

대화형 상품 검색 시스템에서 의미 구조 생성에 관한 연구

정혜경, 배우정, 이용석
전북대학교 컴퓨터정보학과

sysj2@paran.com, topaz.inu@gmail.com, yslee@chonbuk.ac.kr

Generation of a Semantic Structure on a Conventional Goods Search System

Hae-Kyung Jung, Woo-Jung Bae, Yong-Seok Lee
Dept of Computer Infomation, Chonbuk National University

요 약

인터넷 쇼핑물 분야에서 한국어 인터페이스에 대한 필요성이 대두되면서 이와 관련한 연구가 진행되어 왔다. 이러한 기존 연구들의 문제점은 다른 응용 분야에 대한 확장성이 떨어지고, 대화식 질의 문장을 처리하기 위해 필요한 상황 지식을 사용할 수 없다는 것이다.

확장성을 위해 한국어 인터페이스는 내부 구현의 변경 없이 도메인 지식의 교체만으로 다른 분야에 대한 적용이 쉬워야 한다. 아울러, 한국어 질의 문장을 SQL이나 ACL과 같은 여러 응용 분야의 인공 언어에 쉽게 변환할 수 있으려면 모호성이 전혀 없는 의미 구조로 표현되어야 한다. 이렇게 표현된 의미 구조는 상황 지식의 표현과 적용을 용이하게 한다.

본 논문은 Sowa의 개념 구조를 이용하여 의사 의미 구조를 유형 정의, 관계형 정의, 액터와 같은 도메인 지식베이스를 이용하여 표준 의미 구조로 변환하는 시스템을 제안한다. 본 논문은 먼저 필요한 도메인 지식베이스의 종류와 역할 등을 설명하고, 상황 지식을 이용하여 불완전한 의미 구조를 완전한 의미 구조로 변환하는 방법을 보인다. 아울러, 같은 의미를 갖는 여러 형태의 의사 의미 구조가 하나의 표준 의미 구조로 변환됨으로서 시스템의 일관성을 유지하는데 용이함을 보일 것이다.

본 시스템에서 생성한 의미구조는 중간 언어의 역할을 하므로, 제안하는 한국어 인터페이스 시스템은 데이터베이스 분야뿐만 아니라 에이전트 분야, 시맨틱 웹 분야 등에서도 손쉽게 적용할 수 있다는 장점을 가진다.

1. 서론

인터넷 쇼핑물의 급격한 발전으로 수많은 제품 중에서 사용자가 원하는 제품을 빠르고 정확하게 찾는 것이 중요한 이슈가 되었다. 대부분의 쇼핑물은 이를 위해 키워드 기반의 검색 기능을 제공하고 있다. 그러나 일반 사용자가 키워드 검색을 통해 원하는 정보를 찾는 것은 매우 불편하기 때문에 제품 검색을 위한 자연어 인터페이스의 개발에 대한 연구가 활발히 진행되고 있으며[12], 제한적이거나 기존 쇼핑물에 적용되어 사용되고 있다 [13].

기존의 한국어 인터페이스에 대한 연구는 형태소 분석 결과나 구문 분석 결과를 제품 질의를 위한 SQL로 변환하려는 시도로 진행되어 왔다[7,9]. 이러한 연구는 단어나 구문 정보를 해당 SQL로 변환하는 규칙을 기술함으로써 한국어 인터페이스를 구현하려고 노력하였다. 이러한 접근은 특정 응용 분야에 대한 한국어 인터페이

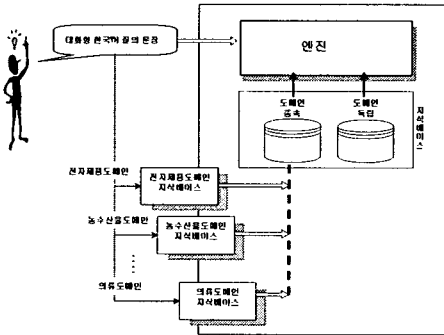
스를 구현하기에는 비교적 용이하나 단어 또는 구문 구조마다 규칙을 따로 기술해야 하기 때문에 다른 응용 분야로 적용하는 확장성이 떨어진다. 또한, 사용자는 상품에 대한 질의를 할 때 상황 지식을 이용할 개연성이 많으나 기존 방법으로는 이를 해결할 방법이 없다. 일례로, "50만원대 TV를 보여줘"라는 질의 후에 "그 보다 싼 것은?"이란 질의를 처리할 수가 없다.

응용 분야에 대한 확장성을 갖고 상황 지식을 사용할 수 있으려면 한국어 인터페이스는 지식베이스를 사용하고 사용자의 질의를 의미 구조로 표현할 수 있어야 한다. 단어가 상황에 따라 그 의미가 다양하고 분야에 따라 의미를 구분하는 것이 애매하기 때문에 범용적인 지식 베이스를 구축하는 것은 불가능에 가깝다.

그러나 인터넷 쇼핑물과 같은 특정 응용 분야로 제한하면 이에 대한 도메인 지식베이스를 완전하게 구축할 수 있다. 따라서, 그림 1에서 예시하는 바와 같이 내부

엔진은 변경하지 않고 응용 분야에 의존적인 도메인 지식만을 교체하여 해당 한국어 인터페이스 시스템을 구현할 수 있다면 응용 분야에 대한 확장성을 가질 수 있을 것이다.

대화 문장에서 생략이나 선행사의 사용은 매우 빈번



(그림 1) 도메인 응용분야에 의존적인 한국어 지식 기반 인터페이스

하지만, 사람은 상황 지식을 사용하기 때문에 다른 사람과의 대화에 어려움을 느끼지 않는다. 사용자가 보다 친숙하게 한국어 인터페이스를 사용할 수 있으려면 시스템은 이러한 대화식 문장을 처리할 수 있어야 한다. 즉, 한국어 인터페이스는 상황 지식을 컴퓨터 내부에 표현하고 선행사 문제나 생략 문제를 해결하려면 의미를 표현하고 연산할 수 있어야 한다. 이를 위해 한국어 인터페이스는 분석 결과를 정확한 의미 구조로 변환할 수 있어야 하며, 각종 모호성을 해결하고, 선행사 문제나 생략 문제를 해결하기 위해 필요한 지식베이스를 사용할 수 있어야 한다. 본 논문에서는 이를 지식 기반 한국어 인터페이스로 정의하였다.

지식 기반 한국어 인터페이스의 또 다른 장점은 하나의 의미가 다양한 문장 형태로 표현되어도 일괄적으로 처리할 수 있다는 것이다. 일례로, “삼성이 만든 TV?”, “제조회사가 삼성인 TV?”, “삼성에서 나온 TV?” 등을 들 수 있다.

본 논문은 Sowa의 개념 구조를 이용하여 한국어 구문 구조를 의사 의미 구조로 생성하고[10], 유형 및 관계형 정의나 정규 그래프, 액터 등과 같은 도메인 지식을 이용하여 이를 표준 의미 구조로 변환하는 시스템을 제안한다. 이렇게 생성한 표준 의미 구조는 상품 질의에 필요한 SQL로 쉽게 변환될 수 있다. 본 논문에서는 표준 의미 구조 변환을 위한 지식 구조의 종류와 역할을 설명하고, 이 지식베이스를 이용하여 표준 의미 구조를 생성하는 방법을 보일 것이다. 아울러, 상황 지식의 사용을 위해 단어 레벨 스택과 문장 레벨 스택을 사용하며, 이를 이용하여 불완전한 질의를 완전한 질의로 변환하는 과정 또한 보일 것이다.

본 논문의 구성으로, 제 2 장은 관련 연구 및 전자제품을 도메인으로 한 여러 종류의 한국어 질의문장들을 나열하고 한국어 질의문장의 특징에 대하여 기술한다. 제 3 장은 표준 의미구조를 생성하는데 필요한 도메인 지식베이스의 종류와 역할에 대해 설명하고, 제 4 장은

의미구조를 생성하는 절차를 기술하고, 제 5 장은 지식 기반 한국어 인터페이스의 구조를 살펴보고, 마지막으로 결론 및 향후 연구 과제를 논한다.

2. 관련 연구 및 기초 지식

2.1 관련 연구

자연어 인터페이스 분야는 전통적으로 데이터베이스를 검색하는데 정형질의어 SQL를 사용하지 않고 자연어를 이용하여 검색할 수 있는 데이터베이스 자연어 인터페이스에 관한 연구로 시작되었다. 현재까지 개발된 데이터베이스 자연어 인터페이스 시스템은 RENDEXVOUS[1], TEAM[2], LADDER[3], EUFID[4], IRUS[5], KDA[6] 등이 있고 국내에서 개발된 한국어 인터페이스 시스템은 NAURI[8], KNLQ[9] 등이 있다.

해외의 경우, 기존에 연구된 자연어 인터페이스는 자연언어의 처리단계를 구문분석 결과나 의미분석 결과뿐만 아니라 로직형태의 중간표현 질의[14,15]로 변환하여 SQL를 생성하려는 연구가 진행되어 왔다. 국내에서, 한국어 인터페이스에 대한 연구는 형태소 분석 결과나 구문 분석 결과를 이용하는 단계에 머무르고 있는 실정이다.

사용자가 입력하는 한국어 질의 문장을 형태소 분석 단계[7]만을 적용하는 경우 자연언어 질의 문장을 처리할 수는 있지만 단어레벨로 처리하므로 같은 구조를 갖는 문장도 어휘마다 예외로 처리할 뿐만 아니라 영역이 확장되는 경우 이를 체계적으로 처리할 수 없게 된다. 연구 [9]는 구문정보를 해당 SQL로 변환하는 규칙을 기술함으로써 한국어 인터페이스를 구현하려고 노력하였다. 그러나, 사용자가 입력하는 다양한 문장 형태를 고려하지 않고 제한된 문장 유형의 구문구조에 따라 규칙을 적용하고 있어 처리할 수 없는 질의문장도 존재한다. 또한, “비교구”와 같은 질의 문장은 같은 구문구조를 갖지만 의미가 달라 똑같은 규칙을 적용할 수 없다는 단점을 안고 있다. 구문정보만을 이용하여 처리하는 경우 앞에서 예로든 다양한 문장 형태는 동일한 의미를 가지므로 일괄적으로 처리할 수 없다는 단점이 있다.

또한, 사용자는 상품에 대한 질의를 대화식으로 할 때 상황 지식을 이용할 개연성이 많으나 기존 방법으로는 이를 해결할 방법이 없다. 일례로, “50만원대 TV를 보여줘”라는 질의 후에 “그 보다 싼 것은?”이란 질의를 처리할 수가 없다. 이와 같은 상황지식을 필요로 하는 질의문장을 처리하기 위한 연구로 [11]이 있으나, 이는 영어권 언어의 CLE시스템을 한국어에 적용한 한글판 CLE 시스템이다. 영어는 대화문장을 처리하기 위해 문법자체에 많은 정보를 제공하는 장치들이 있으나, 한국어의 문법은 느슨하기 때문에 한국어의 문법만으로는 이 시스템을 적용하기에 많은 한계점을 안고 있다. 이를 위해 구문규칙을 이용하기 보다는 한국어를 처리하기에 적합한 지식베이스를 이용한 방법이 모색되어야 한다.

인터넷 응용분야에서 자연어 처리 기술을 이용한 국내

상품검색 엔진에는 다이버, NLIDB, 판도라, 엔써러, AIAGENT 등이 있다[11]. 현재 국내 상품검색 엔진을 적용해서 상용 서비스중인 일부 시스템[12]에서 “삼성이 만든 TV?”라고 입력하면 삼성제품의 TV를 잘 보여준다. 그러나, “딤채 냉장고보다 싼 냉장고?”라고 입력하면 사용자가 원하는 결과를 보여주지 않고 딤채 냉장고를 보여준다. 이는 문장의 의미정보를 파악하기 보다는 형태소 분석 결과로부터 명사를 추출하여 처리하는 것으로 추정된다. 또 한 예로, “90만원대 컴퓨터?”나 “100만원에서 150만원 사이의 컴퓨터”라는 질문에 적절한 결과를 보여준다. 그러나, 같은 유형의 질문이라 할 수 있는 “300만에서 400만 화소대의 디카”라는 질문에는 적절한 결과를 보여주지 못하고 있다. 이는 문장의 의미를 파악하기 보다는 “원”과 같은 가격에 관련한 단어가 검색어에 있으면 이를 예외 처리하는 것으로 보인다.

위 예에서 보여주듯이 현재 인터넷 쇼핑몰은 사용자가 입력한 한국어 질의 문장을 처리하기 위해 단편적으로 자연어 처리 기술을 도입하기 때문에 만족할 만한 결과를 보여주지 못하고 있다.

2.2 한국어 질의 문장

형식언어는 구문의 모호성과 의미의 모호성을 없애기 위해 인공적으로 만든 언어인 반면에 자연언어는 오랜 세월동안 자연 발생적으로 생성, 진화되어온 언어이다. 이에 자연언어는 동의어와 유사어가 발달하였고, 이를 사용함으로써 문장의 구문구조가 서로 다르게 표현될지라도 문장의 의미는 같게 나타낸다.

다음은 동일한 의미를 갖는 다양한 구문구조의 한국어 질의 문장들이다. 한국어의 경우 용언에 따라 문장 구조가 크게 좌우되는데, 어떤 동사와 형용사를 사용하느냐에 따라 문장의 구문구조가 달라진다.

그림 2-1a는 “삼성 TV”를 묻는 여러 유형의 질의 문장이다. 그림 2-1b는 “가격이 저렴한 TV”를 묻기 위해 형용사 “싸다”를 사용한 여러 유형의 구문구조들이다. 그림 2-1c는 50만원 이하의 가격을 묻는 여러 유형의 질의 문장들이다.

- (1) 삼성 TV를 보여줘
- (2) 삼성 TV는?
- (3) 삼성이 만든 TV?
- (4) 삼성에서 나온 TV?
- (5) 제조회사가 삼성인 TV?
- (6) TV중에서 삼성 것?

(그림 2-1a)

- (1) 가격이 싼 TV는?
- (2) 싼 가격의 TV는?
- (3) 싼 TV?
- (4) TV 가격이 싼 것은?
- (5) 어떤 TV가 가격이 싼가?
- (6) 어떤 TV가 싼가?
- (7) TV중에서 가격이 싼 것은?
- (8) TV중에서 싼 것은?
- (9) TV중에서 어떤 것이 싼가?

(그림 2-1b)

- (1) 50만원 이하의 TV?
- (2) 중 저가의 TV?
- (3) 50만원대보다 싼 TV는?
- (4) 가격이 50만원대 보다 싼 TV는?

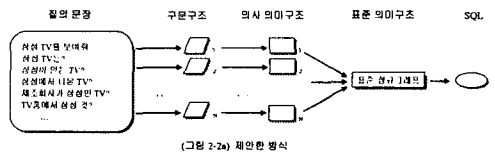
(그림 2-1c)

(그림 2-1) 다양한 유형의 질의문장 묶음들

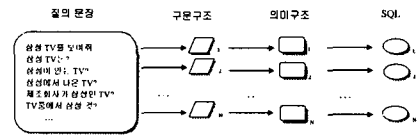
일반 사용자가 질문하는 문장 유형은 다양하지만 앞의 그림의 묶음별로 의미가 동일하므로 결국 하나의 의미구조로 표현할 수 있다. 그림 2-2a는 본 논문이 제안하는

방식으로 그림 2-1a의 질의 문장을 표준 의미구조인 표준 정규그래프로 변환하는 과정을 도식화한 것이다. 그림 2-2a에서 삼성 TV를 묻는 다양한 구문구조의 질의 문장을 표준 정규그래프로 변환한 다음, 이를 데이터베이스 질의어 SQL을 생성하는 후처리부의 입력으로 사용한 다.

그러나, 그림 2-2b는 기존방식으로 이와 같은 질의 문장을 고려하지 않고 각 문장의 구문구조마다 규칙을 적용한 다음 SQL을 생성한다. 기존방식의 의미구조는 질



(그림 2-2a) 제안한 방식



(그림 2-2b) 기존 방식

의 문장의 의미를 표현하기 보다는 SQL 생성에 필요한 릴레이션, 애트리뷰트, 데이터 등을 표현한 것으로 SQL 질의어 생성에 의존적인 정보로 이루어진다.

3. 의미구조 생성을 위한 도메인 지식베이스

이 장은 의사 의미구조를 표준 정규그래프로 변환하는데 필요한 각종 도메인 지식베이스를 기술한다.

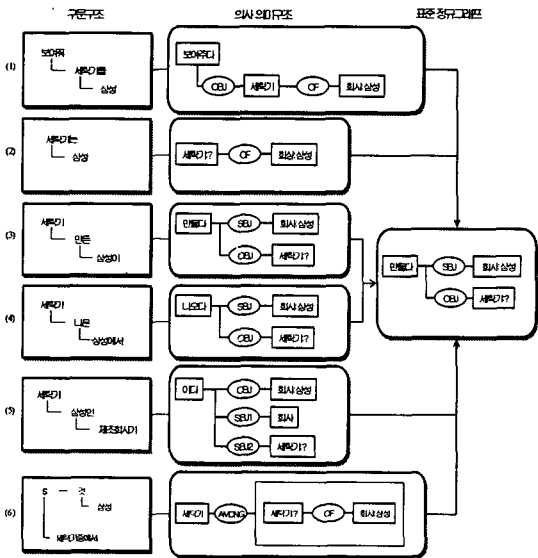
의사 의미구조는 질의문장을 분석하여 이 문장의 표면적인 의미를 나타내는 것으로 이 의사 의미구조는 해결되지 않은 여러 언어적 모호성을 포함할 수 있다. 표 3-1은 한국어의 다양한 질의 유형을 분석하기 위한

질의유형	파스트리 예	도메인 지식	비고
수식구	N(의) M 세탁기 삼성	세탁기 -> V를 -> M 삼성	동일한 의미
술어구	N이 V한 M 세탁기 삼성인 제조회사가	만다 -> C는 -> M 세탁기 -> C는 -> M 삼성인 제조회사가	
	N를 V 보여줘 세탁기를 삼성	만다 -> C를 -> V 보여줘 세탁기를 삼성	
의문형포함	만들었나 어떤 세탁기를 삼성이	만들었다 -> C를 -> V 어떤 세탁기를 삼성이	
		만들었다 -> C를 -> V 어떤 세탁기를 삼성이	

기본단위로 크게 “수식구”, “술어구”, “부사구”, “비교구”로 구분하였다. 첫째, 수식구는 체언이 체언을 꾸미는 경우이다. 둘째, 술어구는 술어를 포함하고 있는 구를 말한다. 셋째, 부사구는 “~중에서”와 같은 부사구를 갖는 경우이다. 넷째, 비교구는 “~보다”와 같은 비교어구를 갖는 경우이다.

표 3-1은 “삼성 세탁기”에 관한 정보를 알기 위해 수식구, 술어구, 부사구를 이용하여 다양하게 표현하고 있지만 의미는 동일하다는 것을 알 수 있다. 이와 반대로 비교구는 똑같은 파스트리를 갖고 있지만 의미는 서로 다르다는 것을 알 수 있다.

따라서, 앞에서 예로 든 이 동일한 의미를 가지는 질의문장들을 패턴매칭을 통하여 그림 3-1과 같이 한 개의 표준 정규그래프로 표현할 수 있다. 이 그림에서 의사의 의미구조는 질의문장의 구문구조를 입력으로 받아 표 3-1에서 제한한 기본 단위의 조합으로 생성한 것으로 이 문장들의 의미가 서로 동일하다는 것을 알 수 없지만 이 문장들의 의미가 서로 동일하므로 이를 하나의 표준 정규그래프로 변환하는데 필요한 도메인 지식들을 기술한다.

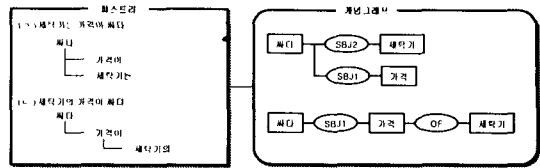


(그림 3-1) 의사의 의미구조를 표준 정규그래프로

3.1 구문구조 변환규칙

한국어는 이종주어를 갖는 술어가 상당히 많다. 예를 들어, 그림 3-2는 이종주어를 취하는 형용사 “싸다”의 질의 문장이다. 문장 (7)은 표층적으로 주격조사 2개를 가지고 있고, 문장 (8)은 주격조사가 1개이다.

그림 3-2에서 보여 주는 문장 (7)과 (8)의 파스트리는 다르지만 문장의 의미는 동일하다. 이를 근거로 (7)의 파스트리로부터 생성된 개념 그래프를 (8)의 개념 그래프로 변환하는 구문구조 변환 규칙을 정의한다. 이 규칙은 이종주어를 취하는 술어를 단일 주어로 단순화하여 표준 정규그래



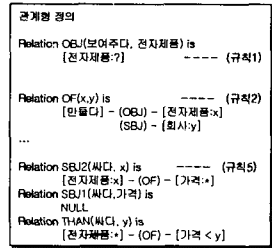
(그림 3-2) 이종 주어 ‘싸다’의 변환규칙 연리

프로 변환을 가능하게 해주는 지식베이스이다.

3.2 관계형 정의

관계형 정의는 실세계 객체들 사이의 관계를 나타낸다. 전자제품 도메인에서 “~ 세탁기를 보여줘”와 같은 유형의 문장은 “~를 보여줘” 대신에 “찾아줘, 알려줘, 원해요, 사고 싶어요, 검색해 줘, 골라 줘, 가 필요해”등으로 정의할 수 있다. 이와 같은 술어들은 대상(OBJ)이 전자제품이므로 이 술어들은 생각하는 규칙한다. 이 경우에 해당하는 변환이 규칙 (1)이다. 예를 들어, “삼성 세탁기를 보여줘”라는 문장은 “삼성 세탁기”를 찾는 것을 의미하기 때문에 “보여줘”는 생각한다.

따라서 앞 문장의 의미구조를 의미변환에 적합한 뒷 문장의 의미구조로 변환할 필요가 있는데 이러한 변환지식을 위한 지식베이스가 관계형 정의이다. 규칙 (2)는 개념 “전자제품”과 개념 “회사”간의 관계를 정의한 것이다. “삼성”과 “세탁기”사이의 “만들다”라는 관계형을 추가하여 정의한 것이다.



(그림 3-3) 관계형 정의의 변환규칙

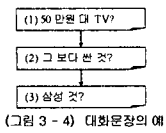
3.3 정규그래프

정규그래프는 실세계에 사용되는 술어와 객체들 사이, 객체와 객체들 사이의 관계를 개념적으로 제약하고 있다. 예를 들어, 표 3-1의 도메인 지식은 정규그래프를 나타낸 것이다. 이중에서 첫 번째 정규그래프는 전자제품과 제조회사는 개념적으로 관계가 있다고 제약하고 있다. 이와 같은 정규그래프는 다음과 같은 경우에 이용된다.

첫째, 부사구를 갖는 질의 문장과 일상생활에서 말하듯이 생각이 많고 지시사를 사용하여 간접하게 질문하는 대화문장을 표준 의미구조로 생성하기 위해 사용한다.

부사구를 갖는 질의문장에서 부사구는 이 부사구를 제외한 문장을 꾸며줄 뿐만 아니라 “것, 곳”과 같은 대응어구의 수반이 상당히 많은 편이다. 대응어구는 앞에서 나온 명사를 대신해서 “것”과 같은 단어로 표현하므로 이것이 가르키는 것을 파악하기 위해 정규그래프를 이용한다.

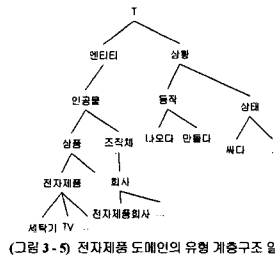
대화문장은 생략현상이 빈번하고 “이”, “그”, “저”등과 같은 대응어구를 포함하고 있다. 예를 들어, 다음과 같은 대화문장을 가정한다. (2)번째 문장의 의미를 파악하



기 위해 지시대명사 “그”와 “것”이 가리키는 것이 무엇인지 알아야 한다. (2)번째 문장에서 지시대명사 “그”와 의존명사 “것”이 가리키는 것을 알기 위해 표 3-1에 있는 “싸다”의 정규그래프를 이용하여 구체화시킨다.

둘째, 트리구조로 이루어진 의사 의미구조를 탐색하면서 매칭 가능한 후보를 찾아 표준 정규그래프로 변환해야 한다. 이때 매칭 가능한 후보를 찾기 위한 근거는 바로 정규그래프이다.

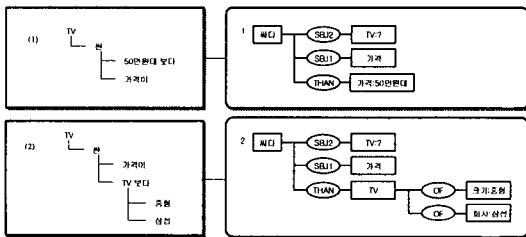
3.4 유형 계층구조



유형 계층구조는 유형간의 상·하위 관계를 보여준다. 지식들의 연산을 위해서는 각 유형들의 상속 관계를 파악할 필요가 있다. 즉, 위 예에서 [세탁기]가 [전자제품]의 하위 유형임을 알 수 있어야 하는데 이때 유형 계층구조를 이용한다.

3.5 액터

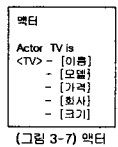
액터를 값을 얻기 위한 일련의 함수처럼 자동으로 실행하는 함수를 말한다. 예를 들어, 비교구를 포함하는 그림 3-6의 문장을 살펴본다.



(그림 3-6) 의사 의미구조

문장 (1)의 의사 의미구조는 표준 의미구조로 변환하기 위해 그림 3-3의 규칙(5)를 적용한다.

그러나, 문장 (2)의 의사 의미구조는 가격이 특정 값으로 정해져 있지 않아서 “삼성 중형 TV”에 해당하는 값을 찾아야 한다. 이 조건에 해당되는 가격을 찾기 위해 TV 테이블에서 크기가 중형이고, 회사가 삼성에 해당하는 가격의 검색이 필요하다. 이와 같이 가격을 검색하기 위한 자동 실행함수를 액터라고 한다.



(그림 3-7) 액터

그림 3-7은 액터 TV를 정의한 것으로 액터 TV는 테이블 이름이고, 테이블의 속성으로 이름, 모델, 가격, 회사, 크기가 있다.

4. 의미구조 생성

이 장은 앞 장에서 기술한 도메인 지식베이스를 이용하여 그림 3-1과 같이 여러 구문구조로 표현된 질의문장의 의사 의미구조를 패턴 매칭을 통해 한 개의 표준 정규그래프로 변환하기 위한 절차를 기술한다.

표준 정규그래프를 생성하기 위한 절차는 다음과 같다.

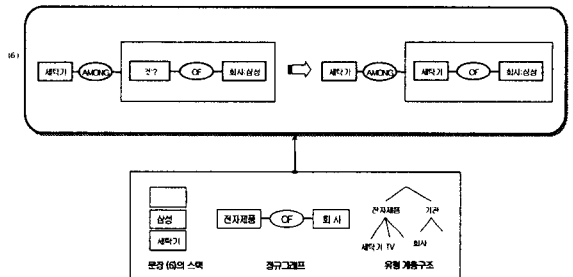
1. 구체화한다 - 지시사 해결과 생략 처리 등
2. 적용 가능한 패턴을 수집한다 - 트리 탐색하면서
3. 패턴의 우선 순위 부여한다
4. 패턴을 하나씩 적용하면서 만일 액터가 존재하면 액터를 실행하고 2 단계로 간다
5. 표준 정규 그래프와 일치하지 않으면 2단계를 처리한다 일치하면 SQL을 생성한다

각 단계가 왜 필요한지를 예를 들어 기술한다. 본 논문은 문장 (3)의 의사 의미구조를 표준 정규그래프로 설정한다.

[1단계] 구체화: 지시사 해결, 생략 처리

“~중에서, ~중”과 같은 부사구를 갖는 질의 문장은 이 부사구를 제외한 문장 전체를 꾸며줄 뿐만 아니라 “것, 곳”과 같은 대용어구의 수반이 상당히 많은 편이다.

대용어구는 앞에서 나온 명사를 대신해서 “것”과 같은 단어로 표현하므로 이 가리키는 것을 파악하기 위해 스택을 이용한다. 예를 들어, 그림 3-1의 문장 (6)에 있는 “것”이 무엇을 가리키는지 알기위해 이 문장의 체언 “세탁기”와 “삼성”을 스택에 넣는다. 그런 다음 정규그래프



(그림 4-1) 문장 (6)에 대한 의미구조 변환과정 중 구체화 단계 예

를 이용하여 스택의 탑에 있는 “삼성”은 “전자제품”이 아니기 때문에 제거되고, “세탁기”는 “전자제품”이므로 그림 4-1과 같은 의미구조를 생성한다.

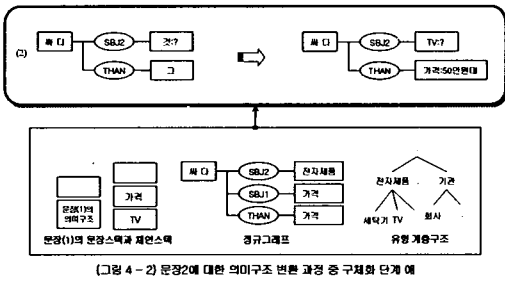
일반 사용자는 대화식으로 간결하게 질문한다. 이러한 대화문장은 생략현상이 빈번하고 “이”, “그”, “저”등과 같은 대용어구를 포함하고 있기 때문에 각 문장이 별개의 의미를 지니기 보다는 문장과 문장간에 의미적으로 연관되어 있다. 단, 문장간의 의미 관련성은 바로 앞에 있는 문장으로만 한정한다.

앞에서 살펴본 그림 3-4의 대화문장을 예로 든다.

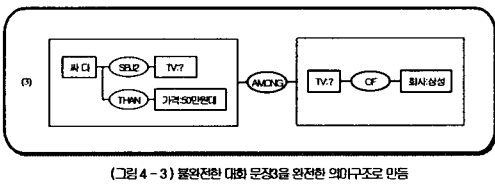
(1)번째 대화문장은 앞에서 기술한 방법과 동일하게 의미구조를 생성한다. 이때, 이 문장에서 품사가 체언인

것을 지식으로 이용한다. 체언에 해당하는 “TV”, “50만원”을 각각 체언스택에 넣는다. 마지막으로, 문장 (1)의 의미구조를 문장스택에 넣는다. 여기서, 스택의 크기는 최대 100문장을 수용할 수 있도록 하였다.

(2)번째 문장인 “그 보다 싼 것?”의 의미를 파악하기 위해 지시대명사 “그”와 “것”이 가리키는 것이 무엇인지 알기 위해 “싸다”의 정규그래프와 스택을 이용한다. “싸다”의 정규그래프는 두 개가 존재한다. 이 중에서 대화문장은 비교구에서 특정가격이 나오지만 첫 번째 정규그래프를 이용한다. 즉, 이 정규그래프를 통해 “그”는 가격을 나타내고, “것”은 “전자제품”에 해당되므로 스택을 이용하여 전자제품 중 TV라는 구체적인 정보를 얻을 수 있다. 그림 4-2는 대화문장에서 의미구조를 생성하기 위한 첫 번째 단계인 구체화 과정을 보여준다.



(3)번째 문장은 “삼성제품의 TV”를 의미하지만, 대화 문장 (2)에 이어 “삼성 것”이라고 질문하면 “50만원대보다 싼 TV중에서”라는 조건이 붙는 대화문장이다. 그러나, 이것은 사람이 판단한 것이고 컴퓨터가 이를 처리하기 위해 바로 앞 문장 (2)의 의미구조를 조건으로 하는 완전한 질의로 만들어야 한다. 즉, 현재문장의 의미구조는 불완전한 질의이다. 그래서, 문장스택에 들어있는 문장 (2)의 의미구조를 관계형 AMONG에 연결하여 그림 4-3과 같은 완전한 질의를 만든다.



[2단계] 트리를 탐색하면서 적용 가능한 패턴 수집
패턴은 정규그래프를 기반으로 매칭 가능한 후보를 말한다. 예를 들어, 그림 4-4에서 점선으로 이루어진 타원형이 하나의 패턴이 된다. 여기서서는 2개의 패턴이 매칭 가능한 후보들이다.

[3단계] 패턴의 우선 순위 부여
매칭 가능한 후보들이 2개이상 존재할 때 어느 후보를 먼저 적용할 것인가 순서를 정하는 것이 패턴의 우선 순위이다.

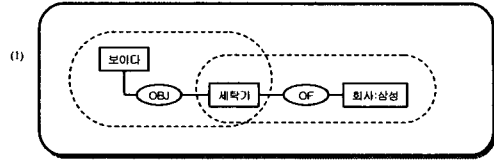
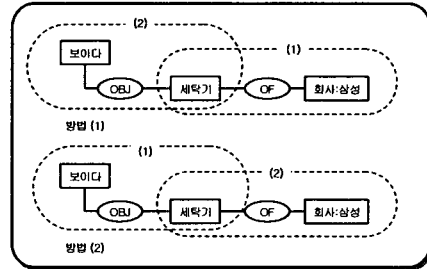
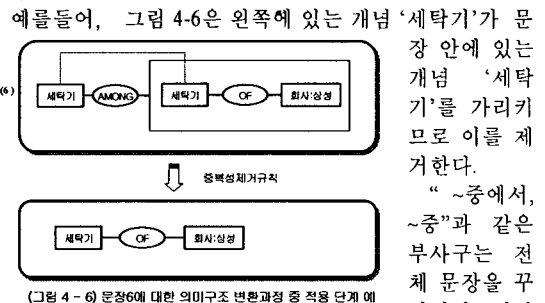


그림 4-5는 패턴이 2개 이상 존재할 때 어느 패턴을 먼저 적용하느냐에 따라 방법1과 방법 2가 존재한다. 방법 1은 체언구를 먼저 한 다음 술어구를 적용한 것이다. 방법 2는 술어구를 적용한 다음 체언구를 적용한 것이다. 여기서, 본 논문은 방법1보다 방법 2를 적용하는 것이 효과적이다. 왜냐하면 방법 2에서 패턴(1)은 관계형 정의 규칙(1)에 의해 생략된다. “보이다”와 같은 술어구는 생각이 가능하므로 개념그래프를 단순화한 다음 규칙을 적용한다.



[4단계] 패턴을 하나씩 적용하면서 만일 액터가 존재하면 액터를 실행하고 2 단계로



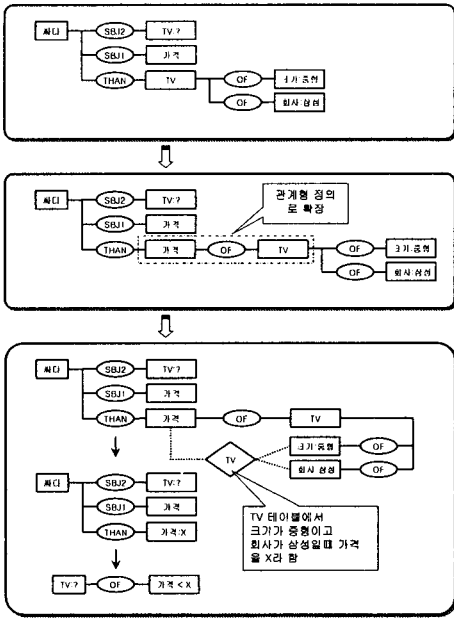
예를들어, 그림 4-6은 왼쪽에 있는 개념 ‘세탁기’가 문장 안에 있는 개념 ‘세탁기’를 가리키므로 이를 제거한다.

“~중에서, ~중”과 같은 부사구는 전체 문장을 꾸미지만 의미적으로 불

전체 문장 중 특정 개념을 한정하기 때문이다.
그림 4-7은 TV 테이블에서 가격을 찾기 위해 액터를 적용한 것이다. 점선은 TV 테이블의 속성을 나타낸다.

[5단계] 표준 정규 그래프와 일치하지 않으면 2단계 처리하고 표준 정규 그래프와 일치하면 SQL 생성한다.
4단계의 결과가 미리 설정한 표준 정규그래프와 비교하여 일치하면 SQL 생성하는 단계로 간다. 그러나, 일치

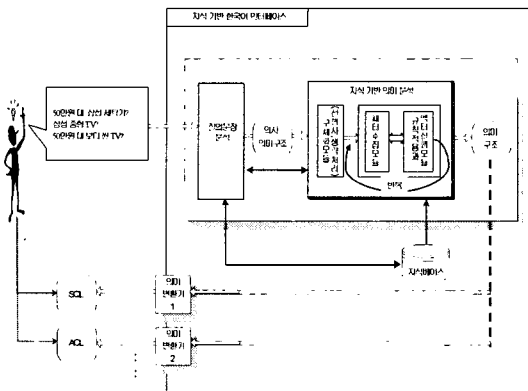
하지 않으면 패턴을 수집하여 다시 적용한다.



(그림 4-7) 역터 적용 단계의 예

5. 지식 기반 한국어 인터페이스의 구조

본 지식 기반 한국어 인터페이스는 질의 문장을 분석하여 문장의 의미를 파악할 수 있는 의미구조를 생성하는 것이다. 그림 5-1은 본 시스템의 기본적인 구조로 질의 문장 분석과 지식기반 의미 분석으로 크게 나눌 수 있다.

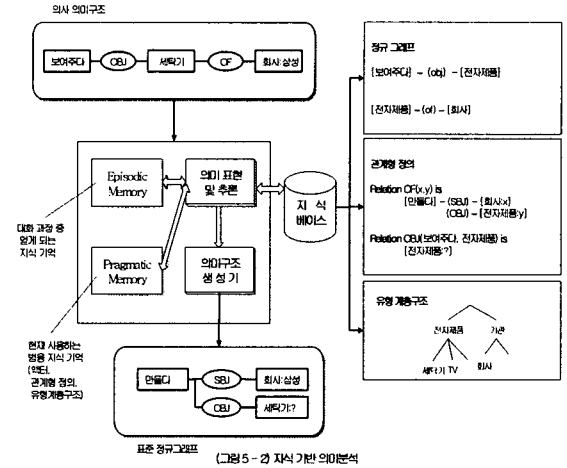


(그림 5-1) 지식 기반 한국어 인터페이스의 구조

질의문장 분석은 문장의 표면적인 의미를 나타내는 의사 의미구조를 생성하는 것으로 이 의사 의미구조는 해결되지 않은 여러 언어적 모호성을 포함할 수 있다. 지식 기반 의미 분석은 도메인 지식베이스를 이용하여 의사 의미구조를 의미구조로 변환한다.

5.1 지식기반 의미분석

표준 정규그래프를 생성하는 지식 기반 의미 분석의 구성이 그림 5-2에 나타나 있다. 그림 3-1의 문장(1)을 예로 들어 보여주고 있다.



(그림 5-2) 지식기반 의미분석

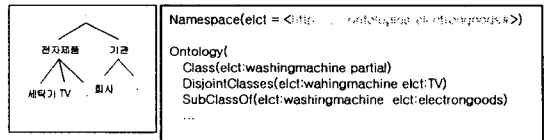
5.2 실험

본 논문은 실험문장으로 학부생으로부터 전자제품을 도메인으로 하는 질의문장을 약 2500여개 수집하였다. 질의문장을 분석해보니 크게 표 3-1과 같이 6가지의 유형으로 분류할 수 있었다.

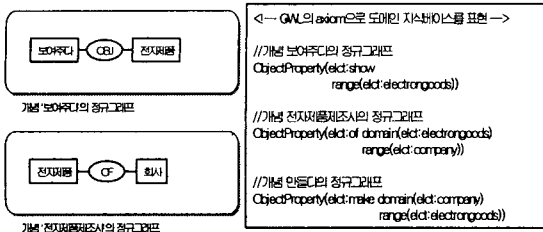
본 논문에서 사용하는 sowa의 개념구조는 자연어 문장의 표현력이 우수하고 술어논리로 변환이 용이하지만 구현이 어렵다는 단점이 있다. 그러나, 온톨로지 기술인 OWL은 술어논리에 기반하고 있어서 논리 기반추론[8]이 가능하며, OWL DL은 계산할 수 있는 한에서 최대의 표현력을 제공한다.

따라서, 본 논문은 온톨로지 저작도구 protege[16]와 Racer[17] 추론엔진을 연결하여 온톨로지 추론을 이용하여 의미구조의 생성을 보인다. 이를 위해 그림 3-1의 문장 (1)에 있는 술어구 질의 문장을 예로 들어 OWL과 추론규칙으로 변환하였다(그림 5-4와 그림 5-5 참조).

그림 5-3은 유형 계층구조의 일부를 온톨로지 OWL 문서로 표현한 것이다. elct는 전자제품 도메인 온톨로지를 의미한다.



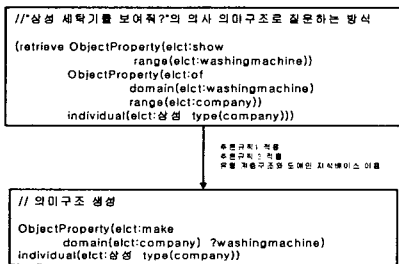
(그림 5-3) 유형계층구조(좌), 온톨로지 OWL문서 일부(우)



(그림 5-4) 도메인 지식베이스(가전, OWL언어서)의 표현



(그림 5-5) 관계형 정의의 규칙(좌), 추론 규칙(우)



(그림 5-6) 의사 의미구조(좌), 의미구조(아래)

6. 결론 및 향후 연구과제

기존의 한국어 인터페이스는 특정 응용 분야에 대한 한국어 인터페이스를 구현하기에는 비교적 용이하나 다른 응용분야로 적용하는 확장성은 떨어지므로 응용 분야에 대한 확장성을 갖기 위한 도메인 지식과 상황 지식을 사용하여 사용자의 질의를 의미 구조로 표현할 수 있는 도메인 지식 기반 한국어 인터페이스에 대하여 기술하였다.

아울러, 하나의 의미가 다양한 문장 형태로 표시되어도 일관적으로 처리할 수 있기 때문에 이를 표준 의미 구조로 변환하였고, 대화문장의 불완전한 의미의 질의문장을 완전한 의미로 채워나가기 위해 상황지식과 도메인 지식을 사용하였다.

앞으로 향후 연구과제로 본 논문에서 생성된 의미구조를 입력으로 하는 데이터베이스 질의어 생성에 관한 연구가 필요하다.

참고문헌

[1] Codd, E., "RENDEZVOUS version1: and experimental English-Language query formulation system for casual users of relation databases," IBM Res. Rep. RJ2144, San Jose, Calif., 1978.

[2] Barbara, J., "TEAM: A Transportable Natural Language Interface System," Artificial Intelligence Center DRI International Menlo Park, CA94025

[3] Hendrix, Sacerdoti, E, Sagalowicz, D and Dlocum, J., "Developing a Natural Language Interface to Complex Data," ACM TODS, pp. 105-147, 1978

[4] Kameny I., Weinor J., Criley M., Burger J., Gates R. and Brill D., "EUFID: The End User Friendly Interface to Data Management Systems," IEEE, 1978, pp. 380-391

[5] M. Bates, M. G. Moser, D. Stallard, "The IRUS transportable natural language database interface," in Expert Database Systems, Benjamin Cummings, 1986

[6] X. Wu and T. Ichikawa, "KDA: A Knowledge-Based Database Assistant with a Query Guiding Facility," Transactions on Knowledge and Data Engineering, vol. 4, No. 5, pp. 443-453, October, 1992.

[7] 김재문, 김재홍, 이상조, 키워드를 이용한 자연언어 질의 시스템 설계 및 구현, 한국정보과학회 '95 봄 학술 발표논문집, 22권 1호, pp. 957-960, 1995

[8] Michael K. Smith, Chris Welty, Deborah L. McGuinness, "OWL Web Ontology Language Guide," WWW Consortium, <http://www.w3.org/TR/owl-guide>, 2003

[9] 채진석, 김성기, 이석호, 한국어 데이터베이스 검색을 위한 질의 시스템의 설계 및 구현, 한국정보과학회논문지, 20권 61호, pp.810-820, 1993

[10] 박인철, 한국어 문형을 이용한 개념그래프 생성에 관한 연구, 박사학위 논문, 1998

[11] 박현석, 한국어 ATIS 질의문의 QLF 표현과 분석방법에 관한 고찰, 한국인지과학회논문지, 10권 2호, pp.17-25, 1999.

[12] <http://www.digquest.com>, <http://www.unisoft.co.kr>, <http://www.guark.co.kr>, <http://www.answercr.co.kr>, <http://ontoline.com>

[13] <http://www.gseshop.co.kr/isp/main.jsp>, <http://kr.vsp.shopping.yahoo.com/vsp/samsungmall/>, <http://www.korca.com/>

[14] I. Androutsopoulos, G. Ritchie, and P. Thanisch, An Efficient and Portable Natural Language Query Interface for Relational Databases, Proceedings of the 6th International conference on Industrial & Engineering Application Conference on industrial & Engineering Applications of Artificial Intelligence and Expert Systems, 1993.

[15] R. Dale, H. Moisl, and H. Somers(Eds.), A Handbook of Natural Language Processing, Marcel Dekker Inc., 2000.

[16] <http://protege.stanford.edu/>

[17] <http://www.sts.tu-harburg.de/~r.f.moeller/racer/>