질의응답시스템의 성능 평가를 위한 테스트컬렉션 구축

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

Construction of Test Collection for Evaluation of Question Answering System

Kyung-Soon Lee, Jae-Ho Kim, Key-Sun Choi
KORTERM, AITRC, KAIST

요약

본 연구에서는 사용자의 질의에 대해 대답을 제시하는 질의응답시스템의 평가를 위한 테스트컬렉션을 구축하였다. 질의응답시스템 평가를 위한 테스트컬렉션은 207,067개의 문서, 90개의 질의, 각 질의에 대한 적합성 판정 집합으로 구성되어 있다. 문서집합은 신문기사로 SGML 형식으로 가공되었고, 질의는 다양한 유형의 질의와 변형질의를 포함한다. 적합성 판정 집합은 각 질의에 대해서 문서에 대답을 포함하는지의 여부에 따라 적합/부적합으로 판정하였고, 적합한 문서에 대해서는 대답을 표시하였다. 본 연구를 통해 구축된 질의응답시스템 평가를 위한 테스트컬렉션은 질의응답시스템의 객관적인 신뢰성 평가를 위한 기반을 마련하였다.

1 서론

사용자의 요구가 다양해짐에 따라 다양한 정보검색시스템이 개발되고 있다. 사용자의 현재의 관심에 따라 일반 문서에 대한 검색을 수행하는 일반문서검색시스템, 대규모의 웹문서에 대한 검색을 수행하는 웹문서 검색시스템, 질의와 문서의 언어가 서로 다른 경우의 검색을 수행하는 교차언어검색시스템, 질의에 대한 구체적인 대답을 제시하는 질의응답시스템, 사용자의 지속적인 관심에 대해 새로운 문서들 중에서 사용자의 관심에 맞는 부분을 걸러주는 정보필터링 시스템, 검색결과에 대한 요약을 하는 문서요약 시스템 등이 개발되고 있다.

다양한 검색시스템들에 대해 사용자의 끊이지 않는 요구는 정보검색시스템 사용의 편리, 신속한 검색과 신뢰성 있는 결과이다. 검색 신뢰도 평가는 같은 질의와 문서집합을 대상으로 하였을 때 객관적인 비교 평가가 가능하다. 이와 같이 정보검색시스템의 객관적인 신뢰도 평가를 위해서는 체계적으로 구축된 테스트컬렉션(test collection)이 필요하다.

현재 개발되어 있는 대부분의 정보검색시스템은 문서단위의 검색을 지원하고 있다. 문서검색시스템은 사용자의 질의에 대해 관련 있는 문서들을 결과로 제시한다. 사용자의 질의가 구체적인 대답을 요구하는 것일 경우에는 문서검색시스템에서의 결과는 사용자가 문서를 읽고 원하는 대답을 찾아야 하는 불편함이 있다. 따라서, 사용자의 질의에 대해 문서단위가 아니라, 문서에서의 구체적인 대답을 검색할 수 있는 시스템에 대한 요구가 증가하고 있다. 사용자의 질의에 대해 구체적인 대답을 찾아주는 시스템을 질의응답시스템이라고 한다.

국제적인 정보검색평가대회인 TREC (Text REtrieval Conference, http://trec.nist.gov)에서는 1990년 초부터 정보검색시스템의 평가를 위한 다양한 테스트컬렉션을 구축해오고 있다. 1999년 TREC-8에서 질의응답시스템의 평가를 위한 테스트컬렉션의 구축 [8]을 시작하였다. 2000년 TREC-9에서 그 수량을 확장하여, 893개의 질의

와 97만여개의 문서를 대상으로 한 대답을 구축하였다.

국내에서는 일반적인 문서검색시스템의 평가를 위한 테스트컬렉션으로는 KT set [3], Kemong set, KRIST set [2], HANTEC [1]이 개발되었다. 그러나, 질의응답시스템의 신뢰성을 평가를 위한 테스트컬렉션이 전무한 상태이어서 한국어 질의응답시스템의 성능을 평가하기 어려운 실정이다.

본 연구에서는 사용자의 질의에 대해 대답을 제시하는 질의응답시스템의 객관적인 평가를 위한 테스트컬렉션을 구축하였다. 테스트컬렉션은 구체적인 대답을 요구하는 질의집합, 질의에 대한 검색대상이 되는 문서집합, 그리고 질의에 대해서 문서가 정답을 포함하고 있는지의 여부에 대한 평가와 정답을 표시하는 적합성 판정집합을 포함하고 있다.

2. 질의응답 시스템 평가를 위한 테스트컬렉션

테스트컬렉션 구축을 위한 작업 절차는 다음과 같다.

- 문서 집합 생성: 질의응답의 대상으로 할 문서를 수집하여, 문서 형식에 따라 가공한다.
- 질의 생성: 다양한 의문유형의 질의와 자연언어 처리 문제를 반영할 수 있는 질의를 생성한다.
- 대답포함문서 후보집합 생성: 질의에 대해 대답이 들어있을 가능성이 높은 문서들을 추출한다. 현재 개발되어 있는 한국어 질의응답시스템이 없기 때문에, 기존에 개발되어 있는 정보검색시스템을 이용하여 질의에 대해 대답포함문서 후보집합을 생성한다.
- 적합성 평가집합 생성: 각 질의에 대해 대답포함문서 후보집합에 있는 문서를 대상으로 질의에 대한 정답이 포함되어 있는지를 판단하여 표시하고, 대답으로 가능한 대답어구를 추출한다.

2.1 문서집합

문서집합은 207,067개(740MB)의 문서로, 1992년에서 1995년까지의 신문기사이다. 문서에 나타나는 내용은 정치, 경제, 증권, 사회, 국제, 정보통신, 문화생활, 스포츠 등 다양한 분야의 내용을 포함하고 있다.

문서는 SGML 형태로 태그를 부착한다. 시작태그와 종료태그로 이루어진다. 다음은 문서에서 나타나는 태그와 태그의 의미를 나타낸다.

<DOC> </DOC> ; 문서의 시작/끝 표시
<DOCNO> </DOCNO> ; 문서의 고유 번호
<TITLE> </TITLE> ; 문서의 제목
<BYLINE> </BYLINE> ; 문서의 저자
<FIELD> </FIELD> ; 문서의 분야(신문기사 면)
<DATE> </DATE> ; 문서가 작성된 날짜
<TEXT> </TEXT> ; 문서의 내용

다음은 테스트컬렉션에 포함된 문서의 가공 예이다.

<DOC>
<DOCNO> HRM920509-22 </DOCNO>
<TITLE> 타지크 공산정권 붕괴 / 나비예프 실각... 혁명평의회 구성 </TITLE>
<BYLINE> 김지석 기자 </BYLINE>
<FIELD> HRM 04면 </FIELD>
<DATE> 1992년 05월 09일 </DATE>
<TEXT> [두샨베 모스크바=외신 종합] 강경 공산주의자인 라흐만 나비예프 타지크 대통령이 7일 실각하고 반정부 6개 단체의 연합체인 인민세력연합이 권력을 장악한 것으로 전해졌다.
<타지크라디오>는 이날 "인민세력연합이 사태를 장악하고 있으며 혁명평의회가 구성됐다"고 보도했다. 이로써 옛 소련 보수파의 쿠데타가 실패한 직후인 지난해 9월부터 불붙은 공산주의세력과 이슬람·자유주의연합세력의 권력투쟁은 공산주의세력의 패배로 일단락됐다.
</TEXT>
</DOC>

2.2 질의 집합

질의는 다양한 대답유형을 요구하는 질문을 포함하였고, 음차 표기/복원 문제, 복합명사 분리문제, 의미중의성 해결 문제 등과 같이 자연언어처리문제를 다룰 수 있도록 하였다. 또한 다양한 분야의 질의를 포함하고 있다.

가. 질의 생성

질의응답 검색을 위한 질의는 문서형태를 요구하는 것이 아니라, 구체적인 대답을 유도하는 질의로 한다. 질의의 수는 90개이다.

질의에 대한 정답을 맞추는 퀴즈문제들의 데이터베이스와 문서내용을 기반으로 하여 300개의 질의를 생성하였다. 그 중에서 다양한 의문유형의 질의, 문서에 질의에 대한 대답이 하나이상 나타나는 것으로 최종 90개를 선택하였다.

다음은 질의응답 검색을 위한 질의 예이다.

이스라엘 전 총리이면서, 노벨평화상을 수상한 사람은?

질의 유형	질의 개수	대답유형	세부 질의 유형
누구	15	사람	저자, 학설주장, 노벨상, 비행사, 암살범, 선수, 감독
어디	20	장소	수도, 국가, 도시, 사물의 위치, 생산국, 구체적 장소
언제	12	시간	년도, 날짜, 계절, 시대, 시기
얼마나	16	수량/크기	크기, 수, 높이, 기간, 인원수, 생산량, 가격, %
이유	2	원인	이유, 원인
무슨/어느	25		동물, 질병, 병명, 사건, 종교, 방법, 물질, ...

표 1. 테스트컬렉션에 포함되어있는 질의 유형 및 개수

한국의 최초 비행사는 누구인가?
이집트의 수도는 어디인가?
프레온가스의 생산과 소비를 규제하는 국제조약은?
'동의보감'의 저자는?
수십년이 지난 유골의 신원을 확인하는 방법은?

각 질의에 대해서 같은 대답을 요구하는 변형된 형태의 질의를 생성하였다. 질의의 변형을 통해서 질의에 나타나는 단어나 표현방식이 달라졌을 경우의 시스템의 검색능력을 평가할 수 있다.

질의의 변형 예)

<질의 10> 제25회 바르셀로나올림픽의 남자 마라톤에서 금메달을 딴 선수는?

<질의 10-1> 제 25회 올림픽의 남자 마라톤을 제패한 선수는?

<질의 10-2> 1992년 올림픽의 남자 마라톤에서 1위를 한 선수는?

나. 질의 유형

질의는 대답으로 사람이름, 시간, 화폐, 병명, 상품명, 길이/크기, 장소 등과 같은 대답을 요구할 수 있도록 한다. 누구(who), 어디(where), 언제(when), 어느(which), 무엇(what), 어떻게/얼마나(how) 등의 의문유형을 만든다. 질의 유형별 분포는 표1과 같다.

다. 질의 형식

질의 형식은 질의 번호, 질의 문장, 질의에 대한 대답의 판단기준을 명확히 하는 설명부분으로 구성한다. 다음은 질의 집합에 나타나는 태그와 태그의 의미를 나타낸다.

<num> 질의 번호

<question> 대답을 요구하는 질의

2.3 대답포함문서 후보집합 생성

질의에 대해 적합한 대답어구가 들어있는 문서를 판정하기 위해 평가자가 모든 문서를 읽고 대답을 추출하는 것이 가장 정확한 방법이다. 그러나, 문서의 개수가 많은 경우에는 아주 많은 시간을 요구하므로 현실적으로 불가능한 방법이다. 현재 대상으로 하고 있는 문서의 수가 20만여개이므로 이것을 모두 읽고 대답을 추출하기는 어렵다. 실용적인 방법으로, 정보검색시스템을 이용하여 질의에 대한 대답이 포함되어 있을 가능성이 높은 후보문서집합을 생성하여, 평가자가 판정할 문서의 수를 줄이는 방법을 이용하였다.

대답포함문서 후보문서집합을 생성하는 방법은 다수의 정보검색시스템을 이용하여 질의에 대해 검색을 수행하고, 각 시스템의 검색결과에서 높은 순위를 갖는 문서들을 조합하여, 그 문서들에 대해서 대답포함 여부를 판단하는 방법[5]이다. 현재 개발되어 있는 한국어 질의응답시스템이 없기 때문에, 기존에 개발되어 있는 문서검색시스템을 이용하여 질의에 대해 대답포함문서 후보집합을 생성하였다.

검색결과 조합 방법 (pooling method)은 특성이 다른 다수의 정보검색시스템에 의해 검색된 문서들 중에서 상위 N개의 검색결과의 합집합이 질의에 대한 모든 관련 문서를 포함한다고 가정한다. 검색결과 조합에 포함되어 있지 않은 문서는 질의에 대한 대답이 포함되어 있지않은 문서로 한다.

검색결과 조합 방법을 적용하기 위해서는 다수의 정보검색시스템이 필요하다. 질의에 대해 다양한 검색결과를 생성하기 위해 불리언 검색을 이용한 한미르 정보검색기 (http://www.hanmir.co.kr), 형태소단위의 색인/bi-그램 방식의 색인, 적합성 피드백 등을 이용한 숭실대 정보검색기, 벡터공간검색방법을 이용한 충남대 정보검색기, 다양한 가중치 기법을 적용한 SMART시스템을 이용하였다.

색인방법, 검색방법, 적합성 피드백 등을 이용하여 16가지의 서로 다른 검색집합을 생성하였다. 그림1은 검색결과 조합 방법을 이용하여 후보문서집합을 생성하는 과

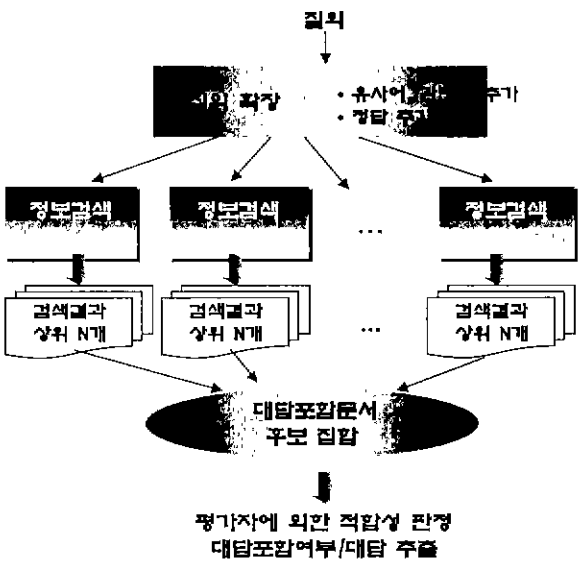

그림 1. 문서검색결과 조합방법을 이용한 대답포함문서 후보집합 생성 과정

정을 나타낸다.

테스트컬렉션에 포함된 각 질의에 대해서 질의에 대한 확장은 질의에 대해 가능한 대답어구를 포함시켰다. 적합성 판정 과정에서 질의에 대한 대답이 들어있는 부분을 추출하는 것이기 때문에 대답을 추가하여 질의를 확장함으로써, 질의와 대답이 모두 포함되어 있는 문서들을 상위로 검색하도록 하기 위한 것이다.

검색결과 조합을 위해서 각 검색결과 집합에서 상위로 순위화된 70개의 문서 (검색결과 조합의 크기 = 70)를 추출한다. K개의 검색결과에는 서로 중복되는 문서가 포함될 수 있으므로, 상위 70개의 문서들을 모두 합하여 유일한 개수의 문서를 추출한다.

이렇게 생성된 대답포함문서 후보집합에 대해 사람이 세세히 읽어보고 대답을 추출하게 된다.

2.4 적합성 평가

질의응답검색에서의 적합성 평가는 사람이 질의에 대해 문서의 내용을 읽고, 질의에 대한 대답이 포함되어 있는지의 여부를 판단하는 것이다. 그리고, 대답으로 가능한 것들을 추출한다. 사람이 모든 문서집합을 다 살펴볼 수 없기 때문에, 검색결과 조합방법을 이용하여 생성된 대답포함문서 후보집합에 있는 문서에 대해서만 적합성을 판단한다. 대답포함문서 후보집합에 포함되어 있지 않은 문서는 잠정적으로 적합하지 않은 것으로 간주한다.

가. 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.

나. 적합성 평가의 단계

적합성 판정은 질의에 대한 대답이 문서에 포함되어 있는지/없는지에 따라서 적합/부적합의 2단계 평가를 한다. 적합성 평가에서 갖는 값은 다음과 같다.

- 적합: 1 (질의와 대답이 포함되어 있는 문서
- 부적합: -1 (질의에 대한 대답이 포함되어 있지 않은 문서)

다. 적합성 평가 결과 형식

적합성 판정 결과형식은 다음과 같다.

[질의번호 문서번호:적합성 판정값 대답*]

- 적합성 판정값: 질의에 대해 각 문서번호에 대해서 적합여부를 1/-1로 표현을 한다.
- 대답문자열: 질의에 대한 대답이 포함되어 있는 문서 (적합 판정값 = 1)인 경우에는 질의에 대한 대답을 {<A> 대답 </A>}+ 형식으로 기술하였다. 대답이 여러개 나타날 경우 <A> 대답 </A> 형식을 반복 기술한다.

적합성 판정을 통해 질의에 대해서 평가자가 적합/부적합의 평가를 수행한 결과는 다음과 같다.

<번호> 11
<질의> 바르셀로나올림픽의 여자 양궁에서 금메달을 딴 선수는?
적합성 판정 결과 일부 예 :
11 DAE920730-2 : -1
11 DAE920801-2 : -1
11 DAE920802-4 : -1
11 DAE920803-17 : 1 <A>조윤정<A>
11 DAE920803-2 : 1 <A>조윤정</A>
11 DAE920803-61 : -1
11 DAE920805-2 : 1
<A>조윤정</A><A>김수녕</A><A>이은경 </A>
11 DAE920805-80 : 1 <A>조윤정</A><A>김수녕</A>
<A>이은경</A>
11 DAE920807-3 : -1
11 DAE920808-4 : -1
...

라. 평가자를 위한 적합성 평가 기준

질의에 대한 대답을 평가할 때 문서내용 중심적으로 대답 여부를 판단한다. 이때, 문서 집합에서 잠정적으로 가정하고 있는 것과 질의 집합에서 가정하고 있는 것을 기반으로 해서 판정을 한다.

가) 문서에서 잠정적으로 가정하고 있는 것

문서의 내용에서 기준이 되는 날짜는 신문기사가 작성된 날짜로 <DATE>에 표기된 날짜로 한다. 문서의 내용에서 '올해', '작년', '지난해' 등은 문서의 <DATE>를 기준으로 하여서, 질의에 대해 대답이 될 수 있는 근거가 되는지를 판단한다.

<질의> 1991년 노벨평화상을 탄 사람은 누구인가?
<문서 내용>
<DATE> 1992년 2월 1일
<TEXT> ... 작년 노벨평화상을 탄 사람은 ...이다.
<평가: 적합> 문서의 날짜인 '1992년'의 시점에서 '작년'은 '1991년'이기 때문에 질의의 내용을 뒷받침한다고 볼 수 있다.

나) 질의에서 잠정적으로 가정하고 있는 것

질의를 던지는 시점을 '현재'의 시간으로 <2000년 8월>로 한다. 문서의 내용에서는 현재인 상태이지만, 질의를 던지는 시점에서 문서의 내용이 현재인 것이라도 과거로 해석을 할 수 있는 것은 과거로 본다.

다) 문서 내용 중심적인 대답 찾기

(1) 대답을 지원하는 정보가 문서에 나타나야 한다. 그렇지 않은 문서에 나타나는 것은 대답이 아니라고 본다.

<질의> 2002년 월드컵을 개최하는 나라는?
<평가: 부적합> 월드컵 개최에 관련된 내용이 하나도 없는 문서에서 '한국', '일본'이 나타났을 경우에는 대답으로 하지 않는다.
<평가: 적합> 2002년 월드컵이 한국, 일본에서 개최된다는 것을 알 수 있는 내용이 같이 나와야 대답으로 한다

(2) 대답이 틀린 경우라도 질의에 대한 대답을 지원하는 정보가 문서에 있으면 대답으로 가능하다. 문서의 내용에서 대답이라고 판단할 수 있는 근거가 있으면 대답으로 한다. 사람이 알고 있는 사실이 아니라, 문서의 문맥에서 답을 판단해야 한다.

<질의> 2002년 월드컵을 개최하는 나라는?
<문서> .. 2002년 월드컵을 개최하는 중국에서는 ...
<평가: 적합> 질의에 대해 알고 있는 사실인 '한국', '일본'이 아니더라도, 문서내용이 질의를 뒷받침해주므로 '중국'을 대답으로 포함시킨다.

(3) 대답을 알지 못하는 사람이 질의를 던졌을 때, 질의와 관련된 문장/문단을 읽어보고 질의에 대한 궁금증을 해소할 수 있어야 대답으로 한다. 즉, 질의에 해당하는 내용과 대답이 포함되어 있는 문서라 할지라도 대답을 유추할 수 없는 경우에는 대답이 아니라고 판단함

<질의> '로미오와 줄리엣'을 쓴 영국의 대문호는 누구인가?
<문서 내용> # 괄순의 노 영문학자가 어린이를 위한 셰익스피어 작품집을 펴냈다. 평생을 교단에서 셰익스피어와 영미 희곡을 가르치는데 바친 김갑순씨(80. 전 이화여대-덕성여대 교수)가 최근 써낸 '이야기 셰익스피어'(정우사 간)는 그의 세 손자에게 주는 글. '로미오와 줄리엣' '햄릿' '리어왕' 등 6편의 비극과 '한 여름 밤의 꿈' '베니스의 상인' '뜻대로 하세요' 등 6편의 희극을 깔끔한 필치로 엮었다.
<평가: 적합> '로미오와 줄리엣'이 셰익스피어의 작품에 해당한다는 내용을 알 수 있으므로, 적합 문서로 판정한다.

<문서 내용> 이 프로그램은 '폴 고갱 스케치여행'을 비롯한 21개 독특한 테마로 구성돼있다. 해당분야 전문가가 동행하면서 자상하게 설명도 해준다. 톱 디자이너의 설명을 들으며 니나리치, 샤넬 같은 세계적 패션회사를 직접 방문해보고 파리 패션가 고급 매장을 둘러보는 패션여행. 영화전문가 안내를 받아 '로마의 휴일'이나 '벤허', '로미오와 줄리엣' 등 기억에 남는 영화가 촬영된 곳을 더듬어보는 시네마기행.
셰익스피어 생가나 T S 엘리어트, 바이런의 시가 탄생한 배경을 체험하는 문학기행등이 눈길을 끈다.
<평가: 부적합> '로미오와 줄리엣'이 셰익스피어의 작품이라는 것을 알수 없으므로, 부적합 문서로 판정한다.

(4) 질의가 두 가지 이상의 조건을 모두 만족해야 하는 것인지를 고려해서 판단한다

<질의> 이스라엘 전 총리이면서, 노벨평화상을 수상한 사람은 누구인가?
<평가기준> "이스라엘의 총리를 지냈다"는 내용과 "노벨평화상을 수상했다" 내용이 모두 나오는 경우 대답으로 함. 둘 중 하나만 나오는 경우는 틀린 대답으로 함

<문서 내용> 지난 77~83년 이스라엘총리를 지낸 메나헴 베긴은 78세로 숨졌다. 이스라엘 건국을 위해 싸웠던 열렬한 시온주의자로서 그는 지난 78년 이집트와의 평화를 일궈내는데 성공해 노벨 평화상을 받기도 했다. ·
<평가: 적합> 조건 만족

<문서 내용> [예루살렘=AFP 연합] 메나헴 베긴 이스라엘 전 총리가 3일 아침 의식을 잃어 병원으로 긴급 이송됐으나 중태이며 현재 인공호흡기구에 의존하고 있다고 의료진들이 말했다.

<평가: 부적합> 조건을 충분히 만족하지 않음(노벨평화상을 받았는지 알 수 없음)

(5) 대답은 하나의 문서에서 하나 이상이 나타날 수 있다.

3. 테스트컬렉션 분석

3.1 적합성 판정 결과

적합성판정 결과에 대해서 각 질의에 대해 대답포함문서 후보집합의 크기, 대답포함문서의 개수, 대답을 포함하는 부분의 개수에 대한 분포는 그림2와 같다. 전체 적합성 판정에 참여한 문서의 수는 22,898개 문서이다.

하나의 문서에는 대답이 1개이상 가능하므로, 여러개의 대답을 포함할 수 있다. 따라서, 대답어구의 개수는 대답포함문서 개수 이상이 가능하다.

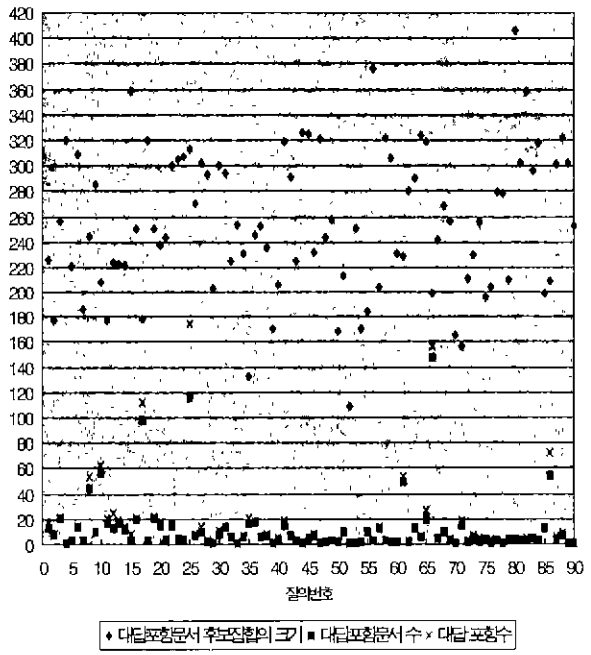

그림 2. 각 질의에 대한 대답포함 후보문서 수, 대답포함문서 수, 대답포함 수

3.2 검색결과 조합에서 상위개수의 변화에 따른 적합문서 후보집합의 크기 비교

정보검색기의 검색결과를 조합할 때, 상위문서의 개수 N을 1, 2, 3, ..., 67, 68, 69, 70로 변화시켰을 때의 변화를 살펴본다. N의 수가 너무 크면, 평가자가 판단해야할 문서의 수가 많아지고, N이 너무 작으면 대답포함문서를 포함하지 않을 수 있기 때문에, 적절한 N을 선택하는 것은 중요하다.

그림3은 상위 N의 변화에 따른 가능한 참여할 수 있는 문서의 크기, 대답포함문서 후보크기와 대답포함문서 수를 나타낸다.

가능한 문서 크기는 정보검색 결과집합 16개 (색인방법과 검색방법의 변화에 따라 생성된 검색결과 집합)에서 제시하는 상위문서 개수를 곱한 값이므로 N*16 이고, 대답포함문서 후보크기는 가능한 문서크기에 포함된 문서들의 유일한 개수이다. 대답포함문서 수는 대답을 포함하고 있는 문서의 개수이다.

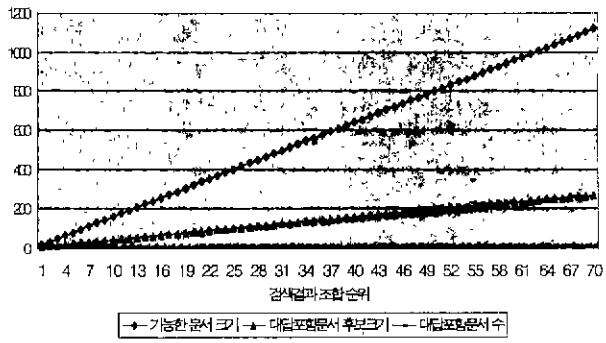

그림 3. 검색결과 조합에서 상위문서 개수의 추가변화에 따른 대답포함문서 후보크기 및 대답포함문서 크기 변화

3.3 검색결과 조합에서 상위문서 개수의 포함 변화에 따른 평균 대답포함문서 추가 분포

문서에 대답이 포함되어 있는지의 여부를 평가할 때, 대답포함문서 후보집합에 대해서 평가를 하였다. 사람이 평가한 문서 수의 제한으로 인해 대답포함문서가 누락될 가능성을 살펴본 것이다.

전체 질의에 대해서 검색결과에 대해서 상위문서를 늘려가면서 조합을 했을 경우에, 대답포함문서가 추가 발견되는 비율을 그림4에 나타내고 있다.

상위문서 수의 개수를 70으로 했을 때, 새로이 추가되는 문서가 거의 나타나지 않음을 알 수 있다. 그러나, 어

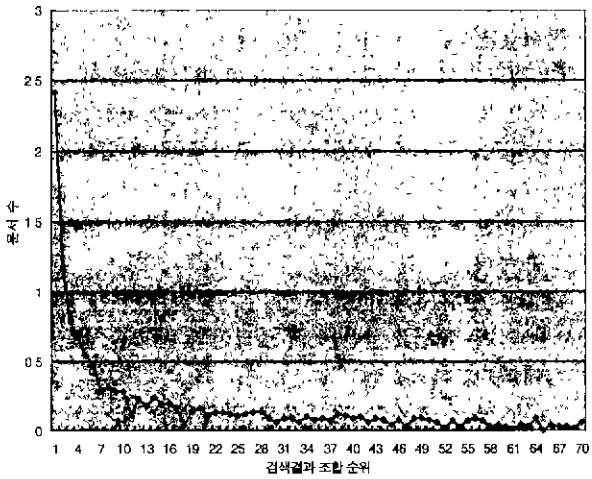

그림 4. 상위문서 수의 변화에 따라 평균적으로 추가되는 대답포함문서 개수

떤 질의에 대해서는 70개 이상의 결과를 조합했을 때, 정답을 포함하는 문서가 나타날 가능성이 있으므로 이러한 경우는 상위문서의 개수를 늘려서 적합성판정 작업을 추가로 하거나, 테스트컬렉션 이용자의 피드백을 반영해서 적합성 판정결과의 수정/보완이 필요하다.

3.4 문서 검색을 통한 질의응답시스템 구축시 정답포함 문서 검색 비율

질의응답시스템을 설계/구축할 때, 문서검색을 통해서 상위의 문서들에 대해서 대답을 추출하는 경우 문서검색 시스템이 대답이 포함되어 있는 문서를 어느 정도 포함하고 있는지 비교하였다.

현재 구축되고 있는 질의응답 시스템들이 대부분 문서검색을 통해 질의에 관련된 문서를 검색하고, 상위에 검색된 문서들을 분석해서 정답을 추출하고 있다. 이때, 문서검색에서 정답을 포함하고 있는 문서를 검색해주는 비율을 그림5에 나타내고 있다.

그림5의 검색 방법들은 검색결과 조합 방법을 적용하기 위해 이용된 정보검색시스템들로, 불리언 검색을 이용한 방법, 형태소단위의 색인/바이그램방식의 색인, 적합성 피드백 등을 이용한 방법, 벡터공간검색방법을 이용한 검색방법 등에 대해서 질의에 대한 검색결과이다. 문서검색시스템의 성능이 우수하다면, 질의응답시스템의 구축시에 문서검색을 통해서 높은 순위를 나타내는 문서에 대해서 대답을 추출할 수 있을 것이다.

그러나, 질의의 유형이나 특성에 따라서 문서검색의 상위문서를 더 포함시켜서 분석을 해야하는 필요성이 있다.

3.5 구축된 테스트컬렉션을 이용한 질의응답시스템의 평가 방법

질의응답 시스템이 질의에 대해 질의응답 시스템을 이용하여 대답이 포함된 부분을 추출한다. 검색결과는 <질의번호, 문서번호, 대답문자열>이다. 대답 문자열은 질의에 대한 대답만을 포함하도록 하거나, TREC-8, TREC-9에서와 같이 50바이트, 250바이트 등으로 가능하다.

질의응답 시스템의 결과에 대한 평가는 구축된 테스트컬렉션을 이용하여 검색된 부분의 문서번호가 적합한 것이고, 대답 문자열에 정답이 포함되어 있는 것일 때 정확한 결과로 판단한다.

질의응답 시스템의 성능 평가는 역순위 평균(Mean Reciprocal Rank)을 이용하여 성능을 평가할 수 있다. 이 방법은 질의응답 시스템에서의 결과로 나타난 것에서 정확한 대답이 포함되어 있는 것이 몇 번째 순위에 나타나는지를 계산해서 그 순위의 역수를 평균한 값이다.

즉, 정답이 1번째의 순위에 나타났으면 1/1, 2번째의 순위에 나타났으면 1/2, N번째의 순위에 나타났으면 1/N으로 점수를 부여하고, 전체 질의에 대한 값을 평균한다.

또, K번째 순위 내에 대답을 포함하고 있는지의 여부에 따라 대답을 포함하는 검색비율을 측정할 수 있다.

이외에도 다른 평가 방법에 대한 성능 평가가 가능할 것이다.

4. 결론

본 연구에서는 질의응답 시스템의 평가를 위한 테스트컬렉션을 구축하였다. 질의응답 시스템 성능 평가를 위한 테스트컬렉션은 207,067개의 문서, 90개의 질의, 각 질의에 대한 적합성 판정 집합으로 구성되어 있다. 질의 90개에 대해 적합성 판정을 한 문서집합은 22,898개 문서이다.

문서는 신문기사로 다양한 분야의 내용을 포함하고 있다. 질의는 사람, 장소, 조직, 시간, 수량 , 병명 등 다양한 대답을 요구하는 질의 유형을 포함하고 있다. 각 질의에 대해서는 같은 대답을 요구하는 질의의 변형집합을 생성하고 있어, 질의에 변형에 따른 질의응답시스템의 성능 변화를 평가할 수 있도록 하였다. 또한, 의미중의성, 음차표기/복원, 복합명사 문제 등과 같이 자연언어처리기

법을 적용할 수 있는 질의로 구성되었다. 적합성 판정 집합은 각 질의에 대해서 문서에 대답을 포함하는지의 여부에 따라 적합/부적합으로 판정하였고, 적합한 문서에 대해서는 대답을 표시하였다.

본 연구를 통해 구축된 질의응답시스템 성능 평가를 위한 테스트컬렉션은 질의응답시스템의 객관적인 신뢰성 평가를 위한 기반을 마련하였다. 문서집합과 질의집합의 크기가 영어권에서 구축하고 있는 질의응답용 테스트컬렉션 수준에 이르고 있다.

구축된 테스트컬렉션은 정보검색분야 연구자 및 개발자들에게 공개함으로써, 한국어 질의응답시스템의 연구에 적용되어 질의응답시스템 구축을 활성화시킬 수 있을 것으로 기대한다.

향후 테스트컬렉션의 확장을 통해서 사람이 궁금해 할 수 있는 다양한 질의들을 포함시킴으로써 현실적인 질의응답시스템의 평가가 가능하도록 해야 할 것이다. 또한, 현재의 테스트컬렉션의 분석을 통해서 비교적 안정적임을 알 수 있지만, 적합성 판정을 한 문서 수의 제한으로 인해 적합성 판정에 포함되지 않은 문서에서 대답이 나올 가능성이 있다. 이는 대답포함문서 후보집합의 생성에서 검색결과의 상위문서 수의 확장을 통해 적합성 평가를 함으로써 보완을 하여야 할 것이다. 또한, 테스트컬렉션을 배포하여 사용자들의 피드백을 반영함으로써 보다 신뢰할 수 있는 테스트컬렉션을 구축하도록 할 것이다.

감사의 글

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

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

[5] 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

[6] 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.

[7] 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

[8] 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

[9] 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