

고속, 고성능의 한국어 질의 응답 시스템

다이퀘스트닷컴

intelligent web dialogue

interface solutions

°김학수*, 이근배**, 서정연***

* ㈜다이퀘스트 연구소

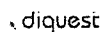
** 포항공대 컴퓨터공학과 자연어처리 연구실

*** 서강대학교 컴퓨터학과 자연어처리 연구실



Contents

- ▼ Introduction
- ▼ Indexing method
 - System architecture
 - Answer-finding
 - Calculation of scores
- ▼ Searching method
- ▼ Experiments
- ▼ Demo
- ▼ Conclusion



Introduction (3/3)

- Representative QA systems

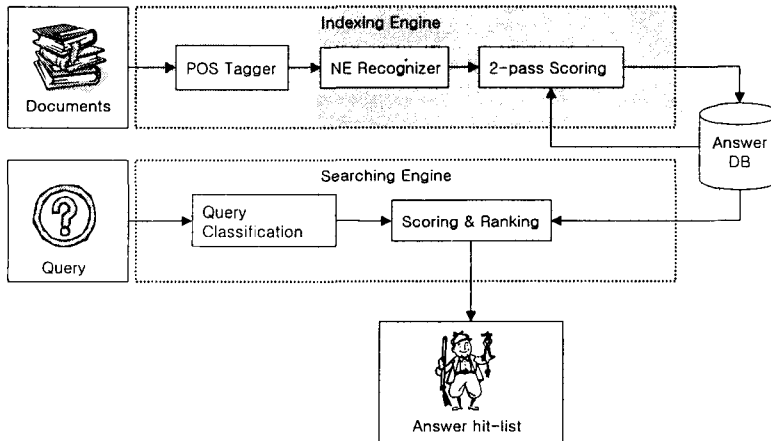
	Strengths	Weaknesses
MURAX	Shallow linguistic analysis (POS tagger, finite-state recognizer)	Handcrafted patterns (Complex and fragile)
GuruQA	Fast response (Predictive annotation)	Use only local information
FALCON	High performance	Domain-specific knowledge like a semantic net.

1. 검색 속도가 IR 시스템에 비해서 상대적으로 많이 느리다.
→ 색인 필요
2. Passage 내의 정답 주변 정보만을 이용한다.
→ Second level 정보 이용 필요

diquest

자연어 처리, 음운론 및 발음, 음향, 베이컨트 시공을 대표하는 연구공급 벤처기업

System architecture (1/2)

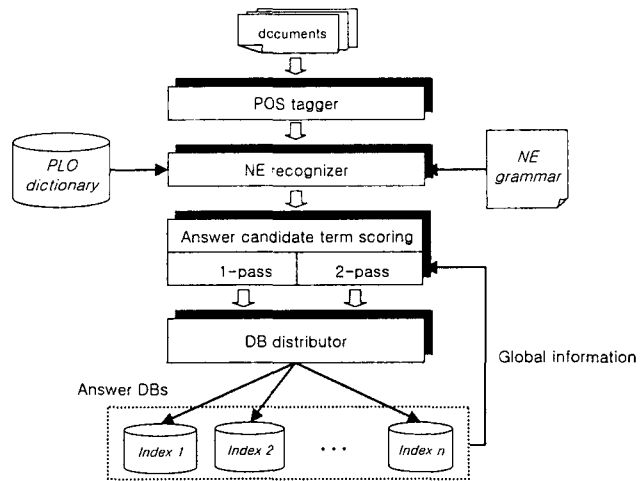


diquest

자연어 처리, 음운론 및 발음, 음향, 베이컨트 시공을 대표하는 연구공급 벤처기업



System architecture (2/2)



diquest

자연어 처리 응용 분야의 글과 음성 패턴 기술을 대표하는 연구개발 벤처기업



Answer-finding (1/2)

- 문맥 윈도우 설정

인터넷 포털 업체들의 새로운 전략

야후코리아(대표 염진섭 www.yahoo.co.kr)는 무료 이메일 서비스를 시작했다.

그 서비스를 사용하는 고객은 6메가의 무료 이메일 공간을 사용할 수 있다.



의미 사전



의미 규칙

diquest

자연어 처리 응용 분야의 글과 음성 패턴 기술을 대표하는 연구개발 벤처기업



Answer-finding (2/2)

- 105개의 의미 분류

계층 1	계층 2				
animal	bird	fish	mammal	person	reptile
plant	grass	tree	flower		
location	address	building	city	continent	country
	state	town			
identification	code	e_mail	nationality	position	tel_num
	URL				
date	day	month	season	weekday	year
time	hour	minute	second		
quantity	age	distance	duration	length	money
	number	power	rate	size	speed
	temperature	volume	weight		

diquest

자연어 처리, 플루언 및 질의 응답 에이전트 시장을 대표하는 연구중심 벤처기업



Calculation of scores (1/3)

야후코리아(사장 염진섭 www.yahoo.co.kr)는 무료 이메일 서비스를 시작했다.
 그 서비스를 사용하는 고객은 6메가의 무료 이메일 공간을 사용할 수 있다.

→ Distance, Term Frequency, TF*IDF

- 거리 가중치 (distance feature 반영)

$$distw_{d,k}(a, w_j) = \frac{c}{\log(dist(i, j)) + c}$$

- 빈도수 (term frequency feature 반영)

$$LS_{d,k}^n(a, w_{pos(n)}) = distw_{d,k}(a, w_{pos(n)}) + (1 - distw_{d,k}(a, w_{pos(n)})) \times LS_{d,k}^{n-1}(a, w_{pos(n-1)}),$$

where $LS_{d,k}^0(a, w_{pos(0)}) = 0$.

diquest

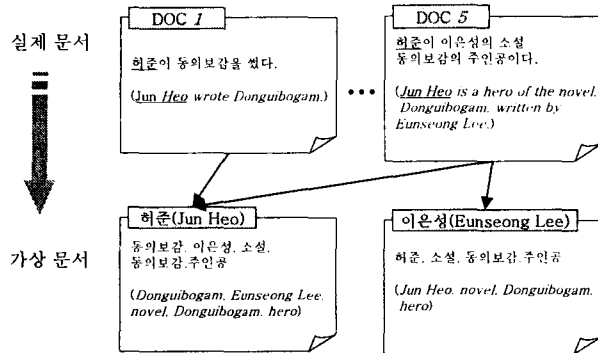
자연어 처리, 플루언 및 질의 응답 에이전트 시장을 대표하는 연구중심 벤처기업



Calculation of scores (2/3)

TF*IDF

$$GS(pseudo_d_w, w) = \begin{cases} \left(0.5 + 0.5 \frac{tf_w}{Max_tf}\right) \frac{\log(N/n)}{\log(N)}, & \text{if } tf_w > 0 \\ 0, & \text{if } tf_w = 0 \end{cases}$$

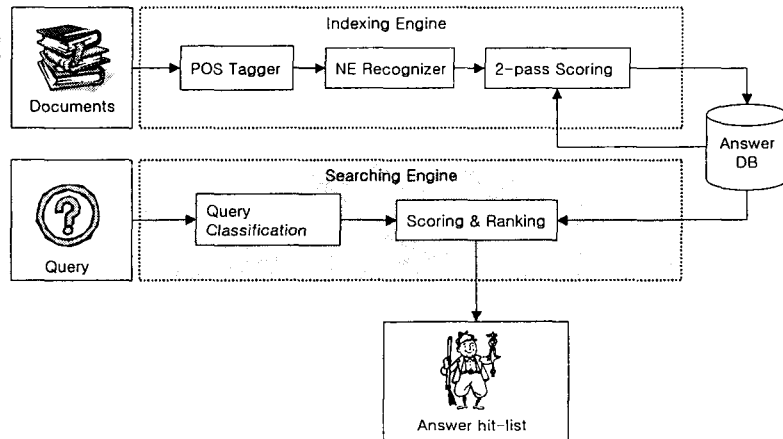


diquest

자연어 처리, 불분명 한 질의 응답 매치먼트 시각화 (대표하는 연구자: 김민기)



System architecture

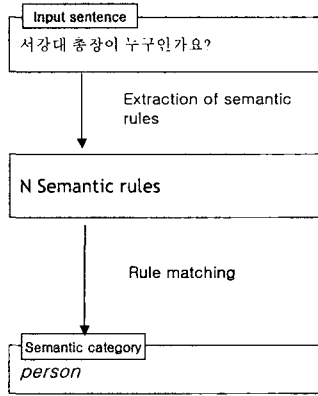


diquest

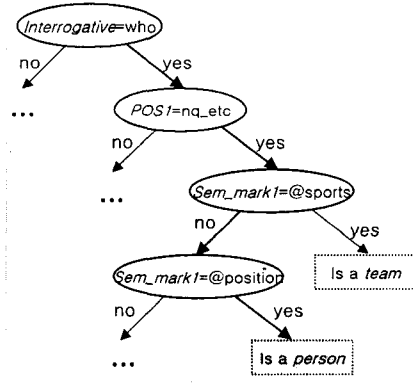
자연어 처리, 불분명 한 질의 응답 매치먼트 시각화 (대표하는 연구자: 김민기)



Searching method (1/2)



Classification Rules



diquesi

자연어 처리 솔루션 및 길의 응답 에이전트 기술을 대표하는 연구중심 벤처기업



Searching method (2/2)

- 정답 후보 점수 계산: p-Norm 모델 (AND)

$$Sim(A, Q_{and}) = 1 - \sqrt[p]{\frac{q_1^p (1 - S(A, w_1))^p + q_2^p (1 - S(A, w_2))^p + \dots + q_i^p (1 - S(A, w_i))^p}{q_1^p + q_2^p + \dots + q_i^p}}$$

- IR, QA 유사도 결합

$$Sim(D, Q) = \frac{\alpha \cdot IRsim(D, Q) + \beta \cdot QAsim_d(A_{DMS}, Q)}{\alpha + \beta}$$

diquesi

자연어 처리 솔루션 및 길의 응답 에이전트 기술을 대표하는 연구중심 벤처기업



Experiments (1/2)

실험 데이터

- KorQATeC 1.0
 - 207,067 balanced documents (368,768 kilobytes)
- WEBTEC 1.0 (WEB TEst Collection)
 - 22,448 documents (43,953 kilobytes)
 - www.sogang.ac.kr (7,869 documents)
 - korea.internet.com (6,452 documents)

diquest

자신의 자리 이름과 및 강의 응답 메타인트 시그를 대표하는 연구용량 벤치마크



Experiments (2/2)

질의 응답 정확률 실험

	Lee2000 (object)	Lee2000 (50-byte)	Kim2001 (object)	실험 시스템
MRAR	0.322	0.456	0.485	0.540
MRAR-1	0.322	0.456	0.539	0.600

질의 응답 및 색인 속도 실험

	Response time per query (seconds)	Indexing time per mega byte (seconds)
기본 IR 시스템	0.026	26.765
실험 시스템	0.048	30.542

diquest

자신의 자리 이름과 및 강의 응답 메타인트 시그를 대표하는 연구용량 벤치마크



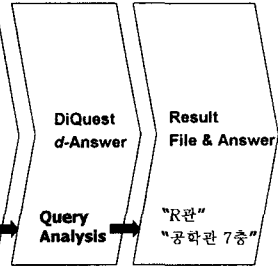
DiQuest d-Answer Enterprise : 지능형 웹 정보검색 에이전트

DiQuest d-Answer의 특징

- 자연어 검색 엔진인 DiQuest IR의 모든 기능에다 고객의 질문에 대한 정답을 추출하여 제공하는 등 이
- 추가된 자체 검색 솔루션
- 고객의 질문에 유사한 문서를 제시하는 일반적인 검색엔진과는 달리 DiQuest d-Answer는 문서 중 에 서
- 질문에 대안이 되는 문장이 있는 문서만을 추려내어 정답 문장을 제시하는 솔루션

DiQuest d-Answer의 질문 예제

“한국 야후의 사장은 누구인가요?”
 “공동구매를 신청했는데, 며칠 만에 배달이 됩니까?”
 “고객 상담센터의 전화번호는?”
 “영업 담당 최진영씨의 이메일은?”
 “수시모집에 관한 안내자료를 찾습니다.”
 “화광과는 어디에 있나요?”



자연어 처리 솔루션 및 질의 응답 에이전트 시장을 대표하는 연구개발 벤처기업



DiQuest d-Answer Preview(1): 웹 문서상에서 검색

검색어: [검색어 입력란]

[검색] [초기화]

검색 결과를 찾아 다음과 같이 추천합니다. 정답에 마우스를 위치시키면 주변 문맥을 확인할 수 있습니다.

정관 (71.1%) · 공학관 (71.1%)
 서강대학교 화학공학과는 1층 1호차 공학
 관 7층에 위치하고 있습니다.

→ **질문에 대한 정답추천**

찾고 싶은 것이 혹시 다음 중에 있으니까?

> 생활의 유익을 확인할 수 있는 캠퍼스 라이프입니다. → **관련 홈페이지 직접 연결**

■ “124.1” 이송로 유선 1층에 대한 문서 검색 결과를 찾았습니다. (1 - 10) / B4

1. 원 서강대학교 화학공학과 홈페이지입니다.
 서강대학교 화학공학과는 1층 1호차 동관 7층에 위치하고 있습니다.

2. 원 서강대학교 화학공학과 홈페이지입니다.

이송로 · 이송시도 124.1 이송로 유선 1층서강대학교 화학공학과 ☎ 02) 705-8474 팩스 02) 711-0439 E-mail: ihb@cs.skogang.ac.kr



자연어 처리 솔루션 및 질의 응답 에이전트 시장을 대표하는 연구개발 벤처기업



DiQuest d-Answer Preview(2) : RDB 내 검색 기능

The screenshot displays a search interface with a search bar containing the text '정문에 대한 정답수신'. Below the search bar, there are several search filters including '만족도' (Satisfaction), '국립대' (National University), '지역' (Region), '연도' (Year), and '분야' (Field). The main content area shows search results with a '정문에 대한 정답수신' section. A '관련 홈페이지 직접 연결' (Direct link to related homepage) button is visible. A 'This' icon is also present on the left side of the results area.

diquest

자연을 편리, 슬루트 및 영인 상담 에이전트 기술을 대표하는 연구 중심 벤처기업



DiQuest d-Answer Preview(3) : 정답문서 추천 및 바로가기 기능

The screenshot shows a search interface with a search bar containing '북학에 대한 정답을 하교 질문요'. Below the search bar, there are search filters for '만족도' (Satisfaction), '매우만족' (Very Satisfied), '만족' (Satisfied), '보통' (Average), '불만족' (Dissatisfied), and '매우불만족' (Very Dissatisfied). The main content area features a section titled '정문에 대한 정답수신' with a '바로가기' (Quick Link) button. Below this, there are two recommendation sections: '1. 북학학 상담' (North Korean Studies Consultation) and '2. 북학학 상담' (North Korean Studies Consultation). A '정문에 대한 관련 홈페이지 추천' (Recommend related homepage for the question) button is also visible.

diquest

자연을 편리, 슬루트 및 영인 상담 에이전트 기술을 대표하는 연구 중심 벤처기업

Conclusion

AS - Is (기존의 검색 시스템)

- 기능 측면에서의 질의 응답 결과
 1. 해당되는 문서를 단순 나열함.
 2. 문서 속에서 정답 검색이 여전히 필요.
 3. 정보 검색 과정에 많은 시간 소요.
 4. 사용자의 문서 활용도가 매우 낮음.
- 정확도 측면에서의 질의 응답 결과
 1. 질문에서 키워드만을 추출.
 2. 키워드에 대한 체계적 검색 결과만 제시.
 3. 정확한 의도 분석이 불가능.

• 질문 : "김낙영 교수님의 전화번호는?"
 1. 김낙영+교수+전화+번호
 문서 내에서 단순히 발견되는 정도 측정.
 "김낙영", "교수", "전화", "번호", "전화번호"
 라는 단어가 들어간 불필요한 모든 문서들을
 제시

To - Be (질의 응답 시스템)

- 기능 측면에서의 질의 응답 결과
 1. 문서에서 정답을 직접 찾아서 제시.
 2. 정답 검색 과정 시간 단축.
 3. 사용자 문서 활용도 증가 → 인지도 향상.
 4. 웹 문서뿐 아니라 데이터베이스까지 검색.
 5. 특정 문서 직접 제시 가능.
- 정확도 측면에서의 질의 응답 결과
 1. 질문의 어휘 구조, 질문 의도까지 파악.
 2. 정확한 질의 분석 → 의미있는 부분만 검색.
 3. 검색 성능 최적화 가능.

• 질문 : "김낙영 교수님의 전화번호는?"
 1. 전화번호에 관한 질문(김낙영 교수)
 우선 전화번호에 관한 질문이라는 것이 파악
 되고, 그 대상이 김낙영 교수라는 것을 인식
 하여 정답 추출

diquesc

지능이 처리 솔루션 및 질의 응답 에이전트 시장을 대표하는 연구중심 벤처기업