DOI: 10.3745/KIPSTB.2008.15-B.4.375

구문 의존 경로에 기반한 단백질의 세포 내 위치 인식

김 미 영†

요 약

단백질의 세포 내 위치를 인식하는 것은 생물학 현상의 기술에 있어서 필수적이다. 생물학 문서의 양이 늘어남에 따라, 단백질의 세포 내 위치 정보를 문서 내용으로부터 얻기 위한 연구들이 많이 이루어졌다. 기존의 논문들은 문장의 구문 정보를 이용하여 정보를 얻고자 하였으며, 언어학적 정보가 단백질의 세포 내 위치를 인식하는 데 유용하다고 주장하고 있다. 그러나, 이전의 시스템들은 구문 정보를 얻기 위해 부분 구문분석기만을 사용하였고 재현율이 좋지 못했다. 그러므로 단백질의 세포 내 위치 정보를 얻기 위해 전체 구문분석기를 사용할 필요가 있다. 또한, 더 많은 언어학적 정보를 위해 의미 정보 또한 사용이 가능하다.

단백질의 세포 내 위치 정보를 인식하는 성능을 향상시키기 위하여, 본 논문은 전체 구문분석기와 어휘망(WordNet)을 기반으로 한 방법을 제안한다. 첫 번째 단계에서, 각 단백질 단어로부터 그 단백질의 위치후보에까지 이르는 구문 의존 경로를 구축한다. 두 번째 단계에서, 구문 의존 경로의 루트 정보를 추출한다. 마지막으로, 단백질 부분트리와 위치 부분트리의 구문-의미 패턴을 추출한다. 구문 의존 경로의 루트와 부분트리로부터 구문태그와 구문방향을 구문 정보로서 추출하고, 각 노드 단어의 의미태그를 의미 정보로서 추출한다. 의미태그로는 어휘망의 동의어 집합(synset)을 사용한다. 학습데이터에서 추출한 루트 정보와 부분트리의 구문-의미 패턴에 따라서, 실험데이터에서 (단백질, 위치) 쌍들을 추출했다. 어떤 생물학적 지식 없이, 본 논문의 방법은 메드라인(Medline) 요약 데이터를 사용한 실험 결과에서 학습데이터에 대해 74.53%의 조화평균(F-measure), 실험데이터에 대해서는 58.90%의 조화평균을 보였다. 이 실험은 기존의 방법들보다 12-25%의 성능향상을 보였다.

키워드 : 단백질 세포 내 위치, 바이오인포매틱스, 구문 관계, 정보 추출, 텍스트 마이닝

Detection of Protein Subcellular Localization based on Syntactic Dependency Paths

Mi-Young Kim†

ABSTRACT

A protein's subcellular localization is considered an essential part of the description of its associated biomolecular phenomena. As the volume of biomolecular reports has increased, there has been a great deal of research on text mining to detect protein subcellular localization information in documents. It has been argued that linguistic information, especially syntactic information, is useful for identifying the subcellular localizations of proteins of interest. However, previous systems for detecting protein subcellular localization information used only shallow syntactic parsers, and showed poor performance. Thus, there remains a need to use a full syntactic parser and to apply deep linguistic knowledge to the analysis of text for protein subcellular localization information. In addition, we have attempted to use semantic information from the WordNet thesaurus.

To improve performance in detecting protein subcellular localization information, this paper proposes a three-step method based on a full syntactic dependency parser and WordNet thesaurus. In the first step, we constructed syntactic dependency paths from each protein to its location candidate, and then converted the syntactic dependency paths into dependency trees. In the second step, we retrieved root information of the syntactic dependency trees. In the final step, we extracted syn-semantic patterns of protein subtrees and location subtrees. From the root and subtree nodes, we extracted syntactic category and syntactic direction as syntactic information, and synset offset of the WordNet thesaurus as semantic information. According to the root information and syn-semantic patterns of subtrees from the training data, we extracted (protein, localization) pairs from the test sentences. Even with no biomolecular knowledge, our method showed reasonable performance in experimental results using Medline abstract data. Our proposed method gave an F-measure of 74.53% for training data and 58.90% for test data, significantly outperforming previous methods, by 12-25%.

Key Words : Protein subcellular localization, bioinformatics, syntactic relation, information extraction, text mining

1. 서 론

생물학 문서의 양이 증가함에 따라, 그 문서들로부터 자동적으로 정보를 추출하려는 연구가 최근 활발히 진행되고 있다. 초기에는 간단한 자질을 기반으로 한 기계학습 방법을 사용하는 연구가 많았으나, 최근에는 문서 내의 언어학적 정보를 이용하는 연구가 활발하다[1]. 기존 연구들은 이

※ 이 논문은 2008년도 성신여자대학교 학술연구조성비 지원에 의하여 연구되었음.
† 정 회 원 : 성신여자대학교 컴퓨터정보학부 전임강사
논문접수 : 2008년 3월 3일
수 정 일 : 1차 2008년 6월 12일, 2차 2008년 6월 19일
심사완료 : 2008년 6월 20일

러한 언어학적 정보가 생물학 데이터로부터 정보를 추출하는 데 유용하다고 주장하고 있다[2-5]. 생물학 데이터를 대상으로 가장 많이 추출되는 정보인 단백질 상호작용을 인식하기 위해, 기존 연구들은 전체 구문분석기로부터 도출된 구문 정보를 이용함으로써 향상된 성능을 보였다[6,7]. 그러나, 생물학 문서를 대상으로 또한 중요하게 추출되는 정보인 단백질의 세포 내 위치를 인식하기 위해, 기존 연구들은 부분 구문분석기만을 사용했다. 이러한 부분 구문분석기는 단지 구의 경계를 인식하고, 생물학 문서에 대해 완전한 구문 정보를 제공하지 않는다. 단백질의 세포 내 위치 정보의 인식을 위해 더 많은 언어학적 자질을 추출하는 것이 필요하다.

단백질의 세포 내 위치를 인식하는 성능을 향상시키기 위해, 본 논문에서는 전체 구문 분석기와 어휘망(WordNet)을 기반으로 한 방법을 제안한다. 전체 구문 분석기로는 D. Lin[16]의 구문분석기를 이용하고. 의미 정보를 얻기 위하여 어휘망의 동의어 집합(synset) 을 이용한다. 실험 결과, 본 논문에서 제안한 방법이 기존의 방법보다 성능이 좋다는 것을 보이고, 사용된 각 정보가 성능에 얼마나 기여했는지를 설명한다.

2. 기존 연구

생물학 분야에서 관계 정보를 추출하기 위한 연구가 최근에 활발히 이루어지고 있다. 현재 연구는 단백질 상호작용[6,8], 세포 내 위치 인식[9], 질병-치료 관계[10], 단백질 배열을 기반으로 한 시스템 등이 있다. 단백질의 세포 내 위치를 인식하는 데 있어서, 초기 연구들은 단백질 배열을 기반으로 한 기계 학습 방법에 주안을 두었다[11-14].

생물학 문서의 양이 증가함에 따라, 몇몇 연구들은 문서 내의 단어 자질을 사용함으로써 정보 추출을 시도했다. B. Stapley 등[9]은 GeneDB 사전을 사용하여 말뭉치로부터 용어들을 추출하였고, 지지벡터기계(Support Vector Machines; SVM)를 기반으로 하여 각 위치에 대한 분류기를 만들었다. H. Shatkay 등[15]은 비록 위치 정보와 직접적으로 연관되어 있지는 않으나 간접적으로 연관된 단어들을 추출했다. 이 용어들은 해당 단백질의 위치기관이 무엇인지 직접적으로 알려주지는 않으나, 그 단백질이 어느 기관에 위치되어 있는지를 논하는 문서에 함께 등장하는 경향이 있기 때문에 도움이 된다. 이 용어들을 기반으로 하여, [15]에서는 각 위치에 대해 지지벡터기계 분류기를 구축하였다. 그들은 단백질 데이터를 기반으로 하여 학습된 4개의 분류기와 생물학 문서로부터의 1개의 분류기를 결합하여서, 문서로부터의 정보 추출(Text mining)이 성능 향상에 기여한다는 것을 보여줬다.

M. Krogel과 T. Scheffer[12] 또한 단백질 위치를 인식하는 데 있어 문서로부터의 정보 추출을 수행하였다. 사전 기반의 접근을 이용하여, 그들은 생물학 문서로부터 용어들을 추출하여 자질로 사용하였다. 정보 추출은 대량의 학습데이터를 요구한다. 그러나 정답이 부착되어 있는 대량의 학습 데이터를 구축하기 위해서는 많은 수작업이 요구되므로, 현재는 정답이 달려있는 소량의 데이터만이 실험용으로 존재한다. 현재 제공되고 있는 소량의 실험용 데이터를 보완하기 위해, [12]에서는 정답이 부착되어 있지 않은 원시 데이터를 함께 사용한 변환적 지지벡터기계(transductive SVM)와 협력학습(co-training) 방법을 수행했다. 그러나, 이러한 기계학습 방법은 더 낮은 성능을 보였다.

몇몇 연구들은 성능을 향상시키기 위한 중요한 열쇠로서 언어학적 정보를 사용했다. M. Craven과 J. Kumlien[2]은 선댄스(Sundance) 부분 구문분석기를 사용하여 언어학적 처리를 수행했다. 단어들을 자질로 사용한 나이브 베이즈(Naïve Bayes) 분류기의 재현율을 향상시키기 위하여 선댄스 부분 구문분석기를 사용했으나, 이 분석기는 단지 구의 경계만을 인식한다. 그들은 구문분석 결과로부터 주어, 목적어 구문관계와 구의 위치정보를 자질로 사용했다. 이러한 언어학적 정보를 나이브 베이즈 분류기와 결합한 결과, 0.21의 재현율과 0.82의 정확률, 그리고 0.34의 조화평균(F-measure) 성능을 얻었다.

D. Page, M. Craven [4]과 M. Skounakis 등[5]은 문맥자질을 사용하여 계층있는 은닉 마코프 모델(HMM) 방법을 이용했다. 그들 또한 선댄스 부분 구문분석기를 사용하여 HMM을 구축한다. 위에서 언급한 바와 같이, 선댄스 부분 구문분석기는 깊은 언어학적 정보를 보여주지 않는 한계가 있다.

M. Goadrich 등[3] 또한 선댄스 부분 구문분석기를 이용하여 251개의 술어(predicate) 정보를 획득했다. 술어 정보들은 주로 구 위치 정보를 나타낸다. 그들은 정확률 0.58, 재현율 0.40, 그리고 조화평균 0.47로서 M. Craven과 J. Kumlien[2]의 실험데이터를 사용한 결과들 중 가장 좋은 성능을 보였다. 실험 전에, 그들은 실험데이터의 잘못된 주석 정보를 수작업으로 수정했다.

이전 연구들을 종합한 결과, 언어학적 정보가 단백질의 세포 내 위치를 획득하는 데 중요하다고 판단하였다. 그러나, 이전 연구들은 부분 구문분석기만을 사용하여 단지 구의 경계 정보만을 이용하였다. 그 결과, 재현율이 좋지 못했다. 그리하여 더 깊이 있는 언어학적 정보가 필요하다고 판단하고, 본 논문에서는 전체 구문분석기와 어휘망을 기반으로 하여, 단백질의 세포 내 위치 정보를 인식하는 방법을 제안한다. 어떤 생물학적 도메인 지식 없이, 본 논문의 방법은 생물학 문서에 이용되어 더 좋은 성능을 낼 수 있다. 다음 장에서, 본 논문에서 제안하는 방법을 상세히 설명한다.

3. 방법론

3.1 언어학적 단서

구문과 의미 정보가 언어학적 단서로서 사용되었다. 구문정보를 위하여, 본 논문에서는 D. Lin[16]의 의존트리 구문분석기인 MINIPAR를 이용한다. MINIPAR는 영어 문장에 대한 대표적인 의존트리 구문분석기로서, 문장에서 각 단어

들의 구문관계를 잘 나타내 준다. 이 구문분석기는 빠르고 설치하기 쉬우며, 결과가 쉽게 처리가능한 장점이 있다. 따라서 질의응답[17], 토픽인식 시스템[18], 의미표현으로부터 구어 생성[19], 단어의 의미 결정 시스템[20], 개체 분류 시스템[21], 유전자 상호작용 인식 시스템[22] 등 많은 응용시스템에서 기본 구문분석기로 사용되어 왔다.

[22]에서 또한 생물학 문서를 대상으로 정보를 얻기 위해 MINIPAR를 사용하고 있다. 하지만 [22]에서는 구문분석 트리의 의존소-지배소 관계 정보 전체를 이용하지 않고 두 유전자 단어 사이에 일차원적인 구문관계 체인만을 추출하여 사용하는 한계가 있다. 또한 의미 정보는 이용하지 않고 있다. 본 논문에서는 의미 정보를 더하여 사용하고, MINIPAR 구문분석 트리의 전체 의존소-지배소 관계 구조 계층도를 그대로 유지한 채 구문 정보를 얻고자 한다.

(그림 1)은 메드라인(Medline)의 문서들 중 한 문장에 대한 MINIPAR 의존트리 결과의 예를 보여준다. 의존트리 구문분석기는 각 단어와 구문적으로 지배소-의존소 관계가 있는 단어들을 분석하고 구문태그를 결정한다. 우리는 MINIPAR 구문트리의 구문태그와 구문 방향 정보를 추출하여 구문 정보로서 이용한다. 의미 정보로서는, 어휘망[23]의 동의어 집합을 사용한다. 동의어 집합은 어휘망[23] 계층에서 각 단어에 붙은 의미코드 숫자를 의미한다. 그 예로, "localize" 단어는 4가지 개념으로 사용될 수 있는데 각각 {02695895}, {02692335}, {02509919}, {01711749}의 의미코드를 가지는 것을 (그림 2)를 통해 알 수 있다. (그림 2)에서 각 의미코드 앞에 쓰여진 숫자는 그 의미로서의 사용빈도를 나타낸다. 여러 의미코드들 중에서, 우리는 가장 자주 사용되는 의미의 코드를 선택한다.

단백질의 세포 내 위치 정보의 인식을 위해 전처리 단계로서 먼저 문장 내의 단백질과 위치 단어를 태깅한다. 단백질 단어의 인식을 위해 정규문법을 사용하여 최소한 하나의 숫자를 포함한 단어를 단백질 단어로 간주한다. 위치 단어의 경우에는, M. Craven과 J. Kumlien[2]의 위치 단어가 태깅되어 있는 데이터를 사용하여 위치(LOCATION)로 태깅되어 있는 단어들을 모아 위치 사전을 구축하여 이용한다. 그리하여, 태깅된 모든 단백질 단어와 위치 단어를 대상으로 (단백질, 위치) 쌍 후보를 추출한다. 이 후보 쌍들은 학습데이터를 대상으로 추출한 것이기 때문에, 정답과 오답이 이미 구분되어 있으므로 이 두 가지를 구분지을 수 있는 특징을 알아낸다. 이 특징을 이용하여 실험데이터로부터 정확

<원시 문장>

We have cloned and sequenced the SPT7 gene and have shown that it encodes a large, acidic protein that is localized to the nucleus.

<구문분석 결과>

NUM	word	stem_word	POS of gov.	NUM type	relation	word of governor	
E2	(()	fin	C	*	)		
1	(We	~	N	3	s	(gov clone))	
2	(have	~	have	3	have	(gov clone))	
3	(cloned	clone	V	E2	i	(gov fin))	
E4	(()	we	N	3	subj	(gov clone)	(antecedent 1))
4	(sequenced	sequence	V	3	lex-dep	(gov clone))	
5	(the	~	Det	7	det	(gov gene))	
6	(SPT7	~	N	7	nn	(gov gene))	
7	(gene	~	N	3	obj	(gov clone))	
8	(and	~	U	3	punc	(gov clone))	
9	(have	~	have	10	have	(gov show))	
10	(shown	show	V	3	conj	(gov clone))	
E5	(()	we	N	10	subj	(gov show)	(antecedent 1))
E1	(()	fin	C	10	fc	(gov show))	
11	(that	~	COMP	E1	c	(gov fin))	
12	(it	~	N	13	s	(gov encode))	
13	(encodes	encode	V	E1	i	(gov fin))	
E6	(()	it	N	13	subj	(gov encode)	(antecedent 13))
14	(a	~	Det	17	det	(gov protein))	
15	(large	~	A	17	mod	(gov protein))	
16	(acidic	~	A	17	mod	(gov protein))	
17	(protein	~	N	13	obj	(gov encode))	
E0	(()	fin	C	17	rel	(gov protein))	
18	(that	~	THAT	E0	whn	(gov fin)	(antecedent 19))
19	(is	be	be	20	be	(gov localize))	
20	(localized	localize	V	E0	i	(gov fin))	
E7	(()	that	THAT	20	obj	(gov localize)	(antecedent 19))
21	(to	~	Prep	20	mod	(gov localize))	
22	(the	~	Det	23	det	(gov nucleus))	
23	(nucleus	~	N	21	pcomp-n	(gov to))	

(그림 1) 의존트리 구문분석 결과의 예

localize:
Sense 1
(42){02695895} place, localize, localise
Sense 2
(42){02692335} localize, localise, focalize, focalise
Sense 3
(41){02509919} localize, localise
Sense 4
(36){01711749} set, localize, localise, place

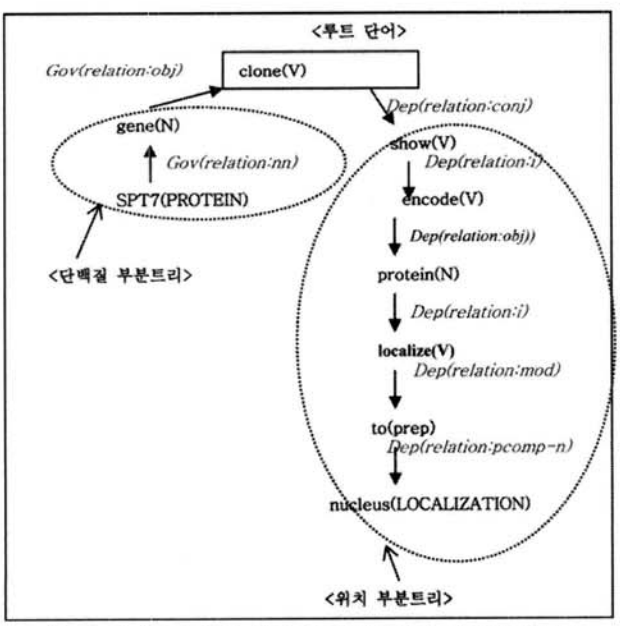
(그림 2) 〈localize〉에 대한 어휘망(WordNet) 정보

한 쌍들만을 추출하도록 한다. 아래의 단계를 통하여 각 쌍 후보로부터 언어학적 특징을 추출한다.

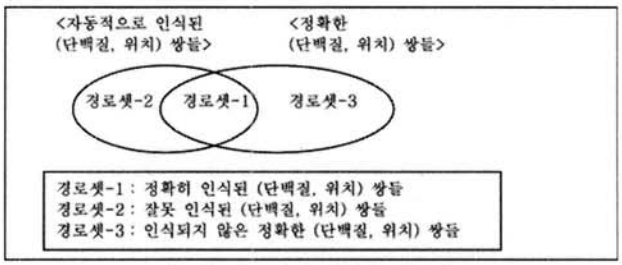
3.2 단계 1: 3가지 타입의 구문 의존 경로 셋 구축

우선 학습 데이터에서 모든 (단백질, 위치) 쌍 후보에 대하여 구문 의존 경로를 구축한다. 예제로, (그림 3)은 (그림 1)의 문장에서 추출된 (단백질, 위치) 쌍인 (SPT7, nucleus)에 대한 구문 의존 경로를 보여준다. (SPT7, nucleus)에 대해 구축된 구문 의존 경로는 <SPT7(단백질) → gene(N) → clone(V) → show(V) → encode(V) → protein(N) → localize(V) → to(prep) → Nucleus(위치)>이다. 이 구문 의존 경로는 (그림 1)의 구문 의존 트리에서 단백질 노드에서 시작하여 위치 노드로 끝나는 경로의 트리구조를 가져온 것이다.

학습데이터에서 자동적으로 인식된 모든 (단백질, 위치) 쌍 후보들 중 정확한 (단백질, 위치) 쌍들의 특징을 판단하기 위하여, 첫 단계에서 3가지 타입의 경로셋을 구축한다. 경로셋-1은 정확히 인식된 (단백질, 위치) 쌍의 구문 의존 경로로 구성되어 있고, 경로셋-2는 잘못 인식된 쌍들의 경로로 이루어져 있다. 경로셋-3은 정확하지만 인식되지 않은



(그림 3) (그림 1)의 문장에 대한 단백질 노드부터 위치 노드까지의 구문 의존 경로의 예



(그림 4) 3가지 타입의 경로셋

쌍들의 경로들로 구성되어 있다. (그림 4)에 각 경로셋이 벤다이어그램으로 표현되어 있다.

(그림 3)의 트리에서, clone(V)이 루트다. 우리는 루트의 왼쪽 부분트리를 '단백질 부분트리'라 부르기로 한다. 단백질 부분트리의 단말노드 단어는 단백질이 된다. 유사한 방법으로, 루트의 오른쪽 부분트리를 '위치 부분트리'라 부른다. 위치 부분트리의 단말노드는 위치 단어가 된다. 구문 의존 경로가 정확한 단백질 위치 정보를 내포하고 있는지 결정하기 위해, 루트와 단백질 부분트리, 위치 부분트리의 특징을 아래와 같이 고려한다.

3.3 단계 2: 루트 정보의 추출

루트 노드는 단백질 부분트리와 위치 부분트리 사이를 연결하는 중요한 열쇠이고, 구문 의존 경로가 단백질의 위치 정보를 담고 있는지를 알려주는 정보가 있다고 우리는 판단한다. 따라서 두 번째 단계에서는 루트 정보를 추출하는 것을 목적으로 한다. 구축된 경로셋-1, 2, 3으로부터 다음과 같은 정보 <루트 단어의 어근, 루트 노드의 구문태그>를 추출하고, 데이터 부족 문제의 해결을 위하여 루트 단어의 어근 대신 의미 정보 또한 추출하여 <루트 단어의 의미, 루트 단어의 구문태그> 정보 또한 구축한다.

루트 단어의 의미로는, 어휘망 내에서의 해당 단어 의미들 중 가장 자주 쓰이는 의미의 코드를 선택한다. 어근과 구문태그 정보는 MINIPAR 구문분석기의 결과에서 추출한다. <표 1>은 루트노드를 대상으로 추출한 정보의 예를 보여준다. 실험데이터의 양이 많지 않기 때문에, 긍정적 정보 대신에 부정적 루트 정보를 구축한다. 그리하여 부정적 정보에 속하는 단백질 위치 후보는 잘못 인식된 후보라고 판단하고 정답셋에서 제외한다. 부정적 루트 정보는 3가지 타입의 구문 의존 경로 중에서 단지 경로셋-2(잘못 인식된 경

〈표 1〉 정확한 경로(경로셋-1과 경로셋-3)의 루트 정보 예

단어의 어근	의미태그	구문태그
colocalize	정보없음	obj2
location	00027167	obj
localization	00155487	pcomp-n
exhibit	02631856	I
bind	01356750	I
synthase	정보없음	S
….	….	….

로셋)에만 존재하는 루트 정보를 추출함으로써 구축하였다. 5-집단 교차검증(5-fold cross validation)을 사용한 실험에서, 학습데이터로부터 평균 325개의 부정적 루트 패턴을 획득하였다.

3.4 단계 3: 단백질 부분트리와 위치 부분트리의 구문-의미 패턴 추출

3번째 단계에서, 3가지 타입의 경로 셋으로부터 단백질 부분트리와 위치 부분트리의 구문-의미 패턴(구문과 의미 정보로 구성된 패턴)을 추출한다. (그림 3)에서 나타나는 구문 의존 경로에서, 루트인 clone(V)는 단백질 SPT7이나 위치 nucleus와 직접적인 구문관계가 없고, gene(N)과 show(V)와 직접적인 연관이 있다. 즉, 단백질 SPT7은 단백질 부분트리 내에서 관계 nn을 거친 후 간접적으로 루트와 관계를 맺고 있고, 위치 nucleus 또한 위치 부분트리에서 관계 I, obj, mod, pcomp-n을 통해 루트노드에 간접적으로 연결되어 있다. 그러므로, 구문 의존 경로가 단백질의 정확한 위치 정보를 담고 있는지를 결정하기 위해, 단백질과 위치 단어가 각각 해당 부분트리에서 적절한 관계들을 통해 루트와 연결되어 있는지를 살펴본다.

위의 3가지 타입의 경로들로부터 두 부분트리의 각 노드에 대한 구문-의미 패턴을 추출한다. 우선 <구문태그, 어근, 구문 방향>을 추출하고, 데이터 부족 문제의 보완을 위해 단어의 어근 대신 사용할 의미태그로서 어휘망에서의 의미코드 또한 추출하여 <구문태그, 의미태그, 구문 방향>을 구축한다. 구문 방향은 현재 노드가 이전 노드의 지배소(gov)인지 의존소(dep)인지를 나타내는 정보로서, gov와 dep, 두 값 중의 하나이다.

<표 2와 3>은 각각 정확한 경로 셋(셋-1과 셋-3)과 잘못된 경로 셋(셋-2)에서의 구문-의미 패턴의 예를 보인다. 우리는 아래와 같은 2가지 타입의 구문-의미 패턴을 구축했다.

(1) 경로셋-1과 셋-3으로부터 추출한 단백질과 위치 부분트리에 대한 정확한 구문-의미 패턴
(2) 경로셋-2로부터 추출한 단백질과 위치 부분트리에 대한 잘못된 구문-의미 패턴

(1)과 (2) 중, 단지 (2)에만 존재하는 구문-의미 패턴을 추출함으로써 단백질과 위치 부분트리에 대한 부정적 패턴을 구축했다. 5-집단 교차검증에 의한 학습에서 단백질 부분트리에 대해 평균 968개의 부정적 패턴을 얻었고, 위치 부분트리에 대해 평균 1133개의 부정적 패턴을 얻었다.

〈표 2〉 정확한 경로(경로셋-1과 경로셋-3)의 구문-의미 패턴의 예

단어의 어근	의미태그	구문의 방향	구문태그
colocalize	정보없음	dep	obj2
colocalize	정보없음	gov	nn
receptor	05608868	gov	pcomp-n
bud	11674914	gov	pcomp-n
location	00027167	dep	conj
encode	00993892	dep	i
…….	……	……	……

〈표 3〉 잘못된 경로(경로셋-2)의 구문-의미 패턴의 예

단어의 어근	의미태그	구문의 방향	구문태그
absence	13960974	gov	s
absent	01847672	dep	pred
insoluble	02265891	gov	pred
defect	14464005	gov	pcomp-n
defective	01752953	dep	
delete	01549187	dep	i
deletion	13524399	dep	pcomp-n
…	…….	…	…

3.5 획득한 정보를 실험데이터에 이용

위의 3가지 단계를 통하여 획득한 정보를 실험데이터에 다음과 같이 이용한다. 학습데이터에서 단백질과 위치 단어 후보를 태깅할 때 사용한 것과 같은 방법으로 실험데이터에서 단백질과 위치 단어 후보를 먼저 태깅한다. 각 실험 문장에서 모든 (단백질, 위치) 쌍 후보에 대해 MINIPAR를 사용하여 구문 의존 경로를 구축한다. 각 구문 의존 경로에 대해, 다음과 같은 절차를 수행하여 정확한 쌍인지를 판단한다.

(1) 구문 의존 경로의 루트 노드가 부정적 루트 정보와 매치되는가.
(2) 경로의 단백질(위치) 부분트리가 부정적 단백질(위치) 부분트리 패턴을 포함하는가.

위의 2가지 조건 중 한 가지라도 만족하면, (단백질, 위치) 쌍 후보를 답에서 제외한다.

4. 결 과

이전 방법들과의 효과적인 비교를 위해, 메드라인 데이터베이스에서 871개 요약문의 7,245개 문장으로 구성된 M. Goadrich 등[3]의 데이터를 이용했다. 실험을 위하여 5-집단 교차검증(5-fold cross validation)을 사용하여 평균값을 구했다.

본 실험은 아래의 내용에 초점을 두고 있다.

1. 본 논문에서 제안한 방법의 성능과 이전 방법들의 성능과의 비교
2. 언어학적 정보가 제거되었을 때의 성능변화

실험 결과들은 다음과 같다.

1. 단백질의 세포 내 위치 정보의 인식을 위한 본 방법은 58.90%의 조화평균을 보였다.<표 4>
2. 본 방법은 이전의 방법들보다 12-25% 더 나은 성능을 보여주었다.<표 5>

〈표 4〉 정보를 사용하지 않을 경우 성능 변화

		모든 정보를 사용하였을 경우	정보의 제거 후 성능			
			구문 정보 제거 후			의미 정보 제거 후
			루트 정보 제거 후	부분트리 정보 제거 후	구문태그 정보 제거 후	의미태그 정보 제거 후
학습데이터에서의 결과	정확률(%)	72.84	68.55	40.95	60.71	36.48
	재현율(%)	76.29	74.48	80.35	13.96	82.44
	조화평균(F-measure)	74.53	71.39	54.25	22.70	50.58
실험데이터에서의 결과	정확률(%)	53.05	49.60	38.89	50.00	36.96
	재현율(%)	66.21	64.74	81.05	12.11	62.63
	조화평균(F-measure)	58.90	56.17	52.56	19.50	46.49

〈표 5〉 본 논문의 시스템과 다른 시스템간의 성능 비교

		Craven(1999)	Goadrich(2004)	Skounakis(2003)	본 논문의 시스템
실험데이터에서의 결과	정확률	0.92	0.58	0.48	0.53
	재현율	0.21	0.40	0.40	0.66
	조화평균(F-measure)	0.34	0.47	0.44	0.59

3. 구문이나 의미 정보가 제거되었을 때, 성능이 급격히 감소되었다.<표 4>

<표 5>에서 보이는 모든 시스템들은 실험을 위해 똑같은 데이터를 사용했고, 그 중 본 논문의 시스템이 정확률 53.05%, 재현율 66.21%, 그리고 조화평균 58.90%의 가장 좋은 성능을 보였다.

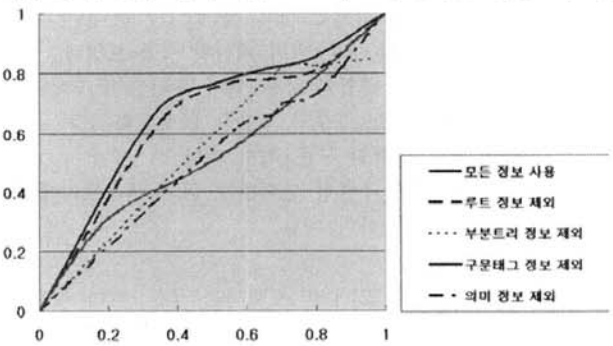
또한, 3장에서 설명한 단계들을 통해 획득한 각 정보가 성능에 어떤 영향을 미쳤는지 알아보기 위해, 본 논문에서 제안한 정보가 제거되었을 때 성능의 변화를 측정했다. <표 4>에서 알 수 있듯이, 구문태그를 사용하지 않을 경우 재현율이 가장 낮았다. 구문태그를 사용하지 않으면 매칭 조건이 더 쉬워지기 때문에, 더 많은 후보들이 부정적 조건에 부합하게 된다. 따라서, 정답 쌍에서 제외되는 경우가 많아지게 되고 재현율이 낮아진다.

의미 정보를 제외하고 실험하였을 경우, 정확률이 가장 낮았다. 구문 의존 경로의 단어와 부합하는 정보가 패턴에 없을 때, 단어의 어근 대신 의미 정보를 사용하게 된다. 따라서 의미 정보를 제외하는 경우, 부정적 패턴에 부합하는 경우가 줄어들게 된다. 결국 도출되는 쌍들이 많아지게 되고, 정확률은 더 낮아지게 된다.

또한 부분트리와 루트 정보를 제외하였을 때의 성능을 측정했다. 마찬가지로 부정적 패턴과 부합하는 경우가 줄어들게 되므로 도출된 쌍이 많아지고, 또한 정확률이 낮아진다.

이 결과들은 루트 정보, 구문 부분트리 패턴들, 그리고 어휘망이 단백질 위치 정보를 인식하는 데 중요하다는 것을 보여준다. 특히, 구문태그 정보는 재현율을 높이는 데 유용하고, 의미태그는 정확률 향상에 도움을 주는 것을 알 수 있다.

위의 결과들을 전체적으로 쉽게 알 수 있도록 하기 위하여, (그림 5)에서 각 실험의 수신자판단특성곡선(이하 ROC 곡선) 결과를 보이고 있다. ROC 곡선은 이진형의 목표변수를 가지는 모형들의 성능을 비교, 평가하는데 매우 유용한 도표로 사용된다. ROC 곡선은 X축(1-특이도)와 Y축(민감도)로 각 분류기준 값에 대해 나타나며, 이러한 결과에서 그래프가 도표의 왼쪽 상단으로 더 가까운 모형을 성능 면에서 우수한 모형으로 판단하면 된다.[24]. 본 논문의 ROC곡선에서 X축은 양성오류율(false positive rate)을 나타내고 Y축은 진양성율(true positive rate)을 나타낸다. 양성오류율은 정답으로 잘못 분류된 오답의 비율이고, 진양성율은 실제 정답들 중 분류 결과 정답으로 바르게 분류된 비율로서 재현율과 같다. ROC 곡선에서 알 수 있듯이 모든 정보를 다 사용한 실험의 경우, 다른 실험들에 비해 전체적으로 위쪽에 있으며, 이는 각각의 분류기준값에서의 민감도와 특이도가 높으며 이 실험의 성능이 가장 높다는 것을 보여준다.



(그림 5) 각 실험에 대한 ROC 곡선

5. 결 론

단백질의 세포 내 위치 정보를 인식하는 성능을 높이기 위하여, 본 논문은 전체 구문분석기와 어휘망을 사용한 방

법을 제안한다. 첫 단계에서 단백질 단어 후보로부터 위치 단어 후보에까지 이르는 구문 의존 경로를 구축한다. 두 번째 단계에서는 부정적 루트 정보를 추출한다. 마지막으로, 구문-의미 패턴을 단백질 부분트리와 위치 부분트리를 대상으로 추출한다. 그리하여 부분트리에 대한 구문-의미 패턴 정보와 루트 정보를 기반으로 (단백질, 위치) 쌍을 인식한다. 실험 결과, 본 논문의 방법이 이전 방법들보다 훨씬 더 좋은 성능인 정확률 53.05%, 재현율 66.21%, 조화평균 58.90%를 보였다. 따라서 본 논문에서 제안하는 방법이 단백질 위치 정보를 인식하는 데 효과적임을 알 수 있다. 더욱이, 구문과 의미 정보가 이 방법의 성능에 중요하다는 것을 증명했다. 추후, 학습 데이터의 양을 늘려서 더 많은 데이터로 실험을 수행하는 것이 필요하다.

참 고 문 헌

[1] C. Blaschke, L. Hirschman and A. Valencia, "Information Extraction in Molecular Biology," Briefings in Bioinformatics, Vol.3, pp.154-165, 2002.

[2] M. Craven and J. Kumlien, "Constructing Biological Knowledge Bases by Extracting Information from Text Sources," Proc. of the 7th Int'l Conf. Intelligent Systems for Molecular Biology, AAAI Press, pp.77-86, 1999.

[3] M. Goadrich, L. Oliphant and J. Shavlik, "Learning Ensembles of First-Order Clauses for Recall-Precision Curves: A Case Study in Biomedical Information Extraction," Proc. of the 14th International Conference on Inductive Logic Programming (ILP), pp.98-115, 2004.

[4] D. Page and M. Craven, "Biological Applications of Multi-Relational Data Mining," ACM SIGKDD Explorations Newsletter, Vol.5, pp.69-79, 2003.

[5] M. Skounakis, M. Craven and S. Ray, "Hierarchical Hidden Markov Models for Information Extraction," Proc. of the 18th International Joint Conference on Artificial Intelligence, pp.427-433, 2003.

[6] F. Rinaldi, G. Schneider, K. Kaljurand, M. Hess, C. Andronis, O. Konstanti and A. Persidis, "Mining of Functional Relations between Genes and Proteins over Biomedical Scientific Literature using a Deep-Linguistic Approach," Artificial Intelligence in Medicine, Vol.39, pp.127-136, 2007.

[7] S. Riedel and E. Klein, "Genic interaction extraction with semantic and syntactic chains," Proc. of ICML05 Workshop on Learning Language in Logic (LLL05), 2005.

[8] M. Goadrich, L. Oliphant and J. Shavlik, "Learning to extract genic interactions using Gleaner," Proc. of ICML05 Workshop on Learning Language in Logic (LLL05), 2005.

[9] B. Stapley, L. Kelley and M. Sternberg, "Predicting the sub-cellular location of proteins from text using support vector machines," Proc. of the Pacific Symposium on Bio-computing, pp.374-385, 2002.

[10] B. Rosario and M. Hearst, "Classifying semantic relations in bioscience texts," Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), pp.430-437, 2004.

[11] M.Krogel, M. Denecke, M. Landwehr and T. Scheffer, "Combining data and Text Mining Techniques for Yeast Gene Regulation Prediction: A Case Study," ACM SIGKDD Explorations Newsletter, Vol.4, pp.104-105. 2002.

[12] M. Krogel and T. Scheffer, "Multi-Relational Learning, Text Mining, and Semi-Supervised Learning for Functional Genomics," Machine Learning, Vol.57, pp.61-81, 2004.

[13] K. Lee, D. Kim, D. Na, D. Lee and K. Lee, "PLPD: Reliable Protein Localization Prediction from Imbalanced and Overlapped Datasets," Nucleic Acids Research, Vol.34, pp.4655-4666, 2006.

[14] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. S. Wishart, B. Poulin, J. Anvik, C. Macdonell and R. Eisner, "Predicting Subcellular Localization of Proteins using Machine-Learned Classifiers," Bioinformatics, Vol.20, pp.547-556, 2004.

[15] H. Shatkay, A. Höglund, S. Brady, T. Blum, P. Dönnes and O. Kohlbacher, "SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data," Bioinformatics, Vol.23, pp.1410-1417, 2007.

[16] D. Lin, "Dependency-based evaluation of MINIPAR," Workshop on the Evaluation of Parsing Systems, 1998.

[17] J. Chen, G. He, Y. Wu and S. Jiang, "UNT at TREC 2004: Question Answering Combining Multiple Evidences," Proc. of TREC. 2004.

[18] C. Lee, G. G. Lee and M. Jang, "Dependency structure language model for topic detection and tracking," Information Processing and Management, Vol.3, No.5, pp.1249-1259, 2007.

[19] R. Higashinaka, R. Prasad and M. Walker, "Learning to Generate Naturalistic Utterances Using Reviews in Spoken Dialogue Systems," Proc. of COLING/ACL, pp.265-272, 2006.

[20] D. Martínez, E. Agirre and L. Màrquez, "Syntactic features for high precision word sense disambiguation," Proc. of the 19th international conference on Computational linguistics, pp.1-7, 2002.

[21] R. Mihalcea and D. Moldovan, "Document Indexing Using Named Entities," Studies in Informatics and Control, Vol.10, No.1, 2001.

[22] 김미영, "구문관계에 기반한 유전자 상호작용 인식", 정보처리학회논문지, Vol.14-B, No.5, pp.383-390, 2007.

[23] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. J. Miller, "Introduction to WordNet: An on-line lexical database," International Journal of Lexicography, Vol.3, No.4, pp.235-244, 1990.

[24] 최종우, 한상태, 강현철, 김은석, 김미경, 이성건, "SAS Enterprise Miner 4.0을 이용한 데이터 마이닝 기능과 사용법", 자유아카데미, 2001.

김 미 영

e-mail : miykim@sungshin.ac.kr

1995년~1999년 포항공과대학교 컴퓨터공학과 학사졸업

1999년~2005년 포항공과대학교 대학원 컴퓨터공학과 박사졸업

2006년~현 재 성신여자대학교 컴퓨터정보학부 전임강사

관심분야 : 자연언어처리, 정보검색, 텍스트마이닝, 바이오인포매틱스 등