이벤트 탐색을 사용하는 일정 영역 질의 응답 시스템의 구현

장두성[**] 오종훈[+] 최기선[+]
[+] 전문용어언어공학연구센터, 첨단정보기술연구센터, 한국과학기술원
[*] 한국통신 연구개발본부
{dschang, rovellia, kschoi}@world.kaist.ac.kr

A Domain-Dependent Question-Answering System

Du-Seong Chang[**] Jong-Hun Oh[+] Key-Sun Choi[+]
[+] KORTERM, AITRC, KAIST
[*] R&D Group, Korea Telecom

요 약

본 논문에서는 한정된 영역을 대상으로 하는 질의응답 시스템에서 사용자의 질의를 해석하고 적당한 대답을 생성하기 위해 백과사전이나 일반사전 등과 같은 지식원에서 추출된 구조화된 지식을 사용하는 과정을 기술한다. 질의응답을 위하여 지식원은 그 단락의 의미에 따라 구조화되고 각 단락은 논리형식으로 변환되었으며, 논리형식 내 각 개체들은 사전 정의문에 따라 확장되었다. 이 구조화된 지식은 입력된 자연언어 질의문에서 질의의 의도를 추출하고, 질의에 포함되어 있는 지식에 의미속성을 부착하기 위해 사용된다. 지식원의 논리형식 변환을 위해 한국어의 논리형식이 도입되었으며, 사용된 지식원은 우리말 큰사전과 계몽백과사전의 30여개 질병정의문이다.

1 서론

다양하게 존재하는 정보에서 인간은 자신에게 필요한 정보만을 선택하여 지식화하고 사용한다. 유통되는 정보의 양이 많아짐에 따라 이들 정보들 중 자신에게 유용한 정보만을 판정하여 주는 정보 여과(information filtering) 기술, 정보를 내부의 지식으로 변환하여 주는 지식습득(knowledge acquisition) 기술, 구축된 지식에서 필요한 부분을 찾아 주는 지식 탐색(knowledge search) 기술 등의 자동화가 필요하게 된다.

본 논문에서는 이러한 지식 습득 및 탐색에 관한 연구의 일환으로 일정 영역을 대상으로 지식을 습득하고, 습득된 지식을 탐색하여 사용자와의 대화를 진행하는 지식습득/탐색 시스템의 프로토타입을 제시하고자 한다.

본 논문에서 구현하고자 하는 질의 응답 시스템은 한정된 영역을 대상으로 지식을 탐색하고자 하는 사용자의 질의를 해석하고, 사용자의 의도에 맞게 내부지식을 탐색하며, 질의에 적당한 대답을 생성한다. 이러한 과정을 위해 백과사전이나 일반사전 등과 같은 지식원에서 정보를 추출하여 구조화된 내부 지식으로 구축하였다.

2장에서는 이 일정 영역 질의응답 시스템의 구조

와 시스템 운용에 대해 기술하고, 3장에서는 내부 지식을 구조화하기 위해 도입된 지식 구조와 한국어 논리 형식(logical form)에 대해 기술한다. 질의응답을 위하여 지식원은 각 구절의 의미에 따라 구조화되고 각 구절은 논리 형식으로 변환되었으며, 논리 형식 내 각 개체들은 사전 정의문에 따라 확장되었다. 이 구조화된 지식은 입력된 자연언어 질의문에서 질의의 의도를 추출하고, 질의에 포함되어 있는 지식에 의미속성을 부착하기 위해 사용된다. 이러한 질의 분석과정은 4장에 기술하며, 질의응답과정은 5장에 기술한다.

2 시스템의 구성

지식탐색을 사용하는 질의응답 시스템의 구조는 그림1과 같다.

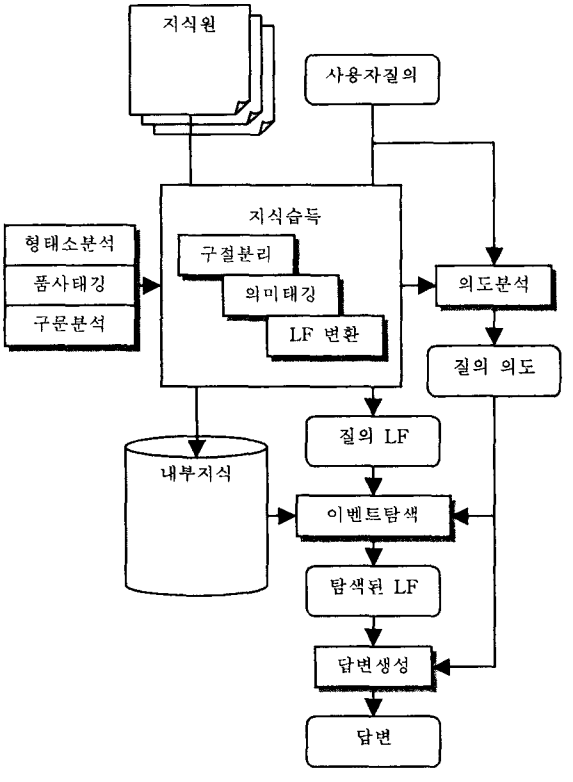

그림 1 시스템 구성도

지식원은 지식습득과정을 거쳐 내부지식으로 변환되며, 형태소 분석 및 구문분석을 사용하는 지식습득과정은 구절 분리, 의미속성 부착, 논리형식 변환을 포함한다. 여기에서 구절은 하나의 이벤트를 포함하는 문장의 한 부분으로 정의되며, 각 구절은 대상 영역에서 어떠한 의미를 가지고 있는지의 의미속성정보가 부착되며, 그 내용은 논리형식으로 변환되어 내부지식으로 저장된다.

사용자 질의에서는 그 질의가 의도하는 바가 추출되고, 그 질의 내용은 지식습득과정을 거쳐 논리형식으로 변환된다. 이 변환된 논리형식과 분석된 질의의도를 통해 사용자가 원하는 정보가 무엇인지를 예측하고 필요한 정보를 구축된 내부지식에서 탐색하여 적절한 대답을 생성한다.

3 내부지식의 구성

한정된 영역에서 습득된 지식은 내부지식으로 변환되며, 습득되는 지식은 두 가지의 정보를 담고 있다. 그 하나는 한정된 영역의 지식구조 그 자체이며, 나머지는 지식구조의 각 부분을 표현하는 지식 정보이다. 현재 설정된 영역은 질병 영역이며, 이 영역의 지식은 백과사전의 질병정의문과 일반사전의 단어 정의문을 통해 습득된다. 백과사전과 사전 정의문을 지식원으로 설정한 이유는 이들 문서가 일정한 영역에서 사용되는 정보를 가장 포괄적으로 포함하여, 다양한 정보를 요약하여 포함하고 있기 때문이다.

현재 질의응답 시스템을 구현하기 위해 사용된 지식원은 계몽백과사전에 정의된 34개 질병부분의 정의문이다. 초기 내부지식을 구축하기 위해 이 정의문을 수동 분석하여 각 단락이 포함하고 있는 질병에 관한 의미속성을 부착하였다. 사용된 의미속성은 질병의 정의, 증세, 원인, 치료방법 등으로 수동 분석된 질병정의문의 예는 표1과 같다.

소화 불량/질병명
소화기에 생기는 병/정의.
피로하거나 운동 부족일 때, 지나치게 많이 마시거나 먹었을 때, 소화가 안되는 음식물을 먹었을 때 등에 생긴다/원인
소화가 충분히 이루어지지 않게 되고, 식욕이 감퇴되며, 위가 아프거나, 구토 또는 설사를 일으킨다/증상.

표 1 의미속성이 수동 부착된 질병 정의문

이와 같이 의미속성이 부착된 질병 정의문은 다시 병렬어구 분리를 통해 각각 하나의 사건만을 포함하는 구절로 분리되었으며, 34개의 질병 정의문에서 생성된 구절은 291개로서, 77개의 원인, 136개의 증세, 44개의 치료방법을 포함하고 있다.

시스템의 지식 확장을 위하여, 논리형식에서 '발열'와 같이 명사의 형태로만 존재하는 사건은 사전을 참조하여 "열이 나다"와 같이 서술형의 사건으로 확장되었다. 또한 우리말 큰사전에서 "... 증세"나 "... 병"으로 정의문이 끝나는 단어를 질병명으로 추가하였다. 이렇게 추가된 단어는 1201개이다.

그림 2는 현재 구축된 질병 영역에서의 내부 지식의 구조 일부분이다. 지식의 구조정보는 각 노드별로 트리 구조로 표현되며, 각 노드는 하나의 질병을 의미한다. 또한 지식정보는 논리형식으로 변환되어 저장된다. 여기에서 사용되는 논리형식은 한국어 의존구조와 의미구조의 중간형태로 정의되며, 형태소 분석 및 구문분석을 거친 결과에서 대명사 참조해결, 병렬어구 분리, 구절 분리을 통해 논리형식 변환과정을 거쳐 자동 구축된다.

논리형식을 정의하기 위해서 의미구조를 전부 나타내기보다는 지식의 탐색과 질의응답에 보다 적합한 형태의 단순화된 논리형식을 고려하였다. 또한 구문 분석된 형태로부터 자동적으로 구축할 수 있도록 구문구조에서 직접 추출할 수 있는 형태의 논리형식을 고안하였다. 이러한 방법은 FALCON[5][6]과 같은 질의응답 시스템에서 사용된 방법이다.

현재 정의된 논리형식은 동사/형용사에 해당하는

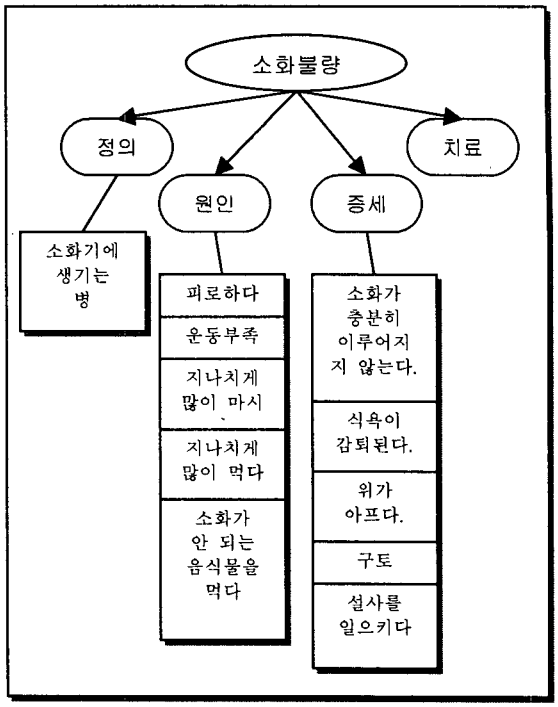

그림 2 내부 지식의 형태

사건형태와, 명사에 해당하는 개체형태, 사건형태와 개체형태간/사건형태간/객체형태간 관계를 나타내는 관계형태가 주요 요소이며, 한국어의 구문구조 특성상 서술격 조사와, '-하-', '-되-', 와 같은 조용사(혹은 보조어간)을 하나의 요소로 하는 조용형태, 그리고 대상영역에서 의미를 갖는다고 여겨지는 일부 관형사/부사에 대해 특성형태를 정의하여 총 5가지 종류의 논리요소를 정의하여 한정 영역에서 사용되는 모든 형태의 구절을 표현하였다.

1) 사건 형태 :
본용언으로 사용된 동사/형용사들은 각각 하나의 사건을 나타낸다고 보고, 용언의 어간에 'e'로 시작되는 사건번호를 붙여 사건형태로 표시하였다.
먹(e1), 아프(e2), 일으키(e3), ...

2) 개체 형태 :

명사나 고유명사는 하나의 사건의 주체나 대상이 되는 존재라고 보고, 'x'로 시작되는 개체 번호를 붙여 개체 형태로 표시하였다.

위(x1), 설사(x2), 운동(x3), 부족(x4), …

3) 관계 형태 :

한국어에서 주로 조사나 어미로 표현되는 사건과 개체간의 관계를 나타내는 논리요소로 관계의 주체가 되는 쪽과 객체가 되는 쪽의 두 사건/개체 번호를 변수로 가진다. 관계형태는 그 관계의 종류에 따라 주어관계, 목적어관계, 수식어관계, 연결관계로 분류하였으며, 변수의 순서는 한국어의 일반적인 어절 순서에 따라 주어졌다.

- 주어관계 : 첫번째 변수는 개체로서 두번째 변수인 사건의 주어 역할을 한다.

  SUB_가(x1, e2) : 위(x1)가 아프다(e2)

- 목적어관계 : 첫번째 변수는 개체로서 두번째 변수인 사건의 목적어 역할을 한다.

  OBJ_를(x2, e3) : 설사(x2)를 일으키다(e3)

- 수식어관계 : 첫번째 변수가 두번째 변수의 수식어가 되는 관계이며, 두 변수는 모두 개체 또는 사건이 될 수 있다.

  MOD_의(x5, x6) : 눈(x5)의 전염병(x6)
  MOD_에(x7, e7) : 각막(x7)에 외상을 입다(e7)
  MOD_ㄴ(e8, x9) : 상한(e8) 음식물(x9)
  MOD_게(e9, e10): 낮게(e9) 하다(e10)
  MOD__(a1, e1) : 많이(a1) 먹다(e1)

- 연결관계 : 체언과 체언이 수식관계 없이 동등한 연결을 가질 때, 혹은 사건과 사건이 병렬적 구조를 가질 때의 관계를 표현한다.

  CON_고(e13, e14) : 배가 아프고(e13) 식욕이 없다(e14)

4) 특성형태 : 일반적으로 관형사/부사는 논리형식의 단순화를 위해 반영하지 않으나 질병영역에서의 '많이', '오래' 등과 같은 일부 단어는 그 영역에서 매우 중요한 의미를 가진다. 이러한 단어들을 사건의 특성을 표현한다는 의미에서 'a'로 시작하는 특성번호를 가지는 별도의 논리요소로 취급하였다.

많이(a1), 오래(a2), …

5) 조용형태 :

한국어에서 명사가 서술어로 쓰일 수 있는 단어의 경우 서술격 조사나 조용보조어간(혹은 조용사)과 결합하여 서술어로 사용되는 경우가 많다. 이러한 결합은 관형격의 수식을 받는 등 체언으로서의 성질과, 주어, 목적어를 가지는 용언으로서의 성질을 같이 보유하게 된다. 이러한 형태를 논리형식에 반영하기 위하여 이러한 결합을 명사와 서술격조사(, 조용사)를 각각 하나의 개체와 조용형태로 분리하여 개체는 체언으로서 수식어 들과의 관계를 유지하게 하고, 조용형태에는 사건으로서 주어나 목적어와의 관계를 유지할 수 있도록 하였다. 이러한 조용형태는 사건과 성질이 동일하므로 'e'로 시작하는 사건번호를 부여하였으며, 사건과는 달리 결합하는 명사와의 관계를 같이 가지고 있어야 하기 때문에 두번째 변수로 개체번호를 가지고 있다.

ZOY_하(e0, x0) : 피로(x0)하다(e0)
ZOY_이(e11, x11) : 세균성(x11)인(e11) 심내막염(x12)

의미속성이 부착된 질병 정의문은 논리형식으로 자동 변환되었으며, 그 결과는 표2와 같다. 질병 정의문의 논리형식에서 개체형태가 사건형태에 비해 약 3배가량 많았으며, 관계형태는 등장하는 횟수는 많았으나 실제 단어는 적었다. 또한 관계형태에서는 수식관계와 주어관계가 주종을 이루고 있었다.

	등장횟수	단어수		등장횟수
개체형태	451	285	주어관계	111
사건형태	167	82	목적관계	45
관계형태	337	37	수식관계	145
특성형태	16	10	연결관계	24
조용형태	32	3	합계	337
합계	1003	415		

표 2 질병 정의문에서 자동 생성된 논리형식 종류

4 질의응답 및 이벤트 탐색

4.1 관련 연구

질의응답시스템이란 사용자의 질의에 근접한 문서만을 찾아주는 문서 검색시스템의 한계를 극복하고자 시작된 연구로서, 문서 내에서 사용자가 원하는 정답을 제시하는 시스템이다. 이러한 질의응답 시스템은 미국 표준기술협회(NIST)[1]에서 주관하는 TREC(Text REtrieval Conference)[2]를 통해 많은 시스템이 발표되고 있다.

이들 질의 응답 시스템은 대부분 질의에서 요구하는 정답의 형태와 질의의 키워드들을 추출한 후, 정보검색기술을 사용하여 문서에서 질의 키워드들이 가장 많이 나오는 단락을 선택한 후 이 안에서 질의가 요구하는 정답의 형태를 찾아 이를 정답으로 제시하는 형태를 취하고 있다.[3][4]

하지만 2000년의 TREC-9과 2001년 TREC-10의 경우 문서내의 문맥을 고려하여 정답을 유추하여야 하는 질문이나 단순히 문장 내 키워드의 분포뿐만 아니라 키워드 간의 관계를 고려해야 하는 질문 등이 포함되어 있으며, 이러한 질문들의 대답을 추출하기 위해서는 문서 내 존재하는 각 사건들이나 개체들간의 관계를 고려할 수 있는 형태의 질의응답 시스템이 필요하다.

FALCON[5][6]은 선택된 단락에 대해 논리형식으로 변환하고 질문의 논리형식과의 교합을 통해 단락이 질의에 적합한 내용인가를 선택하며, 이 과정에서 Wordnet[7][8]을 일반상식의 지식원으로 같이 고려한다. Qanda[9]에서는 추출된 후보문서를 논리형식으로 변환하여 논리형식의 저장소(pool)인 지식베이스를 구성하고 질의의 논리형식과의 비교를 통해 정답을 추출한다.

백과사전을 대상으로 하는 초기의 질의응답 시스템으로는 MURAX[10]이 있으며, 이는 정보검색 기법을 이용하여 사전 정의된 몇 가지 형태의 질의에 대한 정답을 추출하고 있다.

4.2 질의 응답 시스템의 개요

그림1에 보인 질의응답 시스템은 지식습득과정을 통해 습득된 내부지식을 기반으로 사용자의 지식탐색 질의를 분석하고 질의의 의도를 파악하여 지식탐색을 통해 대답을 생성하는 시스템이다. 이 시스템은 질의 응답을 통해 내부지식을 확장해 나가는 자동 지식확장 기능과, 사용자와의 여러 단계의 대화를 통해 내부지식을 탐색하고 사용자의 올바른 의도를 파악하고, 사용자가 내부지식을 파악하는 데 도움이 줄 수 있는 시스템을 목적으로 설계되었으며, 현재 사용자의 질의에 대응하는 답변이나 반문을 하는 1단계의 질의응답과정만이 구현되어 있다.

질의 응답 시스템을 사용하기 위한 초기 내부지식은 3장에서 밝힌 질병관련 백과사전의 정의문에서 반자동 추출되었으며, 질의 분석을 위해 인터넷상에 공개되어 있는 의료상담 질문 1500여건을 사용하였다.[12] 이 질문들은 1~3개의 문장으로 이루어져 있는 의료상담 질문 제목들이다. 질문의 예는 다음

과 같다.

"원형탈모증 같습니다. 어떻게 해야 하나요."
"어디가 안 좋은 거며, 어느 과로 가야 합니까?"
"저희 오빠가 간암 말기입니다."
"소화가 잘 안되서요."
"조울증 같아요. 어떡하죠?"
"몸은 설사를 합니다. 왜 그런지요?"

4.3 구절 분리

구절 분리를 하는 원칙은 하나의 구절에는 하나의 사건만이 포함되도록 한다는 것이다. 이를 위해서 구문 분석된 결과로부터 본용언으로 사용된 용언을 구절 분리의 후보로 하였으며, 용언의 어미를 분석하여 관형형 어미나 부사형 어미와 연결된 용언은 제외하고, '-면/-으면'과 같은 일부 연결어미를 제외한 나머지를 구절분리의 원칙으로 삼았다. 이 과정에서 "-르 수 있다.", "-다고 합니다.", "-해서 그렇다는데."와 같은 두개 이상의 용언으로 이루어진 상용어구는 하나의 어미로 취급하였다.

또한 구절 분리과정에서 중요한 것은 "지나치게 많이 마시거나 먹었을 때"와 같이 하나의 부사어가 여러 개의 용언을 동시에 수식할 때 분리되는 구절에 이들 부사어를 각기 포함시켜줘야 한다는 것이다. 현재 질의 분석과정에서는 간단한 경험 규칙을 이용하여 해결하고 있다.

4.4 구절별 의미속성 부착 및 질의 의도 분석

구절 분리된 사용자 질의는 각 구절별로 질병영역에서 사용되는 의미속성을 부여한다. 이러한 의미속성 부착에 관한 연구로 본 논문과 동일한 영역의 백과사전 정의문을 대상으로 하는 의미 속성 부여 연구[11]가 있었으며, 이 연구에서는 구절 내 주요 단어와 구절어미, 구절 끝 단어를 이용하여 구절의 의미속성 부착을 수행하고 있다. 이 질의응답 시스템에서 의미속성 부여의 대상이 되는 것은 사용자 질의이므로 구절어미와 구절 끝 단어에 대한 정보가 유용하지 않아, 다른 형태의 의미판단 근거를 도입하고자 한다.

이 시스템에서 사용된 의미속성은 질병명, 원인, 치료법, 증세 등의 영역 지식 속성과 질의의 의도이다. 질의 의도를 의미 속성 부여시 같이 처리할 수 있도록 훈련 코퍼스로 사용된 질의문들을 분석하여 각 의도별로 단서 구절들을 추출하였다. 다음은 각 의도별 단서 구절들의 예이다.

사실의 확인:
X(x0) SUB_이(x0, e1) 맞(e1)
현상의 설명:
어떤(a1) MOD__(a1, x0) X(x0)
무슨(a1) MOD__(a1, x0) X(x0)
왜(a1) MOD__(a1, x0) X(x0)
어떻(e1) MOD_게(e1, e0), X(e0)
병명: 병/이상/증상/질병(x0) 그렇(e0)
치료: 치료/치료법/치료방법/수술법/검사/진료(x0)
감당하/극복하/해결하/치료하(e0)
원인: 문제/원인/이유(x0)
일어나/생기/걸리/그렇(e0)

현상의 설명은 사건이나, 개체를 설명하기를 원하는 것이며, 논리형식에서 이 사건/개체가 어떻게 발현되느냐에 따라 병명의 설명을 원하는지, 치료법이나 원인의 설명을 원하는지가 결정된다.

구절의 의미속성 및 질의 의도의 부착은 내부지식에 저장된 각 속성별 논리형식에 발현하는 사건과 개체의 등장횟수와 각 속성별 빈도수, 질의 논리형식과 내부지식에 저장된 구절의 논리형식과의 비교를 기반으로 한다. 구절 p가 표현하는 질병노드를 n이라고 하고 의미속성을 f라고 할 때, 구절 p의 의미속성은 $P(n,f|p)$를 최대로 하는 속성으로 결정된다. 이 확률값은 다시 (1)과 같이 각 구절이 어떠한 의미속성을 가질 것인가의 확률값, $P(f|p)$와 구절의

의미속성이 결정되었을 때 이 구절이 어떠한 질병노드에 속할 것인가의 확률값, P(n|f,p)로 분리할 수 있다.

$$P(n,f|p) = P(f|p) * P(n|f,p) \quad (1)$$

다시 말하면, **"소화가 안되서요."**라는 질의구절이 **'식체'**라는 질병의 **'증세'**라는 의미속성을 말하고 있을 확률값은 이 구절이 **'증세'**라는 의미속성을 가질 확률값과 **"소화가 안된다는 '증세'"**가 **'식체'**의 **'증세'**일 확률값으로 분리될 수 있다는 의미이다.

구절이 어떠한 의미속성을 가질 것인가의 확률 P(f|p)는 내부지식에 저장된 각 속성별 논리형식에 발현하는 사건과 개체의 등장횟수와 내부지식에서의 각 속성별 빈도수를 기준으로 계산된다. 이 확률식으로 각 구절별로 여러 개의 순서화된 의미속성 후보값이 확률값과 같이 추출된다. 하나의 구절에는 하나의 의미속성을 가지는 것을 원칙으로 하나, 여기에서는 질의의도와 의미속성은 하나의 이벤트 안에서 동시에 발현될 수 있다고 보아 이들은 하나의 구절에서 같이 추출될 수 있도록 하였다. 이러한 경우는 **"배가 아픈 이유가 뭡니까"**와 같은 질의에서 나타난다.

또한 의미속성중 질병명은 **"고혈압도 치료되나요?", "천식으로 벤토린을 쓰는데"**와 같이 구절내에 하나의 개체로 존재하므로, 다른 의미속성과 같이 존재할 수 있도록 하였다.

이러한 경우가 아닌 하나 이상의 의미속성이 동시에 후보로 추출되었을 때는 이들 모든 후보를 대상으로 다음 절에서 설명할 P(n|f,p)을 사용하여 질의의 의미속성을 결정함과 동시에 해당되는 질병노드를 선택할 수 있도록 하였다.

4.5 지식탐색과정

구절의 의미속성이 결정되었을 때 이 구절이 어떤 질병노드에 속할 것인가의 확률값은 질의 논리형식과 내부지식내의 논리형식과의 대조를 통해 그 값을 결정하는데, 예를 들어 **"소화가 안되서요."**라는 구절이 **'증세'**라는 의미속성으로 판정되었을 때, 이 증세가 어떤 질병에 해당하는 증세인가를 표현할 수 있는 확률은 질의 논리형식과 내부지식 내 **'증세'**로 의미속성이 부여된 논리형식들 간의 대조를 통해 가장 근접한 논리형식을 가지는 질병노드가 주어진 질의 **"소화가 안되서요."**의 질병노드로 결정된다는 것이다. 이 확률값은 (질의의 논리형식과 부합되는 사건,개체,관계 형태수) / (논리형식의 사건,개체,관계 형태수)로 표현된다.

5 질의 응답

주어진 질의의 의도/질병명/속성이 결정되면, 결정된 내용을 기초로 대답을 생성한다. 질의의 의도가 '사실의 확인'일 경우 질의에 나타난 질병명과 의미속성을 4.5절의 지식탐색과정을 통해 내부지식과 비교하여 사실인지 아닌지에 대한 확인 대답을 생성하며, 내부 지식에 없는 것으로 판명되었을 때 같은 의미속성의 다른 형태를 다시 요구하는 대답을 출력한다.

질의가 '현상의 설명', 즉 자신이 기술한 증세의 병명이나, 치료법등을 원하는 경우, 내부 지식 탐색의 결과로 탐색된 논리형식을 자연언어로 변환하여 생성할 수 있도록 하여야 한다. 현재 이러한 경우 사전 정의문 원문을 사용하여 대답을 출력하고 있으나, 탐색된 논리형식이 **"소화가 안되는 이유가 무엇입니까?"**라는 질문과 같이 '식체'와 '소화불량'의 두 질병노드의 원인을 모두 통합하여 대답해야 하는 경우, 탐색된 지식의 논리형식으로부터 병렬어구를 해결하여 직접 자연언어를 생성하는 방법이 필요하다.

현재 질의의 의도를 파악하는 것이 실패했을 경우 주어진 질의의 의미속성을 가진 질병명을 요구하는 것으로 판단하며, 질의의 의미속성도 파악하지 못하는 경우 사용자에게 다른 이야기를 하도록 유도하는 기초적인 대화시스템을 사용하고 있다.

6 결론

지식을 문서로부터 습득하고 습득된 지식을 대상으로 탐색하여 사용자와의 대화를 진행하는 지식 습득/탐색/질의응답 시스템의 연구를 위해 질병 영역만을 대상으로 프로토타입 시스템을 설계/구현하였다. 지식원으로서는 백과사전의 질명 정의문을 사용하였으며, 일반 사전의 질병 정의문을 사용하여 지식을 확장하였다. 지식원으로부터 지식을 습득하기 위해 형태소분석/구문분석기가 사용되었으며, 초기 지식원의 의미속성구조와 구절별 의미속성부착은 수작업을 사용하였다. 질의응답을 위하여 지식원은 그 단락의 의미에 따라 구조화되고 각 단락은 논리형식으로 변환되었다.

지식원으로부터 지식을 습득하는 과정에서 한국어 의존구조와 의미구조의 중간에 위치하는 논리형식을 정의하여 지식습득과정을 자동화하였고, 사용자 질의를 구절분리하고 의미속성 부착 및 질의의도 추출, 지식 탐색 과정을 구현하였다. 이러한 과정을 자동화하기 위해 내부지식 내 개체 및 사건의 발현도, 논리형식의 유사도를 사용하였다.

현재의 구현된 시스템은 주어진 질의의 의도를 파악하여 이에 해당하는 내용을 내부지식을 탐색하여 보여주고, 탐색에 실패했을 경우, 의도파악에 실패했을 경우, 의미속성 부착이 실패했을 경우 등에 대해 다른 형식의 대답을 생성하는 기초적인 대화시스템을 포함하고 있다. 현재 질의과정에서 습득된 지식을 내부지식으로 재사용하는 방법과 다단계 질의응답과정, 자연언어 대답생성 등의 연구를 진행하고 있다.

감사의 글

본 연구는 전문용어언어공학연구센터에서 수행한 과학 기술부와 KISTEP의 핵심소프트웨어사업 중 "대용량 국어정보 심층처리 및 품질관리 기술개발" 과제의 일환으로 수행되었으며, 부분적으로 첨단정보기술연구센터를 통하여 과학재단의 지원을 받았습니다.

참고 문헌

[1] Nation Institute of Standard and Technology, http://www.nist.gov

[2] Text REtrieval Conference, http://trec.nist.gov

[3] R. J. Cooper, S.M. Rüger, "A Simple Question Answering System", TREC-9 Proceeding, 2000

[4] T. Takaki, "NTT DATA TREC-9 Question Answering Track Report", TREC-9 Proceeding, 2000

[5] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. G rju, V. Rus, P. Morarescu, "FALCON: Boosting Knowledge for Answer Engines," TREC-9 Proceeding, 2000

[6] Dan I. Moldovan and Vasile Rus, "Logic Form Transformation of WordNet and its Applicability to Question Answering", Proc. Of ACL 2001

[7] Christiane Fellbaum, "WordNet: An Electronic Lexical Database", MIT Press

[8] WordNet: a lexical database for the English language, http://www.cogsci.princeton.edu/~wn/

[9] E. Breck, J. Burger, L. Ferro, W. Greiff, M. Light, I. Mani, J. Rennie, "Another Sys Called Qanda", TREC-9 Proceeding, 2000

[10] "MURAX:A Robust Linguistic Approach For Question Answering Using An On-Line Encyclopedia", SIGIR'93

[11] Haeseung Paik, et. al., "Analysis of Linguistic Features for Identifying Information Constituents of a Concept", Proc. of NLPRS01

[12] DOCTOR, http://doctor.co.kr/