개체명 구성 원리를 이용한 교사학습 기반의 한국어 개체명 인식

황이규[0] 이현숙 정의석 윤보현 박상규
한국전자통신연구원 휴먼정보처리연구부
{yghwang, lhs63473, eschung, ybh, parksk}@etri.re.kr

Korean Named Entity Recognition Based on Supervised Learning Using Named Entity Construction Principles

Yi-Gyu Hwang[0] Hyun-Sook Lee EuiSok Chung Bo-Hyun Yun Sang-Kyu Park
Human Information Processing Department, ETRI

요 약

개체명 인식은 질의응답(QA), 정보 추출(IE), 텍스트 마이닝 시스템의 성능 향상에 중요한 역할을 담당한다. 이 논문에서는 교사학습 기반의 한국어 개체명 인식에 대해 설명한다. 한국어에서 많은 개체명들이 하나 이상의 단어로 구성되어 있으며, 개체명을 구성하는 단어 사이에는 의존 관계가 존재하고, 개체명과 개체명 주위의 단어 사이에도 문맥적 의존 관계를 가지고 있다. 본 논문에서는 가변길이의 개체명과 주변 문맥의 학습을 위해 트라이그램을 이용한 HMM을 사용하였으며, 자료 부족 문제를 해소하기 위해 어휘 기반이 아닌 부개체 유형 기반의 학습을 수행하였다. 학습된 개체명 인식 시스템을 이용하여 경제 분야의 신문 기사에 대한 실험 결과, 84.4%의 정확률과 90.9%의 재현률을 보였다.

1. 서론

대용량의 비정형 문서로부터 정보를 추출하는 방법에 대한 연구가 주목을 받으면서, 개체명 인식이 각광을 받고 있다. 개체명 인식은 정보추출을 위한 여러 단계 중 첫 단계로 영어 개체명 인식에 대한 연구 결과가 MUC(Message Understanding Conferences)7에서 보고 되었고[8], 중국어, 일본어 및 스페인어와 같은 비영어권의 개체명 인식 평가대회인 MET(Multilingual Entity Task)2[8]가 있었다. 그리고, 일본에서도 IREX(Information Retrieval and Extraction Exercise)를 통해 일본어에 대한 정보 검색 및 개체명 인식에 대한 평가대회가 있었다[6]. 또한, CoNLL(Computational Natural Language Learning)에서는 언어에 독립적인 개체명 인식에 대한 시스템의 평가를 위하여 스페인어와 독일어에 대한 개체명 인식 시스템을 평가하는 등, 개체명 인식에 대한 꾸준한 관심이 지속되고 있다[3].

개체명이란 문서에서 나타나는 고유한 의미를 가지는 명사나 숫자 표현과 같이 문서의 고유한 성질을 표현하는 개체를 말한다. 개체명은 인명(Person name), 지명(Location name), 기관명(Organization name) 등의 이름표현, 날짜나 시간과 같은 시간 표현, 금액이나 퍼센트와 같은 수치 표현으로 구분할 수 있다. 대부분, 하나 이상의 단어가 결합하여 개체명을 구성한다. 개체명은 일반적으로 아래와 같이 분류할 수 있다.

- 고유명사: 인명, 지명, 기관명
- 숫자 표현: 날짜, 시간, 금액, 비율, 전화번호, 수량

개체명의 인식이 어려운 이유는 첫째, 새로운 개체명이 꾸준히 만들어 지고 있기 때문에 사전에 모든 개체명을 수록할 수 없으며, 둘째, 개체명을 개체형에 따라 인식할 때, 문맥에 따라 다른 개체형으로 해석될 수 있는 개체형 모호성이 발생할 수 있다. 즉, 개체명을 구성하는 단어만으로는 개체명의 유형을 결정할 수 없고, 문맥 지식을 활용해야 하는 경우가 많이 발생한다. 예를 들어, "워싱턴 장군이 태어난 곳이 여기다", "워싱턴 지역은 현재 비가 오고 있다" 와 "워싱턴 당국은 그 사건에 대해 논평을 거부하였다" 의 세 문장에서 "워싱턴" 은 각각, 인명, 지명, 기관명으로 서로 다르게 사용되고 있다. 이러한 문제들을 고려하여 개체명을

인식하는 방법은 크게 두 가지로 나눌 수 있다.

첫째, 규칙에 기반한 방법으로, 개체명 인식을 위해 규칙을 수작업 또는 반자동으로 구축하고, 이를 이용하여 새로운 문서에 대해 개체명을 인식한다[4, 5, 7, 16, 17, 18]. 이 방법은 잘 정의된 고유명사 사전이나 접사 사전, 결합단어 사전 등을 이용한다. 학습 코퍼스를 만들고 이를 이용하여 자동으로 개체명 인식 패턴을 구축하거나, 수동으로 개체명 인식 패턴을 구축하고 이 패턴으로 개체명을 인식한다. 이때, 구문분석이나 격틀, 용언의 하위 범주화 정보, 담화 분석 같은 각종 언어 정보를 이용하기도 한다.

둘째, 통계에 기반한 방법으로, 학습 코퍼스로부터 개체명 인식에 필요한 지식을 학습한다[1, 10, 11, 13, 14]. 이때, HMM(Hidden Markov Model)이나 Maximum Entropy 모델, Decision Tree 모델 등을 이용하는데, 주로 문자형, 철자 정보, 품사, 어휘 정보와 같이 비교적 얻기 쉬운 지식을 이용한다.

한국어 개체명 인식에 대한 연구[12, 16, 17, 18]가 비교적 활발하게 진행되고 있는데, 한국어 개체명의 인식이 다른 언어에 비해 어려운 이유는 대소문자를 구분하는 영어나, 문자형에 대한 정보가 비교적 많은 일본어에 비해 한국어가 문자형에 대한 정보가 부족하기 때문이다.

규칙 기반의 개체명 인식은 소규모 분야에서는 높은 인식률을 보이지만 다른 분야로의 이식성이 낮다. 한 응용분야를 위해 작성된 규칙을 새로운 응용분야에 활용하기 위해서는 많은 양의 규칙을 새로 작성하여야 하는 경우가 빈번하다. 이로 인해 규칙 작성을 위한 비용이 증가한다. 이러한 문제점 때문에 통계기반의 개체명 인식 방법이 활발하게 연구되고 있다.

본 논문에서는 한국어 개체명의 특징을 살펴보고, 이를 바탕으로 한국어 개체명 구성 원리를 이용한 HMM 기반 교사학습 방법을 제안하고 이를 실험하였다.

2. 한국어 개체명의 특성

2.1 개체명의 구조적 특성

한국어 문장에서 발생하는 개체명의 특징과 빈도를 조사하기 위해 경제/공연/여행 관련 신문 기사 및 웹 문서 각각 100 문서를 수집하여 개체명을 부착한 말뭉치를 구축하였다. 각 분야별 개체형 분포는 표 1 과 같은데, 도메인의 특성에 따라 개체형의 비율이 조금씩 달랐다. 경제분야에서는 기관명, 날짜, 지명의 순으로, 공연분야에서는 인명, 날짜, 지명의 순으로 높은 빈도로 나타났으며, 여행분야에서는 지명, 수량, 날짜 등의 순서였다.

또한, 한국어나 일본어가 한자문화권의 단어 특성을 가지고 있어, 개체형이 하나 이상의 단어로 구성된 경우가 많다. 수집된 문서의 분석 결과, 65% 이상의 개체명이 하나 이상의 단어로 구성되어 있으며, 4개 이상의 단어로 구성된 개체명도 약 15% 정도를 차지하는데, 이들은 대부분 지명, 기관명, 시간 및 날짜 표현 등에서 발생하였다.

표 1 개체형 부착 말뭉치의 개체형 분포

개체형	빈도	비율 (경제)	빈도	비율 (공연)	빈도	비율 (여행)
인명	199	8.6%	878	28.8%	121	2.8%
지명	394	17.1%	472	15.5%	2634	61.2%
기관명	567	24.6%	343	11.3%	174	4.0%
날짜	466	20.2%	558	18.3%	224	5.2%
시간	17	0.7%	191	6.3%	68	1.6%
금액	128	5.5%	58	1.9%	165	3.8%
퍼센트	269	11.7%	5	0.2%	5	0.1%
수량	269	11.7%	448	14.7%	757	17.6%
전화번호	0	0%	95	3.1%	159	3.7%

표 2 말뭉치 내의 개체명 길이[1]

개체명 구성 형태소 수	비율
n = 1	33.1%
n = 2	40.0%
n = 3	12.5%
n = 4	7.4%
n = 5	4.4%

개체명이 하나 이상의 단어로 구성될 때, 개체명 주위에 발생하는 단어들과의 관계를 살펴보기 위해 개체명 주위의 품사 분포를 살펴 보았다. 개체명의 앞에서 명사가 약 32% 발생하였으며, 개체명이 문장의 시작인 경우가 15.6%, 기타 조사/어미/심볼 등이 52% 발생하였다. 또한, 개체명의 뒤에서 36% 정도의 명사가 발생하였고, 문장의 끝에서 나타난 개체명도 2.3%, 기타 조사/어미/심볼 등이 61.7% 발생하였다. 따라서, 개체명을 구성하는 단어와 개체명의 앞과 뒤에 발생하는 명사들 사이에 어휘 또는 의미 수준의 상관 관계를 학습한다면 개체명 인식에 도움을 줄 수 있을 것이다.

2.2 개체명 및 주변단어의 분류

표 2에서 살펴본 것처럼 한국어에서 많은 수의 개체명이 하나 이상의 단어로 구성된 복합 단어이기 때문에 개체명을 구성하는 단어들 사이의

[1] 이 자료는 개체명에 대해 가능한 모든 명사를 분해한 경우임. 예를 들어 "한국전자통신연구원"의 경우, "한국+전자+통신+연구+원"으로 구성됨.

상관 관계를 이해하는 것이 개체명의 인식에 도움이 될 수 있다. 개체명을 구성하는 단어와 이들 주위의 단어들이 개체명의 형성 및 인식에 어떠한 영향을 미치는지에 따라 다음과 같은 네 가지 범주로 분류하였다.

- 개체명 독립 단어(Independent Named Entity:IE)
  - 그 자체로 하나의 개체명이 될 수 있는 단어
  - "제품 불매운동과 관련하여 *삼성*은 전자제품에 대한", "*AT&T* 의 새로운 사업은", "명량해협에서 *이순신* 장군은"
- 개체명 구성 단어(Constituent Named Entity:CE)
  - 그 자체로 개체명이 될 수 없지만 다른 단어와 결합하여 개체명이 될 수 있는 단어
  - "대덕*전자*의 이번 결정은", "SK *텔레콤*측의 일정이", "이순신 *기념관* 건립과 관련된", "해운대 *해수욕장*으로 가자", "통영 *여객선 터미널*에서 만나자."
- 개체명 인접 단어(Adjacent Named Entity:AE)
  - 개체명을 구성하지는 않지만 개체명의 주위에 나타나 개체명 인식에 도움을 주는 인접 단어
  - "*대표이사* 강은수*씨*는", "이순신 *장군*이 유배를", "동대문 *지역* 상인들은"
- 개체명과 관련이 없는 단어(Not Entity)
  - 의존명사, 조사, 동사, 어미 등 …

위에서 분류한 범주에 따라 개체명 태깅된 문서로부터 개체명을 구성하는 단어들 사이의 긴밀성에 따른 내부 구성 요소들 사이의 관계를 학습하였다. 이는 2.1에서 살펴본 개체명 외부 단어들과 개체명 사이의 관계와 더불어서 개체명의 인식에 중요한 정보로 활용될 수 있다. 이 논문에서는 개체명 내부 구성 단어들의 의존 관계와 개체명과 개체명 인접 단어들 사이의 의미적 공기 관계가 매우 높다는 분석 결과를 바탕으로 이를 활용한 HMM 기반의 개체명 인식 모델을 제안한다.

3. 한국어 개체명 인식

본 논문에서는 개체명 인식을 위해 교사학습 모델로 HMM을 이용하였다. HMM은 시간에 따라 변화되는 입력열의 다양한 변형을 표현할 수 있는 확률 모델로 음성인식, 품사 태깅과 같은 응용 분야에서 널리 사용되고 있다[9]. 본 논문은 [10]과 [13]의 연구 결과를 기본으로 하여 한국어 개체명의 구조적 특성에 맞게 모델을 수정하였다.

3.1 HMM의 적용

HMM은 모델의 상태수 N, 각 상태에서 일어날 수 있는 서로 다른 관측 심볼의 수 M, 상태전이 확률 A, 특정 상태에서의 관측 확률 B, 초기 상태 확률 π로 구성된다. 상태전이 확률을 계산하기 위해 일반적으로 아래와 같은 바이그램 확률을 이용한다. 어떤 단어 w_{n-1} 다음에 단어 w_n이 발생할 확률을 계산하는 식이다.

$$P(w_n \mid w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)} = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

일반적으로 HMM을 이용한 태깅에서 사용하는 태깅 모델은 다음과 같다. T는 주어진 문장 W의 가장 적합한 태그열을 나타내고, w_i는 단어, t_i는 품사를 나타낸다.

$$P(T \mid W) = \arg\max_T \prod_i P(t_i \mid t_{i-1}...t_1) p(w_i \mid t_i...t_1)$$
$$\approx \arg\max_T \prod_i P(t_i \mid t_{i-1}) p(w_i \mid t_i)$$

개체명 인식은 주어진 문장 W에 가장 적합한 개체형 열인 T를 찾는 과정으로 생각할 수 있다. 본 논문에서는 부개체형[2] 단위의 트라이그램 기반 학습 방법을 이용하여 최적의 개체형 열을 찾는다. 즉, 품사 태깅에서 품사에 대응되는 단위로 각 단어가 가질 수 있는 부개체형에 기반하여 트라이그램을 추출한다. 개체명의 앞과 뒤에 나타나는 문맥에 대한 정보를 활용하고, 다양한 문맥 확률을 반영하기 위해 트라이그램을 사용하였다. 개체명이 하나 이상의 단어로 구성되는 가변길이이므로 n개의 단어로 구성된 개체명 $w_1^{NE}...w_n^{NE}$, 개체명의 좌측 문맥 $w_{-2}^L w_{-1}^L$과 우측 문맥 $w_1^R w_2^R$ 등을 고려해야 한다[10].

(좌문맥)	*(개체명)*	*(우문맥)*
$\Lambda\ w_{-2}^L w_{-1}^L$	$w_1^{NE}...w_n^{NE}$	$w_1^R w_{21}^R \Lambda$

그림 1. 가변길이 개체명을 위한 트라이그램

따라서, 가변길이 개체명을 위한 트라이그램

[2] 논문에서 부개체형은 개체명 독립, 구성, 인접 단어 유형을 포함한다.

모델에서 학습되는 문맥은 아래와 같다.

$$w_{-2}^{L} w_{-1}^{L} w_{1}^{NE}$$
$$w_{-1}^{L} w_{1}^{NE} w_{2}^{NE}$$
$$\Lambda$$
$$w_{n-1}^{NE} w_{n}^{NE} w_{1}^{R}$$
$$w_{n}^{NE} w_{1}^{R} w_{2}^{R}$$

그림 2. 개체명 인식을 위한 학습 문맥

실제, 개체명의 인식 과정은 크게 각 단어가 가질 수 있는 부개체 유형을 찾는 과정과 각 단어들이 결합하여 하나의 개체명을 이룰 때, 개체명의 경계를 인식하는 두 단계로 나눌 수 있으며, 각 단계에서 학습되는 문맥과 정보가 다르다. 또한, 어휘 기반의 트라이그램을 사용할 경우, 자료 부족 문제가 발생하기 때문에, 본 논문에서는 부개체형 및 품사 기반의 트라이그램을 학습하고 이를 이용하여 개체명을 인식하도록 하였다.

부개체형 기반의 학습을 위해 set 함수를 정의하였는데, 각 단어 w_i는 함수 $set(w_i)$에 의해 상태 집합 S로 사상된다. 함수 $set(w_i)$는 단어를 입력으로 하여, 단어가 가질 수 있는 부개체 유형 또는 품사를 반환한다. 예를 들어, "연구원"은 기관명 구성단어(CE_Org) 또는 인명 인접단어(AE_Per)의 두가지 상태를 반환할 수 있다. 만일 단어가 부개체 유형에 포함되지 않으면, 품사를 반환한다.

부개체형 부착을 위한 HMM의 상태전이확률에서 상태 집합 S는 부개체 유형과 품사로 구성된다. 예를 들어, 인명, 지명, 기관명을 인식하고, 8개의 품사를 가지는 모델의 경우에 상태 집합 S는 문장 시작 기호(SOS)와 문장 끝 기호(EOS)를 포함하여 다음과 같다.

S = {IE_Per, CE_Per, AE_Per, IE_Loc, CE_Loc, AE_Loc, IE_Org, CE_Org, AE_Org, POS_1, ···, POS_8, SOS, EOS}

이를 이용한 부개형 부착은 다음과 같다.

$$P(T|W) = \arg\max_{T} \prod_{i=1}^{n} P(set(w_i)|set(w_{i-1})...set(w_1))p(w_i|set(w_i)...set(w_1))$$
$$\approx \arg\max_{T} \prod_{i=1}^{n} P(set(w_i)|set(w_{i-1})set(w_{i-2}))p(w_i|set(w_i))$$

또한, 개체명 경계 인식을 위한 HMM의 상태는 인명, 지명, 기관명과 "개체명_시작(S)", "개체명_계속(C)", "개체명_끝(E)", "개체명_단일(U)"과 같은 개체명 구성 단어의 내부적 순서 관계 정보를 학습하여 경계를 인식하는데 사용하였다.[3] 개체명 경계를 위한 함수 $brf(set(w))$를 이용한 학습 결과는 다음과 같은 상태로 사상된다.

S' = {Per_S, Per_C, Per_E, Per_U, Loc_S, Loc_C, Loc_E, Loc_U, Org_S, Org_C, Org_E, Org_U, Not_NE, SOS, EOS}

트라이그램을 이용한 개체명 경계 인식은 다음과 같다.

$$P(T|W) \approx \arg\max_{T} \prod_{i=1}^{n} P(brf(set(w_i))|set(w_{i-1})set(w_{i-2}))p(w_i|brf(set(w_i)))$$

3.2 개체명 인식 모델

개체명 인식 과정은 개체명을 구성하는 각 단어들을 개체형으로 분류하는 부개체형 분류 단계와 분류된 개체형들을 결합하여 개체명의 경계를 인식하는 개체명 인식 단계로 나눌 수 있다. 이때, 각 단계에서 모호성이 발생할 수 있다. 예를 들어, "LG경제연구원은 현대상사가"라는 문장의 부개체형 분류 단계와 개체명 경계인식 과정을 살펴 보면 다음과 같다(논문에서는 개체형 중 숫자 표현을 제외하고, 인명, 지명, 기관명만을 대상으로 학습하고 실험하였다).

표 3 개체명 인식 과정의 예

부개체형 부착	부개체형 인식	개체명 경계인식
LG:IE_Org	LG:IE_Org	LG:Org_S
경제:CE_Org	경제:CE_Org	경제:Org_C
연구:AE_Per,CE_Org	연구:CE_Org	연구:Org_C
원:CE_Loc,CE_Org	원:CE_Org	원:Org_E
은:jx	은:jx	은:Not_NE
현대:IE_Org	현대:IE_Org	현대:Org_S
상사:AE_Per ,CE_Org	상사:CE_Org	상사:Org_E
가:jc	가:jc	가:Not_NE

개체명 인식의 두 과정에서 각각 모호성이 발생하고, 이를 해결하면서 개체명을 인식하기 위해 서로 다른 통계적 확률을 적용하여 HMM을 구성하였다.

이때, 학습 자료의 부족으로 인한 문제가 발생하며, 부개체형 부착 단계와 개체명 경계 인식 단계 각각에 다른 Back-off 모델을 사용하였다. 부개체형 인식을 위한 상태전이확률에서는 바이그램을 이용하였으며, 개체명 경계 인식을 위한 단계에서는 set(w)의 결과를 다음과 같은 3단계를 거치도록 하였다. 따라서, 세분류한 품사에 따른 문맥이 없을 경우, 대분류한 품사에 기반한 문맥을 활용할 수 있도록 함으로써, 학습

[3] Per_S, Loc_C, 또는 Org_U 등이다.

부족 문제를 해결하였다. 또한, 세가지 각 상태로도 실험을 하여 보았다.

- E0: 3(IE,CE,AE) * 3(P,L,O) + 28 품사 + SOS + EOS = 39 상태
- E1: 3(IE,CE,AE) * 3(P,L,O) + 8 품사 + SOS + EOS = 19 상태
- E2 : 3(IE, CE, AE) * 3(PLO) + 1(Not_SE) + SOS + EOS = 12 상태

4. 실험 및 평가

한국어 개체명 인식을 위해 본 논문에서 제안하는 시스템의 구성은 다음과 같다. 학습 부분은 부개체 유형 부착을 위한 학습과 부개체 유형 경계 부착을 위한 학습이 동시에 진행되며, 개체명을 인식하는 부분은 부개체 유형을 분류하는 부분과 인식된 부개체 유형의 경계를 탐지하는 경계인식 부분으로 나누어진다.

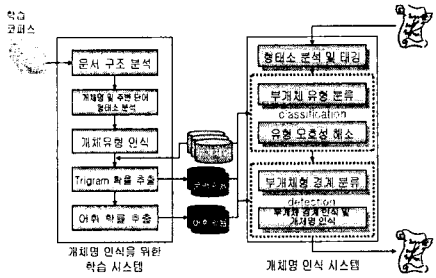

그림 3. 한국어 개체명 인식 시스템

실험을 위해 경제 분야 신문 기사 100문서에서 학습을 위해 90문서를 이용하였고, 10문서를 대상으로 실험하였다. 학습을 위해 사용한 경제 분야의 신문 기사는 문서당 평균 7 문장으로 구성되었으며, 한 문장당 평균 18.7개의 어절을 가지고 있다. 학습을 위한 문서의 일부는 그림 4와 같다.

<S>또 <ORG>재경부</ORG>와 <ORG>금감위</ORG>는 증권거래법 개정을 둘러싸고 다툼을 벌이고 있다.</S>
<S>현행 증권거래법에 따르면 <ORG>증권거래소</ORG>와 <ORG>코스닥</ORG>의 공시·상장·매매에 관한 규정은 <ORG>금융감독위원회</ORG>가 <ORG>재경부</ORG>의 사전협의를 거쳐 승인을 하도록 하고 있다.</S>
<S>그러나 <PERSON>변양호</PERSON> <ORG>재경부 금융정책국</ORG>장은 "신속하게 집행되어야 하는 시장조치가 <ORG>금감위</ORG> 승인절차 때문에 지체되는 부작용이 발생해 승인절차를 <ORG>재경부</ORG>로 일원화하기로 했다"고 발표했다.</S>
...
<S><ORG>대외경제정책연구원</ORG> <PERSON>오동윤</PERSON> (<PERSON>吳東潤</PERSON>) 연구원은"<LOC>대만 </LOC>은 여소야대 정국하의 금융구조조정 등 개혁 실패와 주식시장 붕괴, 내수·수출 침체, 급속한 자본이탈 등 우리나라를 비롯한 <LOC>아시아</LOC>권 전반의 위기 상황을 총체적으로 보여주고 있어 주목된다"며 내달 예정된 입법원선거(총선) 결과가 <LOC>대만</LOC> 경제의 중대 변수가 될 것"이라고 말했다.</S>

그림 4. 학습 문서의 예

학습 문서를 대상으로 다양한 HMM 상태 유형에 따라 부개체형 인식 및 경계 인식을 위한 트라이그램을 추출하였다. 그림 5와 그림 6은 학습 문서를 10문서씩 90문서까지 증가하면서 추출된 트라이그램의 수를 보여주고 있다.

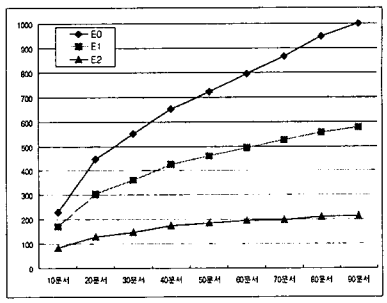

그림 5. 학습된 부개체명 인식 트라이그램

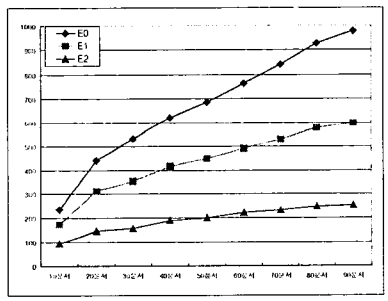

그림 6. 학습된 개체명 경계 인식 트라이그램

그림 5와 그림 6을 통해 개체명 인식을 위한 학습 문서의 양에 따른 트라이그램 수의 수렴 정도를 알 수 있다. 개체명 인식에 사용된 부개체형 사전의 양은 표 4와 같다. 부개체명 사전 중 개체명 독립단어 사전은 고유명사 사전을 기본으로 하여 수동으로 구축하였으며, 학습 문서를 통해 자동으로 보강하였다. 다어절 개체명의 경우, 띄어쓰기 정보를 유지하지 않았다. 따라서, 인식할 대상이 띄어쓰기가 되어있는 다어절 개체명의 경우, 단어의 구성 문맥에 따라 인식되도록 하였다.

표 4 부개체명 사전의 구성

구분	개체명 독립 단어 사전	개체명 구성 단어 사전	개체명 인접 단어 사전
단어수	102,244	241	185

세 가지 학습 유형에 따라 10문서에 대하여

개체명 인식률을 실험하고, E0 상태를 기본으로 자료 부족시에 E1, E2의 순서로 Back-off한 결과(E0*)는 표 5와 같다. 표 5는 부개체형 인식된 각 단어에 대해 개체명의 시작과 끝을 인식하여 개체명을 찾아내는 개체명 인식률을 보여주고 있다.

표 5 개체명 인식 실험

실험 유형	재현률	정확률
E0	71.2%	90.4%
E1	77.3%	87.9%
E2	83.3%	88.7%
E0*	90.9%	84.5%

E0, E1, E2 실험에서 상태수가 많을 때 재현률이 낮은 이유는 여러 단어로 개체명이 구성될 경우, 자료 부족 문제 때문에 발생하며, 비교적 정확률이 높은 이유는, 정확한 문맥을 만족하는 개체명만 찾기 때문이다. 또한, Back-off (E0*)시에 정확률이 떨어지는 이유는, "의존도"나 "외환 시장", "경제 관료 시스템"과 같은 단어를 개체명으로 인식하기 때문이다. 이러한 문제는 상호 배제 어휘 정보와 같은 어휘 레벨의 정보를 통해 해결할 수 있을 것이다. 본 모델이 개선해야 할 점으로는 개체명을 구성하는 단어 또는 개체명 주위에 개체명을 인식할 만한 단서가 없으면, 개체명을 인식하지 못한다는 점이다. 이를 극복하기 위해서는 미지어 추정에 따른 개체명 인식 방법이 보강되어야 할 것이다.

5. 결론 및 향후 연구 방향

본 논문에서는 개체명 인식을 위한 국내외 연구를 간단히 살펴보고, 한국어 문서에서 나타나는 개체형의 유형과 특성, 개체명을 구성하는 한국어의 구조적 특징을 조사하고, 한국어에 적합한 개체명 인식을 위한 방법을 제안하였다. 한국어는 일본어나 영어에 비해, 문자 자체가 가지는 타입 정보가 많이 부족하기 때문에 개체명 사전이나 개체명 구성단어 및 인접 단어 사전의 중요성이 무척 크다. 따라서 사전과 단어의 부개체 유형에 기반해서 개체명의 구성 원리를 이용한 개체명 인식 방법을 선택하였다.

또한, 학습 코퍼스로부터 개체명 인식에 필요한 통계 정보를 자동학습하고, 이를 이용하여 HMM 기반 개체명 인식 시스템을 구축하였다. 한국어 개체형은 개체형 주변 문맥에 의존하며, 개체명 자체의 구성 문맥에도 의존적인 경향이 많으므로, 이들을 자동으로 학습하고 이를 개체명 인식 과정에 반영하였다.

향후 연구 과제로 적극적으로 어휘 수준의 정보를 반영하고, 부개체형 인식 후 개체형을 인식하는 두 단계 접근 방법을 통합하여 하나의 학습 모델에 수용함으로써 불필요한 정보량을 줄이면서 개체명 인식 속도를 향상시키는 것이다.

6. 참고문헌

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

[2] M. Collins and Y. Singer, "Unsupervised Models for Named Entity Classification," EMNLP/VLC-99, pp. 189-196, 1999.

[3] CoNLL, http://cnts.uia.ac.be/conll2002/ner/

[4] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi, "Toward Information Extraction: Identifying protein names from biological papers," In Proc. of the Pacific Symposium on Biocomputing '98 (PSB'98), 1998.

[5] J. Fukumoto, M. Shimohata, F. Masui, and M. Saski, "Description of the Oki System as Used for MET-2," In Proceedings of 7th Message Understanding Conference, 1998.

[6] IREX, http://www.cs.nyu.edu/cs/projects/proteus/irex/index-e.html

[7] A. Mikheev, C. Grover, M. Moens, "Description of the LTG System Used for MUC-7," In Proceedings of 7th Message Understanding Conference, 1998.

[8] MUC7, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

[9] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, Vol. 77, No. 2, pp. 257-286, 1989.

[10] M. Sassano and T. Utsuro, "Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition," Proceedings of the 18th International Conference on Computational Linguistics, pp. 705-711, 2000.

[11] S. Sekine, R. Grishman and H. Shinnou, "A Decision Tree Method for Finding And Classifying Names in Japanese Texts," Proceedings of the Sixth Workshop on Very Large Corpora, 1998.

[12] C. N. Seon, Y. Ko, J. S. Kim, and J. Seo, "Named Entity Recognition using Machine Learning Methods and Pattern-Selection Rules," pp. 229-236, NLPRS 2001.

[13] K. Uchimoto, Q. Ma, M. Murata, H. Ozakum,

and H. Isahara, "Named Entity Extraction Based on A ME Model and Transformation Rules," In Processing of the ACL 2000.
[14] S. Yu, S. Bai, and P. Wu, "Description of the Kent Ridge Digital Labs System Used for MUC-7," In Proceedings of 7th Message Understanding Conference, 1998.
[15] G. D. Zhou, J. Su, "Named Entity Recognition using an HMM-based Chunk Tagger," In Processing of the ACL 2002.
[16] 김태현, 이현숙, 하유선, 이만호, 맹성현, "데이터 집합을 이용한 고유명사 추출", 제 12회 한글 및 한국어 정보처리 학술대회, pp. 11-18, 2000.
[17] 노태길, 이상조, "규칙 기반의 기계학습을 통한 고유 명사의 추출과 분류", 한국정보과학회 가을 학술발표 논문집, Vol.27, No.2, pp. 170-172, 2000.
[18] 이경희, 이주호, 최명석, 김길창, "한국어 문서에서 개체명 인식에 관한 연구", 한글 및 한국어 정보처리 학술대회, pp. 292-299, 2000.