HMM에 기반한 한국어 개체명 인식

황 이 규[†] · 윤 보 현[††]

요 약

개체명 인식은 질의응답 시스템이나 정보 추출 시스템에서 필수 불가결한 과정이다. 이 논문에서는 HMM 기반의 복합 명사 구성 원리를 이용한 한국어 개체명 인식 방법에 대해 설명한다. 한국어에서 많은 개체명들이 하나 이상의 단어로 구성되어 있다. 또한, 하나의 개체명을 구성하는 단어들 사이와 개체명과 개체명 주위의 단어 사이에도 문맥적 관계를 가지고 있다. 본 논문에서는 단어들을 개체명 독립 단어, 개체명 구성 단어, 개체명 인접 단어로 분류하고, 개체명 관련 단어 유형과 품사를 기반으로 HMM을 학습하였다. 본 논문에서 제안하는 개체명 인식 시스템은 가변길이의 개체명을 인식하기 위해 트라이그램 모델을 사용하였다. 트라이그램 모델을 이용한 HMM은 데이터 부족 문제를 가지고 있으며, 이를 해결하기 위해 다단계 백-오프를 이용하였다. 경제 분야 신문기사를 이용한 실험 결과 F-measure 87.6%의 결과를 얻었다.

HMM-based Korean Named Entity Recognition

Yi-Gyu Hwang[†] · Bo-Hyun Yun[††]

ABSTRACT

Named entity recognition is the process indispensable to question answering and information extraction systems. This paper presents an HMM based named entity (NE) recognition method using the construction principles of compound words. In Korean, many named entities can be decomposed into more than one word. Moreover, there are contextual relationships among nouns in an NE, and among an NE and its surrounding words. In this paper, we classify words into a word as an NE in itself, a word in an NE, and/or a word adjacent to an NE, and train an HMM based on NE-related word types and parts of speech. Proposed named entity recognition (NER) system uses trigram model of HMM for considering variable length of NEs. However, the trigram model of HMM has a serious data sparseness problem. In order to solve the problem, we use multi-level back-offs. Experimental results show that our NER system can achieve an F-measure of 87.6% in the economic articles.

키워드 : 개체명 인식(Named Entity Recognition), 정보 추출(Information Extraction), HMM, 트라이그램(Trigram), 가변 길이 개체명(Variable Length Named Entity)

1. 서 론

대용량의 비구조화(unstructured) 또는 반구조화(semistructured)된 문서로부터 정보를 추출하는 방법에 대한 연구가 주목을 받으면서, 개체명 인식이 새로운 연구 분야로 각광을 받고 있다. 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.

개체명 인식이 어려운 이유는 사전에 모든 개체명을 수록할 수 없으며, 새로운 개체명이 꾸준히 만들어지고 있기 때문이다. 또한 개체명을 인식할 때, 문맥에 따라 다른 개체형으로 해석될 수 있다. 즉, 개체명을 구성하는 단어만으로는 개체명의 유형을 결정할 수 없는 경우가 많다. 예를 들어, "**워싱턴** 장군이 태어난 곳이 여기다", "**워싱턴** 지역은 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 수작업으로 구축하고, 이를 이용하여 새로운 문서에 대해 개체명을 인식한다[3-5, 13-15]. 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, 격틀, 용언의

† 정 회 원 : 한국전자통신연구원 선임연구원
†† 정 회 원 : 목원대학교 컴퓨터교육과 교수
논문접수 : 2002년 7월 29일, 심사완료 : 2003년 1월 29일

하위 범주화 정보나 구문분석, 담화 분석 같은 각종 언어 정보 및 언어 분석 도구를 이용하기도 한다.

둘째, 통계에 기반한 방법으로, 학습 코퍼스로부터 개체명 인식에 필요한 지식을 학습한다[1, 7, 8, 10-12]. 은닉 마르코프 모델(hidden Markov model)이나 최대 엔트로피 모델(maximum entropy model), 결정트리(decision tree) 모델 등을 이용하는데, 문자형, 철자 정보, 품사, 어휘 정보와 같이 비교적 얻기 쉬운 지식을 이용한다.

한국어의 경우, 개체명 인식에 대한 연구[9, 13-15]가 비교적 활발하게 진행되고 있는데, 한국어 개체명의 인식이 다른 언어에 비해 어려운 이유는 대소문자를 구분하는 영어나, 문자형에 대한 정보가 비교적 많은 일본어에 비해 한국어는 문자형에 대한 정보가 부족하기 때문이다. 한국어에서는 개체명 인식을 위한 정보를 어휘 그 자체로부터 얻기가 어렵다.

본 논문에서는 이러한 점들을 고려하여 개체명 인식을 위한 국내외의 연구를 살펴보고, 한국어 개체명의 특징을 살펴본다. 한국어 개체명의 구조적 특성을 바탕으로 개체명 및 인접 단어들 사이의 결합 원리를 이용한 HMM 기반 교사학습 방법을 제안하고 이를 평가하였다.

2. 관련 연구 동향

기존의 개체명 인식 관련 연구는 크게 두 가지 방법으로 분류할 수 있다. 첫째는 수동으로 패턴을 구축하고 이를 이용하여 개체명을 인식하는 것이고, 둘째는 통계 모델에 기반한 학습을 이용하여 인식하는 것이다.

2.1 규칙 기반 개체명 인식

규칙 기반의 접근은 정교한 규칙을 수동으로 작성하거나, 수동으로 작성된 규칙을 학습 코퍼스를 이용하여 수정하는 방법이다. 이때, 개체명 인식 시스템은 고유명사 사전이나 접사 사전, 결합명사 사전과 같은 다양한 사전을 이용한다. 규칙 기반의 방법은 적용할 분야의 특성을 활용한 방법[3], 문장에 자주 발생하는 문맥을 활용한 방법[13, 14], 접사 사전과 결합 규칙을 이용한 방법[4], 규칙과 문맥을 다단계로 적용한 방법[15] 등으로 나눌 수 있다. 이중 몇 가지를 살펴 보면 다음과 같다.

[14]에서는 고유명사가 태깅된 코퍼스로부터 실마리어와 문맥 실마리를 학습을 통하여 구축하고, 이를 이용하여 문서에서 회사명을 추출하였다. 이 시스템은 "산업", "투자신탁"과 같은 실마리어가 미등록, 일반명사 또는 고유명사와 결합하여 고유명사를 이룰 수 있는지 학습한다. 이를 통해 "서울산업"이나 "미래투자신탁"과 같은 고유명사를 인식할 수 있다. 또한 'A인 B'와 같은 문맥 실마리를 이용하여 A에 어떠한 유형의 명사가 있을 때 B가 고유명사로 인식이 가능한지 학습하였다. 그러나 이 연구에서 구축한 실마리어나 문맥 실마리는 경제 분야에 종속적이다.

[15]에서는 네 가지 단계를 거치면서 주어진 문서에서 개체명을 추출하였다. 첫 번째 단계는 단어 자체의 구성원리에 대한 규칙을 활용하며, 두 번째 단계에서는 인식 대상 개체명의 주변 문맥정보를 활용할 수 있는 규칙을 이용한다. 세 번째 단계에서는 용언의 하위범주화 정보를 이용하며, 마지막으로 개체명간 관계 정보를 이용하여 개체명을 통합하고 이를 하나의 개체명으로 인식한다. 적용된 규칙의 예를 들면, 고유명사로 단어의 길이가 3이고, 첫 글자가 성씨 사전에 있으면 인명 자질로 판단한다. 이때, 각 패턴의 자질 중요도에 따라 가중치를 할당할 수 있다. 이 연구는 가중치 할당이 경험적 방법에 기반하고 있어 일관성이 결여되어 있고, 각 단계별로 적용한 규칙 작성을 수작업으로 작성해야 하는 어려움이 있다.

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 한다. 이로 인해 규칙 작성을 위한 비용이 증가하며, 이를 극복하기 위해 통계기반의 개체명 인식 방법이 활발하게 연구되고 있다.

2.2 통계 기반 개체명 인식

통계 기반의 접근은 학습 코퍼스의 유무에 따라 교사학습기반[1, 7, 8, 10, 11]와 비교사 학습기반[2]으로 나눌 수 있으며, 학습 방법에 따라 HMM에 기반한 방법[1, 11], 결정트리 기반[7, 8], 최대 엔트로피 모델에 기반한 방법[2, 10] 등이 있다. 주요한 몇 가지 연구를 살펴보면 다음과 같다.

[1]에서는 HMM을 이용하여 영어 문장에서 개체명을 추출하는 방법을 기술하였다. 8 가지 내부 상태(7 가지 유형의 개체명을 위한 상태와 '개체명_아님' 상태), 두 가지 특별한 상태(문장의 시작, 문장의 끝)를 HMM을 구성하기 위한 상태로 가정하고, 단어의 특성을 이용하여 코퍼스로부터 개체명 인식 모델을 학습하였다. 단어의 의미 속성을 배제하고 '두자리 숫자', '모두 대문자', '모두 소문자', '첫 문자만 대문자' 등과 같이 문자나 단어의 철자 속성을 학습에 이용하였다. 철자 속성이 뚜렷한 영어와 같은 언어에서 유용한 방법으로, 한국어의 경우 철자 속성이 뚜렷하게 존재하지 않기 때문에 이를 한국어 개체명 인식에 직접 적용하기 어렵다.

[7]은 트라이그램 모델과 이를 확장한 가변길이 모델을 결정 리스트에 적용하여 일본어 개체명을 인식하였다. 트라이그램 모델이나 가변길이 모델에서 어휘정보, 품사정보, 철자 속성, 개체명 내의 순서 정보 등을 이용하였다. 이는 복합 명사의 단위화와 관련된 연구를 활용하여 다양한 단어로 구성된 개체명을 인식하기 위한 방법으로 도입하였다.

[10]은 많은 개체명이 하나 이상의 단어로 구성되어 있다는 가정에서 단어의 다양한 유형을 바탕으로 학습하였다.

개체명의 경계를 인식하기 위해 개체명을 구성하는 각 단어에 '시작', '계속', '끝', '유일' 표현과 '이전(PRE)', '이후(POST)', '중간(MID)', '이외(OTHER)' 태그 등을 활용하고, 이들 사이의 결합 규칙을 작성하였다. 예를 들어, "여성을 중심으로 '인권을 생각하는 모임'이 열렸다"라는 문장에서, 기관명인 '인권을 생각하는 모임'을 인식하기 위해서 아래와 같은 내부 표현을 사용하였다.

여성 을 중심 으로 ' 인권 을 생각하 는 모임 ' 이
이외 이외 이외 이외 이전 시작 중간 중간 중간 끝 이후 이외

영어의 경우, 통계 기반의 교사 학습이 대부분을 차지하며, 최근에는 비교사 학습에 기반한 연구가 활발히 진행되고 있다. 동양권 언어인 한국어, 일본어 및 중국어는 초기의 규칙 기반의 개체명 인식 방법에서 규칙 기반 접근 방법이 가지는 한계를 극복하기 위해 통계 기반의 학습 방법이 활발히 연구되고 있다.

도메인과 학습 문서의 양 및 평가 문서의 양이 다르기 때문에 직접적인 비교는 어렵지만, 국내외의 개체명 인식 실험 결과를 비교해 보면 <표 1>과 같다.

〈표 1〉 국내외 연구 결과

구 분	영어[5]	일본어[8]	한국어[9]	한국어[15]
학습 도메인	항공 사고 문서	사건/사고 문서	신문기사	신문기사
학습 문서수	100 문서	103 문서	2,555 문장	1,388 문장
평가 문서	100 문서	11 문서	422 문장	190 문장
F-Measure[1)]	93.4	85	84.1	86.8

[5]와 [15]는 규칙 기반의 접근으로, 비교적 상세한 규칙을 수작업으로 작성한 연구이다. [8]은 결정트리 모델에 비교적 자세한 개체명 사전을 이용하였고, [9]는 통계적 방법과 규칙을 결합한 모델로 방대한 고유명사 사전을 이용하였다.

3. 한국어 개체명 인식 모델

본 논문에서는 한국어 개체명의 특징을 먼저 살펴 보고, 한국어 개체명의 구조적 특성에 바탕을 둔 개체명 인식 방법을 제안한다.

3.1 한국어 개체명의 구조적 특성

한국어 문장에서 발생하는 개체명의 특징과 빈도를 조사하기 위해 경제/공연/여행 관련 신문 기사 및 웹 문서에 대해 각각 100문서를 수집하여 분석하였다. 각 분야별 개체형 분포는 <표 2>와 같으며, 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.

〈표 2〉 학습 코퍼스의 개체형 분포

개 체 형	빈 도	비율 (경제)	빈 도	비율 (공연)	빈 도	비율 (여행)
인 명	199	8.6%	878	28.8%	121	2.8%
지 명	394	17.1%	472	15.5%	2634	61.2%
기 관 명	567	24.6%	343	11.3%	174	4.0%
날 짜	466	20.2%	558	18.3%	224	5.2%
시 간	17	0.7%	191	6.3%	68	1.6%
금 액	128	5.5%	58	1.9%	165	3.8%
퍼 센 트	269	11.7%	5	0.2%	5	0.1%
수 량	269	11.7%	448	14.7%	757	17.6%
전화번호	0	0.0%	95	3.1%	159	3.7%

또한, 한국어나 일본어가 한자문화권의 단어 특성을 가지고 있어, 개체형이 하나 이상의 단어로 구성된 경우가 많은데, 이는 복합명사가 개체명으로 많이 이용되기 때문이다. 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 3〉 수집 문서의 개체명 길이

개체명 구성 형태소 수	비 율
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 살펴보았다. 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 단어 또는 개체형에 따른 의미 수준의 단어 분류 정보를 학습한다면 개체명 인식에 도움을 줄 수 있을 것이다.

3.2 개체명 형성 단어의 분류

<표 3>에서 살펴본 것처럼 한국어에서 많은 수의 개체명이 하나 이상의 단어로 구성된 복합 단어이기 때문에 개

1) $F-Measure = \frac{2 * Precision * Recall}{Precision + Recall}$

체명을 구성하는 단어들 사이의 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 단어를 분류하였다.

- **개체명 독립 단어(W_I)** : 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(W_C)** : 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(W_A)** : 개체명을 구성하지는 않지만 개체명의 주위에 나타나 개체명 인식에 도움을 주는 인접 단어
  - "**대표이사** 이건희 **회장**은", "이순신 **장군**이 유배를", "동대문 **지역** 상인들은"
- **개체명과 관련이 없는 단어**
  - 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 3.1에서 살펴본 개체명 외부 단어들과 개체명 사이의 관계와 더불어서 개체명의 인식에 중요한 정보로 활용될 수 있다. 이 논문에서는 개체명 내부 구성 단어들의 구조적 연관성과 개체명과 개체명 외부 단어들 사이의 의미적 공기 관계가 매우 높다는 분석 결과를 바탕으로 이를 활용한 HMM 기반의 개체명 인식 모델을 제안한다.

3.3 HMM을 이용한 개체명 인식

본 논문에서는 개체명 인식을 위해 학습 모델로 HMM을 이용하였다. HMM은 시간에 따라 변화되는 입력열의 다양한 변형을 표현할 수 있는 확률 모델로 음성인식, 품사 태깅과 같은 응용 분야에서 널리 사용되고 있다[6]. 본 논문은 [7]과 [10]의 연구 결과를 기본으로 하여 개체명의 구조적 특성에 맞게 모델을 수정하였다. [7]과 [10]에서는 개체명을 구성할 수 있는 단어와 그렇지 않은 단어, 개체명의 시작과 계속, 끝 등을 학습한 반면, 제안하는 개체명 인식 시스템은 하나의 개체명 내의 여러 단어를 개체명 독립 단어와 개체명 구성 단어로 구분하고, 개체명 주변의 단어도 단순히 "개체명과 관련 없음" 뿐만 아니라, 개체명 인접 단어를 정의하고 학습함으로써 개체명을 형성하는 문맥의 원리를 충실히 반영하도록 하였다.

3.3.1 HMM의 적용

품사 태깅에서 T를 주어진 문장 W의 가장 적합한 태그열로 가정하고, w_i는 문장 W를 구성하는 i번째 단어, t_i는 문장의 품사열 T에서 i번째 단어인 w_i의 품사를 나타낸다고 할 때, 최적의 품사열을 찾는 HMM 기반의 확률 모델은 다음과 같다.

$$\begin{aligned} P(T|W) &= \arg\max_T \prod_i P(t_i|t_{i-1}\cdots t_1)P(w_i|t_i\cdots t_1) \\ &\approx \arg\max_T \prod_i P(t_i|t_{i-1})P(w_i|t_i) \end{aligned} \quad (1)$$

이때, $P(t_i|t_{i-1})$은 상태전이 확률, $P(w_i|t_i)$은 관측확률을 나타내며, 상태집합은 품사 t로 구성되고, 관측심볼 집합은 단어 w로 구성된다. 상태전이 확률을 위한 바이그램 확률의 계산은 다음과 같다.

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}t_i)}{\sum_t C(t_{i-1}t)} = \frac{C(t_{i-1}t_i)}{C(t_{i-1})} \quad (2)$$

개체명 인식은 주어진 문장 W에 가장 적합한 개체형 열을 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 $W_1^{NE}\cdots W_n^{NE}$, 개체명의 좌측 문맥 $W_{-2}^{L}W_{-1}^{L}$과 우측 문맥 $W_1^{R}W_2^{R}$등을 고려해야 한다[7].

(좌문맥)	(개체명)	(우문맥)
$\cdots w_{-2}^{L} w_{-1}^{L}$	$w_1^{NE}\cdots w_n^{NE}$	$w_1^{R} w_2^{R}\cdots$

(그림 1) 가변길이 개체명을 위한 트라이그램 모델

따라서, 어휘를 기반으로 하는 문맥 가변 트라이그램 모델에서 학습되는 문맥은 (그림 2)와 같다.

$w_{-2}^{L} w_{-1}^{L} w_1^{NE}$
$w_{-1}^{L} w_1^{NE} w_2^{NE}$
...
$w_{n-1}^{NE} w_n^{NE} w_1^{R}$
$w_n^{NE} w_1^{R} w_2^{R}$

(그림 2) 학습 문맥

2) 개체명 독립, 구성, 인접 단어를 통칭한다.

그러나, 어휘 기반의 트라이그램 모델은 자료 부족 문제를 일으키기 때문에, 이 논문에서는 부개체 유형 및 품사 기반의 트라이그램을 학습한다. 단어 w_i는 3.2절에서 설명한 부개체 유형에 포함되는 경우에는 부개체 유형을, 개체명과 관련이 없는 경우에는 품사를 이용하여 문맥을 학습하도록 하였다. 즉, 품사 태깅에서 t_i와 대응하는 개념으로 각 단어에 대해 e_i를 이용하여 학습한다. 단어 "상사"의 경우, e_i는 '지명구성단어(W_{C}_Org)' 또는 '인명인접단어(W_{A}_Per)'를 나타내며, 개체명과 관련이 없는 단어의 e_i는 w_i의 품사를 가리킨다.

문장에서 개체명을 인식하는 단계는 크게 두 단계로 나눌 수 있다. 첫 번째 단계는 각 단어에 대해 가능한 부개체 유형을 생성하고, 이에 대한 부개체 모호성을 해소하는 것이다. 앞에서 설명한 "상사"의 경우도 부개체형 모호성이 나타난다. 두 번째 단계는 부개체형 생성된 문장에서 개체명의 시작과 끝을 인식하는 개체명 경계 인식 단계이다. 각 단계별로 학습되는 내용이 다르며, 이에 대한 모델은 다음과 같다.

- 부개체형 생성 단계
  - 인명, 지명, 기관명을 인식하고, 28개의 품사를 가지는 모델의 경우에 HMM 상태 집합 S는 문장 시작 기호(SOS)와 문장 끝 기호(EOS)를 포함하여 다음과 같다. 모든 단어에 대해 e는 아래의 상태 중에 하나이다. 즉, e_i는 문장의 i번째 단어인 w_i의 부개체 유형 또는 품사를 나타내는 임의의 S 중 하나이다.

    S = {W_{I}_Per, W_{C}_Per, W_{A}_Per, W_{I}_Loc, W_{C}_Loc, W_{A}_Loc, W_{I}_Org, W_{C}_Org, W_{A}_Org, POS_1, …, POS_{28}, SOS, EOS}

  - 부개체형 생성 단계에서 E를 주어진 문장 W의 가장 적합한 부개체형 열로 가정하고 부개형 생성 및 모호성 해소를 위한 트라이그램 기반 HMM은 다음과 같다.

$$P(E|W) = \arg\max_{T} \prod_{i}^{n} P(e_i|e_{i-1}\cdots e_1)P(w_i|e_i\cdots e_1) \approx \arg\max_{T} \prod_{i}^{n} P(e_i|e_{i-1}e_{i-2})P(w_i|e_i) \quad (3)$$

- 개체명 경계 인식 단계
  - 개체명 경계 인식을 위한 HMM의 상태는 인명, 지명, 기관명과 개체명의 경계를 나타내는 "시작(S)", "계속(C)", "끝(E)", "단일(U)", "개체명_아님(Not_NE)"과 같은 개체명 구성 단어의 내부적 순서 관계가 결합되어 나타난다. 개체명 경계를 위한 임의의 상태인 b_i는 e_i의 경계를 나타내며, 다음 S'중 하나의 상태를 가리킨다.

    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}

- 개체명 경계 인식 단계에서 B를 주어진 문장 W의 가장 적합한 개체의 경계 열로 가정하고, 개체명 경계 인식을 위한 트라이그램 기반 HMM은 다음과 같다.

$$P(B|W) \approx \arg\max_{T} \prod_{i=1}^{n} P(b_i|e_{i-1}e_{i-2})P(w_i|b_i) \quad (4)$$

3.3.2 개체명 인식 과정

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. 여기에서 "원"은 두 가지 부개체명 유형을 가질 수 있다.

〈표 4〉 개체명 인식의 예[3]

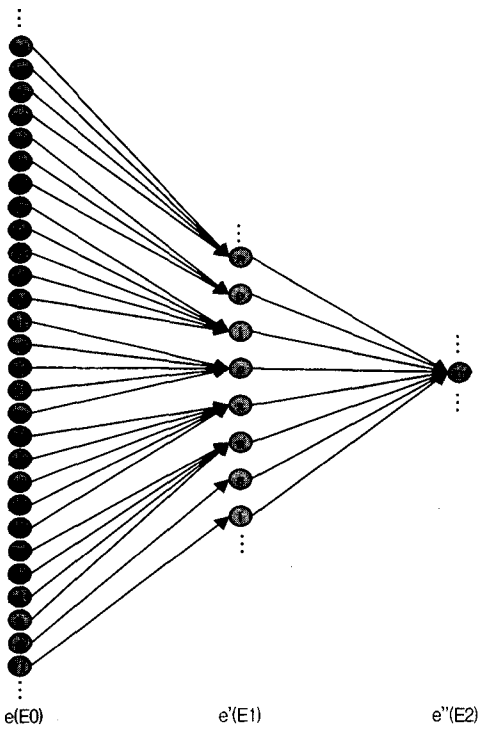
부개체형 생성	부개체형 분류	개체명 인식
set('LG')[4] : W_{I}_Org	LG : W_{I}_Org	LG : Org_S
set('경제') : W_{C}_Org	경제 : W_{C}_Org	경제 : Org_C
set('연구') : W_{A}_Per, W_{C}_Org	연구 : W_{C}_Org	연구 : Org_C
set('원') : WC_Loc, W_{C}_Org	원 : W_{C}_Org	원 : Org_E
set('은') : jx	은 : jx	은 : Not_NE
set('현대') : W_{I}_Org	현대 : W_{I}_Org	현대 : Org_S
set('상사') : W_{A}_Per, W_{C}_Org	상사 : W_{C}_Org	상사 : Org_E
set('가') : jc	가 : jc	가 : Not_NE

3.3.3 백-오프 모델

개체명 인식의 두 과정에서 각각 모호성이 발생하고, 이를 해결하면서 개체명을 인식하기 위해 서로 다른 통계적 확률을 적용하여 HMM을 구성하였다. 충분한 학습 데이터가 없을 경우, 트라이그램 모델은 심각한 데이터 부족 문제를 야기한다. 이를 해결하기 위해 상태전이 확률을 위한 백-오프 모델[16]을 사용하였는데, 부개체형 생성 단계와 개체명 경계 인식 단계 각각에 다른 모델을 사용하였다. 부개체형 인식을 위한 상태전이확률에서는 바이그램 확률을 이용하였다. 예를 들어, $P(e_i|e_{i-2}e_{i-1})$에 대한 확률이 학습되지 않았을 경우, $P(e_i|e_{i-1})$을 이용하였다. 또한, 개체명 경계 인식을 위한 단계에서는 e_i의 상태는 2단계 백-오프를 거치도록 하였다. 즉, 품사의 범주를 단계적으로 축소하는 것이다.

3) W_{I}는 개체명 독립 단어, W_{C}는 개체명 구성 단어, W_{A}는 개체명 인접 단어를 나타낸다. 따라서, W_{I}_Org는 기관명 독립단어, W_{A}_Org는 기관명 인접 단어, Org_E는 기관명의 끝을 나타낸다.

4) set('LG')는 e_i을 나타낸다. 즉, 'LG'의 부개체 유형을 반환한다.



(그림 3) 품사 범주의 축소를 통한 백-오프

- E0 : 3(W_I, W_C, W_A) * 3(P, L, O) + 28 품사 + SOS + EOS = 39 상태
- E1 : 3(W_I, W_C, W_A) * 3(P, L, O) + 8 품사 + SOS + EOS = 19 상태
- E2 : 3(W_I, W_C, W_A) * 3(P, L, O) + 1 품사(Not) + SOS + EOS = 12 상태

〈표 5〉 상태전이 확률을 위한 백-오프

부개체형 생성 단계	개체명 경계 인식 단계
$P(e_i \mid e_{i-2}e_{i-1})$	$P(b_i \mid e_{i-2}e_{i-1})$
↓	↓
$P(e_i \mid e_{i-1})$	$P(b_i \mid e'_{i-2}e'_{i-1})$
	↓
	$P(b_i \mid e''_{i-2}e''_{i-1})$

이는 학습 데이터가 부족함으로써 나타나는 자료 부족 문제를 극복하기 위해 품사 분류의 크기를 점진적으로 축소하는 것이다. 즉, E0 단계를 기본으로 하고 E0 단계의 문맥이 발견되지 않을 때, 완화된 트라이그램 문맥을 E1 단계에서 찾고, 여기에서도 문맥이 없으면 E2 단계에서 문맥을 찾는다. 예를 들어, "-아랍권"의 경우에, E0 단계에서 (-, 아랍, 권)에 대한 문맥인 (sym, W_I_Loc, xsn)을 찾고, 이러한 문맥이 없을 경우, 'xsn'의 제약을 완화하여 E1단계에서 'x' 품사를 포함하는 문맥인 (sym, W_I_Loc, x)를 찾는다[5]. E1 단계에서도 이러한 문맥이 없을 경우에는 다시 'sym'과 'x' 품사 제약을 완화하여 (Not, W_I_Loc, Not)로 백-오프 된다.

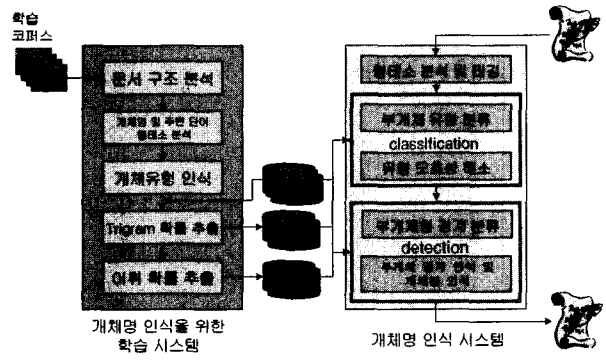
5) sym은 기호, xsn은 명사파생접미사, x는 접사, Not은 부개체형 아님을 나타내는 태그로 사용되었다.

4. 실험 및 평가

경제 분야 기사 100문서에서 학습을 위해 90문서를 이용하였고, 10문서를 대상으로 실험하였다. 학습 문서는 1,075개의 개체명을 포함하고 있으며, 인명 190, 지명 362, 기관명 523개를 포함하고 있다. 또한, 문서당 평균 7 문장으로 구성되었으며, 한 문장당 평균 18.7개의 어절을 가지고 있다. 학습을 위한 문서의 일부는 (그림 4)와 같다.

<S> <PERSON> 강봉균 </PERSON> (<PERSON> 康奉均 </PERSON>) <ORGANIZATION> 한국개발연구원 </ORGANIZATION> (<ORGANIZATION> KDI</ORGANIZATION>) 원장은 <DATE>11일</DATE> "세계적 경기침체가 장기화할 가능성이 없으며, <LOCATION>한국</LOCATION> 경제도 <DATE>내년 1~2분기</DATE>부터는 호전될 것" 이라고 전망했다.</S>
<S><REF> 강 </REF> 원장은 <REF> 이날 </REF> <DATE> 오전 </DATE> <LOCATION> 명동 세종호텔 </LOCATION> 에서 열린 <ARTIFACT> 세종대 세계경영대학 조찬회 </ARTIFACT> 에서 "지난 <DATE> 7, 8월 </DATE> 수출이 <DATE>전년</DATE> 동기 대비 <PERCENT> 20% </PERCENT> 이상 줄어들고 3·4분기 국내총 생산(GDP) 증가율이 <PERCENT> 1% </PERCENT> 미만으로 떨어질 가능성이 높지만, <DATE>내년</DATE> 에는 경제가 회복될 것" 이라고 말했다. </S>
<S> 이밖에도 <REF> 강 </REF> 원장은 " <LOCATION> 중국</LOCATION> 경제가 우리의 <DATE> 1970~1980년 </DATE> 대처럼 박진감 있는 것처럼 보이지만 상업적인 금융시스템이 갖춰져 있지 않고 국영기업의 생산성이 낮은데다가 경영의 투명성과 지배구조의 민주성을 기대하기 어려운 만큼 결코 두려워할 상대는 아니다" 라고 말했다. </S>

(그림 4) 학습 문서의 예



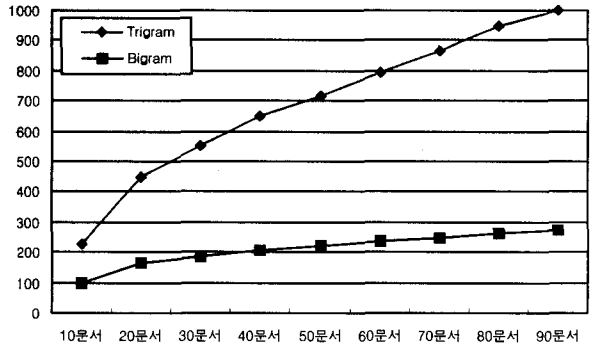
(그림 5) 한국어 개체명 인식 시스템

한국어 개체명 인식을 위해 본 논문에서 제안하는 시스템의 구성은 다음과 같다. 학습 부분은 부개체 유형 인식을 위한 학습 과정과 개체명 경계 인식을 위한 학습이 동시에 진행되며, 이때 상태전이확률과 어휘 확률이 학습된다. 이때, 개개의 단어에 대해 부개체명 사전을 이용하여 트라이그램과 어휘 확률이 추출된다.

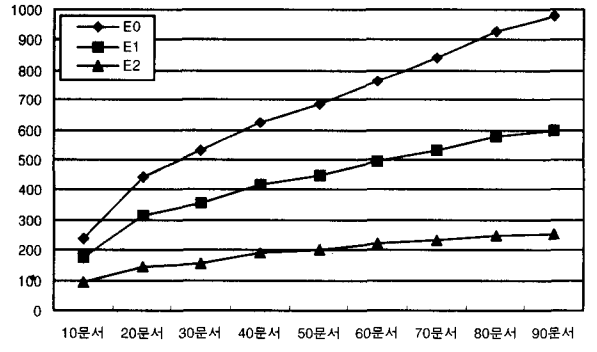
개체명 인식 과정에서는 문장에 대해 형태소 분석과 태깅을 거친 후, 각 단어에 대해 가능한 부개체 유형을 생성하고, 부개체 유형 모호성을 해소하여 최적의 부개체형 열

을 얻는다. 이를 바탕으로, 각 단어에 대해 가능한 개체명의 경계를 생성하고 이를 이용하여 최적의 개체명 경계열을 찾는다.

실험을 위해 학습 문서를 대상으로 부개체형 인식 및 경계 인식을 위한 트라이그램을 추출하였다. (그림 6)과 (그림 7)은 학습 문서를 10문서씩 90문서까지 증가하면서 추출된 문맥들이다. (그림 6)은 부개체형 인식을 위해 학습된 트라이그램 및 백-오프를 위한 바이그램의 수를 나타내며 (그림 7)은 개체명 경계 인식을 위해 학습한 트라이그램의 수를 나타낸다.



(그림 6) 부개체형 인식을 위해 학습한 문맥



(그림 7) 학습된 개체명 경계 인식 트라이그램

개체명 인식에 사용된 부개체형 사전은 다음과 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 6〉 부개체명 사전의 구성

구 분	개체명 독립 단어 사전	개체명 구성 단어 사전	개체명 인접 단어 사전
엔트리 수	102,154	214	185

개체명 인식 시스템의 성능을 평가하기 위해, 학습 문서를 증가시키면서 경제 분야 신문기사 10문서에 대하여 개체명 인식률을 조사하였다. 각 실험은 E0에서 E1, E2로 백-오프를 통해 얻은 결과이다. 개체명 인식률은 학습 문서의 양이 증가함에 따라 인식률이 높아졌다. 기존 연구와 비교해 보면, 상대적으로 적은 학습 문서를 사용하였음에도 유사하거나 높은 인식률을 보였다.

〈표 7〉 개체명 인식률

학습 문서수	정확률	재현률	F-measure
10	82.0%	75.8%	78.7%
40	85.7%	81.8%	83.7%
90	84.5%	90.9%	87.6%

인식오류를 유형별로 분류해보면 개체명 인식 모델의 특성상 과생성이 발생하여, 정확률을 감소시켰다. 예를 들면, "신용+카드(W_{C}_Org)", "미국+산(W_{C}_Loc)", "외환+시장(W_{C}_Loc)", "상품+투자(W_{C}_Org)" 등이 그 예로 보통명사와 개체명 구성 단어들이 결합되어 개체명으로 인식되는 경우가 많았다. 이러한 문제는 단어들 간의 상호 배제 정보와 같은 지식을 통해 해결해야 할 것이다. 또한, 학습 문서의 부족으로 인해, "CJ39 쇼핑"과 같이 비교적 발생 빈도가 낮은 구조를 가지는 개체명 형태를 인식하지 못했다.

5. 결론 및 향후 연구 방향

본 논문에서는 개체명 인식을 위한 국내외 연구를 살펴보고, 한국어 문서에서 나타나는 개체형의 유형과 특성, 개체명을 구성하는 한국어의 구조적 특징을 조사하고, 한국어에 적합한 개체명 인식 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 학습하고, 이를 이용하여 HMM 기반 개체명 인식 시스템을 구축하였다. 한국어 개체명은 개체명 주변 문맥에 의존하며, 개체명 자체의 구성 문맥에도 의존적인 경향이 많으므로, 이들을 자동으로 학습하고 이를 개체명 인식 과정에 반영함으로써 개체명 인식의 성능 향상을 기대할 수 있었다.

본 논문에서 제안한 개체명 인식 시스템은 부개체명 사전에 기반하기 때문에, 개체명 또는 개체명 주변에 개체명을 인식할 만한 단서가 전혀 없는 경우에는 개체명을 인식하지 못하는 단점이 있다. 이러한 문제는 미지어 추정을 통한 해결과 관련이 있으며, 이를 제안한 모델에 포함시키는 방법이 필요하다. 또한, 부개체형 인식 후 개체명을 인식하는 두 단계 접근 방법을 통합하여 하나의 학습 모델에 수

용함으로써 불필요한 정보량을 줄이면서 개체명 인식 속도를 향상시키는 연구가 필요하다.

참 고 문 헌

[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] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

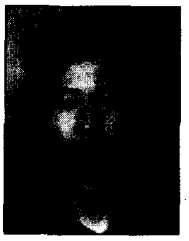
[12] G. D. Zhou, J. Su, "Named Entity Recognition using an HMM-based Chunk Tagger," In Processing of the ACL 2002.

[13] 김태현, 이현숙, 하유선, 이만호, 맹성현, "데이터 집합을 이용한 고유명사 추출", 제 12회 한글 및 한국어 정보처리 학술대회, pp.11-18, 2000.

[14] 노태길, 이상조, "규칙 기반의 기계학습을 통한 고유 명사의 추출과 분류", 한국정보과학회 가을 학술발표논문집, Vol.27, No.2, pp.170-172, 2000.

[15] 이경희, 이주호, 최명석, 김길창, "한국어 문서에서 개체명 인식에 관한 연구", 한글 및 한국어 정보처리 학술대회, pp. 292-299, 2000.

[16] S. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," IEEE Transaction on Acoustic, Speech, and Signal Processing, Vol.ASSp-35, No.3, pp.400-401, 1987.



황 이 규

e-mail : yghwang@etri.re.kr

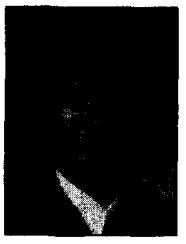
1993년 전북대학교 전자계산학과(학사)

1995년 전북대학교 대학원 전산통계학과 (이학석사)

2001년 전북대학교 대학원 전산통계학과 (이학박사)

2001년~현재 한국전자통신연구원 선임연구원

관심분야 : 자연어처리, 정보추출, 텍스트마이닝



윤 보 현

e-mail : ybh@mokwon.ac.kr

1992년 목포대학교 전산통계학과(학사)

1995년 고려대학교 컴퓨터학과(석사)

1999년 고려대학교 컴퓨터학과(박사)

1999년~2003년 ETRI 지식처리연구팀 팀장

2003년~현재 목원대학교 컴퓨터교육과 전임강사

관심분야 : 자연어처리, 정보검색, 지식처리, 바이오인포매틱스