한국산학기술학회논문지
Vol. 7, No. 2, pp. 182-187, 2006

은닉 마르코프 모델과 계층 정보를 이용한 개체명 경계 인식

임희석[1*]

Named Entity Boundary Recognition Using Hidden Markov Model and Hierarchical Information

Heui-Seok Lim[1*]

요 약 본 논문은 통계 기반 접근 방식인 HMM(Hidden Markov model)과 생물학의 개체명에 관한 온톨로지 정보를 이용한 생물학 문서에서의 개체명(named entity) 경계 인식 방법을 제안한다. 제안하는 방법은 31개의 자질 정보를 이용한 평탄화 기법을 사용하며 생물학 개체명의 계층 정보를 이용하여 HMM의 자료 부족 문제를 완화시킬 수 있도록 하였다. 개체명 경계 인식의 학습과 실험을 위하여 GENIA 코퍼스 ver 2.1을 사용하였으며 개체명 경계 인식 실험을 수행한 결과 모든 부류를 사용한 경우보다 정확도 및 실행 속도가 개선됨을 확인하였다.

Abstract This paper proposes a method for boundary recognition of named entity using hidden markov model and ontology information of biological named entity. We uses smoothing method using 31 feature information of word and hierarchical information to alleviate sparse data problem in HMM. The GENIA corpus version 2.1 was used to train and to experiment the proposed boundary recognition system. The experimental results show that the proposed system outperform the previous system which did not use ontology information of hierarchical information and smoothing technique. Also the system shows improvement of execution time of boundary recognition.

Key Words : Named Entity, Recognition of Named Entity, BioInfomatics, HMM, GENIA corpus

1. 서 론

휴먼 게놈 프로젝트에서 전체 DNA 서열과 다양한 종류의 유전자를 발견함에 따라 유전자의 역할을 알아내려는 연구가 활기를 띠게 되었다. 특히, 단백질과 핵산 그리고 단백질과 단백질의 결합에 대한 많은 문헌들이 쏟아져 나왔다. 이런 막대한 양의 문헌들로부터 일반 PubMed나 Entrez와 같은 검색 엔진을 이용해 필요한 정보만을 골라 찾아내기란 매우 어려운 일이다. 따라서, 자연어처리 기법을 이용한 지능적인 정보 추출(IE) 시스템에 대한 요구가 대두되었다.

개체명 인식은 정보 추출의 전단계로서 문서로부터 개체명을 인식(recognition)하고 분류(classification)하는 작업이라 할 수 있다. 즉, 단백질, DNA, RNA 등과 같은 정보 슬롯에 개체들을 찾아서 채우는 것이라 할 수 있다. 생물학 문서를 대상으로 한 개체명 인식이 특히 어려운 이유는 개체명을 이루는 단어의 수가 많고 그 단어들이 다른 곳에서도 사용될 뿐만 아니라 명명 규칙이 자의적으로 결정되는 경우가 많아 같은 개체명이라 할지라도 여러 가지 형태로 표기될 수 있기 때문이다. 또한 접속사 등으로 다른 개체명과 연결되어 하나의 개체명을 이루는 경우도 있어 인식을 더욱 어렵게 한다.

개체명 인식을 위한 기존의 접근 방법은 크게 4가지로 나눌 수 있다. 첫째, 사전 기반 접근 방법은 사전을 이용하는 방법이다. 이 방법은 성능이 사전의 품질에 의존하고 미등록어를 인식하지 못하는 단점이 있다. 둘째, 규칙 기반 접근 방법[5]은 수작업으로 만든 규칙을 이용하는 방법이다. 이는 사전에 없는 미등록어를 예측할 수 있다는 장점이 있지만 모든 규칙을 수작업으로 얻기 어렵고 규칙 획득에 드는 비용이 크다는 단점이 있다. 셋째, 통계적인 접근 방법은 HMM 모델[1,2], 결정 트리(decision

본 논문은 2006년도 한신대학교 학술연구비 지원에 의하여 연구되었음
[1]한신대학교 컴퓨터정보소프트웨어학부
*교신저자 : 임희석(limhs@hs.ac.kr)

tree) 모델, 최대 엔트로피 모델(maximum entropy model), SVM(Support Vector Machine)[3]과 같은 기계 학습법을 사용하는 방법이다. 이 중에서 HMM은 음성 인식 및 품사 태깅 등에서 자주 이용되어 왔고, 생물정보학 분야에서도 유전자 서열 비교 및 단백질 구조 인식 모델로 사용되어 좋은 성능을 보였다. 넷째, 혼합형 접근 방법[6]은 앞의 방법들을 혼합하여 사용하는 방법이다.

개체명 인식을 위해 HMM 모델을 사용한 Collier[2]는 10개의 부류를 가진 100개의 문서를 대상으로 실험하였으며, 평균 F-score는 72.8로 보고되었다. 이는 적은 수의 학습 문서와 적은 부류를 대상으로 실험하였기 때문에 본 연구와의 직접접인 비교는 불가능하다. 본 연구의 동기가 된 SVM을 사용한 Kazama[3]의 실험은 비교적 좋은 성능을 보였지만, 개체명이 아닌 것도 포함하여 분류를 수행하였기 때문에 휴리스틱한 복잡한 튜닝 방법을 사용해야 했다. 이런 문제점을 피하기 위해서는 개체명들을 대분류로 구분하고 개체명 분류와 개체명 분류가 아닌 것으로 구분하는 대분류를 한 후, 개체명에 해당하는 대분류만을 SVM과 같은 분류기로 세 분류하는 나가는 2단계 방법을 사용하는 것이 보다 합리적일 것이다.

2단계 방법을 사용하는 개체명 인식에 있어서 1단계의 대분류기의 성능이 전체 성능에서 매우 중요한 역할을 차지한다. 본 논문은 2단계 개체명 인식을 위한 1단계 과정으로 HMM과 온톨로지 정보를 이용한 개체명 경계 인식 방법을 제안한다. 본 논문은 개체명 경계는 상대적으로 부류의 영향을 덜 받고 단어의 형태 및 문맥에 더 영향을 받으며, 개체명 분류 작업은 단어의 형태나 문맥보다는 구성 어휘가 무엇이냐에 크게 좌우된다고 보았다. 따라서 개체명의 계층 정보를 이용하여 세 부류들을 대부류로 묶어 부류의 개수를 줄인 뒤 HMM을 사용해 경계 인식 실험을 수행하였다.

2. GENIA 코퍼스

GENIA 코퍼스는 MEDLINE 데이터베이스로부터 얻은 논문 요약을 개체명 태깅한 코퍼스이다. 현재, 1,000여 개의 논문 요약이 생물학 전문가에 의해 개체명 태깅되어 있고, 공개적으로 이용이 가능하다[2]. GENIA 코퍼스의 논문 요약은 "human AND blood cell AND transcription factor"라는 질의를 통해 MEDLINE 데이터베이스에서 검색된 5,000개의 논문 요약으로부터 발췌된 것이다. GENIA 코퍼스는 24개의 서로 다른 개체명 부류가 있으며, 이 부류들의 온톨로지 정보를 이용할 수 있다. 이는 기존 연구[2]에서 사용한 코퍼스에 비해 상대적으로 많은 부류이다. 이는 GENIA 코퍼스로 작업하는 것이 매우 어렵다는 것을 시사한다.

[2] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/를 통해 이용이 가능하다.

표 1. GENIA 코퍼스의 기본 통계

# of sentence	5,781
# of words	161,462
# of named entities	26,170
# of words in Nes	54,191
# of words not in NEs	107,271
Av. Length of NEs	2.07

3. HMM을 이용한 개체명 경계 인식 모델

개체명 경계 인식 모델은 "길이가 N인 단어열(문장) $w_{1,N}= w_1 w_2 \ldots w_N$ 이 주어졌을 때, 가장 확률이 높은 개체명 부류열 $c_{1,N}= c_1 c_2 \ldots c_N$ 을 구하는 것"으로 (식 1)과 같이 정의할 수 있다. (식 1)에서 w_i 는 문장에서 i 번째에 나타나는 단어를 나타내며, c_i 는 i 번째 단어에 할당되는 개체명 부류를 의미한다.

$$T(w_{1.N}) \overset{def}{=} \arg\max_{c_{1.N}} P\ (c_{1.N} | w_{1.N}) \quad (식\ 1)$$

(식 1)은 (식 2)로 풀어쓸 수 있으며 (식 2)에서 $P(w_{1,N})$은 모든 $c_{\{1,n\}}$에 대해서 상수이므로 (식 3)과 같이 쓸 수 있다.

$$T(w_{1.N}) = \underset{c_{1.N}}{argmax} \frac{P(c_{1.N}, w_{1.N})}{P(w_{1.N})} \quad (식\ 2)$$

$$= \underset{c_{1.N}}{argmax}\ P(c_{1.N}, w_{1.N}) \quad (식\ 3)$$

그러나 (식 3)과 같은 모델로 개체명 부류를 인식하기 위해서는 문장 단위의 통계 정보 $P(c_{1,N}|w_{1,N})$를 사용하여야 하나 자료 부족 문제로 인하여 이러한 통계 정보를 얻는 것은 거의 불가능하다.

따라서 각 부류는 현재 단어 w_i와 이전 단어 w_{i-1}, 그리고 이전 부류 c_{i-1}에 의해서만 영향을 받는다고 가정할 경우 개체명 부류 인식 모델은 최종적으로 (식 4)와 같이 정의할 수 있다.

$$T(w_{1,N}) \overset{def}{=} \underset{c_{1,N}}{\operatorname{argmax}} P\ (c_i |\ w_i, w_{i-1}, c_{i-1}) \quad (식\ 4)$$

위 모델에 따르면 현재 어휘(w_i), 이전 어휘(w_{i-1}), 그리고 이전 어휘의 부류(c_{i-1})가 주어졌을 때 현재 단어의 어휘의 부류(c_i)가 관측될 확률을 의미하는 $P(c_i | w_i, w_{i-1}, c_{i-1})$을 구해야 하지만, 이 확률값을 계산하기 위해서는 여전히 자료 문제(sparse data problem)가 발생할 수 있다. 이에 본 논문은 단어를 표현하는 중간 형태로서 자질(f_i)을 이용하는 (식 5)와 같은 선형 보간 방법을 사용한다. (식 5)에서 $\sum \lambda_i = 1.0\ (\lambda_0 \geq \lambda_1 \cdots \geq \lambda_4)$를 만족해야 하며, f_i는 해당 단어의 자질 벡터를 의미한다. 여기서 λ_i는 선형 보간 시 어느 요소의 확률값을 더 중요하게 사용할지를 결정하는 가중치를 의미한다.

$$\begin{aligned}
&P(c_i |\ < w_i, f_i >, < w_{i-1}, f_{i-1} >, c_{i-1}) = \\
&\lambda_0 P(c_i |\ < w_i, f_i >, < w_{i-1}, f_{i-1} >, c_{i-1}) \\
&+ \lambda_1 P(c_i | \leftarrow, f_i >, < w_{i-1}, f_{i-1} >, c_{i-1}) \\
&+ \lambda_2 P(c_i |\ < w_i, f_i >, \leftarrow, f_{i-1} >, c_{i-1}) \\
&+ \lambda_3 P(c_i | \leftarrow, f_i >, \leftarrow, f_{i-1} >, c_{i-1}) \\
&+ \lambda_4 P(c_i |\ c_{i-1}) \quad (식\ 5)
\end{aligned}$$

개체명 경계 인식을 위하여 사용되는 자질의 선택은 다른 패턴 인식 시스템에서와 같이 중요한 요소이다. 일반적인 패턴 인식 시스템의 자질 선정에는 도메인 지식과 시스템 개발자의 사전 지식(prior knowledge)이 매우 중요하게 사용된다. 본 연구에서도 개체명이 태깅된 코퍼스 분석과 연구자의 경험에 의해서 추출된 72개의 자질 후보(feature candidates)를 추출하였고, 자질 후보 집단에 대해서 x^2(chi-square) 검정을 통하여 선정된 31개의 자질을 사용하였다. x^2(chi-square) 검정을 이용한 31개의 자질 선정은 72개 자질 후보의 x^2 검정량을 계산하고, 특정 임계값 이상의 값을 갖는 자질을 선택하였다.

본 연구는 단어의 자질로 31개를 선정하여 사용하였으며 실험에 사용된 자질의 종류와 예가 표 2에 있다. 각 자질은 이진값(binary)으로 표현된다. 즉, 각 단어가 어떤 자질에 해당하면 1을, 그렇지 않으면 0으로 표시하여 0과 1의 벡터 형태로 표현한다. 단어에 대한 자질은 대소문자, 숫자, 특수 문자 등을 주요 특징으로 뽑았다. 그리고 단어에 대한 자질은 중복해서 사용하지 않고 오직 하나만 할당하였다. 자질 STOPWORD는 단어가 불용어이면 1, 아니면 0을 할당했다. 자질 PREFIX, SUFFIX는 코퍼스에서 2개 이상의 단어로 이루어진 개체명의 시작 단어와 끝 단어의 출현 빈도를 세어 추출한 상위 100개의 리스트 중에 단어가 포함되어 있으면 1, 아니면 0을 할당하였다. 자질VOCABULARY는 개체명에서 쓰인 단어들의 빈도를 세어 추출한 상위 100개 리스트 중에 단어가 포

표 2. 자질의 종류와 각 자질의 예

#	자질	예	#	자질	예
1	STOPWORD	of, for	17	SEMOCOLON	;
2	SINGLE_CAP	M	18	APOSTROPHE	'
3	TWO_CAPS	RalGDS	19	OPEN_PAREN	(
4	INIT_CAP	Interleukin	20	CLOSE_PAREN	)
5	IN_CAP	kappaB	21	COMMA	,
6	CAPS_DIGIT	GATA1	22	PERIOD	.
7	ALL_UPPER	HPC	23	QUESTION_MARK	?
8	GREEK_LETTER	kappaB	24	PERCENT	%
9	ALPHA_NUMERIC	p52	25	ETC_SYMBOL	* + ^
10	ALL_LOWER	kinases	26	PREFIX	human, T
11	DIGIT	21, 1999	27	SUFFIX	cells, gene
12	HYPHON	-	28	VOCABULARY	protein, receptor
13	SLASH	/	29	PREVWORD	the, of, in
14	OPEN_SQUARE	[	30	NEXTWORD	is, by, to
15	CLOSE_SQUARE	]	31	POS	NN, JJ
16	COLON	:			

함되면 1, 포함되지 않으면 0을 할당했다. 자질 PREVWORD, NEXTWORD는 각각 개체명 바로 앞에 쓰인 단어와 개체명 바로 뒤에 쓰인 단어의 빈도를 세어 추출한 상위 100개의 리스트 중에 단어가 포함되면 1, 아니면 0을 할당하였다. POS는 단어의 품사가 "NN" 또는 "JJ"인 경우 1, 아니면 0을 할당하였다.

본 연구에서 개체명이 부착되어 있는 기존의 자료를 분석해 본 결과 개체명이 시작되는 경계에 대한 다음 두 가지의 특징을 찾을 수 있었다. 첫째, 개체명이 시작되는 경계는 상대적으로 개체명 부류의 영향을 덜 받고 해당 단어의 형태 및 문맥에 더 영향을 받는다. 둘째, 개체명 분류 작업은 단어의 형태나 문맥보다는 구성 어휘가 무엇이냐에 크게 좌우된다. 이러한 특징을 반영할 때 개체명 경계 인식을 위해서는 개체명의 계층 정보를 이용하여 세 부류들을 대 부류로 묶어 부류의 개수를 줄이는 것이 성능 향상에 기여할 수 있을 것으로 판단하였다. 이에 개체명 부류를 표 3과 같이 4 수준의 부류로 분류한 뒤 각 수준에 해당하는 부류 경계 인식 실험을 수행한다.

4. 실험 및 평가

학습과 실험을 위한 문서 집합은 동경대에서 구축한 GENIA 코퍼스 ver 2.1을 사용하였다. 성능을 평가하기 위한 척도로는 정확률(P), 재현율(R), F-score를 사용하였다. 정확률은 인식 시스템이 결정한 개체명이 정답 문서와 비교해 얼마나 정확한지의 비율이고, 재현율은 정답 문서에 있는 개체명이 얼마나 올바르게 인식되었는지를 나타내는 비율이다. F-score는 정확률과 재현율을 결합한 하나의 척도로 MUC의 평가 척도로 자주 이용된다. F-score의 정의는 (식 6)과 같다.

표 3. 개체명의 수준별 부류

	1-level	2-level	3-level	4-level
ontology	SOURCE	NATURALSOURCE	ORGANISMNATURALSOURCE	MONO_CELL
				MULTI_CELL
				VIRUS
				ORGANISM
			OTHER NATURAL SOURCE	BODY_PART
				CELL_COMPONENT
				CELL_TYPE
				TISSUE
		ARTIFICIALSOURCE	ARTIFICIAL SOURCE	CELL_LINE
				OTHER_ARTIFICIAL_SOURCES
	SUBSTANCE	ORGANICCOMPOUND	AMINO ACID	PROTEIN
				AMINO_ACID_MONOMER
				PEDTIDE
			NUCLEIC ACID	DNA
				RNA
				NUCLEOTIDE
				POLY_NUCLEOTIDE
			OTHER ORGANIC COMPOUND	CARBOHYDRATE
				LIPID
				ORGANIC
				OTHER_ORGANIC_COMPOUNDS
		OTHERSUBSTANCE	OTHER SUBSTANCE	ATOM
				INORGANIC
	OTHER	OTHER	OTHER	TEMP
				OTHER_NAMES

$$F-score = \frac{2 \times P \times R}{P+R} \quad (식\ 6)$$

따로 실험 문서가 존재하지 않으므로 모든 실험은 10-fold cross validation 방식을 사용하여 10번의 실험에 대한 평균을 구하였다. 각 부류에 대한 인식 결과를 표 4에 나타내었다. 여기서 사용된 λ0에서 λ4의 값은 각각 0.95 0.024 0.024 0.001 0.001이다. 이 값들은 반복 실험을 통해 경험적으로 결정되었다.

개체명을 경계 인식을 위하여 각 부류의 시작(begin)과 중간(intermediate)을 표시하는 BI 태그를 사용하였다. 개체명이 아닌 것은 "O" 태그를 사용하였다.

24개의 최하위 부류들의 상위 부류는 GENIA의 온톨로지 정의를 사용하였다. 실험에서는 온톨로지의 정의에 맞게 1수준부터 4수준까지 나누어 각 수준별로 성능을 평가하였다. 표 4에 4개 수준별로 실험 결과를 나타내었다. 표에서 4번째 수준은 최하위 부류를 모두 사용한 결과이다.

표 4. 각 계층 수준별 실험 결과

		1-level	2-level	3-level	4-level
OHTER 부류 포함	P	76.7	75.4	62.1	61.1
	R	47.8	46.1	30.9	28.0
	F	58.9	57.2	41.2	38.4
OTHER 부류 배제	P	78.4	77.1	62.0	61.1
	R	57.9	55.3	32.9	28.0
	F	66.6	64.4	43.0	38.4

표 4에서 볼 수 있듯이 결과 부류의 개수가 적을수록 경계 인식 성능이 좋았다. 부류의 개수가 많아질수록 성능이 낮아지는 것은 자료 부족 문제가 심각해지기 때문인 것으로 보인다. 이를 입증하기 위해 전체 학습 문서(테스트 집합 포함)를 가지고 실험을 수행한 내부 실험 결과 4개 수준 모두 정확률은 약 93~95%, 재현율은 약 83~86%, 평균 F-score는 88~90 정도로 비슷한 성능을 보였다. 이로부터 자료 부족 문제가 매우 심각하게 발생하고 있음을 알 수 있다. 따라서, 개체명의 경계만을 찾고자 할 때는 많은 부류를 사용하지 않는 편이 유리하다고 판단할 수 있다. 개체명 경계 인식 시에 부류를 적게 사용하는 것의 또 다른 장점은 경계 인식의 속도 개선이다. HMM에서 사용되는 동적 알고리즘의 일종인 Viterbi 알고리즘의 복잡도는 O(NT2)이다. N은 문장의 단어 수이고 T는 부류의 수이다. BI 태그를 사용하는 경우 "O"를 제외한 모든 부류의 수가 2배가 되므로 계산량이 4배 가량 늘어나게 된다. 실험 결과 1 수준의 인식 속도와 4 수준의 인식 속도는 거의 10배 이상 차이가 난다. 물론4 수준은 경계 인식과 함께 분류까지 수행하므로, 분류 수행 속도가 빠진 1,2,3 수준과 직접적으로 속도를 비교하는 것은 공정하지 않을 수 있다. 경계 인식 문제만 놓고 보면 부류의 수가 적을수록 좋은 성능을 보이지만, 분류까지 고려해야 할 경우 부류의 수를 너무 적게 사용하면 그것을 분류할 때 수행 시간이 많이 걸릴 수 있다. 따라서, 경계 인식 성능과 분류 성능을 동시에 향상시킬 수 있는 적정 수준을 찾아내어 부류를 나누는 것이 중요하다.

5. 결론 및 향후 과제

본 연구에서는 HMM을 이용하여 생물학 문서로부터 개체명의 경계를 인식하였다. 부류의 수가 적을수록 상대적으로 자료 부족 문제에 덜 영향을 받아 부류의 개수를 많이 사용할 때보다 인식 성능이 좋았다. 그러나, 개체명 인식은 경계 인식뿐만 아니라 분류 작업이 동반되어야 한다. 따라서, 향후 연구로는 1차적으로 계층 정보를 이용해 대 부류를 사용하여 개체명들을 분류한 뒤, 2차적으로 사전이나 다른 기계학습 방법을 통해 이를 세 부류로 분류하고 그 결과가 경계 인식과 분류를 동시에 수행하는 다른 모델보다 좋은 성능을 보이는지에 대한 추가적인 연구가 뒤따라야 할 것이다.

참고문헌

[1] D. Bikel, S. Miller, R. Schwartz, and R. Weischedel. "NYMBLE: A High-Performance Learning Name-finder", In Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 194-201, 1997

[2] N. H. Collier, C. Nobata, and J. Tsujii, (2000), "Extracting the Names of Genes and Gene Products with a Hidden Markov Model", In Proceedings of COLING2000, pp.201-207, 2000.

[3] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii, (2002), "Tuning Support Vector Machines for Biomedical Named Entity Recognition", In Proceeding of the workshop on Natural Language Processing in the Biomedical Domain(at ACL

'2002), pp. 1-8

[4] T. M. Mitchell, Machine Learning, McGraw-Hill companies, Inc., 1997.

[5] K. Pastra, D. Maynard, H. Cunningham, O. Hamza, and Y. Wilks, "How feasible is the reuse of grammars for Named Entity Recognition?", In Proceedings of 3rd Language Resources and Evaluation Conference, 2002.

[6] E. Tjong Kim Sang, "Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition", In Proceedings of CoNLL-2002, 2002.

임 희 석(Heui-Seok Lim) **[종신회원]**

- 1992년 2월 : 고려대학교 컴퓨터학과 (이학사)
- 1994년 2월 : 고려대학교 컴퓨터학과 (이학석사)
- 1997년 9월 : 고려대학교 컴퓨터학과 (이학박사)
- 1999년 3월 ~ 현재 : 한신대학교 컴퓨터정보소프트웨어학부 부교수

<관심분야>
자연어처리, 인공지능, 인지신경계산학, 데이터마이닝