정보검색기반 질의응답 시스템 설계*

김민경*, 안혁주, 김학수
강원대학교 컴퓨터정보통신공학과
e-mail:kmink0817@kangwon.ac.kr, zingiskan12@kangwon.ac.kr,
nlpdrkim@kangwon.ac.kr

Design of a QA System based on Information Retrieval

MinKyoung Kim*, HyeokJu Ahn, Harksoo Kim
Dept of Computer and Communication Engineering,
Kangwon National University

요 약

본 논문에서는 질의유형을 통한 검색기반 질의응답 시스템을 구현하기 위한 설계방법을 제안한다. 이를 위해 위키피디아 문서의 링크 데이터를 이용하여 색인 대상문서와 데이터베이스를 구축하는 색인 모델과 2-포아송 모델을 이용하여 얻은 문서들을 색인 데이터베이스를 통해 필터링하여 정답 후보문장을 추출하는 검색모델, 키워드 패턴 매칭 기반 질의유형 분류 모델을 설계하였다.

1. 서론

기존의 정보검색은 사용자의 질문에 대한 결과로 단순히 키워드를 포함하는 대량의 문서들을 순위화하여 보여준다. 그러나 많은 사용자들은 정보 검색 시스템이 대량의 문서를 찾아주기 보다는 정답을 곧바로 찾아 제시해 주기를 바란다[1]. 이러한 요구를 만족시키기 위하여 질의 응답(question answering, QA)이라는 개념이 출현했다.

정보검색기반 질의 응답(이하 IRQA) 방법론은 크게 텍스트 조각 추출 방법(text-snippet extraction method)과 명사구 추출 방법(noun-phrase extraction method)으로 나눠진다[2]. 텍스트 조각 추출 방법은 질문에 대한 정답을 포함하고 있을 것 같은 텍스트의 단락이나 문장 또는 문장의 일부를 추출하는 것이다. 명사구 추출 방법은 한정된 클래스에 속한 사용자 질문에 대해서 구체적인 정답구(answer phrase)를 찾아주는 것이다. 본 논문에서는 정보검색기를 사용하여 문서를 찾고, 색인 데이터베이스를 통해 필터링하여 정답 후보를 찾는 IRQA 시스템을 설계하였다.

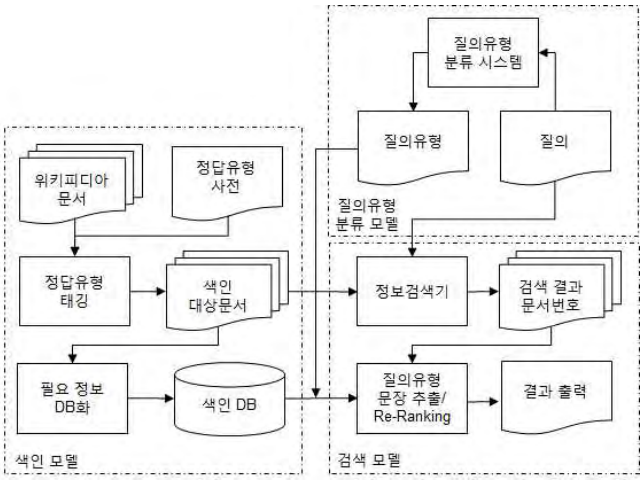
본 논문의 구성은 다음과 같다. 먼저, 2장에서 IRQA 시스템의 정답후보를 찾는 과정을 설명한 후 3장에서 결론을 맺는다.

* 이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2013R1A1A4A 01005074). 또한 본 연구는 LG전자 산학연구용역 과제의 지원을 받아 수행되었음.

2. IRQA 시스템 설계

본 논문에서 설계한 시스템 구조도는 (그림 1)과 같다. 시스템은 색인 모델과 질의유형 분류 모델, 검색 모델로 구성된다.

색인 모델은 색인 대상문서와 색인 데이터베이스를 구축한다. 색인 대상문서를 구축하기 위해 정답유형 사전과 위키피디아(wikipedia) 문서를 이용한다.



(그림 1) IRQA 시스템 구조도

정답유형 사전은 문서에 있는 개체들의 의미 범주를 질의유형에 따라 분류한 것으로 디비피디아 온톨로지의 InstancesType[3]을 이용한다. 위키피디아 문서는 반정형 데이터이며 위키문법 구조를 가지고 있다. 위키문법에서는

가리키고 싶은 문서의 제목 양쪽에 대괄호를 두 개씩 넣음으로써 링크를 표시한다. 본 논문에서는 이 링크 데이터를 개체 후보로 추출한 뒤, 정답유형 사전을 이용해 의미 범주를 할당하여 색인 대상문서를 구축한다. 예를 들어, (그림 2)와 같은 문서에서 두 개의 대괄호로 묶여있는 "1995년", "10월 28일", "미국", "기업인", "하버드 대학교|대학", "폴 앨런", "마이크로소프트"는 모두 개체 후보가 된다.

'''빌 게이츠'''({{lang|en|Bill Gates}}, {{본명|William Henry Gates III}}, [[1955년]] [[10월 28일]] ~)는 [[미국]]의 [[기업인]]이다. 어렸을 때부터 컴퓨터 프로그램을 만드는 것을 좋아했던 그는 [[하버드 대학교|대학]]을 다니다가 자퇴하고 [[폴 앨런]]과 함께 [[마이크로소프트]]를 공동창립했다.

(그림 2) 위키피디아 문서

개체 후보 중 정답유형 사전에 존재하는 "하버드 대학교|대학"과 "마이크로소프트"에 의미 범주가 할당되어 (그림 3)과 같은 색인 대상문서가 생성된다.

<doc>
<id>380</id>
<title>빌 게이츠</title>
<body>빌 게이츠는 미국의 기업인이다.
어렸을 때부터 컴퓨터 프로그램을 만드는 것을 좋아했던 그는 <EducationalInstitution>대학
</EducationalInstitution>을 다니다가 자퇴하고 폴 앨런과 함께 <Company>마이크로소프트
</Company>를 공동창립했다.
</body>
</doc>

(그림 3) 색인 대상문서

마지막으로 색인 대상문서에서 정답후보 문장들을 추출하기 위한 색인 데이터베이스를 구축한다. 색인 데이터베이스의 구조는 <표 1>과 같다.

<표 1> 색인 데이터베이스 구조

테이블	키	데이터
1	정답유형	문서번호
2	문서번호:정답유형	문서번호:문장번호
3	문서번호:문장번호	문장내용

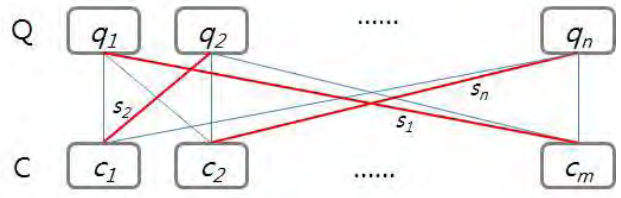
테이블1은 정답유형을 키(key)로 하며 정답유형이 출현하는 문서 리스트를 저장한다. 정보검색 결과문서에서 질의유형이 포함된 문서만을 걸러내는 역할을 한다. 테이블2는 문서번호와 정답유형을 키로 하며 정답유형이 출현한 문서에서 해당 정답유형이 포함된 문장의 위치를 저장한다. 정보검색 결과문서 중 테이블1로 걸러낸 문서에서 질의유형이 포함된 문장위치를 얻는 역할을 한다. 그리고 테이블3은 문서번호와 문장번호를 키로 하며 각 문서에서 특정 위치 문장의 내용을 저장하고 있다. 테이블1과 테이블2를 통해 질의유형이 포함된 문장의 위치를 얻은 후 해당 위치의 문장을 추출하는 역할을 한다.

질의유형 분류 모델은 사용자가 질의를 입력하면 질의에 포함된 키워드패턴을 매칭하고 휴리스틱을 이용하여 질의유형을 분류한 후 검색 모델에 전달한다. 본 논문에서는 질의유형의 의미 범주를 결정하기 위해 디비피디아 클래스 타입 데이터[4]를 참고한다. 각 질의유형을 결과로 출력할 수 있는 질의들을 보고 수작업으로 키워드를 구축한다. 입력 질의에서 해당 키워드가 발견되면 후보 질의유형으로 지정되고, 휴리스틱을 이용해 질의유형을 추출한다. 질의유형이 질의에 존재함에도 불구하고 키워드를 통해 찾지 못하거나 하나의 질의에 여러 개의 질의유형이 존재할 경우 이를 판단해주기 위해 휴리스틱을 이용한다.

검색 모델은 2-포아송 모델을 이용하여 질의를 검색하고, 검색 결과문서에서 색인 데이터베이스를 이용하여 정답 후보문장들을 추출한다. 정답 후보문장을 포함하는 문서번호를 추출하는 과정은 다음과 같다.

- 2-포아송 모델을 이용하여 질의를 검색하여 문서번호 리스트A를 얻는다.
- 질의유형 분류 시스템에서 전달받은 질의유형을 키로 색인 데이터베이스를 검색하여 해당 질의유형이 포함된 문서번호 리스트B를 얻는다.
- 검색결과 문서번호 리스트에서 질의유형이 포함된 문서번호 리스트만을 얻기 위해 두 리스트를 비교하여 공통된 문서번호만을 추출한다.

추출된 문서번호들과 질의유형을 색인 데이터베이스에 검색하여 정답 후보문장을 추출해낸다. 추출된 정답 후보들을 순위화하기 위해 (그림 4)와 같은 방법으로 질의와 정답 후보문장 사이의 유사도를 계산한다.



(그림 4) 유사도 계산 방법

(그림 4)에서 q_i는 사용자가 입력한 질의 Q의 단어이며, c_i는 후보문장 C의 단어이다. 유사도는 수식 (1)에 의해 계산된다.

$$S(Q, C) = \frac{\sum_{i=1}^{n} s_i}{n} \qquad (1)$$
$$s_i = \max_{j=1..m}(q_i, c_j)$$

식 (1)에서 s_i는 q_i와 c_j 사이에 가장 큰 점수를 나타내며, 유사도 S는 s_i의 평균으로 구해진다. 질의와 모든 후보문장과의 유사도를 계산하여 후보문장들을 순위화한다.

3. 결론 및 향후연구

본 논문에서는 정보검색기를 이용해 질의응답 시스템을 개발하기 위한 과정에 대해 설명하였고, 이를 위해 위키피디아 문서의 링크 데이터를 이용하여 색인 대상문서와 데이터베이스를 구축하는 색인 모델과 2-포아송 모델을 이용하여 얻은 문서들을 색인 데이터베이스를 통해 필터링하여 정답 후보문장을 추출하는 검색모델, 키워드 패턴 매칭 기반 질의유형 분류 모델을 설계하였다. 향후에 정답 후보의 문맥과 질의 사이의 의미 유사도를 LDA기반 Word Vector를 이용하여 정답을 특정 지을 계획이다.

참고문헌

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

[2] Vicedo J. L. and Ferrándex A., “Importance of Pronominal Anaphora resolution in Question An swering systems”, In *Proceeding of ACL 2000*, pp. 555-562, 2000

[3] DBPedia instance types, http://wiki.dbpedia.org/

[4] DBPedia Ontology Classes, *http://http://mappings.dbpedia.org/server/ontology/classes/*