한국어 문서에서 개체명 인식에 관한 연구

이경희 이주호○ 최명석 김길창
한국과학기술원 전자전산학과
{khlee, leejh, mschoi}@csone.kaist.ac.kr, gckim@cs.kaist.ac.kr

Study on Named Entity Recognition in Korean Text

Kyung Hee Lee Ju Ho Lee Myung Seok Choi Gil Chang Kim
Dept. of Electrical Engineering & Computer Science, KAIST

요 약

본 논문에서는 개체명 사전과 결합 단어 사전, 그리고 용언의 하위범주화 사전을 이용하는 규칙 기반의 한국어 개체명 인식 방법을 제안한다. 각 규칙은 네 단계로 나누어 적용되는데, 첫번째 단계에서는 어절 내의 단어 정보를, 두번째 단계에서는 제한된 주변 문맥 정보를, 그리고 세번째 단계에서는 용언의 하위범주화 정보와 개체명과의 관계를 이용하고, 마지막으로 네번째 단계에서는 개체명 간의 관계 정보를 고려한다. 본 논문에서 제안한 규칙 기반 개체명 인식기의 성능을 평가하기 위해 실험한 결과 90.4%의 정확률과 83.4%의 재현율을 얻었다.

1 서론

오늘날 정보가 기하급수적으로 증가함에 따라 원하는 정보를 검색하거나 자동으로 정보를 추출하고자 하는 요구가 점점 증가하고 있다. 정보 검색이나 추출과 같은 자연어 처리의 응용 분야에서 중요한 작업 중의 하나는 문서 상의 핵심어를 찾아내는 것이다. 핵심어는 정보 검색에서 주요 검색대상이 되며 정보 추출 시에는 추출할 정보를 구성하는 요소가 될 수 있다. 이러한 핵심어의 대부분은 인명, 지명, 조직명, 시간, 날짜, 화폐 등의 개체명(Named Entity)이다.

이러한 개체명은 대부분 문서에서 중요한 역할을 하지만 사전에 등록되지 않은 고유 명사인 경우가 많다. 고유 명사는 한정된 것이 아니라 계속 만들어지기 때문에 모든 고유 명사를 사전에 등록하는 것은 현실적으로 불가능하다. 또한 사전에 등록된 경우에도 그 범주가 정해진 것이 아니라 문맥에 따라 범주가 달라지기 때문에 사전만 가지고 범주를 구별해 주는 일이 쉽지 않다. 예를 들어 예문 (1)에서는 '청와대'가 조직명이지만, 예문 (2)에서는 장소를 나타내는 지명이다.

(1) 청와대 측은 소문을 부인했다.
(2) 이한동 대표가 청와대를 방문했다.

이와 같이 개체명 인식은 문서에서 개체명을 추출하고 이의 범주를 결정하는 것을 말하며, 개체명의 범주로는 인명, 지명, 조직명, 날짜, 시간, 화폐, 퍼센티지, 상품명 등을 포함하고 있다[16, 15, 8]. 이러한 개체명 인식은 주로 정보 검색 및 추출, 문서 분류, 자동 정렬, 조응 해소 등에서 이용되고 있다[15, 9, 11].

고유 명사를 표기하는데 대문자를 이용하는 영어에서는 개체명의 후보를 비교적 쉽게 추출할 수 있으며, 규칙 기반 방식을 이용하여 높은 수준의 정확률을 보이고 있다. 그러나 한국어에서는 대소문자의 구분이 없기 때문에 개체명 후보를 찾고 이의 범위를 결정하기가 상대적으로 어렵다.

본 논문은 한국어 문서에서 개체명 사전과 용언의 하위범주화 사전을 기반으로 한 규칙 기반 개체명 인식 방법을 제안한다. 규칙은 이용하는 정보에 따라 네 단계로 나누어 순차적으로 적용된다. 첫번째 단계에서는 개체명의 후보가 되는 단어 자체의 사전 정보와 단어 구성 정보를 보고, 두번째 단계에서는 후보 주변에 나타나는 단어에 대한 정보를 이용한다. 세번째 단계에서는 용언의 하위 범주화 사전을 바탕으로 개체명을 제한하는 용언을 고려하고, 네번째 단계에서는 앞의 단계에서 결정된 정보를 바탕으로 어절 간의 관계를 살펴 개체명의 범위와 범주를 결정한다. 각 단계에서는 개체명 후보의 범주를 결정하는 것이 아니라 가중치만을 변경시키므로써 전체 시스템이 좀더 유연한 구조를 이루도록 하였다.

본 논문의 구성은 다음과 같다. 2절에서는 개체명 인식의 관련 연구에 대해서 살펴보고, 3절에서는 한국어 개체명 인식의 네 단계 구조를 보이고 각 단계에서 사용하는 정보와 규칙을 설명한다. 4절에서는 구현 환경과 실험 방법, 한국어 문서에서 개체명 인식의 결과에 대해 기술하고, 각 단계의 결과와 오류에 대해 분석하고 평가한다. 마지막으로 5절에서는 본 논문의 결론을 맺는다.

2 관련 연구

이 절에서는 개체명 인식의 기존 연구에 대해서 살펴보고, 한국어 문서에서 개체명 인식의 문제점에 대해서 살펴본다.

표 1: 문맥 규칙의 예

규칙	범주	예
Xxxx+ is a? JJ* PROF	PRES	Yuri Gromov is a former director,
PERSON-NAME is a? JJ* REL	PRES	John White is beloved brother
Xxxx+, a JJ* PROF,	PRES	White, a retired director,
Xxxx+, ? whose REL	PRES	Nunberg, whose stepfather
Xxxx+ himself	PRES	White himself
Xxxx+, DD+,	PRES	White, 33,
shares of Xxxx+	ORG	shares of Eagle
PROF of/at/with Xxxx+	ORG	director of Trinity Motors
in/at LOC	LOC	in Washington
Xxxx+ area	LOC	Beribidjan area

Xxxx+는 대문자로 시작되는 단어열, DD는 숫자,
PROF는 직업(director, manager, analyst, 등),
REL은 사람 관계(sister, nephew, 등), JJ*는 0이나 하나 이상의 형용사,
LOC는 사전에 있는 지명, PRESON-NAME은 인명을 의미한다.

2.1 영어권의 연구

개체명 인식에 관한 연구는 MUC-6, MUC-7, MET-1, MET-2를 통하여 활발히 진행되어 왔는데, 크게 사전과 규칙을 이용하는 규칙 기반 방식과 사전, 품사 정보, 문자 유형 등을 이용하여 개체명 인식에 필요한 규칙을 자동으로 추출하는 통계 기반 방식으로 나누어 볼 수 있다.

규칙 기반 방식에서는 개체명 인식을 위해 주로 사전과 규칙을 이용한다. 여기서 사전은 인명, 지명, 조직명 등의 개체명에 대한 직접적인 정보를 가지고 있는 개체명 사전과 개체명과 같이 붙어서 나타나는 단어들(Mr., Dr., Co., Ltd., Bank, University, City 등)을 포함하는 결합 단어 사전 등을 말한다. 개체명 인식에서 사용되는 규칙은 개체명의 후보가 되는 단어 자체의 정보를 이용하는 단어 구성 규칙과 문장에서 개체명의 후보가 되는 단어의 주변 정보를 이용하는 문맥 규칙이 있는데 이러한 규칙은 일반적으로 사람이 직접 기술하고 있다.

[14]에서는 사전과 규칙을 기반으로 한 다단계 개체명 인식 방법을 제안했다. 우선 첫번째 단계에서는 확실한 문맥 규칙을 이용하여 개체명을 인식한다. 사용하는 문맥 규칙은 표 1과 같다. 그리고 그 다음 단계에서는 앞단계에서 결정된 개체명을 이용한 부분 매칭을 통해서 새로운 개체명을 인식한다. 즉, 앞에서 나온 단어가 축약되어 사용되는 경우나 부분적인 형태로 사용되는 경우를 개체명으로 인식한다. 예를 들어 "Lockheed Martine Production"가 이미 조직명으로 인식된 경우에, 같은 문서 내에서 나타난 "Lockheed Martine"이나 "Lockheed Production"도 역시 조직을 뜻하는 개체명으로 인식할 수 있다. 이런 식으로 일단 정확하다고 판단되는 개체명을 먼저 인식한 후에 다음 단계에서는 확장 규칙을 이용한다. 예를 들어, 인명 사전에 나타난 단어 뒤에 대문자로 시작하는 미등록어가 나온다면 그것을 합쳐서 인명으로 인식한다. 그 후에 다시 이전 단계에서 인식된 개체명을 이용하여 부분 매칭을 통해 새로운 개체명을 인식한다.

[12]에서도 역시 사전과 규칙을 이용하지만 구문 분석을 통해 문맥 규칙과 용언 정보를 이용하는 차이점이 있다.

규칙에 기반한 방법은 일반적으로 사람이 직접 규칙을 기술해야 하지만 그만큼 정확한 결과를 보여 주고 있다.

그리고 통계 기반 방식은 품사 정보, 문자 유형, 사전 정보를 기반으로 은닉 마르코프 모델(Hidden Markov Model)[8], 최대 엔트로피 모델(Maximum Entropy Model)[10], 결정 트리(Decision Tree)[15] 등을 이용한다.

2.2 한국어 문서에서 개체명 인식

한국어에 대해서는 정보 검색에서 필요한 색인어 추출을 위해서 고유명사의 출현 패턴[1]을 이용하거나 성씨 정보를 이용해 인명일 가능성이 있는 경우를 추출하여 뒤에 호칭이나 지위를 나타내는 말이 나올 경우 색인어로 채택하는 연구가 있었다[2].

그리고 문서에 나타나는 미등록어 추정을 위해 형태소 분석이 실패한 어절을 입력으로 해서 어절 내의 조사와 어미, 접미사, 명사를 고려하여 미등록어를 추출하고, 고유명사인 경우 범주를 추정하는 연구도 있었다[4].

그러나 문서에서 개체명을 찾기 위한 목적보다는 다른 목적을 위한 연구의 일부이거나 인명에 관련된 간단한 규칙을 제시하는 데 그쳤고 조직이나 지명에 대한 체계적인 연구는 거의 이루어지지 않고 있다.

2.3 문제점

1절에서 언급한 바와 같이 개체명 인식은 사전에 등록되지 않은 단어에 대한 문제와 문맥에 따라 다르게 사용될 수 있는 중의성의 문제를 안고 있다.

이에 더하여 본 논문에서 대상으로 하고 있는 한국어는 영어권의 언어와 달리 대소문자 구분 정보가 없기 때문에 생기는 문제들이 있다. 개체명은 대부분이 고유 명사이기 때문에 고유 명사를 대문자로 표기하는 영어권의 언어에서는 대소문자 구분 정보만으로도 개체명의 후보를 쉽게 추출할 수 있다. 또한 이러한 대소문자 구분 정보를 이용하여 개체명의 구성 단어로 이용되는 단어들을 통해 확실한 개체명의 범주를 얻을 수 있다. 예를 들어 영어의 'Bank'의 경우, 한국어의 '은행'으로 볼 수 있다. 그러나 영어권에서는 같은 단어도 고유 명사일 경우는 'Bank'로 사용하지만 보통 명사일 경우는 'bank'로 사용하기 때문에 확실한 정보가 될 수 있다. 하지만 한국어에서는 항상 '은행'으로 사용되기 때문에 보통 명사로 사용되었는지 고유 명사로 사용되었는지 구별해 주어야 한다[2].

또한 한국어는 일반적으로 여러 단어가 합쳐져 복합 명사로 사용되는 특성을 가지고 있기 때문에 단어 자체가 중의성을 가지게 된다. 따라서 개체명을 구성하고 있는 단어의 구별을 통한 인식과 이외의 부가 정보가 필요하다. 예를 들어 인명이 문서 상에서 나타나는 형태를 살펴보면, 한국어에서는 성과 이름이 하나의 단어로 붙어서 나타난다. 따라서 성씨 사전을 가지고 있다 해도 보통 명사의 경우나 다른 고유 명사인 경우와 구별해야 하는 중의성이 존재한다. 즉, '남해안'이나 '하와이'처럼 고유 명사로 태깅되고, 사전에 나타나지 않는 단어인 경우 '남'과 '하'가 성씨 사전에 나타나기 때문에 사람으로 인식될 가능성이 있다.

이처럼 한국어는 개체명의 후보를 찾아주고, 보통 명사와 구별해 주어야 하는 문제와 결합하여 사용되는 단어의 특성으로 인한 단어 자체의 중의성 문제를 더 가지고 있다.

[1] '이름', '이름+조사', '이름+호칭', '이름+호칭+조사', '이름+고유명사+호칭+조사', '고유명사(회사/기관)', '고유명사+조사', '고유명사+호칭', '고유명사+호칭+조사', '호칭', '호칭+조사'

[2] '은행'이라는 단어는 '한빛 은행'과 같이 개체명으로 쓰일 수도 있지만, '시중 은행'과 같이 보통 명사로 사용될 수도 있다.

3 한국어 개체명 인식

본 논문에서는 개체명의 범주를 중의성이 있고 상대적으로 인식이 어려운 인명, 지명, 조직명으로 개체명 인식의 범주를 제한하고자 한다[3]. 이러한 인명, 지명, 조직명의 범주를 결정하는 과정에는 중의성이 많이 존재한다. 그래서 개체명 사전 뿐 아니라 결합 단어 사전, 성씨 사전 등을 이용하여 단어 자체의 범주를 살펴보고, 이를 바탕으로 제한된 주변 문맥 단어 정보와 용언의 하위범주화 정보를 통해 이를 검토하고 새로이 결정하는 과정이 필요하다. 또한 한국어에서는 사전과 규칙을 단계별로 적용하면서 보통 명사와 구별하여 주는 작업도 필요하다. 이러한 작업은 각 단계의 규칙을 통해 이루어진다.

개체명을 인식하는 과정의 전체적인 구성은 그림 1과 같다. 품사 태깅된 문서를 입력으로 하여 각 문장을 어절 단위로 입력 받아 개체명 인식의 네단계가 순차적으로 적용되게 된다. 각 단계는 독립적으로 수행되며, 개체명 후보의 범주를 결정하는 것이 아니라 가중치만을 변경시키므로써 전체 시스템이 좀더 유연한 구조를 이루도록 하였다.

다음으로 본 논문에서 사용하는 사전 정보에 대해서 기술하고, 본 논문에서 제안하는 개체명 인식의 네단계와 각 단계에서 이용하는 규칙에 대해서 설명한다.

3.1 사전 정보

개체명 인식을 위해서 네단계의 규칙을 적용하기 전에 우선 입력으로 들어오는 각 단어들의 사전 정보를 추출해야 한다. 이를 위해 본 연구에서 사용하는 사전의 종류는 표 2와 같다.

표 2: 사전의 종류와 예

사전	표기	엔트리의 예	엔트리 수
인명	PER	김대중	10079
조직명	ORG	공정거래위원회	1026
지명	LOC	서울시	6524
국가명	ORG_GOV	한국, 미국, 일본	180
성씨	PER_LAST	김, 이	178
사람명사	PER_NOUN	지도자, 건설업자	3959
지위 및 호칭	PER_POS	씨, 교수, 회장	110
조직명사	ORG_NOUN	위원회, 은행, 회사	467
지명결합명사	LOC_SUFFIX	시, 도	10
지명명사	LOC_NOUN	지역	10

사전은 크게 두가지로 나누어 볼 수 있다. 우선 인명, 조직명, 지명, 국가명 등의 고유 명사 사전[4]은 개체명을 이루는 전체 단어들로 이루어진 사전으로 이 사전에 나온 단어가 문서 상에 나타난다면 개체명 후보로 인식할 수 있다. 그리고 나머지는 개체명과 함께 사용되는 단어들의 사전[5]이다.

다음으로 본 논문에서 사용하는 규칙 표현에 대해 설명한다. 그림 2를 보면 규칙은 크게 패턴과 값과 범위의 영역으로 나누어진다. 패턴은 규칙이 적용되는 조건을 의미하고, 값은 각 규칙을 만족시키는 경우의 가중치 값과 그 때 부여하는 자질 값을 나타낸다. 가중치는 단계 1에서 단계 4를 통해 단계별로 누적되는 값이며, 단계 4에서는 중심이 되는 단어의 가중치 값을 부여한다. 또한 범위는 개체명으로 인식할 단어의 범위를 의미한다.

[3]이외의 개체명인 날짜, 시간, 화폐, 퍼센티지의 경우 비교적 간단한 규칙만으로 높은 정확률을 보이므로 본 연구에서는 제외하였다.

[4]개체명 사전으로 지칭한다.

[5]결합 단어 사전으로 지칭한다.

3.2 단계 1 : 어절 내 단어 정보

단계 1에서는 어절 내 단어 정보를 이용해서 개체명 인식을 한다. 이 단계는 후보를 찾고 사전을 검색하는 부분과 후보열에 대해 규칙을 적용하는 두 부분으로 나뉜다. 우선 입력으로 품사 태깅된 어절 단위의 문장을 읽어들인 다음, 개체명의 후보를 찾기 위해 고유 명사나 보통 명사의 열을 찾는다. 이렇게 찾은 개체명 후보열에 대해 사전을 검색한다. 후보가 되는 단어열을 개체명 사전에서 검색하여 찾은 경우에는 그 자질 값을 부여하고, 찾을 수 없는 경우에는 각 단어에 대해 다음과 같은 3가지의 검색을 한다.

성씨 사전 검색 고유 명사 한 단어로 이루어지고 2 내지 4글자로 이루어진 경우, 앞글자 한 글자나 두 글자가 성씨 사전에 나타나는 경우, 사람의 성씨를 나타내는 'PER_LAST'라는 자질 값을 부여한다.

부분 단어 검색 호칭이나 지위, 조직이나 지명을 나타내는 보통 명사나 접미어가 다른 단어와 결합하여 사용되는 경우, 결합 단어 사전을 이용하여 단어 내에 나타나는 결합 단어를 찾아내어 해당하는 자질 값을 부여한다. 예를 들어, '최고위원', '전대통령'과 같은 단어의 경우, 사전에 나타나는 '위원'이나 '대통령'과 같은 단어를 보고 지위를 나타내는 'PER_POS'이라는 자질 값을 준다.

축약어 검색 일반적으로 널리 사용되는 조직명이나 지명인 경우, 축약해서 간단한 단어로 사용되는 경우가 많으며, 이러한 축약어는 개체명을 이루는 단어들의 첫글자를 모아서 이용하거나 일부 단어의 전체와 나머지 단어의 첫글자를 모아서 혹은 생략해서 사용한다. 예를 들어 지명의 경우 '서울특별시'는 '서울'이나 '서울시'와 같은 축약된 형태로 이용된다. 그리고 사전에 없는 경우, 문서의 앞문장에서 이미 결정된 개체명 인식의 결과를 가지고 이 단계에서 같은 방법으로 부분 매칭(partial matching)을 시도한다.

다음은 앞에서 찾은 후보와 사전 정보를 이용하여 단어 구성 규칙을 적용한다. 이미 언급한 바와 같이 단어 구성 규칙은 후보가 되는 어절 내의 정보를 이용하는 것이다. 그림 2는 단어 구성 규칙의 일부를 보여준다. 규칙은 그 단어가 만족하는 조건을 제시하며, 규칙에 속한 값의 내용은 그 규칙을 만족시킬 때 각 범주에 대해 부여하는 가중치이다. 이 때 부여하는 가중치는 양의 값일 수도 있지만 음의 값을 가지기도 한다. 예를 들어 규칙 (5)는 고유 명사이거나 보통 명사가 2단어 이상이고 그 뒤에 ORG_NOUN의 자질 값을 가진 단어가 오면 ORG으로 결정함을 의미한다. 즉, 인명에는 가중치 1.0을 빼주고, 지명은 가중치 1.5를 더해주고, 조직명은 가중치 2.0을 더해주는 것이다.

다음의 예문 (3), (4), (5)를 대상으로 품사 정보를 기반으로 후보를 뽑은 후 단어 구성 규칙을 적용시킨 결과는 표 3과 같다.

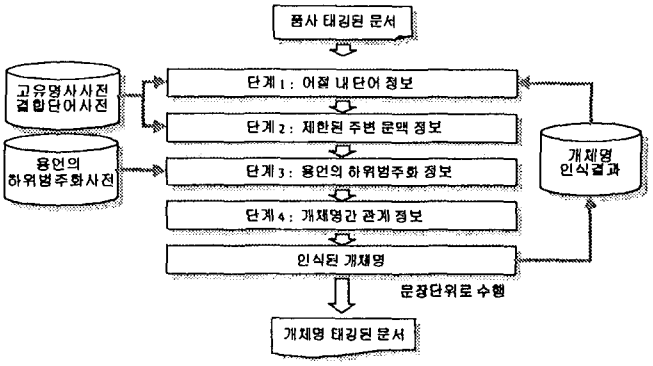

그림 1: 전체 시스템의 구성

패턴	값				범위
	인명	지명	조직명	자질값	
(1) ORG_GOV	-1.0	+1.5	+2.0	ORG_GOV	1-1
(2) Wnq & (length(W) = 3) & PER_LAST	-	-	-	PER_LAST	1-1
(3) Wnq & (LOC_NOUN | LOC_SUFFIX)	-	-	-	LOC	1-1
(4) ((!LOC & !ORG & Wnq) | (PER | PER_LAST)) + PER_POS	+2.0	-1.0	1.0	PER	1-2
(5) (Wnq | Wncn+2) + ORG_NOUN	-1.0	+1.5	+2.0	ORG	1 2

그림 2: 단어구성규칙(일부)

(3) 미국은 캐나다 밴쿠버에서 열린 예선전에서 우세한 경기를 펼쳐 캐나다를 이겼다.
(4) 이한동 대표가 청와대를 방문했다.
(5) 그는 국정의 혁신을 위해서 국정개혁위원회를 설치해야 한다고 주장했다.

표 3: 단계 1의 결과

단어	품사	인명	지명	조직명	자질
미국	nq	-1.0	1.5	2.0	ORG_GOV
캐나다	nq	-1.0	1.5	2.0	ORG_GOV
밴쿠버	nq	0	0	0	NONE
캐나다	nq	-1.0	1.5	2.0	ORG_GOV
이한동	nq	0	0	0	PER_LAST
청와대	nq	0	0	0	NONE
국정 개혁 위원회	ncn+ncn+ncn	-1.0	1.5	2.0	ORG

여기에서 '미국', '캐나다'는 국가명 사전을 이용하여 ORG_GOV라는 것을 알 수 있었고, 규칙 (1)에 따라 각 범주에 가중치를 부여하였다. '이한동'의 경우에는 인명 사전에서 찾지 못하고, 3글자이고, 고유 명사이고, 첫글자가 성씨 사전에 있음을 가지고 규칙 (2)에 따라 PER_LAST로 구분되었다. '청와대'와 '밴쿠버'의 경우에는 사전에서 찾지 못하고, 적용되는 규칙도 없었기 때문에 아무런 정보도 얻지 못했다. 그러나 '국정개혁위원회'의 경우에는 사전 검색에는 실패했으나, 규칙 (5)를 적용하여 ORG의 자질값과 가중치를 부여하였다.

이와 같이 사전과 단어 구성 규칙만을 이용하는 경우 '미국', '캐나다' 외의 단어는 범주를 결정하기가 어려웠다. 또한 이 경우에도 확실히 조직명으로 보기는 어렵기 때문에 단어 구성 자체뿐 아니라 주변 단어들의 정보를 이용해야 한다. 본 논문에서는 이를 문맥 규칙이라고 하고, 다음 3.3절에서 설명한다.

3.3 단계 2 : 제한된 주변 문맥 정보

3.2절에서 언급한 바와 같이 사전과 단어 자체의 정보만으로는 정확한 범주를 추정하기 어렵다. 이 단계에서는 개체명 후보가 되는 단어의 앞이나 뒤에 나타나는 단어를 고려하는 문맥 규칙을 적용시킨다. 이 단계에서 적용하는 규칙의 예는 그림 3과 같다. 문맥 규칙은 규칙 (6)과 같이 개체명 후보의 앞단어를 보는 규칙과 규칙 (7), (8)과 같이 개체명 후보의 뒷단어를 보는 규칙으로 나뉜다. 이 단계에서는 후보가 되는 단어의 앞뒤에 인명이나 지명, 조직명을 나타내는 보통 명사가 나타나는 경우와 고유 명사 뒤에 인명 뒤에 붙을 수 있는 호칭이나 지위 등을 나타내는 단어가 나타나는 경우 등을 포함한다.

패턴	값				범위
	인명	지명	조직명	자질값	
(6) PER_NOUN + W	+2.0	-	-	PER	2 2
(7) ((!LOC & !ORG & Wnq) + PER_POS	+2.0	-	-	PER	1-2
(8) ((Wnq & (words(W) = 1)) | (words(W) >= 2)) + ORG_NOUN	-	-	+2.0	ORG	1-2

그림 3: 문맥 규칙(일부)

3.2절에서 제시했던 예문에 적용한 결과는 표 4와 같다. 규칙 (7)에 따라 단어 '이한동'이 고유명사이고 뒤에 나타난 단어 '대표'가 PER_POS이기 때문에 '이한동 대표'는 PER으로 분류되었다. 그러나 여전히 '밴쿠버'와 '청와대'는 인식되지 않았다. 다음 단계에서는 용언의 하위 범주화 정보와 조사 정보를 이용한다. 이를 본 논문에서는 용언의 선택 제약 규칙이라 하고, 다음 3.4절에서 설명한다.

표 4: 단계 2까지의 결과

단어	품사	인명	지명	조직명	자질
미국	nq	-1.0	1.5	2.0	ORG_GOV
캐나다	nq	-1.0	1.5	2.0	ORG_GOV
밴쿠버	nq	0	0	0	NONE
캐나다	nq	-1.0	1.5	2.0	ORG_GOV
이한동 대표	nq+ncn	2.0	-1.0	-1.0	PER
청와대	nq	0	0	0	NONE
국정 개혁 위원회	ncn+ncn+ncn	-1.0	1.5	2.0	ORG

3.4 단계 3 : 용언의 하위범주화 정보

한국어의 특징 중 하나는 조사가 발달되어 있다는 것이다. 이러한 조사는 명사구에 붙어 동사와의 관계를 잘 드러내 준다[3]. 또한 용언마다 보어로 취할 수 있는 명사들이 한정되어 있다[1]. 이 단계에서는 한국어의 이러한 특징을 이용하여 조사 정보와 용언 정보를 이용하고자 한다. 이를 위해 용언의 하위 범주화 사전[6, 5]을 이용하는 용언의 선택 제약 규칙을 적용한다.

용언의 하위 범주화 사전은 기본적으로 그림 4와 같은 용언의 하위 범주 정보를 제공한다.

용언	문형 및 선택 제약 규칙
방문하다	[인명]이 [인명, 지명]을 방문하다.
열리다	[기타]이 [지명]에(서) 열리다.
이기다	[인명, 조직명]이 [인명, 조직명]을 이기다

그림 4: 단어구성규칙(일부)

먼저 문장에서 나타나는 용언을 찾고, 조사를 이용한 간단한 휴리스틱으로 하위 범주의 제약을 받는 보어가 되는 명사를 찾은 후 명사가 개체명 후보이면 가중치를 부여한다. 그러나 하위범주 정보가 하나의 보어에 대해 하나의 범주만을 갖는 것이 아니므로 앞 단계에서 결정된 정보를 이용하게 된다. 예를 들어, 단어가 사전과 앞단계의 규칙에 의해 조직명이나 지명으로 판단된 경우에는 하위 범주 정보에서 인명에 대한 가중치를 주지 않는다. 그리고 보통 명사인 경우보다 고유 명사인 경우에 높은 가중치를 부여한다. 이러한 하위범주화 정보는 지명과 조직명 사이의 중의성을 효율적으로 구분할 수 있고, 이전 단계에서 결정되지 못한 개체명에 대해서도 범주를 결정할 수 있게 된다.

표 5: 단계 3까지의 결과

단어	품사	인명	지명	조직명	자질
미국	nq	-1.0	1.5	4.0	ORG_GOV
캐나다	nq	-1.0	1.5	2.0	ORG_GOV
밴쿠버	nq	0	2.0	0	LOC
캐나다	nq	-1.0	1.5	4.0	ORG_GOV
이한동 대표	nq+ncn	4.0	-1.0	-1.0	PER
청와대	nq	0	2.0	0	LOC
국정 개혁 위원회	ncn+ncn+ncn	-1.0	1.5	2.0	ORG

3.3절에서 나온 결과에 이 단계의 규칙을 적용시키면 표 5와 같다. '이기다', '열리다'의 선택 제약 정보를 이용하여 '미국', '밴쿠버', 두번째 나타난 '캐나다'를 각각 '조직명', '지명', '조직명'으로 명확히 분류할 수 있었다. 그러나 첫번째의 '캐나다'는 여전히 중의성을 가지고 있고 뒤에 나타난 '밴쿠버'와 분리되어 있다. 따라서 이를 위해 다음 단계에서는 연이어 나타난 개체명 후보에 대한 처리를 통해 범위와 범주를 결정해 줄 수 있는 어절 관계 규칙을 이용한다. 이를 다음 3.5절에서 설명한다.

3.5 단계 4 : 개체명간 관계 정보

이 단계에서는 앞의 세단계에서 결정된 결과를 보고 어절 간의 관계를 이용하여 개체명의 범위를 결정하고, 인접해 나타난 개체명의 범주를 보정하거나 새로이 가중치를 부여한다.

그림 5는 이 단계에서 이용하는 어절 관계 규칙의 일부를 나타낸 것이다. 접속격 조사 '와'의 양쪽에 개체명이 나올 경우 이 두 개체명은 일반적으로 같은 범주를 가지게 된다. 따라서 규칙 (9)는 이런 특성을 이용하여 이미 범주가 결정된 개체명과 아직 결정되지 않은 개체명 후보가 접속격 조사로 결합되어 사용되었을 때 개체명 후보의 범주를 결정해준다. 규칙 (10)은 조직명과 인명 사이의 일반적인 출현 규칙이다. '인명 + 조직명 + 지위'가 나오면 이를 하나로 묶어 인명으로 인식하는 것이다.

패턴	자질값	범위
(9) 개체명 + 와/과[접속격 조사(jcj)] + 개체명후보	개체명의 자질값	3 3
(10) (Wnq | PER) + ORG + PER_POS	PER	1 3
(11) ORG_GOV + LOC	LOC	1 2

그림 5: 어절 관계 규칙(일부)

이러한 규칙 정보와 앞단계에서 결정된 개체명의 범주 정보를 이용하여 중의성이 있는 경우나 결정되지 않은 개체명의 범주를 명확히 할 수 있고 이어서 나타나는 개체명들의 관계를 고려하여 개체명의 범위를 결정할 수 있다.

예를 들어 앞에서 제시한 예문 (3)에서 '캐나다'와 '밴쿠버'는 연이어서 나타나고 각각 조직명과 지명으로 인식되었다. 이 단계에서는 규칙 (11)을 이용하여 '캐나다 밴쿠버'는 하나의 개체명으로, 개체명의 범주는 지명으로 인식한다.

표 6: 단계 4까지의 결과

단어	품사	인명	지명	조직명	자질
미국	nq	-1.0	1.5	4.0	ORG_GOV
캐나다 밴쿠버	nq+nq	0	2.0	0	LOC
캐나다	nq	-1.0	1.5	2.0	ORG_GOV
이한동 대표	nq+ncn	2.0	-1.0	-1.0	PER
청와대	nq	0	2.0	0	LOC
국정 개혁 위원회	ncn+ncn+ncn	-1.0	1.5	2.0	ORG

단계 4까지 각 단계를 거치면서 개체명의 각 범주에 부여한 가중치는 표 6과 같으며, 추출된 개체명은 표 7과 같다.

표 7: 최종 개체명 인식 결과

개체명	범주
미국	조직명
캐나다 밴쿠버	지명
캐나다	조직명
이한동 대표	인명
청와대	지명
국정개혁위원회	조직명

4 실험 및 결과

이 절에서는 실험을 통해 본 논문에서 제시한 사전과 규칙을 기반으로 한 개체명 인식기의 성능을 평가하고 단계별 정보의 기여도에 대해 분석한다.

4.1 실험 환경

본 논문의 실험을 위해 품사 태깅된 "KAIST 말뭉치[6]" 중에서 신문 기사 분야를 선택해서 사용했다.

학습 데이터는 1,338 문장으로 구성되어 있으며, 실험 데이터로는 학습 데이터와는 다른, 정답 개체명이 태깅된 10개의 기사(190문장)를 이용하였다. 본 논문을 위한 학습 데이터와 실험 데이터에 대한 통계 값은 각각 표 8, 표 9와 같으며, 문장당 평균 1.71개의 개체명이 나타나고 있다.

표 8: 학습 데이터의 통계값

항목	수
문장	1338개
어절	19846개
문장당 어절	14.8개

표 9: 실험 데이터의 통계값

항목	수	항목	수
문장	190개	인명	99(0.52)개
어절	3031개	조직명	169(0.88)개
문장당 어절	15.9개	지명	58(0.31)개
		개체명	326(1.71)개

괄호 안의 수는 문장당 각 개체명의 수이다.

개체명 인식을 위해서는 우선 품사 태깅된 코퍼스를 입력으로 받아 품사 정보를 이용하여 어절 단위로 후보를 추출한다. 이와 같이 추출된 후보 단어에 대해 앞에서 제시한 네단계의 규칙을 이용하여 개체명의 범주를 인식한다. 이 때 각 단계에서 적용되는 규칙은 개체명의 각 범주에 가중치를 부여하게 된다. 이를 위해서 문장으로부터 추출된 후보 단어열에 대해 각 범주에 해당하는 점수를 저장할 수 있는 자료 구조를 만들어 이용한다. 따라서 각 단계는 분리되어 있어 단계별 실행이 가능하고, 각 단계는 후보 단어열의 자료 구조로 연결된다.

4.2 실험 결과

본 논문에서 제안한 방법을 이용한 개체명 인식기를 사용하여 4.1절에서 제시한 실험 데이터를 실행시킨 결과는 표 10과 같으며, 실험에 대한 평가 기준으로는 정확률(precision)과 재현율(recall)을 이용하였다. 실험 결과 개체명 전체에 대해 90.4%의 정확률과 83.4%의 재현율을 얻었다.

표 10: 개체명 인식 결과

	정확률	재현율
인명	86.2%	89.8%
조직명	92.2%	84.5%
지명	93.3%	70.0%
전체	90.4%	83.4%

[6]"KAIST 말뭉치"는 문화관광부와 과학기술부, 과학기술정책관리연구소의 지원을 받아 구축되었다.

표 11: 영어 개체명 인식 결과

조건	정확률	재현율
BASE	96%	96%
ALLCAPS	87%	82%

표 11은 [13]에서의 영어 개체명 인식 결과로, 영어권에서 가장 좋은 성능을 보이고 있는 개체명 인식기 중의 하나이다. 표 11의 결과를 보면 BASE의 경우는 96%의 정확률과 96%의 재현율을 보인다. 그러나 대소문자 구분 정보를 무시한 ALLCAPS의 경우에는 87%의 정확률과 82%의 재현율을 보이는데 그쳤다. 이는 한국어와 같이 대소문자 구분 정보가 없는 경우 개체명을 인식하는 것이 훨씬 어려움을 보여준다.

본 논문의 경우 실험을 위한 분야와 데이터가 다르고 특히 제한된 영역(신문 기사)을 대상으로 했기 때문에 객관적인 비교를 할 수는 없지만, 한국어는 영어권 개체명 인식에 비해 대소문자 구분 정보가 없는 등의 어려움이 있다는 것을 고려한다면 본 논문에서 제시한 방법을 이용한 한국어 개체명 인식기는 어느 정도 좋은 성능을 보이고 있음을 알 수 있다.

4.3 오류 분석

개체명 인식의 오류는 개체명을 인식하지 못하거나 잘못된 범주를 인식하는 것으로, 대부분 정보의 부족에 기인하였다. 이는 크게 사전 정보가 부족한 경우와 용언의 하위 범주화 정보가 부족한 경우로 나누어 볼 수 있다.

- **사전 정보의 부족**
  - **약어 표현** '전경련', 'YS'와 같은 약어 표현의 경우 사전에 등록되어 있지 않은 경우가 빈번하게 발생하며 이러한 개체명들이 보충 설명 없이 사용되면 문맥 규칙이나 용언의 선택 제약 규칙을 이용하여 범주를 인식해 주어야 한다. 그러나 이러한 정보들의 부족으로 인해 개체명으로 인식되지 못하는 오류가 발생한다.
  - **용언의 하위범주화 정보의 중의성** 용언의 하위범주 정보의 중의성은 용언이 하위범주화 하는 명사의 자리에 여러 가지 범주를 가질 수 있기 때문에 발생한다. 예를 들어 '조사하다'하는 용언은 '[인명, 조직명]이 [인명, 조직명]를 조사하다'와 같이 인명과 조직명을 동시에 하위범주화 한다. 따라서 어절 내 정보나 주변 문맥 정보가 없는 경우 잘못된 범주를 인식하게 된다.
- **용언의 하위범주화 정보 부족** 용언의 하위범주화 정보는 용언 정보를 통해 문장에서 사용되는 개체명의 범주를 인식하게 함으로써 중의성 해결에 큰 역할을 한다. 그러나 이러한 용언의 하위범주화 정보의 부족으로 다음과 같은 중의성 문제를 갖게 된다.
  - **조직명과 지명, 조직명과 상품명** 일반적으로 조직명과 지명은 각각 주체로 사용되는지 아니면 장소로 사용되

는지에 따라 중의성을 가진 단어가 많이 존재한다. 예를 들어 '한국'이나 '서울시'는 각각 주체로 사용되면 조직명이지만 장소의 의미로 사용되면 지명이다. 또한 'OB맥주'나 '삼보 컴퓨터'와 같은 상품명은 보통 상품을 의미하기도 하지만 조직명으로도 사용된다.

- **결합 단어** 개체명의 범주 인식을 위해 고려하는 정보 중의 하나인 결합 단어 자체가 중의성을 가지고 있는 경우가 있다. 예를 들어 '지사'는 사람의 지위를 나타내기도 하지만 회사를 나타내는 조직명으로 사용되기도 한다.

- **보통 명사의 결합** 대소문자 구분 정보를 이용하는 영어권의 언어에서는 보통 명사의 결합인 경우에도 대소문자 구분 정보를 이용하여 개체명을 구분할 수 있다. 그러나 한국어에서는 이러한 정보가 없기 때문에 개체명의 추출 오류가 발생한다. 예를 들어 '시중 은행'은 '은행'이라는 조직을 나타내는 명사와 같이 사용되었지만 개체명이 아닌 반면에 '국민 은행'은 조직을 나타내는 개체명으로 쓰일 수 있다. 이러한 경우 국민 은행이 개체명 사전에 등록되어 있다면 개체명으로 쉽게 구별할 수 있다. 그러나 '국민 은행'이라는 단어가 보통 명사를 의미하는 단어로도 쓰일 수 있다. 따라서 특정 은행의 이름인지 보통 명사인지 정확히 구별하기 위해서는 본 논문에서 사용하는 정보 외에 다양한 정보가 필요하다.

4.4 단계별 정보의 기여도 분석

본 논문에서는 3절에서 설명한 바와 같이 네단계를 거쳐 개체명을 인식한다. 이러한 네단계에서 사용되는 규칙과 정보가 개체명 인식의 성능에 어느 정도의 영향을 미치고, 각각 어떤 역할을 하는지 분석하기 위해 두 가지의 단계별 실험을 수행하였다.

표 12는 단계 1부터 각 단계까지 수행한 결과이다. 예를 들어 단계 3 실험은 단계 1부터 단계 3까지 수행한 결과이다.

단계 1은 사전 정보를 이용하여 단어가 사전에 있거나 일부 단어가 개체명의 범주를 인식할 수 있는 단어인 경우를 인식한다. 따라서 사전에 개체명 자체가 등록되지 않은 경우에도 '은행'이나 '위원회'와 같이 단어의 구성이 되는 단어를 보고 범주를 판단할 수 있는 조직명의 경우 이 단계에서 높은 재현율을 보인다. 단계 2는 개체명의 주변에 나타나는 단어의 정보를 보고 범주를 인식하기 때문에 지위나 호칭 등의 결합 단어 정보를 많이 가지고 있는 인명의 경우 이 단계에서 높은 재현율을 얻을 수 있다. 반면에 지명의 경우에는 이러한 단어 정보의 중의성이 많아 이용하기 어렵고 조직명의 경우는 거의 나타나지 않는 경향을 보인다. 단계 3은 용언과 보어가 되는 명사의 하위범주 정보를 이용하는 것으로 주로 조직명과 지명의 중의성을 해소하고 사전 정보가 없는 개체명을 인식하는 역할을 한다. 단계 3에서 이러한 용언의 하위범주화 정보를 이용해 중의성을 해소하여 지명과 조직명의 정확률과 재현율이 상승함을 볼 수 있다. 마지막으로 단계 4는 기존에 결정된 개체명의 범주를 이용하여 범위를 결정하고 연이어 나타난 개체명의 범주 결정에 참여한다. 따라서 이 경우에 범위를 결정해주고 앞뒤에 올 수 있는 개체명의 범주를 통해 오류 보정이 일어남으로써 전반적인 성능 상승을 보인다.

표 13은 각 단계의 필요성을 보기 위한 것으로 단계 1부터 단계 4까지 각 단계를 제외하고 수행시킨 결과이다. 즉, 단계 1을 제외한 실험이란 단계 2, 단계 3, 단계 4만을 수행한 결과를 의미한다.

표 13: 각 단계를 제외하고 실험한 결과

	정확률	재현율
단계1 제외	76.1%	15.4%
단계2 제외	77.1%	57.1%
단계3 제외	80.1%	63.1%
단계4 제외	81.4%	73.7%
모든 단계 포함	90.4%	83.4%

단계 1을 제외한 경우는 사전과 단어 자체의 정보를 이용하지 못하기 때문에 극히 낮은 재현율을 보인다. 따라서 기본적인 사전의 구축은 필수적이라 할 수 있다. 또한 단계 2부터 4까지 제외한 경우도 모든 단계를 포함한 실험 결과와 비교해 볼 때 10%이상의 정확률 저하를 보이고 재현율의 경우에는 그 이상의 차이를 보이고 있다. 따라서 각 단계는 개체명 인식을 위해 유용한 정보를 제시하고 있음을 알 수 있다.

5 결론

개체명은 일반적으로 대부분이 사전에 등록되지 않은 고유 명사이고, 사전에 등록된 단어일 경우에도 문맥에 따라 다르게 사용될 수 있다. 또한 한국어는 대소문자 구분 정보가 없고 단어 간에 결합해서 사용하는 특성 때문에 개체명의 후보를 찾고 범위를 결정해야 하는 문제점이 있다.

본 논문에서는 개체명 사전과 결합 단어 사전, 그리고 용언의 하위범주화 사전을 이용하는 규칙 기반 개체명 인식 방법을 제안하였다. 각 규칙은 이용하는 정보에 따라 네단계로 나누어 순차적으로 적용된다. 첫번째 단계에서는 어절 내의 단어 정보를 보고, 두번째 단계에서는 제한된 주변 문맥 정보를 본다. 그리고 세번째 단계에서는 용언의 하위범주화 정보와 개체명과의 관계를 보고, 마지막으로 네번째 단계에서는 개체명 간의 관계 정보를 고려한다.

본 논문에서 제안한 규칙을 이용한 개체명 인식기의 성능을 평가하기 위해 실험을 수행한 결과 90.4%의 정확률과 83.4%의 재현율을 얻었다.

그리고 본 논문에서 제시한 각 단계에서 사용하는 규칙의 기여도 평가를 위해 두가지의 단계별 실험을 하였다. 먼저 단계 1에서 각 단계까지의 실험을 한 결과 어절 내 단어 정보를 이용한 단어 구성 규칙은 조직명을 인식하는데 중요한 역할을 하고 제한된 주변 문맥 정보를 이용한 문맥 규칙의 경우 인명 인식에 큰 영향을 끼쳤다. 그리고 용언의 하위 범주화 정보를 이용한 규칙은 조직명과 지명의 중의성을 해소하고 인식되지 못한 조직명과 지명을 인식하는데 주된 역할을 했다. 마지막으로 개체명간 관계 정보를 고려한 어절 관계 규칙은 개체명의 범위를 결정하고 같이 올 수 있는 범주의 고려를 통해 전반적인 성능 향상을 가져왔다. 또한 각 단계의 필요성을 평가하기 위해 각 단계를 제외한 실험

표 12: 각 단계까지 차례로 실험한 결과

	단계1		단계2까지		단계3까지		단계4까지	
	정확률	재현율	정확률	재현율	정확률	재현율	정확률	재현율
인명	72.9%	54.5%	79.3%	77.8%	80.0%	80.8%	86.2%	89.8%
조직명	78.3%	70.8%	78.3%	70.8%	84.1%	78.6%	92.2%	84.5%
지명	75.9%	36.7%	78.1%	41.7%	88.9%	53.3%	93.3%	70.0%
전체	76.5%	59.6%	78.6%	67.6%	81.4%	73.7%	90.4%	83.4%

을 하였고 이를 통해 각 단계가 개체명 인식을 위해 중요함을 알 수 있었다.

앞으로 개체명 사전과 용언의 하위범주화 사전을 체계적으로 구성하고 다양한 문맥 정보를 확장시킬 뿐 아니라 용언과 명사와의 관계를 알 수 있도록 성능이 좋은 구문 분석기를 이용한다면 더욱 좋은 결과를 얻을 수 있을 것이다. 또한 사람이 규칙을 추출하는 것은 대상 영역에 의존적일 수 밖에 없고 시간이 많이 걸리는 작업이므로 규칙을 자동으로 추출할 수 있는 방법의 연구가 이루어져야 한다. 그리고 인명, 지명, 조직명 외의 기타 개체명 인식에 대한 연구가 필요하다.

참고문헌

[1] 김현진, 박세영, 장명길, 박재득, and 박동인. 용언의 구문 관계를 이용한 명사 분류. In 한글 및 한국어 정보처리 학술발표논문집, pages 111–115, 1997.

[2] 양장모, 김민정, and 권혁철. 언어 정보를 이용한 한국어 미등록어 추정. In 한국정보과학회 봄 학술발표논문집, pages 957–960, 1996.

[3] 유현경 and 이선희. 격조사 교체와 의미역, chapter 격조사 교체와 의미역, pages 129–171. 태학사, 1996.

[4] 정래정 and 김준태. 고유 명사 출현 패턴을 이용한 색인의 성능 향상에 관한 연구. In 한글 및 한국어 정보처리 학술발표논문집, pages 68–72, 1996.

[5] 부산대학교 언어교육연구원. 한국어 문장 분석을 위한 용언의 하위 범주화에 관한 연구. Technical report, 시스템공학연구소, May 1997.

[6] 홍재성 외. 현대 한국어 동사 구문 사전. 두산동아, 1997.

[7] 이경희. 한국어 문서에서 개체명 인식에 관한 연구. Master's thesis, KAIST, 2000.

[8] Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34(1/2/3):211–231, February 1999.

[9] William J Black, Fabio Rinaldi, and David Mowatt. Facile:description of the ne system used for muc-7. In *Proceedings of Seventh Message Understanding Conference (MUC-7)*, 1998.

[10] Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. Nyu : Description of the mene named entity system as used in muc-7. In *Proceedings of Seventh Message Understanding Conference (MUC-7)*, 1998.

[11] Hsin-Hsi Chen, Yung-Wei Ding, Shih-Chung Tsai, and Guo-Wei Bian. Description of the ntu system used for met2. In *Proceedings of Seventh Message Understanding Conference (MUC-7)*, 1998.

[12] J. Fukumoto, F. Masui, M. Shimohata, and M. Sasaki. Oki electric industry : Description of the oki system as used for muc-7. In *Proceedings of Seventh Message Understanding Conference (MUC-7)*, 1998.

[13] George R. Krupka. Sra: Description of the sra system as used for muc-6. In *Proceedings of Sixth Message Understanding Conference (MUC-6)*, pages 221–235, 1995.

[14] Andrei Mikheev, Claire Grover, and Marc Moens. Description of the ltg system used for muc-7. In *Proceedings of Seventh Message Understanding Conference (MUC-7)*, 1998.

[15] Satoshi Sekine, Ralph Grishman, and Hiroyuki Shinnou. A decision tree method for finding and classifying names in japanese texts. In *Proceedings of Sixth Workshop on Very Large Corpora*, pages 171–178, 1998.

[16] Francis Wolinski, Frantz Vichot, and Bruno Dillet. Automatic processing of proper names in texts. In *proceedings, Seventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–30, 1995.