# 구조화된 생물의료 정보의 접근을 위한 자연언어 질의 시스템 [1]

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

# Natural Language Query Interpretation System for Biomedical Database Access

Hodong Lee Jong C. Park
Division of Computer Science and AITrc
Korea Advanced Institute of Science and Technology

요 약

본 논문은 이질적인 데이터베이스에 산재되어 있는 생물의료 정보의 개념적인 접근을 가능하게 하기 위한 자연언어질의 시스템을 설명한다. 이를 위해 본 시스템에서는 질의문을 SQL, OQL, CPL 데이터베이스 정형언어로 변환하는데, 이 과정에서 필요한 질의문의 분석 및 변환과정을 보인다. 제안하는 방법은 구문분석에 의해 도출된 정보를 이용해 직접 다양한 정형언어들로 변환하므로, 시스템의 구조가 간결해지고 모듈화되어 전체 성능과 이식성의 향상을 가져올 수 있다.

## 1 서 론

최근 관심이 늘어나고 있는 생물의료 정보에 대한 연구는 특히 정보에 대한 접근과 통합, 정보의 추출과 유추의 측면에서 활발히 이루어 지고 있다 [1, 2]. 자연언어질의 시스템은 이러한 연구에서 정보의 접근에 대한 수단을 보다 쉽고 자연스럽게 제공하기 위해 개발되고 있다. 이와같은 연구에서는 주어진 질의문을 데이터베이스를 위한 질의 언어인 SQL, OQL, TSQL, CPL 등으로 변환함으로써 이루어 진다 [3, 4, 5, 6, 7, 8].

본 연구에서는 생물의료 도메인에서 결합범주문법을 이용하여 데이터베이스의 정보에 접근하기 위한 정형언어로 질의문을 변환하는 시스템의 구조 및 방법에 대해 다룬다. 대상으로 하는 생물의료 도메인은 그 정보의 양이 방대하고 다양한 데이터베이스 (혹은 정보 저장소) 에 정보들이 산재되어 있으므로 그 환경이 이질적이라는 (heterogeneity) 특징을 내재하고 있다. 이러한 특징은 정보의 접근을 어렵게 하는 요소로 작용하는데, 특히 사용자로 하여금 데이터베이스 구조나 특징에 대한 자세한 정보를 필요로 하는 경우도 있으므로 사용자의 원하는 정보에 대한 개념적인 도메인 지식만으로 정보에의 접근이 힘들게 된다.

(1) UCH37에 작용하는 단백질은?
(Which proteins interact with UCH37 ?)

1과 같은 예문에서 이러한 개념적인 질의에 대한 답을 통합적으로 얻기 위해서는 그 정보가 위치한 데이터베이스들의 이름, 테이블 이름, 속성 이름 등의 구체적인 정보를 알고 있어야 한다. 더욱이 데이터베이스에 대한 폼, 메뉴 기반과 같은 질의 수단이 제공되지 않을 경우, 각 데이터베이스에 적절한 질의 언어까지 알고 있어야 한다. 본 논문에서 제안하는 시스템은 이러한 사용자의 개념에 대해 구체적인 하위 단계의 지식 없이 원하는 정보에 접근할 수 있는 수단을 제공할 수 있다. 또한 관계형 데이터베이스를 위한 질의 언어인 SQL, 객체지향 데이터베이스를 위한 OQL [9], 관계형 및 구조화된 데이터 파일 형태의 데이터베이스 - 정보 저장소 (data repository) 와 같은 - 를 위한 CPL [1]로 변환하므로 이질적 환경의 생물의료 도메인의 다양한 정보원에 접근할 수 있는 수단을 제공할 수 있다.

2장에서는 기존 연구들을 살펴보고, 한국어 결합범주문법을 소개한다. 3장에서 시스템의 구조를 설명하고 4장에서 데이터베이스 질의문으로 변환하는 방법을 보인다.

## 2 관련 연구

### 2.1 결합범주문법

단일화 기반의 어휘문법인 결합범주문법 (CCG)[10]은 각 어휘마다 문법, 의미, 담화 정보를 담는 범주가 할당되는데, 특별한 약정없이 축약규칙만을 통해 문법 정보 외에 의미 정보나 담화 정보까지 한번의 과정으로 유도할 수 있다는 장점을 지닌다 [11, 12]. 축약규칙에서 forward application (>)의 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.

(2)

| UCH37에 | 작용하는 | 단백질은? |
|---|---|---|
| $np$ | $(np/np)\backslash np$ | $np$ |
| | $np/np$ (<) | |
| | $np$ (>) | |

예 2은 예문 1에 대해 간단한 문법정보만으로 유도과정을 보이고 있다. 이 예에서는 backward application 규칙이 명사와 관형사에 적용되고 다시 그 결과가 forward application 규칙을 통해 명사와 결합하여 명사구로 분석된다.

### 2.2 자연언어질의 시스템

자연언어질의 시스템에 대해 최근에 있었던 연구들은, SQL로 변환하는 것을 다루는 연구 [3, 4, 5]와 OQL로의 변환을 다루

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

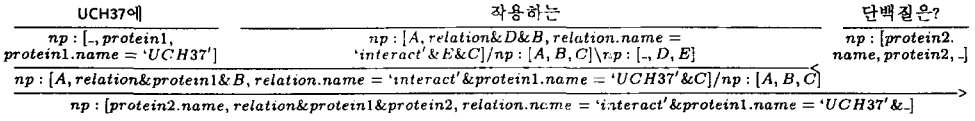


그림 1: 예문 1의 CCG 유도 과정

는 연구 [6], 시제나 시간에 관련된 표현으로부터 이를 표현할 수 있는 질의언어인 SQL/Temporal 또는 TSQL로의 변환을 다루는 연구 [7, 8] 등 다양한 연구들이 있다. 이 중 본 논문에서는 분산된 데이터베이스로의 질의문 처리를 다루는 TAMIC-P 시스템에 대해서 소개한다 [5, 13].

TAMIC-P는 여러 관계형 데이터베이스에 분산 저장되어 있는 오스트리아 사회보장 도메인의 정보에 대하여 명사구 단위의 독일어 질의문으로 접근할 수 있도록 하는 시스템으로 예문 3와 같은 명사구 형태의 질의어로 질의문을 한정하고 있는데 이는 해당 도메인을 이용하는 사무원들이 완전한 문장보다 단순한 형식의 질의문을 선호하기 때문인 것으로 설명된다.

(3) Ersatzzeiten wegen Kindererziehung

*(Exemption times because of child raising)*

특히 예문 3와 같은 명사구 단위의 질의문은 제안된 도메인에서 다양한 형태의 완전한 문장으로 대치될 수 있는데, 사용자가 이러한 특징을 이용하여 관련된 정보를 모두 편리하게 얻을 수 있다는 장점이 있다. 이 시스템에서는 도메인 정보와 개념 정보 구조를 통해 분산된 데이터베이스의 정보를 관리한다. 그러나 이 시스템은 문맥자유문법 수준의 단순한 규칙만을 이용하여 자연언어를 처리하기 때문에 병렬이나 다양한 부사가 포함된 질의문들을 처리하지 못하고 단지 SQL만을 생성함으로써 접근할 수 있는 데이터베이스의 종류가 제한된다. 본 시스템에서는 기존의 연구를 통해 CCG를 이용하여 명사구뿐만 아니라 다양한 언어표현들을 처리하고 다양한 형태의 데이터베이스에 대한 접근이 가능하도록 SQL, OQL, CPL에 대한 처리를 다루고 있다 [4]. 본 연구의 방법은 정형 논리형태의 표현이나 SQL/Temporal과 같은 형태로도 적용이 가능하다 [14].

## 3 시스템 구조

본 시스템은 클라이언트-서버 구조로 서버는 크게 자연언어질의 처리엔진, 데이터베이스 질의언어 생성 모듈, 클라이언트 연결 모듈로 나뉜다. 그림 2는 시스템의 구조와 외부와의 연동 관계를 보이고 있다. 그림 2에서 시스템은 클라이언트로 부터 자연언어질의를 입력받아 자연언어질의 처리엔진에서 분석을 거친 후, 해당 데이터베이스의 질의 언어를 생성하여 이질적인 형태의 정보원에 접근, 정보를 통합하여 다시 해당 클라이언트에게 결과를 반환한다.

시스템의 모듈인 자연언어질의 처리엔진은 자연언어의 처리를 위한 어휘사전과 어휘에 대한 데이터베이스 이름, 테이블 이름, 속성 이름 등의 데이터베이스 정보들의 삽입을 위한 도메인의 클래스 정보들을 이용하여 질의 분석 결과를 생성한다. 그리고 데이터베이스 질의언어 생성 모듈은 질의 처리엔진의 결과를 이용하여 정형언어인 SQL, OQL, CPL을 생성한다.

정형언어의 생성에 있어 각 데이터베이스의 종류에 따라 그 목적이 되는 질의문법의 종류가 결정되므로 이 과정에서는 분산된 데이터베이스의 종류와 사용 질의언어에 대한 문법정보를 이용한다. 또한 분석결과로 생성된 의미정보를 표현하는 문법이

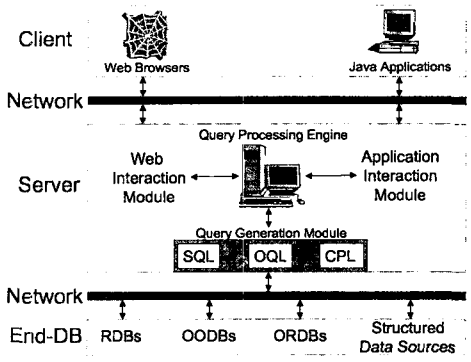


그림 2: 시스템 구조

각 언어의 종류마다 차이가 있으므로 이러한 표현을 위한 도메인 정보도 필요하다. 예를 들어, 데이테베이스에서 테이블 또는 데이터간의 관계를 표현하기 위해 SQL에서는 결합 경로 (join path)가 사용되고 OQL에서는 경로 표현 (path expression)이 사용되며 CPL에서는 OQL과 비슷한 방법으로 복합 객체를 표현한다. 이러한 표현은 각 질의언어의 표현에 있어 핵심이 되는 요소로, 사용되는 도메인 정보는 데이터베이스 스키마 정보의 테이블 관계로부터 추출된다. 이 과정은 질의 처리엔진에서의 공통된 분석 결과를 이용하므로 질의 처리엔진은 한번의 분석단계만을 거치게 되어 시스템의 효율성을 제고할 수 있다.

마지막으로 클라이언트 모듈은 사용자의 입력을 받아 필요없는 입력부분을 제거하는 간단한 필터링 과정과 결과를 클라이언트의 형식에 알맞게 보여주는 포맷팅 과정을 수행한다.

## 4 데이터베이스 언어로의 변환

자연언어질의 처리엔진으로부터 생성되는 결과는 질의문의 문법 및 도메인에 대한 의미정보를 담고 있다. 예문 1은 그림 1와 같은 유도과정을 통해 의미정보를 생성한다. 의미정보는 [A,B,C]와 같은 형태로 구성되는데, A는 사용자가 질의하고자 하는 정보에 대한 테이블 또는 클래스, B는 질의문이 수행되는 테이블 및 클래스, C는 질의문의 조건에 관한 정보가 담기게 된다. 이러한 사항들은 본 연구의 대상이 되는 질의언어인 SQL, OQL, CPL에 모두 공통이 되는 사항이므로 이를 통해 각 질의언어로 변환할 수 있다. 단 앞장에서 설명한 SQL의 결합 경로, OQL의 경로 표현, CPL의 객체 경로 (object path)는 각 데이터베이스의 스키마를 구조를 표현하는 도메인 지식으로부터 유추할 수 있다. 이러한 구조는 SQL의 경우 테이블 간의 내재된 결합 관계로, OQL의 경우 그래프 구조로, CPL의 경우 트리 구조로 표현되는데, 이러한 구조의 탐색을 통해 경로를 찾고 이를 결과질의문에 대치 (substitute) 시킴으로써 원하는 표현을 생성할 수 있다. 이런 과정을 통해 생성된 그림 1에 대한 정

형질의는 SQL, OQL, CPL에 대해 예 4에서 차례대로 보이고 있다.

(4) (1)
```
SELECT p2.name
FROM  relation, protein p1, protein p2
WHERE relation.pid1=p1.pid and relation.pid2=p2.pid
    and relation.name='interact' and p1.name='UCH37'
```

(2)
```
SELECT p2.name
FROM  r in relation, p1 in r.pid1, p2 in r.pid2
WHERE r.name='interact' and p1.name='UCH37'
```

(3)
```
{ z.#name | \x <- relation, \y <- x.#protein1,
            \z <- x.#protein2, x.#name="interact",
            y.#name="UCH37"}
```

변환된 예제 4에서 SQL로 변환하기 위한 예는 기존의 연구에서 연구되었다 [4]. 그러나, 본 연구에서는 SQL에 의존적인 데이터베이스 질의언어 문법이 각 어휘 정보에 포함되지 않고 각 질의언어로 변환되기 위한 의미정보가 담기게 된다. 이러한 점은 OQL 및 CPL로의 확장을 위해 필요한 작업이다. 예를 들어, 질의문에서 '-면', '-중'과 같이 종속문이나 부사구의 중첩관계를 나타내는 표현이 사용될 경우 'condition', 'among'과 같은 의미를 사용하여 이를 나타내고 이 관계들을 통해 각각의 질의언어로 맵핑할 수 있는 방법을 사용한다. 그러나 각 질의언어의 표현력 차이때문에 이러한 관계들이 제대로 표현되지 못할 경우도 있는데, 특히 SQL 같은 경우에는 복합 객체와 같은 리스트와 중집합 (bag)의 개념을 완전히 지원하지 못하기 때문에 'among'과 같은 중첩관계의 표현에 있어 자연스럽게 맵핑이 일어나지 못할 수 있다.

예제 4는 그림 1의 결과인 의미정보에 대해 변환과정을 쉽게 설명하기 위한 단순한 예로 실제 데이터베이스에 맵핑하기 위해서는 각 데이터베이스가 가지는 데이터베이스 이름, 테이블/클라스 이름, 속성 이름 등에 대한 스키마 정보가 사용되어야 한다. 예 1의 의미정보는 한 예제로 생물의료 정보 도메인으로부터 추출된 클라스 정보를 담고 있는데, 이러한 정보는 특정 데이터베이스에 의존적이지 않은 개념적인 정보를 담고 있어야 한다. 이러한 개념적인 정보로부터 각각의 데이터베이스에 대한 맵핑 정보가 구조화 되어 구축되면 생성된 의미정보로부터 실제 데이터베이스에 알맞은 정형질의를 생성하는 과정에서 각 데이터베이스에 대해 질의문이 범위에 맞는 지를 검사하는 과정을 통해 대상 데이터베이스들을 선정하고 이 데이터베이스들에 대해 도메인 정보를 사용해 맵핑하고 알맞은 문법으로 변환함으로써 최종적인 데이터베이스 질의문이 도출된다.

## 5 결론

본 연구에서는 자연언어질의문을 이질적인 환경의 분산된 생물의료 도메인의 데이터베이스에 접근하기 위해서 SQL, OQL, CPL로 변환하는 방법을 수행하였고 이를 위한 시스템의 구성과 변환을 위해 필요한 작업들에 대해 설명하였다. 이러한 방법을 통해 본 연구에서 제안된 시스템은 관계형 및 객체 지향, 객체 관계형 데이터베이스 뿐만 아니라 생물의료 도메인에 많이 존재하는 파일 형태의 정보 저장소와 같은 데이터베이스에 대해서도 자연언어를 통한 개념적인 접근을 제공할 수 있다. 이러한 점은 결합범주문법 체계를 이용한 시스템의 높은 이식성과 모듈화로 인하여 가능한 것으로 생물의료 도메인의 정보가 가지는 특징인 그 구조의 높은 복잡도와 양의 방대함, 이질적 형태의 정보원 등을 극복하여 정보에의 접근 및 통합을 이룰 수 있는 방안을 보이고 있다.

## 참고문헌


[1] L. Wong. Kleisli, a Functional Query System. *Journal of Functional Programming*, 10(1):19–56, 2000.

[2] J. C. Park. Using Combinatory Categorial Grammar to Extract Biomedical Information. *IEEE Intelligent Systems*, 16(6):62–67, 2001.

[3] X. Meng, S. Wang, and K. F. Wong. Overview of A Chinese Natural Language Interface to Databases: NChiql. *Computer Processing of Oriental Languages*, 14(3):213–232, 2001.

[4] H. Lee and J. C. Park. Translating Natural Language Queries into Formal Language Queries with Combinatory Categorial Grammar. In *International Conference on Computer Processing of Oriental Languages*, pages 41–46, 2001.

[5] 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.

[6] J. Chae and S. Lee. Frame-based Decomposition Method for Korean Natural Language Query Processing. *Computer Processing of Oriental Languages*, 11(4):353–379, 1998.

[7] I. Androutsopoulos, G. D. Ritchie, and P. Thanisch. Time, Tense and Aspect in Natural Language Database Interfaces. *Natural Language Engineering*, 4(3):229–276, 1998.

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

[9] R. G. G. Cattell. *The Object Database Standard: ODMG-93*. Morgan Kaufmann, 1996.

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

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

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

[13] J. Matiasek, A. Klein, and H. Trost. TAMIC-P: A System for NL Access to Social Insurance Databases. In *Applications of Natural Language to Information Systems*, pages 209–214, 1999.

[14] H. Lee and J. C. Park. Automatic Augmentation of Translation Dictionary with Database Terminologies in Multilingual Query Interpretation. In *ACL-EACL Workshop on Human Language Technologies and Knowledge Management*, pages 113–120, 2001.