질의응답 시스템을 위한 술어정보 기반 질의분석

김원남, 신승은, 서영훈
충북대학교 컴퓨터공학과
{wonami, seshin}@nlp.chungbuk.ac.kr, yhseo@chungbuk.ac.kr

Predicate-based Question Anlysis for Korean Question-Answering System

Won-Nam Kim, Seung-Eun Shin, Young-Hoon Seo
Dept. of Computer Engineering, Chungbuk National University

요 약

질의 응답 시스템이 정확한 정답을 제시하기 위해서는 사용자가 요구하는 정답의 유형을 결정할 필요가 있다. 질의분석의 일반적인 접근법으로는 의문사 정보, 규칙 그리고 통계 정보에 기반한 방법들이 있다. 본 논문에서는 술어정보를 이용한 질의분석을 제안한다. 먼저 의문사 정보를 이용하여 상위정답유형을 결정하고 질의문의 술어 정보와 구문 구조 정보를 이용하여 초점단어(focus word)를 추출한다. 초점단어란 정답유형을 결정하는데 단서가 되는 단어로써, 추출된 초점단어에 의해 75개의 하위정답유형 중 하나가 결정된다. 실험에 앞서 정답 유형별로 6개의 상위범주와 75개의 하위범주를 정의하였으며, 실험에는 학습 데이터의 일부와 일반 Web에서 수집한 테스트 데이터가 사용되었다. 실험결과 상위범주는 97.6%, 하위범주는 77.8%의 정확도를 보였으며 초점단어는 92.5%의 정확도를 보였다.

1. 서 론

전통적인 정보 검색(IR) 시스템은 사용자의 질의에 대해 정답이 포함된 문서들을 순위화하여 사용자에게 제공한다. 이는 시스템이 제시하는 문서 내에서, 사용자 자신이 원하는 정보를 찾아내야만 하는 별도의 과정을 필요로 한다. 그러나 대다수의 사용자들은 다량의 문서보다는 구체적인 대답을 요구하는 경우가 많다[1]. 이러한 사용자들의 요구로 인해 질의 응답이라는 개념이 등장하게 되었다.

질의응답 시스템이란 사용자가 제시하는 질의를 분석하여 거대한 문서 집합으로부터제한된 길이의 정답을 추출해 내는 시스템이다[2]. 기존 정보 검색 시스템과 질의 응답 시스템의 가장 큰 차이점은 시스템이 제시하는 결과의 형태라 할 수 있다. 질의와 관련된 정확한 정답을 제시하여야만 하는 질의 응답 시스템의 경우 일반 정보 검색 시스템보다 사용자의 의도를분석해내는 것이 매우 중요하다.

이러한 질의응답시스템은 이미 많은 연구들이 TREC (Text REtrieval Conference)[3]을 비롯한 많은 단체를 통해 진행되어 왔다. 지난 몇 년간 TREC QA track에 참가하였던 일반적인 시스템들을 살펴보면 다음 3가지의 단계를 거쳐 정답을 추출한다[4].

1. 질의문에 대한 정답유형을 결정한다.
2. 질의문에 사용된 중심어 그리고 질의어와 관련된 전문용어를 이용하여 정답이 포함되어있을 것으로 예상되는 문서나 문장을 검색한다.
3. 중심어와 검색된문장 사이에 비교 작업을 수행하여 정답을 추출한다.

이와 같이 질의문 분석을 통해 정답추출이 이루어지는 질의응답 시스템은 자연어 질의를 분석해 낼 수 있는 모듈이 필요하다. 이러한 모듈에 의해 질의에 포함된 다양한 정보들이 추출되며 추출된 정보는 정답 선택시 시스템 내에서 이용된다.

현재까지 질의분석을 위한 다양한 접근법이 시도되고 있는데 대표적인방법으로는 의문사 정보를 이용하는 방법[5-8]과 규칙에기반한 방법[2,9], 통계에 기반한 방법[10-12] 등이 있다.

의문사 정보를 이용하여 질의를 분류하는 경우 "who"(PERSON), "where"(LOCATION), "how many"(NUMBER)와 같이 질의문 상에 나타나는 의문사를 통

해 정답유형을 결정하며, "which", "what"과 같이 한가지 정답유형으로 결정할 수 없는 의문사의 경우 질의문에 나타나는 실마리 단어를 이용하여 정답유형을 결정한다. 그러나 의문사 정보만으로 모든 질의문의 정답유형을 결정할 수는 없으며 의문사가 나타나지 않는 질의문의 경우 별도의 해결책이 필요하다.

규칙에 기반한 질의 유형 분류의 경우 필요한 규칙들을 학습 데이터를 통해서 구축한 후, 어휘처리 규칙과 패턴들을 이용하여 질의문에서 의미정보를 추출한다. 일반적으로 규칙에 기반한 시스템은 즉각적인 질의 유형분류가 가능하며 성능향상을 위한 튜닝(tuning)이 용이하다. 그러나 규칙이 다양해질수록 시스템의 성능이 저하되거나 규칙에 벗어나는 질의가 나타날 경우 질의를 분류할 수 없는 단점이 있다[13].

통계적 방법에 기반해 질의를 분류하기 위해서는 수작업을 통해 분류된 대량의 학습 데이터가 필요하며 이를 통해 얻은 통계정보를 이용하여 질의를 분류한다. 통계적 방법을 사용하면 응용영역에 크게 영향을 받지 않을 수 있으며 안정적으로 질의 분류가 가능하다. 더불어 자동화된 통계적 방법을 사용함으로써 시스템 구축이 용이하다는 점도 장점으로 들 수 있다. 그러나 대량의 학습데이터를 구축하는데 많은 노력이 들며, 구조가 유사한 질의문의 경우 쉽게 분류해 내기 어려운 경우가 발생할 수 있다[13].

일반 사용자들은 시스템에 질의시 다양한 언어적 표현을 구사한다. 따라서 규칙에 기반한 질의분석을 위해서는 다양한질의 표현에 대하여 학습데이터를 수집하고 분석하여 규칙들을 정의하여야 하는 어려움이 있다. 이러한 다양한 언어적 표현은 대체적으로 질문의 술어부에서 이루어지므로 술어정보를 이용하면 규칙의다양성을 상당수 감소 시킬 수 있다. 이와 같은 이유로 본 논문에서는 술어정보에 기반한 질의분석에 대해 제안한다.

2. 술어 정보와 구문 구조 정보

2.1 술어 정보

정답 유형 분류를 위해 본 논문에서는 6개의 상위범주와 75개의 하위범주를 정의하였다. [표 1]은 정의된 상위범주와 하위범주의 일부를 보여준다.

일반적인 질의분석 모듈은 정답 유형 분류를 통해 상위범주(Main Category)를 결정해준다. 본 시스템은 상위범주 뿐만 아니라 정답의 하위범주(Sub Catego- ry)까지 결정해 줌으로써 정답 추출시 보다 많은 정보를 제공하게 된다.

[표 1] 상위범주 및 하위범주

상위범주 (Main Category)	하위범주 (Sub Category)
사람	정치가, 예술가, 학자,저자,…
장소	국가, 도시, 바다, 산,…
조직	기업, 정치단체, 교육기관,…
수	길이, 면적, 높이, 속도,…
시간	년, 월, 일,…
기타	동물, 식물, 구체물,…

일반적인 질의문은 다음의 2가지로 분류해 볼 수 있다. 술어가 포함된 질의문과 술어가 생략된 질의문을 들 수 있는데 [표 2]의 A와 B는 술어가 포함된 질의문과 생략된 질의문의 예 이다.

[표 2] 질의문의 예

분류	질의문
A	"동의보감"의 저자는 누구인가?
B	"동의보감"의 저자는?

[표 2] 의 B와 같이 술어가 생략된 질의문의 경우 대부분 마지막 어절에 나타나는 명사를 통해 정답유형의 하위범주를 결정할수 있다. 그러나 술어가 포함된 A와 같은 질의문의 경우 단순히 술어에 기술된 "누구"라는 의문사를 통해 "사람"이라는 상위범주만 결정지을 수 있다. 일반적으로 질의문에는 초점단어 (focus word)가 나타난다. 초점단어란 질의문에서 정답유형을 결정지을 수 있는 단서가 되는 단어를 말하는 것으로 대부분 명사의 형태로 질의문상에 존재한다. [표 2]의 A와 같이 술어가 포함된 질의문의 경우 초점단어의 위치는 술어에 의존적이다.

[표 2]의 A와 같은 질의문에서 "~누구인가?"와 같은 술어의 경우, 초점단어의 문장성분이 주어임을 알 수 있다. 이를 통해 시스템은 초점단어인 "저자"를 추출해낼 수 있다.

학습 데이터는 TREC-8과 9을 통해 수집한 1700여개의 질의문과 일반 웹에서 수집한 300여개의 질의문을 한국어로 번역한 후 정제하여 구축 하였다. 우리는 이렇게 수집한 2000여개의 질의문으로 부터 초점단어의 문장성분을 나타내는 술어정보를 추출, 술어사전을 구축하였다.

[표 3]은 술어정보의 일부를 보여준다. [표 3]의 예외사항은 초점단어의 문장성분에 나타나는 예외적인 경우이다.

[표 3] 술어정보

술어	초점단어(Focus Word)	
	문장성분	예외사항
알고 싶어요	목적어	~에 대하여
누구인가?	주어	
언제인가요?	주어	
몇 개 인가요?	주어	
얼마인가요?	주어	

2.2 구문 구조 정보

본 논문에서는 보다 정확하게 초점단어를 추출하기 위하여 술어정보와 함께 구문구조를 이용하였다. 술어가 생략된 질의문의 경우 술어정보를 이용하기 어렵다. 이러한 경우 명사구 내에 위치한 각각의 명사들을 이용해 초점단어로 활용한다.

[표 4]의 〈NP1〉과 같은 경우 질의문 내에서 각각의 명사들이 일정한 집합을 이루고 있을 때 이러한 명사들 중에서 초점단어를 결정하는 작업이 필요하다. "N_1 N_2 N_3 … N_n"의 순서로 이루어진 명사구의 경우 "N_n 〉〉 … 〉〉 N_3 〉〉 N_2 〉〉 N_1"와 같이 마지막에 위치하는 명사(N_n)가 초점단어일 가능성이 높다.

[표 4]의 〈NP2〉는 명사들 사이에 접속조사가 위치한 경우로 "〈NP1〉$_1$ 초점단어, 〈NP1〉$_2$ 초점단어"와 같이 "〈NP1〉$_1$, 〈NP1〉$_2$" 모두 초점단어로 활용된다.

[표 4] 구문 구조 정보

〈NP1〉	N_1 N_2 … N_n
초점단어	N_n 〉〉 … 〉〉 N_2 〉〉 N_1
질의문	"~ 올림픽 육상 선수는?"
〈NP2〉	〈NP1〉$_1$+〈접속조사〉 〈NP1〉$_2$
초점단어	〈N1〉$_1$ 초점단어, 〈N1〉$_2$ 초점단어
질의문	"~ 의 아들이며 정치가인 사람은?"
〈NP3〉	[〈NP1〉|〈NP2〉|〈NP3〉]$_1$+〈소유격 조사〉 [〈NP1〉|〈NP2〉|〈NP3〉]$_2$
초점단어	[〈NP1〉|〈NP2〉|〈NP3〉]$_2$ 초점단어 〉〉 [〈NP1〉|〈NP2〉|〈NP3〉]$_1$ 초점단어
질의문	"~ 왕의 아들은?"

2.3 자질명사 사전

사용자의 질의에서 초점단어가 추출되면 초점단어를 이용해 사용자가 원하는 정답의 유형을 결정한다. 초점단어만으로는 정답유형을 결정하기 힘들기 때문에 자질명사 사전을 이용하여 질의문의 정답유형을 결정한다.

"미국의 16대 대통령은?"과 같은 질의문에서 '대통령'이라는 초점단어가 추출되었을 경우 자질명사 사전을 통해 '정치가'라는 하위범주를 결정지을 수 있다. 자질명사 사전은 앞서 술어사전 구축시 사용된 학습 데이터에 동의어/유의어 정보를 추가하여 구축하였으며 [표 5]는 정답유형별 자질명사의 예를 보여준다.

[표 5] 정답유형별 자질명사

상위범주	하위범주	자질명사
사람	저자	저자, 작가, 글쓴이, 소설가,…
사람	연예인	가수, 배우, 탤런트, …
사람	정치가	대통령, 부통령, 국회의원, …
사람	학자	과학자, 의사, 철학자, …

2.4 자질용언 사전

자질 명사가 상위범주와 동일하거나 유사한 경우 명사 앞에 위치하는 용언을 이용한다. "세계최초로 비행기를 만든 사람은?"과 같은 경우 자질명사 만으로 하위 정답유형을 결정하기 힘들다. 이러한 경우 자질용언 "만든~"을 이용하면 '제작자'라는 하위범주의 결정이 가능하다. [표 6] 은 이와 같은 자질용언의 내용을 보여주고 있다.

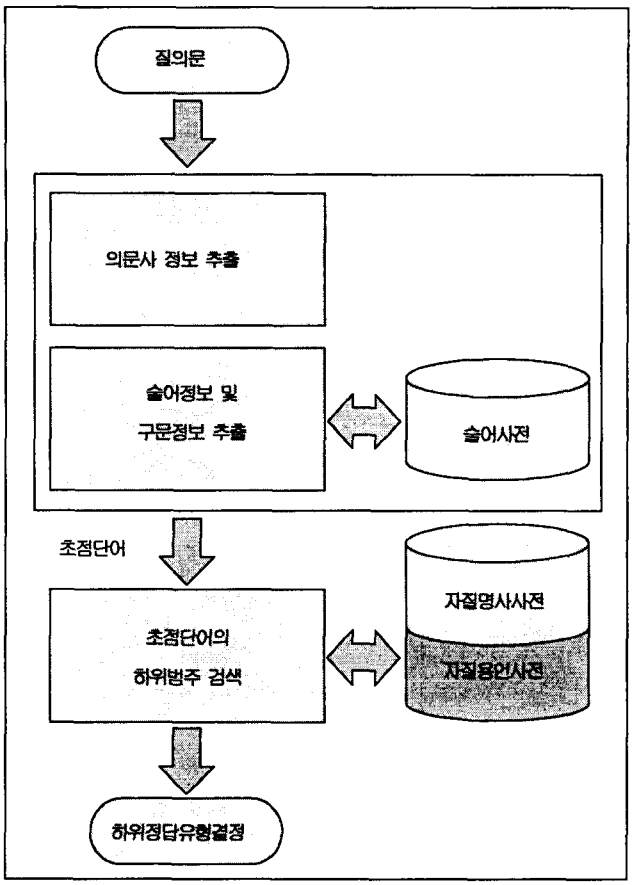
[표 6] 정답유형별 자질용언

상위범주	하위범주	자질명사	자질용언
사람	저자	사람	쓴~, 저술한~, …
사람	제작자	사람	만든~, 제작한~,…

일반적으로 이러한 자질용언을 통해 하나의 정답유형이 결정되어야 하지만 "쓰다"와 같이 "사용하다" (Use)와 "기술하다"(Write) 두 가지 이상의 의미로 사용될 수 있는 용언의 경우 각각의 정답유형을 모두 제시해 준다.

3. 질의 분석

앞서 기술한 여러 단계의 접근법을 통해 질의문을 분석하게 된다. 질의분석은 [그림 1] 과 같은 3단계의 과정으로 이루어진다.



[그림 1] 질의 분석 과정

1. 질의문 입력시 의문사 정보를 이용하여 상위범주를 결정한다.
   질의문 : "동의보감"의 저자는 누구인가?
   상위범주 : 사람(Person)
2. 술어정보 및 구문구조 정보를 이용하여 질의문으로부터 초점단어를 추출한다.
   질의문 : "동의보감"의 저자는?
   구문구조정보 : NP3
   초점단어 : 저자 〉〉 "동의보감"
3. 초점단어의 의미정보를 이용하여, 정의된 하위 범주 중 하나로 결정한다.
   초점단어 : 저자
   자질명사 사전 : 사람 - 저자
   상위범주 : 사람
   하위범주 : 저자

질의문에서 술어가 생략된 경우에는 구문구조 정보만을 이용하여 초점단어를 추출하였다.

[표 7]은 질의분석 결과의 예를 보여준다.

[표 7] 질의분석 결과 예

질의문1	"동의보감"의 저자는 누구인가?	
상위범주	하위범주	초점단어
사람	저자	저자
질의문2	영친왕의 아들은 누구인가?	
상위범주	하위범주	초점단어
사람	가족	아들

4. 실험 및 결과

실험은 인물관련 질의문을 대상으로 이루어졌다. 실험에는 학습 데이터에서 임의로 선정한 320개의 Question1과 일반 Web사이트에서 수집한 100개의 Question2가 사용되었다. [표 8]은 실험에 사용된 질의문 코퍼스들의 정보이다.

[표 8] 질의문 코퍼스

질의문	Question1	Question2
질의문-A (술어포함)	194	58
질의문-B (술어생략)	126	42
합 계	320	100

술어가 포함된 질의문 집합을 질의문-A라 하였고 술어가 생략된 질의문 집합을 질의문-B라 하였다.

[표 9]는 질의문 집합으로부터 추출된 초점단어의 정확률을 보여준다.

[표 9] 초점단어(Focus word) 추출 정확률

초점단어 추출		정확률	
		Question1	Question2
술어정보 사용	질의문 A	86.59%	75.86%
	질의문 B	98.41%	97.61%
술어정보 미사용	질의문 A	85.05%	68.96%
	질의문 B	98.41%	97.61%

[표 10]는 질의문 집합으로부터 분류된 질의문의 정확률을 보여준다.

[표 10] 질의 분류 정확률

질의 분류			정확률	
			Question1	Question2
술어 정보 사용	질의문 A	상위범주	96.90%	91.37%
		하위범주	71.64%	63.79%
	질의문 B	상위범주	98.41%	97.61%
		하위범주	84.12%	78.57%
술어 정보 미사용	질의문 A	상위범주	92.78%	86.20%
		하위범주	67.52%	55.17%
	질의문 B	상위범주	98.41%	97.61%
		하위범주	84.12%	78.57%

술어가 포함된 질의문의 경우 술어정보를 사용할 때 정확률이 향상됨을 알 수 있다. 술어정보를 사용하지 않을 경우 질의문에 일률적으로 단순한 규칙만을 적용하기 때문에 규칙에 예외적인 질의문 출현시 정확한 정답유형 결정이 힘들다. 일반 사용자의 질의는 대부분 술어표현이 다양하기 때문에 이러한 예외적인 질의문의 처리가 필요하다.

질의문-A 보다 질의문-B의 경우가 높은 정확률을 보이고 있다. 이는 술어가 생략된 질의문의 경우 질의의 초점이 명확하여 초점단어가 쉽게 검색되기 때문이다. Question2의 정확률이 낮은 이유는 비교적 정제가 잘 되어있는 Question1 보다 술어부의 표현이 난해하고, 띄어쓰기나 맞춤법 등이 잘 지켜지지 않은 질의문이 다수 포함되어 있기 때문이다. Question2는 일반 Web상에서 이루어지는 질의응답을 고려하여 정제하지 않고 사용하였다.

5. 결 론

본 논문에서는 질의응답 시스템을 위한 술어정보 기반 질의분석을 제안하였다. 먼저 질의문 속에 나타나는 의문사 정보를 이용하여 6개의 상위정답유형 중 하나를 결정한다. 이어서 질의문의 술어 정보와 구문 구조 정보를 이용하여 초점단어를 추출한다. 마지막으로 추출된 초점단어를 통해 75개의 하위정답유형 중 하나가 결정된다.

실험결과 Question1을 기준으로, 술어정보 사용시 71.64%, 술어정보 미사용시 67.52%의 정확도를 보였으며 초점단어의 정확도는 86.59%의 정확도를 보였다.

Question2가 Question1보다 낮은 정확도를 보이는 것은 학습 데이터를 통해 구축된 술어정보가 일반적인 사용자의 질의를 처리할 수 없기 때문이다. 이를 보완하기 위해 술어정보를 추가 할 예정이며 현재 정답 추출에 적용하기 위한 연구를 진행 중이다.

6. 참고 문헌

[1] Ellen M. Voorhees, Dawn M. Tice, "Building a Question Answering Test Collection", In *Proceeding of SIGIR 2000*, pp. 200 207, 2000.

[2] Daisuke Kawahara, Nobuhiro Kaji, Sadao Kurohashi, "Question and Answering System based on Predicate Argument Matching", In *Proceedings of the Third NTCIR Workshop,* 2002.

[3] TREC (Text REtrieval Conference) Overview, http : //trec.nist.gov/overview.html.

[4] Ellen M. Voorhees, "Overview of the TREC 2003 Question Answering Track", In *Proceedings of the Tenth Text REtreival Conference (TREC 2003),*2003.

[5] 김수민, 임해창, "시소러스 범주정보를 이용한 질의응답 시스템", 고려대학교 대학원 컴퓨터학과, 2000.

[6] Yi Chang, Hongbo Xu, Shuo Bai, "TREC 2003 Question Answering Track at CAS ICT", In *Proceedings of the Tenth Text REtreival Conference (TREC 2003),*2003.

[7] Kenneth C. Litkowski, "Use of Metadata for uestion Answering and Novelty Tasks", In *Proceedings of the Tenth Text REtreival Conference (TREC 2003),*2003.

[8] Min Wu, Xiaoyu Zheng, Michelle Duan, Ting Liu and Tomek Strzalkowski, "Question Answering By Pattern Matching, Web Proofing, Semantic Form Proofing", In *Proceedings of the Tenth Text REtreival Conference (TREC 2003),*2003.

[9] E. Nyberg, T. Mitamura, J. Callan, J. Carbonell, R. Frederking, K. Collins Thompson, L. Hiyakumoto, Y. Huang, C. Huttenhower, S. Judy, J. Ko, A. Kupsc, L. V. Lita, V. Pedro, D. Svoboda and B. Van Durme, "The JAVELIN Question Answering System at TREC 2003 : A Multi Strategy Approach ith Dynamic Planning", In *Proceedings of the Tenth Text REtreival Conference (TREC 2003),*2003.

[10] Ittycheriah A., Franz M, Zhu W. and Ratnaparkhi A., "IBM's Statistical Question Answering System", In *Proceedings of the Ninth Text Retrieval Conference(TREC-9).*

[11] Ittycheriah A., Franz M, Zhu W. and Ratnaparkhi A., "Question Answering Using Maximum Entropy Components". In *Proceedings of NAACL,* 2001.

[12] Mann G. S., "A Statistical Method for Short Answer Extraction", In *Proceedings of the ACL Workshop Open Domain Question Answering*, pp.13-30, 2001.

[13] 김학수, 안영훈, 서정연, "한국어 질의응답 시스템을 위한 지지벡터기계 기반의 질의유형분류기", 정보과학회 논문지, 제30권 제 5호, pp.466-475, 2003.