지식기반 (Knowledge-based) 질의응답시스템: 사실 자료 (Fact Database)구축을 중심으로

A Knowledge-based Question-Answering System: With A View To Constructing A Fact Database

신 효 필*
(Hyopil Shin)

요 약 본 논문에서는 질의어 응답시스템에 있어 핵심이 되는 사실 자료 (Fact Database) 구축의 관점에서 지식기반 방법의 중요성과 그 과정에 대해서 논의한다. 지식기반 질의어 시스템은 기존의 이용가능한 자연언어처리의 자원 - 형태소, 구문, 의미분석 등 - 과 온톨로지라는 개념구조망을 이용하는 시스템으로 이 개념을 현실세계의 사실 자료와 연결시켜 개념구조가 지닌 속성과 값의 확장을 통해 그 가능한 응답을 유도해 내는 시스템이다. 이 시스템 구축에 있어 실제 세계의 자료를 수집하고 가공하고 개념화하는 과정은 이 시스템의 성패를 좌우하는 핵심작업으로 아직은 완전히 자동화되기 어렵다. 그러나 지식기반에 기초한 방법은 응용시스템의 질적 향상이라는 측면에서 진지하게 논의될 필요가 있다. 이 글에서는 사실 자료 구축의 관점에서 이런 작업들이 어떻게 행해져야 하는지 그리고 그 방법론이 지닌 특징 및 문제점에 대해 논의한다.

키워드 자연언어처리, 지식기반 질의응답 시스템, 사실 자료, 온톨로지

Abstract In this paper, I describe a knowledge-based question-answering system and significance of the system with a view to constructing a fact database. The knowledge-based system takes advantage of existing NLP-resources such as conceptual structures of ontologies along with morphological, syntactic and semantic analysis. The use of conceptual structures allows us to select right answers through inferences basically made by expansions of concepts. However, the work of constructing factual knowledge requires a great amount of acquisition time in large-scale applications because of the nature of human interference. This is why the procedure of acquiring factual knowledge cannot be fully automated. Apart from efficiency considerations, the knowledge-based system deserves serious consideration. I point out benefits of the system and describe the whole procedure of building the system in terms of a fact database.

1. 서론

지식기초 (knowledge base)로서의 어휘의미론은 현재 자연언어처리에서 중요한 역할을 하고 있다. 지식표현(knowledge representation)에 기초한 다양한 응용시스템이 개발되고 있는데 이런 시스템의 대부분은 노드 (node)와 링크 (link)로 개념들 간의 관계가 네트웍으로 표현되고 있다. [1] 일례로 네트웍과 같은 개념구조를 사용하는 기계번역 시스템은 원시언어의 문장을 언어 중립적인 중간언어 (interlingua)로 나타낸다. 이 중간언어로 사용되는 것이 소위 온톨로지 (Ontology)이다. 온톨로지는 세상에 어떤 "개념(concepts) 구조"가 존재하는지를 나타내고 이것들이 어떻게 연결되는지를 규정하는 전산적 대상(computational entity)이라 할

* 서울대학교 전기공학부
연구세부분야 : 자연언어처리
Tel : 02-880-1765 Fax : 02-882-4657
본 연구는 BK21 서울대/고려대 정보기술사업단의 연구지원을 받았으며, 미국 New Mexico State University의 CRL (Computing Research Laboratory)의 연구를 기반으로 하고 있다.

수 있다. [2] 이런 개념구조인 지식기초 (knowledge base)를 적극 이용하는 응용 시스템들이 현재 기계번역, 정보검색 등과 관련하여 많이 등장하고 있다. [3]

한편 인터넷과 웹의 등장은 엄청난 양의 정보를 손쉽게 얻을 수 있는 기회를 제공하고 있다. 이러한 정보는 지식기초를 위한 데이터베이스 구축에 필요한 자원이 되기도 하지만 그 자체로는 가공되지 않은 자료이기 때문에 그대로 데이터베이스화하기는 어렵다. 이런 자료에서 필요한 정보를 도출해내고 일정한 형태로 가공할 필요가 있는데, 온톨로지를 바탕으로 하는 시스템에서 구현화된 (instantiate) 자료구조인 사실 (fact)이 그 한 형태일 수 있다. 즉 개념구조를 현실세계에 실재하는 대상으로 구현시켜 사실이라는 틀을 만들고 이를 온톨로지의 개념구조와 마찬가지로 일정한 속성과 값으로 표현하여 온톨로지의 계층적인 개념구조에서 나타나는 제약과 연관을 사용할 수 있도록 하려는 것이 지식기반에서 추구하는 사실 자료 (Fact Database)이다. 이 사실자료는 여러 자연언어처리시스템에 기초자료로 사용되며, 질의응답시스템을 포함하는 어떤 종류의 지식기반시스템 (knowledge-based system) 에서는 특정 범위 (domain)의 지식의 체계화된 자료구조라 할 수 있다. [4]

이 글에서는 사실 자료를 중심으로 지식에 기초한 질의응답시스템 (question-answering system)에 관하여 논의하도록 한다. 여기서 추구하는 질의응답시스템은 단순히 키워드만을 입력으로 받아 처리하는 시스템이 아닌 자연언어의 문장을 이해하고 이를 바탕으로 미리 구축해 놓은 데이터 베이스에서 처리된 단어들과 부합되는 개념들을 추론하여 필요한 응답을 이끌어 내는 종합적이고 지능적인 시스템이다.

현재 개발되고 있는 지식기반 방법의 대표적인 것의 하나는 미국 MIT 대학의 AI lab에서 개발되고 있는 START (SynTactic Analysis using Reversible Transformations)이다. 이 시스템도 자연언어를 인지하여 지식기초를 구성하고 웹과 연동되어 그 가능한 답을 제한된 범위 내에서 도출해 낸다. 그러나 이 시스템에서는 실제로 완전히 자연언어처리를 행하고 있지는 않다. 대신 문장을 분석함에 있어 기본적인 작은 단위로 분할하여 그 핵심이 되는 구조들로부터 키워드를 도출하여 질의어를 처리하고 있다. [5]

이 논문에서는 미국 뉴멕시코주립대학 CRL (Computing Research Laboratory)에서 개발되고 있는 질의응답시스템을 기초로 한다. 이 시스템은 MIT의 START에서 한 걸음 더 나아간 기존의 자연언어처리의 모든 자원을 적극적으로 이용하는 그리고 온톨로지라는 개념구조를 이용하는 방법을 근간으로 한다. 여기서는 전체 시스템의 구조와 핵심이 되는 사실 자료의 구축과 이를 바탕으로 응답을 도출해 내는 추론 과정에 초점을 맞추도록 한다.

2. 지식기반 접근방법

어떤 종류의 정보라도 사실 자료를 구축하는 데는 유용하지만 그대로 적용되기는 어렵다. 지식기반 시스템에서 사실 자료는 추출된 정보를 적절한 온톨로지의 개념들의 인스탄스 (instance)로 구현하여 연결시켜 구축된다는 점에서 차이가 있다. 여기서 개념 (concept)과 인스탄스는 구별되는 것으로 정의된다. 개념이란 실제의 대상을 추상화하여 설정된 것인 반면 인스탄스는 이 개념의 현실세계에서 나타나는 실제의 대상이 된다. 즉 '국가'라는 개념에 대해 실제로 구현화 되어 나타난 '한국', '프랑스' 등 실존하는 국가들은 그 개념의 인스탄스가 된다. 따라서 사실 자료는 온톨로지 개념으로부터 실제로 구현화된 현실세계의 정보라고 할 수 있다.

사실 자료를 기존의 온톨로지의 개념구조의 인스탄스로 구축하는 것은 온톨로지가 지닌 전산적, 표현적 장점을 그대로 사용할 수 있게 하고 이를 바탕으로 추론을 가능하게 하기 위해서이다. 온톨로지에서의 개념들은 계층적으로 분류되어 있고 다양한 관계 - *is-a*, *has-a*, *inverse* - 들이 속성과 값으로 표시된 프레임 (frame)으로 표시되어 개념들과의 관계와 제약이 형성된다. 개념들의 인스탄스로서의 사실 자료는 온톨로지의 개념구조가 지닌 연결 관계를 이용할 수 있게 하고 개념들의 확장으로 그 가능한 답들을 정확히 도출해 낼 수 있게 한다.

이런 방법은 비단 질의응답시스템에서 뿐만 아니라 다른 응용시스템에서도 사용될 수 있다. 예를 들어 기존의 정보검색시스템은 검색된 정보가 거의 관련이 없을 때는 사용자로 하여금 질의어나 키워드를 바꾸도록 요구한다. 그러나 온톨로지를 사용하는 지식기반 시스템에서는 그 초점을 관련된 개념으로 옮기게 되어 그 질의어에 대한 더 미세한 정보나 아니면 관련된 더 많은 양의 문서를 도출할 수 있게 한다. [6]

2.1 지식기반 방법의 응용

한편 자연언어는 필연적으로 다의적인 속성을 지니는데, 이로 인해 자연언어처리의 어려움이 나타난다. 따라서 이런 다의적인 단어의 의미를 언어중립적인 중의성이 없는 개념구조로 전환하여 다시 문맥의존적인

(context-sensitive) 어휘의미의 중의성 해결을 시도하기 위해 이 지식기반의 방법이 사용된다. 이런 관점에서 시도되는 방법이 CRL의 Mikrokosmos 프로젝트의 TMR (Text Meaning Representation)이다. 이 TMR을 바탕으로 한 기계번역 시스템에서는 번역대상 언어의 구문분석과 더불어 그 의미구조를 온톨로지를 바탕으로 한 개념구조로 표시한다. 이렇게 표시된 의미구조를 바탕으로 목표언어로의 대응이 이루어지고 최종적인 번역문으로 변환된다.[1]

또한 지식기반을 바탕으로 현실세계의 지식을 데이터베이스화하고 이를 온톨로지의 개념구조와 연결시켜 질의응답시스템을 구축하는 것은 추론을 근간으로 하는 시스템의 질적 향상뿐만 아니라 기존의 이용가능한 모든 자원을 결합하여 사용할 수 있게 한다. 즉 입력문으로 단순한 핵심어가 아닌 자연언어 문장을 받아들이고 이 문장을 분석, 이해하기 때문에 기존의 자연언어처리의 모든 자원들 - 형태소분석, 구문분석, 의미분석- 등의 자원을 그대로 이용하게 한다. 이러한 통합시스템의 구축은 현재 인터넷 및 웹과 관련된 시스템에서 많이 요구되는 방법론이다 [7].

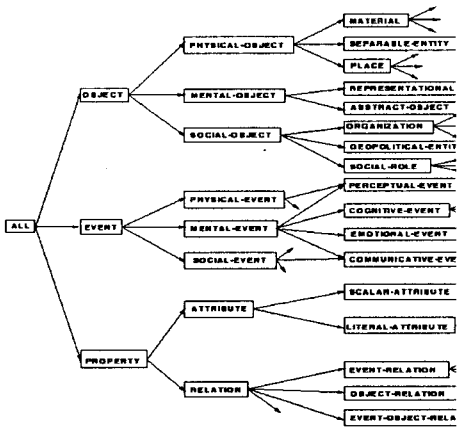
2.2 온톨로지 (Ontology)

지식기반 질의어응답시스템에서 그 자료구조의 근간이 되는 온톨로지는 원래 철학적인 개념으로는 "The branch of metaphysics that studies the nature of existence"라고 정의된다. 그러나 지식기반적인 자연언어 처리에서, 그리고 CRL의 Mikrokosmos에서 온톨로지는 언어-독립적인 (language-independent) 지식 자료로 쓰이는 것으로 어떤 symbol과 그 가능한 관계들로 이루어 진다. 즉 여기서 온톨로지란 어떤 개념들이 실세계에 존재하고 어떻게 그것들이 서로 관련되는가에 대한 지식을 포함하는 전산적 단위, 자원으로 정의된다. 따라서 자연언어처리에서 쓰이는 온톨로지란 세상에 대한 지식체를 구성하는 것이라 할 수 있다. 자연언어처리에서 온톨로지는 어휘부, 통사부, 의미부, 그리고 화용부에 어떤 세상에 대한 지식을 제공하는 것이다. 이런 관점에서 온톨로지는 다음의 정보를 가지고 있는 데이터 베이스라고 할 수 있다 [1]:

(1) 첫째, 어떤 범주들(또는 개념들)이 세상에/어떤 도메인에 존재하는가
(2) 둘째 어떤 속성(property)을 그것들이 지니고 있는가
(3) 셋째, 어떻게 그것들이 서로 연결되어 있는가

1) http://crl.nmsu.edu/New%20Research/research.htm

이 온톨로지는 노드(node)가 개념(concept)으로 되어 있는 방향성이 있는 그래프라고 할 수 있다. 노드들 사이의 링크는 slot과 filler로 표시되며 각 개념들의 속성과 제약이 명시된다. 현재 이 Mikrokosmos 온톨로지에는 대략 5000여개의 개념들이 계층적으로 표시되어 있다. 다음의 그림은 그 상위 몇 층위의 개념구조를 나타낸다.



(그림 1) 온톨로지의 구조 일부

2.3 사실 자료와 인스탄스 (Instance)

이런 개념구조를 바탕으로 사실 자료를 연결시키는 것은 앞에서 지적한 대로 개념의 실제대상으로의 구현(instantiate)된 인스탄스에 의해 이루어 진다. 이 연결은 기본적으로는 온톨로지의 개념구조의 속성 INSTANCE와 그 반대관계인 INSTANCE-OF에 의해 이루어진다. 예를 들어 실제 운동선수들의 자료를 이 개념구조를 바탕으로 구축하기 위해서는 온톨로지 개념 'ATHLETE'를 구현화하여 실제의 운동선수 'ATHLETE-1', 'ATHLETE-2' 등으로 표현한다. 이 인스탄스에는 실제 운동선수들의 이름이 명시된다.

(표 1) 개념구조 예

ATHLETE
DEFINITION: a person trained in exercises or games requiring strength, speed, skills, stamina, etc.
IS-A: SPORTS-ROLE
SUBCLASSES: BASEMAN, CATCHER, CENTER
AGENT-OF: SPORTS-COMPETITION
INSTANCE: ATHLETE-1, ATHLETE-2

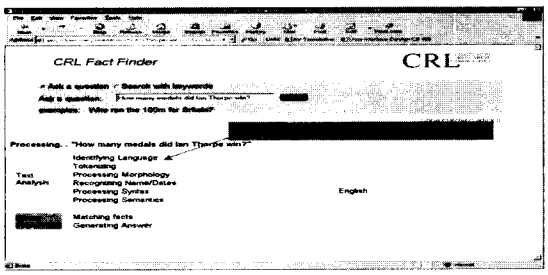
(표 2) 인스탄스 예

Instance: ATHLETE-65	
NAME	Ian Thorpe
AGENT-OF	SPORT-RESULT
RESIDENT-CITY	Sydeney(CITY-8)
MEMBER-OF	TEAM-260

여기서 대문자로 되어 있는 것은 모두 온톨로지에 규정된 개념들이다. 따라서 이런 개념들이 서로 연결되어 필요한 정보를 추출하는데 사용된다. 더 자세한 것은 사실자료 구축에서 논하도록 한다.

3. 질의응답시스템

지식기반 질의응답시스템의 전 과정은 다음과 같은 인터페이스로 도식화될 수 있다. 이 시스템은 '2000년 시드니 올림픽'을 대상으로 구축되었다. 시드니 올림픽과 관련하여 사용자가 여러 질문을 자연언어 문장으로 입력하면 이미 구축되어 있는 사실 자료를 바탕으로 그 답을 도출해 내는 과정으로 되어 있다. 여기서는 질의어 'How many medals did Ian Thorpe win?' 문장으로 설명하도록 한다.



(그림 2) 질의응답시스템 인터페이스

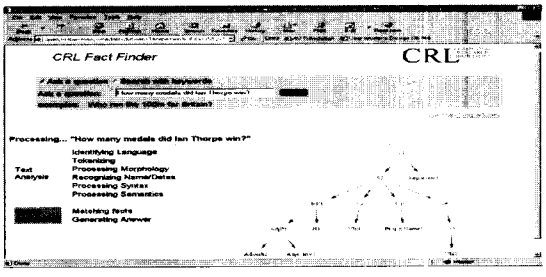
3.1 질의응답시스템의 구조

질의응답시스템에서 해당되는 질문을 자연언어로 입력하면 크게 두 단계를 거치면서 그 답을 도출하게 된다. 첫 단계는 위 그림 2의 왼쪽에 나와 있는 텍스트 분석 (text analysis) 단계로 자연언어처리와 관련한 기존자원을 이용하게 된다. 입력언어를 구별 (identifying language) 하고, 단위별로 분리(tokenizing) 하고, 형태소 분석 (processing morphology) 및 이름/날짜 인식 (recognizing name/dates), 구문분석 (processing syntax), 의미분석 (processing semantics) 등이 이루어진다. 이 분석 단계의 결과로 질의어의 구조와 의미가 파악된다. 이렇게 분석된 문장은 그 다음 단계인 질의어 처리 (query processing) 단계에서 구축해 놓은 사실자료와 분석된 질의어와의 대응 및 개념구조의 확장을 통해 해당되는 값을 찾아 질문의 대답을 도출해 내게 된다. 진행되는 각각의 단계는 왼쪽의 메뉴에서 하이라이트되어 표시된다.

첫 단계인 텍스트 분석에서는 언어 중립적인 (interlingua) 방식에 따른다. 따라서 입력문이 어떤 언어라도 (현재는 터어키어, 러시아어, 스페인어, 한국어, 영어, 이란어, 아랍어, 독일어 자료가 구축) 이 언어가 자동적으로 인식된다. 이 언어의 인식은 해당 언어의 코드 체계 외에 각 언어의 기본적인 어휘 (수사, 고유명사, 년/월/일 등의 표현 등)를 기초로 한다. 각 언어의 형태, 통사 분석이 끝난 후에 언어중립적인 의미구조가 도출된다. 이런 변환으로 언어중립적인 온톨로지의 개념들과 질의어가 대응될 수 있다. 이런 다언어 (multilingual)를 지원하는 특성으로 인해 각 언어의 고유명사 사전, 형태소 분석 사전 등이 필요하다.

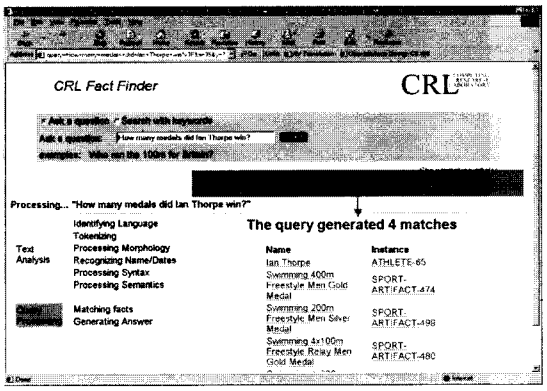
구문분석과 의미분석은 CRL에서 연구되고 있는 MEAT(Multilingual Environment for Advanced Translations)와 TMR(Text Meaning Representation) 모듈에 따른다. 이 구문분석은 차트와 통합연산 (unification operation)에 기반한 방법으로 현재 여러 언어 (이란어에서의 영어로, 한국어에서 영어로) 의 기계번역 시스템에서 사용되고 있다.[2)]

이 텍스트 분석과정은 전체 시스템에서 가장 많은 처리 시간을 요구하는 부분이다. 따라서 각 모듈간의 강건성 (robustness)과 신속성이 요구된다. 현재 각 모듈을 더 견고하게 하는 작업들이 진행되고 있다. 이 결과에 따라 위에서 명시된 각 모듈들은 서로 통합될 수 있다. 다음의 그림은 텍스트 분석에서 구문분석, 의미분석 단계의 결과를 보여주고 있다.



(그림 3) 구문분석 결과

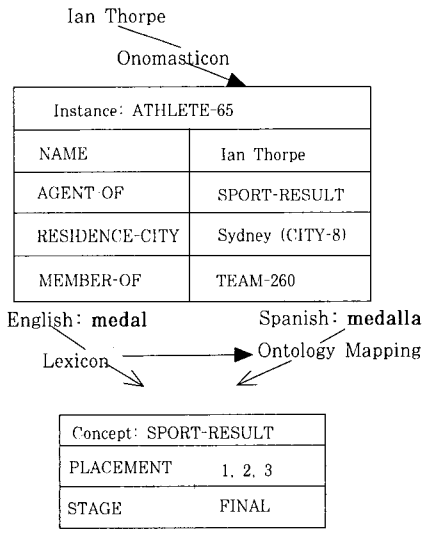
2) 이에 대한 검토는 본 논의의 주제와 벗어나므로 자세한 것은 언급하지 않도록 한다. 세부적인 것은 http://crl.nmsu.edu/New%20Research/research.htm 참조



(그림 4) 의미분석 결과

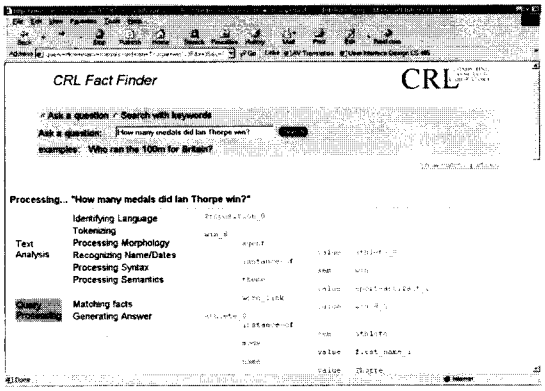
3.2 질의응답 단계

입력문의 형태, 통사, 의미분석이 이루어지고 나면 질의어를 이미 구축해 놓은 사실 자료와 연결시키게 된다. 특히 의미분석이 행해진 후에 다음과 같은 온톨로지에 바탕을 둔 사실 자료와 대응된다. 인명, 지명 등과 같은 고유명사는 따로 구축된 사전을 바탕으로 대응이 이루어진다.



(그림 5) 개념구조와 대응

여기서 영어의 medal과 스페인어 medalla는 각각 온톨로지로 언어중립적인 개념구조와 대응되는데, 이 경우 SPORT-RESULT라는 개념구조와 대응된다. 한편 고유명사 Ian Thorpe는 고유명사 사전(Onomasticon)에 의해 ATHLETE 온톨로지 개념의 한 인스탄스인 ATHLETE-65와 연결되며 그 정보는 사실 자료에 저장되어 있다. 이런 단편적인 사실 정보는 다시 각 개념들의 확장에 의해 추론을 거쳐 관련된 정보를 도출하게 된다. 결과적으로 이 질의응답시스템은 다음과 같은 대응 결과를 화면에 도출하게 된다.



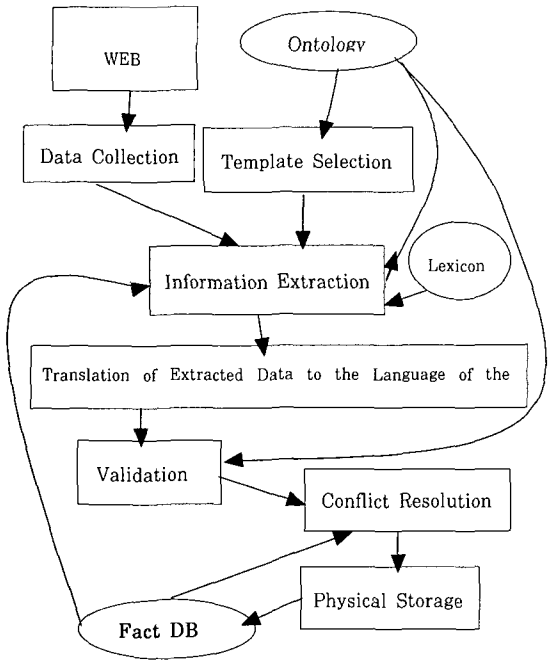
(그림 6) 질의응답 결과

그러나 실제의 질의어응답시스템은 위와 같이 온톨로지에 근거한 개념구조를 다시 자연언어로 변환하여 출력을 하는 과정을 거치게 된다. 위의 그림과 같은 결과는 개념구조와 대응된 결과를 표시할 뿐 실제의 대답이 될 수 있는 자연언어문장은 아니기 때문이다. 이 자연언어 문장 생성은 기계번역시스템의 생성시스템을 바탕으로 한다. 그러나 아직 질의어에 대한 아주 자연스러운 문장은 생성되지 못하고 있다. 이 과정에 있어서도 많은 노력이 요구된다. 4.7에서 이 과정의 자동화 정도와 진행정도에 대해 기술한다.

4. 사실 자료의 구축

4.1 사실 자료 구축과정

지금까지 간략히 살펴 본 질의응답시스템에서 그 핵심과정은 그 자료를 모으고 가공하여 데이터베이스를 구축하는 것이다. 이제 이 과정에 대해 논의하도록 한다. 우선 사실 자료 구축과정은 다음과 같이 도식화할 수 있다.



(그림 7) 사실 자료 구축 과정

사실 자료의 구축과정은 가공되지 않은 웹 자료에서 자료를 추출하고 (data collection) 이를 바탕으로 필요한 개념 판 (template)을 선택하여 웹에서 관련된 정보를 추출한다 (information extraction). 이 추출된 정보를 온톨로지의 개념의 인스탄스가 될 수 있는 구조로 변환하는 작업이 필요하다. 이 과정은 인간의 간섭이 가장 많이 필요한 부분이다. 이렇게 변환된 인스탄스는 다시 기존의 구축된 자료 구조 및 개념들과 부합되는지 조사하는 정당화 (validation) 과정을 거치게 된다. 여기서 나타나는 충돌이 되는 정보는 중의성을 해결한 후 (conflict resolution) 마지막으로 데이터베이스에 기록되어 자료 구조로 사용된다.

4.2 사실 자료 수집 가공

사실 자료의 근간이 되는 자료는 일차적으로 웹에서 얻어진다. 해당 자료를 얻기 위해서 우선 웹 스파이더 (spider)를 사용하여 필요한 자료를 자동적으로 추출하여 저장한다. 한편 이런 자료는 일종의 가공되지 않은 자료이기 때문에 여기서 정보를 자동적으로 추출하기 위해서는 여러 도구 (tool) 들을 개발 할 필요가 있다. 그러나 각 자료들 마다 정형화된 자료 구조를 보이지 않기 때문에 모든 자료에 다 쓰일 수 있는 추출 스크립트 (extraction script)를 작성하기는 어렵다. 따라서 테이블 등의 정형화된 자료들은 HTML의 태그의 정보를 이용하여 일반적인 파서로 분석하여 해당 자료들에서 원하는 내용을 도출해 낸다. 이런 정형화된 구조 외에 각 페이지에 특수한 구조에서 정보를 도출해 내기 위해 페이지 의존적인 파서를 구축할 필요가 있다. 이 연구에서는 Perl 스크립트에 기초한 파서들을 구축하였으며 이러한 파서들로 인해 상당히 많은 자료를 필요한 형태로 빠르게 도출할 수 있다. 실제 스크립트의 간략한 예를 4.4.절에 기술한다.

4.3 Template 선택

각각의 자료들은 특정 범위의 지식에 따라 수집되기 때문에 이미 구축되어 있는 온톨로지에서부터 해당되는 개념의 template를 선택하게 된다 (template 선택 단계). 이 template는 온톨로지 개념구조에서 실제 세계의 자료를 기술하기 필요한 틀이 될 수 있다. 가령 이 '2000년 시드니 올림픽' 사실 자료 구축을 위해서는 ATHLETE, CITY, SPORTS-ACTIVITY, SPORT-RESULT, STAGE 등의 개념구조가 필요하다. 이 경우에 필요한 template 들은 '시드니 올림픽'과 관련된 모든 사실을 기록하기 위한 온톨로지의 개념 구조들이 된다. 즉 운동선수의 신상정보를 위해서 '이름, 성별, 나이, 거주도시, 키, 지난 기록' 등, 운동경기의 정보를 위해서 '경기 종목, 세부구분 (예를 들어 수영의 경우 남자 400m 평형 등), 메달 또는 기록' 등의 실제 자료에 해당되는 온톨로지 개념을 찾고 이를 바탕으로 필요한 사실 자료구조 판을 만들게 된다. 만일 실제의 자료와 부합하는 개념이 없으면, 새로운 개념을 온톨로지에서 획득하거나 가장 밀접한 개념들을 선택한다. 이 단계에서는 숙련된 개념론자들의 선택이 요구된다.

따라서 이 template 들은 실제의 입력 문장을 그대로 반영하여 형성되는 것이 아니라 범위가 정해진 자료에서 가능한 답이 될 수 있는 대상들의 개념화라고 할 수 있다. 이렇게 온톨로지에서 얻어진 개념들을 바탕으로 위 〈그림 5〉에서처럼 ATHLETE, SPORT-RESULT라는 판이 형성되고 실제의 자료에서 해당되는 정보를 채워 넣게 된다. 여기서 대문자로 되어 있는 속성들은 다 온톨로지 개념들이다. 따라서 이 개념들을 연결하여 다른 개념으로 확장되어 필요한 정보들을 획득하게 된다. 다음의 (표 3)은 ATHLETE의 template의 정보들을 보여주고 있다. 여기서 진하게 표시된 왼쪽의 속성들은 온톨로지의 개념들이거나 기

본적인 관리 표시자 (ACTION: Add는 이 자료를 사실자료에 새로 더하라는 것을 의미) 들이다.

(표 3) ATHELTE template

```
ATHLETE
        ACTION: Add
        REFNUM: 1
        NAME: Ian Thorpe
        SRC: /
        http://www.fina.org
        DATALINK:
        http://fina.org/bio thorpe.html
        AGENT-OF: (SWIMMING)
        HAS-COACH: Doug Frost(TRAINER)
        WEIGHT: 96kg
        AGE: 17 yr
        BIRTH-PLACE-CITY: Sydeney(CITY)
        GENDER: MALE
        HEIGHT: 1.95m
        NATIONALITY: Australia(NATION)
        ORIGIN: Australia(NATION)
        RESIDENCE-CITY: Sydeney(CITY)
```

4.4 온톨로지 언어로의 변환

필요한 자료와 해당되는 template 들이 설정되고 난 후 이 판을 채우기 위한 정보가 추출된다. 정보의 추출은 앞에서 설명한 대로 테이블과 같이 정형화된 자료에서는 스크립트로 그대로 추출될 수 있지만 그렇지 않은 경우에는 페이지 의존적인 휴리스틱스에 바탕을 둔 스크립트를 사용한다. 가령 참가선수들의 인적사항을 추출하기 위해서는 '이름, 나이, 출신지, 키, 몸무게 ' 등의 키워드를 기준으로 해당되는 정보를 자동적으로 얻는다. 다음 (표 4)는 키워드를 바탕으로 필요한 정보를 추출하기 위한 Perl 스크립트의 일부로 웹페이지에 나타나는 키워드 'athname' (운동선수이름), 'Sport' (경기종목)를 중심으로 해당되는 정보를 추출한다.

(표 4) 정보추출을 위한 스크립트

```
sub get_name{
if($src =~ /athname₩">(.+?)</){
        $name = $1;
        $name = &name_change($name);
        $name =~ s/(₩w+)/₩u₩L$1/g;
        print "₩tNAME: $name(ATHLETE)₩n";}
}
sub get_sport{
if($src =~ /Sport:.+?₩.html₩">(.+?)</){
        $sport = $1;
        $sport =~ s/ /-/g;
        $sport = uc($sport)}
}
```

그러나 이렇게 추출된 정보는 불규칙한 자료 구조와 파싱으로 인하여 정확하지 않을 수 있기 때문에 이를 인간이 직접 점검하기 위한 정제과정이 필요하고 이 과정 후 다시 온톨로지 언어 (language of ontology)에 부합하도록 변환되어야 한다.

온톨로지 언어로의 변환이란 획득된 자료를 해당 개념 및 사실로 표현하는 것을 의미한다. 예를 들어 'SWIMMING'이라는 개념이 이 사실 자료를 위해 획득되었다면 이 개념이 온톨로지의 개념구조에 이미 존재하는 것인지, 아니면 유사한 개념이 있는지 검토되어야 한다. 만일 이미 존재한다면 더 이상 획득할 필요가 없으나 비슷한 개념이 존재한다면 (예를 들어 SWIMMING-EVENT라는 개념이 존재한다면) 이 개념을 비슷한 개념으로 대응시킨다. 이런 변환은 구체적인 사실 자료를 획득할 때 더 어려움이 나타난다. 가령 일반적으로 같은 대상이 코퍼스에서 다르게 표시되는 경우, 즉 같은 대상을 나타내는 '400mX4 men's freestyle', '400X4 freestyle-men', '4인조 400m 남자 자유형' 등이나, 운동종목에서의 중복된 이름 (수영과 사격 등에서의 freestyle 등) 등의 경우에 이를 표준화할 방법과 중의성을 해결할 필요가 있다. 또한 이 과정은 다음의 획득된 자료의 타당화(validation) 과정과 밀접한 관련을 맺는다.

따라서 이 과정이 질의어 응답시스템에서 사실 자료를 구축하기 위한 과정에서 가장 오랜 시간이 요구되고, 인간의 간섭이 요구되는 부분이다. 앞에서와 같이 자료의존적인 구조로 인해 스크립트 등으로 반자동화할 수 있지만 여전히 인간의 간섭이 필요한 부분이라 대용량의 지식기초를

구축하는데 최대의 장애물이 된다. 획득된 지식이 온톨로지와 기존의 사실 자료와 부합하는지 조사는 일차적으로는 스크립트에 따르지만 그 이후에는 지식기초획득편집기(KBAE: knowledge-base acquisition editor)를 이용한다. 이 편집기는 온톨로지 개념을 입력하고 수정하고 검색할 수 있는 프로그램이다. 따라서 획득된 지식을 이 프로그램을 통해 여러 속성들이 부합하는지를 점검하게 된다. 이 과정은 완전히 자동화될 수 있지 못하기 때문에 이 과정을 얼마나 자동화할 수 있는가가 사실 자료의 확장 및, 다양한 범위의 지식기초를 확립하는데 관건이 된다.

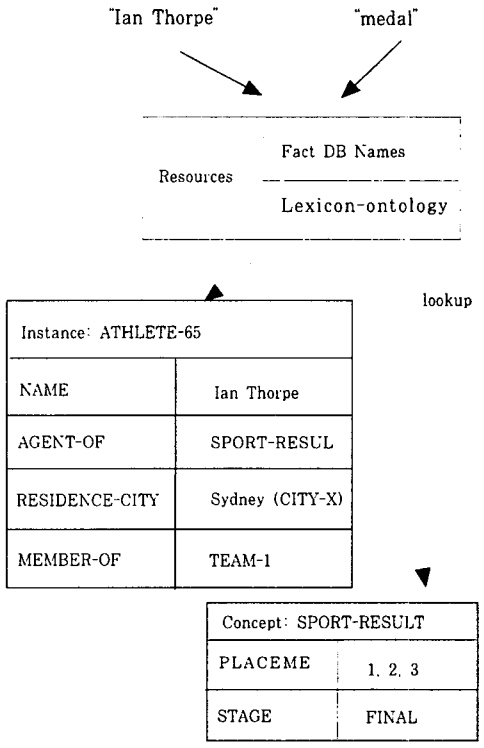
4.5 타당화 및 충돌해결

이렇게 온톨로지로 변환된 자료는 앞 절에서 언급한 대로 편집기를 사용한 타당화 (validation) 단계를 거쳐야 한다. 이 단계에서 명칭의 부합성 그리고 속성들이 domain이나 range에 부합하는지를 점검해야 한다. 그렇지 않고는 개념들의 계층구조와 제약을 나타나는 온톨로지의 특징을 사용할 수 없기 때문이다.

이 타당화 과정에서 기존의 온톨로지와 모순이 되는 결과가 나타나거나 중의성이 생기는 등의 충돌(confliction)이 생기는 경우 이를 해결하는 과정이 필요하다. 예를 들어 획득된 자료가 "Bush"인 경우 이것이 고유명사로 사실 자료에 첨가될 경우 이미 온톨로지의 개념으로 '식물'의 한 유형으로 존재하는 경우 중의성이 생기게 된다. 이런 개념유형 충돌을 해결하기 위해서는 HUMAN 대 PLANT 라는 구별되는 개념구조를 사용하거나 또는 더 자세한 정보를 표시하여 해결할 수도 있다. 예를 들어 'Michael Johnson'이 육상선수일 수도 있고 동명이인으로 하키선수일 경우 그 해당 연결 개념을 RUNNER 대 HOCKEY_PLAYER 등으로 구분하여 사용한다. 이렇게 정제된 사실적 정보는 최종적으로 데이터 베이스 형태로 저장되는데, 이는 XML로 되어 있어 실제 웹 브라우저와 쉽게 연동될 수 있다.

4.6 사실 자료의 확장 및 추론

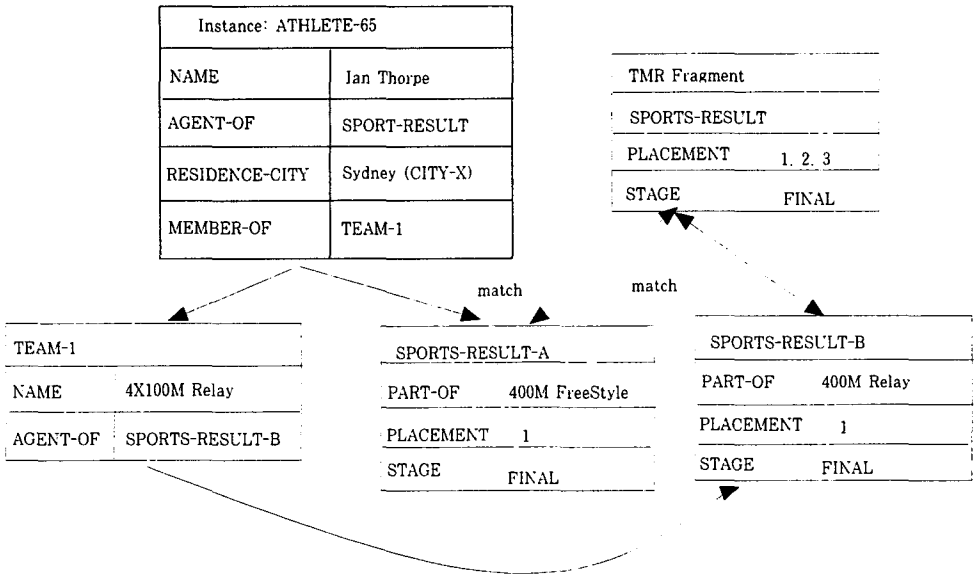
지식기반 질의응답시스템의 가장 큰 장점은 실제세계의 사실 자료를 온톨로지의 개념구조와 연결시켜 추론을 바탕으로 가능한 응답을 도출해 낼 수 있다는 점이다. 이 추론이 개념들의 확장으로 어떻게 이루어지는지 간략히 살펴 보도록 한다. 우선 텍스트 분석 단계에서 형태소분석, 구문분석, 의미분석에 의해 질의어가 무엇을 의미하는지 파악된 후 필요한 정보를 찾기 위해 해당 키워드 설정하고 이를 구축된 자료구조에서 검색하게 된다. 이를 다음과 같이 도식화할 수 있다.



(그림 8) 사실 자료 검색

'Ian Thorpe'는 인명사전에 의해 그리고 medal은 단어와 온톨로지의 사상에 의해 해당되는 사실 자료, ATHLETE-65, SPORT-RESULT로 연결된다. 이렇게 기본적으로 찾아진 사실 자료를 바탕으로 개념의 확장이 된다. 위의 그림에서 ATHLETE-65로 표시되는 "Ian Thorpe"는 SPORT-RESULT-A의 AGENT이며, TEAM-1의 MEMBER라는 것이 표시되어 있다."medal"에 대해서는 FINAL 이라는 STAGE의 정보가 기록되어 있다. 이 속성-값들을 계속 연결해 나가면 다음과 같다.

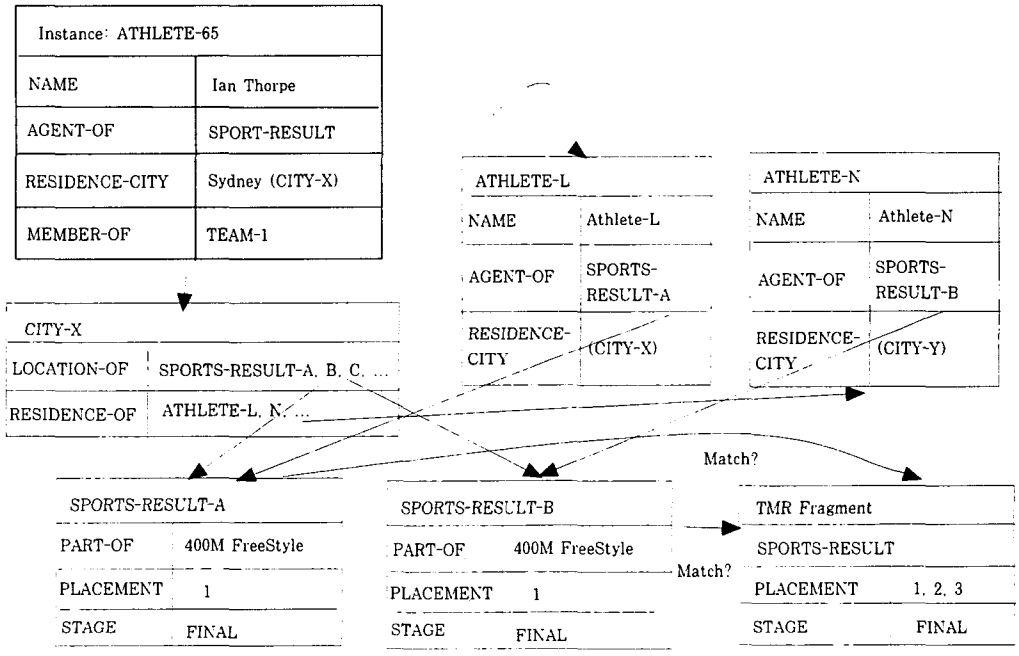
SPORT-RESULT는 각각 SPORT-RESULT-A, SPORT -RESULT-B로 확장되어 400M FreeStyle의 결과와, 4x100M Relay의 결과에 이르게 된다. 이런 확장에 의해 Ian Thorpe와 연관된 팀, 종목 등이 연결되어 질의어의 답이 될 수 있는 '메달의 개수' 및 어떤 종목에서의 메달인지의 정보가 도출될 수 있다.



(그림 9) 개념확장으로 대응 설정

그러나 이러한 확장은 반대로 온톨로지의 계층적인 구조와 inverse link에 의해 무한한 확장과 잘못된 경로로 진행될 수도 있다. 이러한 경우를 배제하기 위해 다음과 같이 inverse 속성에 의해 연결된 속성으로는 확장되지 못하도록 하고 각각의 연결을 타당화해야 한다.

위의 그림에서 'Ian Thorpe'의 인스탄스에서 확장시킬



(그림 10) 잘못된 경로로의 확장

때 만일 RESIDENCE-CITY인 Sydney의 inverse관계인 CITY-X를 먼저 따른다고 생각해 보자. 이 CITY-X는 여러 경기 결과들 (SPORTS-RESULTS-A, B...) 의 장소 (LOCATION-OF) 이기도 하며, 여러 운동선수들 (ATHLETE-L, N, M...) 등의 거주지 (RESIDENCE-OF) 의 정보를 갖고 있다. (그림의 두 번째 단계). 이 개념들과 인스탄스들을 운동경기 결과들을 따라 계속 확장시키게 (세번째 단계) 되어 해당되는 정보를 다 수합하게 되어 최종결승에서 1, 2, 3 위라는 모순된 정보를 도출하게 된다.

이런 잘못된 결과는 'Ian Thorpe'라는 인스탄스에서 개념을 확장시킬 때 inverse 관계인 CITY-OF에서부터 확장을 시켰기 때문이다. 개념구조망에서 어떤 개념으로라도 확장을 시켜 필요한 정보를 도출할 수 있기 때문에 처음부터 이러한 관계로의 확장은 배제되어야 한다.

4.7 시스템의 효율성 및 자동화 정도

지금까지 개략적으로 살펴본 질의응답시스템은 기존의 여러 응용시스템 및 자원을 이용하여 구축되고 있다. 따라서 기존의 이용가능한 자원과 현재 개발되고 있는 모듈과의 결합과 그 자동화 정도가 전체 시스템의 효율성을 평가하는데 중요한 기준이 된다. 지금까지 CRL에서 이용가능한 자원 및 그 자동화 과정은 (표 4)와 같다.

(표 4) 전체시스템 자동화 정도

처리과정	자원들	자동화 정도
Language Recognition	Algorithms Traning Data	100%
Tokenization & Morphology	Lexica	100%
name S Date Recognition	Onomastica MITRE	100%
Syntactic Analysis	Lexica, Grammars, Onomastica	100%
Semantic Analysis	Onotolgy, Lexica	100%
Answer Generation	Ontology, Lexica, FDB	20% (100%)
Ontology Acquisition	Raw & Processed Data	10% (50%)
Lexicon Acquisition	Ontology, Raw & Processed Data	15% (75%)
Fact Acquisition	Ontology, Lexica, FDB, Web, Data	20% (80%)

앞의 (표 4)에서 보듯이 텍스트 분석의 모든 단계는 기존의 자원들을 그대로 사용하기 때문에 100% 자동화 정도를 보인다. 그러나 응답문 생성(answer generation)이나 온톨로지/어휘/사실 정보 획득은 낮은 자동화 단계를 보인다. 이 부분에서는 앞에서 지적한 대로 인간의 간섭이 요구되기 때문이다. 온톨로지, 어휘, 사실 정보의 습득에서 괄호안에 표시되어 있는 비율은 현재 이 시스템에서 구현되어 있는 정도의 비율을 나타낸다.

5. 결론

지금까지 지식기초의 관점에서 질의어응답시스템의 구축에 관해 살펴보았다. 여기서 기술된 시스템은 기존의 자연언어처리의 자원을 충분히 이용하여 단순히 입력문이 키워드가 아닌 자연언어문장을 이해하는 시스템이며, 이 분석을 바탕으로 미리 구축된 사실 자료에서 해당 정보를 사실 자료의 확장으로 인한 개념연결로 가능한 답을 도출해 낼 수 있다. 이런 시스템에서 핵심은 사실 자료를 구축하는 작업이라 할 수 있다. 이를 위해 웹이나 기타 가공되지 않은 자료를 획득하고 이를 개념구조에 부합하도록 정제하는 작업이 필요하며 또한 인간의 간섭이 상당부분 요구되기 때문에 자동적으로 대용량의 시스템을 구축하기 어렵게 한다.

이 과정을 보다 용이하게 하기 위해 자료의존적인 스크립트와 휴리스틱스에 기반한 방법론들이 개발되었다. 또 온톨로지의 개념들과 더 부합되도록 하기위해 여러 에디팅 툴들이 개발되었다. 이런 시스템은 대규모의 자료가 축적된다면 그 시스템의 질을 향상시킬 수 있는 장점이 있지만, 온톨로지의 개념구조가 이미 구축되어 있어야 하며, 실제 자료를 이에 맞게 가공하는데 오랜 시간이 걸린다는 문제점로 인해아직 제한된 범위를 벗어난 대규모의 시스템에는 적용되지 못하고 있다. 그러나 축적된 기술과 방법론 및 도구들의 발전으로 이 과정이 좀 더 자동화되고 사실 자료가 대량으로 개발된다면 효율적으로 그리고 질적으로 향상된 응용시스템을 개발할 수 있다는데 그 의의가 있으며 계속 추구되어야 할 방법론이라 할 수 있다.

참 고 문 헌

[1] Burkert Gerrit. Lexical Semantics and Terminological Knowledge Representation. In Computational Lexical Semantics, Edited by Patric Saint-Dizier and Evelyne Viegas, Cambridge University Press, 1995.

[2] Mahesh Kavi, Ontology Development for Machine Translation: Ideology and Methodology, Technical Report MCCS-96-292, Computing Research Laboratory, New Mexico State University 1996.

[3] Nirenburg Sergei, Carbonell J., Tomita M, and Goodman K, Machine Translation: A Knowledge-Based Approac, Morgan Kaufmann Publishing, 1992.

[4] Shin Hyopil and Spencer Koehler, A Knowledge-Based Fact Database: Acquisition To Application. Proceedings of the International Conference KBCS2000, National Centre for Software Technology, India, 2000.

[5] Katz Boris, From Sentence Processing to Information Access on the World Wide Web, http://www.ai.mit.edu/people/boris/webaccess, 1997.

[6] Nirenburg Sergei, Ontology-based Cross Lingual Information Retrieval, Unpublished Proposal, Computing Research Laboratory, New Mexico State University, 1988.

[7] Shin Hyopil and Bill Ogden, Combining Summarization and Machine Translation Facilities to Build an Interactive Cross-Language Retrieval System. Proceedings in the 19th International Conference on Computer Processing on Oriental Languages, 2000.