새로운 N-ary 관계 디자인 기반의 온톨로지 모델을 이용한 문장의미결정

A Semantic Similarity Decision Using Ontology Model Base On New N-ary Relation Design

김수경(Su-Kyoung Kim)*
안기홍(Kee-Hong Ahn)**
최호진(Ho-Jin Choi)***

초 록

시맨틱 웹 기술의 제안과 더불어 다양한 분야에 온톨로지의 특징을 적용한 기술 개발 연구가 많이 진행되고 있다. 인간이 소유한 개념을 가장 적절하게 표현하기 위해 현재에도 OWL, RDF와 같은 온톨로지 언어의 표현력을 확장시키기 위해 N-ary 관계나 모델-이론 의미론과 같은 개발이 진행되고있다. 본 연구는 한국어에 있어 문장이 내포하는 의미를 정확하게 결정하기 위해 문장의 구조에 따라 달라지는 단어의 의미를 연관할 수 있도록 N-ary 관계와 디자인 기반이 적용된 온톨로지의 지식 표현 방법을 연구하였다. 특히 다양한 지식 영역을 포함하는 다의어(polysemy)와 동의어(synonym)의 특징을 갖는 단어에 있어 각 지식 영역으로 분류되어 각 지식 영역에 있는 유사한 의미를 가진 단어로 확장되어 유사한 의미를 가진 단어가 포함된 문장의 경우 까지도 확장할 수 있는 표현 방법을 연구하였다. 연구의 검증을 위해 사용자가 입력한 병증 문장을 제안된 방법에 따라 구축된 온톨로지내 지식 관계와 의미 결정을 위한 추론 표현 방법을 이용하여 병증의 의미를 결정하고 그에 따른 진단을 제공하는 실험 시스템을 구현하였고, 한국어가 갖고 있는 문장의 유의성, 모호성, 복합성 의 특징을 보유한 증상문들의 실험 결과 의미 결정과 유사 의미 확장에 있어 우수한 성능을 보여주었다.

ABSTRACT

Currently be proceeded a lot of researchers for 'user information demand description' for interface of an information retrieval system or Web search engines, but user information demand description for a natural language form is a difficult situation. These reasons are as they cannot provide the semantic similarity that an information retrieval model can be completely satisfied with variety regarding an information demand expression and semantic relevance for user information description. Therefore, this study using the description logic that is a knowledge representation base of OWL and a vector model-based weight between concept, and to be able to satisfy variety regarding an information demand expression and semantic relevance proposes a decision way for perfect assistances of user information demand description. The experiment results by proposed method, semantic similarity of a polyseme and a synonym showed with excellent performance in decision.

키워드: 의미유사도, 온톨로지, N-ary, 추론, 질의응답시스템
semantic similarity, ontology, N-ary, inference, question-answer system

* 한국정보통신대학교 공학부 연구교수(kimsk@icu.ac.kr) (제1저자)
** 한밭대학교 컴퓨터공학과 교수(khahn@hanbat.ac.kr) (공동저자)
*** 한국정보통신대학교 공학부 교수(hjcho@icu.ac.kr) (공동저자)

■ 논문접수일자 : 2008년 8월 20일 ■ 최초심사일자 : 2008년 8월 25일 ■ 게재확정일자 : 2008년 12월 1일
■ 情報管理學會誌, 25(4): 43-66, 2008. [DOI:10.3743/KOSIM.2008.25.4.043]

1. 서 론

RDF와 OWL과 같은 온톨로지 언어에서, 두 개체들 또는 개체와 값들을 연결하는데 사용하는 속성은 이항관계로 표현된다. 그러나 인간의 소유한 개념이나 문장의 경우는 단순히 이항 관계만으로는 개념들의 표현에 있어 자연스럽고 유용하지 않다. 또한 현재 개념의 상태가 수치나 가중치에 따라 다른 관계를 결정해야 되는 경우에 있어서도 이항 관계만으로는 그에 대한 표현이 어려운 상황이다. 그리고 관계된 개체들 간이 순서를 온톨로지에 표현해야 하는 경우나 "홍길동은 감기 치료를 위해 medicine.co.kr에서 10,000원을 주고 아스피린을 샀다"는 문장은 개체 '홍길동', 엔티티 'medicine.co.kr'과 '아스피린'은 서로 이항의 트리플 관계로 연결되나 이 문장은 목적(감기 치료)와 금액(10,000원)과 같은 다른 구성요소도 표현해야만 한다.

이전 RDF에서 이 같은 복합 관계를 표현하기 위해 RDF reification을 제안하였으나 이 표현법의 경우 다음장에서 기술된 어려움으로 인해 이 방법의 사용을 기피하는 상황이다. 그리고 OWL DL과 Full의 경우 추론을 위한 강력한 어휘를 제공하고 있으나 기존의 온톨로지 모델의 경우 이를 사용할 수 있는 디자인 패턴의 적용이 부족하다. 이는 대부분 온톨로지 모델의 관계 설정을 이항 관계를 기반으로 하기 때문이며 이항 관계의 단점을 극복할 수 있는 표현 방법이 부족하기 때문이다. 이를 위해 W3C는 OWL 1.1 Web Ontology Language: Model-Theoretic Semantic의 Working Draft를 발표했다.

그리고 현재 지식베이스를 활용하는 지식기반시스템이나 정보시스템 또는 진단시스템의 경우 검색어의 경우 문장형태의 질의어에 대한 완전한 처리를 제공하기 위해서는 문장의 의미에 대한 정확한 결정이 매우 중요하다. 그러나 현재 대부분의 시스템은 문장의 의미에 따른 검색 결과를 제공은 어려운 실정이다. 이는 지식기반시스템이나 정보검색시스템에 있어 지식베이스를 구축할 때 모델화한 정보 내용을 사용자가 빈약하게 정의하거나 내재적으로 다의어(polysemy) 같은 광범위한 정의를 가졌을 경우, 특정한 단어를 이용한 검색을 하고 있을 때 연관 페이지가 의미적으로 유사한 동의어(synonym)를 사용하는 경우 정확한 검색 결과의 제공은 더욱 어렵기 때문이다.

이러한 문제를 해결하기 위해 본 연구는 지식베이스로서 개념간 속성관계를 통해 의미를 추론할 수 있는 온톨로지를 적용하고자 한다. 한국어 문장이 갖고 있는 특징이나 어려움을 해결하기 위한 가장 기본적인 방법으로 앞에서 제안한 새로운 N-ary 관계 디자인 패턴과 모델-이론 의미론에 따른 지식 표현에 적용하여 OWL DL과 OWL Full이 갖고 있는 다양한 어휘를 활용하고 이를 기반으로 단계적 추론 공리와 의미 결정을 위한 SWRL 기반의 규칙 표현의 정의를 제공한다.

이같은 방법에 따라 제안된 연구 방법의 검증을 위해 사용자가 입력한 병증 문장을 제안된 방법에 따라 다양한 지식 영역에 대해 온톨로지를 구축하였다. 이를 위한 온톨로지 모델링에 있어 중개자(mediator) 클래스의 적용과 같은 새로운 N-ary 관계 디자인 패턴과, Model-Theoretic Semantic에서 제안한 어휘를 적용한 온톨로지 모델을 제시하였다. 지식 관계와

의미 결정을 위해 추론 표현 방법을 이용하여 병증의 의미를 결정하고 그에 따른 진단을 제공하는 실험 시스템을 구현하였고, 한국어가 갖고 있는 문장의 유의성, 모호성, 복합성의 특징을 보유한 증상문들의 실험 결과 의미 결정과 유사 의미 확장에 있어 우수한 성능을 보여주었다.

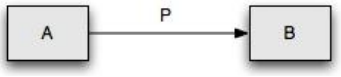
본 논문의 구성은 2장에서는 RDF Reification, OWL Full와 새로운 N-ary 관계 디자인 패턴의 비교, 현재 진단시스템이나 질의응답시스템의 분석을 제공하고, 3장은 사용자의 병증에 대한 간이 진단시스템을 위해 요구되는 온톨로지 구축을 위해 새롭게 제안된 N-ary 관계 디자인 방법을 이용한 온톨로지 모델과 단계적 추론 공리와 유의어 확장을 위한 SWRL 규칙 표현을 제안하고, 4장은 본 연구의 검증을 위해 사용자의 병증에 따른 간이 진단 시스템을 구축하고 다양한 문장에 대한 실험을 제공하고 5장의 결론과 향후 연구 방향으로 마무리한다.

2. 관련연구

2.1 기존 이항 관계 디자인

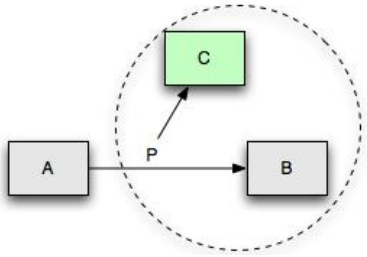
온톨로지 모델링에 있어 표현될 인간의 개념은 단순히 주어(subject) - 술어(predicate) - 목적어(object)와 같은 트리플의 구조만으로 이루어 지지 않는다. 많은 온톨로지 모델에 있어 트리플 들간의 연관성을 표현하기 위해 오브젝트 속성만을 이용하는 경우가 대부분이다. 그러나 복잡한 개념이나, 가중치나 확률의 표현, 다양한 구성요소로 이뤄진 용어적 개념, 관계의 순서화를 위해 이항 관계의 제약점을 극복하고자 하는 연구가 N-ary 관계 디자인 패턴이다.

기본적으로 속성은 〈그림 1〉과 같은 이행관계로, 여기에서 P는 개체 A,B를 연결하는 속성이다.



〈그림 1〉 이항 관계

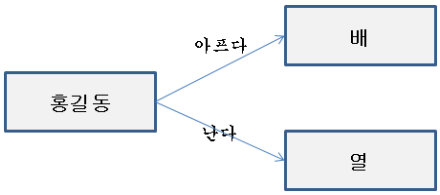
그러나 실제 많은 예에서 〈그림 2〉와 같이 다양한 개체가 추가되고 관련될 필요가 있다. 여기에서 'P'는 'A', 'B' 그리고 'C' 들간의 관계의 인스턴스를 참조한다.(다른 개체 'D', 'E' 그리고 'F'도 가능하다.)



〈그림 2〉 새롭게 추가될 개체 'C'

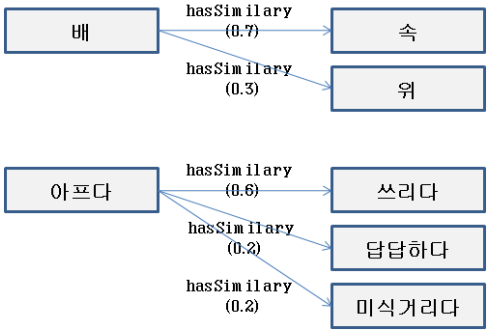
예를 들어 병증에 대한 표현의 경우 "홍길동은 배가 아프고 열이 난다"에서 홍길동은 주어, 배는 목적어, 아프다는 술어, 열은 목적어, 난다는 술어로 가정될 수 있다. 기존에 RDF나 OWL의 표현은 다음 〈그림 3〉과 같은 표현 방법을 적용한다. 이같은 표현을 살펴보면 한 개체에 적용되는 두 개의 술어 '아프다'와 '난다'와 같은 경우는 증상을 나타내는 동사로 구분이 될

수 있으며 진단을 위한 문장의 경우 사용자가 표현한 '~는, ~이, ~가'와 같은 조사에 따라 의미의 결정이 달라지게 되므로 의미 결정을 위해서는 〈그림 3〉과 같은 표현은 많은 제약을 갖는다.



〈그림 3〉 병증의 표현

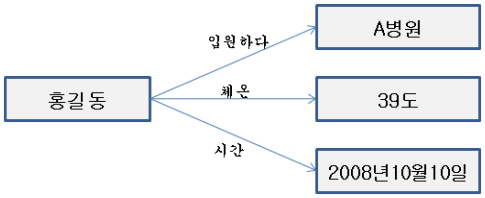
문장 '배가 아프다'와 '속이 쓰리다'는 현실에서 유사한 의미를 갖는 문장이다. 그러나 현재 대부분 온톨로지 모델에서 속성 관계의 가중치나 값을 표현하는데 〈그림 4〉와 같은 그래프로 표현된다.



〈그림 4〉 유사 단어의 연결

이같은 문장의 표현에 있어 속성 'hasSimilary' 아래 기술한 숫자에 따라 의미적 연관도가 결정된다고 할 때 속성에 이를 표현하고 각 객체로 연결하는 표현 방법은 매우 어렵다.

또한 시간이나 공간 그리고 사용자의 현재 상황을 표현한 문장은 '2008년10월10일 홍길동은 A병원에 입원했고 현재 체온은 39도이다'의 문장을 분석하면 사람, 시간, 병원, 체온과 같은 서로 다른 지식 영역의 단어들로 구성되어 있다. 이 문장을 이항 관계 속성을 이용해 표현하면 〈그림 5〉와 같다.

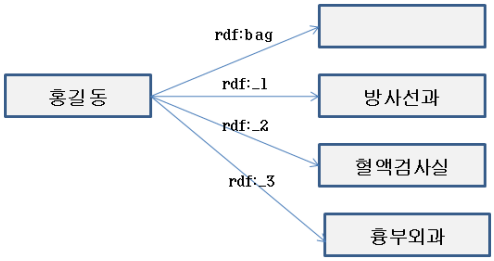


〈그림 5〉 홍길동의 다양한 상황

〈그림 5〉에서 홍길동의 공간은 'A병원' 시간은 '2008년10월10일', 상황은 '체온이 높다'이다. 그러나 상황에서 '체온'이란 단어는 다양한 경우로 변경될 수 있다. 또한 이 문장은 현재까지 RDF Reification 방법을 이용하나 이 방법은 다음과 같은 이유로 사용을 자제해야 한다.

1) RDF Reification 어휘는 statement을 언급하기 위해 설계 되었다.
2) RDF Reification은 트리플에 대한 추가적인 정보를 넣는데 포함하나 인수의 처리에 대한 추가적인 정보를 공급하지 않는다.

마지막으로 '홍길동은 진단를 위해 방사선과, 혈액검사실, 흉부외과를 들려야 한다'는 문장에서 홍길동이 들려야하는 순서를 꼭 지켜야 한다고 할 때 RDF collection을 사용하여 〈그림 6〉과 같이 표현될 수 있다.



〈그림 6〉 RDF에서 순서 표현

2.2 질의 응답 시스템

정보검색 기술의 처리 결과로서 문서 혹은 문단과 같이 텍스트 덩어리 집합을 제공한다. 사용자는 이러한 집합을 통해서 포괄적인 정보를 얻기도 하고 필요한 답을 찾아내기도 한다. 이때 제공된 결과는 답으로 이뤄지는 것이 아니고 다시 문서의 내용을 사용자가 일일이 확인해야 하는 번거로움이 있다. 이에 반해 질의응답형 정보검색 기술은 사용자 질문의 의도를 세밀하게 파악한 후, 검색 대상 문서로부터 답을 찾아 제공하기 때문에 사용자가 일일이 제공된 텍스트 덩어리 집합을 재검색하는 번거로움이 줄어들게 된다.

2.2.1 질의 응답 시스템 개요

질의 응답 검색은 사용자의 자연어 질문과 검색 대상 문서의 의미를 파악하기 위한 고정밀 자연어처리 기술과 대상 문서로부터 답을 추출하기 위한 정보추출 기술을 필요로 하며, 많은 후보 문서들로부터 답을 포함하는 문서를 걸러주는 역할을 위해 기존의 문서검색 기술도 사용한다. 질의 응답 기술은 현재 정형 데이터베이스에 담겨 있는 상품 정보나 기업 정보를 찾아주는 데이터베이스 질의 응답과 게시판 혹은 전자메일 정보를 찾아주는 FAQ(Frequently Asked Question) 질의 응답 등에 실제로 활용되고 있다.

정보검색의 기술적 한계에 직면한 몇몇 검색 포털 업체들은 새로운 검색 서비스 모델로 "지식검색"이라는 개념을 도입하였는데, 2003년도부터 국내 검색시장의 대표적인 서비스로 자리매김하게 되었다. 지식검색은 알고리즘에 의존하는 기존 검색방법과 달리 포탈 사용자의 참여를 활용한다. 사용자의 질문에 다른 사용자가 답을 달고, 많은 사용자들이 이 지식의 유용성에 대해 평가함으로써, 네티즌들이 지식을 공유하는 수단으로 크게 각광을 받고 있다. 대부분의 국내 인터넷 포털 사이트는 지식검색을 중요한 서비스의 하나로 제공하고 있는데, 네이버의 '지식iN', 엠파스의 '지식거래소', 세이클럽의 '세이테마', 야후의 '야후! 지식검색' 등이 있다.

이러한 지식검색 서비스의 장점은 실생활에 도움이 되는 실용적인 질문과 답변이 많고, 정보의 시사성에 민감하다는 것이다. 이런 장점에도 불구하고 지식검색이 가지고 있는 문제점은 다음과 같다. 첫째, 정답을 정의하는 것이 명확하지가 않다. 즉 네티즌 다수가 인정하면 '정답'으로 채택되고, 질문자가 답변을 선택하게 되면 그것이 마치 실제 정답인 것처럼 왜곡될 수 있다. 둘째, 방대한 양의 질문과 답변 중에서 신변잡기적인 내용이 많으며, '퍼오기'를 거듭하면서 출처가 불분명해진 정보들은 지식검색의 신뢰도를 크게 떨어뜨린다. 셋째, 흥미를 끌기 힘든 주제의 질문에 대해서 충실한 답변을 기대하기 어렵다.

Ellen은 질문에 대한 정답을 사람이 제시하

는 지식검색과는 달리, 질의 응답 기술에서는 사용자의 질문에 대해 질의 응답 시스템이 문서로부터 정답을 추출하여 사용자에게 제시한다. 질의 응답시스템의 목적은 질문에 대해 이를 만족하는 문서의 리스트를 찾기보다는 정답을 추출하는 것이다. 이는 단순히 키워드에 기반을 둔 정보검색이 아니라, 사용자 질문의 정교한 분석, 검색 엔진을 통한 문서 및 정답후보 단락 추출, 단락에서 원하는 정답을 추출하는 과정으로 이루어져 있다. 또한, 질의에 대한 정답이 사실에 기반을 둔 정확한 정답인지를 검증하기 위한 방법도 연구되고 있다.

질의 응답 시스템과 일반 정보검색 시스템의 큰 차이점 중 하나는 자연어로 된 질문을 입력하여, 문서를 검색하는 것이 아니라 정답을 찾는 것에 있다. 이를 위해 질문의 처리과정에서 사용자가 원하는 정답이 무엇인지 질의의도를 파악할 수 있는 질의유형이나 키워드 등의 정보를 질의로부터 추출한다. 또한, 기존의 정보검색에 의해 질의와 유사한 문서를 추출하고, 문서에서 다시 정답을 포함할 가능성이 있는 단락을 추출한 후, 단락에서 질의유형과 동일한 개체를 찾아내어 사용자에게 정답으로 제시한다. 최근에는 단답(factoid) 뿐만 아니라 서술형 정답, 나열형 정답 등에 대한 연구도 활발히 진행 중에 있다.

일반적인 질의 응답 시스템의 다음과 같은 모듈로 구성되며 각 모듈의 역할은 다음과 같다.

- 질문분석: 질문이 요구하는 정답의 유형을 결정하고, 검색엔진을 위한 키워드 추출
- 문서검색: 키워드를 이용하여 관련 문서를 검색엔진을 통해 검색
- 단락검색: 검색된 문서에서 키워드와 관련이 있는 단락 검색
- 정답추출: 단락에서 질문이 요구하는 정답유형과 관련이 있는 개체를 정답으로 추출하여 사용자에게 제시

2.2.2 질의 응답 검색 조건

정보검색의 다양한 기술 및 알고리즘을 객관적으로 평가하기 위해 평가컬렉션을 개발하고 매년 참여하는 시스템을 새로운 데이터를 사용해서 평가한 후 그 결과를 공유하는 정보검색의 주요한 학술대회인 TREC QA의 기술로드맵에 따르면, 시스템이 제시해야 할 정답은 단순히 문장의 일부분이 아니라 정의나 설명문과 같은 서술형 문장(descriptive answer), 여러 문서로부터 추출된 다중 정답(list answer) 등을 포함하여야 하고 이러한 기능을 평가할 수 있도록 평가의 범위를 확장해 나갈 예정이다. 더 나아가 문장들 사이의 추론이 필요한 질의를 제시하는 기능, 주어진 정답에 대한 배경설명, 정답의 정당성 검증, 정답의 모호성 해결, 전문가 수준의 의견제시, 이질적 정보의 통합을 통한 정답의 제시 등 점점 질의응답 시스템의 난이도를 높여 갈 계획이다. 다음은 Q/A에서 가능한 질문의 형태들이다.

- "AMTRAK의 금융적인 상황은 어떠한가?"
- "주몽의 어머니는 누구인가?"
- "대한민국의 2002년도 대통령선거에서 선출된 사람은 누구인가?

이 같은 문장형 질의에 대해 TREC 2002는 정답을 포함한 문자열을 제시하는 것이 아니라 실제 정답을 제시하도록 요구하였고, CWS(Con-

fidence Weighted Score)를 시스템에 대한 새로운 평가 척도로 사용하였으며, 자연어가 갖고 있는 의미론에 입각한 처리의 중요성을 강조하였다.

2.2.3 주요 질의 응답 시스템

(1) FAQ 시스템

질의 응답 기술은 현재 인터넷 쇼핑몰에서 데이터베이스에 담겨있는 상품정보 등을 찾아주거나 고객의 소리와 같은 게시판 또는 FAQ 등을 검색하는데 부분적으로 실용화되고 있다. 문서검색이 문서로부터 필요한 정보를 찾는 것임에 반해, 데이터베이스에서 사용자가 원하는 자료를 검색하는 데이터베이스 자연어 질의처리 기술은 DB의 SQL과 같은 질의처리 언어를 자연어로 대신하도록 하는 것이다. 초창기의 대표적 DB를 위한 자연어 인터페이스로는 BASEBALL과 LIFER가 있다. BASEBALL은 야구 통계에 대한 질의처리 시스템으로 "Who did the Red Sox lose to on July 5?"와 같은 자연어 질문을 처리할 수 있다. 현재 이러한 DB 질의처리 기술이 사용자의 자연스러운 대화체 질의 문장을 형태소, 의미, 구문 분석하여 사용자의 요구에 맞는 최적의 상품을 검색해 주는 상품검색 및 추천 등에 이용되고 있다. 이러한 시스템은 텍스트에서 답을 찾아야 하는 근래의 질의 응답 시스템과는 기술적인 면에서 근본적으로 상이하다.

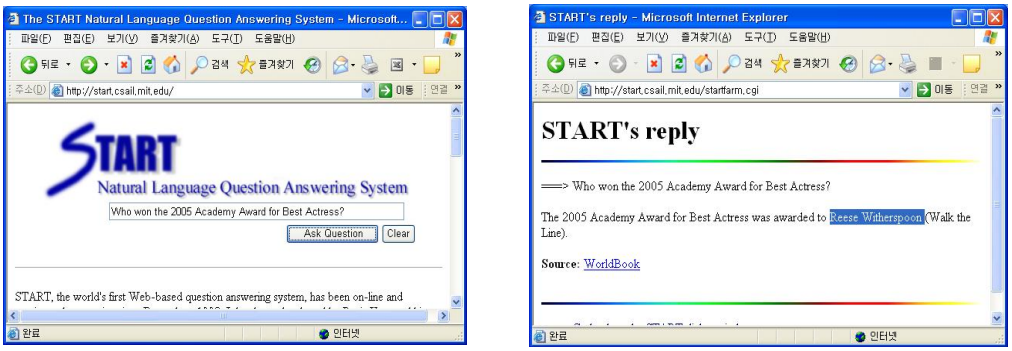
(2) START

START는 1993년부터 제공된 온라인 질의 응답 시스템으로 구조화 또는 반 구조화된 웹 데이터들을 통합한 가상의 데이터베이스인 Omnibase를 이용한 자연어 질의 응답 시스템이다. 이 시스템은 다수의 웹 사이트에 대한 단일한 질의 응답 인터페이스로 볼 수 있는데, World Factbook을 정보의 원천으로 하여 각국에 대한 정치, 경제, 지리적 정보에 대한 질의응답을 제공하며, Biograph.com 사이트로부터 25,000여명의 인물에 대한 정보를, Internet Movie Database 사이트로부터 수십만 편의 영화에 대한 전반적인 정보를, 그리고 Merriam-Webster 사전을 이용하여 각종 정의와 관련된 정보 등을 제공한다. 이 시스템은 반 구조화된 웹 페이지를 대상으로 한다는 면에서 근래의 질의 응답 시스템에 보다 가까이 접근해있다. 다음 〈그림 7〉은 START를 이용해 2005년도 아카데미 여우주연상에 대한 답을 구한 화면이다. 그러나 START의 경우도 영화의 'star'와 하늘의 'star'과 같은 동음이의어에 대한 정확한 답을 제공하지 못하는 단점이 있다.

(3) AnyQuestion 3.0

AnyQuestion은 ETRI에서 개발한 백과사전 대상의 질의 응답 시스템으로, 이 시스템은 '파스칼대백과'를 기반으로 14개 분야의 총 100,373 표제어에 관한 사용자 질문에 대해 질문의도를 파악하여 정답을 제시한다. AnyQuestion은 추출된 정답뿐 아니라 정답을 추출하게 된 근거가 되는 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. 주요한 특징은 다음과 같다.

- 사용자 질문 의도에 따른 다양한 정답 추출 및 제시



〈그림 7〉 START를 이용한 2005년도 아카데미 여우주연상 검색화면

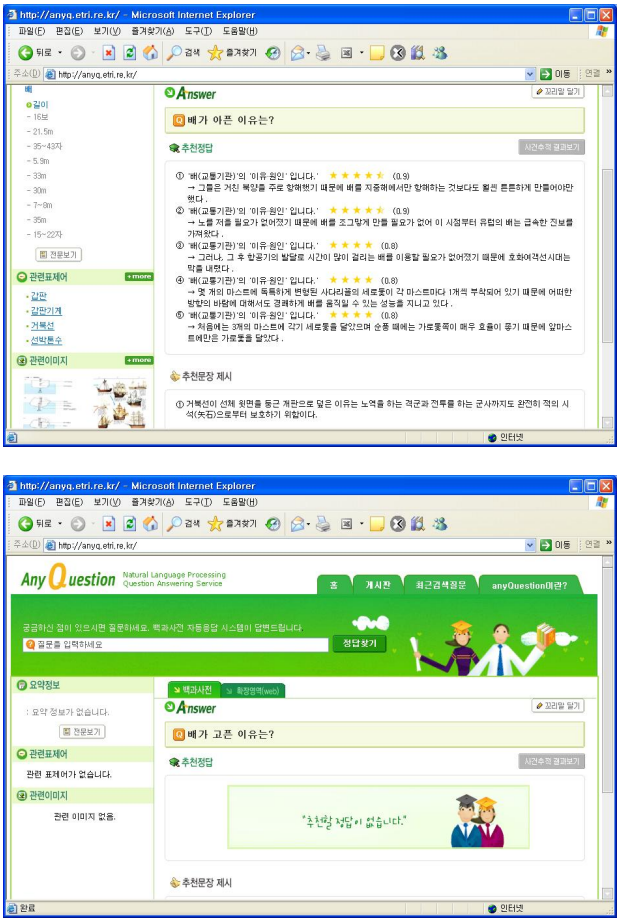
- 백과사전 지시베이스의 반자동구축
- 정답 관리자를 통한 복잡한 질문 해결 및 정답 필터링 기능

위 기능을 수행하기 위해서는 기본적으로 고정밀 언어 분석을 위한 고성능 형태소 분석기, 개체명 인식기, 구문분석기 등이 필요하다. 뿐만 아니라 백과사전 문장에서 공통된 지식을 추출하여 구조화된 지식베이스 추출을 위해서는 정보 추출(Information Extraction) 기법이 필요하고, 이를 학습하기 위한 기계학습 기술과 학습에 필요한 수작업을 줄이기 위한 반자동 구축 도구 등이 필요하다.

AnyQuestion은 일반적인 단답형 질의 응답뿐만 아니라 서술형 질의에 대한 정답 및 나열형 정답도 제시하는데, 서술형 질의란 정답이 한 단어 즉 단답으로 제시할 수 없는 경우로 예를 들면 "가을 하늘은 왜 높은가요?", "비타민 C가 많은 과일은 무엇인가요?"와 같은 질문이 그것이다. 서술형 정답을 추출하기 위해서는 서술형 문장에 나타나는 구문 및 의미 패턴을 인식하고 이를 적용하는 패턴 매치 기술이 필요하다. 나열형 정답을 추출하기 위해서는 병렬형 구문을 분석하여 생략된 구문을 파악하고, 여러 문서에 중복으로 나타난 정답을 통합 관리하는 기술이 필요하다. 다음 〈그림 8〉은 사용자 질문에 대한 AnyQuestion의 화면으로, "배"라는 단어가 "사람의 배"와 "운송 수단의 배", "먹는 배" 등의 다중의미가 존재하는데도 불구하고 이에 대한 적절한 해답을 제공하지 못하고 있다. 이는 AnyQuestion 또한 단어나 문장이 갖는 중의성 또는 유사성에 대한 의미도 결정이 부족하기 때문이다.

2.3 의미 유사도 관련 연구

P.Selvi는 용어들 의미론적 유사도와 거리를 측정하기 위한 방법으로 개념의 정보 내용과 코퍼스로 주어진 지식 베이스의 정보 내용을 추출한 Comm Spec에 패스 거리(path length)를 이용한 유사도 측정 방법을 제안하였고, Zahra는 유사도 계산과 OWL 온톨로지 정렬을 위해 벡터 공간 모델을 제안하였으나 개념들 간의 유사도를 matrix 기반으로 계산함으로써 필요 이상의 계산이 요구되는 경우가 있다. 특히 근래는 Vector Space를 이용하여 온톨로지 유사도



〈그림 8〉 AnyQuestion의 문장 검색

를 계산하는 연구가 많이 진행되며 Ce Zhang, Rajesh는 온톨로지 정렬과 온톨로지 정합을 위해 Vector space를 사용하였다. 안우식은 OWL 속성을 이용하여 온톨로지 간의 의미 유사도 측정 방법을 제안하였다. 두 엔티티가 가지는 오브젝트의 집합에 각 술어에 대한 부분 유사도를 적용하는 방법을 제안하였으나 술어에 대한 명확한 관련성의 명시와 일대일 관계만을 정의하여 한 개체가 갖는 다양한 개체와의 관련성을 명시하는데는 어려움이 있다.

3. 온톨로지 모델링

3.1 디자인을 위한 지식 영역 분석

본 연구는 의미 추론 규칙의 적용을 위해, 언어학적 특징 중 하나인 동음이의어에 관련된 문장들을 분석한 결과 단어 '배'는 결합 동사나 목적어에 따라 다양한 의미가 포함됨에 주목하였다. '배'는 ① 탈것의 배, ② 신체의 배 ③ 과일의 배 ④ 술이나 음료의 잔 ⑤ 허리를 굽히는

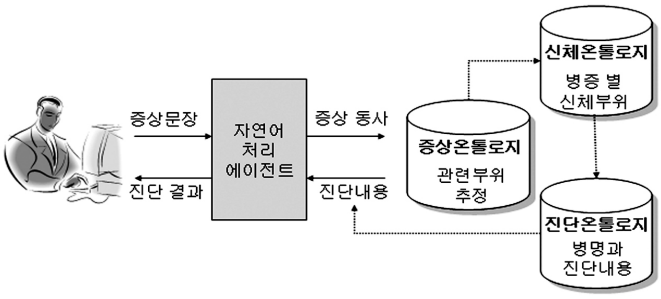
인사 ⑥ 기타 이름 성씨 등을 표현하며, 관련되는 동사(술어) 또한 '먹다', '타다', '갈다', '하다', '올리다', '자르다', '부서지다', '구입하다'.... 등등 상대 개체에 따라 결합되는 동사(술어)가 다양한 특징이 있다. 본 연구는 온톨로지의 개념간 의미 추론 규칙을 이용하여 사용자 정보 요구를 서술하기 위해 병의 다양한 증상에 대한 표현과 이에 대한 진단을 표현하는 영역으로 도메인을 한정하였다.

특히 문장형 사용자 질의에 대한 의미 결정 후 결과를 제공하기 위해 신체의 '배'와 관련된 병증을 중심으로 구문 규칙과 의미 추론 규칙을 연구한다. 다음 〈표 1〉은 '배'와 유의어인 '속'을 포함한 병증 표현을 정리한 것이다. 〈표 1〉의 의미에서 보여주듯이 '배'와 관련하여 유사한 증상과 의미가 다양하게 표현될 수 있음을 알 수 있다. 이같은 내용에 따라 증상표현에 따른 용어와 유사한 의미 그리고 관련 신체 부위가 다를 수 있다.

다음 〈그림 9〉는 각 온톨로지들간의 지식 정보 전달에 대한 개요를 나타낸 그림에서 사용자로부터 입력받은 증상 표현 문장은 시스템의 자연어처리 에이전트를 통해 증상에 관련된 동사들을 추출하고 추출된 동사들은 증상 온톨로지의 동의어나 다의어들을 추론 규칙에 의해 검색하고 검색된 결과는 온톨로지내에 정의된

〈표 1〉 병증표현과 그 의미

병증 표현	설 명
속이 거북하다	몸이 찌뿌드드하고 괴로워 움직임이 자연스럽지 못하거나 자유롭지 못함
속이 아프다	몸의 어느 부분이 다치거나 맞거나 자극을 받아 괴로움을 느끼다.
속이 더부룩하다	소화가 잘 안되어 배 속이 거북하다.
속이 메스껍다	먹은 것이 되넘어 올 것같이 속이 몹시 울렁거리는 느낌이 있다.
메시껍다	[형용사]'메스껍다'의 잘못된 표현.
메시꼽다	[형용사]'메스껍다'의 잘못된 표현. 속이 울렁거리다
속이 매슥거리다	먹은 것이 되넘어 올 것같이 속이 몹시 울렁거리는 느낌이 있다.
속이 끓다	소화가 잘 안되어 배 속이 편안하지 아니하다
속이 니글니글하다	먹은 것이 내려가지 아니하여 곧 게울 듯이 속이 매우 메스껍다
속이 니울니울하다	먹은 것이 내려가지 아니하여 곧 게울 듯이 속이 매우 메스껍다
속이 늬울늬울하다	- 속이 메스껍다 - 라는 뜻의 제주도 사투리
속이 미싱그리다	[메스껍다][구토증세가 있는]의 의미로 사용되는 경상도 사투리
속이 울렁거리다	속이 메슥메슥하여 자꾸 토할 것 같아지다. ≒울렁대다.
속이 군성거리다	속이 메슥메슥하여 자꾸 토할 것 같아지다. ≒울렁대다.
속(이) 뒤집히다	비위가 상하여 욕지기가 날 듯하게 되다
설사하다	설사 상태의 변을 보다. ≒사하다·설하다
변비가 있다	정상적으로 배변이 이루어지지 않는다.
게우다	먹은 것을 삭이지 못하고 도로 입 밖으로 내어 놓다. ≒토하다
도라내다	≒ 게아내다. 먹은 것을 삭이지 못하고 도로 입 밖으로 내어 놓다.
체하다	소화가 잘 안되어 배 속이 편안하지 아니하다.
설사하다	설사 상태의 변을 보다
구토하다	먹은 음식물을 토하는 증상으로 구토에 시달리다와 같은 표현도 사용



〈그림 9〉 온톨로지간 지식 전달 체계 개요

신체 온톨로지의 신체부위와 함께 신체 온톨로지로 전달된다. 신체 온톨로지로 전달된 지식은 예를 들어, 미릉골통의 경우는 증상이 나타나는 부위가 머리(head)와 눈(eye)이다. 이때 신체 온톨로지는 증상 온톨로지로부터 전달받은 정보를 분석하여 병증별 정확한 신체부위를 추론하고 이 값을 다시 진단 온톨로지로 전달한다. 진단 온톨로지는 전달받은 값들을 추론하여 입력된 병증에 적합한 병명과 진단내용을 제공한다.

이러한 지식 전달 체계 중 "연관부위"에 대한 지식 전달은 다음과 같은 기술 논리로 정의할 수 있다.

연관부위(증상, 신체부위)
→ *relationPart*(미릉골통, 머리⋂눈)

즉, 병증은 신체부위의 연관된 부위를 가지며 미릉골통의 경우는 머리와 눈이 같이 아픈 증상이라는 표현이다.

증상 표현 용어와 유사어 의미와 관련 부위에 대한 설명은 〈표 2〉와 같이 정리될 수 있다. 정리된 내용을 기초로 예상되는 병증에 대한 질문과 진단은 〈표 3〉에 정리하였다.

〈표 3〉의 진단 결과는 구현될 시스템의 성능을 검증하는데도 사용되며 온톨로지의 지식 내용을 지속적으로 관리할 수 있는 문서 역할을 하게 된다.

3.2 새로운 N-ary 디자인 패턴

2.1절에서 고찰한 온톨로지 표현의 이항관계의 어려움이나 문제점을 해결하기 위해 본 절에서는 새로운 N-ary 디자인 패턴을 적용한 온톨로지 모델 방법을 기술한다.

(1) 관계에 대한 중개자 클래스 도입

"홍길동은 배가 아프고 열이 난다"와 같은 문장의 경우 〈그림 10〉과 같이 중개자(mediator) 역할을 하는 클래스를 이용하여 한 개체에 관계하는 다양한 관계와 개체를 관련지을 수 있다. 〈그림 10〉에서 Symptoms_Relation 클래스는 중개자 역할을 하는 클래스로 신체 부위나 증상 등 사람(환자)의 증상을 더욱 다양하게 표현하는 방법을 제공할 수 있다.

〈그림 11〉은 〈그림 10〉의 모델을 OWL에서 정의한 코드이다. 이때 두 속성 symptomsPart와 symptomsValue는 함수적으로 정의된다.

〈표 2〉 증상 표현 용어와 유사어 등의 정의

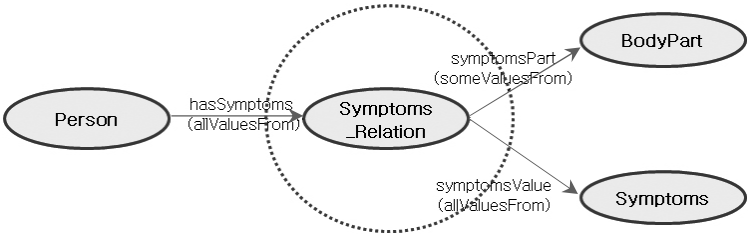
상위 증상	용어	유사어	의 미	관련부위
아픔	아프다	통, 통증	몸이 어떤 부분에 괴로운 느낌을 받는 상태	배(유사 = 속 = 강자) (하위 = 위, 장) 머리(하위 = 얼굴)
소화	거북하다	더부룩하다 소화가 안된다	몸이 찌뿌드드하고 괴로워 움직임이 자연스럽지 못하거나 자유롭지 못하다.	배(유사 = 속)
소화	더부룩하다	거북하다	소화가 잘 안되어 배 속이 거북	배(유사 = 속)
소화	메스껍다	메스껍다 메시꼽다 매슥거리다 울렁거리다	먹은 것이 되넘어 올 것 같이 속이 몹시 울렁거리는 느낌이 있다	배(유사 = 속)
소화	끓다	소화가 안된다	소화가 안되어 배 속이 편안하지 아니하다	배(유사 = 속)
소화	니글니글	니울니울 늬울늬울	먹은 것이 내려가지 아니하고 곧 게울 듯 속이 매우 메스껍다.	배(유사 = 속)
소화	늬울늬울	메스껍다	메스껍다의 제주도 사투리	배(유사 = 속)
소화	미싱그리다	메스껍다	메스껍다의 경상도 사투리	배(유사 = 속)
소화	뒤집히다	울렁거리다	비위가 상하여 속이 편안하지 아니하다.	배(유사 = 속)
복통	설사하다	사리 설리 설하	변에 포함된 수분의 양이 많아져서 액상으로 된 변을 보다	배(하위 = 장 = 대장, 소장)
복통/ 배앓이 소화	구토	토하다 도라내다 게아내다	먹은 음식을 토함 먹은 것을 삭이지 못하고 도로 입 밖으로 내어 놓다.	배(유사 = 속) (하위 = 위)

〈표 3〉 진단 결과

배(복통)에 관련된 증상과 진단 결과	
증 상	진단 결과
1. 배(속)가 답하다.	소화불량 ,따뜻한 차로 몸을 따뜻하게 해주고, 가벼운 운동이나 반신욕과 족욕도 좋다
2. 배(속)가 미식거리고 아프다.	복통, 복통약을 먹는다
3. 울렁거리고 메스껍다	구토, 한방약으로는 반하가 있다
머리(두통)에 관련된 증상 문장	
1. 머리가 지끈거리다.	두통, 편두통,
2. 골이 아프다	두통, 두통약을 먹는다.

(2) 속성의 가중치나 확률의 표현

"배가 아프다"와 "속이 답답하다"와 같은 문장의 유사함을 표현하기 위해 '배'는 '아프다'와 '타다'가 동일한 속성명인 'hasSimilarity'에 속성의 주석으로 가중치를 부여하는데 단순히 '아프다'나 '타다'가 어느 단어와 관련하여 그같은 가중치를 갖는지에 대한 명시적인 표현이 없기 때문에 유사도의 결정의 애매함을 갖게



〈그림 10〉 중개자를 이용한 n-ary 관계 표현 모델

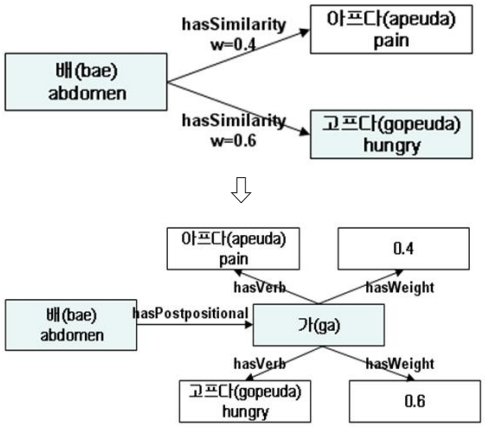
된다. 그러나 새로 제안한 n-ary 관계는 '배'와 관련된 조사들 중 '~가'가 연결될 때 hasVerb 속성은 '먹다'이고 hasWeight 속성은 0.9가 됨을 표현할 수 있다.

```
Symptoms_Relation
  a      owl:Class ;
  rdfs:subClassOf
    [ a       owl:Restriction ;
      owl:someValuesFrom:BodyPart;
      owl:onProperty:symptomsPart
    ] ;
  rdfs:subClassOf
    [ a       owl:Restriction ;
      owl:allValuesFrom:Symptoms ;
      owl:onProperty
            :symptomsValue
    ] .
:Person
  a      owl:Class ;
  rdfs:subClassOf
    [ a       owl:Restriction ;
      owl:allValuesFrom
         :Symptoms_Relation ;
      owl:onProperty
         :hasSymptoms
    ] .
```

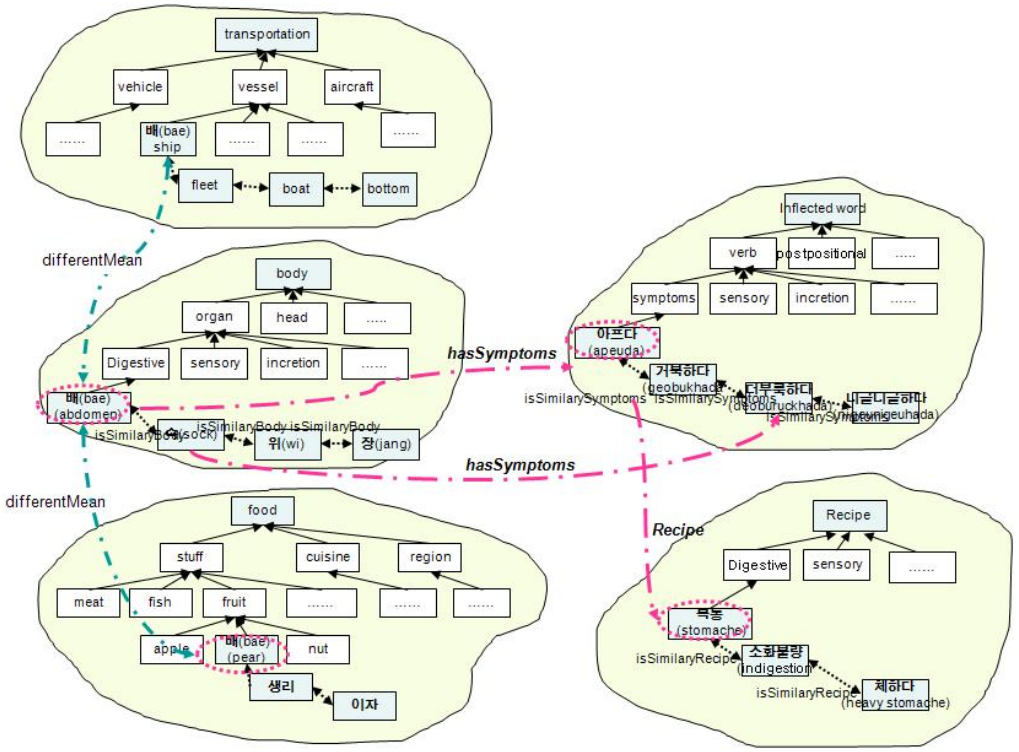
〈그림 11〉 중개자 코드의 정의

속성간 가중치나 확률과 같은 수치 표현은 〈그림 10〉의 클래스 모델과 유사하게 적용할 수 있다. 특히 사용자가 입력하는 한국어 문장의 경우 "주어+조사+목적어+조사+동사"와 같은 구조가 일반적인 형태이다. 이 때 '조사'는 다음에 연결될 목적어를 결정하는 중요 요인이다. 만약 조사에 연결될 목적어가 다른 지식 영역으로 구분될때 조사에 따라 분기될 수 있도록 목적어나 동사에 가중치를 부여한다면 문장의 의미적 결정이 더욱 정확해질 수 있다.

〈그림 13〉은 동음이의어와 이음동의어 등의 특징이 현재 제안한 중개자 클래스를 기반으로 다양한 지식 영역으로 구분될 수 있음을 표현한 그림이다. '배'는 앞서 기술한 바와 같이 '탈 것', '먹을 것', '신체' 등의 지식 영역으로 구분되고 이는 문장에 연결되는 '조사'에 따라 이에 적합한 동사나 활용어로 연결될 수 있음을 화살표로 표현한 것이다.



〈그림 12〉 '조사' 관계를 이용한 문장 구성과 결정



〈그림 13〉 동음이의어와 이음동의어의 다중 지식 영역의 연결

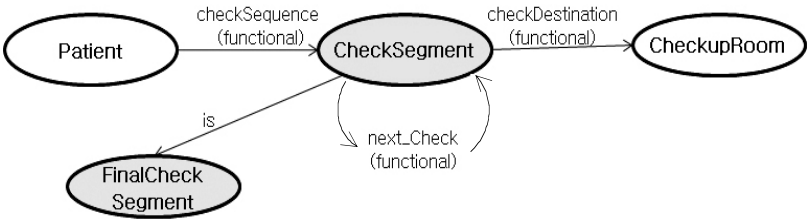
(3) 관계 리스트 표현

"홍길동은 진단를 위해 방사선과, 혈액검사실, 흉부외과를 들려야 한다"는 표현에서 순서가 중요하다고 한다면 여기에서 속성은 홍길동과 홍길동이 방문해야 하는 진료 순서를 차례대로 명시하는 것이다. 이 관계는 인수가 다른 많은 수들 간의 관계를 유지해야 하고, 각 진료실에 관련된 환자를 구분하는 속성의 집합으로 분리하는 자연스러운 방법으로 표현해야 한다.

이를 위해 어떤 관계에 따라 순서로 이 인수들을 연결하는 것과 이 순서에 있어 첫 번째 개체를 연결시키는 것이 중요하다. 〈그림 14〉는 검사 구간(Check Segment) 클래스의 인스턴스들 간에 순서화된 관계를 표현하는 것이다. 각 검사는 그 구간의 진료실에 대한 속성을 갖는다. nextCheck 속성에 최대 카디널리티 0으로 검사 구간에 대한 고유의 서브클래스인 FinalCheckSegment를 추가하는 것이 검사 순서의 마지막을 나타낼 수 있다. 〈그림 15〉는 〈그림 14〉의 OWL 순서를 이용하여 표현한 OWL 온톨로지의 일부이다.

3.3 단계적 추론 공리와 규칙 표현

〈표 4〉은 조사에 의해 증상과 신체 부위 그리고 진단에 관련된 용어들의 유사어들에 대한 관계를 정의하는 속성 관계 정의이다. 〈표 2〉의 선언적 형식 중 '배'와 '속'이 유사한 관계임을



〈그림 14〉 CheckSegment를 이용한 OWL 순서의 표현

```
:Patient
   a owl:Class .
:check_sequence
   a owl:ObjectProperty,
         owl:FunctionalProperty ;
   rdfs:domain:Patient ;
   rdfs:range:CheckSegment .
:CheckSegment
   a owl:Class ;
   rdfs:subClassOf owl:Thing ;
   rdfs:subClassOf
         [ a owl:Restriction ;
            owl:cardinality "1";
            owl:onProperty:destination
         ] ;
   rdfs:subClassOf
          [ a owl:Restriction ;
             owl:allValuesFrom:Airport ;
             owl:onProperty:destination
          ] .
   :next_checkt
      a owl:ObjectProperty ,
         owl:FunctionalProperty ;
      rdfs:domain:FlightSegment ;
       rdfs:range:FlightSegment .

:FinalCheckSegment
   a owl:Class ;
   rdfs:subClassOf:CheckSegment ;
   rdfs:subClassOf
         [ a owl:Restriction ;
            owl:maxCardinality "0";
            owl:onProperty:next_check
          ] .
   :CheckRoom
       a owl:Class .
   :destination
      a owl:ObjectProperty , owl:FunctionalProperty ;
      rdfs:domain:CheckSegment .
```

〈그림 15〉 OWL 순서 클래스 표현

선언한 규칙은 [정의 1]과 같다. [정의 1]의 이행적 속성을 이용하여 다음과 같이 [정의 2]와 [정의 3] 같은 규칙을 정의할 수 있다.

모든 증상은 신체의 부위와 관련되어 있다. 따라서 사용자가 질의문을 '배가 메스껍다'로 서술하였을 때나 '속이 메슥거리다'를 하였을 때도 유사한 증상으로 결정하여 동일한 진단을 하여야 하므로, 먼저 [정의 4]로 신체부위와 관련 증상을 정의하였다. [정의 4]의 선언적 규칙은 〈표 4〉에서 기술한 isSimDescription을 통해 [정의 5]와 같이 확장되고 '배가 메스껍다'는 문장은 '속이 메슥거리다'와 유사하다는 결정을 할 수 있다. 증상에 대한 진단을 위한 추론 규칙은 [정의 6]과 같다. 여기에서 "메슥거리다"는 증상 또는 다른 증상들을 'or' 조건으로 연결하여 조건이 성립하면 "서술", "진단"의 개념으로 연결된다. 사용자가 입력한 증상에 해당하는 각 신체 부위와 신체 부위를 종합하여 진단 온톨로지의 진단으로 추론하는 과정을 '미릉골통'을 예로 들어 설명하면 다음과 같다. 미릉골통의 증상은 "머리가 아프며 동시에 눈이 콕콕 찌르거나 쑤시는 듯 한 통증"을 나타낸다. 이때 신체 부위는 "머리", "눈"이 되며 증상은 "아프다", "찌르다", "쑤시다" 등이 된다. 이를 기준으로 한 진단 추론 규칙은 [정의 7]에서 제시하였다.

[정의 1]	*isSimBody(신체:배, 신체:속) ....*
[정의 2]	*isSimBody(신체:배, 신체:속) ∧ isSimBody(신체:속, 신체:위) → isSimBody(신체:배, 신체:위)..*
[정의 3]	*isSimSymptoms(증상:기침, 증상:해수) ∧ isSimSymptoms(증상:해수, 증상:기츰) → isSimSymptoms(증상:기침, 증상:기츰)*
[정의 4]	*hasRelationSim(신체:배, 증상:메스껍다)*
[정의 5]	*hasRelationSim(신체:배, 증상:메스껍다) ∧ isSimBody(신체:배, 신체:속) ∧ isSimSymptoms(증상:메스껍다, 증상:메슥거리다) hasRelationSim(신체:속, 증상:메슥거리다) ∨ isSimDescription(서술$_1$, 서술$_2$)*
[정의 6]	*hasRelationSim(신체:속, 증상:메슥거리다) ∨ isSimDescription(서술$_1$, 서술$_2$) → hasDiagnose(서술, 진단)*
[정의 7]	*hasRelationSim(증상.아프다, 신체.머리) ∧ hasRelationSim(증상.쑤시다, 신체.눈) ∨ hasRelationSim(증상.찌르다, 신체.눈) → hasDiagonis(증상.아프다, 진단.미릉골통)*

〈표 4〉 유사 의미 관계의 속성 정의

선언적 형식	Domain	Range	의 미
isSimBody	신체	신체 증상	신체에 대한 유사한 표현
isSimSymptoms	증상	증상	증상에 대한 유사한 표현
hasRelationSim	신체	증상	신체 부위별 관련 증상
isSimDescription	서술문$_1$	서술문$_1$	서술문$_1$과 서술문$_2$는 유사한 표현
hasDiagnose	서술문	진단문	서술문에 해당되는 진단문

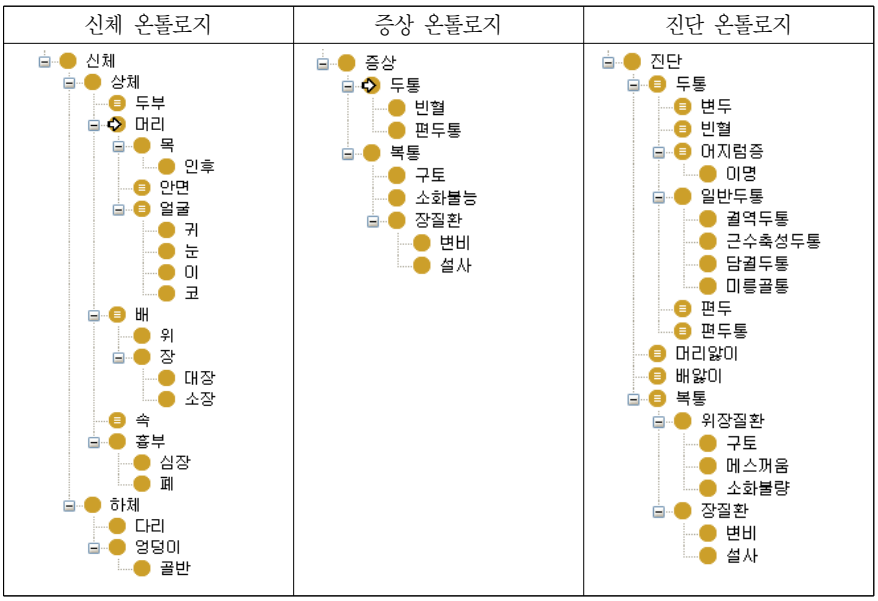
3.4 온톨로지 구축

정의된 용어와 동의어를 바탕으로 개념(클래스)을 정의하고 개념 계층화를 진행한다. 이에 따라 작성된 개념과 개념간 계층은 〈그림 15〉와 같다.

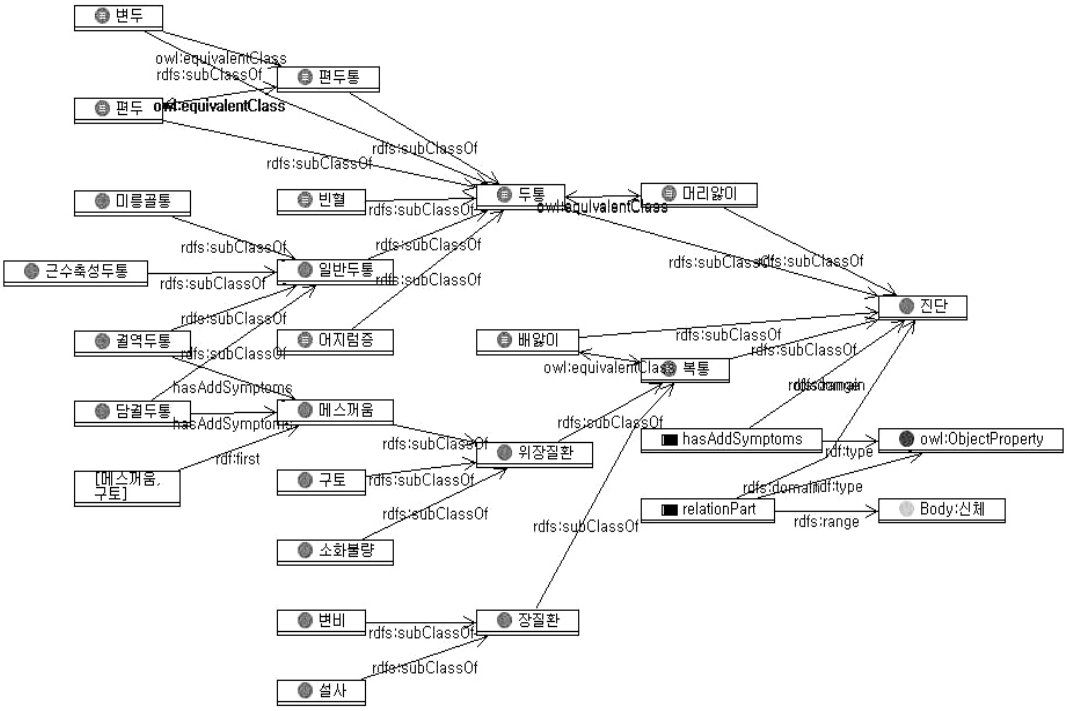
〈그림 15〉에서 증상 온톨로지의 경우 진단 온톨로지의 개념과 많은 차이가 없는 이유는 증상 온톨로지에 대한 표현은 대부분 동사로 이뤄진다. 이같은 동사들을 개념으로 정의하는 것은 온톨로지의 특징상 무리가 있으므로 증상을 구체화하는 동사들은 대략 증상에 해당하는 개념의 인스턴스로 처리한다. 예를 들어 빈혈에는 "뱅뱅거리다", "뱅뱅대다", "어지럽다", "희끗희끗하다", "어질어질하다" 등의 동사가 인스턴스로 표현된다.

온톨로지 표현 구축 과정에 따라 완성된 각 온톨로지들의 그래프중 〈그림 16〉은 진단 온톨로지를, 〈그림 17〉은 증상 온톨로지를 나타낸 그림이다.

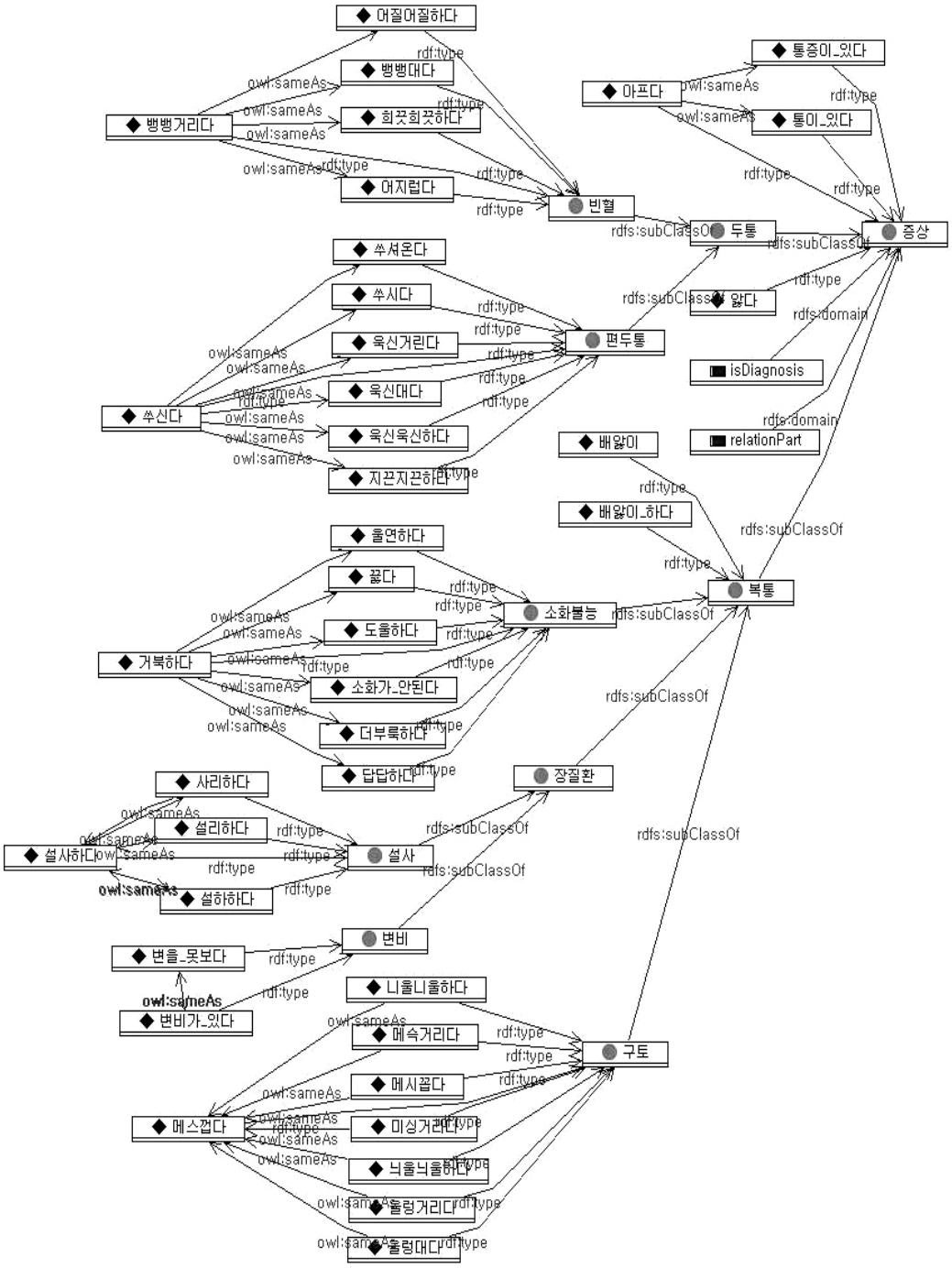
〈그림 18〉은 〈그림 10〉에 의해 제안된 n-ary 관계로 표현된 온톨로지의 일부이다. '조사'를 중개자 클래스로 도입하여 단어 '배'가 증상 온톨로지에 있는 '아프다'로 연결되며 '아프다'는 hasWeight 속성을 통해 가중치를 표현하였다. 또한 〈그림 19〉는 SPARQL을 이용해 〈그림 18〉에 표현된 관계를 추론한 결과로 '배'는 활용어 온톨로지의 '아프다'는 동사와 연결될 때 가중치는 '0.8'임을 단계적 추론(cascade inference)을 통해 추론할 수 있음을 보여주는 화면이다.



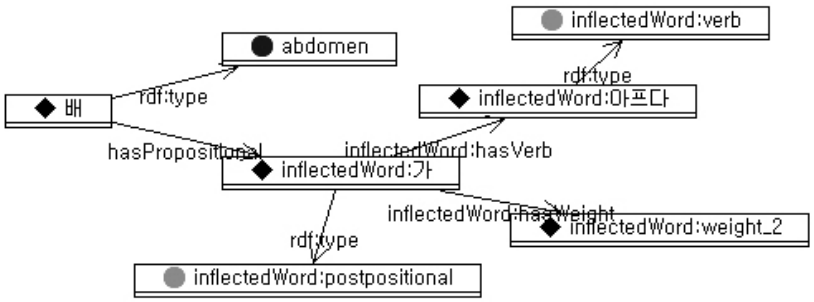
〈그림 15〉 주요 구축 온톨로지 클래스 계층



〈그림 16〉 진단 온톨로지 그래프



〈그림 17〉 증상 온톨로지 그래프



〈그림 18〉 새로운 중개자 클래스를 이용한 각 개체의 연결

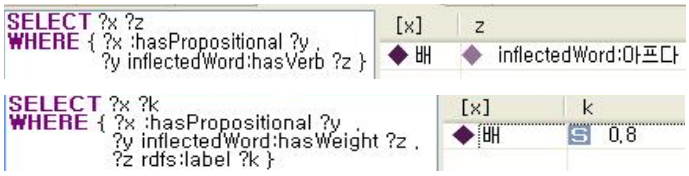
4. 실험 및 검증

4.1 실험 시스템 설계

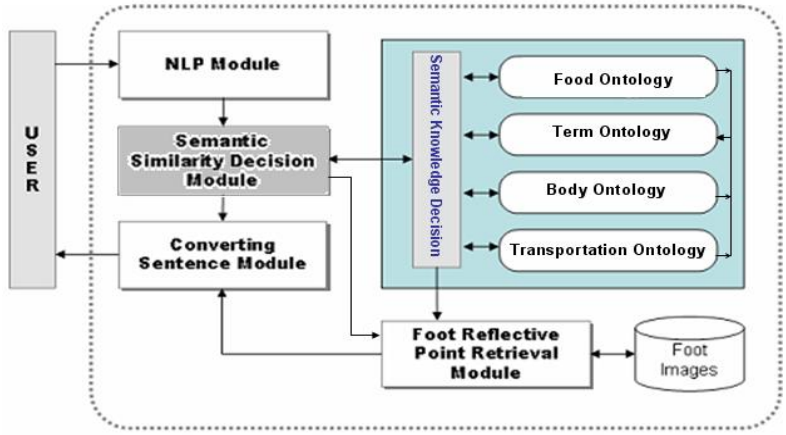
제안된 방법의 실험을 위해 본 연구는 사용자가 서술한 병증에 대한 문장을 진단하여 그에 대한 처방을 출력하는 실험시스템을 구현하였다. 실험시스템은 사용자로부터 입력된 질의 문장을 파싱하는 "자연어 처리 모듈"과 추론 규칙과 벡터 모델에 의해 온톨로지내 유사 의미들을 추출하는 "추론 엔진" 그리고 각 온톨로지 간의 지식 전달 모듈과 결정된 진단에 따른 "발 반사구 이미지 검색 모듈", 마지막으로 온톨로지에서 전달 받은 진단 내용을 사용자가 이해할 수 있는 문장 형태로 변환하는 "진단 문장 제공 모듈"로 구성된다. 실험 시스템의 구조는 〈그림 20〉과 같다. 〈그림 21〉에서 사용자로부터 입력된 증상 문장을 처리하기 위한 "자연어 처리 모듈"의 구문 분석 과정은 다음과 같다.

(1) 일반적인 증상문 처리

- ▸ 속이 답답하다 ☞ 속→신체＃속≒배, 답답하다→증상＃답답하다≒.
  - ◦ 속이: 속→신체＃속≒배(최종단어)
  - ◦ 답답하다: 답→단어 없음, 답답→단어 없음, 답답하→단어 없음, 답답하다→증상＃답답하다(최종단어)
- ▸ 머리가 아프다 ☞ 머리→신체＃머리, 아프다→증상＃아프다



〈그림 19〉 '배'의 연관 동사와 가중치 추론



〈그림 20〉 실험 시스템 구조

◦ 머리가: 머→단어 없음, 머리→신체＃머리(최종단어)
◦ 아프다: 아→단어 없음, 아프→단어 없음, 아프다→증상＃아프다(최종단어)
▸ 답답하다 ☞ 증상＃답답하다(최종단어)

(2) 유의어 처리
▸ 속이 도울하다 ☞ 신체＃속, 증상＃도울하다≒답답하다(최종단어)
▸ 속이 답답해 ☞ 신체＃속, 증상＃답답하다

(3) 모호한 의미를 가진 증상문
▸ 위가 쓰리다 ☞ 신체＃배＃위, 증상＃아프다＃쓰리다
▸ 눈 위가 아프다 ☞ 신체＃머리＃입, 위치＃위, 증상＃아프다

(4) 복합 증상문 처리
▸ 머리가 아프고 눈이 쑤신다 ☞ 신체＃머리 & 신체＃머리＃눈, 증상＃아프다 & 증상＃쑤신다
▸ 속이 쓰리고 가슴이 답답해 ☞ 신체＃배≒속 & 신체＃상체＃가슴, 증상＃쓰리다 & 증상＃답답해

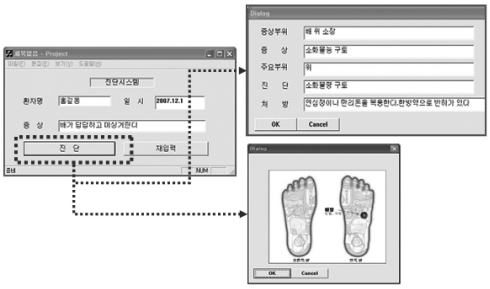
이 같은 원칙에 따라 입력된 증상 문장들은 파싱되고 파싱된 단어는 증상을 저장하는 연결리스트, 신체 부위를 저장하는 연결리스트 자료구조로 분류되어 추론엔진으로 전달된다. 추론엔진은 먼저 증상 온톨로지의 인스턴스들을 비교하고 인스턴스의 포함관계를 추론하여 일반적인 부위를 정한다.

결정된 부위와 전달된 신체 연결리스트의 데이터를 이용하여 *hasRelationSim* 규칙을 이용하여 정확한 신체 부위를 결정한다. 결정된 신체 부위 정보는 다시 진단 온톨로지로 전달되어 $isDiagonisis \wedge hasRelationSim$ 추론 규칙에 의해 진단 결과를 추출한다.

추출된 진단 지식은 다시 진단 지식 문장화 모듈로 전달되어 사용자가 이해할 수 있는 문장의 형태로 구성되며 이에 적합한 발 반사구의 위치를 나타내는 이미지를 검색하여 그 결과를 사용자에게 전달하게 된다.

4.2 실험 시스템 검증

실험 시스템에 다양한 증상 문장들을 입력하여 그에 대한 진단 결과를 통해 제안한 의미유사도 결정 방법의 성능을 실험하였다.

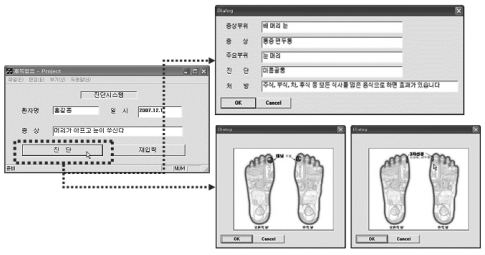


〈그림 21〉 일반 증상에 대한 진단

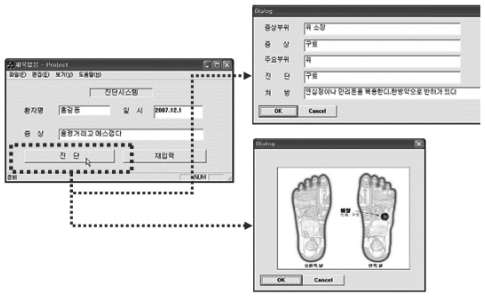
〈그림 21〉은 "배가 답답하고 미싱거린다"는 증상 문장을 입력한 결과로 진단 결과에 대한 자세한 설명과 설명을 기반으로 해당되는 발의 반사구의 위치가 나타난 이미지를 출력한다. 이 문장은 동사가 두 개이나 두 증상 모두 증상의 부위가 '배'와 관련된 것으로 추론되며 따라서 그 결과 "소화불량"이란 진단을 나타낸다.

〈그림 22〉는 "머리가 아프고 눈이 쑤신다"이다. 문장에 입력된 신체 부위는 '머리'와 '눈'이고 증상은 '아프다'와 '쑤신다'이다. 이때 '눈'은 머리의 하위 개념이며, '아프다'와 '쑤신다'는 모두 머리의 통증을 표현하는 동사로 많이 사용된다. 이 두 신체부위와 두 동사들을 만족하는 병증에 대한 추론과 제안된 유사도 결정 알고리즘에 의해 '미릉골통'이 진단되었고 해당하는 이미지를 보여준다.

〈그림 23〉은 "울렁거리고 메스껍다"는 문장이다. 이 문장은 신체부위에 해당하는 단어가 전혀 입력되지 않았다. 제안 시스템은 이러한 경우 추론 규칙에 있어 '울렁거리고'와 '메스껍다'가 거의 '복통'과 관련된 증상이라는 것을 추출하고 이를 신체부위의 배(속)에만 해당하는 증상으로 판단하여 구토에 해당하는 진단 결과를 보여준다.



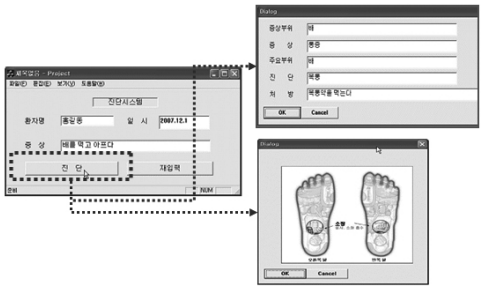
〈그림 22〉 복합 증상에 대한 진단



〈그림 23〉 신체 부위 추론 진단

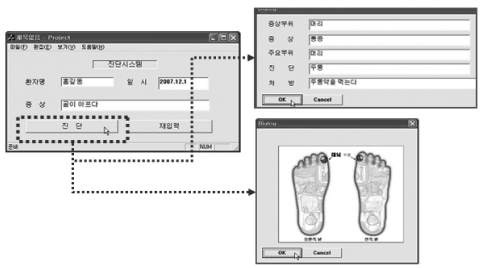
다음 〈그림 24〉는 동음이의어의 처리 예이다. 동음이의어는 같은 음절은 가진 단어가 다른 의미를 갖는 경우로 단어 '배'의 경우 '먹는 배', '사람의 배', '탈 것의 배'와 같은 다양한 의미가 존재하며 이의 정확한 의미 파악은 다음 단어와의 결합을 통해 판단할 수 있다. 다음 문장은 '배를 먹고 아프다'고 입력하였을 경우이다. 이때의 '배'는 '먹고'라는 단어로 인해 음식물로 결정하며, 다음 단어인 '아프다'로 인해 음

식물과 결합되어 통증을 일으킬 수 있는 신체 부위인 '배'를 결정하게 되고 그에 적합한 진단 결과를 제시한다.



〈그림 24〉 동음이의어 대한 진단

〈그림 25〉는 동의어나 유사어에 대한 처리이다. 장년층에서 '머리가 아프다'는 표현 대신 '골이 아프다'고 표현하거나 '배'라는 부위 대신 '속'이라는 표현도 많이 한다. 그리고 '메식꼽다'는 동사는 '미식거린다'를 비롯해 많은 유사어들이 있다. 이 같은 단어들의 특정을 포함한 '골이 쑤시고 속이 메식꼽다'는 문장이 입력되었을 때, '머리'≒'두부'≒'골'과 '배'≒'속'은 '[정의 4]'의 선언적 규칙에 의해, '메식꼽다'≒'미식거린다'는 '[정의 5]'의 선언적 규칙과 제안된 유사도 결정 알고리즘에 의해 적절한 유사어를 결정하여 진단 결과를 제공한다.



〈그림 25〉 유사어에 대한 진단

다양한 경우에 따른 증상 문장의 처리 결과를 살펴본 결과 본 실험 시스템은 온톨로지내에서 정의되고 표현된 지식이나 규칙들과 유사도 결정 알고리즘에 의해 적합한 의미를 결정하여 그에 따른 진단 결과를 보여주었다. 사용자 정보 요구 서술에 있어 제안된 의미 유사도 결정 방법이 정보 검색에 있어 문장의 의미 유사성을 결정하는데 좋은 방법임을 증명하였다.

5. 결 론

본 연구는 현재 지식베이스를 활용하는 지식기반시스템이나 정보시스템 또는 진단시스템의 경우 검색어의 경우 문장형태의 질의어에 대한 완전한 처리를 제공하기 위해서는 문장의 의미에 대한 정확한 결정이 매우 중요하다. 그러나 현재 대부분의 시스템은 문장의 의미에 따른 검색 결과를 제공은 어려운 실정이다. 이는 지식기반시스템이나 정보검색시스템에 있어 지식베이스를 구축할 때 모델화한 정보 내용을 사용자가 빈약하게 정의하거나 내재적으로 다의어(polysemy) 같은 광범위한 정의를 가졌을 경우, 특정한 단어를 이용한 검색을 하고 있을때 연관 페이지가 의미적으로 유사한 동의어(synonym)를 사용하는 경우 정확한 검색 결과의 제공은 더욱 어렵기 때문이다.

이러한 문제를 해결하기 위해 본 연구는 지식베이스로서 개념간 속성관계를 통해 의미를 추론할 수 있는 온톨로지를 적용하고자 한다. 한국어 문장이 갖고 있는 특징이나 어려움을 해결하기 위한 가장 기본적인 방법으로 앞에서 제안한 새로운 N-ary 관계 디자인 패턴과 모델

-이론 의미론에 따른 지식 표현에 적용하여 OWL DL과 OWL Full이 갖고 있는 다양한 어휘를 활용하고 이를 기반으로 단계적 추론 공리와 의미 결정을 위한 SWRL 기반의 규칙 표현의 정의를 제공한다.

이같은 방법에 따라 제안된 연구 방법의 검증을 위해 사용자가 입력한 병증 문장을 제안된 방법에 따라 다양한 지식 영역에 대해 온톨로지를 구축하였다. 이를 위한 온톨로지 모델링에 있어 중개자(mediator) 클래스의 적용과 같은 새로운 N-ary 관계 디자인 패턴과, Model-Theoretic Semantic에서 제안한 어휘를 적용한 온톨로지 모델을 제시하였다. 지식 관계와 의미 결정을 위해 추론 표현 방법을 이용하여 병증의 의미를 결정하고 그에 따른 진단을 제공하는 실험 시스템을 구현하였고, 한국어가 갖고 있는 문장의 유의성, 모호성, 복합성의 특징을 보유한 증상문들의 실험 결과 의미 결정과 유사 의미 확장에 있어 우수한 성능을 보여주었다.

그러나 병증의 표현에 따른 진단 시스템 들과 본 연구에서 표현된 문장과 구조와 유사한 연구 시스템의 부재로 인해 본 연구와 비교 실험이 어려운 문제점이 있었으며, 더욱 폭 넓은 지식 영역에 대한 표현과 의미 결정을 위해서는 기존에 구축된 온톨로지들과의 결합이나 선언적 규칙에 대한 자동화된 정의 방법등의 연구가 요구되었고, 특히 다양한 문장들을 분석하여 동적으로 온톨로지내 지식에 대한 추론 규칙을 생성하고 정합할 수 있는 연구가 필요하다고 사료된다.

참 고 문 헌

노상규, 박진수. 2007. 인터넷 진화의 열쇠, 온톨로지. 『gods Toy business』, 94-104.

박영택, 최중민. 2006. 온톨로지 추론 개요와 연구 동향. 『한국정보과학회 학회지』, 24(4): 17-23.

안우식, 박정은, 오경환. 2006. OWL 속성을 이용한 온톨로지 간 의미 유사도 측정 방법, 『한국컴퓨터종합학술대회논문집』, 33(1)(B).

이현자, 심준호. 2005. Description Logic을 이용한 전자카타로그 온톨로지 모델링. 『한국정보과학회 논문지, 데이터베이스』, 32(2).

한국전산원. 2004. 『웹 환경에서의 지식교환/유통을 위한 지식표현 및 추론기술 연구』, 65-80.

Baader, F., D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, eds. 2003. "The Description Logic Handbook." Cambridge University Press,

Bernardo Cuenca Grau. 2008. "OWL 1.1 Web Ontology Language: Model-Theoretic Semantics." W3C Working Draft, 2.

Bright, M. W., A. R. Hurson, and S. H. Pakzad. 1994. "Automated resolution of semantic heterogeneity in multi-databases." *ACM Transaction on Database Systems*19: 212-253.

Ce Zhang, Yu-Jing Wang, Bin Cui, and Gao Cong. 2008. "Semantic Similarity Based on Compace Concept Ontology." WWW2008, April, Beijing, China.

Ellen M. Voorhees. 1999. "The TREC-8 Question Answering Track Report." In Proceedings of TREC-8.

G. Salton and M. E. Lesk. "Computer evaluation of indexing and text prcessing." *Journal of the ACM*, 15(1): 8-36, January, 1968.

G. Salton. "The SMART Retrieval System - Experiments in Automatic Document Processin." Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

Guarino, N., C. Masolo, and G. Verete. 1999. "Ontoseek: Content-based access to the web." *IEEE Intelligent Systems*, 3: 70-80.

Hollink, L., G. Schreiber, J. Wielemaker, and B. Wielinga. "Semantic Annotation of Image Collections." In Proceedings of Knowledge Capture - Knowledge Markup and Semantic Annotation Workshop, 2003.

Lee, J., M., Kim, and Y. Lee. 1993. "Information retrieval based on conceptual distance in is-a hierarchies." *Journal of Documentation*, 2: 188-207,

Natasha Noy, Alan Rector. 2006. "Defining N-ary Relations on the Semantic Web." W3C Working Group Note 12,

P. Selvi, and N. P. Gopalan. 2008. "The New Semantic Similarity Measure Using Ontology and Corpus." *The Icfai Journal of Computer Science*, Vol.2, No1., pp.29-37, Jan.

Rajesh Thiagarajan. 2008. "Computing Semantic Similarity Using Ontology." ISWC.

Resnik, P. 1999. "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language." *Journal of Artificial Intelligence Research*, 11: 95-130.

Sharon L. Greene, Susan J. Devlin, Philip E. Cannata, and Louis M. Gomez. "No IFs, ANDs, or Ors: A study of database querying." International Journal if Man-Machine Studies, 1990.

Staab, S., and R. Studer, eds. "Handbook on Ontologies." International Handbooks on Information Systems, Springer, 2004.

Zahra Eidoon, Nasser Yazdani, and Farhad Oroumchian. "Ontology Matching Using Vector Space." ECIR 2008, LNCS 4956, pp.472-481, 2008.