KorQATeC2.0: 질의/응답 시스템의 성능 평가를 위한 평가집합 구축

김재호, 이경순, 오종훈, 장두성, 최기선
전문용어언어공학연구센터, 첨단정보기술연구센터, 한국과학기술원
{jjaeh, kslee, rovellia, dschang, kschoi}@world.kaist.ac.kr

KorQATeC2.0: Construction of Test Collection for Evaluation of Question Answering System

Jae-Ho Kim, Kyung-Soon Lee, Jong-Hoon Oh, Du-seong Chang, Key-Sun Choi
KORTERM, AITRC, KAIST

요 약

본 논문에서는 질의/응답 시스템의 평가를 위해 구축된 평가집합 (Korean Question Answering Test Collection 2.0: KorQATeC2.0)에 대하여 기술한다. KorQATeC2.0은 총 120개의 질의와 207,067개의 문서로 구성되어 있으며, 120개의 질의는 질의에 대한 정답을 제시하는 방식에 따라 기본 과제 질의, 나열 과제 질의, 문맥 과제 질의, 요약 과제 질의로 나누어진다. 또한 KorQATeC1.0과는 달리 여러 문서를 참조하여 정답을 구성하는 질의와 문서집합에 정답이 존재하지 않는 질의를 포함시킴으로써 질의/응답 시스템의 평가를 다양하게 할 수 있도록 하였다. 본 논문에서 기술하는 평가집합은 질의/응답 시스템의 객관적 평가를 가능하게 한다는 점에서 그 의의가 있다.

1 서론

정보의 양이 많아짐에 따라 원하는 정보를 보다 빠르고 정확하게 찾기 위해 사용자의 질의에 대해서 문서가 아닌 구체적인 대답을 제시하는 질의/응답 시스템의 필요성이 증가하고 있다. 이러한 질의/응답 시스템을 연구하고, 개발된 시스템의 객관적인 신뢰도 평가를 위해서는 체계적으로 구축된 평가집합 (test collection)이 필요하다.

TREC (Text REtrieval Conference) [11]에서는 TREC-8 [13]부터 질의/응답 시스템의 평가를 위한 평가집합 구축과 질의/응답 시스템의 평가를 수행하고 있다. TREC-8 [13]에서는 200여 질의만을 구축하였지만 TREC-9 [14]에서는 97만여 개의 문서를 대상으로 893개의 질의로 수량을 확장하였고, TREC-10 [15]에서는 질의/응답 과제를 기본 과제 (Main task), 나열 과제 (List task), 문맥 과제 (Context task)의 세 가지 세부 과제로 나누어 여러 문서에서 답을 추출하는 질의와 이전 질의의 내용과 해당 정답을 참조하여 현재 질의에 대한 정답을 구하는 질의 등 수준이 높아진 질의를 대상으로 평가집합을 구축하였다. 또한 TREC에서는 TREC-10부터 TREC-14까지의 질의/응답 평가집합의 구축에 있어 향후 5년 간의 평가 내용에 대한 계획을 제시하였다 [9]. 이들 5년 간의 주된 평가 내용은 문맥 내의 질의/응답, 정답의 요약 및 생성, 전문지식을 이용한 전문가 수준의 질의에 대한 정답 생성 등이다.

국내에서는 KorQATeC1.0 [3]이 구축되어 한국어 질의/응답 시스템의 평가집합이 구축되었으며, 이를 토대로 질의/응답 시스템에 대한 연구 [4]와 새로운 평가집합의 구축이 진행되고 있다.

본 논문에서는 KorQATeC1.0에 기반한 새로운 평가집합인 KorQATeC2.0에 대하여 기술한다. KorQATeC2.0

은 크게 다음과 같은 3가지 특성을 가지고 있다.

- ▶ 네 가지 종류의 세부 과제 (기본 과제, 나열 과제, 문맥 과제, 요약 과제)
- ▶ 질의에 대한 정답이 문서 집합에서 나타나지 않는 무응답 질의 포함
- ▶ 12가지 세부분류에 따른 질의의 다양성

평가집합은 구체적인 대답을 요구하는 질의집합, 질의에 대한 검색대상이 되는 문서집합, 질의에 대해서 문서가 정답을 포함하고 있는지의 여부에 대한 평가와 정답을 표시하는 적합성 판정집합으로 구성되어 있다.

2 평가집합 구축 과정

평가집합 구축은 문서 집합 생성, 질의 생성, 대답포함 문서 후보집합 생성, 적합성 평가집합 생성의 4단계로 이루어져 있다.

2.1 문서집합 생성

정보검색의 대상으로 할 문서를 수집하여 형식에 따라 가공한다. 문서집합은 207,067개의 문서로, 1992년에서 1995년까지의 신문기사이다. 정치, 경제, 증권, 사회, 국제, 정보통신, 문화생활, 스포츠 등 다양한 분야의 내용이 포함되어 있다. KorQATeC2.0에서는 KorQATeC 1.0에서 생성한 문서집합과 동일한 문서집합을 사용하였다.

2.2 질의 생성

질의/응답에 적합한 구체적인 대답을 유도할 수 있는 질의를 만든다. 1992년에서 1995년 사이의 주요 사건과 웹에서 질문에 대한 정답을 맞추는 퀴즈문제들의 데이터베이스와 문서내용을 기반으로 하여 질의를 생성하였다.

KorQATeC1.0에서는 육하원칙(누가, 언제, 어디서, 무엇을, 왜, 어떻게)의 6가지 질의 유형에 질의가 고루 분포하도록 질의를 생성하였는데 KorQATeC2.0에서는 질의/응답 시스템이 더 다양한 질의 형태를 다룰 수 있도록 하기 위해 Arthur Graesser 의 18가지 질의 분류 [9]를 기준으로 하여 그 중 12개의 분류에 대해 질의를 생성하였다. 질의 유형에 대해서는 3장에서 자세히 설명한다.

질의는 우선 260개를 만들어 이 중 질의의 유용성과 난이도를 고려하여 120개를 최종적으로 선택하였다. 유용성은 질의가 실제로 사용될 것 같은지를 판단하는 척도로 높을수록 좋은 질의이며, 난이도는 질의/응답 시스템이 질의에 대한 정답을 찾기 쉬운지를 나타내는 척도로 너무 한쪽으로 치우치지 않아야 한다.

2.3 대답포함문서 후보집합 생성

질의에 대한 적합한 대답어구가 문서에 들어있는지에 대한 적합성을 판정하기 위해서는 모든 문서를 읽고 적합성을 판단하는 것이 가장 정확한 방법이다. 그러나 문서의 개수가 많은 경우에는 이것을 모두 읽고 적합성을 판단하기란 매우 어려운 일이기 때문에 KorQATeC2.0에서는 정보검색시스템을 이용하여 질의에 대한 대답이 포함되어 있을 가능성이 높은 후보문서집합을 생성하여, 평가자가 판정할 문서의 수를 줄이는 방법을 이용하였다.

다수의 정보검색시스템을 이용하여 질의에 대해 검색을 수행하고, 각 시스템의 검색결과에서 높은 순위를 갖는 문서들을 조합하여 적합문서 후보문서집합을 생성하고 각 문서들에 대해서 적합성 여부를 판단하는 방법[7]으로 특성이 다른 다수의 정보검색시스템에 의해 검색된 문서들 중에서 상위 N개의 검색결과의 합집합이 질의에 대한 모든 관련문서를 포함한다고 가정한다.

검색결과 조합 방법을 적용하기 위해서는 다수의 정보검색시스템이 필요하다. 질의에 대해 다양한 검색결과를 생성하기 위해 형태소단위의 색인/바이그램 색인/적합성 피드백 등을 이용한 숭실대 정보검색기, 2-포아송 모델로 가중치를 할당한 바이그램 색인을 이용한 날리지큐브(http://www.kcube.co.kr), 다양한 가중치 기법을 적용한 SMART 시스템 (ftp://ftp.cs.cornell.edu/pub/smart)으로 색인, 검색방법, 적합성 피드백 유무 등을 달리하여 18가지의 서로 다른 검색집합을 생성하였다. 이 검색결과 집합에서 각각 상위 70개의 문서를 추출한 후 중복되는 문서를 제거하여 대답포함문서 후보집합을 생성하였다.

2.4 적합성 평가

2.3장에서 생성된 적합문서 후보문서집합에 있는 문서에 대해서 적합성을 판단한다. 질의/응답 검색에서의 적합성 평가는 사람이 질의에 대해 문서의 내용을 읽고, 질의에 대한 대답이 포함되어 있는지의 여부를 판단하고 대답으로 가능한 것들을 추출한다. 대답포함문서 후보집합에 포함되어 있지 않은 문서는 잠정적으로 적합하지 않은 것으로 간주한다.

1) 적합성 평가를 위한 평가자

적합성 판정을 위한 평가자는 10명으로 구성되었고, 각 질의에 대해서 2명이 상호 독립적으로 같은 질의에 대해 평가를 수행하였다. 두 사람의 판정이 다른 경우에 대해서는 최종 의견 교환을 통해서 확정지었다.

2) 평가자를 위한 적합성 평가 기준

질의에 대한 대답을 평가할 때 평가자의 사전 지식이 아닌 문서내용을 중심으로 대답 여부를 판단한다. 고려할 사항은 다음과 같다.

- ▶문서의 내용에서 기준이 되는 날짜는 <DATE>에 표기된 신문기사가 작성된 날짜로 한다.
- ▶대답을 지원하는 정보가 문서에 나타나 있어야 한다.
- ▶대답이 틀린 경우라도 질의에 대한 대답을 지원하는 정보가 문서에 있으면 정답이다.
- ▶질의에 나와 있는 조건을 모두 만족하여야 정답이다.
- ▶정답은 하나의 문서에서 하나 이상이 나타날 수 있다.

3) 적합성 평가 결과 형식

적합성 판정 결과형식은 다음과 같다.

[질의번호 문서번호 : 적합성 판정값 대답문자열*]

- ▶적합성 판정값: 해당 문서에 질의에 대한 정답을 포함하고 있으면 1, 정답이 포함되어 있지만 불확실한 정보이거나 애매한 경우는 0, 그렇지 않으면 -1 값을 준다..
- ▶대답열 문자열: 적합성 판정값이 1인 경우 질문에 대한 대답을 표 1과 같이 4가지로 나누어 기술한다.

<A>대답</A>	질의가 요구한 대답
<PA>대답</PA>	나온 대답들의 비교 과정이 필요한 부분 대답, 나열 과제 대답, 요약과제 대답
<DA>대답</DA>	문서에 표현된 형태가 아니라 유추한 대답
<CA>대답</CA>	문서에 표현된 상태 그대로의 대답

표 1. 적합문서에 사용된 대답열 태그

{<PA>, </PA>}는 질의가 요구하는 대답이 아닌 정보가 나타나 있어 대답을 만들기 위한 추가적인 과정이 필요한 경우나 해당 대답이 전체 대답의 일부만을 나타내는 경우에 그에 해당하는 대답을 실제 질의가 요구한 대답을 나타내는 {<A>, </A>}와 구분해 주기 위하여 사용된다.

<질문> 우리나라보다 사망률이 낮은 나라는?
문서: HRM930313-11
<TEXT> …. 우리나라의 사망률은 1천명당 10명으로 세계 평균 97명보다 훨씬 낮으며 동아시아-태평양 지역의 14개 개발도상국가중 홍콩(7명), 싱가포르(8명)에 이어 3위인 것으로 나타났다. … </TEXT>
대답 평가:
HRM930313-11 : 1
<PA> [1993년] 우리나라 1천명당 10명 </PA>
<PA> [1993년] 홍콩 7명 </PA>
<PA> [1993년] 싱가포르 8명 </PA>
<A> [1993년] 홍콩 </A>
<A> [1993년] 싱가포르 </A>

<질문> 석유를 수출하는 나라를 5개 적으시오.
문서: HRM930315-15
<TEXT> …. 이라크의 주 수출품은 석유이다. … </TEXT>
대답 평가:
HRM930315-15 : 1
<PA> 이라크 </PA>

표 2. {<PA>, </PA>}의 사용 예

{<DA>, </DA>}, {<CA>, </CA>}는 문서에서 대답은 실제 질의가 요구한 대답이 아니지만 문서의 내용을 통하여 유추를 할 경우 질의가 요구한 대답이 될 수 있는 답을 표시할 때 사용된다.

<질문> XXX가 노벨평화상을 수상한 해는?
문서: HRM940713-61
<DATE> 1993년 3월 2일 </DATE>
<TEXT> 지난해 노벨 평화상을 받은 XXX ……. </TEXT>
대답 평가:
HRM940713-61: 1
<CA> 지난해 </CA> <DA> 1992년 </DA>

표 3. {<CA>, </CA>}, {<DA>, </DA>}의 사용 예

3 질의 유형

질의가 요구하는 정답의 유형에 따라 12가지로 나누어진다. Arthur Graesser [9]의 18가지 질의 분류 중 중요하다고 생각되는 것을 고려하여 12가지만 선택하였다.

3.1 개념의 완성 (Concept completion)

개념의 완성은 '누가', '언제', '어디서', '무엇'을 요구하는 질의 유형을 말한다. 영어로 표현할 경우 'who', 'when', 'where', 'what' 등의 의문사로 표현되는 질의가 이러한 유형에 속한다. 대부분의 질의가 여기에 속한다.

예) 동의보감의 저자는?

3.2 양 (Quantification)

어떤 것의 수, 양을 묻는 질문으로 높이, 수량, 크기, 기간, 수치, 사람 수, 거리, 순위, 가격 등을 요구하는 질의이다.

예) 미얀마의 카렌족 인구수는 얼마인가?

3.3 확인, 검증 (Verification)

사실의 진실성여부나 사건의 발생여부를 '예'나 '아니오' 로 대답해야 하는 질의이다.

예) 국회의원의 임기는 5년인가?

3.4 양자택일 (Disjunctive)

둘 혹은 셋 이상에서 하나 이상을 선택하는 질의이다.

예) 서울의 강수량은 겨울에 증가합니까 감소합니까?

3.5 가능하게 하는 것 (Enablement)

'어떤 객체(자원)가 그런 행동을 가능하게 하는가?'를 묻는 질의가 이 유형에 속한다.

예) 이 고기를 굽기 위해 무엇이 필요한가?

3.6 비교 (Comparison)

'X는 Y와 얼마나 비슷한가, X는 Y와 얼마나 다른가?' 와 같은 질의가 이 유형에 속한다.

예) 어떤 면에서 한국은 중국과 비슷한가?

3.7 설명, 정의 (Definition)

사물이나 사실의 정의를 질의의 정답으로 요구하는 질의이다.

예) CPU 란 무엇인가?

3.8 예 (Example)

질의가 요구하는 정답이 어떠한 사물이나 사실의 예인 경우의 질의를 가리킨다.

예) 민속놀이의 종류에는 어떤 것들이 있는가?

3.9 선행 원인 (Causal antecedent)

어떠한 사건이나 사물의 원인을 요구하는 질의가 이 유형에 속한다.

예) 우주선 충돌에 대한 주요 원인은 무엇인가?

3.10 원인 결과 (Causal consequence)

어떠한 사건이나 상태의 결과를 묻는 질의가 이 유형에 속한다.

예) 인플레이션의 결과는 무엇인가?

3.11 목적성 (Goal orientation)

어떠한 행동의 동기나 목적을 정답으로 요구하는 질의가 이 유형에 속한다.

예) 시내의 택시들의 목적은 무엇인가?

3.12 해석 (Interpretation)

특정 사건에 대한 설명이나 요약, 해석 등을 요구하는 질의가 이 유형에 속한다.

예) 이 그래프는 A에 대한 주된 효과를 보여주는가?

4 과제 분류

KorQATeC2.0 에서는 질의에 대한 대답을 제시하는 방식에 따라 질의를 기본 과제, 나열 과제, 문맥 과제, 요약 과제 네 가지로 나누고 그 과제의 특성에 맞게 평가 집합을 구축하였다.

4.1 기본 과제 (Main task)

질의는 하나의 대답을 요구한다. 질의에 대한 대답이 여러 개가 나온다 할 지라도 그 중에 하나만 제시하면 된다. 단일문서에서 간단히 정답을 찾을 수 있는 것도 있지만 여러 문서에서 얻은 정보를 이용하여 대답을 만들어야 하는 경우도 있다. 앞서 설명한 다양한 질의 유형이 포함되어 있다. 문서집합에 정답이 나타나 있지 않은 질의도 있어서 이러한 질의에 대해서는 "문서집합 내에 정답 없음"이라고 제시하여야 한다. 질의는 92개이며 다양한 유형의 질의로 구성되어 있다.

4.2 나열 과제 (List task)

질의는 하나의 대답만을 요구하지 않고 정해진 수의 정답을 요구하거나 가능한 정답을 모두 제시하라고 요구

한다. "석유를 수출하는 나라를 5개 적으시오." 와 같이 5개의 정해진 수의 정답을 요구하는 질의가 여기에 속한다. 요구하는 정답들은 한 문서에서 다 뽑을 수도 있지만 보통 여러 문서에서 해당하는 정답을 모아 정해진 수를 채워야 한다. 10개의 질의로 구성되어 있으며 보통 서너 개의 정답을 요구한다.

4.3 문맥 과제 (Context task)

주어진 주제에 관한 연속된 질의가 주어진다. 질의에 포함된 대명사 등을 조응해소하고 앞선 질의의 내용과 앞서 구한 정답을 이용하여 현재 질의에 대한 정답을 찾아 나간다. 다음은 질의의 예이다.

- 평화의 댐의 초기건설 목적은 무엇인가?
- 설계된 규모는 어느 정도인가?
- 이 정도 규모는 우리나라의 댐 중 얼마나 큰 것인가?

두 개의 주제, 총 9개 질의로 구성되어 있다.

4.4 요약 과제 (Summarization task)

어떤 것의 설명이나 이유를 요구하는 것과 같이 단답형으로 간단하게 나타낼 수 없는 질의들이다. 문서에서 질의에 관련된 정보를 추출하고 문장을 요약 생성하여 필요한 정보만을 간단명료하게 보여주어야 한다. "금융실명제의 실시 목적은 무엇인가?" 와 같은 질의가 여기에 속하고 총 9개의 질의로 구성되어 있다.

5 평가집합 분석

5.1 적합성 판정 결과

2.3장에서 얻어진 대답포함문서 후보집합 중 2.4장의 적합성 판정결과 실제 대답을 포함하고 있는 것으로 판정된 대답포함문서의 비율을 구하였다. 그 분포는 그림 1과 같고 비교 평가를 위해 KorQATeC1.0에서의 분포도 같이 나타내었다.

대답포함문서의 비율이 높다는 것은 후보문서 중에 많은 문서에 대답이 나타나 있다는 뜻이다. 질의에 적합한 대답이 많은 문서에 나타나 있으므로 그만큼 정답을 포함한 문서를 찾기 쉽다는 것을 의미한다. 반면 대답포함문서의 비율이 너무 낮으면 후보문서에서 적합한 문서가 거의 없어서 정답은 커녕 정답을 포함한 문서조차 찾기 어렵게 된다. KorQATeC2.0은 KorQATeC1.0 보다 대답포함문서의 비율이 조금 높음을 볼 수 있는데 대답포함문서가 많아 나열 과제, 요약 과제 질의와 같이 여러 문서에 걸쳐 있는 정보를 처리하여 정답을 만들어 내기에 좀 더 적합하다고 할 수 있다. 다양한 질의 유형과 위에서 언급한 여러 과제에 속한 질의를 다루기 위해서는 대답 포함 문서 내에서 특별한 정답추출 기법이 요구되는데 KorQATeC2.0은 대답을 포함한 문서 자체를 찾는 어려움을 줄여 이에 더 적합한 평가집합이라고 볼 수 있다.

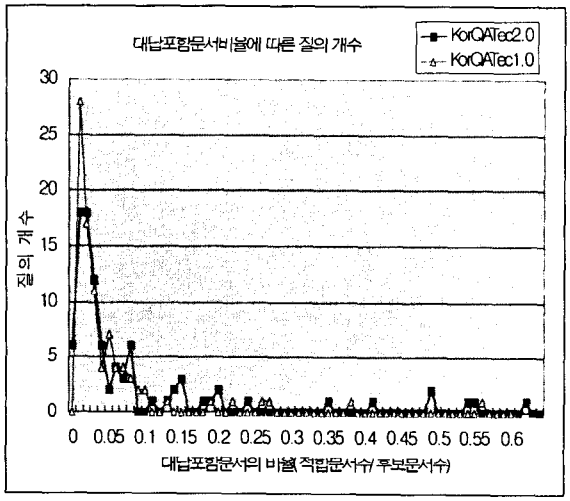

그림 1. KorQATeC1.0과 KorQATeC2.0의 후보문서 내 정답포함문서의 비율

5.2 질의유형별 정답포함문서 검색 비율

질의/응답 시스템은 주어진 질의에 대하여 문서를 검색하고, 상위 문서를 분석함으로써 정답을 추출한다. 따라서 검색된 상위 문서에 정답이 존재하는지는 질의/응답 시스템의 정답 추출에 있어 중요하다. 따라서 질의의 난이도는 문서 검색 결과에서 정답을 포함한 문서의 검색 비율을 비교함으로써 분석 가능하다.

질의유형에 따른 문서 검색 비율을 알아보기 위해 각 질의유형에 대해 검색조합깊이를 5, 10, 15, 20으로 변화시키면서 정답포함문서를 검색할 비율을 살펴보았다. 그림 2는 각 질의유형별 적합문서 검색 비율을 나타낸다. 그림에서 R(5)는 검색조합깊이 5에서의 재현율을, R(10)은 검색 조합깊이 10에서의 재현율을 나타낸다. 원인결과 질의의 경우 비교적 재현율이 낮게 나타난 반면 목적성질의나 선행원인 질의는 R(5)에서도 높은 재현율이 나타났는데 이는 목적성과 선행원인에 대한 질의는 질의 자체만을 가지고도 쉽게 적합문서를 검색할 수 있다는 것을 의미한다.

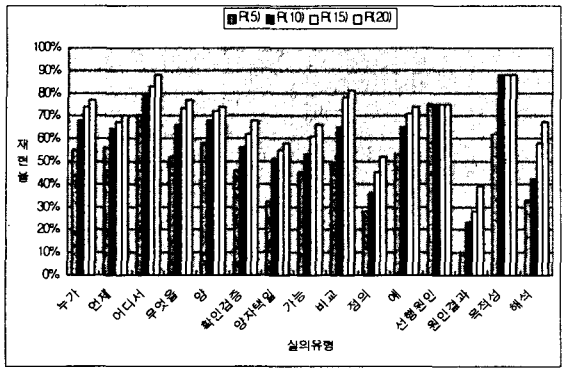

그림 2. 질의유형별 정답포함문서 검색 비율

5.3 과제별 정답포함문서 검색 비율

그림 3은 각 과제별 적합문서 검색 비율을 나타낸다. 문맥 과제의 경우 기본 과제에 비해 높은 정답문서 검색 비율이 나타났는데 이는 질의에서의 조응해소와 이전 질의의 답 등을 질의에 추가하여 검색을 하였기 때문이다. 실제 질의를 그대로 사용할 경우 매우 낮은 재현율을 보일 것이다. 요약 과제의 경우 비교적 낮은 적합문서 검색 비율이 나타났는데 이는 요약 과제의 질의가 정답생성은 물론 관련 문서 추출도 어렵다는 것을 보여 준다.

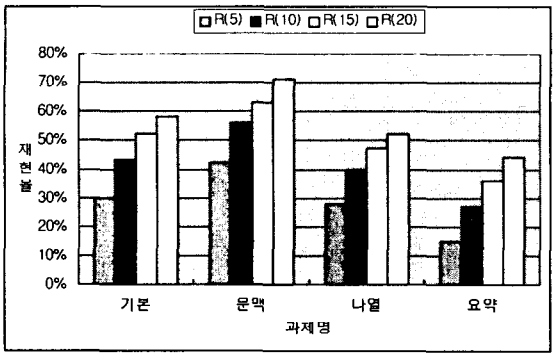

그림 3. 과제별 정답포함문서 검색 비율

5.4 나열과제 질의 및 정답 분석

1) 문서에 나타난 정답과 질의가 요구한 정답 비율 분석

나열 과제에서 질의가 요구한 정답과 문서에 나타난 정답 비율을 분석함으로써, 나열 질의의 난이도를 분석할 수 있다. 즉, 질의가 요구한 정답보다 문서에 많은 개수의 정답이 포함될수록 질의/응답 시스템이 해당 정답을 찾기가 보다 쉬울 것이며, 질의가 요구한 정답과 같은 개수의 정답만이 문서에 존재할 경우, 질의/응답 시스템이 정답을 추출하기가 보다 어려울 것이다. 표 4는 나열 과제의 질의가 요구한 정답의 개수와 문서에서 나타난 실제 정답의 개수를 나타낸다.

질의 번호	질의가 요구한 정답의 개수	문서에서 나타난 실제 정답의 개수
301	3	4
302	4	8
303	4	8
304	6	12
305	5	7
306	4	4
307	4	4
308	5	14
309	4	4
310	3	3
평균	4.2	6.8

표 4. 나열 과제의 질의가 요구한 정답의 개수와 문서에서 나타난 실제 정답의 개수

2) 정답을 추출하기 위한 문서의 개수

나열 과제에서 정답은 하나의 문서에서 모두 나타나는 경우도 있지만 일반적으로 여러 문서에서 부분적으로 존재하는 경우가 많다. 따라서 해당 질의에 대한 정답이 몇 개의 문서에 걸쳐 나타나는가는 질의/응답 시스템이 정답을 추출하는데 중요한 요소가 될 수 있다.

표 5는 정답을 추출하기 위한 최소 문서 수와 적합문서 수를 나타낸다.

질의 번호	정답을 추출하기 위한 최소 문서 수	적합문서 수
301	1	8
302	4	41
303	1	3
304	2	11
305	1	98
306	2	8
307	1	111
308	1	91
309	4	9
310	3	15
평균	2.0	39.5

표 5. 나열 과제의 질의에서 정답을 추출하기 위한 최소 문서 수와 적합문서 수

5.5 요약 과제 질의 및 정답 분석

요약 과제의 목적은 산재되어 있는 정보를 사용자가 요구하는 질의에 따라 정리하여 제시해주는 것을 목적으로 한다. 따라서 요약 과제에 있어 고려되어야 할 사항은 요약문의 길이, 정답을 추출하기 위한 문서의 개수, 질의가 필수적으로 요구하는 정보의 개수이다.

1) 요약문의 크기

표 6은 각 질의에 대한 요약문의 크기를 나타낸다. 요약문은 평균 2문장 40어절 정도로 요약된다.

질의번호	401	402	403	404	405	406	407	408	409	평균
문장 수	2	3	1	2	2	4	1	5	4	1.89
어절 수	38	46	19	20	30	46	33	67	64	40.33

표 6. 요약문의 크기

2) 정답을 추출하기 위한 문서의 개수

요약문을 구성하기 위한 정보가 여러 문서에 산재되어 있다면 단일 문서에 모든 정보가 포함되어 있는 경우보다 요약문을 구성하기가 어려울 것이다. 따라서 정답을 추출하기 위한 문서의 개수는 질의의 난이도를 유추하기 위한 지표로 사용될 수 있다.

표 7은 요약 질의에 대한 요약문을 구성하기 위해 필요한 문서 수를 나타낸다. 질의 번호 403의 경우 적합문서 수가 200여개로 가장 많은 반면 정답을 구성하기 위한 문서는 2개뿐이다. 따라서 많은 적합문서에서 나타나는 중복되는 정보나 모순되는 정보를 처리하는 기법이 필요하다.

질의 번호	적합문서수	정답을 추출하기 위한 최소 문서수
401	3	2
402	26	1
403	223	2
404	85	8
405	78	4
406	18	2
407	57	2
408	7	4
409	13	2

표 7. 요약 과제의 질의에서 정답을 추출하기 위한 최소 문서수와 적합문서수

3) 정답을 구성하기 위한 필요 요소 수

질의가 필수적으로 요구하는 정보를 완벽하게 제시하는 것은 요약문의 정확성을 판단하는 기준이 될 수 있으며, 이는 요약 과제의 질의/응답 시스템을 평가하는데 중요한 기준이 된다. 질의가 요구하는 사항이 적다면 질의/응답의 난이도가 비교적 낮을 것이고, 질의가 요구하는 사항이 많다면 질의/응답의 난이도가 보다 높을 것이다.

표 8은 각 질의가 요구하는 필수 정보 수를 나타낸다. 괄호 안의 수는 필수 정보 중 유추를 통하여 생성해야 하는 정보 수로 질의/응답 시스템의 유추 기능에 대한 평가를 하는데 사용될 수 있다.

질의번호	401	402	403	404	405	406	407	408	409	평균
필수 정보 수	3 (0)	3 (0)	2 (0)	5 (0)	5 (1)	5 (1)	3 (1)	4 (1)	3 (1)	3.67 (0.56)

표 8. 요약 과제의 질의가 요구하는 필수 정보 수

6 결론

본 논문에서는 질의/응답 시스템의 평가를 위한 평가집합을 구축하였다. 질의/응답 시스템 성능 평가를 위한 평가집합은 207,067개의 문서, 4개의 과제 120개의 질의, 각 질의에 대한 적합성 판정 집합으로 구성되어 있다. 질의 120개에 대해 적합성 판정을 한 문서집합은 22,898개 문서이다.

문서는 신문기사로 다양한 분야의 내용을 포함하고 있다. 평가집합은 질의에 대한 대답을 제시하는 방식에 따라 기본 과제, 나열 과제, 문맥 과제, 요약 과제 네 가지로 나누어진다. 또한 질의가 요구하는 정답의 유형에 따라 12가지의 다양한 질의 유형을 고려하여 평가집합을 구축하였다. 적합성 판정 집합은 각 질의에 대해서 문서에 대답을 포함하는지의 여부에 따라 적합/부적합/애매함으로 판정하였고, 적합한 문서에 대해서는 대답을 표시하였다.

본 논문을 통해 구축된 질의/응답 시스템 성능 평가를 위한 평가집합은 보다 다양한 질의에 대한 평가를 가능하게 하였으며, 영어권에서 구축하고 있는 질의/응답용 평가집합 수준에 이르고 있기 때문에 향후 질의/응답 시스템의 연구에 효율적으로 사용될 수 있을 것이다.

구축된 평가집합은 정보검색분야 연구자 및 개발자들에게 공개함으로써, 한국어 질의/응답 시스템의 연구에 적용되어 질의/응답 시스템 구축을 활성화시킬 수 있을 것으로 기대된다.

감사의 글

본 논문은 전문용어언어공학연구센터에서 수행한 과학기술부와 KISTEP의 핵심소프트웨어사업 중 "대용량 국어정보 심층처리 및 품질관리 기술개발" 과제의 일환으로 수행되었으며, 부분적으로 첨단정보기술연구센터를 통하여 과학재단의 지원을 받았습니다.

참고문헌

[1] 맹성현, 장동현, 송사광, 김지영, 이석훈, 이준호, 이용봉, 서정현, 1999. "정보검색 테스트 컬렉션 구축 및 유효성 평가", 한글 및 한국어 정보처리학회.

[2] 박영찬, 최기선, 김영환, 김재군. 1996. "한국어 정보검색 연구를 위한 시험용 데이터 모음 2.0(KTSET 2.0) 개발. 한국어정보과학회 인공지능연구회 춘계학술 발표. pp.59~65.

[3] 이경순, 김재호, 최기선, 2000. "질의/응답 시스템의 성능 평가를 위한 테스크컬렉션 구축", 한글 및 한국어 정보처리학회. pp.190~197

[4] 이경순, 김재호, 최기선, 2001. "KorQuA: 질의응답에서 자료유형을 고려한 대답검색과 대답해석", 한국인지과학회 춘계 학술대회. pp.73~78

[5] 이준호, 최광남, 한현숙, 김종원, 남성원, 1995. "정보검색을 위한 KRIST 테스트 컬렉션의 개발", 한국정보과학회.

[6] CLEF. 2000. Cross-Language Evaluation Forum http://galileo.iei.pi.cnr.it/DELOS/CLEF/clef.html

[7] Harman, Donna. 1995. "Overview of the Fourth Text REtrieval Conference(TREC-4)", In proceedings of TREC-4. http://trec.nist.gov/pubs/trec4/t4_proceedings.html

[8] Hersh, William. 1994. "OHSUMED:An Interactive Retrieval Evaluation and New Large Test Collection for Research". In Proceedings of the 17th Annual International ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.192-201.

[9] John Burger, Claire Cardie, Vinay Chaudhri et. al. 2001. "Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)"

[10] Kando, Noriko and Kuriyama, Kazuko. 1999. Toshihiko Nozue, "The NTCIR Workshop(NTCIR-1)", In Proceedings of the 22nd Annual International ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.299-300

[11] TREC, homepage. http://trec.nist.gov

[12] Voorhees, Ellen M. and Harman, D. 1998. "Overview of the Seventh Text REtrieval Conference(TREC-7)", In proceedings of TREC-7. http://trec.nist.gov/pubs/trec7/t7_proceedings.html

[13] Voorhees, Ellen M. and Tice, D. 1999. "The TREC-8 Question Answering Track Evaluation". In Proceedings of the TREC-8. http://trec.nist.gov/pubs/trec8/t8_proceedings.html

[14] Voorhees, Ellen M. 2000. "Overview of the TREC-9 Question Answering Track". In Proceedings of the TREC-9.

[15] TREC, 2001, "TREC 2001 Question Answering Track Guidelines", http://trec.nist.gov/act_part/guidelines/qa_track_spec.html