기계학습 알고리즘에 근거한 단백질 이름의 자동 추출

김정호° 백은옥 이공주
서울시립대학교 기계정보공학과, 이화여자대학교 분자생명과학부/약학대학
redscene@sidae.uos.ac.kr paek@uos.ac.kr kjl@ewha.ac.kr

A Machine Learning Approach for Automatic Protein Name Extraction from Journal Articles

Jung-Ho Kim°, Eunok Paek, and Kong-Joo Lee
Department of Mechanical and Information Engineering, University of Seoul
Center for Cell Signaling Research, Division of Molecular Life Sciences, Ewha Womans University

요약

생물학 분야의 문헌으로부터 유전자 및 단백질 이름을 추출하는 기술은 바이오 텍스트 마이닝 분야의 기반 기술로 그 중요성이 점차 증대되고 있다. 이 연구에서는 생물학 분야 문헌의 초록으로부터 하나의 토큰으로 구성된 single gene name은 물론 여러 개의 토큰으로 이루어진 multi gene name까지 유전자나 단백질의 이름을 자동으로 추출하는 시스템 TagGeN(Tagger for Gene Name)을 제안한다. TagGeN은 기존의 태거와 달리, 문자나 숫자 이외의 기호를 포함한 유전자나 단백질 이름의 품사 처리에 있어 개선 방안을 제시하고, 여러 토큰으로 이루어진 이름의 인식에 있어 나란한 두 토큰이 갖는 태그 정보를 이용한 조건부 확률에 근거하여 Markov 모델을 도입한다. 위와 같은 개선방안을 구현한 TagGeN은 성능면에서 기존의 유사시스템에 비해 recall 20.8%, precision 4.7%의 성능향상을 보임으로써 본 연구에서 제안한 방법론의 효과를 입증한다

1. 서론

1.1 개요

게놈 프로젝트가 성공적으로 마무리된 이후 생물학 분야의 연구가 활발히 진행되고 있고, 그 결과 생물학 관련 문헌의 양이 끊임없이 증가하고 있다. 따라서 생물학 분야의 문헌에 대한 정보 추출 (Information Extraction) 기술 및 텍스트 마이닝 (Text Mining) 기술의 중요성이 점점 증가하고 있다. 그 중에서도, 문헌 상에서 단백질 혹은 유전자 개체의 이름을 인식하는 것은 텍스트 마이닝에 있어서의 기반 기술이라고 할 수 있는데, 이러한 기반 기술의 개발이 선행되어야만 단백질 이름의 동의어에 관한 연구 등의 보다 의미 있는 정보를 얻기 위한 연구가 진행될 수 있기 때문이다 [1].

본 연구에서는 문헌 상의 단백질 및 유전자 개체 이름을 자동으로 추출해 주는 프로그램 TagGeN (Tagger for Gene Name)을 개발하였다. 이 프로그램은 문헌 상에서 하나의 토큰으로 이루어진 Single Gene Name은 물론, 두 개 이상의 토큰으로 이루어진 Multi Gene Name까지 찾을 수 있도록 고안되었다.

1.2 유전자/단백질 이름

문헌에서 발견되는 단백질과 유전자의 이름은 몇 개의 토큰으로 이루어졌는지에 따라, 하나의 토큰으로 이루어진 이름인 single gene name과, 두 개 이상의 토큰으로 이루어진 이름인 multi gene name으로 나누어 생각할 수 있다.

Single gene name의 경우, 기본적으로 Brill의 품사 태거를 이용하여 문헌 상에서 하나의 토큰으로 이루어진 단백질 이름을 찾는다 [2]. 여기에, Brill 태거를 사

용함에 있어 발생하는, 기호를 포함한 단어 처리에 있어서의 문제점을 해결하기 위하여 post-processing 단계를 도입하였다. Multi gene name의 태깅에 있어서는 여러 토큰으로 이루어진 단백질 이름에 포함될 가능성이 가장 높다고 여겨지는 토큰을 인식하고, 이를 시작점으로 삼아 multi gene name의 범위를 결정하는 과정을 거치게 되는데, 이 때 확률적인 방법론을 사용한다.

2. 연구 방법

2.1 개요

단백질 이름은 하나 혹은 두 개 이상의 토큰으로 이루어져 있다. "Acetyltransferase", "ATPase", 혹은 "TNF" 등은 하나의 토큰으로 이루어져 있는 단백질 이름의 예이고, "NF-kappa B"나 "Tumor necrosis factor"는 두 개 이상의 토큰으로 이루어져 있는 단백질 이름의 예이다.

이렇게 찾고자 하는 이름이 하나의 토큰으로 구성되어 있는지 혹은 두 개 이상의 토큰으로 구성되어 있는지에 따라, 서로 다른 접근 방법이 필요하다. 먼저 하나의 토큰으로 이루어져 있는 단백질 이름인 single gene name을 찾은 후에 이 결과를 이용하여 두 개 이상의 토큰으로 이루어져 있는 단백질 이름인 multi gene name을 찾는 방법을 취한다.

2.2 Brill의 품사 태거

Single gene name tagging 과정에서 사용한 기본 프로그램은 Brill의 품사 태거이다. Brill의 품사 태거는 사용하고자 하는 도메인에 대해서 태거를 학습시킨 이후에 사용한다. 학습 단계에서는 Lexicon(단어사전)과 품사가 태깅 되어 있는 해당 분야의 말뭉치(corpus)를 입력으로 사용하여 어휘규칙과 문맥규칙을 생성한다.

Brill의 품사 태거를 생물학 분야에 맞게 학습시키기 위해서는 우선 Lexicon을 수정하여야 한다. 이를 위해 품사 태그에 새로운 품사인 **GENE** 태그를 추가하고, **GENE** 태그를 품사로 가질 수 있는 어휘의 목록을 Lexicon에 추가하였다.

Brill 태거의 학습을 위해서는 Lexicon에 어휘들을 추가하는 것과 더불어, 태깅이 완성된 말뭉치가 필요하다. 생물학 분야에서 사용할 수 있는 말뭉치로는 GENIA corpus가 대표적이다 [3].

2.3 Single Gene name 태깅

Brill의 품사 태거를 생물학 도메인에 적응시킨 후에도 이를 그대로 생물학 문서에서 단백질 이름을 찾아내는 데에 사용하는 데에는 몇 가지 문제점이 존재한다. 그 중에서도 가장 주목해야 할 문제점은 Brill의 품사 태거가 "-" 혹은 "/"와 같이 기호를 포함한 단어를 처리하는 방식이 gene name 태깅에 적합하지 않다는 것이다. 단백질이나 유전자 이름의 경우에는 -, /, (,), ', + 와 같은 기호가 매우 빈번히 이름의 일부로서 사용되고 있으므로 이에 대한 알맞은 처리가 반드시 필요하다.

하나의 토큰 내에 기호를 포함하는 single gene name을 인식하기 위하여, Brill 태거에 의한 품사 태깅 이후에 post-processing 단계를 두어서 Brill 태거가 갖는 한계를 극복하고자 하였다. Post-processing 과정에서는 기호 "-"나 "/"을 포함하는 토큰의 경우, 해당 기호를 구분자(delimiter)로 사용하여 다시 한번 더 작은 토큰으로 나눈 다음, 나뉘어진 각 토큰에 대하여 다시 Brill의 태거로 품사를 태깅하였다.

또한 Brill의 품사 태거가 가지는 또 다른 제약 요건인, Lexicon에 들어있는 어휘와 약간의 변형된 단어가 문장에 나타나는 경우에 이것을 처리하지 못하는 점을 보완하기 위해서, 어휘들의 변형을 고려하여 정규 표현식을 만든 후, 문장에서 패턴이 일치하는 unknown word를 찾으면, Lexicon의 어휘와 unknown word 사이의 부분 매치를 통해 품사를 수정할 수 있도록 하였다.

2.4 Multi Gene Name 태깅

Multi gene name은 2개 이상의 토큰으로 이루어진 단백질 이름을 말한다. Single gene name 태깅 단계에서 **GENE** 품사를 가지는 것으로 결정된 단어가 주변의 다른 단어와 함께 하나의 multi gene name을 구성하는 경우가 흔히 있다. 그러나 "tumor/**NN** necrosis/**NN** factor/**NN**"의 경우처럼 **GENE** 태그를 갖지 않는 단어들만으로 multi gene name이 구성되는 경우도 존재한다. 따라서 multi gene name을 찾는 단계에서는 이 두 가지 경우를 모두 고려한다.

여기서는 "protein", "gene", "factor", "receptor"

등과 같이 단백질 이름에 자주 등장하는 단어들을 "seed form"으로 정의하고 앞서 **GENE** 품사의 단어들과 함께 탐색의 시작점으로 하여 multi gene name의 범위를 결정한다. Multi gene name을 찾을 때에는 Multi gene name을 포함할 것이 거의 확실시되는 부분을 "gene name chunk"로 추출한다. 그러나, 모든 gene name chunk가 multi gene name을 포함하는 것은 아니기 때문에, 확률 모델을 고안하여 주어진 gene name chunk 안에서 multi gene name의 범위를 결정 하는 데에 사용하였다.

여기서 제안하는 확률 모델에서는, gene name chunk 안에서 seed token의 위치를 찾은 후에 seed token에서부터 이웃한 단어를 차례로 조합하여 multi gene name의 범위를 결정할 때, 이웃한 단어들의 태그 값에 따른 조건부 확률을 사용하였다.

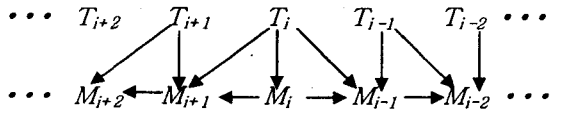

그림 1. Multi gene tagging의 확률 모델
(T = 태그 값, M = Multi gene의 여부를 나타내는 값)

그림 1은 Multi gene name tagging에서 사용한 확률 모델을 그림으로 표현한 것이다. 그림에서와 같이 seed token에서부터 시작하여 좌우로 multi gene의 범위를 확장하였으며, 이때 multi gene에 속하는지의 여부는 해당하는 단어의 바로 앞 단어의 태그값과 multi gene인가의 여부, 그리고 해당 단어의 태그값을 고려하여 조건부 확률을 적용하였으며 그 확률값을 기준으로 multi gene 여부를 결정하였다.

3. 실험 결과

TagGeN의 성능을 평가하기 위해서 7-fold cross validation을 실시하였다. Brill의 품사 태거를 생물학 분야에 맞게 적응시키고, multi gene name 태깅에 사용하는 조건부 확률을 구하기 위해 무작위로 선택한 GENIA corpus의 논문 초록 700개를 7개로 나누어 cross validation을 실시했고, 또한 위의 700개와는 별도로 생물학자가 독립적으로 gene name을 태깅한 새로운 60개의 논문 초록을 테스트 데이터로 사용하였다.

표 1. TagGeN의 실험 결과

TagGeN	Exact match		Inclusive match		Overlap Match	
Performance measure	R	P	R	P	R	P
7-fold Average	90.0%	80.1%	91.9%	82.5%	93.1%	83.6%
Independent test set	87.3%	76.1%	92.1%	80.1%	93.9%	82.0%

4. 결론

최근의 생물학 연구의 증가로 인해 생물학 관련 문헌의 양은 끊임없이 증가하고 있으며, 문헌들을 기반으로 새로운 연구가 진행된다. 따라서 이 수많은 문헌들에 대한 보다 구조적인 접근 방법의 중요성은 날로 증가하고 있다.

이번 연구가 기존의 연구와 다른 점은 두 가지를 들 수 있다. 하나는 단백질 이름에 나오는 문자 중에서 알파벳이나 숫자가 아닌, 기호의 처리에 관한 부분이고, 또 하나는 문장에서 multi gene name의 범위를 결정하는 데 확률 모델을 사용했다는 것이다. 이 두 가지의 개선점은 기존의 연구보다 성능 면에서 많은 향상을 가져오게 되었다.

이번 연구는 단백질이나 유전자 이름 사이의 동의어 연구 혹은 단백질 상호작용 정보를 텍스트 마이닝을 통해 얻고자 하는 연구와 같은 경우에 기반 기술로서의 역할을 담당할 수 있으리라고 기대한다.

참고 문헌

[1] Hong Yu and Eugene Agichtein "Extracting Synonymous Gene and Protein Terms from Biological Literature". *Bioinformatics*, 19, pp. i340-i349. (2003)

[2] Eric Brill. "Some Advances in Transformation-Based Part of Speech Tagging". In *Proceedings of the National Conference on Artificial Intelligence. AAAI press.* pp. 722-727. (1994)

[3] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii "GENIA corpus—a semantically annotated corpus for bio-textmining". *Bioinformatics*, 19, pp. i180-i182. (2003)