

# 질의 응답형 정보검색 기술 동향

한국전자통신연구원 오효정 · 장명길  
한국정보통신대학교 맹성현\*

## 1. 서 론

인터넷 기술의 발전과 함께 축적된 방대한 정보로 인해 사용자는 원하는 정보를 이전보다 더욱 정확하고 빠르게, 더불어 보다 편리하게 얻을 수 있는 새로운 정보검색 기술을 기대하고 있다. 정보검색 기술의 범주에는 전통적인 키워드 기반, 교차언어검색, 자동분류, 사전탐지/추적, 자동요약 등에 걸쳐 다양한 기술이 포함되나 이들은 모두 처리 결과로서 문서 혹은 문단과 같이 텍스트 덩어리 집합을 제공한다. 사용자는 이러한 집합을 통해서 포괄적인 정보를 얻기도 하고 필요한 답을 찾아 내기도 한다. 이에 반해 질의 응답형 정보검색 기술(이하 질의 응답 기술)은 사용자 질문의 의도를 세밀하게 파악한 후, 검색 대상 문서로부터 답을 찾아 제공한다. 정보의 홍수 속에 살고 있는 사용자들로서는 정확도가 높은 질의 응답 시스템의 출현을 기대하고 있고 이에 부응하여 새로운 기술이 개발되고 있다.

질의 응답 기술은 사용자의 자연어 질문과 검색 대상 문서의 의미를 파악하기 위한 고정밀 자연어처리 기술과 대상 문서로부터 답을 추출하기 위한 정보추출 기술을 필요로 하며, 많은 후보 문서들로부터 답을 포함하는 문서를 걸러주는 역할을 위해 기존의 문서검색 기술도 활용한다. 질의 응답 기술은 현재 정형 데이터베이스에 담겨 있는 상품 정보나 기업 정보를 찾아주는 데이터베이스 질의 응답과 게시판 혹은 전자메일 정보를 찾아주는 FAQ(Frequently Asked Question) 질의 응답 등에 실제로 활용되고 있다. 하지만, 백과사전이나 전자 매뉴얼 혹은 신문 등의 일반적인 텍스트로부터 사용자 요구에 적합한 정답을 찾아 주기 위한 질의 응답 시스템은 고정밀 자연어 처리 및 정교한 정보추출 등의 텍스트 마이닝 기술이 요구되므로 현재 국내외적으로 활발한 연구가 진행 중이다.

본 논문에서는 언어공학적 기술 관점에서 매우 다양한 고급 기술이 적용되어야 하고 그 응용분야가 광범위한 질의 응답 기술의 동향을 살펴보고, 기술적인 특성을 비교 분석할 수 있도록 현재 개발된 주요 질의 응답 시스템을 간략히 소개한다. 마지막으로 모바일 환경에 적합한 차세대 질의 응답 기술인 음성 질의 응답 기술과 대표 시스템을 소개한다.

## 2. 질의 응답 기술 동향

### 2.1 질의 응답 기술

기존의 정보검색은 대량의 문서에서 사용자의 질의에 적합한 관련된 정보를 찾아내는 문서의 검색에 초점을 맞추어 왔다. 그러나 검색 대상 문서의 양이 많아짐에 따라, 검색의 결과로 나타나는 문서의 양은 검색 시스템 사용자가 필요한 정보를 얻어 내는데 큰 부담이 되었다. 따라서 가장 적합하고 정보의 가치가 높은 문서를 상위에 올려 주기 위해 문서 간의 하이퍼링크를 활용하는 기술(예: 구글의 PageRank<sup>tm1)</sup>)이나 논문, 뉴스, 상품, 인물 등의 전문 분야 검색기술을 통해 사용자의 편의성을 제공해 왔다.

정보검색의 기술적 한계에 직면한 몇몇 검색 포털 업체들은 새로운 검색 서비스 모델로 "지식검색"이라는 개념을 도입하였는데, 2003년도부터 국내 검색시장의 대표적인 서비스로 자리매김하게 되었다[1]. 지식검색은 알고리즘에 의존하는 기존 검색방법과 달리 포털 사용자의 참여를 활용한다. 사용자의 질문에 다른 사용자가 답을 달고, 많은 사람들이 이 지식의 유용성에 대해 평가함으로써, 네티즌들이 지식을 공유하는 수단으로 크게 각광을 받고 있다. 대부분의 국내 인터넷 포털 사이트 업체는 지식검색을 중요한 서비스의 하나로 제공하고 있는데, 네이버(www.naver.com)의 '지식iN', 엠파스(www.empas.com)의 '지식거래소', 세이클

\* 종신회원

1) [http:// www.google.com/technology/](http://www.google.com/technology/)

럽(www.sayclub.com)의 '세이테마', 야후(www.yahoo.co.kr)의 '야후!지식검색' 등이 널리 알려져 있다.

이러한 지식검색 서비스의 장점은 실생활에 도움이 되는 실용적인 질문과 답변이 많고, 정보의 시사성에 민감하다는 것이다. 이런 장점에도 불구하고 지식검색이 가지고 있는 문제점은 다음과 같다[2]. 첫째, 정답을 정의하는 것이 명확하지가 않다. 즉 네티즌 다수가 인정하면 '정답'으로 채택되고, 질문자가 답변을 선택하게 되면 그것이 마치 실제 정답인 것처럼 왜곡될 수 있다. 둘째, 방대한 양의 질문과 답변 중에서 신변잡기적인 내용이 많으며, '펼(펼오기2)'을 거듭하면서 출처가 불분명해진 정보들은 지식검색의 신뢰도를 크게 떨어뜨린다. 셋째, 흥미를 끌기 힘든 주제의 질문에 대해서 충실한 답변을 기대하기 어렵다.

질문에 대한 정답을 사람이 제시하는 지식검색과는 달리, 질의 응답 기술에서는 사용자의 질문에 대해 질의 응답 시스템이 문서로부터 정답을 추출하여 사용자에게 제시한다. 질의 응답 시스템의 목적은 질문에 대해 이를 만족하는 문서의 리스트를 찾기보다는 정답을 추출해내는 것이다[3]. 이는 단순히 키워드에 기반한 정보 검색이 아니라, 사용자 질문의 정교한 분석, 검색 엔진을 통한 문서 및 정답후보단락 추출, 단락에서 원하는 정답을 추출하는 과정으로 이루어져 있다. 또한, 질의에 대한 정답이 사실에 기반한 정확한 정답인지를 검증하기 위한 방법도 연구되고 있다.

질의 응답 시스템과 일반 정보검색 시스템의 큰 차이점 중 하나는 자연어로 된 질문을 입력하여, 문서를 검색하는 것이 아니라 정답을 찾는 것에 있다. 이를 위해 질문의 처리과정에서 사용자가 원하는 정답이 무엇인지 질의의도를 파악할 수 있는 질의유형이나 키워드 등의 정보를 질의로부터 추출한다. 또한, 기존의 정보 검색 방법에 의해 질의와 유사한 문서를 추출하고, 문서에서 다시 정답을 포함할 가능성이 있는 단락을 추출한 후, 단락에서 질의유형과 동일한 개체를 찾아내어 사용자에게 정답으로 제시한다. 최근에는 단답(factoid) 뿐만 아니라 서술형 정답, 나열형 정답 등에 대한 연구도 활발히 진행 중에 있다. 일반적인 질의 응답 시스템의 구조는 그림 1과 같다.

각 모듈이 하는 일은 다음과 같다.

- 질문분석: 질문이 요구하는 정답의 유형을 결정하고, 검색엔진을 위한 키워드를 추출
- 문서검색: 키워드를 이용하여 관련 문서를 검색

2) 인터넷에서 다른 사람의 글이나 자료를 그대로 가져옴을 의미하는 용어로, '펼'의 줄임 말이다.

진을 통해 검색

- 단락검색: 검색된 문서에서 키워드와 관련이 있는 단락을 검색
- 정답추출: 단락에서 질문이 요구하는 정답유형과 관련이 있는 개체를 정답으로 추출하여 사용자에게 제시

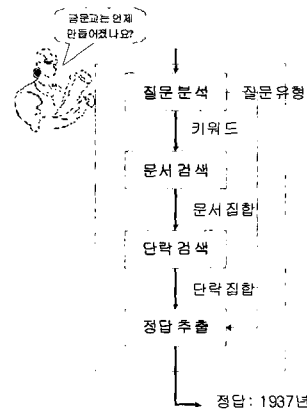


그림 1 일반적인 질의 응답 시스템의 구조

TREC<sup>3)</sup> QA(Question Answering)의 기술 로드맵[4]에 따르면, 시스템이 제시해야 할 정답은 단순히 문장의 일부만이 아니라 정의나 설명문과 같은 서술형 문장(descriptive answer), 여러 문서로부터 추출된 다중 정답(list answer) 등을 포함하여야 하고 이러한 기능을 평가할 수 있도록 평가의 범위를 확장해 나갈 예정이다. 더 나아가 문장들 사이의 추론이 필요한 질의를 처리하는 기능, 새로운 문장을 생성하여 이를 정답으로 제시하는 기능, 주어진 정답에 대한 배경 설명, 정답의 정당성 검증, 정답의 모호성 해결, 전문가 수준의 의견 제시, 이질적 정보의 통합을 통한 정답의 제시 등 점점 질의 응답 시스템의 난이도를 높여 갈 계획이다. 이때 가능한 질문의 형태는 아래와 같다.

- “AMTRAK의 금융적인 상황은 어떠한가?”
- “유로화에 대한 덴마크인들의 의견은 무엇인가?”
- “최근에 많은 사람들이 사륜구동차량을 구입하는 이유는 무엇인가?”
- “연방준비제도이사회가 다음 모임에서 금리를 인상할 것 같나요?”

다음 절에서는 질의 응답 시스템을 평가하기 위한 척도와 기준에 대해 다루기로 한다.

## 2.2 질의 응답 시스템 평가

질의 응답 시스템의 성능을 평가하는 대표적인 장소

3) TREC(Text REtrieval Conference)은 정보검색의 다양한 기술 및 알고리즘을 객관적으로 평가하기 위해 평가컬렉션을 개발하고 매년 참여하는 시스템을 새로운 데이터를 사용해서 평가한 후 그 결과를 공유하고 논의하는 학술대회임

로 미국 NIST(National Institute of Standards and Technology)가 주관하는 질의 응답 평가대회 TREC QA Track을 꼽을 수 있다. TREC에서 주관하는 다양한 분야 중 1999년부터 시작된 QA Track에서는 질의 응답 시스템을 평가하기 위해 대량의 평가 집합을 구축하고, 성능평가를 위한 평가척도를 개발하였다. QA의 평가를 시작한 첫 대회인 TREC-8 [3]에서는 약 53만 문서의 코퍼스로부터 수작업으로 작성한 200질문에 대해서 50바이트 및 200바이트 크기의 정답을 포함한 단락을 제시하도록 하였다. TREC-8에서 제시된 질문을 유형에 따라 분류하면 표 1과 같은 분포를 가진다.

표 1 TREC-8의 질문유형 분포

질문유형	질문수	질문유형	질문수
What	64	Which	10
Who	47	Name	4
How	31	Why	2
Where	22	Whom	1
When	19		

TREC-8에 제시된 실제 질문의 예는 다음과 같다.

“Who is the author of the book, ‘The Iron Lady: A Biography of Margaret Thatcher’”

“What was the monetary value of the Nobel Peace Prize in 1989?”

“What does the Peugeot company manufacture?”

“How much did Mercury spend on advertising in 1993?”

TREC-9에서부터는 검색엔진의 실제 로그를 이용하여 질문을 작성하였는데, TREC 2001의 경우, 약 25% 정도가 정의(definition)에 관련된 질문이었다. 질의 응답 시스템의 성능을 평가하기 위한 척도로 MRR(Mean Reciprocal Rank)을 사용하였다. MRR은 질의 응답 시스템에서 결과로 나타난 것에서 정확한 답이 포함되어 있는 것이 몇 번째 순위에 나타나는지를 계산해서 그 순위의 역수를 평균한 것이다. 아래 식에서  $rank_i$ 는  $i$ 번째 질문에 대한 응답으로 제시된 것들 중에서 첫 번째로 정답인 것의 순위이고,  $n$ 은 질문의 수이다.

$$MRR = \frac{1}{n} \left( \sum_i \frac{1}{rank_i} \right)$$

TREC 2002에서는 정답을 포함한 문자열을 제시하는 것이 아니라 실제 정답을 제시하도록 요구하였고,

CWS(Confidence-Weighted Score)를 시스템에 대한 새로운 평가척도로 사용하였는데, 이의 정의는 아래와 같다[5].

$$\frac{\sum_{i=1}^n i_c / i}{n}$$

여기서  $i_c$ 는 첫  $i$ 번째 질문에서의 정확한 답의 수를 의미한다. 표 2는 2002년 TREC QA 경쟁 부문에 참가한 시스템의 성능을 CWS로 평가한 결과를 예로 보여준다.

표 2 TREC-11의 성능 평가 결과

Run Tag	Confidence weighted Score	Correct Answers # %		Number Inexact	NIL Accuracy Prec Recall	
		#	%		Prec	Recall
LCCmain2002	0.856	415	83.0	8	0.578	0.804
exactanswer	0.691	271	54.2	12	0.222	0.848
pris2002	0.610	290	58.0	17	0.241	0.891
IRST02D1	0.589	192	38.4	17	0.167	0.217
IBMPSQACYC	0.588	179	35.8	9	0.196	0.630
uwmtB3	0.512	184	36.8	20	0.000	0.000
BBN2002C	0.499	142	28.4	18	0.182	0.087
isi02	0.498	149	29.8	15	0.385	0.109
lmsiQalir2	0.497	133	26.6	11	0.188	0.196
ali2002b	0.496	181	36.2	15	0.156	0.848
ibmsqa02c	0.455	145	29.0	44	0.224	0.239
FDUT11QA1	0.434	124	24.8	6	0.139	0.957
aranea02a	0.433	152	30.4	36	0.235	0.174
nuslamp2002	0.396	105	21.0	17	0.000	0.000
pqas22	0.358	133	26.6	11	0.145	0.674

### 3. 주요 질의 응답 시스템

#### 3.1 FAQ 시스템

질의 응답 기술은 현재 인터넷 쇼핑물에서 데이터베이스에 담겨있는 상품정보 등을 찾아주거나 고객의 소리와 같은 게시판 또는 FAQ 등을 검색하는데 부분적으로 실용화되어 있다. 문서검색이 문서로부터 필요한 정보를 찾는 것임에 반해, 데이터베이스에서 사용자가 원하는 자료를 검색하는 데이터베이스 자연어 질의처리 기술은 DB의 SQL과 같은 질의처리 언어를 자연어로 대신하도록 하는 것이다. 초창기의 대표적인 DB를 위한 자연어 인터페이스로는 BASEBALL[6]과 LIFER[7]가 있다. BASEBALL은 야구 통계에 대한 질의처리 시스템으로 “Who did the Red Sox lose to on July 5?”와 같은 자연어 질문을 처리할 수 있었으며, LIFER는 인사통계에 대한 질의처리 시스템으로 “How many professors are there in the computer science department?” 등에 대한 질문을 의미적 수준에서 파싱하고 이를 직접 DB 질의어로 변환하였다. 현재는 이러한 DB 질의처리 기술이 사용자의 자연스러운 대화체 질의 문장을 형태소, 의미, 구문 분석하여 사용자의 요구에 맞는 최적의 상품을 검색해 주는 상

품검색 및 추천 등에 이용되고 있다. 이러한 시스템은 텍스트에서 답을 찾아야 하는 근래의 질의 응답 시스템과는 기술적인 면에서 근본적으로 상이하다.

### 3.2 START

START[8, 9]는 1993년부터 제공된 온라인 질의 응답 시스템으로 구조화 또는 반 구조화된 웹 데이터들을 통합한 가상의 데이터베이스인 Omnibase를 이용한 자연어 질의 응답 시스템이다. 이 시스템은 다수의 웹사이트에 대한 단일한 질의 응답 인터페이스로 볼 수 있는데, World Factbook<sup>4)</sup>을 정보의 원천으로 하여 각국에 대한 정치, 경제, 지리적 정보에 대한 질의 응답을 제공하며, Biograph.com 사이트<sup>5)</sup>로부터 25,000여명의 인물에 대한 정보를, Internet Movie Database 사이트<sup>6)</sup>로부터 수십만 편의 영화에 대한 전반적인 정보를, 그리고 Merriam-Webster 사전<sup>7)</sup>을 이용하여 각종 정의와 관련된 정보 등을 제공한다. 이 시스템은 반 구조화된 웹페이지를 대상으로 한다는 면에서 근래의 질의 응답 시스템에 보다 가까이 접근해 있다.

START에서 질문 "Who won the Nobel Peace Prize in 2000?"에 대한 정답으로 그림 2를 제시하고 있다.

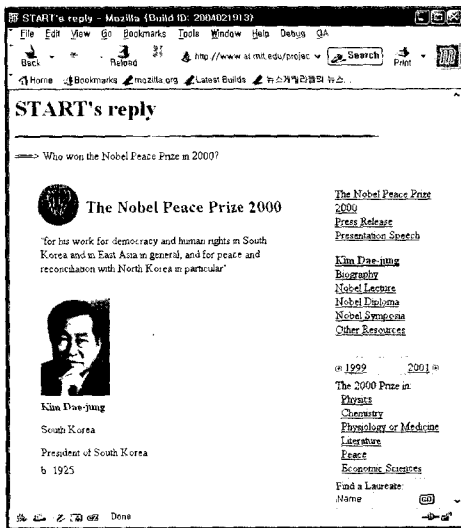


그림 2 START의 정답제시 화면

### 3.3 FALCON

FALCON[10, 11]은 TREC-8 및 TREC-9 QA

4) CIA - The World Factbook, [www.cia.gov/cia/publications/factbook/](http://www.cia.gov/cia/publications/factbook/)  
 5) Biography.com, <http://www.biography.com>  
 6) Internet Movie Database, <http://www.imdb.com>  
 7) Merriam-Webster Dictionary, <http://www.m-w.com>

Track에서 텍스트로부터 단답형 정답을 찾아내는 작업에 있어 우수한 성능을 보여준 시스템으로 그림 3과 같은 구조를 가지고 있다.

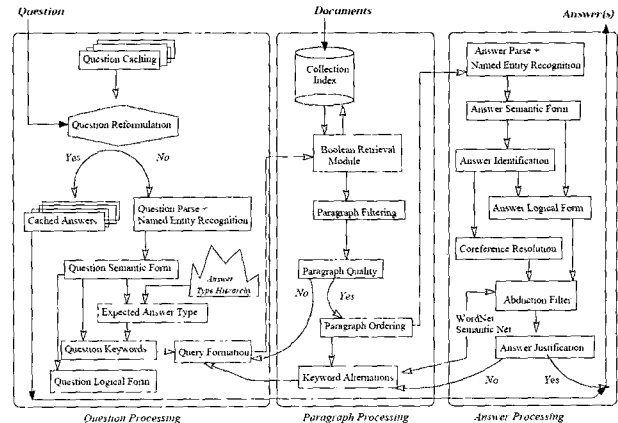


그림 3 FALCON의 구조

이 시스템의 주요 특징은 정답의 추출을 위해 질문 분석을 통해 정답유형을 정의한 것인데, 질의처리 과정에서 1) 질문의 유형 정의, 2) 예상되는 정답의 유형 정의, 3) 정답의 포커스 선택, 4) 검색 엔진을 위해 질문을 질의로 변환하는 역할을 수행한다. 예를 들어, "Which city has the oldest relationship as a sister-city with Los Angeles?"라는 질문의 분석 과정에서 얻어진 정보는 다음과 같다.

질문유형: Which-Where  
 정답유형: Location  
 포커스: city  
 키워드: city, oldest, relationship, sister-city, Los Angeles

정답처리과정에서는 정답유형의 대상이 되는 개체명(named entity), 화폐단위, 날짜, 시간/장소 표현, 상품명 등을 인식하기 위해 Lexico-Semantic 정보를 이용하였다. 이를 통해 얻어진 정답후보들을 대상으로 다양한 가중치 함수를 두고 이를 결합하여 정답을 추출하였다. 이때, 질문을 분석한 논리 형태(logic form)와 정답 문장을 분석한 논리 형태를 비교하여 정답을 추출하게 된다. 예를 들면,

질문: Who was the first Russian astronaut to walk in space?  
 질문 논리 형식:  $\text{first}(x) \wedge \text{astronaut}(x) \wedge \text{Russian}(x) \wedge \text{space}(x) \wedge \text{walk}(y, z, x) \wedge \text{HUMAN}(x)$   
 정답문장: The broad-shouldered but paunchy Leonman to walk in space, signed autographs.  
 정답 문장 논리 형식:  $\text{paunchy}(y) \wedge \text{shouldered}(e1 y x) \wedge \text{broad}(x) \wedge \text{Leonov}(x) \wedge \text{first}(z)$

$\sim$  man(z)  $\wedge$  space(t)  $\wedge$  walk(e2 t z)  $\wedge$   
 became(e3 z u x)  $\sim$  1965(u) autographs  
 (v)  $\wedge$  signed(e4 v x)  $\wedge$  HUMAN(x)  $\wedge$  DATE(u)

LCC(Language Computer Corp.)에서는 1999년 TREC-8에는 LASSO라는 이름으로 출전하여 MRR 0.58(정답문장50 byte 평가 결과)로 1등을 차지하였고[10], 2000년 TREC-9에 FALCON이라는 이름으로 출전하여 MRR 0.57 (단답 평가 결과)로 2등을 기록하였다[11]. 이후 보다 고성능 개체명 인식기와 정밀한 구문 분석 파서 개발, 정답 추론 기법 도입 등을 통해 매년 성능향상을 보여 2003년 단답형 질의 응답 부분에서 0.70을 획득하여 1등을 차지하였다[12].

### 3.4 NUSCHUA

앞서 말한 TREC QA 기술 로드맵[4]에 따라, 초창기 TREC이 단답형(factoid) 질의 응답에 집중하였다면 2001년부터는 나열형(list) 정답이나 정의형(definitional) 정답과 같은 다른 형태의 질의 응답 트랙을 제시하였다. 나열형 정답은 여러 정답 문서에서 복수개의 정답을 추출해서 제시하는 경우를 말하는데, 예를 들면 "G8 국가는 어디어디 인가?"에 대한 정답으로 8개 국가명을 제시하는 경우이다. 정의형(definitional) 정답이란 어떤 개념이나 개체의 정의문, 설명문을 정답으로 제시하는 것으로, "헬렌 켈러가 누구예요?"와 같은 질문에 대한 정답을 찾는 경우이다. 초기인 2001년도에는 참가한 시스템이 별로 없었지만, 2003년부터 각 기관의 활발한 참여로 2004년, 2005년 확대되어가고 있는 추세이다.

싱가폴 국립 대학에서 개발한 NUSCHUA[13]은 2004년 나열형 정답 및 서술형 정답 부문에 참가하여 각각 2위와 1위를 차지한 시스템으로, 그 후 2005년에도 높은 순위에 랭크 되었던 시스템이다[14]. NUSCHUA는 정답을 찾기 위해 정답 문장에 나타나는 용어간의 의존 관계(dependency relation)를 참조하여 정답간의 유사도를 비교하였는데, 두 단어 간의 의존 관계를 다음과 같은 트리플(triple) 형태로 표현하여 사용하였다. 여기서 패스는 구문 분석 결과 트리에서의 두 단어 간의 경로를 의미하는 것으로, 아래 예제는 'Benedict Arnold'와 'West Pont'의 관계가 주어와 목적어 관계에 있음을 표현한다.

〈Slot1, Path, Slot2〉

정답 문장: "Benedict Arnold's plot to surrender West Point to the British."

의존 관계 예) 〈Benedict Arnold, pass s obj, West Point〉

NUSCHUA는 정의형 질의 응답을 위해 외부 자원을 활용하였는데, Biograph.com<sup>8)</sup>이나 Wikipedia<sup>9)</sup>, Google Glossary<sup>10)</sup> 등에 기술된 설명문에서 주어진 주제에 해당한 문장을 검색하는 방법으로 정의문을 추출하여 정답으로 제시하고 있다.

### 3.5 AnyQuestion 3.0

AnyQuestion[15]은 ETRI에서 개발한 백과사전 대상의 질의 응답 시스템으로, 본 절에서는 2005년도에 개발된 백과사전 및 확장영역 질의 응답 시범 시스템 AnyQuestion 3.0<sup>11)</sup>을 소개한다.

AnyQuestion 3.0은 '파스칼대백과(Pascal<sup>tm</sup>)<sup>12)</sup>'를 기반으로, 14개 분야의 총100,373 표제어에 관한 사용자 질문에 대해 질문의도를 파악하여 정답을 제시한다. AnyQuestion 3.0은 추출된 정답뿐 아니라 정답을 추출하게 된 근거가 되는 문장을 함께 제시하여 사용자가 제시된 정답의 정확성 유무를 확인할 수 있게 하며, 현재 질문과 관련된 추가 질문들을 추천하는 기능 및 백과사전 인물에 대한 주요한 정보를 간략히 제시하는 기능 등을 제공하고 있다. 주요한 특징은 다음과 같다.

- 사용자 질문 의도에 따른 다양한 정답 추출 및 제시
- 백과사전 지식베이스(KB) 반자동 구축
- 정답관리자를 통한 복잡한 질문 해결 및 정답 필터링 기능 강화

위 기능을 수행하기 위해서는 기본적으로 고정밀 언어 분석을 위한 고성능 형태소 분석기, 개체명 인식기, 구문 분석기 등이 필요하다. 뿐만 아니라 백과사전 문장에서 공통된 지식을 추출하여 구조화된 지식베이스 추출을 위해서는 정보 추출(Information Extraction) 기법이 필요하고, 이를 학습하기 위한 기계학습(machine learning) 기술과 학습에 필요한 수작업을 줄이기 위한 반자동 구축 도구 등이 필요하다.

AnyQuestion은 일반적인 단답형(factoid) 질의 응답뿐 만 아니라 서술형(descriptive) 질의에 대한 정답 및 나열형(list) 정답도 제시하는데, 서술형 질의란 정답이 한 단어 즉 단답으로 제시될 수 없는 경우로 TREC에서 제시하고 정의형(definitional) 정답에만 국한되는 것이 아니라, 사건이 일어나게 된 경위, 현상에 대한 설명문과 같이 서술형 문장으로 정답을 요구

8) <http://www.biograph.com>

9) [http://en.wikipedia.org/wiki/Main\\_Page](http://en.wikipedia.org/wiki/Main_Page)

10) Search by "define:<term>" in Google

11) AnyQuestion, <http://anyq.etri.re.kr>

12) Pascal tm, [www.epascal.co.kr](http://www.epascal.co.kr)

하는 질문을 말한다. 예를 들면 “하늘은 왜 파랗요?” 나 “감기에 걸리지 않으려면 어떻게 해야 하죠?”과 같은 질문이 그것이다. 서술형 정답을 추출하기 위해서는 서술형 문장에 나타나는 구문(syntactic) 및 의미(semantic) 패턴을 인식하고 이를 적용하는 패턴 매치(pattern matching) 기술이 필요하다. 나열형 정답을 추출하기 위해서는 병렬형 구문을 분석하여 생략된 구문을 파악하고, 여러 문서에 중복으로 나타난 정답을 통합 관리하는 기술이 필요하다. 이밖에 AnyQuestion의 또 다른 특징은 정답관리자를 통해 보다 복잡한 질문을 해결하기 위한 연구가 시도되었는데, 예를 들면 “한국인으로 노벨상을 수상한 사람은?”의 질문의 경우, 노벨상 수상자를 먼저 찾고, 이후 국적이 한국인인 사람을 찾아서 최종 ‘김대중 전 대통령’이라는 답을 제시한다. 또한, 정보추출 기술을 활용한 지식베이스(KB)를 미리 구축함으로써 사용자에게 보다 빠르고 정확한 정답을 제시한다.

그림 3은 사용자 질문에 대한 AnyQuestion의 정답제시 화면이다.

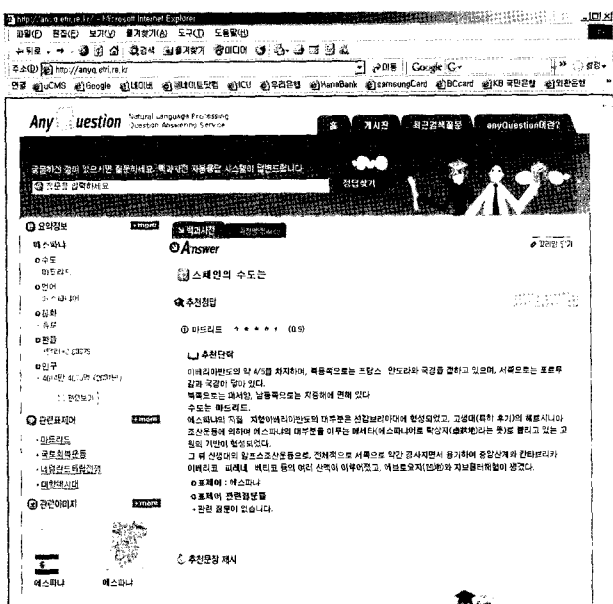


그림 3 AnyQuestion 시스템 정답 제시 화면

### 3.6 국내 주요 질의 응답 시스템

[16]에서는 여러 개의 관계형 데이터베이스(relational DB)로부터 추출하는 다국어 질의 응답 시스템에 대해 소개하고 있는데, 다국어 질의 응답 시스템이란 사용자가 입력한 질의를 외국어로 번역한 후 이를 외국어 문서에서 정답을 검색하는 시스템을 말한다. 사용자에게 제시되는 정답은 다시 원 국어로 번역하여 제공하기도 한다. 이 시스템에서는 자연어 질의를 분석하여 개체명 사전과 Lexico-Syntactic Pattern을 이용

하여 개체명을 인식하고, 이를 데이터베이스 질의(Data Base Query)로 변환, 이를 번역하여 정답DB에서 정답을 찾아준다. 자연어 질의 전체를 번역하는 방법에 비해 데이터베이스 질의만을 변환하여 사용함으로써 확장성(portability) 향상을 꾀하였다.

MAYA[17]라는 질의 응답 시스템에서는 개체명 사전과 LSP(Lexico-Syntactic Pattern)을 이용하여 개체명을 인식하고, 이를 질의 응답 시스템의 정답 가능한 후보로 미리 색인하였다. 사용자의 질의유형을 105가지의 의미범주로 구분하고, 이에 따라 정답유형을 분류하였으며, Lexico-Syntactic Parser를 이용하여 사용자의 질의유형을 분석하여 색인된 정답DB에서 정답후보를 순위화하고 이를 정답으로 제시하였다.

[18]에서는 한국어 질의 응답을 위한 시스템으로, 질의해석을 통해 낱자, 범위, 핵심어, 중요어, 질의유형을 파악하고, 질문해석 과정에서 얻어진 정보를 이용하여 문서에서 정답을 포함할 것 같은 단락을 검색하며, 대답 추출에서 여러 언어자원을 이용하여 개체명을 추출한 후, 비사실 문맥에서 나타나는 정답제거, 정답 확인 및 정답해석을 통해 정답을 찾아낸다. 주요한 특징으로 메타-사실 프레임의 통해 대답을 확인하는 부분이 있는데, 질문에 나타난 어휘를 모두 포함하지 않더라도, 메타-사실 프레임의 만족하는 패턴을 이용하여 정답에 대한 확신을 증가시킨다. 예를 들어, “허준의 동의보감”은 ‘<PERSON>의 <BOOK, PEOM, WRITING>’이라는 프레임을 통해 “허준이 ‘동의보감’을 쓴 저자임을 검증하는 것이다.

[19]에서는 정의형(definitional) 질문에 특화된 시스템을 제안한다. 정의형 질문의 의도는 “질문 대상에 대한 정의”에 집중된 질문으로, 이에 대한 정답 문장의 특징은 정의형 표현을 내포하고 있다는 점이다. 이러한 언어 모델 특징을 확률 모형으로 모델링하고, 정의형 표현의 문장 패턴을 활용함으로써 정의형 질의 응답 시스템의 성능 향상을 꾀하였다. 예를 들면, 영어의 관계 대명사 구문이 주어 수식하는 구문이라는 특성에 착안하여 다음과 같은 문장을 ‘Copland’를 설명하는 정답 문장으로 추출하게 된다.

질문: Who is Copland?”

정답: Copland, who was born in Brooklyn, would have turned 100 on Nov. 14, 2000.

### 4. 모바일 환경을 위한 음성 질의 응답 기술 동향

기존 질의 응답 기술은 텍스트 형태의 사용자 질문 입력을 대상으로 발전되어왔다. 그러나 최근 휴대폰이나 PDA와 같은 개인 단말기 보급이 폭발적으로 증가

함에 따라 이러한 휴대장치를 이용하여 인터넷에 접속하고 정보를 얻으려는 사람들의 수도 급속도로 증가하고 있다. 그러나 사용자들에게 제공되는 정보검색 기술들은 여전히 모바일 환경의 특성을 반영하지 못하고 있는 실정이다. 모바일 환경에 적합한 정보검색 기술을 개발하는데 있어서 선결해야 할 기술적 요소는 정확한 음성 입력 인식과 실시간 정보 제공, 효율적인 정답제시로 사용자 만족도를 높이는 데 있다[20]. 이러한 요구사항을 만족시키기 위해서는 높은 정확도를 갖는 음성 질의 응답 기술 개발이 필요하다.

일반적인 질의 응답 시스템은 의도분석, 개체명 인식을 비롯한 다양한 언어처리 기술을 이용하여 질의어와 의미적으로 일치하는 정답 후보들을 찾아 제시해준다. 그러나 기존의 질의 응답 시스템들이 모바일 정보검색에 활용되기 위해서는 다음과 같은 두 가지 문제점들을 해결해야 한다. 첫째, 입력 공간의 제약으로 인하여 질의어가 짧을 가능성이 매우 커져서 질의 의도를 파악하는 것이 상대적으로 어렵다. 이러한 문제점을 해결하기 위해서는 음성으로 질의를 입력 받는 것이 고려될 수 있다. 그러나 70~80%의 음성인식 성공률을 보이고 있는 현재 기술 수준에서 기존의 음성인식기와 질의 응답 시스템을 단순히 결합하는 것은 오히려 나쁜 결과를 초래할 수 있다. [21]에서는 30%의 단어 인식 오류를 가지고 있는 음성인식 시스템과 76%의 정확률을 보이는 질의 응답 시스템을 단순 결합하였을 때 전체 시스템 성능은 7% 밖에 되지 않는다는 것을 보였다. 그러므로 질문 유형이나 구문, 의미, 담화 정보 등을 이용하여 음성인식 오류를 후처리하는 다양한 방법들이 연구되어야 한다[21,22]. 둘째, 출력 공간의 제약으로 인하여 질의 응답 시스템이 추천한 정답 후보들을 포함하는 문맥이나 원문을 효과적으로 보여주기 어렵다. 이러한 문제를 해결하기 위해서는 문서요약 기술을 활용하여 정답과 관련된 핵심 문장들을 찾아 제시하는 것이 필요하다. 다음은 음성인식 기술과 질의 응답 기술을 접목하여 차세대 음성 질의 응답 시스템을 개발한 사례를 다루기로 한다.

#### 4.1 VAQA

VAQA(Voice-Activated Question Answering) [21]는 음성인식(voice recognition) 기술과 기존 텍스트 기반 질의 응답 기술을 통합한 대표적인 시스템으로, 웹을 통해 입력된 음성 질의를 텍스트 질의로 변환하여 정답을 추출한 후 이를 음성 합성(speech synthesis) 기술을 활용하여 다시 음성으로 전달하는 시스템이다.

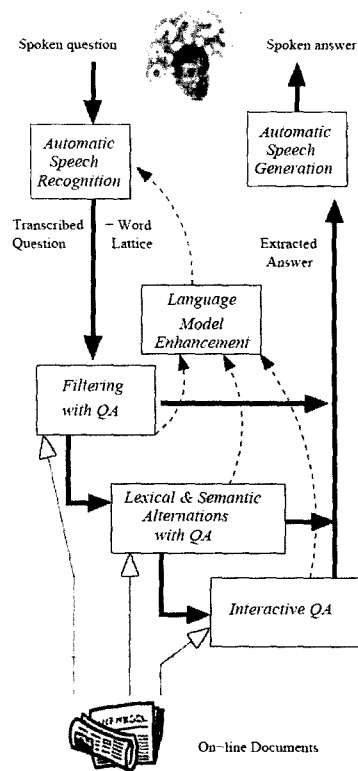


그림 4 VAQA 시스템 흐름도

이 시스템에서는 음성 인식 결과의 신뢰도가 낮은 경우, 이를 사용자에게 검증하도록 하는 간단한 대화 기능을 제공함으로써 음성 인식 오류를 최소화하였고 사용자와 시스템이 상호 교류(interactive)할 수 있도록 하였다. TREC-8[3]에서 제공하는 데이터를 대상으로 시스템을 평가하였다. 그림4는 VAQA의 시스템 흐름도이다.

#### 4.2 SPIQA

SPIQA(SPoken Interactive Question Answer) [22]은 일본에서 개발된 음성 질의 응답 시스템으로, 사용자 질의문은 문장의 의존 구조(dependency structure)를 활용해 템플릿(template) 형태로 분석된다. 주요한 특징으로는 적합한 정답을 찾을 수 없는 경우나 입력된 음성 질의를 분석한 결과에 애매성(ambiguity)이 발생한 경우, 이를 분석된 템플릿을 통해 재 질의를 생성한 후 이를 제시함으로써 사용자가 직접 오류를 걸러낼 수 있는 스크린 필터(screening filter) 방식을 제공하였다. 그림 5는 SPIQA의 시스템 구조와 데이터 흐름도를 나타낸다. 먼저 사용자의 질문이 음성으로 입력되면, 음성인식을 통해 질의문을 추출하고 적합한 정답을 추출하기 위해 애매성이 발견했는지를 판단한다. 애매성이 발견된 경우, 추가 정보를 활용해 질의를 재정의하고 정답을 추출하고, 추출된

정답을 기반으로 자연스러운 정답문장을 생성하여 음성합성을 통해 사용자에게 전달한다.

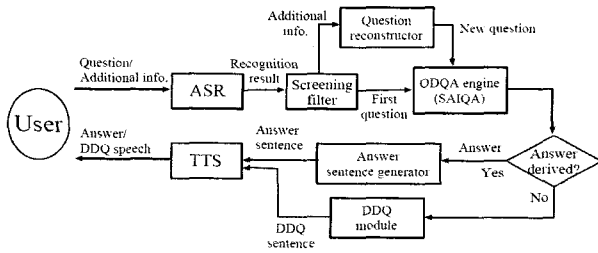


그림 5 SPIQA 시스템 흐름도

### 4.3 JUPITER

JUPITER(23)는 MIT에서 음성 정보 시스템 개발 프로젝트인 GALAXY의 일환으로 개발된 음성 질의 응답 시스템으로, VAQA나 SPIQA가 일반 모든 영역(open-domain)에 대한 질의 응답을 대상으로 한 반면, JUPITER는 특화된 영역(specific-domain)을 대상으로 한 질의 응답 시스템이다. 이 시스템은 전화를 통해 전 세계 500여개 도시의 기상 정보를 제공하는데, 1997년 5월부터 2년 동안 약 3만여 통의 전화 내용을 통해 18만 발화문을 수집하였다. 전화를 통해 입력된 음성을 분석하기 때문에 실생활에서 활용되는 축약어, 사투리 등에 대한 오류를 최소화하였으며, 제한된 영역의 질의 응답 기술을 통해 정확도(precision)이 매우 높은 정답을 자연어 문장으로 제시함으로써 시스템과 사용자가 대화할 수 있도록 한다. 그림 6은 사용자와 JUPITER 간의 대화문 예를 보여 준다.

### 5. 결 론

질의 응답은 기존의 문서 검색과는 달리 사용자의 질문에 대하여 정확히 답을 찾아 제공한다. 이를 위해서는 기본적으로 사용자의 다양하고 복잡한 질문 유형을 적절히 파악하고 검색 대상 문서의 의미적 내용을

정확히 분석하기 위하여 방대한 언어지식을 이용한 고 정밀 자연어처리 기술이 필요하며, 단답형, 문장형, 나열형 등의 여러 형태의 정답을 제시할 수 있어야 한다.

개인 단말기 보급과 유비쿼터스 환경에 대한 관심이 높아지면서 모바일 정보 검색 기술에 대한 요구도 높아져가고 있다. 모바일 환경에서의 실용적인 음성 질의 응답 기술 개발을 위해서는 정밀한 음성 인식 기술, 빠른 응답 속도, 편리한 정답 제시 방법 등이 연구되어야 한다.

앞으로 질의 응답 기술은 백과사전이나 전자 매뉴얼을 대상으로 한 질의 응답 외에도, 방대한 양의 웹 정보를 대상으로 한 웹 질의 응답, 다국어 텍스트 대상의 다국어 질의 응답, 텍스트 뿐만 아니라 멀티미디어 정보에 대한 멀티미디어 질의 응답, 그리고, 목적 지향의 대화 진행을 통한 대화형 질의 응답 등의 종류로 계속 연구가 진행되어, 다양한 분야에서 차세대 정보검색기술로서 사용자에게 보다 큰 만족을 제공할 것으로 기대 된다.

### 참고문헌

- [1] 황이규, 김현진, 장명길, "질의 응답 기술 개발", 정보처리학회지, 제11권, 제2호, pp. 48-56, 2004.
- [2] 이강룡, "지식검색이라는 자유공간", <http://www.hani.co.kr/section-001055000/2004/01/001055000200401081917067.html>
- [3] Ellen M. Voorhees, "The TREC-8 Question Answering Track Report," In Proceedings of TREC-8, 1999.
- [4] J. Burger, C. Cardie. et.al. "Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)", NIST DUC Vision and Roadmap Documents, <http://www-nlpir.nist.gov/projects/>

JUPITER:	Hi! Welcome to JUPITER, the MIT Lab for Computer Science weather information system. This call is being recorded for system development. You may hang up or ask for help at any time. How can I help you?
User:	CAN YOU TELL ME WHAT'S THE WEATHER LIKE IN BOSTON TODAY
JUPITER:	In Boston this afternoon. Partly sunny early. Then becoming cloudy with 30 percent chance of showers toward evening. Tonight: rain. Possibly heavy at times. Near 100 percent chance of rain. Can I help you with something else?
User:	YES WHAT ABOUT SAN FRANCISCO
JUPITER:	In San Francisco today: patchy fog near the ocean morning. Otherwise mostly sunny. Tonight: areas of coastal fog and low clouds. Otherwise fair. Is there something else I can do for you?
User:	YES HOW COLD IS IT IN DALLAS TEXAS TODAY
JUPITER:	In Dallas today: high upper 40s. Tonight: lows near 32. What else would you like to know?
User:	I'D LIKE TO KNOW WHETHER IT'S SNOWING IN DETROIT
JUPITER:	I'm sorry, I didn't understand you. Please rephrase your query.
User:	WHAT IS THE WEATHER IN DETROIT
JUPITER:	In Detroit this afternoon: occasional light rain and drizzle. Tonight: rain showers likely. Becoming snow showers by late tonight. 70 percent chance of precipitation.

그림 6 사용자와 JUPITER간의 대화 예제



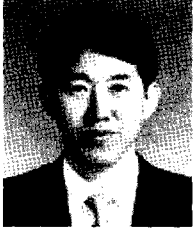
- duc/roadmapping.html, 2001.
- [5] Ellen M. Voorhees, "Overview of the TREC 2002 Question Answering Track." In Proceedings of TREC-11, 2002.
- [6] Bert Green, Alice Wolf, Carol Chomsky, and Kenneth Laughery, "BASEBALL: An Automatic Question Answerer," In Proceedings of the Western Joint Computer Conference, 1961.
- [7] Gary G. Hendrix, "Human Engineering for Applied Natural Language Processing," In Proceedings of the 5th International Joint Conference on AI(IJCAI-1977), 1977.
- [8] Boris Katz, Jimmy J. Lin, "START and Beyond," In Proceedings of 6th World Multiconference on Systemics, Cybernetics, Informatics(SCI 2002), 2002.
- [9] Jimmy Lin, Boris Katz, "Question Answering Techniques for the World Wide Web," In proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2003), 2003.
- [10] D. Donovan, S. Harabaqiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju, and V. Rus, "LASSO: A Tool for Surfing the Answer Net," In Proceedings of TREC-8, pp. 175-184, 1999.
- [11] S. Harabaqiu, D. Donovan, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju, and V. Rus, "FALCON: Boosting Knowledge for Answer Engines," In Proceedings of TREC-9, pp. 65-74, 2000.
- [12] S. Harabaqiu, D. Donovan, M. C. Clark, M. Mowden, J. Williams, J. Bensleny, "Answer Mining by Combining Extraction Techniques with Adductive Reasoning," In Proceedings of TREC-12, pp.65-74, 2003.
- [13] H. Cui, K. Li, R. Sun, T-S. Chua, M-Y. Kan, "National University of Singapore at the TREC-13 Question Answering Main Task," In Proceedings of TREC-12, 2004.
- [14] Ellen M. Voorhees, "Overview of the TREC 2005 Question Answering Track," In Proceedings of TREC-14, 2005.
- [15] H. J. Oh, C. H. Lee, C. K. Lee, and M. G. Jang, "Heterogeneous Answer Acquisition methods in Encyclopedia QA," LNCS vol. 4224(Intelligent Data Engineering and Automated Learning - IDEAL 2006) ,pp 346-354. 2006.
- [16] H. Jung, G G. Lee, W. Choi, K. M., and J. Seo. "Multi-lingual question answering with high portability on relational databases", IEICE transactions on information and systems, Vol E86-D, No 2, pp306-315, 2003.
- [17] 김학수, 서정연, "2-패스 색인 기법과 규칙 기반 질의처리 기법을 이용한 고속, 고성능 질의 응답 시스템", 정보과학회논문지: 소프트웨어 및 응용, 제29권 제11호, pp. 795-802, 2002.
- [18] 이경순, 김재호, 최기선, "KorQuA: 질의 응답에서 자료유형을 고려한 대담검색과 대담해석", 한글 및 한국어 정보처리학술대회, 2000.
- [19] K. S. Han, Y. I. Song, S. B. Kim, and H. C. Rim, "A Definitional Question Answering System Based on Phrase Extraction Using Syntactic Patterns," IEICE Transactions on Information and Systems, Vol. E89-D, No. 4, pp.1601-1605, April 2006.
- [20] 김학수, 장명길, "유비쿼터스 환경에서의 모바일 정보검색 기술의 동향", 정보과학회지, 제24권, 제1호, pp. 48-55, 2006.
- [21] Harabagiu S., Moldovan D., and Picone J., "Open-domain voice-activated question answering," In Proceedings of COLING, Vol. 1, pp. 1-7, 2002.
- [22] Hori C., Hori T., Tsukada H., Isozaki H., Sasaki Y., and Maeda E., "Spoken interactive ODQA System: SPIQA," In Proceedings of ACL-2003 Interactive Posters and Demonstrations, 2003.
- [23] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. J. Hazen, and L. Hetherington.. "Jupiter: A telephone-based conversational interface for weather information," IEEE Transactions on Speech and Audio Processing, vol. 8, no. 1, 2000.

오 호 정



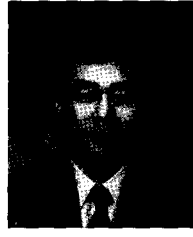
1998 충남대학교 컴퓨터과학과(학사)  
2000 충남대학교 컴퓨터과학과(석사)  
2000~현재 한국전자통신연구원 지식마이닝  
연구팀 연구원  
관심분야: 자연어처리, 정보검색, 질의 응  
답, 지식 및 대화 처리, 문서  
분류, 기계 학습, 지식 추출  
E-mail : ohj@etri.re.kr

장 명 길



1988 부산대학교 계산통계학과(학사)  
1990 부산대학교 계산통계학과(석사)  
2002 충남대학교 컴퓨터과학과(박사)  
1990~1997 시스템공학연구소 연구원  
1998~1999 한국전자통신연구원  
선임연구원  
2000~현재 한국전자통신연구원 지식마이닝  
연구팀 팀장(책임연구원)  
관심분야: 자연어처리, 정보검색, 질의 응  
답, 지식 및 대화 처리, 미디어 검  
색 및 관리, 시맨틱웹, 온톨로지  
E-mail : mgjang@etri.re.kr

맹 성 현



1983 미국 캘리포니아 주립대학(학사)  
1987 미국 Southern Methodist  
University (SMU) 석사 및 박사  
미국 Temple University 조교수  
Syracuse University 종신교수  
충남대학교 교수 역임  
2002 ACM SIGIR Conference  
Program Committee Chair  
현 재 한국정보통신대학교(ICU) 공학부  
교수

AIRS 2004 Program Committee Chair  
AIRS 2005 조직위원장  
Information Processing & Management  
Journal of Natural Language Processing  
Journal of Computer Processing of Oriental Languages  
편집위원  
관심분야: 정보검색, 텍스트마이닝, 디지털 도서관, 시맨틱웹  
E-mail : myaeng@icu.ac.kr

• 한국소프트웨어공학 학술대회 •

- 일 자 : 2007년 2월 22~24일
- 장 소 : 용평리조트
- 내 용 : 논문발표 등
- 주 최 : 소프트웨어공학연구회
- 상세안내 : <http://selab.skku.ac.kr/~kcse07>