질문 규칙을 이용한 기록정보 질의-응답 시스템[1)]

오수현, 안영민, 박희근, 이충희¶, 서영훈
충북대학교 컴퓨터공학과
한국전자통신연구원 지식마이닝연구팀¶
{ceany, maniac, pinetree}@nlp.chungbuk.ac.kr, forever@etri.re.kr¶,
yhseo@chungbuk.ac.kr

Record Information Question-Answering System Using Question Rules

Su-Hyun Oh, YoungMin Ahn, Hee-Geun Park, Chung-Hee Lee¶, Young-Hoon Seo

Dept. of Computer Engineering, Chungbuk National University
Electronics and Telecommunications Research Institute(ETRI)¶

요 약

본 논문에서는 기네스 기록정보, 즉 기록적 가치가 있는 기록정보에 대한 질의를 처리하는 시스템에 대하여 기술한다. 기록정보 질의의 경우 일반적으로 정형화된 형태로 나타나며 이 형태를 규칙으로 사용하여 질의에 해당되는 정답을 추출하게 된다. 기록적 가치가 있는 문장에서 해당 문장이 기록 문장임을 나타내어 주는 부사를 기록부사로 정의하고, 예로 가장, 제일, 최고의, 최대의, 최소의, 최초의, 최초로 등을 들 수 있다. 기록정보 질의의 경우 용언의 포함여부에 따라 기록부사는 두 가지 유형으로 분류된다. 기록부사는 질의문 내의 지역정보 및 정답유형과 함께 정답 추출의 중요한 요소로 사용되고, 용언정보는 기록 부사의 유형, 질의문 내의 용언 포함 여부에 따라 정답 추출의 요소로 결정되어진다. 제안한 시스템은 질의분석을 통하여 정답 추출을 위한 단서를 찾고 이를 이용하여 후보 문서와 후보 문장을 검색한 후 정답 추출 규칙을 이용하여 정답을 추출하게 된다.

1. 서론

정보검색 시스템과 질의-응답 시스템의 가장 큰 차이점은 사용자가 원하는 정답을 바로 제공하느냐 제공하지 않느냐에 있다. 정보검색 시스템은 사용자가 입력한 자연언어 질의와 관련된 문서를 순위화하여 제공해주는 시스템으로 추가적으로 사용자가 작업을 하여 자신이 원하는 정보를 추출해야 하는 반면, 질의-응답 시스템은 사용자의 자연언어 질의를 입력받아 사용자가 요구하는 정답을 사용자에게 제공해주는 시스템으로 정보검색 시스템보다 편의성이 높으므로 그 요구가 증가되고 있는 추세이다[1].

질의-응답 시스템에 관한 연구는 TREC(Text REtrieval Conference)을 중심으로 활발히 진행되고 있다[2].

질의-응답 시스템은 질의를 분석하는 부분과 정답을 추출하는 부분으로 나눌 수 있다. 질의를 분석하는 부분은 사용자가 입력한 질의에 대하여 질의유형, 정답유형 등을 결정하고, 정답 추출 부분에서는 질의에 사용된 단어나 질의와 관련된 어휘를 사용하여 정답이 포함되어 있을 만한 문장이나 문단을 검색하고 정답 추출 규칙 등을 이용하여 정답을 추출하게 된다.

본 논문에서는 기네스 기록정보, 예를 들어 "세계 최초의 금속활자는?", "세계에서 가장 높은 산은?" 등의 기록적 가치가 있는 질의를 처리하여 질의에 맞는 정답을 찾아주는 시스템에 대해 기술한다.

기록정보 질의-응답 시스템으로 현재 서비스되고 있는 시스템으로 한국전자통신연구원의 AnyQ System[3]이 있다. 이 시스템은 문장의 문맥정보를 정답 색인 템플릿이라는 특정 형태로 표현, 이를 이용하여 기록정보를 색인하고 정답으로 제시해주는 시스템이다[4].

정답 색인 템플릿을 이용한 시스템의 경우 템플릿이 잘 정의되어 있으면 정확한 정답 색인이 가능하기 때문에 시스템이 높은 성능을 낼 수 있다는 장점이 있으나, 템플릿을 작성하는 데 많은 시간이 소요되고 도메인이 변경되었을 경우 재색인이 필요하게 되는 문제가 있다.

이에 템플릿을 사용하지 않고 일반적인 방법으로 기록정보에 대한 질의-응답을 처리할 수 있는 시스템을 제안한다.

1) 본 연구는 2005년 한국전자통신연구원의 용역과제로 수행한 "패턴기반 질의응답을 위한 백과사전 및 웹 뉴스 패턴 지식 구축"의 자료를 바탕으로 수행되었다.

2. 기록정보 질문규칙과 기록부사

기록정보란 특정 분야에서의 기록적 가치가 있는 문장을 뜻하는 말로 문장 내에 기록부사라 불리는 부사를 포함한다. 기록부사[2]는 기록문장의 뜻을 명확히 나타내주는 부사로 해당 문장이 기록문장인지 아닌지의 여부를 처음으로 결정짓게 해주는 요소이며, 그 예로 "가장, 제일, 최초로, 처음, 처음으로, 최초의, 최대의, 최소의, 최고의, 제일의" 등의 단어를 들 수 있다.

그러나 기록부사가 포함되어 있는 문장이 모두 기록정보가 될 수는 없고, 기록문장 내에 필수 정보를 포함하는 문장만이 기록정보가 될 수 있다. 필수 정보란 기록문장 내의 지역정보, 정답유형을 말하며, 이 두 가지 정보 및 기록부사가 모두 존재하는 기록문장을 기록정보로 정의하였다. 또한 기록부사에 따라 기록정보는 용언정보를 추가적으로 포함할 수 있다[4].

예를 들어 "(가재) 달팽이나 곤충의 유충, 벌레, 올챙이 등을 잡아먹고 밤에 **가장** 활동적이다." 의 문장은 기록부사 "가장" 을 포함하고 있으나 필수 정보인 지역정보 및 정답유형을 문장 내에서 찾을 수 없으므로 기록정보에서 제외하였다. 또 다른 문장 "기독교 방송은 1954년 12월 15일 창립된 한국 최초의 민간방송이다." 에서 기록부사는 "최초의" 가 되고, 지역정보는 "한국", 정답유형은 "민간방송" 이 되며, 이 문장을 질의문 형태로 나타내면 "한국 최초의 민간방송은" 의 형식으로 표현할 수 있다. 다른 형태로 용언이 포함되는 질의문의 경우 기록문장 "세계에서 가장 높은 산은 에베레스트산이다." 에서 기록부사는 "가장" 이 되고 지역정보는 "세계", 정답유형은 "산", 그리고 용언정보는 "높은" 이 되어 질의문 "세계에서 가장 높은 산은?" 의 형태로 표현할 수 있다.

이처럼 기록정보 질의는 질문을 형성하는 단어의 개수가 적고 일정한 규칙으로 이루어지는 경우가 많다. 그리고 표1과 같이 두 그룹으로 나눌 수가 있다.

[표 1] 기록부사 및 질문 예

기록부사		예문
그룹1	최초의	한국 **최초의** 신부는?
	최대의	세계 **최대의** 자동차 회사는?
	최고의	중국 **최고의** 부자는?
그룹2	가장	세계에서 **가장** 높은 산은?
	제일	타이완에서 **제일** 큰 담수호는?
	최초로	세계 **최초로** 실용화된 계산기는?

그룹1의 기록부사는 '최초의', '최대의', '최소의', '최고의', '제일의' 로 질의문 내에 서술적인 의미를 포함하기 때문에 별도의 용언이 필요하지 않고, 그룹2의 기록부사는 '가장', '제일', '최초로', '처음', '처음으로' 등이 있으며 질의문 내에 용언을 필요로 한다. 그룹1의 질의문은 '지역정보|기록부사|정답유형' 의 정형적인 형태를 보이고, 그룹2의 질의문은 '지역정보|기록부사|용언정보|정답유형' 의 정형적인 형태를 보이거나 목적어, 부사구 등을 동반하게 된다.

기록정보 질의-응답 시스템은 일반적인 질의-응답 시스템의 서브시스템으로 사용되어, 기록정보에 관련한 질의가 입력되면 기록정보 질의-응답 시스템에서 처리하게 된다. 이때, 기록부사를 중심으로 '지역정보', '정답유형', '용언정보' 등의 정답 추출에 필요한 필수 정보를 추출하며, 이 정보들을 추출할 때 기록정보 질의의 정형적인 문장 형태를 이용하게 된다.

본 논문에서 제안하는 기록정보 질의-응답 시스템의 실험을 위해 파스칼 백과사전[5]에서 기록 정보 질의에 대한 응답이 가능할 것으로 판단되는 16000여 문장에 대한 분석결과[3] 기록부사 "가장, 제일" 을 포함하는 문장은 5383개, "최초로, 처음, 처음으로" 를 포함하는 기록문장은 4476개, 그리고 "최초의, 최대의, 최소의, 최고의, 제일의" 를 포함하는 문장은 4502개였다.

이 중 기록부사 그룹 1을 포함하는 4502문장을 분석한 결과 '지역정보|기록부사|정답유형(단일어)' 의 순서로 나오는 문장이 1898개로 상당히 많은 수의 문장이 추출되었고 일부는 '기록부사|지역정보|정답유형' 의 형태로 나타나기도 하였다. 기록부사 그룹1을 포함하는 4502문장 중 기록적 가치가 없다고 판단된 문장은 1605문장이고, 중심어휘가 2개 이상인 문장이 27개, '한국 최초의' 처럼 지역정보와 기록부사가 붙어 다중태그를 가지는 문장이 27개 추출되었고, 기록정보인지 아닌지 판단하기 모호한 문장은 804개였다.

그룹 2를 포함하는 9859 문장의 분석결과는 1043개의 문장이 기록적 가치가 있는 문장으로 추출되었고, 여기서 828개의 문장이 "지역정보|기록부사|용언정보|정답유형" 의 순서로 나타남을 확인하였다. 기록적 가치가 없다고 판단된 문장은 7193개, 중심어휘가 2개 이상인 문장이 13개, 다중태그를 가지는 문장은 113개, 2개 이상의 용언을 가지는 문장은 16개 추출되었고, 기록정보로 판단이 모호한 문장은 1481개였다.

본 논문에서는 "지역정보|기록부사|정답유형" 의 형태와 "지역정보|기록부사|용언정보|정답유형" 의 형태를 가지는 질의문을 대상으로 하였다.

3. 질의 분석 및 정답 추출

3.1 질의 분석

본 논문에서 제안하는 시스템은 일반적인 질의-응답 시스템의 서브 시스템으로, 사용자의 질의가 기록정보

2) 기록문장 내에 쓰인 부사 중 최상급을 나타내주는 어휘에 대해 기록부사라 정의하였다.

3) 기록문장의 분석과 규칙의 작성은 약 2개월의 시간이 소요되었다.

와 관련된 질의일 경우 이에 대한 처리를 한다.
일반 질의-응답 시스템에 사용자의 자연언어 질의가 입력되면 첫 번째로 형태소 분석이 수행된다. 질의문 내에서 기록부사가 발견되면 기록정보 질의 처리 모듈로 넘겨져 처리를 하게 된다.
또한 형태소 분석 과정에서 용언이 발견될 경우 용언 정보도 함께 기록정보 질의 처리 모듈로 넘겨지게 된다.
기록정보 질의 처리 모듈로 넘겨진 질의문은 개체명 인식기를 통해 기록부사를 중심으로 지역정보와 정답유형을 결정하게 된다. 지역정보와 정답유형은 동의어 사전 및 시소러스를 이용하여 어휘 확장이 된다. 예를 들어 지역정보가 한국일 경우 '대한민국', '국내', '우리나라' 등으로 확장된다. 확장된 어휘는 정답 추출 시 정답 후보 문서를 검색하는 데 사용된다.

질의문 구조는 표 2에 예시하였고 시스템 구성도는 그림 1에 예시하였다.

[표 2] 질의문 구조

질의문 : 한국 최초의 동물원은?			
기록부사	**지역정보**	**용언정보**	**정답유형**
최초의	한국	없음	동물원
질의문 : 세계에서 가장 높은 산은?			
기록부사	**지역정보**	**용언정보**	**정답유형**
가장	세계	높	산

질의분석이 완료되면 각 정보를 정답 추출 모듈로 넘겨주게 되며 정답 추출 모듈에서 정답을 추출하게 된다.

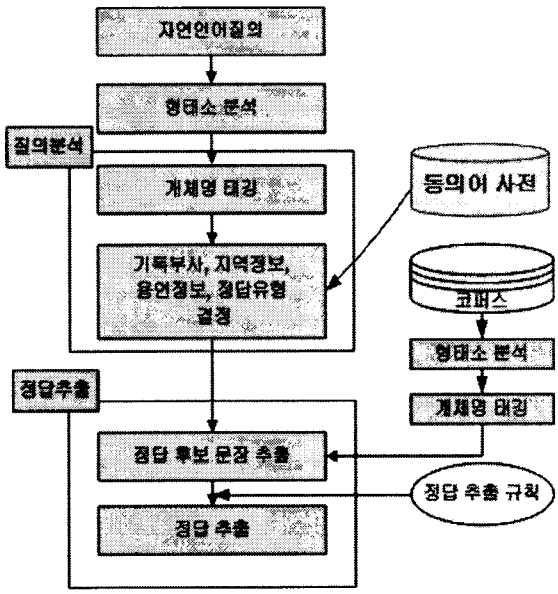
3.2 사용 자원

본 논문에서 제안한 시스템에 사용된 자원으로 형태소 분석기는 본 연구실의 CBKMA[6] V3.0을 사용하였다. CBKMA는 시스템 사전 112,669어절, 문법 형태소 사전 38,019어절, 고빈도 어절 기분석 사전 8,044어절로 구성된 사전을 사용한다. 또한 개체명 인식기는 본 연구실의 CBKNE V1.0을 사용하였다. 개체명 인식기는 인명 사전 22만 엔트리, 국가명 사전 210엔트리, 회사명 사전 2287엔트리, 지명사전 445엔트리로 구성되어 있으며, 본 시스템에서는 정답유형의 인식을 위해 "~산", "~방송" 등의 정보를 추가하여 사용하였다.
동의어 사전은 13689개의 엔트리로 구성되어 있다.

3.2 정답 추출

정답 추출 모듈에서 질의분석 모듈에서 추출된 지역정보, 정답유형, 용언정보 및 확장된 어휘를 이용하여 정답 후보 문서를 검색, 추출하게 된다.

[그림 1] 시스템 구성도

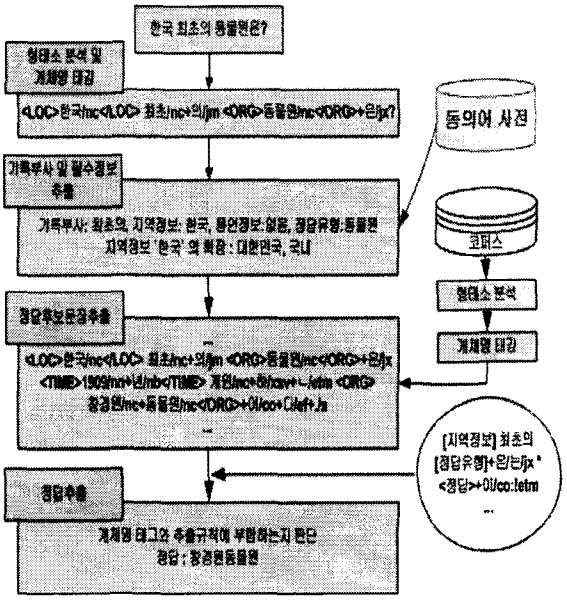


정답 후보 문서 검색 시 기록부사 그룹1의 경우 "지역정보|기록부사|정답유형"의 순서를 가지는 문장을 우선적으로 추출하게 되며, 기록부사 그룹2의 경우는 "지역정보|기록부사|용언정보|정답유형"의 순서를 가지는 문장을 우선적으로 검색하게 된다. 해당 규칙의 순서에 맞지 않을 경우 순서에 관계없이 "기록부사", "지역정보", "정답유형", "용언정보"의 출현 여부로 추출하게 된다.
정답 후보 문장을 추출한 후에는 각각의 정답 후보 문장에 정답 추출 규칙을 적용하여 해당 규칙에 맞는지를 검토하고 개체명 태깅을 통하여 정답유형과 같은 유형의 개체명 태그를 가지는 단어를 정답으로 제시한다.
정답 추출의 과정은 그림 2에 예시하였다.
정답 추출 규칙은 [7]의 정의문 규칙을 적용하여 작성하였으며, 기록정보를 갖는 문장을 분석한 결과를 토대로 하여 규칙을 생성하였다. 기록부사 그룹1의 규칙은 23개, 기록부사 그룹2의 규칙은 20개를 작성하였으며 출현 빈도수가 높은 순서대로 우선순위를 적용하였다.

다음은 기록부사 그룹1의 "최초의" 및 기록부사 그룹2의 "가장"에 대한 정답 추출 규칙의 일부이다.

"최초의" 정답 추출 규칙
1. [지역정보] 최초의 [정답유형]+은/는/jx * <정답>+이/co:!etm
2. [지역정보] 최초의 [정답유형]+이/co+ㄴ/etm <정답>+을/를/jc
3. <정답>+이/가/jc [지역정보] 최초의 [정답유형]+이/co:!etm
4. ...

[그림 2] 정답 추출 과정



"가장" 정답 추출 규칙
1. [지역정보]|+에/에서/jc 가장 [용언정보]/pa|pv+etm [정답유형]+은/는/jx * <정답>+이/co:!etm
2. [지역정보]|+에/에서/jc 가장 [용언정보]/pa|pv+etm [정답유형]+이/co+ㄴ/etm * <정답>+이/co:!etm
3.<정답>+이/가/jc [지역정보]+에/에서/jc 가장 [용언정보]/pa|pv+etm [정답유형]+이/co:!etm
4. ...

기록부사 그룹 1을 포함한 정답 후보 문장 "한국 최초의 동물원은 1909년 개원한 창경원동물원이다."는 기록부사 "최초의"의 첫 번째 정답 추출 규칙을 적용할 수 있고, 또 다른 후보 문장 "1883년(고종 20) 박영선(朴永善)과 함께 박문국(博文局)을 설치하고 일본인 이노우에 가쿠고로[井上角五郎(정상각오랑)]를 초빙하여 한국 최초의 신문인 한성순보를 간행하였으며, ..."에는 "최초의"의 두 번째 정답 추출 규칙을 적용할 수 있다.

기록부사 그룹 2를 포함하는 정답 후보 문장 "세계에서 가장 높은 산은 에베레스트산이다."에서 "가장"의 정답 추출 규칙 중 첫 번째 규칙을 적용할 수 있다. "한국에서 가장 큰 새인 황새는 두루미에 비해 크며,..."의 문장은 "가장"의 정답 추출 규칙 중 두 번째 규칙을 적용하여 정답을 추출할 수 있다.

정답 추출 규칙 중 '*' 기호는 한 문장 안에서 임의의 단어들이 올 수 있다는 의미이며, ':!etm'은 관형형 어미가 붙을 수 없다는 제약조건이다. 용언정보 형태소분석 태그 'pa' 또는 'pv' 뒤에 붙은 '+etm'은 용언에 관형형어미가 붙을 수 있다는 것을 의미한다.

4. 실험 및 결과

실험은 기록정보 질의-응답 시스템용으로 작성된 800여개의 질의문 셋[4] 내에서 기록부사 그룹 1에 해당하는 질의문 250여개 중 90개를 이용하였고, 기록부사 그룹 2에 해당되는 질의문 550여개 중 80개를 이용하였다.

각 기록부사 그룹별로 각각의 질문에 대해 정답을 포함하는 문서 5개, 정답을 포함하지 않는 문서 3개를 해당 질문에 대한 말뭉치로 구축하였다[5].

기록부사 그룹1의 90개 질의문 중 50개, 기록부사 그룹2의 80개 질의문 중 50개는 각 질의문에 대한 정답 후보 문서와 함께 시스템 구축 및 정답 추출 규칙을 작성하는데 사용된 학습 말뭉치이고, 나머지 40개 및 30개의 질문과 각 질문의 정답 후보 문서들은 시스템의 성능을 측정하는데 사용된 비학습 말뭉치이다.

실험에 사용된 정답 후보 문서들은 파스칼 전자 백과사전에서 기록정보 질의에 응답이 가능한 문장과 인터넷 웹 검색을 통하여 수집하였다.

[표 3] 제안한 시스템의 실험 결과

기록부사	말뭉치	질문수	정답 제시	정답	재현율(%)	정확률(%)
그룹1	학습	50	45	41	90.0	**91.1**
	비학습	40	24	16	60.0	66.7
그룹2	학습	50	40	34	80.0	**80.5**
	비학습	30	17	10	56.7	58.8

비학습 말뭉치의 실험 결과는 학습 말뭉치에 대한 결과보다 좋지 않은 결과를 보였다.

기록부사 그룹 1의 오류의 예로 "(거짓말쟁이) ..., 사건 본위에서 완전히 벗어난 것은 아니나, 프랑스 최초의 성격 희극의 걸작이다."에서 정답 유형이 여러 어절에 걸쳐서 나타나는 경우 즉, 복합명사나 구의 형태로 이루어진 경우 정답유형의 처리가 이루어지지 않아 올바른 정답을 추출할 수 없었다.

다른 예로, 질문 "한국 최초의 종합경기장은?"에 대한 정답 후보 문서 검색에서 "국민들은 몬트리올 올림픽에서 양정모 선수가 한국 최초의 금메달을 땄을 때만큼이나 ...벌써부터 나는 또 다시 태극기를 들고 종합경기장 사거리로 나가 목이 쉬도록 '대~한민국'을 ..."의 문장을 보면 정답 추출에 필요한 '기록부

4) 한국전자통신연구원에서 제공한 기록부사 그룹1에 대한 질의문 215개, 기록부사 그룹2에 대한 질의문 200개를 확장하여 만든 질의문 셋(Set)

5) 인터넷 검색을 통해 정답 문서를 수집하였으나 질문당 많은 정답 문서가 발견되지 않았다. 각 질문당 정답문서의 개수는 1~7개로 규칙적이지 못하였으며, 이에 각 질문당 추출된 정답문서 평균값인 5를 정답문서의 개수로 정하였다.

사', '지역정보', '정답유형'이 하나의 문장 내에 함께 출현하지 않고 문장 또는 단락을 달리하여 나타는 경우 정답을 제시하지 못하거나 오답을 출력하였다.

기록부사 그룹 2의 오류의 예로 "(건고) 지상 높이 4.15m, 지름 1.6m, 길이 1.49m로 한국의 국악기 가운데 가장 크고 화려하다." 문장의 경우 용언정보가 "크고 화려하다"의 두 가지가 나오게 되어 이에 대한 올바른 처리가 이루어지지 않았고, 다른 예로 "(울름대성당) 독일 고딕건축의 대표작 가운데 하나로, 성당 서쪽 정면에 있는 세계에서 가장 높은 첨탑으로도 유명하다." 의 문장에서 정답유형은 '첨탑'이 되지만 그에 해당하는 정답이 없어 오답을 제시하였다.

또한 다른 이유로 정답 추출 규칙의 적용 우선순위로 인해 적합한 추출 규칙이 아닌 다른 정답 추출 규칙이 적용되어 오답이 제시되거나 정답 추출 규칙이 없는 경우, 정답 추출 규칙의 제약 조건이 부족하여 오답을 추출하였다.

5. 결론 및 향후 연구

본 논문에서는 일반적인 질의-응답 시스템의 서브시스템으로 사용될 수 있으며, 기록정보 질의에 질문 규칙을 이용하여 정답을 추출하는 기록정보 질의-응답 시스템에 대하여 기술하였다.

기록정보 질의는 일정한 형태로 구성되며 "지역정보|기록부사|정답유형", "지역정보|기록부사|용언정보|정답유형"의 순서가 가장 많았다. 용언이 포함되는 질의문의 경우 목적어나 장소를 나타내는 부사구가 포함되기도 하였으며 이에 대한 처리는 향후 연구로 남긴다.

질의 분석을 통해 기록부사, 지역정보, 정답유형 및 용언정보를 결정하고 정답 추출 모듈에서 질의 분석 모듈로부터 넘겨받은 정보를 이용하여 정답 후보 문서를 검색한 다음 정답 추출 규칙을 이용하여 정답을 찾아낼 수 있었다. 따라서 제안한 시스템을 일반 질의응답 시스템의 서브시스템으로 사용한다면 질의-응답 시스템의 성능을 조금 더 향상시킬 수 있을 것이다.

향후연구로 정답 추출 규칙과 제약의 확장, 수식어구(복합명사)의 처리, 용언을 포함하는 질의의 경우 목적어와 장소를 나타내는 부사구 처리에 대한 연구가 진행되어야 할 것이다.

참고 문헌

[1] 강유환, 안영민, 서영훈, "개념 기반 질의-응답 시스템에서의 개념 규칙을 이용한 해답 추출", 제17회 한글, 언어, 인지 학술대회 발표논문집, pp.184-187, 2005

[2] TREC(Text REtrieval Conference) : http://trec.nist.gov

[3] ETRI AnyQ QA System : http://anyQ.etri.re.kr

[4] 이충희, 오효정, 김현진, 장명길, 템플릿에 기반한 기록정보 QA, 2005 한국컴퓨터종합학술대회 발표논문집, 2005

[5] 파스칼 전자 백과사전 : http://www.epascal.com

[6] 김남철, 서영훈, "한국어 형태소 분석기 CBKMA와 색인어 추출기 CBKMA/IX, 제11회 한글 및 한국어 정보처리 학술대회 및 제1회 형태소 분석기 및 품사태거 평가 워크숍 발표 논문집, pp.50-59, 1999

[7] 고병일, 강유환, 신승은, 서영훈, "질의응답시스템을 위한 서술형 정답 추출", 제16회 한글, 언어, 인지 학술대회 발표논문집, pp.303-307, 2004