# 한국어 질의 응답 시스템을 위한 초점단어 기반 질의분석

## Question Analysis based on Focus-words for Korean Question-Answering System

김원남, 신승은, 서영훈
충북대학교 컴퓨터공학과

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

**요약**

질의 응답 시스템은 사용자의 질의를 분석하여 제한된 길이의 정답을 제시해 주는 시스템이다. 질의 응답 시스템은 정확한 정답을 추출하기 위해 사용자의 질의를 분석하는 과정을 필요로 한다. 본 논문에서는 초점단어(focus-word)를 이용한 질의분석을 제안한다. 초점단어란 정답유형을 결정하는데 단서가 되는 단어로써, 추출된 초점단어에 의해 75개의 하위정답유형 중 하나가 결정된다. 실험에는 학습 데이터의 일부와 일반 Web에서 수집한 테스트 데이터가 사용되었다. 실험결과 상위범주는 97.18%, 하위범주는 95.31%의 정확도를 보였다.

**Abstract**

Question-Answering (QA) system has to analyze user's intention correctly to respond correct answer for user's question., This paper proposes a focus-word-based question analysis approach for Korean QA system to analyze user's intention correctly. focus-word is a clue-word which selects question type. The question type is determined to one in 75 subcategories using semantics of focus-words. the proposed system accomplished 97.18% accuracy for the main category and 95.31% accuracy for the subcategory in the question classification.

## Ⅰ. 서 론

기존 정보 검색 시스템(IR System)은 사용자의 질의에 대해 정답이 포함된 문서들을 순위화 하여 제공한다. 이는 시스템이 제시하는 문서 내에서, 사용자 자신이 원하는 정보를 찾아내야만 하는 별도의 과정을 필요로 한다. 그러나 대다수의 사용자들은 문서의 형태보다는 짧은 길이의 명확한 정답을 시스템이 제시해 주길 원한다.[1]. 이러한 사용자들의 요구로 인해 질의 응답이라는 개념이 등장하게 되었다.

질의응답 시스템이란 사용자가 제시하는 질의를 분석하여 거대한 문서 집합으로부터 제한된 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 의도를 분석해내는 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 포함되어 있을 것으로 예상되는 문서나 문장을 검색한다.
3. 중심어와 검색된 문장 사이에 비교 작업을 수행하여 정답을 추출한다.

이와 같이 질의문 분석을 통해 정답추출이 이루어지는 질의 응답 시스템은 자연어 질의를 분석해 내는 과정이 필요하다. 이러한 과정을 통해 질의에 포함된 다양한 정보들이 추출되며 추출된 정보는 정답 추출에 사용된다.

현재까지 질의분석을 위한 다양한 접근법이 시도되고 있는데 대표적인 방법으로는 의문사 정보를 이용하는 방법[5-8]과 규칙에 기반한 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 통해 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 다양해 질수록 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 사용자들이 정보를 얻고자 할 때 질의를 통해 시스템에 정보를 요구한다. 사용자의 질의문 속에는 사용자의 질의 의도가 담긴 초점단어들이 존재하고 있으므로 정확한 초점단어들을 추출해 내는 것은 질의분석에 있어 매우 중요하다. 그러나 일반 사용자의 질의는 다양한 언어적 표현으로 이루어지기 때문에 정확한 초점단어를 추출하기 위해서는 학습데이터를 수집하고 분석하여 규칙들을 정의하는 과정이 필요하다. 이러한 다양한 언어적 표현은 대체적으로 질의문의 술어부에서 이루어지며 술어정보를 이용하면 규칙의 다양성을 상당수 감소 시킬 수 있다.

이러한 과정을 통해 추출된 초점단어들은 정답의 하위유형을 결정하는데 사용된다. 정답의 유형을 결정하는 것은 문서내에 나타나는 정답의 형태가 유형별로 일정한 패턴을 보인다는 점에서 중요하다. 이렇게 정답 유형을 결정하게 되면, 정답 추출시 문서내에 나타나게 되는 패턴 정보를 활용 할 수 있으며, 색인어 정보만으로 정답 추출이 어려운 경우도 보완할 수 있다. 본 논문에서는 술어정보와 함께 구문 구조 정보, 축약어 정보를 이용하여 초점단어를 추출하는 방법과 추출된 초점단어를 통해 정답의 하위범주를 결정하는 방법을 제안한다.

## II. 초점단어 추출

### 2.1 축약어 정보

질의 응답 시스템에서 일반적인 사용자들의 질의는

자연어로 이루어진다. 이는 기존 정보 검색 시스템보다 질의의 표현이 자유롭고 사용자 자신이 원하는 정보에 대한 보다 많은 정보를 전달 할 수 있는 이점이 있다. 그러나 표현이 자유롭고 다양한 만큼 문법적 오류나 일상적으로 많이 사용되지 않는 인터넷 용어들이 자주 등장하게 된다.

일반 Web에 기반한 질의 응답 시스템 구축 시 이러한 문제점은 고려되어야 한다. 본 시스템에서는 이러한 문제점 중 인터넷에서 사용되는 축약어에 관련된 정보를 수집하여 질의 분석에 사용 하였다.

사용자의 질의가 입력된 경우 축약어 정보를 이용하여 질의문내에 축약어가 사용되었는지 검사하게 되는데, 축약어가 사용되었을 경우 동일한 의미를 가지는 술어로 변환 함으로써, 정확한 초점단어 추출을 가능하게 한다.

[표 1]은 축약어 정보의 예 이다.

[표 1] 축약어 정보

| 축약어 | 축약어 복원 |
|---|---|
| ~ 질문요 | ~ 질문이요 |
| ~ 알려주셈 | ~ 알려 주세요 |
| ~ 갈쳐주셈 | ~ 가르쳐 주세요 |
| ~ 누구져? | ~ 누구지요? |

## 2.2 술어 정보

정답 유형 분류를 위해 본 논문에서는 6개의 상위범주와 75개의 하위범주를 정의하였다. [표 2]는 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 Category)까지 결정해 줌으로써 정답 추출시 보다 많은 정보를 제공하게 된다.

[표 2] 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 3]의 A와 B는 술어가 포함된 질의문과 생략된 질의문의 예 이다.

[표 3] 질의문의 예

| 분류 | 질의문 |
|---|---|
| A | "햄릿"의 지은이는 누구인가? |
| B | "햄릿"의 지은이는? |

[표 3]의 B와 같이 술어가 생략된 질의문의 경우 대부분 마지막 어절에 나타나는 명사를 통해 정답유형의 하위범주를 결정할 수 있다. 그러나 술어가 포함된 A와 같은 질의문의 경우 단순히 술어에 기술된 "누구"라는 의문사를 통해 "사람"이라는 상위범주만 결정 지을 수 있다.

일반적으로 질의문에는 초점단어 (focus-word)가 나타난다. 초점단어란 질의문에서 정답유형을 결정지을 수 있는 단서가 되는 단어를 말하는 것으로 대부분 명사의 형태로 질의문상에 존재 한다. [표 3]의 A와 같이 술어가 포함된 질의문의 경우 초점단어의 위치는 술어에 의존적이다.

[표 3]의 A와 같은 질의문에서 "~누구인가?"와 같은 술어의 경우, 초점단어의 문장성분이 주어임을 알 수 있다. 이를 통해 시스템은 초점단어인 "지은이"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 질의문으로부터 초점단어의 문장성분을 나타내는 술어정보를 추출, 술어사전을 구축하였다.

[표 4]는 술어정보의 일부를 보여준다. [표 3]의 예외사항은 초점단어의 문장성분에 나타나는 예외적인 경우이다.

[표 4] 술어정보

| 술어 | 초점단어(focus-word) | |
|---|---|---|
| | 문장성분 | 예외사항 |
| 알고 싶어요 | 목적어 | ~에 대하여 |
| 누구인가? | 주어 | - |
| 언제인가요? | 주어 | - |
| 몇 개 인가요? | 주어 | - |
| 얼마인가요? | 주어 | - |

## 2.3 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 5]의 <NP1>과 같은 경우 질의문 내에서 각각의 명사들이 일정한 집합을 이루고 있을 때 이러한 명사들 중에서 초점단어를 결정하는 작업이 필요하다. "N1 N2 N3 ... Nn"의 순서로 이루어진 명사구의 경우 "Nn >> ... >> N3 >> N2 >> N1"와 같이 마지막에 위치하는 명사(Nn)가 초점단어일 가능성이 높다.

[표 5]의 <NP2>는 명사들 사이에 접속조사가 위치한 경우로 "<NP1>1 초점단어, <NP1>2 초점단어"와 같이 "<NP1>1, <NP1>2" 모두 초점단어로 활용된다.

[표 5] 구문 구조 정보

| | |
|---|---|
| <NP1> | N1 N2 ... Nn |
| 초점단어 | Nn >> ... >> N2 >> N1 |
| 질의문 | "2004년 마라톤에서 금메달을 획득한 올림픽 육상 선수는 누구인가?" |
| <NP2> | <NP1>1+<접속조사> <NP1>2 |
| 초점단어 | <N1>1 초점단어, <N1>2 초점단어 |
| 질의문 | "김좌진 장군의 아들이자 정치가인 사람은 누구인가?" |
| <NP3> | [<NP1>\|<NP2>\|<NP3>]1+<소유격 조사> [<NP1>\|<NP2>\|<NP3>]2 |
| 초점단어 | [<NP1>\|<NP2>\|<NP3>]2 초점단어 >> [<NP1>\|<NP2>\|<NP3>]1 초점단어 |
| 질의문 | "영친왕의 아들은 누구인가?" |

# III. 정답 유형 결정

## 3.1 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 결정 지을 수 있다. 자질명사 사전은 앞서 술어사전 구축시 사용된 학습 데이터에 동의어/유의어 정보를 추가하여 구축하였으며 [표 6]은 정답유형별 자질명사의 예를 보여준다.

[표 6] 정답유형별 자질명사

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

### 3.2 자질용언 사전

자질 명사가 상위범주와 동일하거나 유사한 경우 명사 앞에 위치하는 용언을 이용한다. "세계최초로 비행기를 만든 사람은?"과 같은 경우 자질명사 만으로 하위정답유형을 결정하기 힘들다. 이러한 경우 자질용언 "만든~"을 이용하면 '제작자'라는 하위범주의 결정이 가능하다.

[표 7]은 이와 같은 자질용언의 내용을 보여주고 있다.

[표 7] 정답유형별 자질용언

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

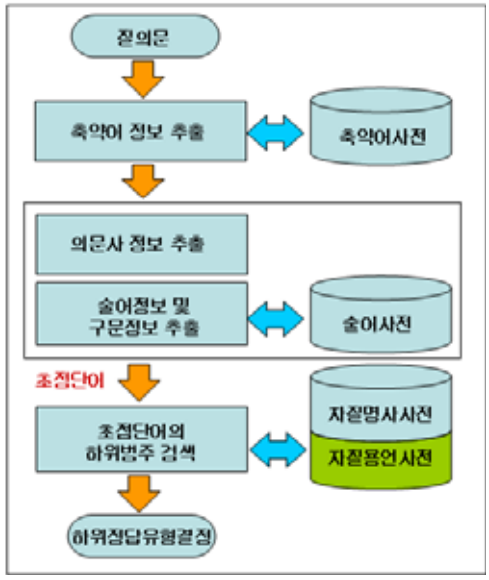
일반적으로 이러한 자질용언을 통해 하나의 정답유형이 결정되어야 하지만 "쓰다"와 같이 "사용하다"(Use)와 "기술하다"(Write) 두가지 이상의 의미로 사용될 수 있는 용언의 경우 각각의 정답유형을 모두 제시해 준다.

## IV. 질의 분석

앞서 기술한 여러 단계의 접근법을 통해 질의문을 분석하게 된다. 질의분석은 [그림 1] 과 같은 4단계의 과정으로 이루어진다.

1. 질의문 입력시 축약어 정보를 이용하여 질의문의 원형 복원.
   - 질의문 : "햄릿"의 지은이는 누구져?
   - 원형복원 : "햄릿"의 지은이는 누구지요?

2. 질의문 입력시 의문사 정보를 이용하여 상위범주를 결정한다.
   - 질의문 :"햄릿"의 지은이는 누구지요?
   - 상위범주 : 사람(Person)




[그림 1] 질의 분석 과정

3. 술어정보 및 구문구조 정보를 이용하여 질의문으로부터 초점단어를 추출한다.
   - 질의문 :"햄릿"의 지은이는?
   - 구문구조정보 : NP3
   - 초점단어 : 지은이 >> "햄릿"

4. 초점단어의 의미정보를 이용하여, 정의된 하위범주 중 하나로 결정한다.
   - 초점단어 : 지은이
   - 자질명사 사전 : 지은이 -> 저자
   - 상위범주 : 사람, 하위범주 : 저자

질의문에서 술어가 생략된 경우에는 구문구조 정보만을 이용하여 초점단어를 추출하였다.

[표 8]은 질의분석 결과의 예를 보여준다.

[표 8] 질의분석 결과 예

| 질의문1 | 동방견문록을 지은 사람은 누구인가? | | |
|---|---|---|---|
| 상위범주 | 하위범주 | 초점단어 | 자질용언 |
| 사람 | 저자, 제작자 | 사람 | 지은~ |
| 질의문2 | 이성계의 아들은 누구인가? | | |
| 상위범주 | 하위범주 | 초점단어 | 자질용언 |
| 사람 | 가족 | 아들 | - |

## V. 실험 및 결과

실험은 인물관련 질의문을 대상으로 이루어졌다. 실험에는 학습 데이터에서 임의로 선정한 320개의 Question1과 일반 Web사이트에서 수집한 100개의 Question2가 사용되었다. [표 9]는 실험에 사용된 질의문 코퍼스들의 정보이다.

[표 9] 질의문 코퍼스

| 질의문 | Question1 | Question2 |
|---|---|---|
| | (학습데이터) | (Web 데이터) |
| 질의문 수 | 320 | 100 |

[표 10]은 질의문 집합으로부터 추출된 초점단어의 정확률을 보여준다.

[표 10] 초점단어(focus-word) 추출 정확률

| 정확률 | |
|---|---|
| Question1 | Question2 |
| 95.93% | 90.00% |

Question1의 경우 학습 데이터에서 실험 데이터를 추출하였기 때문에 Question2보다 높은 정확률을 보이고 있다.

Question2는 일반 Web에서 수집한 실험 데이터로 문법적 오류가 다수 포함되어 비교적 낮은 정확률을 보였다

[표 11]은 질의문 집합으로부터 분류된 정답유형의 정확률을 보여준다.

표 11. 질의 분류 정확률

| 질의 분류 | 정확률 | |
|---|---|---|
| | Question1 | Question2 |
| 상위범주 | 97.18% | 95% |
| 하위범주 | 95.31% | 89% |

Question2의 정확률이 낮은 이유는 앞서 기술한 바와 같이 비교적 정제가 잘 되어있는 Question1 보다 술어부의 표현이 난해하고, 띄어쓰기나 맞춤법 등이 잘 지켜지지 않은 질의문이 다수 포함되어 있기 때문인데, Question2는 일반 Web상에서 이루어지는 질의응답을 고려하여 정제하지 않고 사용하였다.

실험 결과의 세부적인 사항을 살펴 볼때 술어가 포함된 질의문의 경우 술어정보를 사용할 때 정확률이 향상됨을 알 수 있었다. 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 본 시스템과 같이 이러한 예외적인 처리가 필요하다.

반면에 술어가 생략된 질의문의 경우 술어가 포함된 질의문 보다 높은 정확률을 보였다. 이러한 경우 사용자가 자신의 의도를 짧고 명확하게 표현하게 되는데, 질의의 초점이 명확하여 초점단어가 쉽게 검색되기 때문이다.

## VI . 결 론

본 논문에서는 질의응답 시스템을 위한 초점단어 기반 질의분석을 제안하였다. 먼저 질의문 속에 나타나는 축약어를 원형으로 복원한 후, 의문사 정보를 이용하여 6개의 상위정답유형 중 하나를 결정한다. 이어서 질의문의 술어 정보와 구문 구조 정보를 이용하여 초점단어를 추출한다. 마지막으로 추출된 초점단어를 통해 하위 정답 유형을 결정한다.

실험결과 Question1을 기준으로, 상위범주는 97.18%,

하위범주는 95.31%의 정확도를 보였으며 초점단어 추출은 95.93% 의 정확도를 보였다.

Question2가 Question1보다 낮은 정확도를 보이는 것은 학습 데이터를 통해 구축된 술어정보 및 축약어 정보에 예외적인 질의가 다수 포함되어 있기 때문이다. 현재 이를 보완하기 위한 작업과 정답 추출에 적용하기 위한 연구가 진행 중이다.

## ▌참고문헌▐


[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 질의응답시스템", 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 Approachith 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", 정보과학회논문지, 제30권 제 5호, pp.466-475, 2003.