과 장단점?		
정답 유형1	정답 유형2	중요 키워드	일반 키워드
일반		차이점, 장단점	고정IP, 유동IP

그림 7. 구문 구조를 이용한 질의분석 결과

4. 실험 및 평가

제안한 질의분석을 평가하기 위해 웹 상에서 수집한 인물 관련 질의 중 100개의 질의를 임의 추출하여 실험하였다. 실험은 정답 유형 결정과 키워드 추출의 정확률을 측정하였다. 표 1은 실험 질의 집합에 대한 정보이다.

표 1. 실험 질의 집합

전체 질의 수	100
술어로 끝나는 질의 수	49
술어가 생략된 질의 수	51

표 2는 실험 질의 집합에 대한 정답 유형 결정과 중요 키워드 추출의 정확률을 측정한 결과이다. 세부 정답 유형 결정에 대한 정확률과 중요 키워드 추출의 정확률이 낮은 원인은 정답 유형 자질 명사와 수식 구조를 포함하는 구문 때문이다. 정확률 향상을 위해 정답 유형 자질 명사의 확장과 수식구조를 포함하는 구문 구조 정보의 추가가 필요하다.

표 2. 질의분석 실험 결과

실험 내용		정확률
정답 유형 결정	정답 유형 1	98%
	정답 유형 2	59%
중요 키워드 추출		66%

5. 결론 및 향후 연구

본 논문에서는 한국어 질의응답시스템에서 정확한 정답 추출을 위한 구문 정보에 기반한 질의분석을 제안하였다. 질의분석을 위한 구문 정보는 술어 유형 정보와 구문 구조 정보이며, 세부 정답 유형 결정을 위해 정답 유형 자질 명사를 이용하였다. 실험을 통해 제안한 질의분석은 정답 유형 결정에서 98%(정답유형1), 59%(정답유형2), 중요 키워드 추출에서 66%의 정확률을 보였다. 다소 낮은 정확률을 보이나 정답 유형 자질 명사와 다양한 구문 구조 정보의 확장을 통해 정확률 향상을 기대할 수 있다. 세부 정답 유형을 결정하고 질의의 초점을 나타내는 중요 키워드를 추출하는 질의분석을 통해 정답추출 단계에서 보다 정확한 정답을 추출할 수 있을 것이며, 이에 대한 실험이 필요하다.

참고 문헌

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

[2] 김수민, 백대호, 김상범, 임해창, " 시소러스 범주를 이용한 질의응답시스템" , 제12회 한글 및 한국어 정보처리 학술대회, pp.179~183, 2000.

[3] 이경순, 김재호, 최기선, " 한국어 질의응답 시스템에서 개체인식에 기반한 대답추출" , 제12회 한글 및 한국어 정보처리 학술대회, pp.184~189, 2000.

[4] 강승식, 이하규, 손소현, 문병주, 홍기채, " 자연언어 질의문장의 용어 가중치 부여기법" , 제14회 한글 및 한국어 정보처리 학술대회, pp.223~227, 2002.

[5] 장문수, 장명길, 김현진, 오효정, 이재성, " 인터넷 질의/응답을 위한 지식베이스 구축" , 제12회 한글 및 한국어 정보처리 학술대회, pp.198~204, 2000.

[6] 허정, 옥철영, " 사전 뜻풀이말에서 추출한 의미정보에 기반한 의미 중의성 해결" , 제12회 한글 및 한국어 정보처리 학술대회, pp.269~277, 2000.