제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

생물학적 개체명 사전을 위한 구축 및 관리 도구에 관한 연구

장현철, 김태현, 이현숙, 박수준, 박선희
한국전자통신연구원 컴퓨터시스템연구부 바이오정보연구팀
e-mail : {janghc, heemang, lhs63473, psj, shp}@etri.re.kr

A Study on Construction and Management Tools for Biological Named Entity Dictionary

Hyunchul Jang, Taehyun Kim, Hyunsook Soojun Park, Seonhee Park
Bioinformatics Research Team, Computer System Research Dept.,
Electronics and Telecommunications Research Institute

요 약

바이오 텍스트 마이닝을 위한 정보 추출의 첫 단계는 생물학적 문헌으로부터의 유전자, 단백질, 세포조직 등과 같은 생물학적 개체명의 인식이다. 생물학적 개체명의 명명법상 특징이 매우 다양하고 저자의 개성에 의해 쉽게 좌우되어 단순히 규칙이나 학습 방법 만으로는 쉽게 개체명들을 인식할 수 없다. 또한, 생물학 관련 문헌에 나오는 가능한 모든 개체명과 이들의 모든 변형을 수록하는 것은 현실적으로 불가능하므로 이를 해결하기 위해 이미 알려진 개체명에 대해서 기본적으로 사전을 탐색하고 알려지지 않은 용어들을 규칙과 통계 기반 방법을 통하여 인식하는 것이 효과적이다. 그러나 만족할 만한 수준의 양질의 사전을 구축하는 것은 쉽지 않을 뿐만 아니라 많은 비용이 소요되며, 어느 순간 만족할 만한 성능을 낼 수 있는 사전을 구축했다 할지라도 유지 관리 하는 것이 결코 쉬운 일이 아니며 마찬가지로 많은 비용을 필요로 하게 된다. 따라서, 잘 구축된 자원으로부터 필요한 정보를 추출하여 적절한 사전을 자동으로 구축하여 활용하는 방법을 사용할 경우, 사전 구축 및 관리에 드는 많은 비용을 줄이면서도 상당히 효과적인 성능을 얻을 수 있을 것이다. 본 연구에서는 바이오 텍스트 마이닝 엔진을 위한 생물학적 개체명 사전을 자동으로 구축하고 이를 쉽게 관리하도록 하는 도구를 개발하였다.

1. 서론

생물학 관련 문헌의 급속한 증가 추세로 인해 생물학 관련 분야의 논문이나 지식 데이터베이스 등에 있는 텍스트로부터 최대의 부가가치를 창출하기 위해 유용한 지식 정보를 자동적으로 추출해 내는 기술에 대한 연구가 진행되고 있는데 이것이 바이오 텍스트 마이닝이다. 바이오 텍스트 마이닝 기술은 생물학 관련 문헌을 대상으로 자연어 처리 및 정보 추출 기술을 적용하여 생물학적 개체들에 대한 정보와 이러한 개체들 간의 관계를 추출함으로써, 이들의 기능과 상호관계를 유추해 내는 기술이다.

이를 위한 가장 기본적인 기술이 생물학적 요소들의 개체명, 즉 단백질이나 유전자 등의 이름을 추출하고 인식하는 기술이다. 이는 단백질, 유전자, 약물, 질병 등의 생물학적 개체들과 관련된 이름들을 개체명 사전, 명명규칙, 개체명 추정 기법 등을 이용해 인식함으로써, 정보의 주체들을 추출하는 기술이다.

개체명 인식을 위한 접근방법은 크게 규칙 기반의 개체명 인식과 통계 기반의 개체명 인식, 그리고 두 가지 방법을 통합한 하이브리드 방식이 있다. 규칙 기반의 방법은 생물학적 개체명 인식을 위한 규칙을 수작업으로 구축하고, 다양한 사전을 이용하여 개체명을 추출하는 방법이다. 규칙 기반의 연구는 가장 높은 성능을 내지만 응용 도메인에 따라 사전 및 규칙이 수동으로 바뀌어야 하므로 비용이 많이 든다는 단점이 있다. 통계에 기반한 방법은 학습 데이터로부터 개체명 인식에 필요한 지식을 자동적으로 학습하는 방법

으로 결정 트리, 은닉 마르코프 모델(HMM), 최대 엔트로피를 이용한 방법이 대표적이다. 이 방법은 규칙 기반의 방법에 비해 유연성이 있지만, 대량의 학습 데이터를 구축하는데 많은 비용과 시간을 필요로 한다. 하이브리드 방식은 규칙 기반의 방법과 통계 기반의 방법을 통합하여 좀 더 나은 성능을 얻기 위한 목적으로 통계 기반의 모델에 규칙이나 어휘 정보, 사전 정보 등의 다양한 지식들을 결합하는 방식이다[K. Fukuda, Denys Proux]. 생물학적 개체명은 규칙이나 통계에 의해 쉽게 추출하기가 쉽지 않다. 또한, 통계 기반을 위한 대용량의 학습 데이터를 구축하기도 쉽지 않다. 본 연구에서는 이를 해결하기 위해 이미 알려진 개체명에 대하여 잘 구축된 리소스를 활용하여 사전으로 구축하고 사전을 탐색하여 개체명을 인식하는 방법을 기본으로 하고 규칙 및 통계 기반 방식을 병행하는 방법을 사용한다.

2. 관련연구 및 배경

생물학적 개체명은 명명법 상 특징이 복잡한데 생물학적 개체명은 대소문자를 구별하여 사용되며 대소문자를 구분하지 않고 문자열을 다루는 경우 서로 다른 개념들을 지칭하는 경우가 발생할 수도 있다. 일반적인 용어에는 포함되지 않는 알파벳이 아닌 문자, 기호나 숫자들이 사용된다. 일반적인 텍스트 처리에서는 기호로 분리될 것들이 개체명에 포함되어 있는데, '+', '-'와 같은 기호는 물론 '(', ')'와 같은 괄호들도 개체명에 포함될 수 있으며 전치사나 접속사도 개체명에 포함될 수 있다.

사용법 상 동일한 개체를 다양한 형태의 이름으로 표현하는데 'epidermal growth factor receptor', 'EGF receptor', 'EGFR'나 'c-Jun', 'c-jun', c jun'이 모두 같은 개체명이다. 텍스트 마이닝을 위해서는 다르게 표현된 한 개체명을 동일한 의미로 인식할 수 있어야 한다. 이는 개체간의 관계를 인식하였을 때 같은 개체에 다르게 표기된 한 동일한 개체의 한가지 기능을 여러 개체가 같은 기능으로 이 개체에 작용한 것으로 분석되면 안되기 때문이다.

그리고 형태상(morphologically) 다양한 이름을 가지기도 하는데, 복수형이나 품사의 다양성으로 'enzyme activity', 'enzyme activities', 'enzymatic activity'가 동일한 개체명이며 문장 구성상 다양한 이름을 가질 경우는 'enzyme amidolytic activity', 'activity of enzyme', 'enzyme and bactericidal activity'가 동일한 개체명이 된다.

또한, 표현된 category 에 따라 전혀 다른 의미나 다른 기능을 가지게 되는 개체도 존재하게 되므로 단순히 용어와 문법적인 쓰임 외에 category 등 복잡한 구조를 가지게 된다.

이상과 같이 바이오 텍스트 마이닝을 위한 사전이 일반적인 자연어처리나 정보검색을 위한 사전과는 다른 복잡한 특성이 있다. 따라서, 생물학적 문헌에서 개체명 인식을 위해 사용된 사전의 품질은 개체명을 성공적으로 인식하기 위한 기본 조건이 될 수 밖에 없다. 생물학 관련 문헌에 나오는 가능한 모든 개체명과 이들의 변형(약어, 복수형 등)이 수록되어 있다면 개체명 인식 모듈의 역할 중 사전 탐색을 제외한 기능은 사소한 것들이 될 것이다. 그러나 현실적으로 이러한 사전을 구축하는 것은 거의 불가능하며[Daniel Hanisch], 특정 분야에 적합한 한 분류에 적절한 사전을 따로 구축하는 것도 어렵다. 또한 어느 순간 만족할 만한 성능을 낼 수 있는 사전을 구축했다 할지라도 유지 관리 하는 것이 결코 쉬운 일이 아니며 막대한 비용을 필요로 하게 된다.

따라서, 잘 구축된 자원으로부터 필요한 정보를 추출하여 적절한 사전을 자동으로 구축하여 활용하는 방법을 사용할 경우, 사전 구축 및 관리에 드는 많은 비용을 줄이면서도 상당히 효과적인 성능을 얻을 수 있다.

UMLS(Unified Medical Language System)는 다양한 생물의학 정보 자원들로부터 얻은 정보를 검색하고 통합하는 일을 용이하게 하기 위한 목적으로 시작된 프로젝트로서 생물의학 문헌의 명세, 임상 자료, 원시 데이터뱅크(Factual Databanks), 지식 기반 시스템, 인간과 생물의 디렉토리 등의 자원들을 대상으로 하고 있다. 이들 자원을 통합하여 메타시소러스(Metathesaurus), 시멘틱 네트워크(Semantic Network) 그리고 SPECIALIST 사전(lexicon)을 제공한다. 메타시소러스는 생의학적 개념들에 대한 다양한 이름과 상호 관계 등과 같은 의미 정보를 포함하고 있으며 많은 표준 임상 및 생의학 어휘들도 포함하고 있다. 시소러스, 분류, 코딩 시스템 등 다양한 기관들에서 개발되고 관리되는 통제어들의 목록 등을 이용해서 만들어졌다. 시멘틱 네트워크는 메타시소러스에 있는 개념들이 포함되는 분류상의 구분 및 시멘틱 타입들에 대한 정보 네트워크이며, SPECIALIST 사전은 생의학 용어들에 대한 구문론적 정보를 제공한다. UMLS 는 이처럼 여러 생물학 리소스에서 사용하는 어휘를 개념(Concept) 중심으로 통합하는 프로젝트로서 자동으로 사전 생성하기 위한 리소스로 매우 적당하다[UMLS].

CUI	LAT	TS	LUI	STT	SUI	STR	LRL
Unique identifier for concept	Language of Term	Term status	Unique identifier for term	String type	Unique identifier for string	String	Least Restriction Level

Sample Records

C0002871|ENG|P|L0002871|PF|S0013742|Anemia|0|
C0002871|ENG|P|L0002871|VP|S0013787|Anemias|0|
C0002871|ENG|P|L0002871|VC|S0352787|ANEMIA|0|
C0002871|ENG|P|L0002871|VC|S0414880|anemia|0|
C0002871|ENG|P|L0002871|VO|S0470197|Anemia, NOS|3|
C0002871|ENG|S|L0280031|PF|S0803242|Anaemia|3|

CUI	TUI	STY
Unique identifier for concept	Unique identifier of Semantic type	Semantic type. The valid values are defined in the Semantic Network.

Sample Records

C0002871|T047|Disease or Syndrome|

그림 1. UMLS 메타시소러스의 MRCON(상)과 MRSTY(하)

그림 1 은 UMLS 의 메타시소러스에 있는 파일의 구조이다. 위의 테이블이 MRCON 이고 아래의 테이블이 MRSTY 이다. MRCON 은 메타시소러스에 있는 유일한 각각의 문자열(string)의 의미를 기술하기 위한 파일이며, MRSTY 는 각 개념(Concept)에 할당된 각 시멘틱 타입(Semantic Type)을 기술하기 위한 파일이다.

GO(Gene Ontology)는 생물학적 프로세스(Biological Process), 분자 기능(Molecular Function), 세포 조직(Cellular Component) 등에 관한 통제 어휘를 제공하기 위한 목적으로 다양한 단체들이 자율적으로 참여하여 만들고 있는 온톨로지(Ontology) 공동체(Consortium)이다. GO 에 의해 제공되는 통제 어휘(GO term)들은 모두 유전자 생성물질(Gene Product)의 속성(Attribute)으로 사용될 수 있다[GO]. GO term 들은 주로 주어진 유전자 생성물질이 가담할 수 있는 프로세스와 기능 들을 명시하는 데 사용된다. 이는 유전자 생성물질들이 이런 관계를 가질 때 특정한 프로세스나 기능들을 명확히 기록할 수 있도록 한다[David Dehoney]. 따라서, GO term 들이 관계 인식을 위한 사전 생성의 리소스로 적당하다 할 수 있다.

3. 사전의 자동 생성

본 연구에서는 사용자가 생성한 리소스에 UMLS 자원과 GO 자원을 더하여 사전을 구축한다. 특히 생물학적 개념 어휘를 통합한 UMLS 의 Metathesaurus 를 활용한다. UMLS Metathesaurus 의 각 개념의 이름들과 시멘틱 네트워크(semantic network) 체계를 활용하여 생물학적 어휘들을 자동으로 분류한 후, 분류된 어휘들로부터 자동으로 사전을 생성한다. 여기에 추출된 개체들간의 관계를 인식하기 위하여 GO 에서 추출된 용어들을 추가한다. 이 용어들은 주로 생물학적 프로세스나 기능들을 식별하는데 사용된다.

본 연구에서는 특히 생물학적 개념 어휘들을 통합한 UMLS Metathesaurus 에 초점을 두었다. UMLS Metathesaurus 의 개념 이름과 시멘틱 네트워크를 분석하고 생물학적 어휘들을 자동으로 분류한 후 분류된 어휘들로부터 사전을 구축하였다. 여기서 추출된 어휘들을 개체명 인식 모듈에서 개체명을 찾는데 사용하도록 하였다.

UMLS 메타시소러스의 MRCON 과 MRSTY 의 CUI(Unique identifier for Concept)는 각 개념들에 부여된 식별번호로 사전은 기본적으로 CUI 를 기준으로 필요한 각 파일들을 통합한 후, SUI 단위로 사전으로 구축하였다.

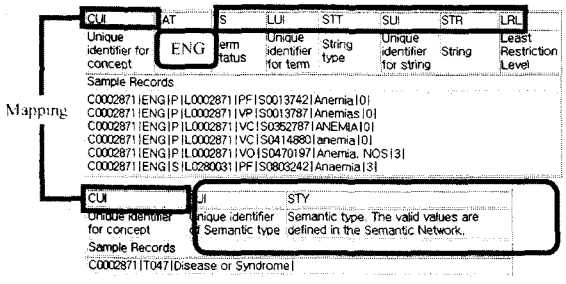

그림 2. UMLS 메타시소러스 MRCON 과 MRSTY 의 관계와 사전 추출

위와 같이 사전을 구축하게 되면 UMLS 메타시소러스에 구축된 어휘들은 그림 3 과 같은 개념들로 쉽게 분류할 수 있게 되는데 MRSTY 파일의 TUI(Unique identifier of Semantic Type)에 의해 이미 분류가 되어 있기 때문이다. 이는 대상으로 하는 생물학적 연구 목적에 따라 표 1 과 개념들 중에서 적당한 개념들을 선택하면 그에 해당하는 어휘들을 쉽게 추출하고 이들을 바탕으로 자동으로 사전을 구축할 수 있게 한다.

```
Entity
  Physical Object
    Organism
      Plant
        Alga
      Fungus
      Virus
      Rickettsia or Chlamydia
      Bacterium
      Archaeon
      Animal
        Invertebrate
        Vertebrate
          Amphibian
          Bird
          Fish
          Reptile
          Mammal
            Human
    Anatomical Structure
      Embryonic Structure
      Anatomical Abnormality
        Congenital Abnormality
        Acquired Abnormality
  ...
```

표 1. UMLS Semantic Type Hierarchy 의 일부분

이렇게 구축된 개체명 사전에 추출된 개체들 간의 관계를 인식하기 위하여 GO term 들을 추가하였다. GO term 들은 주로 조직체들의 분자 기능과 생물학적 프로세스들을 지칭하는데 사용된다.

본 연구에서는 텍스트 파일로 저장된 리소스로부터 개체명 사전은 MySQL 을 사용하여 데이터베이스 테이블로 저장하고 검색이 필요한 부분에 색인을 생성하도록 하였으며, 개체명 인식 모듈에서 개체명 인식을 위해 빈번한 검색이 필요한 내용들은 필요할 경우 B-Tree 를 따로 사용토록 하였다.

4. 사전 구축 및 관리 도구

지금까지 설명한 생물학적 개체명 사전의 자동 생성 및 관리를 위하여 본 연구에서는 활용 가능한 자원으로부터 자동으로 사전을 구축하고 유지할 수 있는 도구를 제작하였다. 사전 구축 및 관리 도구는 앞에서 기술된 개체명 사전의 구축은 물론 검색 및 유지 보수하기 위한 관리를 지원하기 위한 도구이다. 본 연구에서는 UMLS 자원에 GO 를 추가적으로 사용하고 개발된 도구를 사용하여 두 자원을 통합 관리 할 수 있도록 하였다. UMLS 와 GO 는 서로 다른 저장 형식과 구조, 분류 체계를 가지므로 이를 통합적으로

관리 운영할 수 있도록 구축되고 색인 되었으므로 관리 도구는 이를 지원하기 위한 기능을 지원하고 사전의 구조 변경 시에도 동적으로 이를 관리할 수 있어야 한다. 이를 위해 사전 구축 및 관리 도구는 다음과 같은 세부 기능을 만족하도록 개발하였다.

4.1 사전 구축 기능

(1)개체명 사전 생성 - 사용자가 생성한 리소스와 UMLS, GO 사이트로부터 다운로드 받은 리소스로부터 자동으로 개체명 사전을 구축한다.

(2)개체명 사전 갱신 - 갱신된 전체 리소스나 변경 사항이 담긴 파일(Update files)을 사용하여 원본 리소스의 갱신 사항을 사전에 반영한다. UMLS 나 GO 는 새로 갱신된 부분에 대하여 따로 파일로 제공하는데 이 파일들을 이용하여 기 구축된 개체명 사전에 쉽고 편리하게 추가할 수 있는 기능을 제공한다. 또한, 갱신된 전체 리소스를 사용하여 갱신된 내용을 판별하고 변경 사항을 반영한다.

4.2 사전 관리 기능

(1)백업(Backup).

(2)분리(Split) 및 통합(Merge) – 기 구축된 사전에서 특정 조건에 일치하는 일부분을 분리하거나 추출하여 또 하나의 사전으로 구축할 수 있다. 또한 분리된 사전을 기 구축된 사전에 통합할 수 있다.

(3)들여오기/내보내기 – 위의 기능에 부가적으로 사전의 전체 및 일부분을 파일(Plain Text, XML)로 내보내거나 이러한 파일을 사용하여 들여오기가 가능해야 하며 이러한 기능은 새로운 사전을 생성할 때에도 가능해야 한다. 이 기능은 각 생물학적 정보 시스템이나 바이오 데이터 마이닝 시스템 간에 효율적인 정보 교환 및 공유가 가능하도록 한다.

(4)색인 – 검색 및 개체명 인식의 빠른 수행을 위하여 저당된 데이터베이스를 이용하여 색인을 생성하거나 별개의 탐색 구조를 구축할 수 있다.

(5)사전의 편집 – 리소스의 변경과 관계 없이 사전의 내용을 변경(추가/갱신/삭제)할 수 있어야 하고 다음 리소스를 사용한 갱신 시 편집된 내용이 보전되거나 무시될 수 있어야 한다.

(6)오류 보고 - 사전의 구축, 변경 및 검색 과정에서 발생하는 다양한 상황을 이해하고 제어하기 위해 로그 파일 형태로 처리 정보 및 오류 정보를 기록한다.

(7)통계 – 구축된 사전의 통계 정보를 제공한다.

(8)검색 – 구축된 사전의 내용을 검색할 수 있다. 구축에 사용된 각 리소스 별로 제공하는 기본 질의 방식을 포함하여 사전을 조건 별로 검색한다. 데이터베이스에 저장하였을 경우 색인되지 않은 필드를 검색하거나 테이블 형태의 출력을 제공하기 위해 데이터베이스의 검색 기능을 이용한다. 검색 결과는 백업, 분리 및 통합, 들여오기 및 내보내기 등에 사용될 수 있다. 일반적인 검색 기능을 사용하며 일치 검색(start with, end with, contain, match)과 대소문자 구별(case sensitive) 검색을 제공한다. 사용자 GUI 를 사용하여 트리뷰(Tree View)를 이용한 계층적 검색이 가능하도록 한다. 특히, 검색 기능은 사용자 응용프로그램과 웹에서 편리하게 사용할 수 있도록 자바 패키지화 하였다.

개발 플랫폼으로 JAVA 를 도입함으로써 UMLS 에서 제공되는 MetamorphoSys, Lexical 툴 및 라이브러리를 본 영역에 쉽게 활용할 수 있게 되어 시스템 통합성 및 개발 기간 단축에 도움이 될 수 있었다.

개체명 사전을 웹 브라우저를 통해 쉽고 빠르게 검색할 수 있도록 검색 기능의 일부를 Tomcat4 서버와 JSP 를 연동하여 구축하였다.

5. 결론 및 향후 연구 방향

본 연구에서는 생물학적 문헌의 텍스트 마이닝을 위한 개체명 사전을 구축하고 관리하는데 드는 비용을 줄이고 양질의 사전을 자동으로 구축하는 방법과 도구를 개발하였다. UMLS 와 GO 와 같은 잘 구축된 자원을 활용할 경우 많은 장점을 얻을 수 있으며 쉽게 사전을 구축할 수가 있다.

생물학적 문헌의 텍스트 마이닝을 위해서는 개체명의 인식 외에 개체간의 관계 추출이 필수적인데 이를 위해서는 개체명간의 관계를 인식하기 위한 관계 용어 사전 구축과 생물학 관련 키워드 사전을 구축해야 한다. 또한 생물학적 용어들과 일반 용어들을 구분하고 각 용도로 쓰일 수 있는 공통의 용어에 대하여 문헌 내에서 어떤 의미로 쓰였는지 판별해 낼 수 있어야 한다.

보다 다양한 자원을 활용할 수 있는 연구가 필요하며 이를 위해 개체명 사전과 관리 도구의 범용성을 높일 수 있도록 사전 구조를 설계하고 관리 도구를 개발하여 기능 및 내용 추가가 용이하도록 하는 연구가 필요하다.

참고문헌

[K. Fukuda] K. Fukuda, T. Tsunoda, A. Tamura, T. Takagi, "Toward IE: Identifying protein names from biological papers," Proceedings of the Pacific Symposium on Biocomputing (PSB1998).

[Denys Proux] Denys Proux, Francois Rechenmann, Laurent Julliard, "Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction," Genome Informatics, 9, pp. 72-80, 1998.

[UMLS] http://www.nlm.nih.gov/research/umls/ UMLS Project Web Site

[GO] http://www.geneontology.org/ Gene Ontology Consortium Web Site.

[David Dehoney] David Dehoney, Rachel Harte, Yan Lu, and Daniel Chin, "Using Natural Language Processing and the Gene Ontology to Populate a Structured Pathway Database," IEEE CSB 2003 poster paper, 646-647, 2003.

[Daniel Hanisch] Daniel Hanisch, Juliane Fluck, Heinz-theodor Mevissen, Ralf Zimmer, "Playing Biology's Name Game: Identifying Protein Names in Scientific Text," In Proceedings of the Pacific Symposium on Biocomputing. pp. 403-414. 2003.