개념어의 습득을 위한 지식기반 질의응답 시스템

울산대학교 컴퓨터정보통신공학과
이재홍 · 최호섭 · 옥철영

Knowledge-Based Question Answering System for Aquisition of Concept Word

Jae-Hong Lee, Ho-Seop Choe, Cheol-Young Ock
Department of Computer Engineering and Information Technology, University of Ulsan, Ulsan, Korea

요 약

본 논문에서는 현실 세계가 가지고 있는 지식이 어느 정도 체계적으로 정제되어 있는 국어사전, 백과사전 등을 중심으로, Hybrid Method를 이용한 통계(Statistics)기반 지식베이스와 어휘분류(Lexicon Classification)기반 지식베이스를 효율적으로 구축하여 질의응답시스템에 활용한다. 또한 특정한 문서를 보여주는 일반적인 질의응답시스템과는 달리, 이러한 지식베이스를 이용하여 사용자에게 정확한 개념어(정답어)를 습득하게끔 해주고, 사용자의 인지 체계 속에 어렴풋이 내포되어 있는 개념적 지식을 더욱더 표면적으로 확장해 나갈 수 있는 질의응답시스템을 구축하는 방안을 제시한다.

서 론

지식사회, 정보사회가 대두되면서, 컴퓨터로 하여금 사용자의 질의에 의한 정확하고 의미 있는 정답을 보여주는 지식기반 정보검색 시스템을 요구하게 되었고(Voorhees E & Ticd DM, 2000), 많은 연구들이 AAAI(AAAI Fall Symposium on Question Answering)와 TREC(TREC(Text REtrieval Conference) Overview)을 중심으로 수행되어 왔다.

기존의 정보검색 시스템이 단순 키워드에 의한 문서의 검색에 초점을 맞추었다면, 최근의 질의응답시스템은 사용자의 자연언어 질의에 대해 언어처리 기술을 이용하여 1차적인 언어 분석을 한 다음, 2차적으로 현실 세계의 지식을 체계화한 지식베이스를 통해 의미 분석 및 추론을 하고, 마지막으로 정확한 정답어나 정답문서를 제시하는 시스템으로 발전하고 있는 단계이다.

그러나 현재 질의응답 시스템이 제공하는 것은 언어처리 기술을 이용하여 질의를 분석하여 특정한 문서를 제시하는 수준에 그치고 있다. 이런 의미에서 보다 정밀적이고 도메인 제약적인 효율적인 시스템에 대한 연구가 활성화될 필요가 있다. 즉 질의응답 시스템이 하나의 소규모 대화형 시스템이라는 점을 감안한다면, 사용자와 컴퓨터간에 지식을 서로 주고 받을 수 있는 의미 있는 시스템이 개발되어야 한다.

본 논문에서는 특정한 문서를 보여주는 일반적인 질의응답시스템과는 달리, 현실 세계가 가지고 있는 지식이 어느 정도 체계적으로 정제되어 있는 국어사전, 백과사전 등을 이용하여 지식베이스를 효율적으로 구축하여 사용자에게 정확한 개념어(정답어)를 습득하게끔 해주고, 사용자의 인지 체계 속에 어렴풋이 내포되어 있는 개념적 지식을 더욱더 표면적으로 확장해 나갈 수 있는 질의응답시스템의 구축 방안을 제시하고자 한다. 이를 위해 도메인을 '교통기관'으로 정하였으며, 기존의 질의응답 시스템에서 사용하는 일반적인 언어처리 기법을 이용하여 질의를 처리함과 동시에, 의미 있는 정답 제시를 위해 어휘 분류 구조, 통계기반 지식베이스, 어휘분류기반 지식베이스, 구문 정보 등을 이용한 질의응답 시스템을 제시하고자 한다.

질의응답 시스템을 위한 기초 작업

1. 개념어 선정

1) 개념어의 정의

'개념'이란 일반적으로 어떤 사물 현상에 대한 일반적인

E-mail : jahlee@mail.ulsan.ac.kr
E-mail : hoseop@mail.ulsan.ac.kr
E-mail : okcy@mail.ulsan.ac.kr

지식을 말하며, 철학적으로는 여러 관념 속에서 공통된 요소를 뽑아내어 종합하여 얻은 하나의 보편적인 관념을 말한다. 이러한 보편적이고 일반적인 지식 또는 관념이 언어로 표현되는데, 이것이 하나의 개념어가 되는 것이다. 본 논문에서는 어떤 사물을 사람이 인지할 경우, 어떤 사물을 지칭하는 어휘를 개념어로 판단하고, 그 사물과 관련하여 연상되는 단어들을 의미 정보로 판단한다.

2) 개념어에 해당하는 기초어휘 선정

기초어휘는 개별언어의 여러 사용 영역에서 공통되는 어휘로서 일상생활을 영위하는 데 필수적이고 절대적인 어휘를 가리키는 것으로, 본 논문에서는 <교과서의 어휘 분석 연구> 서종학(1999)이 제시한 빈도순 어휘 목록(상위 3000위)에서 선정하도록 한다.

3) 도메인 선정 및 해당명사 목록 선정

본 논문에서는 일상생활에서 가장 많이 접할 수 있는 도메인인 '교통' 중 '교통수단' 과 관련된 명사 어휘들을 대상으로 한 질의응답시스템을 구현하고자 하였다.

이를 위해 '교통수단' 과 관련된 명사 어휘 목록을 확보하기 위하여, 먼저 기초어휘를 중심으로 1차적으로 선별하고, 다음으로 <한국어 명사의 의미 계층 구조>(조평옥, 1996)를 이용하여 교통수단과 관련된 어휘들을 확장하였다. 그 결과 '교통수단' 과 관련된 283개의 어휘 목록을 확보하였다.

2. 언어자원

지식베이스 형성과 질의집합 형성을 위해, 금성국어사전, 계몽백과사전, 세종150만 어절 말뭉치를 언어자원으로 사용하였다. 이 언어자원들은 지식베이스를 형성하는 체언(명사, 고유명사), 용언(동사, 형용사)으로 구성된 의미 정보 추출에 이용되었으며, 질의응답시스템에 실험 평가용 질의집합에도 이용되었다.

3. 개념어의 어휘 분류 구조

개념어들을 일정한 규칙에 의해 계층적 분류 구조를 형성시켜, 질의응답시스템에 활용하고자 하였다. 최상위 노드에 위치한 '차', '배', '항공기' 에 대한 각각의 어휘 분류 구조는 다음과 같다.

차의 경우는 Fig. 1과 같이 총 108개의 단말 노드로 구성이 되어 있으며, 레벨(level)은 최고 8레벨까지 형성되어 있다.

배의 경우는 Fig. 2과 같이 총 111개의 단말 노드로 구성되어 있으며, 레벨은 최고 8레벨까지 형성되어 있다. 뜻풀이에 기술된 상위어 정보를 이용하여 어휘 분류 구조를 자동 구축할 경우에, '배' 의 하위 노드가 많이 생성되는 문제가 발생하기 때문에, "Class A"라는 새로운 노드를 설정하여 군함, 상선, 어선 등에 속하지 않고, 뜻풀이에서 배를 상위어로 가지는 것들을 "Class A"의 하위노드로 분류하였다.

항공기의 경우는 총 64개의 단말 노드로 구성이 되어 있으며, 레벨은 최고 7레벨로 형성되어 있다.

4. 구문 정보(Syntactic Information)

용언을 중심으로 한 격 관계(case relation)를 이용하여 질의응답 시스템에서 구문 정보로 활용하였다. Table 1에서와 같이 어휘간의 의미적 상호 연관 정보를 사전의 뜻풀이와 백과사전의 개요 정보, 그리고 세종 말뭉치에서 자동으로 추출하였다.

여기에서 중요한 점은 동사, 형용사 중심의 용언을 비롯

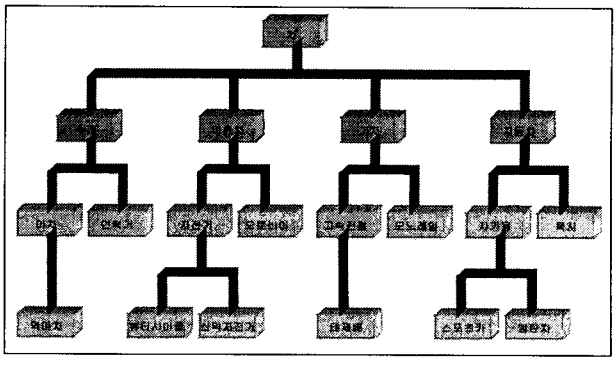

Fig. 1. '차(Car)'의 어휘 분류 구조.

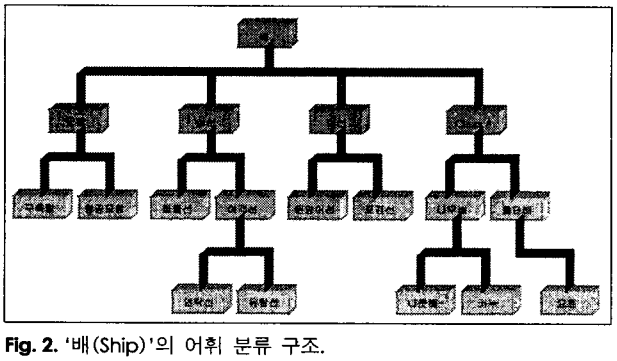

Fig. 2. '배(Ship)'의 어휘 분류 구조.

Table 1. '구문 정보' 테이블

개념어	Case relation	Fmpm	Bmpm
비행기	N_N	양력	발생
배	N_B_V	물	뜨다
자전거	N_O_V	페달	밟다
침몰선	V_M_N	가라앉다	배
달구지	N_S_V	소	끌다

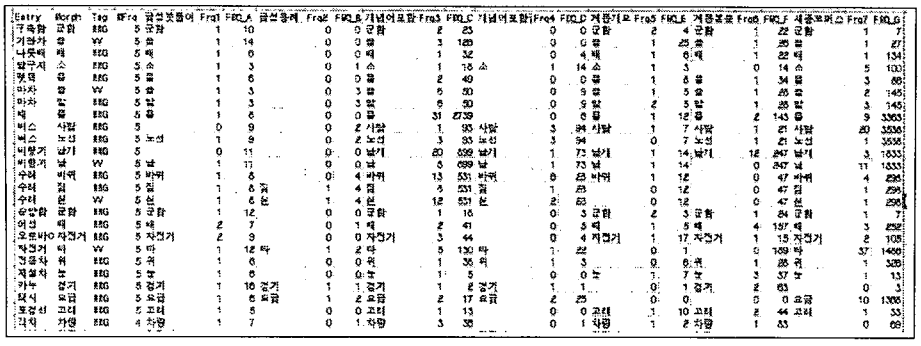

Fig. 3. 통계기반 지식베이스의 일부.

하여 용언적 성격을 가지고 있는 서술성 명사(동작 명사, 상태 명사)들도 구문 정보로 추출하였다는 것이다. 즉 Table 1과 같이 어휘간의 구문 정보를 추출하여 질의응답시스템에서의 질의 처리에 이용하고자 하였다.

5. 지식베이스 구축

1) 통계기반 지식베이스

통계기반 지식베이스는, 먼저 품사 태그 및 의미 태그가 부착된 언어자원(사전 및 말뭉치)으로부터 개념어를 구성하는 의미 정보들을 추출하고, 다음으로 각 언어자원마다의 빈도정보를 이용한 확률값을 가지도록 한 지식베이스이다.

Fig. 3과 같이 통계기반 지식베이스에는, 본 논문에서 사용한 언어자원을 7가지로 나누어 지식베이스를 구성하고 있다. 먼저 '금성국어사전'은 개념어의 뜻풀이와 용례, 그리고 개념어가 포함된 뜻풀이와 용례로부터 의미정보를 추출하였으며, '계몽백과사전'은 개요 부분과 본문 부분을 나누어 문장순위 가중치를 적용하여 의미 정보를 추출하였다. 다음으로 '세종 150만 어절 말뭉치(Corpus)'는 개념어들과 공기하는 의미 정보를 추출하였다. 여기에서 추출된 의미 정보들 중 불필요한 의미 정보들을 정제하기 위하여 매칭(Matching) 빈도를 적용하였다. 여기서의 매칭 빈도는 분류된 7가지의 구조에서 개념어를 표현하는 의미 정보들이 7가지 구조에 몇 번 존재하는가를 나타내는 것으로, 7가지 구조 중 과반수 이상으로 개념어에 대한 의미 정보를 가졌을 경우에는 필요한 의미 정보로 판단하기 위해 사용된 것이다. 즉, 일반적으로 매칭 빈도가 높은 의미정보는 중요도가 커지게 된다.

위의 7가지의 구조를 가진 통계기반 지식베이스는 각 구조별로 중요도를 부여하여, 각 구조별 가중치를 실험을 통하여 가지게 된다.

2) 어휘분류기반 지식베이스

어휘분류기반 지식베이스는, 도메인 '교통수단'과 관련된 어휘들을 의미 분류 방식을 이용하여 계층적으로 분류한 구조를 가짐과 동시에, 각 개념어들이 가지는 의미 정보를 공통 의미 정보와 개별 의미 정보로 나누어 의미 정보의 상속 구조를 가지도록 한 지식베이스이다. 이러한 어휘분류기반 지식베이스가 가지는 장점은, 분류 정보를 이용한 질의 제약이 가능하며, 의미 정보를 계층적 분류 구조를 이용하여 공통 의미 정보와 개별 의미 정보로 구분하여 사용할 수 있어, 의미 정보를 분류적으로 정제할 수 있다. 또한 분류 구조가 가지는 의미 정보의 상속이라는 측면을 통하여, 하위 개념어가 가지지 못하는 중요한 의미 정보를 공통 의미 정보를 통해 보충해주는 역할을 할 수 있다. 어휘분류기반 지식베이스의 구조와 의미 정보 상속에 대하여 Fig. 4를 통해 확인할 수 있다.

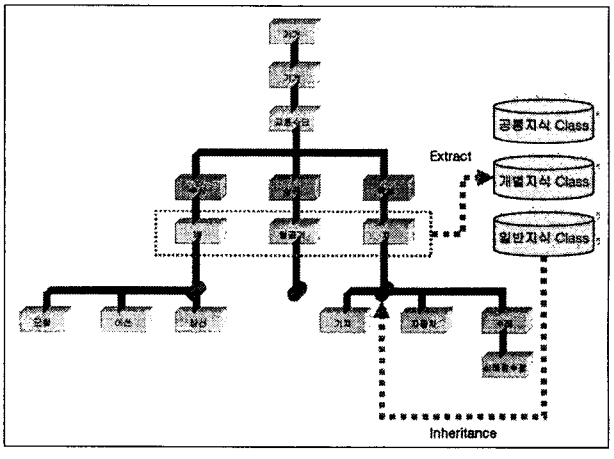

Fig. 4. 어휘분류기반 지식베이스와 의미 정보 상속.

어휘분류기반 지식베이스에서 사용된 공통 의미 정보와 개별 의미 정보, 일반 의미 정보에 대한 각각의 설명은 다음과 같다.

- 공통 의미 정보 Class : 형제 노드에서 공통으로 가지는 의미 정보로, Base-Line을 50%로 설정하여 실험한 결과, 70% 이상 일치되는 의미 정보를 공통 의미 정보로 설정하였다.

• 개별 의미 정보 Class : 형제 노드에서 개별적으로 가지는 의미 정보로, 실험을 통하여 20% 이하만 일치되거나 한번만 출현하는 의미 정보들을 개별 의미 정보로 설정하였다.

• 일반 의미 정보 Class : 위의 공통 및 개별 의미 정보 Class에서 제외된 의미 정보로, 상대적으로 중요도를 낮게 가지게 된다.

의미 정보의 상속은 일부의 경우, 상위어(Hypernym)의 의미 정보를 하위어(Hyponymy)가 상속받지 않아야 하는 경우가 생긴다. 예를 들어 '자전거'의 경우, "바퀴가 있다"라는 구문 의미 정보를 하위어인 '뷰티사이클'이 상속받게 되는데, '뷰티사이클'은 미용기구로 바퀴가 없는 경우이다. 이와 같이 상속이 이루어질 때, 상속 제약이 필요하게 되는데, 이것은 구문 정보를 활용하여 제약을 하도록 한다.

개념어 습득을 위한 질의응답 시스템

1. 시스템 전체 개요도

개념어 습득을 위한 질의응답 시스템은 Fig. 5에서 보듯이 사용자의 질의에 대한 의미 있는 정답을 제시하기 위하여 3단계의 과정을 거치게 된다.

우선 자연어 질의가 입력되면 형태소 분석과 단어 중의성 해결(WSD)이라는 기본적인 언어처리 과정을 거치게 되며, 3.2절에서 언급할 질의 분석을 통하여 생성된 질의 유형과 질의 제약 정보를 통해 질의문의 분석이 이루어진다.

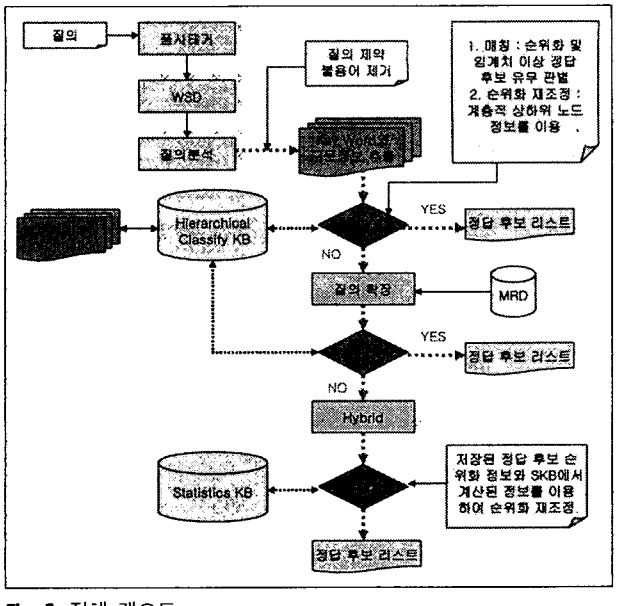

Fig. 5. 전체 개요도.

• 1단계 과정 : 질의분석을 통해 수집된 정보를 구축되어진 구문정보와 어휘분류 기반 지식베이스와 매칭을 시키는 단계로 먼저 정답후보를 순위화하고 실험을 통한 임계치 이상의 후보의 유무를 판별하여 정답후보 리스트가 생성이 되면 3.3절의 어휘 분류 구조를 이용한 순위화의 재조정과정을 거친다.

• 2단계 과정 : 1단계 과정을 통해 정답 후보 리스트가 생성이 안 될 경우에 사용자가 입력한 질의문에 등장하는 Key Word의 동의어, 유의어와 뜻풀이를 이용하여 확장을 하게 된다. 질의확장은 울산대 "한국어처리연구실"이 보유한 기계가독사전과 유의어사전을 이용하여 자동으로 이루어진다.

• 3단계 과정 : Hybrid Method를 이용하여 1,2 단계를 통해 생성된 정답후보 순위화 정보와 통계기반 지식베이스를 통해 생성된 순위화 정보를 통합하여 최적화 된 정답 후보 리스트를 생성하게 된다. 위의 단계에서 정답 오차범위를 실험을 통해 결정하여 한 개의 정답개념어 또는 정답개념어 리스트를 보여주게 된다.

2. 질의분석

사용자가 원하는 정확한 개념어를 제시하기 위해 질의에서 사용자의 질의 의도를 분석할 필요가 있다. 이러한 질의분석을 수행함으로서 질의의 유형과 질의의 제약을 설정할 수 있다.

1) 질의유형 및 질의제약

질의 분석을 수행한 결과 6개의 질의유형을 설정할 수 있었고 내용은 아래와 같다.

• 단문에 의한 질의 : 대부분 사용자가 여러 개의 정답어를 요구하는 경우로 예를 들어 "사람이 타고 다니는 것은?"과 같은 질의는 사람이 타고 다니는 모든 교통수단을 보여주기를 원하는 경우이다. 또는 사용자가 차(car)에 대한 질의를 위와 같이 하더라도 정답어 리스트 중에 원하는 정답어가 포함되도록 하였다.

• 장문에 의한 질의 : 단문에 비하여 보다 많은 정보가 질의 내에 존재함으로, 사용자가 하나의 정답어를 요구하는 경우이다.

• 개념어의 정의를 묻는 질의 : "자동차란 무엇입니까?"와 같이 사용자가 정답어를 원하는 경우가 아니라 개념어의 정의를 묻는 경우로 개념어의 뜻풀이를 제시해 주도록 한다.

• 질의제약이 주어지는 질의 : 해당 개념어의 하위분류 집합에서 정답을 원하는 경우로 질의에 포함된 개념어의 하위

집합만을 정답집합으로 간주함으로써 질의처리의 부담을 줄임과 동시에 처리가 어려운 질의가 입력되더라도 오분석을 방지하여 근사한 정답을 제시해 주게 된다.

- 말뭉치에서 해당 개념어를 괄호로 대체한 질의 : 신문, 뉴스, 백과사전 본문, 웹 페이지 등을 활용하였다.
- 부정확한 질의 : 개념어를 부분적으로 잘못 인지하고 있거나 개념어를 표현하는 일부 단어를 모르는 경우이다.

3. 어휘분류 구조를 이용한 정답 개선

어휘 분류 구조에서 생성된 의미 정보들은 통계기반에 비하여 정보의 희소성(Data Sparseness) 문제가 발생하게 되는데 이를 어휘분류 구조를 이용하여 해결할 수 있다.

다음은 항공기에 관한 질의의 예이다. "사람이 타서 핸들을 잡고 날아다니는 것은 무엇입니까?" 위의 질의에서 "핸들을 잡고"라는 표현보다는 "조정하여"라는 질의가 항공기에는 적합한 질의이지만 사용자가 조정이라는 단어를 모르거나 사용자가 항공기에 대해 할 수 있는 질의 수준이 다양하다는 점을 고려해야만 한다.

"핸들, 잡다"라는 단어는 자전거가 일반적으로 가지는 의미정보이며 "사람, 타다"라는 단어는 공통적으로 가지는 의미정보이다. 그러므로 위의 질의는 의미정보를 많이 가지는 자전거를 오답을 내보낼 가능성이 있다.

그러나 Fig. 6에서와 같이 어휘분류 구조의 레벨 정보를 이용하면 정답어를 개선할 수 있다. 즉 Level 4에 있는 "핸들, 잡다"라는 정보보다는 Level 2에 있는 "날아다니다"가 더 지배적이라는 것이다. 상위 Level에 있는 정보는 하위어들이 공통적으로 가지는 의미정보로 그만큼 질의에서 지배적인 요소로 작용한다. 날아다닐 수 없는 자전거의 경우에 항공기와 동일한 Level을 가지는 차(car)로부터 상속받는 정보가 없으므로 정답어는 항공기를 제시해주게 된다.

Table 2. 질의제약 정보와 질의 예

질의 제약 정보	질의 예
-개념어/NNG(+-)?	불을 끄는 차는?
-개념어/NNG 중/NNB+-	비행기 중에 빠른 것은?
-무슨/MM 개념어/NNG+-	무슨 배가 고기를 잡나요?
-어느/MM 개념어/NNG+-	화물을 싣는 것은 어느 차?
-어떤/MM 개념어/NNG+-	어떤 차가 레일을 달리지?

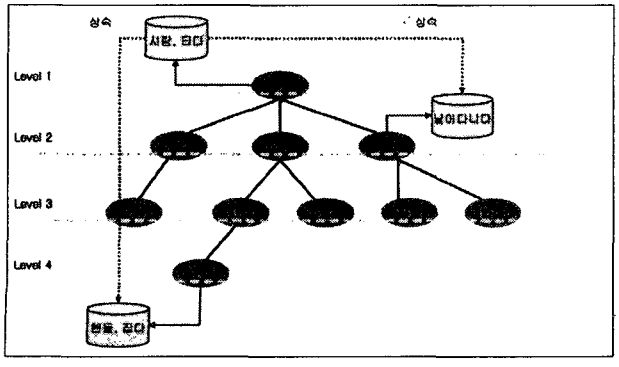

Fig. 6. 어휘분류 구조의 레벨 정보를 이용한 정답 개선.

실험 및 평가

1. 질의 집합

중학생 63명을 대상으로 한 설문지를 통하여 630개의 질의집합을 수집하였는데, 어휘 분류 구조의 총 283개의 '교통수단' 개념어 중에 각 레벨별로 다양하게 선별한 50개의 '교통수단' 개념어를 제시해 주고, 사용자가 원하는 10개의 개념어를 선정하여 개념어와 관련된 질의를 작성하도록 하였다.

Table 3에서 분류된 질의 유형의 예는 아래와 같으며 질의 유형 'A'는 대부분의 일반인이 정답어을 제시한 경우이고, 질의 유형 'B'에서는 일반인도 정답어를 제시하지 못하는 경우가 간혹 있었다.

- 은유적인 표현으로 된 질의 : "차위에 조그만 모자를 씌웠고 사람이 없으면 모자에 불이 들어오는 것은? 택시".
- 2차 지식까지 요구되는 질의 : "얼마 전 대구에서 이것을 타고 화재로 많이 죽었는데 이것은 무엇인가요? 지하철", "드라마 '요조숙녀'에서 김희선이 아버지의 직업을 이것을 탄다고 속였음? 윈양어선".
- 일반인도 정답어 제시를 못한 질의 : "내가 한번 타 보고 싶은 차? 리무진".

2. 실험 결과

질의 집합을 시스템에 적용한 결과는 다음과 같다.

Table 4의 결과에서 전체 질의에 대한 정확률은 일반인

Table 3. 질의 유형별 구성 비율

기호	질의 유형	비율
A	비교적 적합한 질의	71.61%
B	은유적인 표현이나 2차 지식까지 요구되는 질의	12.28%
C	일반인도 정답어 제시를 못한 질의	16.11%

Table 4. 질의 유형별 시스템 성능 평가

질의 유형	정확률(%)
전체 질의(A+B+C)	71.61%
일반인도 정답어 제시를 못한 질의 제외(A+B)	81.64%
비교적 적합한 질의(A)	90.16%

도 정답어 제시를 못한 질의를 포함하므로 질의 유형 'C'를 제외한 정확률 81.64%가 개념어의 습득을 위한 질의응답 시스템의 평가 기준이 될 것이다.

결 론

본 논문은 개념어의 습득을 위한 질의응답시스템을 제안하고 구축하였다. 이 시스템은 사용자의 인지체계 속에 어렴풋이 내재되어 있는 개념어에 대한 지식을 표면적으로 체계화시키기 위하여 개념어(정답어)의 제시와 함께 개념어의 뜻풀이말과 어휘분류 구조의 상하위 계층 정보까지 제시하여 준다. 사용자는 개념어를 습득한 다음에 제시한 정보를 바탕으로 자연스럽게 2차 질의를 하게 되며, 이러한 과정을 거치면서 개념어에 대한 명확한 이해와 개념적 지식의 확장이 이루어지게 된다.

향후에는 인간과 컴퓨터사이의 인터페이스 기능을 더욱 강화한 대화형 시스템 구축에 관한 연구와 은유적인 표현이나 2차 지식까지 요구되는 질의를 해결하기 위해 신문, 뉴스 등의 말뭉치를 효율적으로 활용하는 방안에 대한 연구가 계속 진행될 것이다.

REFERENCES

Voorhees E, Ticd DM(2000) : *"Building a Question Answering Test Collection", In Processing of SIGIR 2000, pp200-207*

AAAI Fall Symposium on Question Answering, http://www.aaai.org/Press/Reports/Symposia/Fall/fs-99-02.html

TREC(Text REtrieval Conference) Overview, http://trec.nist.gov/overview.html

김영택 등(2001) : 자연언어처리, 생능출판사

조평옥(1996) : "한국어 명사의 의미 계층 구조", 울산대 석사학위논문

김수민(2000) : "시소러스 범주정보를 이용한 질의응답시스템", 고려대 석사학위논문

이경순, 김재호, 최기선(2000) : "질의응답시스템의 성능평가를 위한 테스트컬렉션 구축", 제 12 회 한글 및 한국어 정보처리 학술대회 논문집, *pp190-197*

이경순, 김재호, 최기선(2000) : "한국어 질의응답시스템에서 개체인식에 기반한 대답 추출", 제 12 회 한글 및 한국어 정보처리 학술대회 논문집, *pp184-189*