자연어처리를 이용한 교육행정의 질의응답시스템

이미나° 윤성대
부경대학교 교육대학원 전산교육학과° 부경대학교 전자계산학과
minari@mail1.pknu.ac.kr° sdyoun@pknu.ac.kr

Question and Answering System of Educational Administration Using Natural Language Processing

Mi-Na Lee° Sung-Dae Youn
Dept. of Computer Education, Pukyung National University°
Dept. of Computer Science, Pukyung National University

요 약

정보통신 기술의 발달로 일반기업체 뿐만 아니라 공공기관 등 행정업무가 필요한 곳에서는 대부분 웹사이트를 통해 사용자에게 원하는 정보를 제공해 주고 있다. 그러므로 대부분의 상업용 사이트들은 사용자에게 보다 편리하게 정보를 제공해 주기 위하여 다양한 정보검색의 접근 방법을 사용하고 있다. 그러나 현재 교육행정의 업무처리 분야에서 정보제공은 웹사이트의 단순 키워드검색을 통하여 사용자가 직접 정보를 찾는 방식으로 이루어지고 있다. 본 논문에서는 자연어처리를 사용한 교육행정의 질의응답시스템을 제안한다. 사용자 질의의 의도를 분석하여 기본사전과 매칭한 후에 추출된 사용자 질의정보를 통해 자동으로 정답 데이터뷰를 생성하여 사용자 의도에 알맞는 정확한 정답을 제공하도록 하였다. 또한 동적인 FAQ 관리기능인 히스토리를 통해서 한번 질의한 정답을 신속히 제공하도록 하였다. 제안한 시스템의 효용성을 검증하기 위해 교육행정정보를 제공하는 간단한 질의응답시스템을 구현하여 적용해본 결과 일반 키워드 검색에서보다 정확하게 정답을 제공해 주는 것을 확인할 수 있었다.

1. 서론

정보통신 기술과 인터넷이 발달하면서 사용자는 인터넷을 통해서 더 많은 정보를 찾고 가상공간에서 서로의 의견을 교환하여 기존의 정보를 확대-재생산하고 있다. 하지만 이러한 인터넷의 발달은 정보의 폭발적인 증가를 가져왔고 사용자는 거대한 정보의 홍수 속에서 자신에게 필요한 정보만을 찾기 위해서 많은 시간과 노력을 투자해야 하는 문제를 가지게 되었다. 이를 해결하기 위해 먼저 전자상거래 시스템에서 사용자와 사업자 모두의 요구를 충족시킬 수 있는 효과적인 정보 검색 도구의 필요성이 제기 되었다[1,2]. 이에 따라 여러 가지 검색 도구들이 개발되었고 초기 복잡한 메뉴 기반의 네비게이션이나 지루한 키워드 기반검색에서 벗어나 사용자와 친숙할 수 있는 자연어 처리를 이용한 검색도구까지 발달하게 되었다[3].

현재 자연어를 사용한 검색도구는 사용자의 편의가 우선적으로 적용되는 상업용 사이트와 학술검색용 사이트에서 주로 사용되고 있다. 하지만 아직 일반 행정기관 및 행정업무를 담당하는 곳에서는 이러한 시스템이 미비한 실정이다. 따라서 본 논문에서는 교육행정업무에서 보다 쉽고 정확하게 사용자질의에 응답할 수 있는 효율적인 정보제공자로서의 질의응답 시스템을 위해 사용자 질의의도에 따른 답변매칭방법을 제안하고 이를 구현하여 검증하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로서 자연어처리를 적용한 질의응답시스템에 대하여 간략하게 소개를 하고 현재 행정처리 시스템의 구성 방식과 문제점을 지적해 본다. 3장에서는 본 논문에서 제안하는 시스템의 구성과 핵심내용에 대하여 서술한다. 4장에서는 제안한 시스템에 대한 실험 및 평가를 하고, 끝으로 5장에서는 결론 및 향후 연구 과제를 제시한다.

2. 관련 연구

2.1 자연어처리를 사용한 질의응답시스템

질의응답시스템의 목적은 다수의 문서를 검색해주는 정보검색과는 달리 사용자가 원하는 정확한 정답만을 제공해주는 것이다. 인터넷을 통해서 정보를 얻는 일반 사용자에게 보다 편리한 사용자 인터페이스 환경을 제공하기 위하여 질의응답시스템에서는 자연어처리를 이용한 연구가 활발히 진행되고 있다. 자연어 처리기법은 자연어 질의에 대하여 띄어쓰기나 비속어 등의 전처리과정을 거친 후 형태소 단위로 분석하여 사전에 있는 품사 정보 등을 출력해 주는 형태소 분석, 형태소 분석 결과와 문법 규칙등을 바탕으로 문장의 구조를 분석하는 구문 분석[5,6], 이 결과를 바탕으로 사용자의 질의 의도를 파악하는 의미 분석의 과정을 거쳐 최종적으로 사용자에게 응답을 제공한다. 주요 질의응답시스템으로는 KorQuA[7], LASSO[4], Any-Question[8] 등이 있다.

2.2 행정처리 시스템

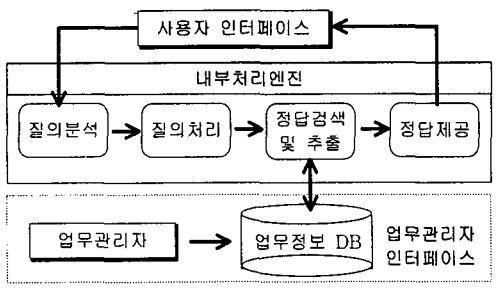
행정처리 시스템은 크게 행정정보를 제공하는 업무담당자와 정보를 제공받는 사용자의 입장으로 나눌 수 있다. 현재의 행정처리 시스템의 연구는 이 두 가지 입장 중에서 주로 업무담당자에게 초점이 맞추어져 있다. 현재 활성화 되어 있는 행정기관 사이트 중 중앙행정기관 20곳, 지방행정기관 15곳, 교육행정기관 20곳을 방문하여 행정에 관한 가장 일반적인 정보를 검색해본 결과 게시판 형태의 일차원적 관리가 이루어지는 사이트가 대부분들이었다. 또한 웹사이트 상에서 찾을 수 없는 정보를 얻기 위해서는 직접 전화를 하거나 게시판에 글을 남긴 후 답변을 기다릴 수밖에 없어 사용자 편의를 보장하고 있지 않는 사이트가 대부분이었다.

행정업무의 특징상 업무가 분산되어 있어서 온라인상에서나

오프라인상에서도 업무담당자를 알기도 쉽지 않았다. 행정업무 분야의 하나인 교육행정업무에서도 상황은 마찬가지이다. 현재 행정분야에서는 정보를 제공받는 사용자입장에서 이루어진 연구가 미비하며 검색된 사이트 중에서는 질의응답시스템을 사용하고 있는 곳은 없었다. 이에 담당자와 사용자가 보다 효율적으로 함께 사용할 수 있는 None-Stop Service의 필요성이 증대되고 있다.

3. 시스템 구성

제안하는 시스템은 크게 세부분으로 나누어 볼 수 있다. 그림1에서 보는 것과 같이 사용자가 스스로 질문을 하고 답변을 받는 사용자 인터페이스, 사용자 질의를 처리하여 답변을 제공해주는 내부처리엔진과 마지막으로 업무담당자가 데이터베이스의 내용을 관리하는 업무관리자 인터페이스로 이루어져 있다.



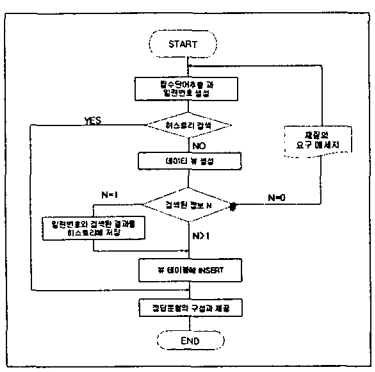
(그림 1) 시스템 구성도

3.1 사용자 인터페이스

사용자 환경은 사용자가 질의를 하고 답변을 받는 인터페이스로서 내부처리엔진의 과정을 통하여 얻어진 정답을 볼 수 있는 곳이다.

3.2 업무관리자 인터페이스

사용자에게 정보를 제공해주기 위해서 대단위의 행정정보가 필요하다. 본 논문에서는 이 행정정보를 업무 담당자가 직접 데이터베이스에 저장할 수 있도록 하였다. 관리자 이름으로 행정정보가 저장이 됨과 동시에 현재의 업무관리자를 자동으로 구성하여 업무내용만으로도 업무관리자를 쉽게 찾을 수 있도록 하였다.



(그림 2) 내부처리엔진 흐름도

3.3 내부처리엔진

내부처리엔진은 그림 2에서 보는 것과 같이 사용자 질의를 행정업무에 맞게 일반문형으로 질의를 변형한 후 질의분석과정을 거친다. 분석된 질의를 처리하여 업무정보 데이터베이스에서 정답을 검색 및 추출한 후 최종적으로 동적 정답 문형의 구성으로 사용자에게 정답을 제공해주는 처리과정이다.

3.3.1 질의분석

사용자 질의는 정형화된 형태가 있는 것이 아니라 개인의 특성에 따라 여러 형태로 이루어진다. 질의응답시스템에서는 이러한 정형화되지 않은 질의를 분석하여 사용자의 의도를 추출해 내야한다. 사용자의 의도가 정확히 추출되지 않는다면 정답의 정확성을 보장 할 수가 없기 때문이다. 본 논문에서 사용자 질의 분석은 공개된 형태소 분석기를 사용하여 논문의 특성에 맞게 불용어 제거, 복합어구성을 추가하여 적용하였다.

3.3.2 질의처리

사용자 질의가 질의분석 단계에서 질의의도에 맞게 추출되면 질의처리 단계에서는 이 추출된 질의의 정답을 제공해 줄 수 있는 상태로 만든다.

<표1> object_base 테이블

object_no	data
501	WHEN
502	WHERE
503	WHO
504	WHAT
505	HOW
506	WHY

<표2> topic_base 테이블

topic_no	data
4001	오리엔테이션
4002	입학식
4003	학생증발급
4004	수강신청
4005	논문제출
...	...

분석된 질의는 두 단계를 거쳐서 고유의 일련번호를 가지게 된다. 사용자의 의도가 담긴 의문사로 구성된 표1과 행정업무의 내용이 담긴 표2와의 매칭을 거쳐서 object_data와 topic_data값을 가진다. 본 논문에서는 웹사이트의 FAQ와 같은 기능을 가진 히스토리를 검색하기 위해서 별도의 일련번호(result_no)생성이 필요하다. 일련번호는 [object_no· topic_no]로 구성하였다.

3.3.3 정답의 검색 및 추출

정확한 정답을 제공하기 위하여 정답의 검색 및 추출과정에서는 데이터뷰 와 히스토리 두 가지 과정을 거친다. 먼저 히스토리는 한번 검색된 내용이 저장되어 있는 곳으로서 질의처리 과정을 거쳐서 생성된 각 질의의 고유 일련번호를 통해서 히스토리에 있는 데이터들의 정답 여부를 검색한다. 행정업무는 각 시기마다 특정적인 업무가 있고 그 시기에는 반복된 질문들이 많기 때문에 히스토리에 한번 질의된 내용은 반복질의 때 빨리 응답해 줄 수 있다는 장점이 있다. 최초로 질의된 질문은 데이터뷰를 통해서 검색된다.

뷰는 물리적 공간을 차지하지 않을 뿐 아니라 자료검색을 단순화하여 원하고자 하는 데이터만을 볼 수 있기 때문에 사용자 의도에 알맞은 정답을 제공해 줄 수 있다. 이 과정에서는 검색된 정답의 개수를 파악하여 정답제공과정에 영향을 준다. 검색된 정답의 개수를 N 이라고 하면, $N=0$일 때는 검색된 정보가 없기 때문에 시스템은 사용자에게 재 질의요구 메시지를

보낸다. $N=1$일 때는 검색된 정보를 뷰에 저장을 하고 정보가 하나라는 것은 정확한 정보라는 것을 의미하기 때문에 히스토리에 저장을 한다. $N>1$일 때는 검색된 정보를 모두 뷰에 저장을 하지만, 히스토리에는 추가하지 않는다.

3.3.4 정답제공

질의응답시스템들은 정답을 제공해 주기 위하여 미리 일정유형의 정답문형을 정의해 놓고 그 유형에 맞추어 정답을 제공해 준다. 하지만 본 논문에서 이러한 정답문형의 정의가 필요하지 않고 데이터 뷰의 컬럼명을 통해서 정답 문형을 구성하여 사용자에게 제공해준다. 표3에서 보듯이 데이터뷰에서 ["topic_data"은(는) "result_data"입니다. 담당자 : "staff_name" ("staff_tel")]라는 정답 문형을 구성하게 되고 정답으로는 [입학식은(는) 3월 2일입니다. 담당자: 이미나(620-6390)]을 제공해준다. 히스토리에 정답이 있는 경우도 같은 형식으로 정답을 제공해준다. 사용자에게 정답을 제공해주고 나면 그 사용자 질의에 맞추어 생성된 데이터 뷰는 삭제를 한다.

<표3> 정답제공을 위한 뷰

no	term	staff_name	staff_tel	objec_data	topic_data	result_data
1	2005-1	이미나	620-6390	when	입학식	3월 2일

4. 실험 및 평가

본 논문에서 제안한 시스템의 유용성을 평가하기 위하여 Windows XP Pro, IIS 5.0 , PHP 5.0 , Mysql 로 구성된 시스템을 구현하였다. 행정업무정보는 2005학년도 1학기 부경대학교 교육대학원 전산교육전공의 업무내용을 참고하였고, 사용자 질의는 30자이내의 단문형식으로 자주 묻는 질의 100개를 사용하여 질문세트를 구성하였다. 질의응답시스템을 평가하기 위하여 질의응답시스템의 성능평가를 위한 척도로 사용되고 있는 응답의 평균역순위(MRR:Mean Reciprocal Rank)과 단일키워드, 이중키워드, 자연어질의로 비교한 시스템 적중률(HR:Hit Rate)을 사용한다.

$$MRR=\frac{1}{n}\left(\sum_{i=1}^{n}\frac{1}{rank_i}\right) \quad (1)$$

- $rank_i$: i 번째 질문에 대한 응답으로 제시한 것들 중에서 첫 번째로 정답인 것의 순위
- n : 질문의 수

$$HR=\frac{1}{n}\sum_{i=1}^{n} r_i \quad (2)$$

- r_i : i번째 질문에 대한 응답 중 정확한 응답의 포함 확률

<표4> 시스템 적중률의 비교

질문형식	단일키워드	이중키워드	자연어질의
HR(%)	30.4 %	79.9 %	88.5 %

평가 결과 MRR은 0.87를 얻었고 시스템 적중률(HR)은 표 4와 같은 결과를 얻을 수 있었다.

5. 결론 및 향후 연구

인터넷이 보편화됨에 따라 상업적 용도뿐 아니라 사용자에게 정보를 제공해 주기 위해 많은 홈페이지들이 만들어지고 있다. 행정업무 역시 예외가 아닐 수 없다. 하지만 현재 행정업무에 사용되고 있는 홈페이지에서는 단순 정보만 제공할 뿐 사용자의 질문에 대하여 신속하게 답변을 주지 못하고 있다. 이에 본 논문에서는 교육행정업무에 적용한 질의응답시스템을 제안하였다.

제안한 시스템은 신속 정확하게 정답을 제공하기 위하여 자연어로 된 사용자 질의의도를 추출하고 각 사용자에 맞추어 데이터뷰를 생성한 후 정확한 응답을 제공 하도록 하였고 반복적인 질문은 동적인 FAQ 관리기능인 히스토리를 통해서 신속히 제공 하도록 하였다. 제안한 시스템의 성능을 평가한 결과 단일키워드와 이중키워드에 대한 시스템 적중률(HR)이 각각 30.4% 와 79.9% 를 보였다. 자연어질의의 시스템 적중률(HR)이 88.5% 인 것을 비교해 볼 때 본 논문에서 제안한 시스템이 단일 키워드에 비해서 58.1% , 이중키워드에 비해서는 8.6% 뛰어난 시스템 적중률(HR)을 보임을 알 수 있었다.

향후 연구과제로는 단순히 정답만을 제공해주는 시스템이 아닌 목적형대화가 가능한 질의응답시스템의 설계 및 구현이다.

참고문헌

[1] Aggarwal C., Wolf J., and Yu P., *A frame-work for the potimizing of WWW advertising*, In W. Lamersdorf and M. Merz(Eds), Trends in Distributed Systems for Electronic Commerce, LNCS 1402, Berlin: Springer, 1998.

[2] Muller J. and Pischel M., "Doing business in the information marketplace," In Proccedings of the 1999 International Conference on Autonomous Agents, pp.139-146, 1999.

[3] Chai J., Lin J., Zadrozny W., Ye Y., Stys-Budzikowska M., Horvath V., Kambhatla N., and Wolf C., The role of a natural language conversational Interface in online sales: A case study, International Journal of Speech Technology, Vol.4, pp.285-295, 2001.

[4] D. Donovan, S. Harabaqiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju and V. Rus, "LASSO: A Tool for Sur -fing the Answer Net", In Preceedings of Trec-8, pp.65-74, 1999.

[5] 강승식, "한국어 정보처리의 현황 및 발전 방향" , 한국음성과학회 제 6차 학술발표대회 학술논문집, 1999.

[6] 박미화, 원형석, 이근배, "구문 분석에 기반을 둔 한글 자연어 질의로부터의 불리언 질의 생성" , 정보과학회 논문지(B),26권,10호, pp.1219-1229, 1999.

[7] 이경순, 김재호, 최기선, " KorQuA: 질의응답에서 자료유형을 고려한 대답검색과 대답해석, 한국인지과학회 춘계학술대회 학술논문집, 2000.

[8] 황이규, 김현진, 장명길, "질의응답 기술 개발" , 한국정보처리학회,Vol.11, pp.48-56, 2004.