Journal of Digital Contents Society Vol. 11 No. 3 Sep. 2010(pp. 375-383)

문서 말뭉치 기반 질의응답 시스템

김한준*, 김민경**, 장재영***

요 약

질의응답시스템을 구축하는데 있어서 사용자 질의로 입력된 자연어 문장을 문법적 또는 의미적으로 완벽하게 분석하는 작업과 그 질의에 대한 정확한 답변을 찾아내는 작업은 쉬운 일이 아니다. 본 논문에서는 질의응답시스템 구축의 난제를 극복하기 위해, 문서 말뭉치에 기반하여 질의문을 자동 생성, 저장하여 이를 키워드로 검색하는 새로운 방식의 시스템을 제안한다. 질의문 생성을 위한 기본 아이디어는 수집 문서의 주요 문장에 대해 고유명사인식 기술을 활용하여 사람, 사물, 장소, 시간 등의 고유명사를 인식한 후, 각 고유명사에 해당하는 자연어 질의문을 생성하는 것이다. 질의문은 두가지 유형인 단순형 및 문장구조유지형 질의문으로 구분한다. 시스템은 이렇게 준비된 질의문 데이터베이스를 가지고 입력된 검색 키워드에 대하여 관련 질의문과 답변을 쉽게 얻을 수 있다. 본 연구의 관건은 생성된 질의문이 명확한 해답을 도출할 수 있는 의미있는 질의문을 생성하는 것이다. 이를 위해 본 연구에서는 질의문의 원천이 되는 평서문장을 선별하는 원칙과 선별된 평서문으로부터 의미있는 질의문을 생성하는 방법론을 제시한다.

Text Corpus-based Question Answering System

Han-joon Kim*, Min-Kyoung Kim**, Jae-young Chang***

Abstract

In developing question-answering (QA) systems, it is hard to analyze natural language questions syntactically and semantically and to find exact answers to given query questions. In order to avoid these difficulties, we propose a new style of question-answering system that automatically generate natural language queries and can allow to search queries fit for given keywords. The key idea behind generating natural queries is that after significant sentences within text documents are applied to the named entity recognition technique, we can generate a natural query (interrogative sentence) for each named entity (such as person, location, and time). The natural query is divided into two types: simple type and sentence structure type. With the large database of question-answer pairs, the system can easily obtain natural queries and their corresponding answers for given keywords. The most important issue is how to generate meaningful queries which can present unambiguous answers. To this end, we propose two principles to decide which declarative sentences can be the sources of natural queries and a pattern-based method for generating meaningful queries from the selected sentences.

Keywords: information retrieval, question answering systems, text corpus

1. 서론

최근 기존 정보검색 엔진의 한계를 극복하기 위해 자연어 질의를 지원하는 질의응답시스템(Question Answering Systems, QA) 기술을 병합하기 위한 노력이 진행되고 있다[1][2]. 이는 기존 정보검색 엔진이 문서 단위의 결과를 반환하기 때문에 사용자의 구체적인 정보 요구에 대

※ 제일저자(First Author): 김한준
접수일:2010년 08월 16일,수정일:2010년 10월 01일,
완료일:2010년 10월 01일
* 서울시립대학교 전자전기컴퓨터공학부 (교신저자) khj@uos.ac.kr
** 아시아투데이 인터넷부 comengs@asiatoday.co.kr
*** 한성대학교 컴퓨터공학과 jychang@hansung.ac.kr
▣ 본 연구는 서울형산업지원개발사업 (과제번호: NT080624, 연구과제명: 집단지성 기반 사용자 인지 웹검색시스템 개발)의 지원에 의해 수행되었으며, 또한 2010년도 한성대학교 교내연구비로 지원되었음.

해서 정확한 답변을 제공하지는 못하기 때문이다. 질의응답시스템은 자연어로 된 질의 문장에 대하여 문서수준이 아닌 단답형 수준의 구체적 답변을 제시하는 시스템이다. 질의응답시스템에 관한 연구는 자연어 질의문의 해석, 답변 검색 등에 있어서 그 정확도를 높이기 위해 다양한 시도가 이루어져 왔지만, 기본적으로 질의문 해석의 오류 가능성을 최소화 하기는 어려운 문제이다. 즉, 명확한 답변을 도출하기 위해 질의문 분석의 정확율이 높아야 하는데, 질의문의 형식이 복잡하거나 정형성이 부족하다면 질의문 해석의 오류율이 높아질 수밖에 없다. 이는 결국 시스템의 신뢰도가 제한을 받게 됨을 의미하고, 합당한 수준의 신뢰도를 만족하기 위해서는 사용자가 작성하는 질의문의 자유도가 높아지기 어렵다.

관련 시스템으로서 IBM사의 Jennifer ChuCarroll 연구원이 개발한 'Piquant' 시스템[3][4]은 추론과정을 자동화하여 자연어 질의에 대한 답변의 정확도를 높였다. 예를 들어, "Who is Canada's prime minister?" 라고 질의했을 때, 수상 직함과 수상 성명이 하나의 문장 내에 존재하지 않아도 캐나다 수상의 성명이 반환될 수 있다. 이는 자연어 문장을 처리하는데 있어 기존의 시스템보다 인공지능적 접근방법을 제시했다는데 의미는 크지만[5], 복잡도가 높은 자연어 문장에 대한 분석 오류 문제와 자연어 질의 작성의 자유도가 높지 못한 문제는 극복하지 못하고 있다.

본 논문에서는 이러한 문제를 근본적으로 해결할 수 있는 새로운 형태의 질의응답시스템을 제안하다. 제안 시스템은 질의문과 그 답변을 미리 생성하여 데이터베이스에 저장해놓고 질의문을 사용자가 키워드 방식을 검색할 수 있다. 핵심적인 사항은 자연어 질의문을 생성하는 원리와 '의미있는' 질의문을 생성하기 위한 방법론이다. 기본적 원리는 고유명사인식 기술을 적용하여, 인식된 고유명사를 질의하는 자연어 질의문을 생성하는 것이다. 그리고 '의미있는' 질의문을 생성하기 위해 2가지 형태의 질의문 생성 방식을 제안한다.

2. 관련 연구

질의응답시스템은 주어진 질의에 대해 구체적 수준의 응답을 주는 시스템이다. 전통적인 질의응답시스템의 단계는 다음과 같다. 사용자로부터 검색어가 입력이 되면 문장 분석을 통해 질의문을 특성에 맞게 분류한다. 예를 들면 질의문의 답변이 될 수 있는 지명, 인명, 날짜 등에 따라 분류될 수 있다. 질의문의 분석결과를 이용하여 시스템은 수집문서에서 질의문의 답변과 매칭되는 문장을 찾아서 사용자에게 제공한다. START*[7], Ask Jeeves**, Mulder[8], Webclopedia[9], MURAX[10] 등은 잘 알려진 질의응답시스템이다.

전형적인 질의응답시스템의 전체적인 구조는 그림1에 나타나 있다***. 우선 수집문서가 정보검색기술을 사용하여 색인되어 있어야 한다. 자연어 질의문은 대체로 적당히 문맥 해석이 될 수 있도록 일관된 문장으로 가정한다. 주어진 질의문은 Link Parser, CONTEXT와 같은 문맥해석기에 의해 분석된다. 그리고 나서 축적(또는 학습)된 지식(또는 모델)을 활용하여 질문분류 및 답변 유형을 결정한다. 그 유형에 따라 답변이 될 수 있는 문서내의 특정 부분인 후보 세그먼트를 추출한다. 여기서 세그먼트는 질의응답시스템의 출력이 되는 단위를 의미하며, 명사 또는 명사구가 그 예이다. 그리고 후보세그먼트는 정보검색 기술에 의존하여 선택된 문서 내에서 추출되는 것이다. 후보세그먼트들은 질의문의 정보와 견주어서 점수가 매겨진다. 응답처리 단계에서는 질의문 분석에 의해서 응답형태가 결정되었으면 세그먼트 매칭(segment-matching) 단계에서 근사검색을 수행하여 적합한 순서를 정하여 답변 리스트를 사용자에게 전달한다. 기존 시스템은 유형은 두 가지로 분류할 수 있다. 하나는 Shallow 유형이고, 다른 하나는 Deep 유형이다. Shallow 유형 시스템은 키워드 기반 기술을 이용하는 것으로서, 주어진 키워드가 포함되어 있는 문장 또는 절을 검색하여, 그것들의 문법적 구조 및 키워드 순서에 따라 순위를 매기는 방식이다[7][8]. 비교하여 Deep 유형 시스템은 검색 대상이 되는 문서에 포함된 문장들에 대해서 고유명사 인식, 상호참조연결 인식, 논리적 의미관계 추론 등의 기술을 이용하여 단어, 절, 문장

* http://www.ai.mit.edu/projects/infolab
** http://ask.com
*** 그림1에 나타난 구조도는 Webclopedia와 Mulder 시스템에 따름

들간의 관계 구조를 미리 생성하는 방식이다 [3][11]. 이를 위해서는 VerbNet*과 같은 온톨로지를 활용해야 하는 단점이 있다.

이러한 전통적 시스템의 추구하는 구조에서는 질의문의 분석과 세그먼트 추출 과정에서 그 정확도가 시스템의 성능을 결정짓는다. 그것의 정확도를 개선하기 위해 자연어 처리, 인공지능, 논리, 통계 등의 요소를 병합하는 노력을 기울이고 있지만 확실한 개선책이 되지는 못하고 있다. 기존 시스템의 접근방식이 사용자의 질의에서 출발하는 것이라면, 본 연구는 질의의 대상이 되는 문서말뭉치로부터 시작한다고 할 수 있다. 다시 말해서, 본 연구가 지향하는 질의응답시스템은 문서말뭉치 자체를 지식베이스로 간주하여 이로부터 사용자가 질의할 수 있는 질의문을 자동생성하고 그것의 명확한 답변을 미리 준비하는 것이다.

(그림 1) 전통적 질의응답 시스템의 구조

3. 문서말뭉치 기반 질의응답 시스템

3.1 질의문 자동생성을 위한 기본 원리

질의문을 자동으로 생성하기 위해서는 주어진 평서문에서 질의 대상이 될 수 있는 고유명사를 찾는다. 이를 위해 '고유명사인식기술(Named Entity Recognition, NER)' [6]을 활용한다. 여기서 인식하는 고유명사의 유형은 인물, 위치 (또는 지위), 지역, 시간, 기관, 기타로 분류되며, 각 유형에 대한 태그명은 *<PERSON>*, *<POSITION>*, *<LOCATION>*, *<TIME>*, *<ORGANIZATION>*, *<MISCELLANY>* 이다.

예를 들어, 아래와 같은 문장이 있다고 하자.

> "Dr.Martin Luther King was assassinated on April 4, 1968, in Memphis, Tennessee."

위 문장에 고유명사인식기술을 적용하면 아래와 같은 결과를 얻을 수 있다.

> Dr.Martin Luther King : *<PERSON>*
> April 4, 1968 : *<TIME>*
> Memphis, Tennessee : *<LOCATION>*

고유명사인식기술을 적용한 결과를 보고 각 고유명사마다 관련된 의문사를 적용하여 질의문을 생성한다. *<PERSON>*에는 'Who', *<TIME>*에는 'When', *<POSITION>*에는 'Where/What', *<LOCATION>*에는 'Where', *<MISCELLANY>*에는 'What'의 의문사를 적용하여 질의문을 생성한다. 그래서 위의 예에서 나올 수 있는 질의문은 다음과 같다.

> When was Martin Luther King. assassinated?
> Where was Martin Luther King assassinated?
> Who is Martin Luther King?

그리고 고유명사 인식결과가 동급이고 'of', 'and', ','(쉼표)로 연결된 고유명사는 그것들을 합쳐 하나의 대답이 가능하다. 예를 들어, 문장 "George Walker Bush is the forty-third and current President of the United States of America."을 NER에 적용하면 George Walker Bush : *<PERSON>*, President : *<POSITION>*, United States of America : *<LOCATION>* 라는 결과를 얻게 된다. 여기서, 'President'와 'United States of America'가 동급으로서 'of'로 연결되어 있어서 'President of the United States of America'를 최종 답변으로서 제시할 수 있다. 또 다른 예로서, 명사구 'German-Swiss poet, novelist, and painter'와 같은 유형도 하나의 답변으로서 제시되어야 한다.

3.2 의미있는 질의문 생성을 위한 원칙

기본 원리에 따라 생성된 질의문 중에는 의미없는 질의문도 포함된다. 여기서 '의미없는' 질의문이란 자동 생성된 질의문이 특정의 해답을 요

* http://verbs.colorado.edu/verb-index/

구할 수 없거나 답변이 너무 과다할 수 있는 문장을 의미한다. 예를 들어 다음과 같은 어떤 문서에 다음과 같은 문장이 있다고 가정하자.

"Martin Luther King was killed in 1968."

위 문장에 대해 고유명사인식기술을 적용하면 아래와 같은 결과를 얻을 수 있다.

Martin Luther King : <*PERSON*>
in 1968 : <*TIME*>

여기서 3.1절에서 기술한 질의문 생성원리에 따라 주어에 해당하는 고유명사(즉, Martin Luther King : <*PERSON*>)를 질의하는 질의문 "Who was killed in 1918?" 이 가능하다. 하지만 이 질의문은 그 답변의 수가 너무 많아 의미성이 거의 없는 질의문이다. 이와 반대로 '의미있는' 질의문은 질의하는 대상에 대한 명확하고 분명한 의미가 담겨있고, 그 답변의 개수가 제한적인 문장이 되어야 한다.

자동 생성된 질의문은 당연히 사용자가 직접 입력하는 질의문과 같이 '의미있는' 질의문이 되어야 한다. '의미있는' 질의문은 생성된 질의문이 단독으로 사용되었을 때 그 의미가 무엇을 질의하는지 - 즉, 특정 인물, 시간, 장소 등- 분명한 문장을 의미한다. 위의 예에서 <*TIME*>을 묻는 "When was Dr. Martin Luther King killed?"는 한정된 답변 (즉, 1968)을 도출하는 의미있는 질의문이 된다. 질의문의 의미성을 결정하기 위해 사람의 경험 또는 지식베이스에 의지하는 것은 실용적이지 못하여, 본 연구에서는 의미성을 가지는 문장들의 패턴을 분석한 결과, 두 가지 유형의 '의미있는' 질의문 생성 방안을 찾아냈다.

3.2.1 의미있는 질의문의 생성을 위한 원칙

질의문의 의미성을 분별하기 이전에, 질의문 생성에 원천이 되는 평서문의 요건을 먼저 따져봐야 한다. 원천 평서문의 요건은 다음과 같다

① 동사를 기준으로 주어부와 서술부에 대명사가 존재하지 않아야 한다.

② 동사를 기준으로 주어부 또는 서술부에 고유명사가 1개 이상 포함되어야 한다.

<표 1> 명사단독형 질의문의 생성

원 문	Dr.Martin Luther King was assassinated on April 4, 1968, in Memphis, Tennessee.
고유명사 인식	*Dr. Martin Luther King* : <*PERSON*> *April 4, 1968* : <*TIME*> *Memphis, Tennessee*: <*LOCATION*>
↓	
자동생성 질의문	답변
When was Dr.Martin Luther King,Jr. assassinated ?	April 4, 1968
Where was Dr.Martin Luther King,Jr. assassinated ?	Memphis, Tennessee

원 문	Albert Einstein received the Nobel Prize in physics in 1921 for his services to Theoretical Physics.
고유명사 인식	*Albert Einstein* : <*PERSON*> *Nobel Prize* :<*PERSON*> *in 1921* : <*TIME*> *Theoretical Physics* : <*MISCELLANY*>
↓	
자동생성 질의문	답변
When did Albert Einstein received the Nobel Prize?	in 1921
For what did Albert Einstein received the Nobel Prize?	his services to Theoretical Physics

위에서 제시한 원천 평서문의 요건을 만족하는 평서문장에 대해 고유명사인식 및 파싱 과정을 수행하여 얻은 정보를 토대로 두 가지 유형, 즉 '명사단독형' 질의문, '구조유지형' 질의문을 생성한다. 여기서 편의상 동사를 기준으로 왼쪽 부분은 주어부, 우측 부분은 서술부라 명명한다.

'명사단독형' 질의문은 주어 역할을 하는 고유명사에 대해서 서술부에 포함된 고유명사를 각기 질의하는 질의문을 의미한다. 서술부에 존재하는 모든 고유명사에 대해서 인식된 고유명사의 유형에 따른 의문사를 사용함으로써, 최대 서술부에 존재하는 고유명사 개수만큼의 질의문을 생성하게 된다. 단, 동사가 목적어를 취하는 타동사의 경우에는 그 속성상 목적어를 포함한 질의문이어야 한다. 표1은 명사단독형 질의문을 생성한 예를 보여준다. 두 번째 원문에 대한 질의문 "For what did Albert Einstein received the Nobel Prize?"에 대한 답변은 원칙적으로 고유명사 <MISCELLANY>에 해당하는 'Theoretical Physics'이나 세밀한 문장분석을 통해 'his services to Theoretical Physics'로 확장할 수 있다.

'구조유지형' 질의문은 서술부에 명사(보통명사와 고유명사를 포함)를 포함하는 목적어(또는 보어)가 존재할 때*, 서술부 구조를 유지하면서

* be동사인 경우, 목적어 대신 보어 개념을 사용한다.

주어부의 고유명사를 질의하는 질의문을 의미한다. 이 원칙에 따라 서술부의 구조를 유지하면서 주어에 해당하는 고유명사를 질의하는 의미있는 문장을 만들 수 있다. 표2는 구조유지형 질의문을 생성한 예를 보여준다.

구조유지형 질의문의 경우, 하나의 원천 평서문으로부터 하나의 질의문이 생성된다. 여기서 명사를 포함하는 목적어가 존재하지 않는 경우는 의미없는 질의문이 생성되는 예가 많아 이런 경우는 배제하였다.

한편 문장구조를 유지하지 않은 채 서술부의 고유명사만을 가지고-즉, 고유명사 주위에 존재하는 부사구 및 수식구를 제외하여 구조를 유지하지 않고- 주어를 질의하는 의미있는 문장을 만들 수는 있다. 사실 질의문의 의미성은 서술부에 존재하는 명사의 수에 따라 결정된다. 명사의 수가 1개인 경우에는 의미없는 질의문의 될 가능성이 크다. 예를 들어, 평서문 “Hermann Hesse received the Nobel Prize in literature in 1946.”이 주어질 때, 서술부에 존재하는 고유명사 ‘Nobel Prize’ 만을 이용하여 만든 질의문 “Who received the Nobel Prize?”는 명확한 답변을 제시할 수 없는 것이다. 하지만 서술부 구조를 모두 활용한 질의문 “Who received the Nobel Prize in 1946?”은 의미가 있다. 다른 예로서, 평서문 “Einstein found the theory of relativity in 1905.”이 주어질 때, ‘relativity’라는 명사만을 사용하여 작성한 질의문 “Who found the theory of relativity?”은 충분히 의미가 있다. 이처럼 동사의 의미에 따라서 질의문의 의미성이 결정되어서, 서술부의 고유명사의 개수를 고정하는 것은 어렵지만 실험적으로 명사의 수가 2개 이상이 되면 질의문의 의미성이 강해지는 것으로 나타났다.

3.3 질의문 검색

본 연구에서는 이미 질의응답을 위한 질의문을 준비해놓기 때문에, 전통적인 질의응답시스템에서 사용자가 직접 질의문의 작성을 요구하는 방식은 적합하지 않다. 그리고 대부분의 경우 사용자가 작성할 수 있는 질의문은 명사단독형 질의문에 가까울 것이다 (표1 참조). 비교해서 구조유지형 질의문은 실제 사용자가 작성하기에는 어려운 형태가 될 수 있다 (표2 참조).

그러므로, 제안 시스템의 실용성을 높이기 위해서는 자동 생성된 질의문 자체에 대한 검색기능이 포함되는 것이 바람직하다. 그리고 실제 시스템 구현시 답변이 이루어지고 나서 그 답변과 관련된 구체적 문서 내용을 확인할 수 있는 서비스가 제공되도록 하였다. 이와 관련한 검색 인터페이스에 대한 설명은 4.3절로 미룬다.

<표 2> 구조유지형 질의문의 생성

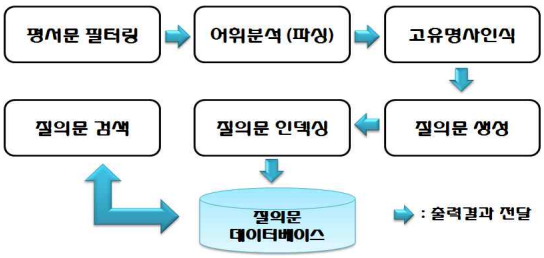
원문	George Walker Bush is the forty-third and current President of the United States of America.
고유명사 인식	George Walker Bush: <PERSON> United States of America: <LOCATION>
↓	
자동생성 질의문	답변
Who is the forty-third and current President of the United States of America ?	George Walker Bush

원문	Park Ji-Sung is a professional South Korean footballer who plays for the English football club Manchester United in the Premier League.
고유명사 인식	Park Ji-Sung: <PERSON>, South Korean, English: <LOCATION>, Manchester United, Premier League: <MISCELLANY>
↓	
자동생성 질의문	답변
Who is a professional South Korean footballer who plays for the English football club Manchester United in the Premier League?	Park Ji-Sung

4. 시스템 구현

4.1 시스템 구조

3.2.1절에서 설명한 의미있는 질의문 생성을 위한 두 가지 원칙을 가지고, 우리는 그림2와 같은 형태의 질의응답 시스템을 구현하였다. 생성된 질의문은 향후 검색을 위해서 데이터베이스에 인덱싱된다.



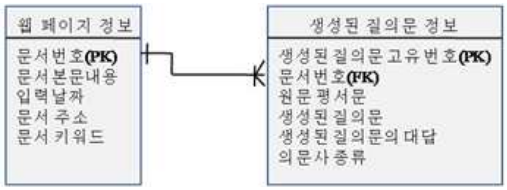
(그림 2) 말뭉치 기반 질의응답시스템 구조도

- 평서문 필터링: 말뭉치로터 자동질의문 생성의 대상이 되는 평서문을 골라내는 작업을

수행한다 (3.2.1절 참조).

- 어휘분석(파싱): 주어진 문장에서 품사 및 주어부, 서술부를 감지하는 수준에서 shallow parsing을 수행한다.
- 고유명사 인식: 고유명사의 유형을 판별하는 작업을 수행한다.
- 질의문 생성: 명사단독형 질의문과 구조유지형 질의문을 생성한다 (3.2.1절 참조).
- 질의문 인덱싱: 생성된 질의문은 그것의 원천이 되는 평서문과 속한 문서 정보를 데이터베이스에 저장한다.
- 질의문 검색: 사용자가 입력한 키워드를 포함한 질의문을 찾아내고, 관련 문서의 상세 정보를 보여준다.

4.2 데이터 모델링



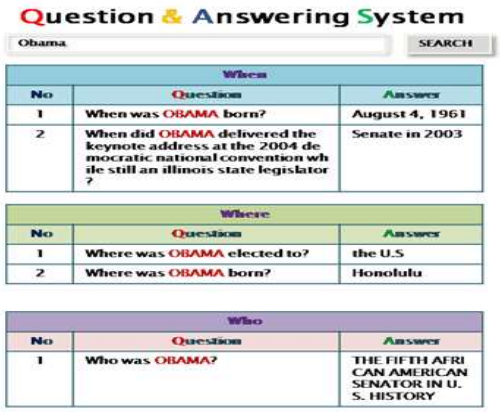
(그림 3) 말뭉치 기반 질의응답을 위한 데이터모델

자동생성된 질의문을 활용하여 질의응답을 지원하기 위해 그림3과 같은 형태의 데이터모델을 작성하였다. 자동 생성된 질의문, 해당 답변, 원본 평서문, 원본문서 등의 정보를 데이터베이스에 저장한다.

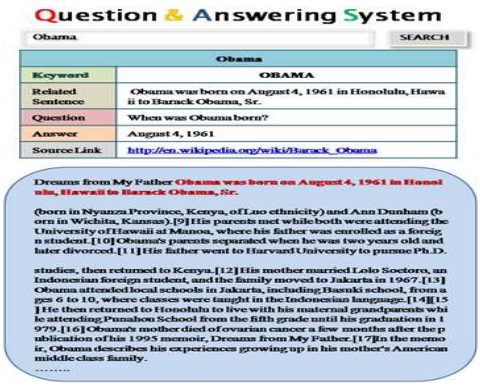
구체적으로, '웹 페이지 정보' 테이블은 원본 문서의 메타데이터인 '문서번호', '문서내용', '입력날짜', '문서주소', '문서키워드' 등을 저장한다. '문서번호' 컬럼은 시스템에서 부여하는 고유번호로 생성된 '질의문 정보' 테이블의 문서번호와 연결되는 주키 역할을 한다. '문서내용' 컬럼은 수집 문서의 본문 내용이고, '문서주소' 컬럼은 웹 URL(Uniform Resource Locator)주소이다. '문서 키워드' 컬럼은 문서의 요약을 위해 제공되는 정보이다. 또한 '자동생성된 질의문 정보' 테이블은 수집문서로부터 선별된 원본평서문, 자동생성된 질의문, 답변, 의문사 유형등의 정보를 저장한다. '질의문 번호' 컬럼은 시스템에서 부여하는 고유번호이다. '원문평서문' 컬럼에는 의미있는 질의문을 생성을 위해 선별된 평서문 문장, '생성된 질의문' 컬럼에는 원본 평서문으로부터 자동 생성된 질의문 정보를 저장한다. '질의문 답변' 컬럼은 관련 질의문의 답변을 의미한다.

4.3 질의문 검색 인터페이스

그림 4는 데이터베이스에 미리 저장되어 있는 질의문을 사용자가 질의를 통해 확인할 수 있는 검색 인터페이스를 보여준다. 그림에서 보는 바와 같이 사용자가 검색어를 입력하면 이를 포함한 질의문이 의문사별로 구분되어 화면상에 출력된다. 이를 통해 데이터베이스에 저장된 질의문과 해당 답변 정보는 검색 및 확인이 가능하다. 사용자는 원하는 질의문을 조회할 수 있도록 하여 분명하지 않은 질의 의도를 가지고 있음을 감안하여 가능한 모든 질의문을 표시하였다.



(그림 4) 질의문 검색 및 검색결과 인터페이스



(그림 5) 질의문의 답변이 포함된 문서의 상세 정보

그리고 질의문과 해당 답변을 바로 보여줌으

로써 질의문과 대답을 쉽게 볼 수 있도록 화면을 구성할 수 있다. 사용자가 만약 답변에 대한 상세한 정보를 알고 싶을 때에는 해당 답변을 선택하여, 답변과 관련된 상세한 원문 내용을 확인할 수 있다. 그림5의 상세정보 화면은 질의문을 생성한 평서문 문장과 원본 문서의 전체내용과 해당 웹 페이지의 주소를 보여준다.

5. 실험

5.1 실험환경

본 시스템의 유용성을 증명하기 위해'백과사전' 성격을 띠는 위키피디아(Wikipedia) 문서 데이터를 사용하였다. 이는 위키피디아가 질의 대상에 대한 정의와 관련 내용이 명확하게 기술되어 있어 자동으로 질의문을 생성하는데 적합하기 때문이다. 위키페이지의 웹 페이지로부터 소스를 제공받아 3만개 이상의 평서문 데이터를 사용하였다.

실제 구현 측면에서, 위키피디아 문서 데이터로를 평서문 선별 모듈에 적용하여 질의문 생성이 가능한 문장만을 추려낸 후 이를 고유명사인식 기술을 적용하여 그 결과를 XML(eXtensible Markup Language)문서로 얻어낸다. 본 연구에서는 고유명사인식도구로서 Alias-i사의 'LingPipe'*를 이용하였다. LingPipe 도구가 출력하는 XML문서를 분석하여 질의문을 자동 생성한 것이다. XML문서는 자바 XML 파싱(Parsing) 모듈을 이용하였다. 그리고 질의문 데이터베이스를 구축하기 위한 DBMS로서 관계형 데이터베이스인 MySQL 5.0을 사용하였다.

5.2 실험내용

본 실험에서는 두 가지 측면을 검증하고자 한다. 하나는 본 연구에서 제안한 질의문 생성 방안이 충분히 의미성을 가지고 있는지 검증하는 것이고, 다른 하나는 구조유지형 질의문의 경우, 서술부에 존재하는 명사의 개수에 따라서 생성되는 질의문의 의미성 여부를 실험적으로 측정하고자 한다.

* Alias-i.com 사에서 개발한 텍스트 처리 도구이며, 자연어 처리를 수행하기 위한 API를 제공한다.

<표 3> 명사독립형 질의문 생성

원문 1	Sejong the Great was the fourth king of the Joseon Dynasty of Korea.
↓	
자동생성 질의문	답변
What Sejong the Great was the fourth king of the Joseon Dynasty of Korea.	Sejong the Great

원문 2	The Beatles were an English rock band, formed in Liverpool in 1960
↓ (서술부 고유명사: Liverpool<LOCATION>, in 1960<TIME>	
자동생성 질의문	답변
Where was formed The Beatles formed?	in Liverpool
When was The Beatles formed?	in 1960

원문 3	Albert Einstein received the Nobel Prize in physics in 1921 for his services to Theoretical Physics.
↓ (서술부 고유명사: Nobel Prize<MISCELLANY>, Theoretical Physics<MISCELLANY>	
자동생성 질의문	답변
When did Albert Einstein received the Nobel Prize?	in 1921
For what did Albert Einstein received the Nobel Prize?	his services to Theoretical Physics

원문 4	Carnegie Mellon University is a prestigious private research university in Pittsburgh, Pennsylvania, founded by Andrew Carnegie in 1900.
↓ (서술부 고유명사: Pittsburgh, Pennsylvania <LOCATION>, Andrew Carnegie <PERSON>, in 1900<TIME>	
자동생성 질의문	답변
Where is Carnegie Mellon University?	Pittsburgh
By whom was Carnegie Mellon University founded?	Andrew Carnegie
When was Carnegie Mellon University founded?	in 1900

원문 5	Lincoln married Mary Todd , On November 4, 1842.
↓ (서술부 고유명사: Mary Todd <PERSON>, November 4, 1842 <TIME>)	
자동생성 질의문	답변
Whom did Lincoln marry ?	Mary Todd
When did Lincoln marry?	November 4, 1842

5.2.1 질의문의 의미성 검증

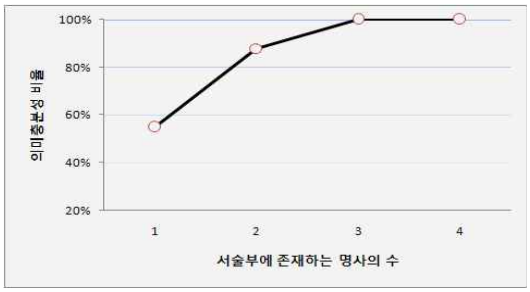
명사독립형 질의문을 생성하는 원칙에 따라 생성된 질의문의 의미성이 충분한지를 평가하기 위해 수천개의 문장에 대해 확인 작업을 수행하였으며, 모든 명사독립형 질의문은 그 의미성이 충분한 것으로 나타났다. 표3에서 대표적인 예를 나열한다.

다만, 명사독립형 질의문을 구성하기 위한 요건을 약간 약화시킬 수 있는 부분이 있다. 명사

독립형 질의문을 생성하는 요건에서 원천 평서문에서 목적어가 필요한 타동사가 존재하는 경우, 목적어를 포함한 질의문의 생성이 기본 원칙이나, 주변 수식어구 및 부사구의 도움없이 타동사에 대한 목적어가 분명하게 지정되는 경우에는 목적어 자체를 묻는 질의문이 가능할 수 있다. 이에 해당하는 동사는 *'marry', 'invent', 'create',' kill'* 등이다. 표3의 원문5 부분은 *marry* 동사를 사용한 문장에서 목적어 자체를 질의하는 질의문을 생성한 예를 보여준다. 그래서 목적어 자체를 질의할 수 있는 질의문을 생성하기 위해서는 타동사를 별도 관리하는 것이 바람직할 수도 있다. 하지만 이 방법 역시 동일한 동사를 포함한 문장들에서 문맥에 따라서 목적어를 질의하는 질의문이 가능한 경우와 그렇지 않은 경우가 존재하여 의미론적 접근이 필요한데, 이는 본 연구범위에서 제외하였다.

5.2.2 구조유지형 질의문의 의미성 분석

3.2.1절에서 기술한 바와 같이, 동사의 의미에 따라 질의문의 의미성이 결정되기 때문에 서술부의 고유명사의 개수를 고정하는 것은 바람직하지 않기에, 그 개수에 따른 질의문의 의미성 분석을 수행하였다. 본 실험에서는 서로 다른 동사를 포함한 200여개의 문장에 대하여 구조유지형 질의문 생성 원칙에 따라 질의문을 생성한 결과, 그 의미가 충분한 개수에 비율을 측정하였다. 여기서 이를 '의미충분성 비율'이라 부르겠다.



(그림 6) 구조유지형 질의문의 의미충분성 비율

그림6은 생성된 구조유지형 질의문들에 대하여 의미충분성 비율을 보여준다. 서술부에 존재하는 명사의 수가 3개 이상인 경우는 모든 질의문이 의미성이 충분하였으며, 명사의 수가 2개인 경우는 90%에 가까운 비율로서 의미성이 강한 질의문이 작성되었다. 이 실험을 근거로 구조유지형 질의문을 생성하는 원칙을 수립하는데 있어서 약 10%의 오류를 감안하고 서술부에 존재하는 명사의 수가 2개 이상이면 질의문의 의미성이 강한 것으로 간주하였다.

5.3 평가

본 논문에서는 질의응답시스템을 이용함에 있어 사용자가 키워드만으로 원하는 질의문을 선택할 수 있고, 이를 위해 질의문을 자동생성하는 시스템을 제안하였다. 본 연구는 기존의 질의응답시스템에서 사용자가 감수해야 했던 질의문 생성에 대한 부담감과 정확한 문장 분석의 어려움 등을 해결 할 수 있었을 뿐만 아니라. 자동생성된 질의문에 해당하는 정확한 답변을 자연스럽게 얻을 수 있기 때문에, 기존 질의응답시스템의 문제점을 쉽게 극복할 수 있다.

제안 시스템은 질의문 검색시스템이라 할 수 있는데, 이를 정보검색 기술의 관점에서 평가해보자. 기존 검색시스템의 목적은 키워드 입력을 통해 원하는 문서를 찾는 것이다. 제안 시스템의 검색 개념은 주어진 키워드에 대해 답변을 이미 가지고 있는 질의문을 검색하는 것이다. 이는 사용자가 원하는 질의문과 답변을 문서 수준만이 아닌 문장과 문서 수준에서 동시에 접근할 수 있는 '다차원' 검색엔진이라 평가할 수 있다.

6. 결론

본 논문에서 제안하는 질의응답시스템은 수집한 문서집합으로부터 의미있는 질의문과 해당 답변을 미리 생성함으로써, 기존 질의응답시스템 구축의 난제를 극복할 수 있는 가능성을 열었다는데 큰 의의가 있다. 구체적으로 질의문을 생성할 수 있는 평서문을 선별하는 원칙과 '의미있는' 질의문을 생성하는 원칙을 적용하여 명사단독형 및 구조유지형질의문을 구분 생성하며, 본 시스템의 유용성을 높이기 위해 질의문 검색의 개념을 제안하였다.

본 논문이 제안한 기법은 현재 실험적으로 위키피디아와 같은 백과사전식 문서에 적용하였으며, 이를 일반 문서로 확장하기 위해서는 의미있는 질의문을 생성할 수 있는 패턴을 세밀하게 도출하는 작업이 필요하다. 이를 위해 중요한 점이 문장 내 품사의 순차적 패턴과 수식어 등의

상호 관계를 파악하는 것이며, 이는 향후 중요한 연구 이슈로 삼을 것이다.

참 고 문 헌

[1] R. Srihari, W. Li, "A Question Answering System supported by Information Extraction", Proceedings of the 6th conference on Applied Natural Language processing, pp. 166 - 172, 2000

[2] S. Dumais, M. Banko, E. Brill, J. Lin, A. Ng, Web Question Answering: is more always better?", Proceedings of the 25th ACM SIGIR, pp. 291-298, 2002

[3] J. Chu-Carroll, J. Prager, C. Welty, K. Czuba, and D. Ferruci, "A multi-strategy and multi-source approach to question answering", Proceedings of the 10th Text Retrieval Conference (TREC), pp.281-288, 2002

[4] J. Prager, J. Chi-Carroll, K. Czuba, C. Welty, A. Ittycheriah, R. Mahindru, "IBM's PIQUANT", Proceedings of the 11th Text Retrieval Conference (TREC), 2003

[5] W. Salloum, "A Question Answering System based on Conceptual Graph Formalism", Proceedings of the 2nd International Symposium on Knowledge Acquisition and Modeling (KAM 2009), pp. 383-386, 2009

[6] R. Florian., "Named Entity Recognition as a House of Cards: Classifier Stacking", Proceedings of CoNLL2002, pp.175 - 178, 2002

[7] B. Katz. "From sentence processing to information access on the World Wide Web", Proceedings of AAAI Spring Symposium on Natural Language Processing for the World Wide Web, pp.77-94, 1997

[8] C. Kwok, O. Etzioni, and D. S. Weld, "Scaling Question Answering to the Web", World Wide Web journal, Vol.10, pp.150-161, Hong Kong, 2001

[9] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C.-Y. Lin. "Question Answering in Webclopedia", Proceedings of the 9th Text Retrieval Conference (TREC), 2001

[10] J. Kupiec. "MURAX: A robust linguistic approach for question answering using an online encyclopedia", Proceedings of the 16th ACM SIGIR, pp.181-190, 1993

[11] W. Salloum, "A Question Answering System based on Conceptual Graph Formalism", International Symposium on Knowledge Acquisition and Modeling (KAM 2009), pp.383-386, 2009

김한준

1994년: 서울대학교 계산통계학과 (공학사)
1996년: 서울대학교 전산과학과 (공학석사)
2002년: 서울대학교 컴퓨터공학부 (공학박사)
2002년~현재: 서울시립대학교 전자전기컴퓨터공학부 부교수
관심분야 : 정보검색, 텍스트마이닝, 데이터베이스, 기계학습, e-비즈니스 기술

김민경

2005년: 목원대학교 컴퓨터공학과 (공학사)
2009년 : 서울시립대학교 전자전기컴퓨터공학부 대학원 (공학석사)
2009년~현재: 아시아투데이 인터넷부 과장
관심분야: 정보검색, 데이터마이닝

장재영

1992년: 서울대학교 계산통계학과 (이학사)
1994년: 서울대학교 계산통계학과 (이학석사)
1999년: 서울대학교 계산통계학과 (이학박사)
2000년~현재: 한성대학교 컴퓨터공학과 부교수
관심분야 : 데이터베이스, 정보검색, 데이터마이닝