결합범주문법을 이용한 자연언어 인터페이스 [1]

이호동○ 박종철
한국과학기술원 전산학전공 및 첨단정보기술연구센터
{hdlee,park}@nlp.kaist.ac.kr

Natural Language Interface with Combinatory Categorial Grammar

Hodong Lee○ Jong C. Park
KAIST, Computer Science Division and
Advanced Information Technology Research Center

요 약

본 연구에서는 전자상거래 데이터베이스를 대상으로 결합범주문법을 이용한 자연언어질의 인터페이스를 구현한다. 이를 위해 질의문을 분석하고 표현 방법을 논의한다. 또한 SQL 형식언어로 변환하기 위한 어휘 표현 및 유도 방법을 보인다. 제안하는 방법은 구문분석 과정에서 SQL 형식의 질의문을 직접 유도하는 것으로 기존 연구에서 제안됐던 중간논리언어 변환단계를 거치지 않으므로 과정이 간결해져 시스템의 성능향상을 가져올 수 있다. 시스템은 웹 기반과 client/server 구조로 구현된다.

1 서 론

자연언어질의 인터페이스는 정보의 제공을 쉽고 자연스럽게 하기 위한 목적으로 개발되고 있다 [3, 6, 8]. 이와 같은 연구에서는 데이터베이스에 대한 저장된 정보의 접근 수단으로 QUEL, QBE, SQL 등과 같은 형식언어 대신 자연언어를 이용하고자 한다.

본 연구는 쇼핑몰과 같은 전자상거래 도메인에서 결합범주문법[10]을 이용하여 대상 데이터베이스 질의언어인 SQL로 질의문을 분석하고 변환하는 작업을 다룬다. 대상 도메인은 불특정 다수로부터 특정 분야에 대한 다양한 요구가 발생하는 특징을 가지고 있어 실험 도메인으로 선정되었다. 결합범주문법은 어휘중심 문법으로 결합 규칙을 이용하여 넓은 언어현상의 범위를 다룰 수 있다는 장점을 지니는데 본 연구에서는 한국어를 위한 결합범주문법[1, 2, 5]을 직접 개발하여 질의문을 분석하였다. 결합범주문법에서는 통사, 의미적인 자연언어 분석에 필요한 정보들이 어휘(lexical entry)에 할당되어 문장을 분석하게 되므로 어휘사전의 구축이 중요한 작업이다. 따라서 본 논문에서는 실험 도메인의 질의문에서 SQL 표현이 유도될 수 있게 하는데 필요한 어휘의 통사, 의미 정보 표현 형식에 대해 중점적으로 설명한다. 특히 본 연구에서는 질의문 분석을 통하여 중간단계인 논리언어 표현 및 변환과정이 없이 데이터베이스 질의언어를 유도하므로 질의문 변환과정의 투명성을 개선할 수 있고 어휘정보의 사용을 특정 과정에 한정함으로써 어휘에 대한 모듈화를 제공할 수 있다. 또한 중간단계 표현에 필요한 데이터베이스 정보나 도메인 데이터에 관한 정보를 따로 필요로 하지 않으므로 이와 관련된 정보 구축 작업을 줄일 수 있다는 장점을 가진다.

2장에서는 기존 연구들을 살펴보고, 한국어에 대한 결합범주문법을 소개한다. 3장에서 시스템의 구조를 설명하고 4,5장에서 질의문을 처리하여 SQL로 유도하는 방법을 보인다.

[1]본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

2 관련 연구

2.1 자연언어질의 인터페이스

데이터베이스에 대한 자연언어질의 인터페이스에 관한 연구는 1960년대부터 계속되어 왔다. 초기에는 간단한 패턴 매칭의 방법을 사용하여 질의문을 처리하였지만 최근 연구에서는 넓은 범위의 자연언어 질의문을 처리하기 위해 문법규칙을 사용하여 시스템을 구축하고 있다. [6]에서는 학과성적 관리 도메인을 대상으로 하고 있는 객체지향 데이터베이스에 대하여 한국어 질의를 객체지향 데이터베이스 질의 언어인 OQL로 변환하는 시스템을 소개하였다. 이 연구에서는 한국어 질의문의 구성요소를 정의된 기본패턴으로 분류하여 질의문들의 논리적 구조를 파악할 수 있도록 하였다. [3]은 오스트리아 사회보장 도메인에 대하여 명사구 단위의 독일어 질의문으로 데이터베이스에 질의할 수 있도록 하는 연구를 소개하고 있다. 이 연구에서는 질의문을 말뭉치로부터 유도한 문맥자유문법 규칙들로 분석하여 Quasi-Logical Form 형태의 의미 표현으로 변환하여 처리한다. [8]는 시간 데이터베이스에 시간 표현을 담는 절이나 구로 수식된 영어 질의문 유형을 처리하도록 하였다. 이 연구에서는 자연언어 질의문을 Type Logic Grammar로 구문 분석하여 시간 표현이 가능한 데이터베이스 질의 언어인 SQL/Temporal로 변환한다. 이들 연구에서는 공통적으로 질의문을 미리 정의된 중간단계 논리언어/표현형식을 거친 후 대상 데이터베이스 형식언어로 변환하고 있다. 이러한 방법은 시스템에서 도메인에 관련된 부분을 분리할 수 있어 DBMS에 독립적인 시스템을 구축할 수 있고 어휘 정보와 도메인 정보를 분리시켜 새로운 지식 도메인에 대한 이식성을 높일 수 있다는 장점이 있다 [4]. 그러나 이러한 방법은 따로 도메인 정보를 구축하는 작업을 필요로

하고, 도메인에 한정되지 않는 중간 단계 표현언어를 정의해야 되는 부담을 갖는다. 본 연구에서는 어휘사전이 도메인 의존적인 정보를 담고 있으므로 시스템은 DBMS에 독립적일 수 있고, 도메인 정보를 구축하는 작업을 어휘 사전 구성만으로 해결하므로 충분한 이식성을 확보할 수 있다. 또한 중간 단계 언어를 정의하는 노력이 필요 없고 시스템의 과정이 간결해 짐으로써 질의문 변환과정의 투명성을 개선하여 작업의 유연성을 증가시킨다.

2.2 결합범주문법

단일화 기반의 어휘문법인 결합범주문법(CCG)[10]은 각 어휘마다 범주가 할당되는데, 범주는 문법 정보 외에 의미 정보나 담화 정보까지도 포함할 수 있다. 이러한 범주는 축약규칙에 의해 처리된다[2, 5]. Forward application의 경우, functor 'X/Y'의 오른쪽에 논항 'Y'가 나타나면 축약규칙이 적용되어 결과로 'X'를 내어준다. 여기서 사선 '/'는 오른쪽에서 논항을 받고 '\'는 왼쪽에서 논항을 받는다는 정보를 제공한다. Composition은 functor와 functor를 합성하는 연산이고 type-raising은 논항끼리 합성해야 할 필요가 있을 경우에 functor의 범주로 변환하여 합성을 가능하게 만드는 규칙이다. Coordination은 접속사를 중심으로 좌우의 통사범주가 동일한 성분들을 묶어주는 규칙으로, 병렬구문을 처리하는데 필요한 규칙이다. CCG를 사용하여 문법을 기술하였을 때의 이점은 다른 문법체계에서 처리할 수 없거나 특별한 약정을 사용하여 해결하는 문형을 결합자를 통한 축약규칙만으로 처리할 수 있다는 것이다. 또한 한번의 유도과정으로 통사적 분석과 의미적 분석을 얻어낼 수 있고 담화정보인 정보구조를 동시에 얻어내는 것도 가능하다.

3 시스템 구조

본 논문에서 제안하는 자연언어질의 처리 엔진은 그림 1과 같이 크게 파서, 응답 생성 부분과 어휘 사전, 데이터베이스로 구성된다.

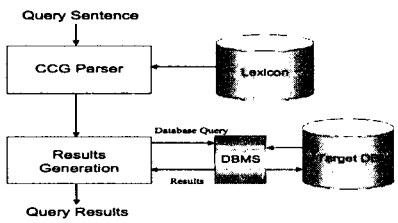

그림 1: 자연언어질의 처리 엔진 구조

어휘 사전은 파싱과정과 복합명사 처리의 효율성, 제한된 도메인의 특성을 고려하여 어절단위로 구성하였다. 이러한 사전으로부터 CCG 파서는 입력된 문장을 어절단위로 처리하는데 복합명사와 같은 경우는 등록된 최장일치 어휘를 찾아 해당 범주를 통해 의미를 추출한다. 파서는 차트 기반의 CKY 알고리즘을 사용하였다.

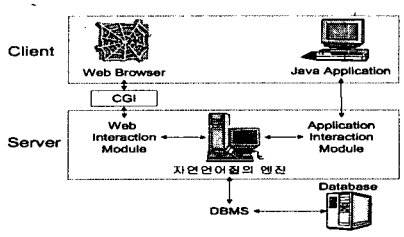

그림 2: 전체 시스템 구조

추출된 리스트 형태의 의미는 SQL로 변환되어 응답생성 모듈을 통해 해당 DBMS에서 결과를 얻어온다. 이 밖에 응답생성 모듈에서는 DBMS를 통해 질의된 결과에 해당되는 데이터를 받아서 사용자가 보기 쉽도록 가공해주는 작업을 수행한다. 이러한 작업에는 시스템의 인터페이스에 따라 부합되는 형식으로 변환하는 작업과 데이터베이스에서 고유하게 정의된 특수한 code나 ID 등과 같은 데이터들을 사용자가 이해할 수 있는 해당 이름으로 변환해 주는 작업이 해당된다.

현재 구현되고 있는 전체 시스템은 그림 2과 같은 클라이언트/서버 구조를 가지고 있다. 자연언어질의 시스템의 엔진부분은 주로 SICStus Prolog와 Java로 구현되어 있으며 2만개의 어휘로 구성된 어휘사전을 지닌다. 또한 서버는 웹브라우저와 Java 응용 형태의 클라이언트와 통신하기 위한 모듈을 지닌다.

4 자연언어질의 의미표현

자연언어질의 시스템에서 질의로 표현되는 문장들은 한정된 도메인에 대한 것이기 때문에 데이터베이스에서 의미 있는 응답을 얻기 위한 어휘도 한정된다[7, 9]. 그러나 표현되는 언어현상은 일반 단문에서와 같이 다양하게 나타날 수 있다.

4.1 질의문 예제의 분석

본 연구에서 대상으로 하는 쇼핑몰과 같은 도메인의 전자 제품에 대해 대학원 학생들을 대상으로 질의문을 조사하였다. 질의문은 총 200개로 질의문에 합당치 않은 개인적 주관이나 SQL문으로 표현하기 힘든 연산 표현이 있는 문장, 대상 도메인의 범위를 넘는 문장을 제외한 147문장에 대해 분석하였다. 이 중 주어, 목적어, 보어, 서술어 등의 필수격에 간단한 부사어나 수식어 등이 첨가된 절이 포함되지 않은 예제가 32문장으로 '벨로체 디지털 피아노의 제조사는?', '디오스의 가격은 얼마인가?', '삼성 TV의 리스트를 보여주세요.' 등과 같이 간단한 질의를 표현하는데 사용된다. 절이 포함된 단문 예제가 가장 많이 나타났고 이 중 관형절이 59개로 가장 많이 나타났다. 관형절은 주로 명사로 표현된 제품을 설명하는데 제약이나 조건을 나타내는 구조로 사용되어 질의문에 자주 발생하는 것으로 보인다. 이러한 관형절을 처리하기 위해 관형격 어미에 'np/np\s'와 같은 CCG 범주를 할당함으로써 처리할 수 있다. 병렬구문이 포함된 예제는 39개가 나타나 질의문에서 병렬구문의 처리가 중요함을 알 수 있다. 이 밖에 예제에 나타난 문형은 표 1에 나타나 있다.

분 류	문장 갯수
절이 포함되지 않은 단문	32
절이 포함된 단문(관형절, 명사절)	61
병렬 구문	39
종속 관계	9
담화 형식	1
철자 오류	5
잘못된 질의	53
계	200

표 1: 질의문의 문장 유형

4.2 도메인에 기반한 어휘의 표현

예제 데이터베이스는 판매하는 상품에 대한 정보를 중심으로 하는 Product라는 한 개의 테이블로 구성되며 테이블은 제품 관련 정보와 쇼핑몰, 제조사에 관한 정보로 구성되어 있다. 구현하는 시스템은 질의문에 대한 한번의 CCG 파싱으로 SQL2 형식의 질의를 직접 유도하므로 이러한 작업을 위해서는 어휘에 통사정보와 SQL을 유도할 수 있는 의미를 부여하는 작업이 필요하다. 각 어휘는 예 (1)과 같은 형식으로 표현된다.

(1) lex('가격이', np:[_,_,'가격'=A]).

lex('냉장고의', np:[_,Product,'분류'='냉장고'&B]/[_,A,B]).

(1) 형식의 사전에서는 어휘, 형태소명, 통사범주/의미정보의 순서쌍으로 이루어진다. 통사범주/의미정보는 CCG를 적용하기 위해 ':' 왼편에 나타나는 구문범주와 ':' 오른편에 나타나는 대괄호 '[ ]'안의 의미정보로 구성된다. 의미정보를 표현하는 대괄호 표현은 SQL의 SELECT/FROM/WHERE 질의 각 내용을 구성하는 속성, 테이블, 조건에 해당되는 의미를 담게 된다. 형태소명은 같은 단어에 대해 문장에서 다르게 사용되는 형태소와 구분하여 처리하기 위하여 필요한 정보이다. 위의 방법과 같은 어휘 표현 방법에서는 특정 데이터베이스 도메인에 대한 어휘 사전의 정보 의존도가 크기 때문에 대상으로 하는 데이터베이스 및 질의 언어에 대해 어휘 사전을 새롭게 구축해야 된다는 단점이 있지만, 이러한 점은 중간단계 표현언어를 가지는 다른 방법에서도 나타나는 문제점이다. 반면에 이 방법에 비해 필요한 데이터베이스 관련 정보 및 대상 질의언어로의 변환 과정에 필요한 데이터베이스로의 변환 정보를 따로 구축할 필요가 없다는 장점이 있다.

5 SQL 질의문으로의 변환

SQL질의문은 2.2절에서의 축약규칙을 적용하여 수행된다. 입력된 질의문에서는 '$' 기호를 삽입하여 문장의 종결을 표시한다. 이 기호는 질의문에 대한 최종적인 의미를 부여할 수 있도록 올바른 문장에 대해 (2)와 같은 범주가 할당된다.

(2) sql:[[Select,A],[From,B],[Where,C]]\s:[A,B,C]

(2)의 범주는 질의문의 입력이 종결될 때 SQL 질의언어 형식으로 질의문의 정보들을 통합하는 역할을 한다. 질의문이 명사구일 경우를 처리하기 위해서 '$'에 (3)의 범주를 할당한다.

(3) sql:[[Select,A],[From,B],[Where,C]]\np:[A,B,C]

이러한 범주를 포함하여 질의문을 파싱하여 최종적인 SQL의 SELECT/FROM/WHERE의 의미를 같는 의미 리스트를 생성해 낼 수 있다. 본 논문에서는 지면상 자세한 유도 과정은 보이지 않는다.

6 결론

자연언어질의 인터페이스를 구현하기 위해 결합범주문법 기반의 파서를 구현하고 어휘사전을 구축하였다. 도메인은 쇼핑몰과 같은 전자상거래 사이트의 전자제품 대상으로 하고 있으며 이에 대한 실험 데이터베이스가 구축되고 있다. 어휘사전은 한국어 질의문에서 나타나는 언어현상에 대한 분석을 기반으로 SQL 문장으로 변환될 수 있도록 구축되었다. 제안된 데이터베이스 질의문 변환 방법은 데이터베이스 구조 정보 및 도메인 정보가 어휘사전에 통합되어 표현됨으로써 질의문 분석을 통해 SQL 문장으로 변환되도록 하였다. 이러한 점은 중간단계 논리언어의 도움을 받는 기존은 방법들과 달리 논리언어 변환과정을 거치지 않음으로써 성능향상을 기대할 수 있는 이점을 가진다.

참고문헌

[1] 조형준, 박종철. 한국어 병렬문의 통사, 의미, 문맥분석을 위한 결합범주문법. **정보과학회논문지**, 27(4):448–462, 2000.

[2] 조형준. **한국어 병렬구문과 결합범주문법에서의 구문분석**. 석사학위논문, 한국과학기술원 전산학과, 2000.

[3] A. Klein, J. Matiasek, and H. Trost. The treatment of noun phrase queries in a natural language database access system. In *COLING-ACL'98 workshop on the computational treatment of nominals*, pages 39–45, 1998.

[4] I. Androutsopoulos, G. D. Ritchie, and P. Thanisch. Natural Language Interfaces to Databases - An Introduction. *Natural Language Engineering*, 1(1), 1995.

[5] J. C. Park and H. J. Cho. Informed Parsing for Coordination with Combinatory Categorial Grammar. In *COLING*, pages 593–599, 2000.

[6] J. Chae and S. Lee. Identifying Basic Patterns of Korean Natural Language Query. In *NLPRS*, pages 606–611, 1995.

[7] M. Mosny. Semantic Information Preprocessing for Natural Language Interfaces to Databases. In *ACL*, pages 314–316, 1995.

[8] R. Nelken and N. Francez. Querying Temporal Databases Using Controlled Natural Language. In *COLING*, pages 1076–1080, 2000.

[9] M. Rayner. Natural-Language Database Interfacing From First Principles. In *2nd Symposium on Logical Formalizations of Commonsense Reasoning*, 1993.

[10] Mark Steedman. *The Syntactic Process.* MIT Press, 2000.