개념 기반 질의-응답 시스템에서 개념 규칙을 이용한 해답 추출

강유환, 안영민, 서영훈
충북대학교 컴퓨터공학과
{eric, maniac}@nlp.chungbuk.ac.kr, yhseo@chungbuk.ac.kr

Answer Extraction using Concept Rules in Concept-based Question-Answering System

Kang yu-hwan, Ahn young-min, Seo young-hoon
Dept. of Computer Engineering, Chungbuk National University
{eric, maniac}@nlp.chungbuk.ac.kr, yhseo@chungbuk.ac.kr

요 약

본 논문에서는 개념 기반 질의-응답 시스템에서 개념 규칙을 이용하여 해답을 추출하는 방법에 대하여 기술한다. 개념 기반 질의-응답 시스템은 질의문의 각 유형별 개념 정보를 이용하여 질의문을 분석하고 해답을 추출하는 시스템이다. 질의문의 키워드들을 개념에 따라 분류하고, 질의 유형별로 공통적으로 나타나는 개념들을 이용하여 개념 프레임을 정의한다. 또한, 개념 정보와 해답이 들어 있는 문장과 문단에서 공통적으로 나타나는 구문 특성을 이용하여 해답 추출을 위한 규칙을 작성한다. 개념 규칙은 형태 정보와 구문 정보를 포함하며, 질의 유형별로 따로 작성한다. 작성된 규칙을 이용하여 문서로부터 해답이 들어 있는 문장과 문단을 추출한 후 질의문의 해답 유형에 해당하는 개체를 해답 후보로 제시한다. 실험 결과 개념 규칙을 이용한 해답 추출의 정확도가 매우 높게 나타났다.

1. 서론

정보검색 시스템은 사용자에게 질의와 관련이 있는 문서를 순위화하여 제공한다. 그러나 검색 대상 문서의 양이 많아짐에 따라 검색의 결과로 나타나는 문서의 양은 사용자에게 큰 부담이 되었다. 따라서 사용자의 질의에 대해 구체적인 해답을 제공해 줄 수 있는 질의-응답 시스템에 대한 요구가 증가하고 있다.

질의-응답(Question-Answering) 시스템은 대용량의 데이터 모음으로부터 사용자의 다양한 자연어 질문을 입력으로 받아 문서가 아닌 해답을 제공해주는 시스템이다[1]. 질의-응답 시스템에 대한 많은 연구들이 AAAI[2]와 TREC[3]을 중심으로 활발히 수행되고 있다. 일반적인 질의-응답 시스템은 질의문으로부터 질의 유형이나 해답 유형, 키워드 등을 추출한다. 또한, 기존의 정보 검색 방법을 이용하여 질의문과 유사한 문서를 검색하고, 검색된 문서 내에서 다시 해답을 포함할 가능성이 높은 단락을 추출한다. 마지막으로 해답 유형과 관련이 있는 개체를 해답으로 추출한다.

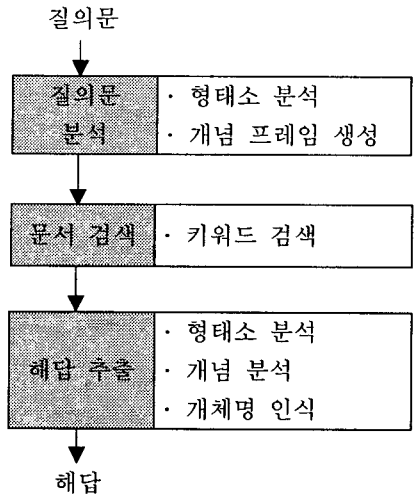
해답 추출을 위한 기술에는 질의문의 키워드를 이용하는 방법[4,5,6]과 구문 정보와 같은 자연어 처리 기술을 이용하는 방법[7,8,9] 등이 있다. 키워드 정보를 이용한 방법은 검색 속도 면에서 빠르다는 장점이 있지만 단지 키워드와 제한된 정보만을 이용하기 때문에 정확한 해답을 추출하기 어렵다. 구문 정보와 같은 자연어처리 기술을 이용하는 방법은 키워드 정보를 이용한 방법보다 보다 정확한 정답을 추출할 수 있다. 그러나 질의문의 구문 정보가 문서에서 다양한 형태로 나타나기 때문에 해답 추출을 위한 구문 패턴 작성에 어려움이 있고, 구문 정보만으로는 해답이 들어 있는 문장이나 문단의 특성을 표현하기 어렵다.

본 논문에서는 이러한 문제점을 해결하기 위해 개념 정보를 이용하여 해답을 추출할 수 있는 개념 기반의 질의-응답 시스템을 제안한다. 제안한 방법은 구문 정보와 개념 정보를 이용하여 해답 패턴을 작성함으로써 보다 정확한 해답 추출이 가능하도록 하고, 일반적인 질의-응답 시스템과 병합하여 사용할 수 있도록 한다.

본 논문의 구성은 다음과 같다. 2장에서는 질의 유형에 따른 개념 프레임의 정의와 개념 정보를 이용한 해답 추출 방법에 대해 살펴본다. 3장에서는 제안한 시스템에 대한 실험 및 결과에 대해 토의한다. 끝으로 4장에서는 결론 및 향후 연구에 대해 기술한다.

2. 개념 기반 질의-응답 시스템

그림 1은 본 시스템의 전체적 구성을 간략히 보여준다. 본 시스템은 크게 질의문 분석, 문서 검색, 해답 추출의 세 단계로 구성된다. 입력된 질의문에 대해 형태소 분석과 개념 분석을 수행한다. 질의문 분석의 결과로 질의문에 대한 개념 프레임을 얻을 수 있는데, 개념 프레임에는 질의 유형/해답 유형, 개념 정보 등이 포함된다. 개념 프레임의 개념 항목은 질의문의 키워드와 확장 키워드에 의해 채워진다. 문서 검색 단계에서는 개념 프레임에 들어 있는 키워드를 이용하여 관련 문서를 검색한다. 해답 추출 단계에서는 관련 문서들에 대해 형태소 분석, 개념 분석, 개체명 인식을 수행한 후, 해답 추출용 개념 규칙을 이용하여 해답이 들어 있는 문장과 문단을 추출하고, 질의문의 해답 유형에 대항하는 개체를 해답 후보로 제시한다.



[그림 1] 시스템 구성도

2.1 개념 프레임

질의문 분석 단계에서는 질의문의 질의 유형과 해답 유형을 결정하고, 질의문에 대한 개념 프레임을 구성한다. 현재 질의문의 질의 유형은 인명에 대한 질의 유형 30개, 지명에 대한 질의 유형 14개 등 총 79개의 질의 유형을 분류하였다. 개념 프레임은 각 질의 유형마다 정의되며, 질의 유형별로 공통적으로 나타나는 개념을 이용하여 정의한다. 질의문 분석과 해답 추출에 개념을 이용하면 보다 정밀하고 함축적인 규칙 작성이 가능하고, 구문 정보만으로는 추출하기 어려운 해답을 추출할 수 있다. 다음은 인명에 대한 질의 유형 중 저자와 가족에 대한 개념 프레임의 개념 목록을 보여준다.

- 저자 : 저서명, 국적, 장르, 저자부연, 필명, 저서부연, 시대, 시간
- 가족 : 기준인물, 기준인물부연, 관계, 관계부연, 역관계

표 1은 저자 질의문에 대한 개념 프레임의 구성

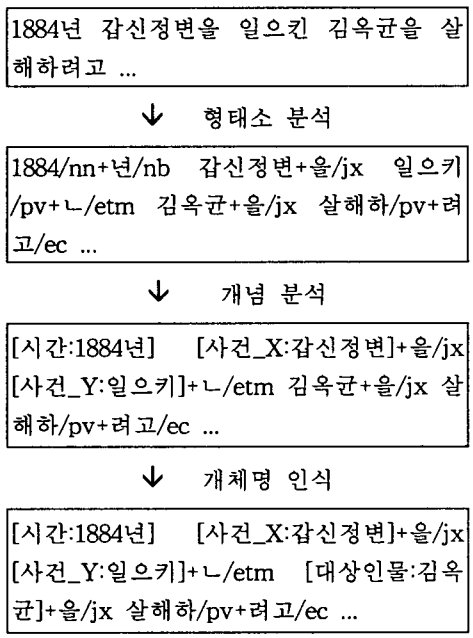
예를 보여준다.

[표 1] 저자 유형 질의문에 대한 개념 프레임 구성 예

질의문	1991년 소설 '개미'를 출간한 프랑스의 작가는?	
개념 프레임	질의 유형	저자
	해답 유형	인명
	장르	소설
	저서명	개미
	국적	프랑스
	시간	1991년, 91년

2.2 해답 추출

해답 추출 단계에서는 형태소 분석, 개념 분석, 개체명 인식, 해답 후보 순위화 등을 수행한다. 해답 추출을 위한 개념 규칙은 문서 검색 단계에서 추출된 관련 문서로부터 해답을 포함하고 있는 문장이나 문단을 추출하는데 이용된다.



[그림 2] 해답 추출을 위한 문서 분석 과정 예

해답 추출기는 관련 문서 집합에 대해 형태소 분석을 수행한 후, 문서에 들어 있는 단어들 중 개념 프레임의 개념 정보와 일치하는 단어에 대해 개념 태그를 부여한다. 그런 다음 질의문의 해답 유형에 부합하는 개체명에 대해 개체 태그를 부여한다. 그림 2는 해답 추출을 위한 문서 분석 과정을 보여준다.

마지막으로 해답 추출기는 해답 추출을 위한 개념 규칙을 이용하여 해답이 들어 있는 문장과 문단을 추출한다. 다음은 저자와 가족, 정치가 유형에 대한 해답 추출을 위한 개념 규칙의 예를 보여준다.

· 해답 추출을 위한 개념 규칙의 예

저자 :

[대상인물]+는/jc [*] [저서명]+를/jc <저자_V>

[대상인물]+가/jc [*] <저자_V>+ㄴ/etm [저서명]

가족 :

[기준인물]+의/jm [관계]+ㄴ/etm [대상인물]

[대상인물] [기준인물]+의/jm [관계]+!etm

정치가 :

[시간] [사건_X]+를/jc [사건_Y]+ㄴ/etm [대상인물]

위 개념 규칙에서 <저자_V>는 저자와 관련된 용언 즉, '쓰다', '저술하다', '출판하다' 등의 용언이 올 수 있음을 의미한다. '*'는 규칙을 매칭시킬 때 여러 단어를 건너뛸 수 있음을 의미한다.

'1884년 갑신정변을 일으킨 조선후기의 정치가는?' 이라는 질의문에 대해 관련 문서 중에 '1884년 갑신정변을 일으킨 김옥균을 살해하려고 ...' 라는 문장이 들어 있다면, 정치가에 대한 첫 번째 규칙에 의해 해당 문장이 해답 문장으로 추출되고, 해답 유형에 해당하는 '김옥균'을 최종 해답으로 추출하게 된다. 해답 후보가 여러 개일 경우에는 해답 후보 순위화에 의해 가장 높은 순위의 해답을 제시한다. 만일 문서 내에 개념 규칙과 일치하는 문장이 없을 경우에는 일반적인 검색 방법을 이용하여 해답 후보를 추출한다.

3. 실험 및 분석

본 연구에서는 인명에 대한 질의 유형 30개에 대해 개념 프레임과 개념 규칙을 정의하고, 이중 '저자', '정치가', '가족', '수상자', '연예인' 의 5가지 질의 유형에 대한 해답 추출 실험을 수행하였다. 실험에 사용된 문서 집합은 일반 정보 검색 시스템을 이용하여 수집한 문서 집합을 사용하였으며, 질의문 당 15개의 문서를 실험 문서로 구축하였다. 표 2는 질의 유형별 질의문 수와 상위 5위 안에 들어 있는 정확한 해답의 수를 보여준다.

[표 2] 상위 5위 안에 들어 있는 정확한 해답의 수

질의 유형	질의문 수	해답 수
저자	5	4.2
정치가	4	3
가족	3	3.7
수상자	4	2.8
연예인	4	2.3
총	20	3.2

실험에 사용된 질의문은 각 질의 유형을 대표할 수 있는 형태의 질의문을 선정하였으며, 20개의 질의문에 대해 상위 5위 안에 들어 있는 정확한 해답의 수는 평균 3.2개를 보였다. '저자' 유형의 경우에는 문서 내에 질의문에 대한 해답을 포함하는 문장이 많이 들어 있었고, 또한 개념 규칙과 일치하는 문장이 여러 개가 추출되었기 때문에 많은 수의 해답이 정확하게 추출되었다. 반면에 '연예인' 의 경우에는 문서 내에 들어 있는 해답의 수 자체가 적었고, 해답 문장의 구문과 어휘가 다양한 형태로 나타나는 특성을 보였다. 개념 규칙을 이용하여 추출된 해답들은 대부분 정확한 해답들이었다.

4. 결론 및 향후 연구

본 논문에서는 개념 기반 질의-응답 시스템에서 개념 규칙을 이용한 해답 추출 방법을 제안하였다. 의미적으로 유사한 질의문을 하나의 질의 유형으로 분류하고, 각 질의 유형마다 질의문에서 공통적으로 나타나는 개념을 이용하여 개념 프레임을 정의하였다. 또한 개념 정보와 해답 문장의 구문 특성을 이용하여 해답 추출을 위한 개념 규칙을 정의하였다. 실험 결과 개념 규칙을 해답 추출에 적용함으로써 해답을 보다 정확하게 추출할 수 있었으며, 구문 정보만으로는 추출할 수 없는 해답 또한 추출할 수 있었다. 제안한 방법은 일반 질의-응답 시스템과 병합하여 사용함으로써 해답 추출의 정확도를 높일 수 있다.

향후 연구로는 질의 유형에 대한 개념 규칙을 수정, 보완하고, 아직 정의되지 않은 다른 질의 유형에 대해서도 실험을 계속할 예정이다.

참 고 문 헌

[1] Ellen M. Voorhees, The TREC question answering track, *Natural Language Engineering*, 7(4), pp. 361-378, 2001

[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] S. Abney, M. Colins, A. Singhal, Answer Extraction, *In 6th Applied Natural Language Processing Conference*, 2000

[5] G.G Lee, J. Seo, S. Lee, H. Jung, B. Cho, C. Lee, B. Kwak, J. Cha, D. Kim, J. Ann, H. Kim, K. Kim, SiteQ:Engineering High Performance QA System Using Lexico-Semantic Pattern Matching and Shallow NLP, *In 10th Text REtrieval Conference*, pp. 437-446, 2001

[6] A. Ittycheriah, M. Franz, W. Zhu, A. Ratnaparkhi, IBM's Statistical Question Answering System, *In 9th Text REtrieval Conference*, pp. 229-334, 2000

[7] S. Buchholz, W. Daelemans, Compelx Answers: a case study using a WWW question answering system, *Natural Language Engineering*, 7(4), pp. 301-323, 2001

[8] S. Harabagiu, M. Pasca, S. Maiorano, Experiments with open-domain with open-domain textual question

answering, *In COLLING-2000*, pp. 292-298, 2000

[9] Valdo Keselj, Question Answering using Unification-based Grammar, *Advanced in Artifical Intelligence, AI 2001, volume LNAI 2056 of Lecture Notes in Computer Science*, Ottawa, Canada, Springer, 2001