3단계 정답 추출 방법을 이용한 백과사전 인물분야 질의응답 시스템 구현 : AnyQuestion1.0

김현진,오효정,왕지현,이충희,장명길
한국전자통신연구원 음성/언어정보연구부 지식마이닝연구팀
{jini, ohj, jhwhang, forever, mgjang}@etri.re.kr

The 3-step Answer Processing Method for Encyclopedia Question-Answering System : AnyQuestion1.0

Hyeon-Jin Kim, Hyo-Jung Oh, Ji-Hyun Wang, Chung-Hee Lee, Myung-Gil Jang
Knowledge Mining Research Team
Electronics and Telecommunications Research Institute (ETRI)

요 약

본 논문은 3단계 정답 추출 방법을 통해 백과사전 인물분야 질의응답 시스템을 구현하는 방법을 제안한다. 논문에서 제안한 3단계 정답 추출 방법은 1) 백과사전 문서 내에서 정형화 될 수 있는 지식들을 추출한 백과사전 KB 기반 정답 추출 방법, 2) 문장을 언어분석 하여 LF(Logical Form)구조를 추출하여 색인한 LF 기반 정답 추출 방법, 3) 각 문장을 주제 태깅을 하여, 주제별로 묶어 의미적 단락으로 구분하고, 단락 검색을 기반으로 정답을 추정하는 의미적 단락 기반 정답 추출 방법으로 구성되어 있다. 이러한 방법론은 백과사전이라는 문서 도메인의 특성을 반영하고, 사용자 질문의 난이도 또는 형태에 따라서 정답을 제공할 수 있는 백과사전 인물분야 질의응답 시스템에 적합하다.

1. 서 론

기존의 정보검색은대량의 문서에서 사용자의 질의에 적합한 문서를 찾아내는 문서검색에 초점을 맞추어 왔다. 그러나 인터넷의 발달과 더불어 검색 대상 문서의 양이 많아짐에 따라, 검색의 결과로 나타나는 문서의 양은 검색 시스템 사용자에게 큰 부담이 되었다. 따라서 정보검색의 기술적 한계를 직면한 많은 검색 포털 업체들이 "지식검색"이라는 새로운 개념의 검색 서비스를 시작했는데, 지식검색은 사용자의 질문에 다른 사용자가 답을 달고, 많은 사용자들이이 지식의 유용성에 대해 평가함으로써, 네티즌들이 지식을 공유하는 수단으로 크게 각광을 받고 있다[1]. 즉, 많은 사용자들이검색 서비스가 대량의 문서를 찾아주기 보다는 자신이 원하는 질문에 대한 답변을 제시해 주기를 바란다는 것을 증명하고 있다. 이러한 요구를 만족시키기위해서, 질의응답기술이 많이 연구되어지고 있다. 본 논문에서는 ETRI 지식마이닝연구팀에서 2003년부터 개발 중인 백과사전 대상의 질의응답 시스템인 AnyQuestion1.0 (이하 AnyQuestion 1.0) (http : //anyQ.etri.re.kr)을 소개하고자 한다.

AnyQuestion 1.0은 ㈜두산이 공동연구로 참여하여 백과사전 컨텐츠를 제공하고 있으며, 백과사전 인물분야 질의응답 시스템은 인물분야 약 25,000 표제어에 관한 사용자 질문에 대해 질문의도를 파악하여 단답형의 정답을 제시한다. AnyQuestion1.0은 정답이 본문의 문장을 통해서 추출되었을 경우에는 해당 문장을 함께 제시하여, 사용자가 제시된 정답의 정확성 유무를 확인할 수 있게 하며, 현재 질문과 관련된 추가 질문들을 추천하는 기능 및 백과사전 인물에 대한 주요한 정보를 간략히 제시하는 기능 등을 제공하고 있다. AnyQuestion1.0에서는 사용자의 질문의 난이도 또는 형태에 따라서 3단계의 정답 제시 알고리즘을통해 정답 또는 정답을 포함한 문장을 제시하고 있다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 질의응답 시스템에 관한 국내외 연구를 살펴보고, 3장에서는 AnyQeustion1.0의 시스템 구조와 모듈별 기능에 대해서 간단히 살펴본다. 4장과 5장에서는 논문에서 제안하는 3단계 정답 색인 방법론과 정답 제시 방법론을 설

명하고, 6장에서는 시스템의 성능을 평가하고 분석하며, 7장에서 결론을 맺는다.

2. 관련 연구

국외(Trec 등)에서 연구되고 있는 QA기술은 크게 두 그룹으로 나뉠 수 있는데, IE(Information Extraction)기법을 이용한 방법론과 기존 정보검색 엔진을 응용한 단락 검색 시스템을 사용하는 방법론이다. [2]는 IE 기법을 이용한 대표적인 QA 시스템으로 각 entity(예 : person entity)들에 대해서 일종의 template(예 : name, birth_time, what, when 등)을 정의하고, 정보추출을 통해서 각 template 값을 채우는 방식을 제안하고 있다. 이러한 IE 기법을 이용한 QA 시스템은 정답 제시에 속도가 빠르고, 신문 기사, 백과사전 등 특정 제한된 도메인에서는 효과적이라는 평가를 받고 있으나[3][4], entity 또는template 정의에 있어서 일부 수작업의 노력이 필요하기 때문에, 확장성 면에서 문제가 제기되고 있다. 따라서 많은 국외 QA 시스템에서 따르는 방법론은 기존 정보검색 시스템에서 검색 단위를 문서가 아닌 단락 또는 문장 단위로 하여, 정답 후보가 포함된 일부를 검색 한 후에, 실시간으로 언어분석 등을 통해서 정답으로 추정되는 단어 또는 어구를 추출하는 단락 검색 응용 방법론을 채택하고 있다. [5]는 이러한 방법론의 대표적인 논문으로 키워드 검색 방법론을 이용하여 문서에서 주요한 단락을 검색한 후, Lexico-syntactic information 또는 NLP 기술을 응용하여 해당 단락에서 비교적 정확한 정답을 제시해 주고 있다. 그러나 이러한 방법론은 사용자 질문이 입력된 후 실시간으로 언어분석을 통해서 문장들을 분석해 내기 때문에 응답 시간이 매우 길다는 단점을 가지고 있다.

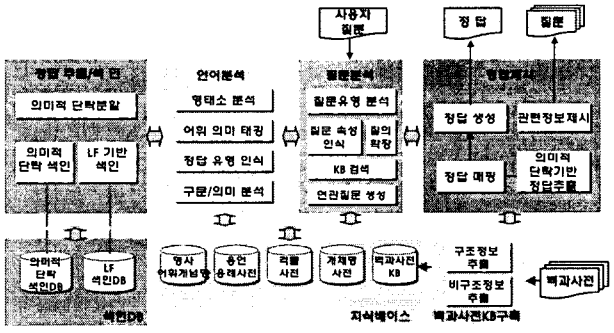
국내의 연구 내용으로는 [6]는 한국어 질의응답을 위한 시스템으로, 질의해석을 통해 날짜, 범위, 핵심어, 중요어, 질의유형을 파악하고, 단락검색에서 질문해석 과정에서 얻어진 정보를 이용하여 문서에서 대답을 포함할 것 같은 단락을 검색하며, 대답추출에서 여러 언어자원을 이용하여 개체명을 추출한 후, 비사실 문맥에서 나타나는 대답제거, 대답확인 및 대답해석을 통해 정답을 찾아낸다. [7]는 MAYA라는 질의응답 시스템을 개발하였는데, 응답 속도를 개선하는데 초점을 맞춘 시스템으로 개체명 사전과 LSP(Lexico-Syntactic Pattern)을 이용하여 개체명을 인식하고, 이를 질의응답 시스템이 정답 가능한 후보로 미리 색인하였다. 사용자의 질의유형을 105가지의 의미범주로 구분하고, 이에 따라 정답유형을 분류하였으며, Lexico-Syntactic Parser를 이용하여 사용자의 질의유형을 분석하여 색인된 정답 DB에서 정답후보를 순위화하고 이를 정답으로 제시하였다.

본 논문에서는 특정 도메인(백과사전 분야)을 대상으로 질의응답시스템을 구현하므로, 응답 속도 등을 고려하여 IE 기법을 응용한 방법론과 함께, 단락 검색을 개선한 의미적 단락 검색을 응용한 방법론을 단계적으로 적용하는 3단계 정답 추출 방법론을 제시하고자 한다.

3. 시스템 구조

AnyQuestion1.0의 전체 흐름은 크게, 1) 백과사전 문서를 입력으로 해서 사용자의 질문을 예측하여 미리 정답이 될 만한 부분을 색인하고, 정답을 포함한 문장들을 효과적으로 저장하는 부분과 2) 사용자 질문을 실시간으로 입력 받아서, 질문자의 의도를 파악하고, 색인된 데이터베이스를 중심으로 정답으로 추정되는 부분을 추출해서 제시하는 부분으로 나뉠 수 있다.

시스템의 전체 구성도는 [그림1]과 같다.



[그림 1] 시스템 구성도

각 모듈별로 간단한 기능을 설명해 보면 다음과 같다.

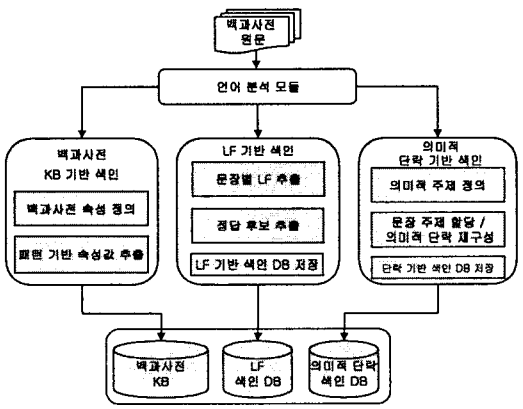
- 질문분석 모듈 : 사용자의 질문을 입력으로 해서 사용자의 의도를 파악하고, 질문 문장을 정답을 추출하기 위한 단위로 분석하는 모듈. 질문의 유형(5W1H 유형)과 사용자가 얻고자 하는 정답의 유형(예 : 인명, 지명, 날짜 등 70 여가지)을 인식하여 제공
- 언어분석 모듈 : 질문분석 모듈과 정답색인 모듈 등에서 필요한 형태소 분석, 어휘 의미 태깅, 구문 분석, 정답유형 인식(확장된 개체명 인식기의 일종) 등의 언어처리 기능을 수행하는 모듈
- 정답색인 모듈 : 백과사전 문서를 입력으로, 3가지 방법론의 정답 제시를 위해 각각의 정답색인을 하는 모듈. 1) 백과사전 문서 내에서 정형화 될 수 있는 지식들을 반자동으로 추출하여 백과사전 지식베이스에 저장하고, 2) 특정 용언이 포함된 문장을 언어분석하여 LF(Logical Form)구조를 추출한 후 DB에 저장

하고, 3) 각 문장을 주제 태깅을 하여, 주제별로 묶어 의미적 단락으로 구분하고, 각 단락별로 주요한 색인어를 추출하며, 정답으로 활용 될 수 있는 부분을 색인하는 모듈로 구성

- 정답제시 모듈 : 질의분석 모듈에서 처리된 질문을 입력으로, 정답색인 모듈에서 저장된 색인 DB(3가지 모델)에서 정답으로 추정되는 부분을 추출하고 제시하는 모듈. AnyQuestion1.0에서는 3단계의 정답제시 방법을 이용하는데, 1) 백과사전 KB 기반 정답 제시, 2) LF 기반 정답 제시, 3) 의미적 단락기반 정답 제시 방법론을 단계적으로 적용하여 사용자에게 적합한 정답을 제공
- 지식베이스 모듈 : 타 모듈들에서 공통적으로 또는 특수하게 사용하는 전자사전, 어휘 개념망, 격틀 사전 등의 정보를 관리하는 모듈. 이 중 명사 어휘 개념망은 상하위 관계가 트리 형태로 구성되며, 유의, 동의, 반의 부분전체 관계 등의 정보가 포함되어 노드수는 약 5만여 일반 어휘이고, 개체명 단위의 어휘를 포함한 백과사전 어휘를 합산하면 29만 어휘로 구성

4. 정답 색인

본 논문에서는 앞에서 언급했듯이 3단계 정답 제시를 위해 각각 3가지 방식의 정답색인 방법론을 제안하고 있다. 첫번째는 백과사전 원문에서 정형화 될 수 있는 정보들을 정보추출 방법을 이용하여 반자동으로 추출하는 방식인 백과사전 KB 기반 색인 방법론과 두번째는 725개의 용언을 포함한 문장의 구문분석 결과를 이용하여 LF(Logical Form)을 생성하여 입력문장에서 who, when, where, what, why, with 등의 단위를 추출, 색인하는 LF기반 색인 방법론, 마지막으로는 일반적으로 QA 방법론에서 많이 응용하는 단락기반 색인을 개선한 의미적 단락기반 색인 방법론이다. [그림 2]는 정답 색인 방법론의 구조도이다.



[그림 2] 정답 색인 구조도

각각의 방법론을 설명하면 다음과 같다.

4.1. KB 기반 정답 색인

KB기반 정답 색인 방법론은 백과사전의 본문에는 해당 표제어와 직접 관련이 있는 정보들이 산재해 있는 점에 착안하여, 이러한 표제어 관련 정보를 정형화하여 지식베이스로 구축한 후 색인 데이터 또는 표제어의 요약 정보로 제공하는데 초점을 맞추고 있다. 본 논문에서는 이러한 표제어 관련 정보를 각 표제어의 속성이라고 정의하였는데, 표제어가 해당되어 있는 그룹에 따라서 공통 속성과 개별 속성이 존재하게 된다. 공통 속성이란 표제어가 해당 되는 그룹에서 공통적으로 발견되는 속성으로 예를 들어, 인물 분야에서는 출생(출생일, 출생지 등)이나 사망(사망원인, 사망일 등) 등과 같이 인물이면 공통적으로 발생되는 주요한 정보를 말하고, 개별 속성은 해당 인물 분야에서도 세부적으로 정치인(당선일, 정당명 등) , 발명가(발명품 등), 군인(부대명, 군계급 등) 등의 세부 그룹별로 추가적으로 발견되는 개별적인 정보 중에서 빈번히 나타나는 정보들을 위주로 정의되어 있다.

AnyQuestion1.0에서는 인물분야에 대해 인물의 공통 속성(21개)과 개별 속성(31개)을 먼저 정의하고, 각 속성별로 속성값을 결정할 수 있는 속성 패턴(일종의 FSA)을 이용해서 정보를 추출하게 된다. 다음은 인물에서 정의된 공통 속성과 개별 속성의 일부를 나타낸 표이다.

[표 1] 인물 분야의 공통 속성과 개별 속성

공통 속성	개별 속성
출생일, 출생장소, 사망일, 사망장소, 사망원인, 별칭, 졸업학교 , 저서명, 주요 업적 등 21개	건축물, 연구분야, 개발품,개발일, 발견물, 발견일, 대회 성적, 대회일, 군계급, 주요 평가 등 31개

이렇게 정의된 속성태그를 가지고 백과사전 본문을 대상으로 속성 태깅 데이터를 구축하게 된다. 속성 태깅 데이터는 총 1,000개의 문서(표제어)에 대해서 수작업으로 이뤄지고, 태깅 데이터를 기반으로 각 표제어의 속성별 패턴 FSA가 생성되며, 속성값 추출기를 통해서 각 속성에 대한 속성값들이 매핑되어 지식베이스를 구축하게 된다. 다음은 문장에서 속성 "출생장소"를 추출하는데 필요한 속성 패턴과 추출된 속성값의 예이다.

[속성 패턴 예1]
출생일 : = "<person> "[" @<date> "~" <date> "]"
사망일 : = "<person> "[" <date> "~" @<date> "]"

본문 예) *박정희[1917.11.14~1979.10.26]*
분석 예)
박정희(@person) [1917.11.14(@date)~1979.10.26(@date)]
출생일 : "1917.11.14"
사망일 : "1979.10.26"

[속성 패턴 예2]
출생장소 : = @<Location>에서 태어나다
출생장소 : = @<Location> 출생.
출생장소 : = <person>은 @<Location>에서 출생하…
출생장소 : = <person>의 출생지는 @〈Location〉…
출생장소 : = <person>의 고향은 @〈Location〉…
출생장소 : = <person>이 출생한 곳은 @〈Location〉…

4.2. LF 기반 정답 색인

앞 절에서 설명한 KB 기반 정답색인 방법론에서는 몇 가지 문제점을 가지고 있는데, 먼저 각 본문에서 속성을 수작업으로 정의하여야 하고, 정의된 속성 이외의 정보는 추가로 추출하기 힘들다는 것이 대표적이다. 즉 도메인 이식성 면에서 한계점을 가지고 있다. 또한 패턴방식으로 속성값을 추출함에 따라, 패턴의 한계로 적절한 속성값을 추출함에 있어서 부족한 부분이 드러난다. 따라서 이 절에서는 도메인이 달라지더라도 어느 정도 문장이 가지는 특성만으로 쉽게 정형화된 데이터를 추출 할 수 있는 방법으로 LF기반 정답색인 방법을 제안한다. LF기반 정답 색인은 백과사전 인물분야 문장에서 주요한 역할을 하는 용언들을 중심으로 LF(Logical Form)을 추출하고, 각 용언별 LF에서 나타나는 용례를 기반으로 who(누가), when(언제), where(어디서), what(무엇을), whom(누구를), with(누구와 함께), why(왜) 등의 정보로 매핑하여 색인하는 방법이다 AnyQuestion1.0에서는 백과사전 텍스트와 일반 코퍼스의 용언을 대상으로 빈도와 중요도를 계산하여 총 725개의 용언을 선정하였다. 선정된 용언의 예를 보면 다음과 같다.

가르치다, 개발하다, 개최하다, 결혼하다, 귀국하다, 낳다, 멸망하다, 발견하다, 발명하다, 사망하다, 살해하다, 설립하다, 암살하다, 졸업하다, 죽이다, 합격하다 등

이러한 용언을 포함한 문장을 입력으로, 용언 격틀을 이용하여 생성한 구문트리에서 각 용언과 구조적으로 연결된 노드들을 의미구조로 생성하는데 이를LF(Logical Form)로 정의하였다. 다음은 예문에 대해서 LF를 생성한 결과이다.

[예문] 박정희는 1961년 516군사정변을 주도하였다.
[LF 분석 결과]
주도하(〈subj : 박정희〈PERSON〉가〉 〈obj : 5 · 16군사정변〈EVENT〉를 〉 〈adverb : 1961년〈DATE〉)

용언의 LF결과 용례와 기본 격틀을 입력으로 해서, 각 용언별로 정답 색인 규칙을 자동으로 생성하게 되는데, 다음은 용언 "주도하다"의 경우 생성된 색인 규칙의 예이다. AT_PERSON, SEM_사람 등의 코드는 각 정답유형 인식기와 어휘 의미태깅기에서 부여한 의미코드들의 일종이다. 한 예로, 용언 "주도하다"에서 adv(부사어) 위치에 SEM_지리(지리를 나타내는 명사류)나 AT_LOCATION(지리를 나타내는 개체명)가 나타나면 "where"로 매핑하여 색인한다는 의미를 나타내고 있다.

verb 주도하
who subj &가 &AT_PERSON | &가 &SEM_사람
what obj &를
where adv &에서 &SEM_지리 |
&에서 &AT_LOCATION
when adv &AT_DATE

이러한 색인 규칙에 따라서 정답 색인 DB를 구축하면 다음과 같다.

verb 주도하다
who 박정희
what 5 · 16군사정변
when 1961년

4.3. 의미적 단락 기반 색인

단락 기반 색인 방법론은 일반적으로 QA 시스템에서 색인 기법으로 많이 사용하는 것인데, 보통은 특정 길이 또는 일정 범위(윈도우)의 문장들을 단락으로 구성하여 키워드 색인을 하게 된다. 본 논문에서는 각 문장이 개별적인 주제 또는 정보를 가진 경우가 대부분인 백과사전과 같은 특징적인 문서를 효과적으로 활용할 수 있는 방법으로 문장 주제 할당 기법을 이용한 의미적 단락 분할 및 색인 방법론을 제안한다.

의미적 단락을 생성하기 위해 먼저, 백과사전의 도메인 특성을 반영하는 문장 주제를 정의하였는데, AnyQuestion1.0에서는 인물분야에 대해서 8개의 대분류와 35개의 소분류 주제를 가진 계층적 문장 주제를 정의하였다. [표 2]는 대분류와 소분류 주제의 일부분

을 나타낸 것이다.

[표 2] 문장 주제표 예

대분류	중분류	소분류
출생	국적	
사망		
활동	업적	설립, 연구, 발견 등
	교육	졸업, 유학, 수학 등

이렇게 정의된 문장 주제를 기반으로 문장주제 할당기를 통해서 각 문장마다 주제를 부여하게 된다. 다음 예는 표제어 "박정희"의 일부 문장에 문장주제 할당기를 통해 문장 주제를 할당한 결과를 보여주고, 이를 통해 주제별 단락을 재구성한 결과를 보여준다.

[원문 주제 할당 예]
경북 선산(善山) 출생[출생]. 가난한 농부인 박성빈(朴成彬)과 백남의(白南義) 사이에서 5남 2녀 중 막내로 태어났다[출생]. 1937년 대구사범학교를 졸업하고[졸업], 3년간 초등학교 교사로 근무하다가[역임], 만주의 신경(新京 : 現 長春)군관학교를 거쳐 1944년 일본육군사관학교를 졸업하였으며[졸업], 8 ·15광복 이전까지 주로 관동군 배속되어 중위로 복무하였다[역임].

[주제별 단락 재구성 예]
[출생] 경북 선산(善山) 출생 가난한 농부인 박성빈(朴成彬)과 백남의(白南義) 사이에서 5남 2녀 중 막내로 태어났다.
[졸업] 1937년 대구사범학교를 졸업하고, 3년간 초등학교 교사로 근무하다가, 만주의 신경(新京 : 現 長春)군관학교를 거쳐 1944년 일본육군사관학교를 졸업하였으며…

주제별 단락으로 재구성된 의미적 단락을 기존의 키워드 방식을 개선한 방법으로 색인을 한다. 명사 키워드 위주의 색인 단위를 확장하여, 명사구(속격 어구 포함) 단위 및 용언 논항 단위의 색인어를 추출하여 저장하게 된다.

5. 정답 제시

정답제시 모듈은 사용자의 질문이 들어오면, 질문분석 모듈을 거쳐서 나온 결과를 입력으로 정답색인 모듈에서 저장한 색인 DB에서 해당 질문에 맞는 정답을 추출하고 제시하는 기능을 수행한다. 정답제시 모듈에서는 크게 다음의 3단계의 정답추출 모듈을 단계별로 거치면서 질문에 대한 정답을 추출한다.

- [1단계]KB 기반 정답 추출 : 백과사전 KB에 저장되어 있는 정답으로 매핑 되는 경우 정답 추출
- [2단계]LF 기반 정답 추출 : 질문 문장이 용언을 기반으로 구조화 될 수 있는 경우, LF 기반 정답 색인 DB를 검색해서 매핑 되는 경우 정답 추출
- [3단계]단락검색 기반 정답 추출 : 앞의 두 단계에서 정답을 추출하지 못하는 경우에는, 단락검색 방식을 이용하여 실시간으로 의미적인 단락을 검색하여, 정답을 추출

5.1. 1단계 : KB 기반 정답 추출

KB 기반 정답추출은 백과사전 KB에 미리 구축된 지식을 검색하는 방법으로, 질문에서 매핑에 필요한 속성정보를 추출한 뒤 KB 매핑 과정을 거쳐 정답으로 제시한다. 즉, 질문에 해당하는 질문 튜플을 추출하여 KBQ를 생성하게 된다. 이때 질문 튜플의 스키마는 아래의 예처럼 구성된다.

[질문 튜플 스키마]
{표제어, 속성명, @정답유형}
{?표제어, 속성명, 속성값}

"표제어"는 KB 기반 정답 추출의 경우에는 해당 정보가 모두 표제어와 직접 관련이 있는 정보를 추출하게 되므로, 질문이 표제어의 속성을 묻는 질문일 수도 있고, "표제어"의 속성값을 물어보는 경우로 나눠질 수 있기 때문에 필요하다. 두번째 "속성명"은 앞서 색인파트에서 정의한 52개의 속성 태그명을 KB 메타데이타 정보를 활용하여 선정하게 된다. 마지막 "정답유형" 또는 "속성값"은 해당 튜플을 가지고 검색하였을 때 최종답이 될 수 있는 유형을 제약하는 정보이다. 다음은 질문 튜플과 KB 메타데이타 정보를 활용하여 질문 튜플을 생성하는 과정을 보여준다.

[질문1] 박정희는 어디에서 태어났나?
[질문 튜플] {박정희, 출생장소, @LOC}
[질문2] 1979년에 암살된 대통령은?
[질문 튜플]
{?Person, 사망일, 1979년} & {?Person, 직위, 대통령}

생성된 질문 튜플을 이용해서 KBQ를 생성한 후 KB 매핑기를 거쳐 정답을 추출하게 된다. 만약 해당 질문에서 질문 속성 튜플을 생성하기에 적합하지 않은 경우에는 다음 단계인 LF 기반 정답 추출 모듈을 실행하게 된다.

5.2. 2단계 : LF 기반 정답 추출

LF 기반 정답 추출은 입력된 질문에 대해 언어분석을 하여, 질문 LF 구조를 생성하고 이를 4.2절에서 언

급했던 LF 기반 정답 색인 DB를 검색하여 정답을 제시하는 방법이다. 질문 LF 구조는 색인과정에서 생성했던 문장의 LF 구조와 비슷한데, 해당 질문의 LF 구조에서 검색의 대상이 되는 정답 추출 대상(@Search_Pont)을 표기한 점이 차이점이 된다. 다음은 질문을 질문 LF 구조로 변경한 예이다.

```
[질문1 ] 516 군사정변을 주동한 사람은?
[질문 LF] 주동하{subj : 사람〈@search_point〉, obj :
          5·16군사정변 〈EVENT〉를 〉 }
```

위 예문의 질문 LF에서 "사람은?"이 질문의 특성이 고려되어서 subj로 분석이 되고, LF 색인 DB에서 찾아야 하는 대상이 되므로 @search_point로 선택되게 된다. 정답 추출 대상으로 선정된 위치와 질문분석에서 분석한 질문 유형을 고려하여, LF 색인 DB의 어떤 슬롯과 매핑해야 하는지 결정하게 된다. 다음은 LF 색인 DB의 매핑을 위한 슬롯 결정을 하는데 도움이 되는 규칙의 일부이다.

```
switch (질문 LF Structure[i])
  case : subject → who
  case : object → verb type ? whom : what
  case : adverb
    switch(adverb[j])
      case (AT type=TIME/DATE/extra time
      expression) → when : how-long
      case (AT type = Location/Organization/extra
          location expression)  where
                    :
```

위 규칙에 따라서 질문1의 경우에는 용언 "주동하다"에 대해서 what 슬롯에는 {5·16군사정변}이 채워지고, who 슬롯을 정답으로 찾게 되는 매핑구조가 생성되는데, 이때 "주동하다"는 유사어 확장에 따라서 "주도하다""선동하다" 등의 용언으로 확장하게 되고, 이에 따라 4.2절의 LF 색인 DB를 참조하면 "박정희"라는 정답을 제시할 수 있게 된다. LF 기반 정답 추출 방법론의 장점으로는 문장 분석이 가능하므로 피동/사동형의 문장의 정답 추출도 가능한데, 다음은 "암살당하다" 용언의 확장에 따른 정답 추정의 예를 보여주고 있다.

```
[예문] 안중근이 이토 히로부미를 사살하였다.
[LF 색인 DB]
  verb : 사살하다
  who : 안중근
  what : 이토 히로부미
```

```
[질문2] 안중근에게 암살당한 사람은?
[질문 LF]
  암살당하(subj : 사람〈@search_point〉,  adv : 안중근
  에게)
  verb : 암살당하다
  who : @search_point
  whom : 안중근
[질문 LF 격 전이 확장]
  verb : 암살하다
  who : 안중근
  what : @search_point
```

즉, 질문에서의 "암살당하다"에서의 who가 "암살하다"의 동사로 변경되면서 what으로 격전이가 일어나게 된다. 따라서 이러한 피동/사동 규칙을 이용하고, 용언 "암살하다"는 "사살하다"와 용언 유사어 확장에 의해서 정답인 "이토 히로부미"를 추정하게 되는 것을 알 수 있다.

5.3. 3단계 : 의미적 단락 기반 정답 추출

2단계인 LF 기반 정답 추출 방법론은 일반적인 문장에서 문장 분석으로 정답을 추정하는 기술이나, 질문이나 또는 색인 대상 문장에 해당 용언이 포함되어 있지 않거나 언어분석이 오분석되는 경우에는 적합한 정답 추정을 하기 힘든 단점을 가지고 있다. 또한 앞서 두 단계는 정형화된 형태로 색인된 DB에서 직접 정답 부분을 매핑하는 방식인데 두 단계에서 정답을 추출하지 못하는 경우를 위해서 본 논문에서는 기존 단락검색 방식을 이용하여 정답을 추정하는 모델인 의미적 단락 기반 정답 추출을 제안한다. 즉, 입력된 질문과 유사한 의미적인 단락을 검색하여, 실시간으로 언어분석을 해서 정답 추정 방식을 이용하여 정답을 추출하게 된다.

질문 분석 모듈에서는 의미적 단락 검색을 위해, 다음의 정보들을 추출하게 되고, 각각의 유사도 측정 규칙에 의해서 단락을 검색하게 된다.

표제어질문에 표제어가 출현한 경우, 특정 표제어의 단락을 중심으로 정답 추정하기 위함	
단락 주제	단락 색인 과정에서 태깅한 주제가 질문에 할당되는 경우 동일한 주제를 가진 단락의 유사도를 높여줌.
정답 유형	질문분석 시 요구되는 정답 유형이 존재하는 단락의 유사도를 높여줌.
문장 정보	같은 문장에 동시에 색인 어휘가 나타나는 경우 단락의 유사도를 높여줌.
키워드 가중치	2포아송 모델을 이용한 색인어 가중치 부여

추출된 후보 단락을 중심으로, 각 문장의 언어분석을 수행하게 되고, 문장에서 정답유형으로 태깅된 정답 대상어들을 중심으로 아래의 자질값을 추정하여 정답 후보를 순위화 하게 된다.

$$Score(A_i) = \sum_i aw_i * af_i$$

$af_1 = Dist,\ af_2 = Count,\ af_3 = Score(s_i)$

aw_i = Answer feature weighting

- Dist : 정답과 질의 word 간의 평균거리
- Count : 정답 후보 문장간의 중복 횟수
- Score(Sj) : 정답 후보가 포함된 문장의 점수.

$$Score(S_i) = \sum_i sw_i * sf_i$$

$sf_1 = LF\_mapping,\ sf_2 = Q\_AT,\ sf_3 = Q\_word,\ sf_4 = P\_score$

sw_i = Sentence feature weight

- LF_mapping : 질문에 나오는 LF 구조와 문장 LF 구조와의 매핑 유사도
- Q_AT : 질의에 포함된 정답 유형 여부
- Q_word : 질의어 키워드 매핑 유사도 값
- P_score : 단락 검색 유사도 값

6. 성능 평가

AnyQuestion1.0을 객관적으로 평가하기 위해서, 평가 질의와 정답쌍으로 구성된 백과사전 인물분야 질의응답 평가셋(총 402개)을 구축하였다. 실제 사용자가 질의할 수 있는 다양한 형식의 질문과 난이도에 따라 정답 추출 능력을 평가할 수 있는 테스트컬렉션을 구축하기 위해서, 시스템 개발자가 아닌 일반인 6명이 25개 인물 세부 범주(예 : 인물 직업별, 시대별, 지역별 그룹 등)에 고루 분포한 질문을 선정하고 문장 내에서 정답을 추천하여 구축하게 하였다. 다음은 평가셋 중 한 예이다.

```
〈Q_ID〉1〈/Q_ID〉
〈Q〉 나이팅게일상을 만든 기관은? 〈/Q〉
〈Q_type〉 What 〈/Q_type〉
〈Answer〉
〈Ans〉 국제적십자〈/Ans〉
〈A_type〉Organization/PublicOrganization 〈/A_type〉
〈/Answer〉
〈Ans〉 국제적십자〈/Ans〉
〈A_type〉Organization/PublicOrganization 〈/A_type〉
〈/Answer〉
〈Passage〉
〈Sent〉 국제적십자에서는 '나이팅게일상(賞)'을 마련하여 매년 세계 각국의 우수한 간호사를 선발, 표창하고 있다. 〈/Sent〉
〈title〉 나이팅게일 〈/title〉
〈/Passage〉
```

질의응답 시스템을 평가하기 위해 사용한 평가지수(measure)는 정답 역순위 평균(MRAR : Mean Reciprocal Answer Rank)과 사용자 정답 만족도를 사용하였다. 정답 역순위 평균이란 사용자가 원하는 정답이 몇 번째에 나타났는가에 대한 순위를 평가하는 방법으로, 순위를 1/n의 가중치로 반영한다. 본 실험에서는 질의응답 평가셋 중 200개의 평가셋으로 상위 5등까지의 순위를 평가하였다. 사용자 정답 만족도란 시스템에서 제공한 정답(상위 5위 제공) 뿐만 아니라 시스템에서 제공하는 부가정보(요약, 문장, 단락 정보 등)를 통해 사용자가 원하는 정답을 찾을 수 있는 지의 여부를 평가한 것이다. 시스템 평가 결과는 [표 3]과 같다.

[표 3] 전체 성능 평가 결과

	1등	2등	3등	4등	5등	사용자 정답 만족도
개수	94	111	114	115	115	149
MRAR	0.47	0.51	0.51	0.51	0.51	0.74

시스템 성능 평가결과는 MRAR 0.51, 사용자 만족도 0.74의 수치를 기록했다. 그리고 본 시스템에서 정답추출을 위해 사용한 3단계 정답제시 방법론에서 정답 추출에 대한 각 단계별 질문 분포 및 방법별 정확도를 보면 다음 [표 4]와 같다.

[표 4] 정답 제시 방법별 질문 분포 분석

정답 제시 방법론	KB 기반 정답 추출	LF 기반 정답 추출	의미적 단락 기반 정답 추출
질문 분포	20%	11%	69%
사용자 만족도	0.85	0.90	0.67

[표 4]를 분석해 보면, KB 및 LF 기반의 정답 추출 방법론은 정확도 수치는 높은 반면, 전체적으로 질문을

처리할 수 있는 능력은 30%정도인데 반해, 질문이 입력된 후에, 단락 검색을 통해 실시간으로 정답을 추정하는 모듈인 의미적 단락기반 정답 추출은 다소 낮은 정확도 수치를 나타내기는 하나, 앞의 두 단계가 처리 못한 70%정도의 질문을 처리한 것으로 분석된다. 따라서, 앞으로의 연구 방향은 빠르고 비교적 정확한 정답 제시 기능을 가진 KB 및 LF 기반의 정답 제시 방법에서의 질문 처리 능력을 향상시키고, 많은 질문을 처리할 수 있는 단락기반 정답 추출에서의 속도와 정확도를 향상 시킬 수 있는 방법론을 연구하고 있다.

7. 결론 및 향후 연구 방향

본 논문에서는 백과사전 인물분야 질의응답 시스템 구현을 위한 3단계 정답 추출 방법론을 제안하였다. 백과사전 KB 기반 정답 추출 방법론은 정보추출 방법을 이용하여 백과사전 본문에서 정형화될 수 있는 정보 추출하는 것으로, 빠르게 정답을 제시하는 방법 뿐 아니라, 구축된 백과사전 지식베이스를 각 표제어의 요약 정보로도 활용할 수 있는 장점이 있다. 그러나 일부의 수작업으로 속성 구축하고 패턴을 보정하는 점에서 도메인 이식성 면에서 한계점을 가지고 있으므로, 향후 수작업을 최소한으로 줄이고, 패턴 방식의 한계점을 극복하기 위해 통계적인 방법론을 적용하는 것으로 현재 연구를 진행 하고 있다. LF 기반 정답 추출 방법론은 특정한 도메인에 관계없이 문장 구조의 특성과 한국어 언어분석 기술을 이용하여 비교적 정확한 정답을 추출할 수 있다는 장점을 가지고 있다. 그러나 이 방법론 역시 언어분석 기술 자체의 한계점과 문장 구조에서 고정된 형식의 정보만을 추출한다는 점에서 근본적인 문제점을 안고 있으므로, 언어분석 기술을 향상시키고 많은 질문을 처리할 수 있는 방법으로 개선할 필요가 있다. 마지막으로 의미적 단락기반 정답 추출은 문장 주제 할당을 이용하여, 각 단락을 의미적으로 재 구성하여 색인하여, 비교적 높은 정확률과 함께 많은 질문에 대해서 정답을 제시하였다는 점에서 장점을 가지고 있으나, 실시간으로 언어분석을 해야만 하는 점은 실용적인 질의응답 서비스를 하는 부분에 있어서 한계점으로 남아 있다. 따라서 정답 추출 속도를 개선하는 데에 노력을 기울여야 한다.

본 연구팀에서는 AnyQuestion1.0 에서의 한계점과 개선점을 파악하여, 백과사전의 인물분야 뿐 아니라 전체 분야 16만 표제어에 대해서 질의응답 서비스를 하는 것을 연구 목표로 하여 현재 각 모듈들을 개발하고 있다. AnyQuestion1.0 에서는 단답형의 정답을 위주로 개발되어 있으나, 향후에는 서술형(정의/원인/방법/종류 등 10가지 타입) 정답과 나열형(한 문장 이상에서 추출된 정답)의 정답까지도 처리 할 수 있는 실용적인 한국어 백과사전 질의응답 시스템을 개발할 예정이다.

8. 참고 문헌

[1] 황이규, 김현진, 장명길, "질의응답 기술개발", 정보처리학회지, 제11권 제2호, 2004.

[2] Wei Li, Rohini K. Srihari, "Extracting Extract Answers to Questions Based Structural Links", Coling-2002

[3] Ellen M. Voorhees, "The TREC-8 Question Answering Track Report"

[4] Julian Kupiec, "MURAX : A Robust Linguistic Approach for Question Answering Using On-line Encyclopedia, SIGIR 93

[5] Sanda M. Harabagiu, Steven J. Maiorano, "Finding Answers in Large Collections of Texts : Paragraph Indexing+Abductive Inference", AAAI-1999.

[6] 김학수, 서정연, "2-패스 색인 기법과 규칙 기반 질의처리 기법을 이용한 고속, 고성능 질의 응답 시스템", 정보과학회논문지 : 소프트웨어 및 응용, 제29권 제11호, pp. 795-802, 2002

[7] 이경순, 김재호, 최기선, "KorQuA : 질의응답에서 자료유형을 고려한 대답검색과대답해석", 한글 및 한국어 정보처리학술대회, 2000.