위키백과를 이용한 질의응답 시스템의 구현

박영민○, 서정연*
서강대학교 컴퓨터공학과, 서강대학교 컴퓨터공학과/바이오융합기술협동과정*
pymnlp@gmail.com, seojy@sogang.ac.kr

Implementation of Question-Answering System using Wikipedia

Young-Min Park○, Jung-Yun Seo*
Department of Computer Science and Engineering, Sogang University○
Department of Computer Science and Engineering, and Interdisciplinary Program of Integrated Biotechnology Sogang University*

요 약

본 논문에서는 위키백과를 이용한 지식DB구축의 예로서 연예인 관련 정보들을 자동으로 추출한다. 우리는 위키백과의 연예인 문서로부터 생년월일, 학력, 본명 등 총 9가지 정보들을 추출하고 이를 지식DB로 구축한다. 또한 추출된 지식DB를 이용하여 질의응답 시스템을 구현하여 유용함을 입증하였다. 질의응답 시스템은 어휘의미패턴 방법으로 질의를 분석하고, 템플릿 기반의 문장생성 방법으로 정답을 자연어문장으로 생성한다. 성능 평가결과 총 6471명의 연예인 정보들을 추출하였고, 95%에 해당하는 질의분석 성능을 제공하였다.

주제어: 위키피디아, 질의응답 시스템, QA, 지식DB

1. 서론

질의응답 시스템(Question-Answering System)은 사용자가 입력한 질의를 분석하고 질의에 대한 정답에 해당하는 응답을 제공하는 시스템이다.[1] 기존의 질의응답 시스템은 정답에 대한 내용이 있는 문서집합에서 해당 문서를 찾아내는 방법이 많이 사용되었다. 하지만 이러한 시스템은 문서집합을 구축하는데 많은 비용을 소비하게 된다.

위키백과(www.wikipedia.org)는 위키미디어 재단이 운영하는 백과사전 프로젝트로 세계최대 규모의 인터넷 백과사전이다. 위키백과는 누구나 작성, 편집에 참여할 수 있고, 구축된 정보를 자유롭게 사용할 수 있기 때문에 거대한 지식 공급원으로 사용할 수 있다.[2][3]

본 논문에서는 위키백과를 이용한 질의응답 시스템을 제안한다. 질의응답 시스템은 지식DB를 구축하기 위해 위키백과로 필요한 지식들을 자동으로 추출한다. 사용자 질의가 입력되면 어휘의미패턴 기반의 질의분석을 통해 질의를 분석하고 그에 대한 답변을 지식DB로부터 추출한다. 마지막으로 템플릿(template)기반의 문장생성 방법으로 자연어 응답을 출력한다.

2. 위키백과를 이용한 질의응답 시스템

본 논문에서 제안하는 질의응답 시스템은 (그림 1)과 같이 위키백과로부터 정보를 추출하는 색인 시스템과 질의를 분석, 추출 및 문장을 생성하는 QA시스템으로 구성한다. 질의영역은 연예인 관련 정보로 제한하여 진행한다. 질의의 분류체계는 (표 1)과 같다.

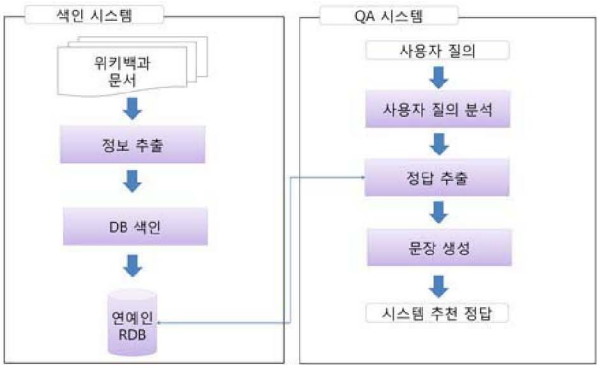

그림 1. 질의응답 시스템

분류	예문
생년월일	“한가인이 언제 태어났니?”
고향	“최진실이 태어난 곳이 어디지?”
데뷔	“아이유가 무슨 노래로 데뷔했니?”
학력	“이경규가 동국대 나온거 맞니?”
가족관계	“유재석 부모님 이름이 뭐지?”
소속사	“박정현 소속사가 어디냐?”
출연작품	“김수현이 출연했던 작품들이 뭐지?”
본명	“보아는 본명이 보아야?”
기본정보	“문소리가 누구냐?”

표 1. 질의 분류체계

2.1 정보 추출 및 색인

위키백과 문서 “문소리 (배우)”의 구성은 (그림 2)와 같다. 우리는 연예인 문서들만 수집하기 위해 분류 정보

에 연예인에 해당하는 단어가 존재할 경우 연예인으로 판단하고 추출한다. 또한 드라마 표제어 사전을 구축하기 위해 분류 정보에 "드라마" 단어가 있는 표제어들을 추출한다.

연예인의 생년월일, 고향, 데뷔, 학력, 가족관계, 소속사, 본명 정보는 DB정보(Info Box)로부터 각 속성에 해당하는 정보들을 추출한다. DB정보의 속성에 해당하는 태그들이 동일하지 않기 때문에 수동으로 사전을 구축하였다. 학력정보의 경우 본문에 학력 단락을 따로 작성한 경우가 많기 때문에 "= 학력 =" 의 문자열이 있는 경우 학력 단락에서 추출한다. 출연작품 정보는 문서 내에서 드라마 사전에 있는 표제어가 아웃링크로 표시된 경우 출연작으로 추출한다. 기본정보는 위키백과 문서의 요약 텍스트 부분을 추출한다. 추출된 연예인 정보는 RDB 형태로 저장하여 "문소리(연예인) + 생년월일(속성) -> 값(1974년 7월 2일)" 의 형태로 검색할 수 있도록 색인한다.

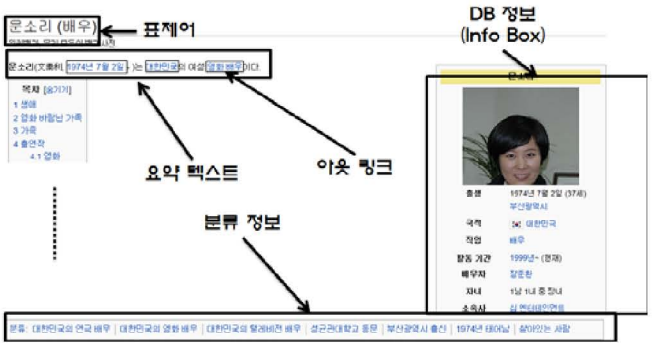

그림 2. 위키백과 문서의 구성

2.2 어휘의미패턴을 이용한 질의 분석

우리는 사용자 질의 분석을 하기위해 어휘의미패턴(LSP : Lexico Semantic Pattern)을 사용한다. 어휘의미패턴은 질의응답 시스템에서 사용자 질의를 미리 정의된 질의패턴과 일치(matching)하는 방법이다.[4][5]

우리는 학습데이터로부터 필요한 어휘의미패턴과 의미사전을 수동으로 구축하였다. 본 실험에서 사용자 질의를 어휘의미패턴으로 변환한 예는 다음과 같다.

질의 : 한가인은 생일이 언제지?
LSP : @entertainer.*@birth_date.*@when.*
질의 : 김수현이 나온 작품들 알려줘.
LSP : @entertainer.*나오.*@work.*알리.*

입력된 질의 문장은 구축된 어휘의미패턴들 중 일치하는 패턴으로 분류하여 질의 분류를 하고, 질의 문장에서 가장 먼저 나온 이름을 대상 연예인으로 분석한다.

2.3 문장 생성

색인 시스템에서 구축된 RDB검색을 통해 대상 연예인, 속성, 값에 대한 결과를 출력한다. 본 논문에서는 질의에 대한 응답을 자연어로 출력하기 위해 템플릿 기반의 문장생성 방법을 사용한다[6]. 문장생성의 예는 다음과 같다.

대상 -"김윤석" 분류 -"학력" 값 -"혜광고등학교, 동의대학교"
템플릿 - "@(target)@(이/가) 나온 학교는 @(value) 입니다."
생성 결과 - "김윤석이 나온 학교는 혜광고등학교, 동의대학교 입니다."

대상 -"보아" 분류 -"본명" 값-"권보아"
템플릿 - "@(target)@의 진짜 이름은 @(value) 입니다."
생성 결과 - "보아의 진짜 이름은 권보아 입니다."

3. 성능 평가

우리는 위키백과 문서로부터 연예인 문서를 추출하기 위해 분류정보에 "배우", "가수", "연극인", "음악가", "희극인", "모델", "방송인", "연예인" 의 단어가 있는 경우의 문서들을 추출하였다. 추출 결과 총 6471명에 해당하는 연예인 문서들을 수집하였다. 위키백과 문서는 계속해서 작성, 편집이 되고 있기 때문에 시간이 지남에 따라 더 많은 정보들을 추출할 수 있다.

우리는 질의분석의 성능을 평가하기 위해 각 질의 분류별로 75개 총 675개의 문장을 수집하였다. 수집된 데이터 중 질의 분류별로 50개의 문장으로 어휘의미패턴을 구축하였고 25개의 문장을 성능평가에 사용하였다. 질의분석 성능은 (표 2)와 같다.

분류	정확률
생년월일	0.92
고향	0.92
데뷔	0.96
학력	0.88
가족관계	0.93
소속사	1.0
출연작품	1.0
본명	0.93
기본정보	1.0
평균	0.95

표 2. 질의분석 성능

질의분석 성능은 범위를 연예인 정보 9가지로 제한하였고 질의문장패턴이 대부분 비슷하였기 때문에 전체적으로 높은 성능을 보여주었다. 학력질의의 경우 가장 낮은 성능을 보여주었는데, 고향 질의로 오분석 하는 경우가 많았다. 예를 들어 "박은빈이 서강대 출신이지?" 라는 질의는 "서강대" 가 의미사전에 구축되지 않았기 때

문에 고향으로 오분석하게 된다.

4. 결론

위키백과는 많은 사람들에 의해 계속해서 작성, 편집될 뿐 아니라 저작권에서 자유롭게 사용할 수 있기 때문에 적은 비용으로 최신 지식들을 추출할 수 있다. 본 논문에서는 위키백과를 이용하여 자동으로 지식DB를 구축하는 방법을 제안하였다. 또한 연예인 정보를 대상으로 하는 질의응답 시스템을 구현하여 스마트 TV환경에서 위키백과가 유용하게 사용될 수 있음을 보였다. 향후 우리는 위키백과로부터 더욱 다양한 지식을 추출하는 방법을 연구하고 이 결과를 여러 분야에 응용할 계획이다.

"본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음"(NIPA-2012-(H0301-12-3004))

참고문헌

[1] L. Hirschman and R. Gaizauskas, "Natural Language Question Answering: the View from here," Natural Language Engineering, vol. 7, no. 4, pp. 275-300, 2002.

[2] http://en.wikipedia.org/wiki/Wikipedia:Database_download, (Wikipedia:Database download)

[3] L. Denoyer and P. Gallinari. "The Wikipedia XML Corpus," SIGIR Forum, 2006.

[4] Sanda Harabagiu, Dan Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan Bunescu, Roxana Girju, Vasile Rus and Paul Morarescu "The Role of Lexico-Semantic Feedback in Open-Domain Textual Question-Answering," in Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001), Toulouse France, pp.274-281, 2001.

[5] Gary Geunbae Lee, Jungyun Seo, SeungwooLee, Hanmin Jung, Bong-Hyun Cho, Changki Lee, ByungKwan Kwak, Jeongwon Cha, Dongseok Kim, JooHui An, Harksoo Kim, Kyungsun Kim. "SiteQ: Engineering High Performance QA System Using Lexico-Semantic Pattern Matching and Shallow NLP ," In Proceedings of the 10th Text Retrieval Conference(TREC-10), 2001.

[6] E. Reiter and R. Dale, "Building Applied Natural Language Generation Systems," Natural Language Engineering, vol.3, no.1, pp.57-87, 1995.