제 18 회 한국정보처리학회 추계학술발표대회 논문집 제 9 권 제 2 호 (2002.11)

# Cotraining 학습을 이용한 한국어 개체명 인식

이현숙*, 정의석*, 황이규*, 윤보현*
*한국전자통신연구원 휴먼정보처리연구부 지식처리연구팀
e-mail : {lhs63473, eschung, yghwang, ybh}@etri.re.kr

# Korean Named Entity Recognition using Cotraining-based Learning

Hyun-Sook Lee*, Euisok Chung*, Yi-Gyu Hwang*, Bo-Hyun Yun*
*Knowledge Processing Research Team, Human Information Processing Dept.
Electronics and Telecommunications Research Institute

**요 약**

본 논문에서는 정보추출 및 정보검색, 문서요약과 같은 자연어처리 응용에서 중요한 역할을 하는 개체명 인식 모델을 제안하였다. 기존의 한국어 개체명 인식에 관한 연구는 규칙 기반 연구의 경우 수동으로 생성한 규칙이나 어휘사전에 매우 의존적이고, 통계기반의 연구의 경우 개체명이 태깅된 대량의 학습데이터를 필요로 하므로 새로운 도메인으로의 이식성 관점에서 한계가 있다. 이를 극복하기 위해 본 논문에서는 개체명이 태깅되지 않은 학습데이터를 이용하여 Cotraining 기반 학습을 수행함으로써 개체명 인식을 위한 규칙과 사전을 자동적으로 확장하였다. 실험 결과, 경제분야 문서에 대해 87.6%의 정확률을 보였다.

## 1. 서론

개체명 인식은 문서에서 고유한 의미를 가지는 개체명을 찾아내고 인명, 지명, 기관명 등과 같은 개체명 범주를 할당하는 것이다. 이는 정보추출 뿐만 아니라 정보검색, 기계번역과 같은 다양한 자연어처리 응용을 위한 전처리 과정으로도 중요하다.

개체명의 범주는 인명, 지명, 기관명과 같은 고유명사와 날짜, 시간, 화폐 등의 수치 표현이 있다. 수치 표현은 비교적 단순한 문법에 의해 효과적으로 인식될 수 있다. 그러나 고유명사는 새로운 고유명사가 계속적으로 만들어지고 그 형태가 가변적이기 때문에 사전에 등록되지 않은 개체명을 인식하기 어렵다. 또한 동일한 개체명이 문맥에 따라 다른 범주로 사용될 수 있다는 문제점이 있다.

한국어에서의 개체명 인식에 관한 연구는 대부분 규칙 기반으로 이루어졌고, 통계 기반 연구의 경우에도 개체명이 태깅된 학습데이터를 이용하는 교사 학습 방법이 주로 이용되었다. 그러나 규칙 기반의 연구는 수동으로 규칙을 생성하고 어휘 사전에 의존적이므로 새로운 도메인으로의 이식성 관점에서 제약을 받는다. 교사학습 기반의 연구는 개체명이 태깅된 대량의 학습데이터의 생성이 요구되고, 영어와 달리 대문자와 같은 문자형 정보가 부족하므로 한국어에 적절한 자질을 찾아내기 어렵다는 단점이 있다.

본 논문에서는 한국어 개체명 인식에서의 이러한 한계점을 극복하기 위해 Cotraininig 기반 학습을 이용한 개체명 인식 방법을 제안한다. Cotraining 모델은 동일한 데이터를 서로 다른 뷰(View)로 나누고 각 뷰를 위한 학습기가 다른 뷰를 지원하도록 하는 학습 기법이다. 즉, 하나의 분류기에 의해 학습된 인스턴스(Instance)들이 다른 분류기에 의해 사용되도록 함으로써 학습의 범위를 확장시키는 상승효과를 일으킨다. 본 연구에서는 개체명이 태깅되지 않은 학습데이터를 이용하여 Cotraining 학습을 수행함으로써 수동으로 개체명 태깅 학습데이터를 생성하는 부담을 줄이고, 개체명을 위한 규칙과 사전을 자동으로 확장함으로써 어휘사전과 규칙에 의존적인 규칙기반의 개체명 인식의 단점을 해결하고자 한다. 2 장에서는 개체명 인식의 관련연구에 대해서 살펴보고, 3 장에서는 Cotraining 기반의 학습을 통해 개체명을 인식하는 모델을 설명한다. 4 장은 실험 결과에 대해 기술하고, 5 장에서는 결론과 향후 연구를 보인다.

## 2. 관련연구

개체명 인식을 위한 접근방법은 크게 규칙 기반의 개체명 인식과 통계 기반의 개체명 인식, 그리고 두 가지 방법을 통합한 Hybrid 방식으로 나누어 볼 수 있다. 규칙 기반의 방법은 개체명 인식을 위한 규칙을 수작업으로 구축하고, 개체명 사전, 개체명 구성단어 사전, 개체명 문맥단어 사전 등을 이용하여 개체명을 추출하는 방법이다[1]. 통계에 기반한 방법은 학습 데이터로부터 개체명 인식에 필요한 지식을 자동적으로

학습하는 방법으로, 주로 철자, 품사, 형태소로부터 얻어낸 정보를 이용하여 개체명 인식을 위한 규칙을 학습한다.[2] 대표적인 방법으로는 결정트리를 이용한 방법, HMM 을 이용한 방법, 최대 엔트로피를 이용한 방법이 있고 개체명이 태깅되지 않은 코퍼스를 이용하는 비교사 학습 방식이 있다. Hybrid 방식은 규칙 기반의 방법과 통계 기반의 방법을 혼합하여 좀더 나은 성능을 얻기 위한 목적으로 통계 기반의 모델에 규칙이나 어휘 정보, 사전 정보 등의 다양한 지식들을 결합하는 방식이다.[3]

한국어 개체명 인식에 관한 연구로는 다양한 사전과 규칙을 이용하는 연구와 기계학습을 이용한 연구가 있었다. 연구 [4]에서는 개체명 사전, 개체명의 구성단어 사전, 개체명의 문맥단어 사전, 용언의 하위범주화 사전 등의 다양한 사전과 개체명의 구성 규칙, 문맥 규칙, 개체명 간의 결합규칙, 용언과 개체명 간의 규칙 등을 이용하였다. 이 연구는 90.4%의 정확률과 83.4%의 재현율을 보인다. 그러나 어휘 사전과 규칙에 매우 의존적인 방법으로 이식성 문제와 어휘 사전 및 규칙이 고정적이거나 혹은 수동으로 확장되어야 한다는 단점이 있다. 연구 [5]에서는 새로운 도메인으로의 이식성을 위해 사전과 비교적 단순한 패턴 규칙을 이용하여 학습 코퍼스에서 개체명의 내부 어휘 또는 문맥 어휘를 학습하였다. 이 연구는 이미 정해진 단순한 패턴에 의해 학습이 이루어지고 개체명이 태깅된 대량의 학습데이터를 필요로 한다는 단점이 있다. 연구 [6]은 다양한 사전과 패턴 규칙을 이용하여 개체명의 후보를 선택하고 미등록어 처리를 위해 최대엔트로피 모델을 이용하여 신경망으로 모호성을 해소하고자 하였다. 이 연구에서는 F-measure 80.27%의 결과를 보이지만 개체명이 태깅된 학습데이터의 생성이 요구되고, 세분화된 어휘 사전에 의존적이라는 단점이 있다.

## 3. Cotraining 기반 개체명 인식

Cotraining 기반 학습의 특징은 개체명이 태깅되지 않은 학습 코퍼스를 이용하여 기본적인 Seed[1] 데이터만으로 자동 학습이 가능하다는 것이다.개체명 인식을 위한 Cotraining 기반의 학습 방법은 개체명의 자질을 두 가지 뷰로 나누고 각 뷰에 대한 학습기를 확장해 나가는 것이다. 학습기의 입력은 몇 개의 Seed 데이터와 개체명이 태깅되어 있지 않은 학습데이터이다. 학습기의 규칙 집합은 Seed 데이터로 초기화된다. 학습 단계에서는 먼저, 하나의 뷰에 대한 학습기가 현재의 규칙 집합을 이용하여 학습데이터의 개체명을 그것의 범주로 분류한다. 그 결과로부터 즉, 학습데이터에서 의미 범주가 정해진 개체명으로부터 또 다른 뷰에 해당하는 자질들을 추출한다. 그 중 가장 신뢰할만한 자질을 선택하여 그 뷰에 대한 학습기의 규칙 집합을 확장한다. 하나의 뷰에 대한 학습이 완료되면 또 다른 뷰에 대해 위와 동일한 방법으로 학습을 수행한다. 이와 같은 과정을 반복하여 더 이상 새로운 규칙이 추가되지 않을 때까지 점차적으로 학습기를 확장하고 최종적으로 두 학습기를 결합한다.

### 3.1 어휘 정보를 이용한 Cotraining 기반 개체명 인식

한국어 개체명의 특징은 첫째, 삼성전자, 계룡산, 서울문화회관과 같이 개체명(삼성, 계룡, 서울)과 개체명 인식의 단서가 되는 단어(전자, 산, 회관)로 이루어진 개체명이 많다는 것이다. 둘째, 개체명의 문맥이 개체명 인식의 단서를 제공한다는 것이다. 예를 들면, '홍길동씨', '낙동강 일대'의 경우, 개체명 '홍길동'의 문맥인 '씨'가 인명의 단서가 되고, 개체명 '낙동강'의 문맥인 '일대'가 지명의 단서가 된다. 이러한 특징을 이용하여 어휘 정보를 이용한 Cotraining 기반의 개체명 인식을 수행할 수 있다. 이를 위해서는 먼저 개체명의 자질을 철자자질과 문맥자질로 나눈다[7]. 철자자질은 개체명 후보의 문자열을 대상으로 문자형 정보 또는 전체문자열, 부분문자열 등을 자질로 추출한다. 문맥자질은 개체명 후보의 왼쪽 또는 오른쪽 문맥을 대상으로 문맥 단어의 위치, 문자열 등이 자질로 추출된다. 예를 들면, '최근 소설가 최인호씨는 그의 작품인 ...'에서 개체명의 문자열인 '최인호'를 대상으로 철자자질인 'Full_String=최인호', 'Contain(최)', 'Contain(인호)'를 추출하고, 개체명의 문맥인 '소설가'나 '씨'를 대상으로 문맥자질인 'Left_Context(소설가)', 'Right_Context(씨)'를 추출할 수 있다. 이러한 철자자질과 문맥자질을 번갈아 적용하여 개체명을 인식함으로써 규칙을 확장하고 점점 더 많은 개체명을 인식해 나간다.

어휘 정보를 이용한 Cotraining 기반의 개체명 인식이 한국어 개체명 인식에 어느 정도 적절한지를 알아보기 위해 HANTEC 신문기사 38,000 문서와 Seed 규칙 50 개를 이용하여 실험을 수행해 보았다. 최종 학습 결과로 나온 규칙을 이용하여 1000 개의 개체명에 대해 평가한 결과 Correctness[2] 43%를 얻었다.

위의 실험을 통해 분석해 본 결과, 어휘 정보만을 이용하는 개체명 인식에서 다음과 같은 문제점들을 확인할 수 있었다. 첫째, 연속된 명사, 접사의 배열로 개체명 후보를 선택하여 학습하기 때문에 많은 일반 명사가 개체명 후보로 선택되어 정확한 학습이 이루어지지 않는다. 둘째, 한국어에서의 문맥은 위치 가변적인 경우가 많아서 특정 문맥 단어의 출현빈도가 여러 위치에 분산되어 나타난다. 셋째, 자질 하나로는 개체명을 인식하는 확실한 규칙을 만들기 어렵다. 이러한 문제점들을 극복하기 위해 다음 절에서는 패턴 규칙을 이용하는 Cotraining 기반의 개체명 인식 방법을 제시하고자 한다.

### 3.2 패턴규칙을 이용한 Cotraining 기반 개체명 인식

#### (1) 패턴규칙의 필요성

개체명 인식을 위한 학습에서 패턴규칙을 이용하는 것은 어휘정보를 이용한 개체명 인식의 한계를 해결하는데 도움을 준다. 첫째, 개체명의 후보를 결정할

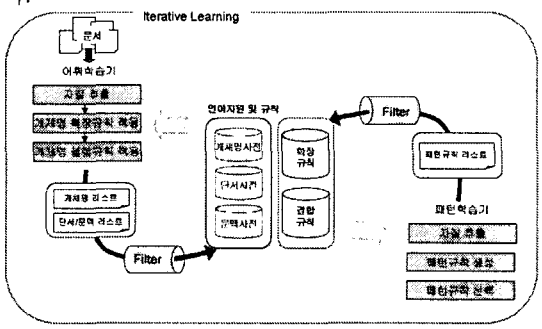
[1] 학습 이전에 미리 주어지는 데이터

[2] Correctness = (바르게 인식된 개체명)/(전체 개체명)

때 패턴규칙을 적용함으로써 보다 양질의 후보를 선택할 수 있다. 둘째, 위치 가변적인 문맥을 패턴규칙으로 인식함으로써 규칙의 모호성을 해결하는데 도움을 준다. 예를 들어, '선경증권 박도근 사장'과 '럭키 사장 성재갑씨'의 경우, 인명인 '박도근', '성재갑'의 문맥으로 출현하는 '사장'이 인명의 오른쪽 또는 왼쪽에 위치한다. 이러한 경우, 앞의 예는 패턴규칙 [<ORG>+<PER>+<POS>][3]에 의해, 뒤의 예는 패턴규칙 [<ORG>+<POS>+<PER>+<HOR>][4]에 의해 개체명을 인식할 수 있다. 셋째, 패턴규칙은 문맥의 단어 대신 문맥의 의미범주를 고려함으로써 개체명 인식의 범위를 확장할 수 있다. 예를 들면, '강봉균 한국개발연구원 원장'에서 패턴규칙 [<ORG>+<PER>+<POS>]에 의해 '원장' 대신 '사장' 혹은 '과장' 등 직위에 해당하는 여러 다른 단어가 출현할 때도 개체명을 인식할 수 있다. 넷째, 특정 어휘에 의존적인 개체명 인식이 가능하다. '정태기 전 GT 웹코리아 사장'과 같이 개체명 인식의 단서라고 보기 어려운 '전' 과 같은 특정 어휘와 개체명이 함께 등장하는 경우에 패턴규칙 [<PER>+<STRING [5] (전)>+<ORG>+<POS>]를 이용하여 개체명을 인식할 수 있다.

### (2) 시스템 구성도

본 논문에서 제안한 시스템의 구조는 [그림 1]과 같다.




[그림 1] 시스템 구성도

어휘학습기에서는 여러 문서가 입력으로 주어지면 문서로부터 자질을 추출하고 패턴규칙을 적용하여 개체명 및 특정 의미범주에 해당하는 어휘들을 인식한다. 즉, 사전에 있는 개체명 및 단서단어, 문맥단어들을 문서에서 찾아낸다. 이렇게 인식된 개체명과 어휘들 중 신뢰도가 높은 것들을 선택하여 개체명사전, 단서사전, 문맥사전을 확장한다. 패턴학습기에서는 앞의 단계에서 확장된 사전을 이용하여 새로운 패턴규칙을 탐색한다. 새로 발견된 각 패턴규칙이 추출하는 어휘와 의미범주가 얼마나 정확히 일치하느냐에 따라 확장될 패턴규칙을 선택한다.

입력문서에 두 모듈을 번갈아 적용하면서 Cotraining 학습을 통해 패턴규칙과 어휘사전 사이에 상승효과를 일으키면서 개체명 인식의 범위를 확장해 나간다. 즉, 개체명 인식을 위한 패턴규칙을 이용하여 어휘사전을 확장하고, 어휘사전을 이용하여 새로운 패턴규칙을 생성해 나간다.

### (3) 자질 유형

본 논문에서 제안한 모델에서의 자질은 크게 개체명 내부 자질과 외부 자질로 나눌 수 있다. 개체명의 내부 자질에는 문자형 자질, 단서자질, 사전자질이 있다. 문자형 자질은 개체명이 숫자나 한자, 특수문자를 포함하는지 여부를 나타내는 자질로 [표 1]과 같다. 단서자질은 [표 2]과 같이 개체명을 구성하는 형태소 중 개체명 인식의 단서가 되는 자질이다. 사전자질은 인명, 지명, 기관명 등의 개체명 사전에 해당하는 자질로서 [표 3]와 같다.

| Word Feature | | Example | Description |
|---|---|---|---|
| Chinese | OneChinese | 英, 中 | Country Name |
| | ThreeChinese | 崔錫泰 | Person Name |
| | ContainsOneChineseAndLetter | 삼성物 | Organization Name |
| Alphabet | ContainsAlphaAndLetter | LG경제연구원 | Organization Name |
| | AllCapitalization | IBM, OECD | Organization Name |
| Letter | TrenLetters | 정주영 | Person Name |

[표 1] 문자형 자질

| NE Category | Clue Feature | Example | Description |
|---|---|---|---|
| 지명(73) | DistrictClueLOC | 시, 군, 특별시, 광역시, … | 행정구역을 나타내는 형태소나 형태소의 결합 |
| | ClueLOC | 강, 산, 암, 령, 회관, … | 지명으로 분류하는데 결정적인 형태소나 형태소의 결합 |
| 기관명(140) | ClueORG | 위원회, 건설, 텔레콤, … | 기관명으로 분류하는데 결정적인 형태소나 형태소의 결합 |

[표 2] 단서자질

| NE Category | Dictionary Feature | Example | Description |
|---|---|---|---|
| 인명 | DicPERSON(43,300) | 김달러오, … | 유명인, 인명 등 |
| 지명 | DicLOC(16,471) | 한국, 미국, 서울, … | 나라이름, 행정구역, 산 이름 등 |
| 기관명 | DicORG(1,740) | 삼성, 대우증공업, LG, … | 기관명 이름 |

[표 3] 사전자질

개체명의 외부 자질은 지역문맥자질과 전역문맥자질이 있다. 지역문맥자질은 [표 4]와 같이 개체명의 문맥에 출현하면서 개체명 인식의 단서가 되는 자질로서 '송병락 교수'에서 '교수'나 'LG 텔레콤 사장'에서 '사장' 등이 그 예이다. 전역문맥자질은 전체 문서에서 이미 인식된 개체명을 이용하는 자질로서 시스템에서 인식된 개체명 리스트를 유지해야 한다.

| NE Category | Local Context Feature | Example | Description |
|---|---|---|---|
| 인명 | PosPERSON(105) | 부장, 차장, 원장, 위원, 이사장 | 직함 또는 지위 |
| | RelPERSON(150) | 아들, 형수, 남편, 친구, … | 관계 |
| | JobPERSON(310) | 매니저, 작가, 교수, 가수, … | 직업 |
| 지명 | LContLOC(45) | 근교, 기슭, 방면, … | LOC 인접명사 |
| 기관명 | LContORG(30) | 회원, 단원, 등문, … | ORG 인접명사 |

[표 4] 지역문맥자질

### (4) 패턴규칙 유형

패턴규칙 유형은 개체명 확장규칙과 개체명 결합규칙으로 구분할 수 있다.

개체명 확장규칙은 개체명의 구성에 관한 규칙으로 (1)의 기관명을 구성하는 규칙의 예를 보면, 기관명이 기관명 사전에 있는 단어와 기관명의 단서가 되는 단어로 이루어짐을 알 수 있다. '대우전자', 'LG 텔레콤',

[3] <ORG> : 기관명, <PER> : 인명, <POS> : 지위
[4] <HOR> : 호칭
[5] <STRING> : 특정 어휘

'삼성물산'등이 그 예이다.

<ORG> := <ORG, DIC> + <ORG, CLUE>[6] (1)

개체명 결합규칙은 개체명과 문맥 간의 의존 관계를 나타낸 규칙으로 개체명 간 결합규칙, 병렬형 결합규칙, 어휘기반 결합규칙이 있다.

<ORG> + <PER> + <POS> (2)
[<LOC> + ',']$^{+}$ + <LOC> + <STRING(등)> (3)
<LOC> + <STRING(태생)> (4)

개체명 간 결합규칙은 '서울대 송병락 교수'의 예에서처럼 인명, 지명, 기관명 등의 개체명들 간의 결합 패턴을 나타낸 것으로 (2)와 같은 형태이다. 병렬형 결합규칙은 (3)과 같이 '미국, 일본, 독일 등' 개체명이 나열된 경우를 인식하기 위한 규칙이다. 어휘기반 결합규칙은 '싱가포르 태생'과 같이 의미범주가 정해지지 않은 어휘에 의해 개체명이 인식되는 경우의 규칙으로 (4)와 같은 형태이다.

### (5) 패턴규칙과 어휘사전의 확장

새로 발견된 각 패턴규칙의 신뢰도를 계산하기 위해서는 추출하는 어휘와 레이블(Label)[7]이 얼마나 정확히 일치하는지를 측정한다. 즉, 패턴규칙이 인식한 레이블의 어휘가 바르게 추출된 것이 많을수록 패턴의 신뢰도가 높다. 각 패턴에 대한 신뢰도는 다음과 같이 계산된다.

$$p(l \mid pattern_i) = \frac{M_i + \alpha}{N_i + k\alpha}$$

$k$ : 가능한 레이블의 개수
$\alpha$ : smoothing parameter

$M_i$: 패턴 $i$에 의해 추출된 레이블 $l$의 멤버 개수
$N_i$ : 패턴 $i$에 의해 추출된 명사의 개수

어휘사전의 확장을 위해서는 어휘를 추출한 모든 패턴규칙에 대해 각 패턴규칙이 어휘의 레이블에 해당하는 어휘를 어느 정도 정확하게 추출했는지를 계산함으로써 결정한다. 즉, 해당 어휘를 추출한 패턴규칙들이 레이블에 해당하는 어휘를 정확하게 추출할수록 어휘의 신뢰도가 높다. 어휘에 대한 신뢰도는 다음과 같이 계산된다.

$$p(l \mid word_i) = \frac{\sum_{j=1}^{n} p(l \mid pattern_j) / \sum_{k=1}^{N} p(l \mid pattern_k)}{\frac{n}{N}}$$

$n$ : word $i$를 추출한 패턴의 개수
$N$ : 패턴의 총 개수

## 4. 실험 및 결과

본 논문에서 제안한 시스템의 성능을 평가하기 위해 경제, 여행, 공연 분야에서 각각 50 개 문서를 대상으로 인명, 지명, 기관명에 대해 개체명 인식을 수행하였다. [표 5]은 각 분야별, 개체명 범주별로 시스템 성능을 평가한 결과를 나타낸 것이다.

[표 5]에서 볼 수 있듯이 경제 분야의 문서들이 높은 성능을 보이고 공연 분야의 문서들은 정확도가 낮은 편이다. 공연 분야의 문서들은 '베르테르를 사랑하는 모임', '상상 21' 등 형태가 다양한 기관명이 많이 출현하고, 인명으로는 '만프레드 뮈사우어', '이브힘 페레' 등 인식하기 어려운 외국 예술인의 이름이 많이 등장하기 때문이다. 또한, 인명이 기관명과 지명에 비해 인식하기 어려운 것을 볼 수 있다. 이는 인명은 개체명 자체에서 얻을 수 있는 단서가 매우 부족하기 때문이다.

| | 경제 | | 여행 | | 공연 | | 경제/여행/공연 | |
|---|---|---|---|---|---|---|---|---|
| | 개수 | % | 개수 | % | 개수 | % | 개수 | % |
| 인식된 개체명 | 528 | 87.6 | 1074 | 80.6 | 564 | 59.3 | 2166 | 75.1 |
| 총 개체명 수 | 603 | 100 | 1333 | 100 | 951 | 100 | 2887 | 100 |

| 150문서 결과 | 기관명 | | 지명 | | 인명 | | 기관명/인명/지명 | |
|---|---|---|---|---|---|---|---|---|
| | 개수 | % | 개수 | % | 개수 | % | 개수 | % |
| 인식된 개체명 | 437 | 76.4 | 1342 | 81.5 | 387 | 57.8 | 2166 | 75.1 |
| 총 개체명 수 | 572 | 100 | 1646 | 100 | 669 | 100 | 2887 | 100 |

[표 5] 개체명 인식 결과

## 5. 결론 및 향후 연구

본 논문에서는 한국어 개체명 인식의 한계점인 새로운 도메인으로의 이식성을 고려하여 패턴규칙과 사전을 이용하는 cotraining 기반 학습의 개체명 인식 모델을 제안하였다. 이 모델의 가장 큰 장점은 개체명이 태깅되지 않은 학습데이터를 이용해 어휘사전과 규칙을 자동으로 확장함으로써 사람의 개입을 최소화할 수 있다는 것이다. 실험 결과, 경제분야에서는 87.6%의 정확률을, 경제, 여행, 공연 분야를 통합했을 때에는 75.1%의 정확률을 보였다.

향후연구로는 잘못된 패턴규칙이나 어휘가 사전에 추가되었을 때 영향이 매우 크므로 이들에 대한 정교한 에러 필터링이 요구된다. 또한, 새로운 패턴을 추출할 때, 이미 정의한 의미범주 외에 새로운 의미범주가 등장했을 때도 고려해야 할 것이다.

## 참고문헌


[1] Yu, S. Bai, S. and Wu, P., "Description of the Kent Ridge Digital Labs System Used for MUC-7", MUC-7, 1998.

[2] GuoDong Zhou, Jian Su, *"Named Entity Recognition using an HMM-based Chunk Tagger,"* 2002

[3] Rohini Srihari, Cheng Niu and Wei Li, *"A Hybrid Approach for Named Entity and Sub-Type Tagging,"* 2000

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

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

[6] Choong-Nyoung Seon, Y. K., J. K. and J. S, *"Named Entity Recognition using Machine Learning Methods and Pattern-Selection Rules,"* pp. 229-236, NLPRS 2001.

[7] Michael Collins, *"Unsupervised Models for Named Entity Classification,"* 1999


[6] DIC : 사전자질, CLUE : 단서자질

[7] 인명, 지명, 기관명, 직위, 직업 등의 의미범주