지역, 전역 정보를 이용한 정답 후보 색인 방법

김학수°, 김경선, 서정연

서강대학교 컴퓨터학과 자연어처리 연구실

{hskim, kksun}@diquest.com, seojy@ccs.sogang.ac.kr

A Predictive Answer Indexing Method Using Local and Global Information

Harksoo Kim°, Kyungsun Kim, Jungyun Seo

Natural Language Processing Lab., Dept. of Computer Science, Sogang University

요약

본 논문은 2-패스에 걸쳐 지역, 전역 정보를 추출하고 이 정보들을 이용하여 효과적으로 정답 후보들을 색인하는 방법을 제안한다. 제안한 정답 후보 색인 방법은 다음과 같다. 먼저, 대상 문서에 포함된 모든 정답 후보들을 추출한다. 그리고, 지역 정보(한 문서 내에서 정답 후보와 주변 내용어 사이의 관계)를 이용하여 각 내용어에 점수를 부여한다. 다음으로 전역 정보(모든 문서를 대상으로 하여 정답 후보와 공기(co-occurrence)하는 내용어 사이의 관계)를 이용하여 각 내용어에 이미 할당되어 있는 점수를 변경한다. 마지막으로 데이터베이스에 각 정답 후보와 점수가 부여된 내용어들을 역과일 형태로 저장한다. 이러한 색인 방법은 빠른 응답 시간과 비교적 높은 정확률을 필요로 하는 실용적 질의 응답 시스템에 적합하다.

1. 서론

기존의 정보검색(information retrieval, IR)은 사용자의 질문에 대한 응답으로 대량의 문서를 검색하고 순위화하는데 초점을 맞추어 왔다. 그러나, 많은 사용자들은 명확한 의도를 가지고 질문을 하며, 정보 검색 시스템이 대량의 문서를 찾아주기 보다는 정답들을 곧바로 찾아 제시해 주기를 바란다[1]. 이러한 요구를 만족시키기 위하여 질의 응답(question answering, QA)이라는 개념이 출현했으며, 많은 연구들이 AAAI[2]와 TREC[3]을 중심으로 수행되어 왔다. 질의 응답 시스템은 대용량의 문서를 검색하고, 검색된 텍스트로부터 부적당한 구(phrase)나 문장을 제거한다. 이러한 필터링(filtering) 과정 덕분에 사용자는 검색된 문서를 모두 읽지 않고도 빠른 시간 내에 자신이 얻고자 하는 정답에 접근할 수 있다. 그러나 지금까지의 질의 응답 시스템에 대한 연구는 WWW(world wide web)와 같은 실제 필드(field)에서 나타날 수 있는 다음과 같은 문제를 간과해 왔다.

- 사용자들은 가능한한 빨리 정답을 찾고 싶어 한다. 만약 질의 응답 시스템이 수 초 내에 응답이 없다면 사용자들은 해당 시스템의 유용성에 의심을 품을 것이다.
- 사용자들은 다양한 어휘 및 구문 형태(syntactic form)를 이용하여 자신의 의도를 표현한다. 이러한 이유로 질의 응답 시스템이 응용 영역에 상관없이 안정된 성능을 보인다는 것은 매우 어려운 일이다.

이러한 문제를 해결하기 위해서 본 논문에서는 실용적 질의 응답 시스템에 적합한 효과적인 정답 후보 색인 방법을 제안한다.

본 논문의 구성은 다음과 같다. 먼저, 2장에서 질의 응답 시스템에 대한 기존의 연구를 살펴보고, 3장에서는 실용적인 질의 응답 시스템을 위한 효과적인 정답 후보 색인 방법을 제안한다. 4장에서 정답 후보 색인 방법의 유용성을 평가하기 위한 실험을 한 후, 5장에서 결론을 맺는다.

2. 관련 연구

질의 응답 방법론은 크게 텍스트 조각 추출 방법(text-snippet extraction method)과 명사구 추출 방법(noun-phrase extraction method)으로 나뉘어 진다[4]. 텍스트 조각 추출 방법은 질문에 대한 정답을 포함하고 있을 것 같은 텍스트의 단락이나 문장 또는 문장의 일부를 추출하는 것으로 지난 TREC QA 부문에 참가한 시스템들[5, 6]이 TREC의 평가 기준에 따라 일반적으로 사용한 방법이다. 명사구 추출 방법은 한정된 클래스(closed-class)에 속한 사용자 질문에 대해서 구체적인 정답구(answer phrase)을 찾아주는 것이다. 이 방법은 일반적으로 단답형 정답(noun-phrase answer)만을 찾아서 제시할 수 있다는 단점을 가지고 있다.

ExraAns[7]는 텍스트 조각 추출 방법을 사용하는 대표적인 질의 응답 시스템으로 정답을 포함하는 문서 내의 절이나 구를 찾아 준다. 그러나, ExtraAns는 제한된 영역의 구문적(syntactic), 의미적(semantic) 정보를 사용하기 때문에 응용 영역을 바꾸기가 쉽지 않다. FALCON[8]도 역시 대표적인 텍스트 조각 추출 시스템 중에 하나이다. 그 시스템은 구문, 의미, 화용(pragmatic) 지식 등을 이용하여 매우 정확하게 정답을 추천해 준다. FALCON이 비록 높은 정확률을 보이지만 의미망(semantic net)과 같은 영역 의존적 지식을 구축하는 것이 현실적으로 매우 어렵기 때문에 실용적인 질의 응답 시스템으로는 적합하지 않다.

MURAX[9]는 대표적인 명사구 추출 시스템으로 품사 태거(POS tagger), lexico-syntactic 패턴 매칭(pattern matching)을 위한 유한 상태 인식기(finite-state recognizer)와 같은 비교적 저급의 언어 지식(shallow linguistic knowledge)을 이용하여 정답을 추천한다. 유한 상태 인식기는 사용자의 질의 의도를 파악하고 올바른 정답 후보를 결정하는데 사용된다. TREC에 참가한 몇몇 시스템들은 사용자의 질의 의도를 파악하기 위해서 MURAX가 사용한 것과 비슷한 유한 상태 인식기를 사용한다[4]. 그러나, 이러한 시스템들은 검색 시에 정답 후보들을 찾아서 점수를 부여하고 상위 몇 개를 추천하기 때문에 응답 시간이 매우 길다는 단점이 있다. 이런 단점은 극복하기 위해서 GuruQA[10, 11]는 검색 전에 미리 정답 후보들을 찾아 색인하는 방법(predictive annotation)[10, 11, 12]을 사용한다. 비록 GuruQA가 빠른 시간에 정답을 제시하고 좋은 성능을 보이지만 여러 문서에 걸쳐서 관찰될 수 있는 유용한 정보들을 이용하지 못한다. 즉, GuruQA는 색인의 범위를 최소 한 문장에서 최대 한 문서 내로 한정을 하기 때문에 동일한 정답이 여러 문서에서 나타났을 경우에 각각의 문서에 존재하는 정보를 서로 공유하지 못한다.

3. 정답 후보 색인 방법

정답 후보 색인 과정은 후보 찾기 단계와 점수 부여 단계로 이루어진다. 정답 후보를 찾기 위해서 본 논문에서는 사용자의 질의 유형을 105개의 의미 범주(semantic category)로 구분하고, 그것에 따라 정답 유형(answer type)을 분류한다. 본 논문에서는 105개의 의미 범주를 결정하기 위하여 TREC에 참가한 질의 응답 시스템들의 의미 범주를 참고하였고, 상업용 정보 검색 시스템[13]으로부터 수집한 질의 로그(log)를 분석하였다.

문서로부터 각각의 의미 범주에 속하는 정답 후보들

을 추출하기 위해서는 품사 태거와 개체명 인식기(named entity recognizer)를 이용한다. 개체명 인식기는 PLO 사전이라고 불리어지는 개체명 사전(named entity dictionary)과 패턴 매처(pattern matcher)로 구성된다. PLO 사전은 사람, 국가, 도시, 기관의 이름 외에도 센티미터(cm), 킬로그램(kg)과 같은 단위들을 포함한다. 정답 후보를 추출하는 과정은 다음과 같다. 먼저, 품사 태거가 입력된 문장에 형태소 분석을 한다. 그리고, 개체명 인식기가 그 결과를 입력으로 하여 개체명을 추출하고 의미 범주를 할당 한다. 개체명 추출을 위해서는 정규 표현(regular expression) 형태의 lexico-syntactic 패턴이 이용되고, 의미 범주 할당을 위해서는 PLO 사전이 이용된다. 예를 들어, "야후코리아(사장 염진섭, www.yahoo.co.kr)은 무료 이메일 서비스의 용량을 6메가로 늘렸다." 라고 하는 문장이 있을 때, 개체명 인식기는 그 문장으로부터 4개의 정답 후보를 추출한다. PLO 사전을 이용하여 '야후코리아' 는 company로, '염진섭' 은 person으로, '6메가' 는 size로 의미 범주가 결정된다. 그리고, 'www.yahoo.co.kr' 과 URL은 패턴 매처에 의해서 의미 범주가 결정된다. 패턴 매처는 정규 표현(regular expression) 형태의 lexico-syntactic 패턴을 이용하여 telephone number, email address와 같은 정형화된 정답 후보들을 추출한다. 또한, 패턴 매처는 '~ 때문에', '~하려면' 과 같은 단서 구문을 이용하여 이유(reason)나 방법(method)과 같은 서술형 정답의 후보들도 추출한다. 그러나, 서술형 정답은 그 경계(boundary)를 추출하는 것이 매우 어려우므로 패턴과 매치(match)된 문장 전체를 정답 후보로 추출한다. 정답 후보를 추출한 후 같은 문맥(context)에 존재하는 내용어들(content words)에게 점수를 부여한다. 본 논문에서는 문맥의 범위를 정답 후보가 속한 문장을 기준으로 앞의 한 문장, 뒤의 한 문장으로 한다. 다시 말해서, 색인의 범위가 되는 문맥 윈도우(context window)의 최대 크기는 3문장이 되는 것이다. 문맥 윈도우의 크기는 정답 후보가 포함된 주변 문장에 어휘 체인이나 대용어가 포함되었는지의 여부에 따라 동적으로 변한다. 예를 들어, 정답 후보를 포함한 문장이 이전 문장과 어휘 체인을 가지고 있고, 다음 문장과는 그렇지 못할 경우에 문맥 윈도우의 크기는 2로 줄어든다. [그림 1]은 문맥 윈도우의 크기 변화를 예로 보여준다.

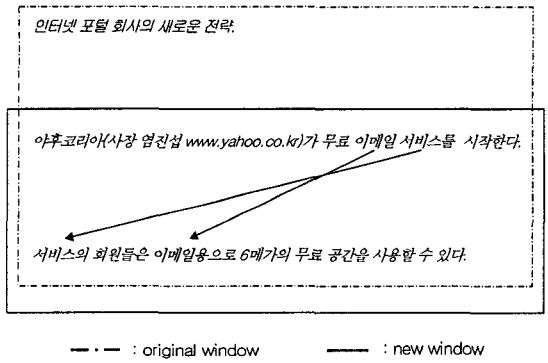


그림 1. 문맥 윈도우의 크기 변화

문맥 윈도우의 크기가 결정되면 2-패스 점수 부여 방법에 따라 윈도우에 속한 내용어들에게 점수를 부여한다. 첫 번째 패스에서 각 내용어에 지역 점수(local score)를 할당한다. 지역 점수는 한 문서의 해당 문맥 윈도우 내에서 정답 후보와 내용어 사이의 연관 관계를 의미한다. 예를 들어, "야후 코리아(www.yahoo.co.kr)가 새로운 서비스를 시작한다." 라는 문장이 있고 정답 후보가 '야후코리아' 일 때, 'www.yahoo.co.kr' 은 '서비스' 보다 '야후코리아' 와 더 밀접한 관계를 맺고 있다. 본 논문에서는 이렇게 하나의 문맥 윈도우 내에서 정답 후보와 내용어 사이에 존재하는 연관 관계를 지역 정보(local information)이라고 부르고, 그것을 점수화한 것을 지역 점수라고 부른다. 지역 점수를 할당하기 위해서는 다음과 같은 2개의 특징(feature)을 이용한다.

- 빈도수(frequency): 문맥 윈도우에 속한 내용어들의 빈도수. 빈도수가 높은 내용어에 더 높은 점수를 부여한다. 예를 들어, [그림 1]에서 '이메일' 이 '회원' 보다 높은 점수를 받는다.
- 거리(distance): 정답 후보와 내용어 사이의 거리. 정답 후보와 가까운 거리에 있는 내용어에게 높은 점수를 부여한다. 예를 들어, [그림 1]에서 '염진섭' 이 정답 후보일 때, '사장' 이 '서비스' 보다 높은 점수를 받는다.

제안하는 색인 방법은 단어들 사이의 IS-A 관계나 문장 성분(grammatical role)과 같은 정보를 사용하지 않는다. 왜냐하면 실제 응용 영역에서는 웹 문서와 같이 테이블(table)이나 이미지(image)를 포함한 문서들이 많이 존재하며, 문장 분할도 어려울 만큼 자유롭게 기술된 문서들이 많아서 위와 같은 고급 정보를 추출하는 것이 상대적으로 어렵기 때문이다.

지역 점수를 계산하기 위해서는 2단계 계산 과정을 거친다. 먼저, [식 1]을 이용하여 정답 후보와 내용어 사이의 거리 가중치(distance weight)를 계산한다.

$$distw_{d,k}(a_i, w_j) = \frac{c}{log(dist(i,j)+c} \qquad (1)$$

[식 1]에서 $distw_{d,k}(a_i, w_j)$는 문서 d의 k번째 문맥 윈도우에 j번째 위치한 내용어 w의 거리 가중치이다. $dist(i,j)$는 i번째 위치한 정답 후보 a와 j번째 위한 내용어 w사이의 거리이다. c는 실험 상수이다. 거리 가중치 계산이 끝나면, 동적 프로그래밍 기법(dynamic programming method)에 따라 기술된 [식 2]를 이용하여 문맥 윈도우에 속한 동일한 내용어들의 거리 가중치를 모두 더한다. [식 2]를 이용하여 빈도수가 높은 내용어들이 더 높은 점수를 부여 받는다.

$$LS^n_{d,k}(a_i, w_{pos(n)}) = distw_{d,k}(a_i, w_{pos(n)}) + (1 - distw_{d,k}(a_i, w_{pos(n)})) \times LS^{n-1}_{d,k}(a_i, w_{pos(n-1)}), \quad (2)$$
$$where\ LS^0_{d,k}(a_i, w_{pos(0)}) = 0$$

[식 2]에서 $LS^n_{d,k}(a_i, w_{pos(n)})$은 문서 d의 k번째 문맥 윈도우에 동일한 형태의 내용어가 n개 존재할 때 n번째 내용어의 지역 점수이다. $pos(n)$은 n번째 내용어의 위치이다. [식 2]를 재귀적으로 풀고 나면 i번째 정답 후보 a_i와 k번째 문맥 윈도우에 속한 내용어 w 사이의 지역 점수 $LS_{d,k}(a_i, w)$가 계산된다. [그림 2]는 지역 점수 계산 과정의 예를 보여준다.

DOC 1

인터넷 포털 회사의 새로운 전략.

야후코리아 (사장 염진섭 www.yahoo.co.kr)가 무료 이메일 서비스를 시작한다.

서비스의 회원들은 이메일용으로 6메가의 무료 공간을 사용할 수 있다.

정답후보	처리 과정
야후코리아	1. '야후코리아'와 인접한 두 문장에 존재하는 '서비스' 사이의 거리를 계산한다. dist(1, 7)=6, dist(1, 9)=8 2. 거리 가중치를 계산한다. *distw(Yahoo Korea1, service7)* =1/(log(6)+1)=0.358 *distw(Yahoo Korea1, service9)* =1/(log(8)+1)=0.325 3. 두 거리 가중치를 합한다. *LS(Yahoo Korea1, service)* =0.358+(1.0-0.358)*0.325=0.567

그림 2. 지역 점수 계산의 예

두 번째 패스는 가상 문서(pseudo-document) 생성, 전역 점수(global score) 계산, 그리고 점수 합산 단계로 나뉘어 진다. 첫 번째 단계에서는 정답 후보를 기준으로 가상 문서를 생성한다. 가상 문서는 모든 문서를 대상으로 하여 동일한 정답 후보를 포함하는 문맥 윈도우들에 속한 내용어들로 구성된다. 가상 문서는 해당하

는 정답 후보로 구분되어 이름 붙여진다. [그림 3]은 가상 문서의 예를 보여준다.

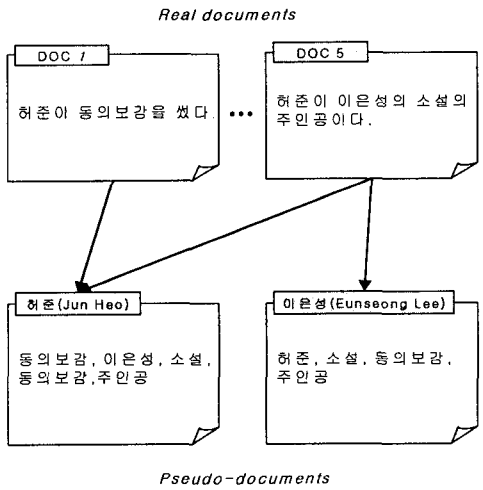
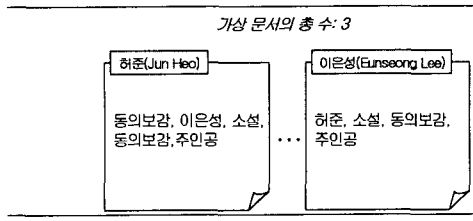


그림 3. 가상 문서의 예

다음 단계에서는 [식 3]을 이용하여 각 정답 후보들(가상 문서들)의 전역 점수를 계산한다. 전역 점수는 각 정답 후보와 그 후보를 포함하는 여러 문맥 윈도우들에 속한 내용어들 사이의 연관 관계를 점수화한 것이다. 본 논문에서는 이러한 연관 관계를 전역 정보(global information)라고 부른다.

$$GS(pseudo\_d_a, w) = \begin{cases} \left(0.5 + 0.5\frac{tf_w}{Max\_tf}\right)\frac{log(N/n)}{log(N)}, & \text{if } tf_w > 0 \\ 0, & \text{if } tf_w = 0 \end{cases} \quad (3)$$

[식 3]은 대상 문서가 실제 문서가 아니라 가상 문서라는 것을 제외하고는 잘 알려진 $TF \cdot IDF$ 수식[14]과 동일하다. 그러므로, TF 컴포넌트인 $(0.5+0.5 \cdot (tf_w/Max_tf))$는 가상 문서 $pseudo_d_a$에 포함된 내용어 w의 정규화된 빈도수를 의미한다. IDF 컴포넌트인 $log(N/n)/log(N)$은 내용어 w를 포함하는 가상 문서들의 정규화된 역문헌 빈도수를 의미한다. $GS(pseudo_d_a, w)$는 정답 후보 a와 내용어 w 사이의 전역 점수를 의미한다. 좀더 자세히 설명하면, tf_w는 가상 문서 $pseudo_d_a$에 포함된 내용어 w의 빈도수이고, Max_tf는 가상 문서 $pseudo_d_a$에 포함된 내용어들의 최대 빈도수이다. n은 내용어 w를 포함하는 가상 문서들의 수이고, N은 가상 문서의 총 수이다. [그림 4]는 전역 점수 계산 과정의 예를 보여준다.



정답후보	처리과정
허준	GS(*허준, 동의보감*) =(0.5+0.5*(2/2))*(log(3/2)/log(3))=0.369 GS(*허준, 이은성*) =(0.5+0.5*(1/2))*(log(3/2)/log(3))=0.277 GS(*허준, 소설*) =(0.5+0.5*(1/2))*(log(3/2)/log(3))=0.277 GS(*허준, 주인공*) =(0.5+0.5*(1/2))*(log(3/2)/log(3))=0.277

그림 4. 전역 점수 계산의 예

마지막 단계에서는 [식 4]를 이용하여 지역 점수와 전역 점수를 합한다.

$$S_{d,k}(a_i, w) = \frac{\alpha \cdot LS_{d,k}(a_i, w) + \beta \cdot GS(pseudo\_d_a, w)}{\alpha + \beta} \quad (4)$$

[식 4]에서 $LS_{d,k}(a_i, w)$는 정답 후보 a_i와 문서 d의 k번째 문맥 윈도우에 포함된 내용어 w 사이의 지역 점수이고, $GS(pseudo_d_a, w)$는 전역 점수이다. α와 β는 가중치 인수(weighting factor)이다. 두 점수를 합한 후에 정답 후보의 위치, 점수와 함께 내용어들을 역파일 형태로 데이터베이스에 저장한다.

4. 실험 및 평가

4.1 실험 데이터

실험을 위해서 본 논문에서는 두 종류의 테스트 컬렉션(test collection)을 이용하였다. 하나는 실제 웹사이트(www.koreainternet.com과 www.sogang.ac.kr)로부터 수집된 테스트 컬렉션이고, 다른 하나는 한국어 질의응답 시스템 성능 평가 테스트 컬렉션인 KorQATeC 1.0[15]이다. 본 논문에서는 전자를 WEBTEC이라고 부르고, 후자를 KorQATeC이라고 부른다. WEBTEC은 22,448 문서(110,004 KB)로 구성되어 있으며, KorQATeC은 207,067 문서(368,768 KB)로 구성되어 있다. 그리고, 각각의 컬렉션은 50개의 질의 응답 셋(set)을 포함한다.

색인 방법의 성능을 평가하기 위해서 본 논문에서는 [12]에서 사용한 규칙 기반의 질의 처리 방법과 [식 5]를 이용하여 질의어와 정답 후보 사이의 유사도를 계산하는 실험용 질의 응답 시스템을 구성하였다. 그리고, 이 실험용 질의 응답 시스템의 성능을 이용하여 간접적으로 제안한 색인 방법의 성능을 평가한다. [식 5]는 p-Norm 모델[16]의 AND 오퍼레이션(operation)이다.

$$Sim(A, Q_{and}) = 1 - \sqrt[p]{\frac{q_1^p(1-t_1)^p + q_2^p(1-t_2)^p + \ldots + q_i^p(1-t_i)^p}{q_1^p + q_2^p + \ldots + q_i^p}} \quad (5)$$

[식 5]에서 A는 정답 후보이고, t_i는 그 정답 후보의 문맥 윈도우에 포함된 i번째 내용어의 점수이다. q_i는 질의에 포함된 i번째 내용어이고, p는 p-Norm 모델의 p-value이다.

질의 응답 시스템의 성능을 평가하기 위해서 TREC에서 사용하는 것과 마찬가지로 각 질의에 대한 첫 번째 정답의 RAR(reciprocal answer rank)을 계산하고, [식 6]과 같이 평균을 구한다[3, 17].

$$MRAR = 1/n\left(\sum_i 1/rank_i\right) \quad (6)$$

[식 6]에서 $rank_i$는 i번째 질문에 대해 응답으로 제시한 것들 중에서 첫 번째로 정답인 것의 순위이다. n은 질의의 총 수이다.

4.2 실험 결과 분석

제안한 색인 방법은 [식 4]와 같이 지역 점수와 전역 점수의 합을 이용하여 정답 후보와 내용어 사이의 점수를 계산한다. 본 논문에서는 실험을 통하여 [식 4]의 α와 β를 0.1과 0.9로 설정한다. 실험용 질의 응답 시스템의 성능을 평가하기 위해서 이경순2000[18]과 Kim2001[12]을 비교 대상으로 삼았다. 비교 대상 시스템들이 WEBTEC에 대한 결과를 제시하지 않기 때문에 KorQATeC만을 이용하여 실험하였다.

표 1. 질의 응답 시스템의 성능 비교

System	이경순2000	이경순2000 (50-byte)
MRAR	0.322	0.456
MRAR-1	0.322	0.456
System	Kim2001	실험용 질의 응답 시스템
MRAR	0.485	0.540
MRAR-1	0.539	0.600

[표 1]에서 보듯이 실험용 질의 응답 시스템의 성능은 다른 비교 대상 시스템보다 좋다. 이 사실은 제안한 색인 방법에서 사용한 점수 부여 방법이 단순하지만 매우 유용하다는 것을 보여준다. [표 1]에서 이경순2000 (50-byte)은 정답 그 자체를 제시하는 것이 아니라 정답을 포함하는 주변 50-byte를 제시하는 시스템이다. MRAR-1은 정답을 제시하지 못한 것들을 제외한 MRAR이다.

[표 2]는 기존 정보 검색 시스템[19]과 실험용 질의 응답 시스템의 검색 속도를 비교한 것이다. [표 2]에서

보듯이 정보 검색 시스템의 평균 검색 시간은 듀얼(dual) CPU의 펜티엄(pentium) III 서버에서 0.026초이고, 실험용 질의 응답 시스템은 0.048초이다. 이것을 바탕으로 실험용 질의 응답 시스템과 정보 검색 시스템과의 속도 차이는 그렇게 크지 않음을 알 수 있다.

표 2. 검색 및 색인 속도 비교

	질의당 응답 시간(초)	메가 바이트당 색인 시간(초)
IR system	0.026	26.765
실험용 질의 응답 시스템	0.048	30.542

[표 3]은 정답 후보의 수가 많은 상위 20개의 의미 범주에 대한 원본 문서 1메가 바이트(mega byte) 당 정답 후보수를 보여준다. [표 3]을 분석한 결과, 제안한 정답 후보 색인 방법이 원본 문서 1메가 바이트를 색인하는데 의미 범주 당 0.102 메가 바이트의 디스크 공간을 필요로 하는 것을 알 수 있었다.

표 3. 의미 범주 당 정답 후보의 수

의미 범주	정답 후보 수	의미 범주	정답 후보수
person	2029.72	rate	270.56
country	1907.92	age	139.17
number	1302.64	address	110.28
city	723.75	tel_num	104.72
year	577.22	URL	88.19
day	569.03	continent	81.53
date	458.19	state	80.42
school	435.97	e_mail	74.58
time	319.72	sports	57.08
month	274.03	money	55.28

5. 결론

본 논문에서는 실용적 질의 응답 시스템을 위한 효과적인 정답 후보 색인 방법을 제안하였다. 제안한 정답 후보 색인 방법은 정답 후보와 인접한 내용어들을 추출하고, 2-패스 점수 부여 방법에 따라 각 후보와 내용어들 사이의 연관 관계를 계산한다. 이러한 방법을 이용한 실험용 질의 응답 시스템은 빠른 검색 시간과 함께 높은 정확률을 보여 주었다.

6. 참고 문헌

[1] Voorhees E. and Tice D. M., "Building a Question Answering Test Collection", In *Proceedings of SIGIR 2000*, pp. 200-207, 2000

[2] AAAI Fall Symposium on Question Answering, *http://www.aaai.org/Press/Reports/Symposia/Fall/fs-99-02.html*

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

[4] Vicedo J. L. and Ferrándex A., "Importance of Pronominal Anaphora resolution in Question Answering systems", In *Proceeding of ACL 2000*, pp. 555-562, 2000

[5] Moldovan D., Harabagiu S., Paşca M., Mihalcea R., Goodrum R., Gîrju R. and Rus V., "LASSO: A Tool for Surfing the Answer Net", In *Proceedings of The Eighth Text REtrieval Conference (TREC-8)*, from http://trec.nist.gov/pubs/trec8/t8_proceedings.html, 1999

[6] Prager J., Radev D., Brown E. and Coden A., "The Use of Predictive Annotation for Question Answering in TREC8", In *Proceedings of The Eighth Text REtrieval Conference (TREC-8)*, from http://trec.nist.gov/pubs/trec8/t8_proceedings.html, 1999

[7] Berri J., Molla D., and Hess M., "Extraction a

utomatique de réponses: implémentations du sy stéme ExtrAns" , In *Proceedings of the fifth c onference TALN 1998*, pp. 10-12, 1998

[8] Harabagiu S., Moldovan D., Pasca M., Mihalcea R., Surdeanu M., Bunescu R., Gîrju R., Rus V. and Morarescu P., "FALCON: Boosting Knowl edge for Answer Engines" , In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, from http://trec.nist.gov/pubs/trec9/t9_proce edings.html, 2000

[9] Kupiec J., "Murax: A Robust Linguistic Appro ach for Question Answering Using an On-line Encyclopedia" , In *Proceedings of SIGIR'93*, 19 93

[10] Prager J., Radev D., Brown E., and Coden A., "The Use of Predictive Annotation for Questio n Answering in TREC8" , In *Proceedings of th e Eighth Text REtrieval Conference (TREC-8)*, http://trec.nist.gov/pubs/trec8/t8_proceeding s.html, Gaithersburg, Maryland, 1999.

[11] Prager J., Brown E. and Coden A., "Question -Answering by Predictive Annotation" , In *Pro ceedings of SIGIR 2000*, pp. 184-191, 2000

[12] Kim H., Kim K., Lee G. G. and Seo J., "MA YA: A Fast Question-answering System Based On A Predictive Answer Indexer" , In *Proceedi ngs of the ACL Workshop Open-Domain Quest ion Answering*, pp. 9-16, 2001

[13] diquest.com, *http://www.diquest.com*

[14] Fox E. A., *Extending the Boolean and Vector Space Models of Information Retrieval with P-norm Queries and Multiple Concept Types*, Ph. D. Thesis, CS, Cornell University, 1983

[15] 이경순, 김재호, 최기선, "질의응답시스템의 성능 평가를 위한 테스트컬렉션 구축" , *제 12 회 한글 및 한국어 정보처리 학술 대회 논문집*, pp. 190-197, 2000

[16] Salton G., Fox E. A. and Wu H., *Extended Bo olean Information Retrieval*, Communication of t he ACM, Vol.26, No.12, pp. 1022-1036, 1983

[17] Voorhees E. and Tice D. M., "The TREC-8 Question Answering Track Evaluation" , In *Pro ceedings of the Eighth Text REtrieval Confere nce (TREC-8)*, from http://trec.nist.gov/pubs/tr ec8/t8_proceedings.html, 1999

[18] 이경순, 김재호, 최기선, "한국어 질의응답시스템 에서 개체인식에 기반한 대답 추출" , *제 12회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 184-18 9, 2000

[19] Lee, G., Park, M., and Won, H., "Using syntactic information in handling natural la nguage queries for extended boolean retrie val model" , In *Proceedings of the 4th inte rnational workshop on information retrieval with Asian languages (IRAL99)*, pp. 63-7 0, 1999