UMLS를 이용한 언어자원 구축 및 생물학적 개체명 인식 시스템

이현숙*, 김태현*, 장현철*, 박수준*, 박선희*
*한국전자통신연구원 바이오정보연구팀
e-mail : lhs63473@etri.re.kr

Biological Language Resource Construction and Named Entity Recognition System using UMLS

Hyun-Sook Lee*, Tae-Hyun Kim*, Hyun-Chul Jang*, Soo-Jun Park*, Seon-Hee Park*
*Bioinformatics Research Team,
Electronics and Telecommunications Research Institute

요 약

본 논문에서는 생물학적 문헌으로부터 유의미한 정보를 추출하는 바이오 텍스트 마이닝의 기본 단계인 생물학적 개체명 인식 모델을 제안하였다. 기존의 생물학적 개체명 인식은 규칙 혹은 코퍼스 구축뿐만 아니라 개체명 인식에 요구되는 기본 자원을 구축하는데만도 많은 시간과 비용이 요구되므로 한정된 도메인을 대상으로 연구가 진행되어 왔다. 본 논문에서 제안하는 개체명 인식 방법은 이러한 비용 문제 및 새로운 도메인으로의 이식성 문제를 극복하기 위해 UMLS 로부터 통계적인 방법으로 정보를 추출해 기본적인 언어자원을 구축하고 이를 이용해 규칙을 생성함으로써 개체명 인식을 수행한다. 본 연구에서 제안하는 방법은 바이오 텍스트 마이닝 연구의 도메인 한정적인 문제를 해결하는데 기여할 수 있을 것으로 기대된다.

1. 서론

최근 바이오 산업의 고속 발달로 인해 다양한 형태의 생물학적 데이터가 대량으로 생성되고 있으며 각종 연구보고서 및 논문 등과 같은 생의학 관련 문헌의 양 또한 빠른 속도로 증가하고 있다. 따라서, 방대한 생물학적 문헌으로부터 유의미한 정보를 자동으로 추출하는 바이오 텍스트 마이닝 기술에 관심이 집중되고 있다.

생물학적 개체명 인식은 바이오 텍스트 마이닝에서 가장 기본이 되는 단계로 생물학적 문헌들에서 단백질명, 유전자명 등과 같이 정보의 주체가 되는 생물학적 개체들을 지칭하는 이름들을 추출하는 것이다. 생물학적 개체명들은 명명법상 대문자, 숫자, 혹은 알파벳이 아닌 문자들을 포함할 수 있으며, 사용상에 있어서 동일 개체에 대해 매우 다양한 형태의 이름이 혼용되고 있다. 또한 구성상으로 볼 때, 전치사나 접속사를 포함하거나 개체의 기능 또는 범주를 나타내는 exchange factor, receptor, protein 등과 같은 어휘를 포함하기도 한다. 따라서, 생물학적 개체명 인식을 위해서는 위와 같은 다양한 특징을 고려한 개체명 인식이 수행되어야 한다.

생물학적 개체명 인식 방법은 크게 사전 및 개체명 인식 규칙을 패턴 매칭 형태로 적용하는 규칙 기반 방식과 기구축된 학습 코퍼스로부터 다양한 통계적 학습 알고리즘을 적용해 개체명을 인식하는 통계 기반 방식으로 나누어 볼 수 있다. 규칙 기반 개체명 인식의 경우는 전문가가 규칙 및 각종 언어 자원을 생성하는데 많은 비용이 들고, 새롭게 출현하는 개체명들이나 변형된 형태의 다양한 개체명들을 인식하지 못한다는 단점을 갖는다. 반면에 통계 기반 개체명 인식의 경우는 대용량 학습 코퍼스를 구축하는데 많은 시간과 비용이 소모된다는 단점이 있다. 두 가지 접근 방법의 공통적인 문제는 새로운 도메인으로의 이식성 관점에서 제약을 받는다는 것이다. 생물학적 개체명 인식에 관한 지금까지의 연구는 특정 도메인만을 대상으로 개체명 인식을 수행한 것으로 대상 도메인이 바뀔 경우 좋은 인식 결과를 기대할 수 없다는 문제가 있었다.

본 연구에서는 위에서 간략하게 살펴 본 생물학적 개체명 인식에서의 한계점을 극복하기 위해 UMLS 를 이용하여 전문가의 도움없이 자동으로 규칙을 추출하고 이를 적용하여 개체명을 인식하는 도메인 독립적인 방법을 제안한다.

UMLS(the Unified Medical Language System)는 다양한 생물의학 정보 자원들로부터 얻은 정보를 검색하고 통합하는 일을 용이하게 하기 위한 목적으로 시작된 프로젝트이며 100 개 이상의 정보 자원으로부터 획득한 약 200 만개 이상의 생의학 어휘들을 포함하는 방대한 metathesaurus 를 제공한다. 이러한 어휘들은 현재 135 개의 의미 범주(semantic type)로 분류되어 있다. UMLS 의 metathesaurus 는 특정 도메인을 위한 데이터

베이스가 아니므로 기존의 연구들의 가장 큰 단점인 도메인 이식성 문제를 해결하는데 중요한 역할을 할 수 있을 것이다.

본 논문의 구성은 다음과 같다. 2 장에서는 생물학적 개체명 인식의 관련연구에 대해서 살펴보고, 3 장에서는 UMLS 를 이용하여 개체명을 인식하는 모델을 설명한다. 4 장은 결론과 향후 연구를 보인다.

2. 관련연구

[Campbell99]는 "Pathological Process", "Finding"의 두 가지 의미범주에 대해 수동 생성된 규칙을 이용하여 개체명 인식을 수행한다. 일반적인 문법 정보와 UMLS 의 metathesaurus 를 이용하여 코퍼스로부터 추출한 [SemanticType + Connector + SemanticType][1] 형태의 의미-문법 패턴을 검토하여 최종 규칙을 정의한다. [Irena03]은 코퍼스와 UMLS 를 이용해 생성한 동사-보어 패턴을 활용해 개체명을 인식한다. 동사-보어 패턴은 도메인에 적합한 동사(e.g. activate, stimulate, etc.)와 함께 출현하는 개체명 후보들을 수집해 구성한다.

[Fukuda98]은 개체명을 이루는 문자열의 외형적 특성과 기정의된 생물학적 개체명에 사용되는 명명법만을 이용하여 생물학 분야에 의존적인 사전이 불필요한 방식의 개체명 인식을 수행하였다. 이 방식은 외형적 특성에 크게 의존하기 때문에 개체명들의 세부 분류를 결정짓기 어렵다는 단점을 갖는다. [Proux98]에서는 문장 단위로 어휘분석과 문맥분석을 통해 초파리와 관련이 있는 유전자명들을 인식하였다. 어휘분석 단계에서는 도메인 특화된 정보를 이용함으로써 오류를 교정하고, 문맥분석 단계에서는 어휘-문맥 패턴을 적용함으로써 새로운 개체명들을 추정해낸다.

[Kazama02]는 GENIA 코퍼스 (corpus) 및 기계학습 방식을 이용해 생물학적 개체명 인식을 수행하였다. 이 기법에서는 용어들의 품사정보를 이용해 비개체명(non-entity) 클래스를 여러 개의 부클래스(subclass)들로 분할하였다. 그리고, 이러한 클래스들과 GENIA 코퍼스에 제시되어 있는 개체명 클래스들을 대상으로 SVM(Support Vector Machine)을 적용함으로써 개체명 인식을 수행하였다. [Gaizauskas00]은 MUC[2]에서의 개체명 인식 방식을 응용하여 EMPathIE[3] 및 PASTA[4] 프로젝트에서 생물학적 개체명 인식을 수행하였다. 형태소 분석 후 다양한 생물학 정보 자원들에서 제공되는 어휘들과 각 단어들을 매칭시키고, 문법 규칙을 적용해 세부 구성요소들로 분할한다. 마지막으로 약어 인식과 수동 구축된 부가규칙을 적용한다. 이 방식은 각 개체명 클래스마다 규칙을 따로 정의해야 한다는 한계가 있다.

[1] e.g. [Diagnosis + of + Pathological Process/Finding]

[2] Message Understanding Conference

[3] The Enzyme and Metabolic Pathways Information Extraction (http://www.dcs.shef.ac.uk/research/groups/nlp/ funded/empathie.html)

[4] The Protein Active Site Template Acquisition Project (http://www.dcs.shef.ac.uk/research/groups/nlp/pasta)

3. 접근방법

본 연구에서 제안하는 개체명 인식 방법은 UMLS 의 metathesaurus 로부터 통계적인 방법을 사용해 생물학적 정보를 자동으로 획득하여 기본 언어자원 구축하고 이를 활용해 규칙을 생성한다. 이 방법은 전문가에 의해 정제된 언어자원 구축 및 규칙 생성이 요구되지 않고 도메인 변화에 유연하게 적응할 수 있다는 장점을 갖는다. 이러한 방식은 기존의 UMLS 를 이용한 생물학적 개체명 인식 연구에서 UMLS 의 자원을 단순히 사전으로 이용한 방법과는 구분된다.

본 시스템은 크게 기본 자원 구축, 규칙 구성, 개체명 인식의 세 단계로 나누어진다.

3.1 기본 자원 구축

Metathesaurus 는 UMLS 의 주요 구성요소로서 생의학 분야에서 사용되는 다양한 통제어휘들과 분류 등에서 한 번 이상 나타난 개념들에 대한 정보를 포함하고 있는 데이터베이스이다. Metathesaurus 는 동일 개념에 대한 대체이름과 다양한 관점(view)들을 모두 연결하고, 서로 다른 개념들 사이의 관계 정보를 제공한다. Metathesaurus 는 개념어(concept names)와 이를 의미적으로 분류하기 위한 Semantic type 정보를 제공하기 때문에 생물학적 개체들을 지칭하는 개체명을 추출하고 의미범주를 부여하는데 유용하게 활용될 수 있다.

본 연구에서는 metathesaurus 를 이용하여 기본 자원을 구축하기 위해 개념어와 Semantic type 을 각각 개체명과 의미범주로 매핑한다. Semantic type 을 이용해서 개념어를 의미범주로 나누고 각 의미범주를 특징지울 수 있는 정보를 수집하기 위해 다음과 같은 과정을 거친다. 개념어(concept name)를 토큰 단위로 분리하고 각 토큰의 가중치를 계산하여 단일어(Single Term)와 범주키텀(Keyterm)을 추출한다. 단일어는 단독으로 개체명이 되는 단어를 의미하고, 범주키텀은 특정 범주에서 주로 출현하여 개체명을 구성하는데 있어 중요한 역할을 하는 단어들을 말한다.

개체명 인식에서 인식의 대상이 되는 각 범주별로 단일어와 범주키텀을 구축하는 것은 전문가가 아니면 어려운 일이고 시간과 비용이 많이 드는 작업이다. 그러나 UMLS 를 이용하여 자동으로 자원을 구축하게 되면 비용면에서의 절감 효과뿐만 아니라 대상 도메인이 바뀌어도 손쉽게 적응할 수 있다는 장점이 있다.

3.2 규칙 구성

본 연구에서는 개체명 인식을 위한 규칙 생성을 위해 다음의 [표 1]과 같이 7 가지의 자질을 사용하여 개체명을 이루는 각 토큰의 특징을 추출한다. 각 토큰의 의미적인 특징을 얻기 위해 Single 및 Keyterm 을 자질로 사용하였고, 외형적인 특징을 얻기 위해 Caps 및 Alpnum 자질을 사용하였다. 또한 여러 토큰으로 이루어지는 개체명의 구성상의 특징을 얻기 위해 Precnj 및 Spchar 자질을 사용하였다.

자질명	의 미
Single	단독으로 쓰여 개체명을 나타내는 단어 (단일어)
Keyterm	특정 카테고리에서 주로 사용되는 단어 (범주키텀)
Caps	토큰을 이루는 문자열의 대소문자 구성 특징
Alpnum	숫자를 포함하는 토큰
Precnj	전치사 및 접속사
Spchar	특수 문자
Others	그 밖의 경우

[표 1] 자질의 종류

각 자질은 그 특성에 따른 서브타입을 갖는다. *Single* 및 *Keyterm* 자질의 경우에는 *Semantic type ID* 를 서브타입 값으로 갖게 되고, 그 밖의 자질의 경우에는 다음의 [표 2]와 같은 서브타입들을 갖는다.

자질명	서브타입	예
Single	선택한 의미범주의 ID 들	Bacteria, Virus, ...
Keyterm	선택한 의미범주의 ID 들	Bacteria, Virus, ...
Caps	Start	Abcde
	End	abcdE
	All	ABCDE
	Mixed	abCDe
	Romnum	IX
Alpnum	NumOnly	120
	Year	1900~2999
	NumUnit	10bp
	NumGreek	alpha10
	NumOther	10abc
Precnj	108 개	about, and, ...
Spchar	36 개	TAB, !, &, ...

[표 2] 자질 별 서브타입

자질 추출은 개체명을 이루는 토큰을 기본 단위로 이루어진다. 자질추출의 결과는 다음의 [그림 1]과 같은 방식으로 표현된다. 즉, 대상이 되는 토큰에서 추출된 모든 자질들을 결합한 FEATNUM 을 우선적으로 기술하고, 그 뒤에 각 자질들의 서브타입을 나타내는 TYPENUM 을 나열한다.

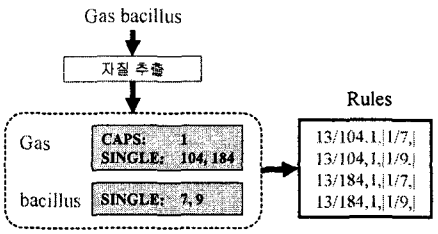
TOKEN := FEATNUM / TYPENUM, (TYPENUM)*
FEATNUM := A number that represents the combination of features.
TYPENUM := A number that represents each of the feature's subtype.

[그림 1] 자질추출 결과의 표현 방식

위와 같이 개체명을 단어 단위로 토큰화하고, 각 토큰을 대상으로 다양한 자질들을 추출한 후 그 결과를 다음의 표현과 같은 방식으로 결합하여 규칙을 구성한다.

RULE := TOKEN | (TOKEN)*

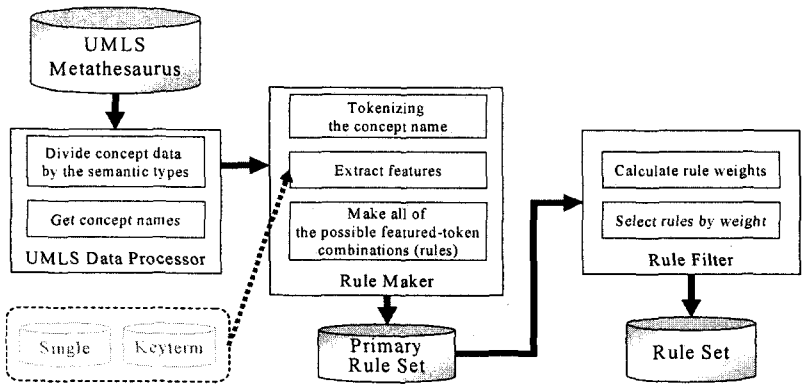
하나의 규칙은 여러 개의 토큰을 포함할 수 있고, 각 토큰에서 추출된 자질이 여러 개의 서브타입을 가질 수 있다. 따라서, 위에서 정의한 바와 같이 각 자질 별로 하나의 서브타입을 갖는 형식으로 규칙을 표현하기 위해서는 이들의 모든 조합을 고려한 개수만큼의 서로 다른 규칙을 생성해내야 한다. [그림 2]에서 "*Gas bacillus*"라는 개체명은 2 개의 각 토큰이 2 개의 서브타입을 갖기 때문에 최종적으로 4 개의 규칙이 만들어지게 된다.



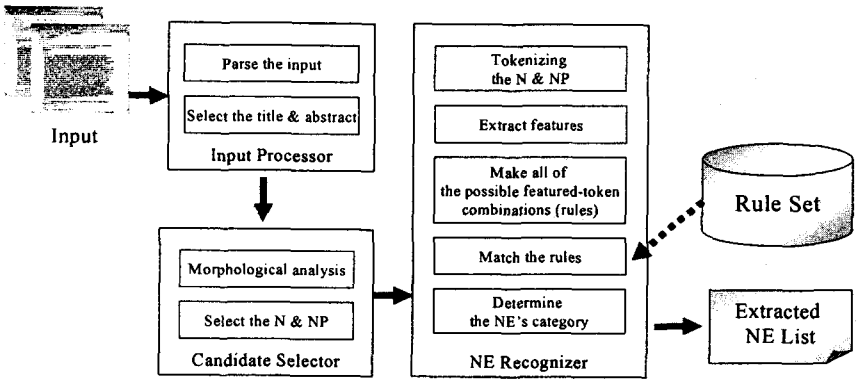
[그림 2] 규칙 생성 예

규칙이 생성된 후, 각 규칙이 semantic type 별로 가지는 가중치를 계산하여 필터링을 수행한다.

위에서 설명한 규칙 구성 과정은 [그림 3]과 같다.



[그림 3] 규칙 구성 과정



[그림 4] 개체명 인식 과정

3.3 개체명 인식

본 연구에서 제안한 개체명 인식 방법은 [그림 4]와 같이 크게 입력처리, 개체명 후보 선택, 개체명 인식의 세 단계로 구성된다.

입력처리 단계에서는 본 연구의 개체명 인식 대상이 되는 XML 문서 집합을 처리한다. 이 문서 집합은 MEDLINE 으로부터 수집한 의학 문헌 초록들로 이루어진다. XML 파서를 이용하여 문서를 파싱하고 그 결과로부터 제목과 요약문을 추출해 다음 단계의 입력으로 넘겨준다.

개체명 후보 선택 단계는 입력을 형태소 분석한 후 이로부터 명사 및 명사구를 추출한다. 추출한 결과에 다양한 휴리스틱을 적용하여 생물학적 개체명이 될만한 후보들을 선택한다.

개체명 인식 단계에서는 각 개체명 후보를 토큰화하고 3.2 절에서 설명한 자질 추출 및 규칙 구성 방법을 적용하여 규칙을 생성한다. 이와 같이 생성된 개체명 후보의 규칙을 규칙집합에 있는 규칙들에 매치하여 유사규칙들을 선별한다. 선별된 유사규칙들의 가중치를 계산하여 가장 높은 가중치를 가지는 규칙의 의미범주를 개체명의 최종 범주로 결정한다. 가중치는 본 연구에서 대상으로 하는 의미범주의 특성 및 규칙의 부분매칭을 고려한 tf·idf 수식을 활용해 계산한다.

4. 결론 및 향후 연구

본 연구에서는 생물학적 개체명 인식의 한계점인 새로운 도메인으로의 이식성을 극복하기 위해 UMLS 를 이용한 개체명 인식 모델을 제안하였다. 이 모델은 UMLS 를 이용하여 전문가의 도움없이 통계적인 방법으로 기본적인 언어자원 구축 및 개체명 인식을 위한 규칙 생성이 가능하다. 따라서 비용의 절감 효과뿐만 아니라 새로운 도메인에도 용이하게 적용할 수 있다는 것이 가장 큰 장점이다. 현 과제에서는 MEDLINE 으로부터 수집한 "apoptosis"에 관한 문서를 대상으로 개체명 인식에 관한 연구를 수행하고 있다.

바이오 텍스트 마이닝에 관한 연구는 아직까지 한정된 도메인만을 대상으로 이루어지고 있는 실정이다. 바이오 텍스트 마이닝의 기본 단계인 개체명 인식이 도메인의 확장 및 변화에 유연하게 대처할 수 있을 때, 바이오 텍스트 마이닝 기술도 빠르게 성장할 수 있을 것이다.

참고문헌

[Campbell99] David A. Campbell and Stephen B. Johnson, *"A Technique for Semantic Classification of Unknown Words Using UMLS Resources"*, Proceedings of American Medical Informatics Association Symposium, 1999, pp. 716-720.

[Irena03] Irena Spasic, Coran Nenadic and Sophia Ananiadou, *"Using Domain-Specific Verbs for Term Classification"*, Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, Sapporo, Japan, July, pp. 17-24.

[Fukuda98] Fukuda, K., Tamura, A., Tsunoda, T. and Takagi, T., *"Toward IE: Identifying protein names from biological papers"*, Proceedings of the Pacific Symposium on Biocomputing (PSB98).

[Proux98] Denys Proux, Francois Rechenmann and Laurent Julliard, *"Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction"*, Genome Informatics, 9:72-80, 1998.

[Kazama02] Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta and Jun'ichi Tsujii, *"Tuning Support Vector Machines for Biomedical Named Entity Recognition"*, 2002

[Gaizauskas00] Robert Gaizauskas, George Demetriou and Kevin Humphreys, *"Term Recognition and Classification in Biological Science Journal Articles"*, Proceedings of the Computational Terminology for Medical and Biological Applications Workshop of the 2nd International Conference on Natural Language Processing (NLP-2000), Patras, Greece, June 4, pp. 37-44.