개념 기반 질의-응답 시스템에서의 정답 추출

Answer Extraction of Concept based Question-Answering System

안영민, 오수현, 강유환, 서영훈
충북대학교 컴퓨터공학과

Ahn Young-Min, Oh Su-Hyun, Kang Yu-Hwan, Seo Young-Hoon
Chungbuk National University

요약

본 논문에서는 개념 기반 질의-응답 시스템에서의 정답 추출 방법에 대하여 기술한다. 개념 기반 질의-응답 시스템은 개념 정보를 이용하여 해답을 추출하는 시스템을 말하며, 질의분석을 통해 분류되고 추출된 개념 그에 따른 정답 추출 규칙을 이용하여 정답을 추출하는 방법과 시스템에 대하여 연구하였다. 질의에 대한 정답이 들어 있는 문서들을 분석하여 정답 추출 규칙을 작성한다. 규칙은 개념과 구문정보를 포함하고 있으며 작성된 규칙을 통하여 문서로부터 정답후보를 생성하고 정답을 선택한다.

Abstract

In this paper, we describe a method of answer extraction on a concept-based question-answering system. The concept-based question answering system is a system which extract answer using concept information. we have researched the method of answer extraction using concepts which analyzed and extracted through question analysing with answer extracting rules. We analyzed documents including answers and then composed answer extracting rules. Rules consist of concept and syntactic information, we generated candidates of answer through the rules and then chose answer.

I. 서 론

1. 질의-응답

정보검색 시스템은 사용자의 질의에 대하여 관련이 있는 문서들을 자체 scoring 알고리즘을 통하여 순위가 높은 순서대로 문서를 제공한다. 사용자는 제공된 문서 내에서 자신이 원하는 정보를 찾기 위하여 문서의 내용을 읽거나 관련 내용이 없으면 다른 방식으로 질의를 하거나 하는 별도의 과정을 수행하여야 한다.

검색 대상 문서의 양이 많아짐에 따라서, 검색의 결과로 나타나는 문서의 양은 사용자에 큰 부담을 줄 수가 있다. 이에 따라 사용자의 질의에 대해 구체적인 해답을 제공해 줄 수 있는 질의-응답 시스템에 대한 요구가 증가하고 있다.

질의-응답 시스템(Question-Answering System)은 대용량의 데이터 모음으로부터 사용자의 다양한 자연어 질의를 입력으로 받아 문서가 아닌 정답을 제공해 주는 시스템이다[1]. 질의를 분석하거나 정답을 추출하는 방법에 대하여 많은 연구들이 AAAI[2]와 TREC[3] 등에서 진행되어 왔다.

정답 추출을 위한 기존 연구에는 키워드를 이용하는 방법[4,5,6]과 구문 정보 등 자연언어처리 기술을 이용하는 방법[7,8,9] 등이 있다.

질의-응답 시스템은 크게 질의를 분석하는 부분과 정답을 추출하는 부분으로 나눌 수 있다. 본 연구실에서 진행하고 있는 개념 기반 질의-응답 시스템의 연구 중에서 정답 추출하는 부분을 본 논문의 대상으로 한다. 따라서 질의 분석의 결과로 나오는 정답 유형과 개념 그리고 개념 프레임(개념들의 모음)은 존재한다고 가정하고 정답 유형과 개념 프레임을 이용하여 문서에서 정답을 추출하는 부분만을 기술하기로 한다.

"강희제의 아버지는?"이라는 질의에 대하여 질의 분석을 거치면 "인물"이라는 대유형과 "가족"이라는 세부 유형이 결정된다. 그리고 "강희제"라는 기준이 되는 인물과 "아버지"라는 관계가 개념으로 추출된다.

'기준인물:강희제', '관계:아버지', '역관계:아들'이라는 개념프레임이 생성된다. 또한 가족 관계를 나타내는 문서들을 분석하여 얻어진 정답 추출 규칙을 개념 프레임과 함께 적용하여 정답을 추출한다. 이렇게 개념을 적용하면 "강희제", "아버지"라는 키워드만으로는 찾을 수 없는 "순치제의 아들인 강희제는 ~ "이라는 문장에서도 정답을 추출해 낼 수 있다.

II. 정답 추출

1. 규칙

정답 추출에 사용될 규칙을 추출하기 위해서는 각 대유형과 세부 유형에 따른 질의에 대하여 정답을 포함하고 있는 문서들을 수집한다. 수집된 문서들에서 정답을 포함하고 있는 문장이나 문장들을 분석하여 개념 프레임과 연관하여 규칙을 작성한다. 규칙들은 개념, 형태소 그리고 구문정보를 이용하여 구축되어 있다. 또한 규칙들은 문서들에서 많이 발견된 빈도와 경험적인 내용을 고려하여 우선순위를 부여 받는다.

아래의 표는 대유형 인물에서 세부유형 가족에 대한 정답 추출 규칙의 예이다.

[표 1] 대유형 인물의 세부유형 가족에 대한 정답 추출 규칙의 예

순서	규칙
1	[기준인물]+[*]+의|와|과/j_ [관계]+?etm [대상인물]
2	[대상인물] [1L] [기준인물]+[*]+의/j_ [관계]+?!etm
3	[기준인물]+은|는|이|가|의/j_ [관계]+?etm [대상인물]
4	[기준인물]+[*]+의/j_ [관계]+?etm [*] [대상인물]
5	[대상인물]+의/j_ [역관계]+[*] [기준인물]+?!etm
6	[기준인물] [10L-] [대상인물]+의/j_ [*] [역관계]

정답 추출 규칙에 표현된 '[기준인물]', '[대상인물]', '[관계]', '[역관계]' 등은 세부유형 가족에 대해 정의된 개념들이고, '[1L]'은 '[기준인물]'을 기준으로 그 이전 문장까지를 정답추출 대상으로 포함시키라는 의미이다. 6번 규칙에서 '[10L-]'은 기준인물을 대상으로 10문장 이내에서 규칙이 매칭되는 가를 판단하라는 의미이며 '[10L-]'의 앞부분인 '[기준인물]'은 표제어 형태여야 한다는 암시적인 의미도 가지고 있다. 좌우에 아무것도 연결되지 않은 [*]는 문장 내에서 패턴이 매칭될 때까지 몇 어절을 건너뛰어도 관계없음을 나타낸다. 좌우에 '+'기호로 연결된 '[*]'는 명사구 또는 3어절 내에서 매칭이 되어야 한다는 것을 나타낸다. 규칙에 '[*]'과 '[1L]' 및 '[10L]' 등을 사용하여 기술함으로써 아주 엄격하게 규칙을 적용하는 방법이 아닌 소프트 패턴 매칭의 의미를 일부 표현하였다.

2. 정답 추출

정답 추출 과정은 우선 정답을 추출할 문서집합에 대하여 품사 태깅을 수행하고, 개념 프레임 정보를 이용하여 개념 태깅을 수행한다. 개념 태깅을 수행한 후에는 개체명 인식기를 이용하여 해답 유형에 해당하는 개체를 태깅한다. 아래 표는 태깅 전후의 문장의 예를 보여준다.

[표 2] 개념 태깅의 예

원문	8세의 어린 강희제는 아버지 순치제의 유언에 따라...
태깅 후	8세의 어린 [기준인물:강희제]+는/jx [관계:아버지] 순치제/nh+의/jm 유언에 따라...

문서에 대한 태깅 작업이 끝난 후에는 정답 추출 개념 규칙을 이용하여 정답이 들어 있는 문장과 문단을 추출한다. 정답으로 추출되는 문장과 문단은 개념 규칙과 일치하면서 해답 유형에 해당하는 개체가 포함된 문장과 문단이어야 한다.

정답 후보는 문장/문단 추출 단계에서 생성된 정답 문장/문단 집합으로부터 추출한다. 정답 문장/문단 집합에는 정답 유형에 해당하는 개체가 이미 태깅되어 있으므로 정답 후보를 손쉽게 꺼낼 수 있다. 정답 추출 개념 규칙에는 규칙마다 우선순위가 정해져 있기 때문에 정답 후보는 적용된 규칙의 우선순위에 따라 차별화된 가중치를 부여 받는다.

적용된 규칙의 우선순위에 따라 정답 후보를 1차적으로 순위화하고, 정답 후보의 출현 빈도수를 이용하여 순위를 재조정 한다. 최종적으로 상위에 포함되어 있는 정답 후보를 해답으로 제시한다.

III. 실험 및 분석

실험에서는 인명 관련 질의 20개에 대해 상위 5위 안에 들어 있는 정확한 정답의 수를 평가하였다. 인명 관련 질의문의 세부유형 수는 34개이며, 10개의 세부 유형에 대해 개념 프레임과 개념 규칙을 정의하였다. 실험에서는 이 중 '저자', '정치가', '가족', '수상자', '연예인'의 5개 세부유형에 대해 실험하였다.

본 연구의 목적은 정답을 추출하기 위하여 질의에 관련된 문서를 수집하고 수집된 문서를 분석하여 정답 추출 개념 규칙을 작성한 다음 작성된 규칙을 적용하여 정답이 제대로 추출되는 지를 판단하는 것이다. 5개의 세부유형에 대해 작성된 규칙을 가지고 그 규칙을 작성하는데 참고로 사용이 되었던 문서들을 적용하여 정답을 추출해보니 99%의 정답이 추출되었다. 당연한 결과로 규칙을 생성하기 위하여 사용된 문서를 대상으로 정답을 추출하였으니 100%가 나오지 않은 것이 더 이상한 결과다. 정답이 제대로 추출되지 않은 경우는 개체명 인식기를 통해 나온 결과가 약간의 오류를 포함하는 경우였다.

IV. 결론 및 향후연구

실험 결과는 개념 프레임이 잘 정의되고 정답을 포함하는 문장을 잘 분석하여 정답 추출 규칙을 구성한다면 그것들을 가지고 올바른 정답을 추출해 낼 수 있다는 것을 보여주는 것이다.

향후로 질의-응답 시스템을 구성하기 위해서는 질의 분석 부분과 정답을 추출하는 부분의 통합이 이루어져야 하겠고 현재 인물의 세부유형 10개에 대해서만 정의되어 있는 개념 프레임과 규칙들을 더 확장하고 다른 범주의 질의에 대해서도 연구가 진행되어야 할 것이다. 일반 정보검색에서 질의를 분석한 후 검색을 통하여 얻어온 문서들에서 정답을 바로 찾아주는 연구나 또는 찾아진 문서들에서 정답을 포함할 가능성이 높은 문서들을 상위에 랭크시켜주는 연구도 개념 기반의 질의분석 및 정답추출을 이용하여 진행될 수 있을 것으로 생각된다.

▌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.