# 웹을 이용한 개체명 부착 말뭉치의 자동생성과 정제

안주희[0] 이승우 이근배

포 항 공 과 대 학 교 컴 퓨 터 공 학 과

{minnie, pinesnow, gblee}@postech.ac.kr

# Automatic Generation of Named Entity Tagged Corpus using Web Search Engine

JooHui An[0] SeungWoo Lee Gary GeunBae Lee

Dept. of Computer Science, Postech

요 약

최근 정보 추출, 질의응답 시스템 등의 고정밀 자연어처리 어플리케이션이 부각됨에 따라 개체명 인식의 중요성이 더욱 커지고 있다. 이러한 개체명 인식을 위한 학습에는 대용량의 어휘자료를 필요로 하기 때문에 충분한 학습 데이터, 즉 개체명 태그가 부착된 충분한 코퍼스가 제공되지 못하는 경우 자료희귀문제(data sparseness problem)로 인하여 목적한 효과를 내지 못하는 경우가 많다. 그러나 태그가 부착된 코퍼스를 생성하는 일은 시간과 인력이 많이 드는 힘든 작업이다. 최근 인터넷의 발전으로 웹 데이터는 그 양이 매우 많으며, 습득 또한 웹 검색 엔진을 사용해서 자동으로 모음으로써 다량의 말뭉치를 모으는 것이 매우 용이하다. 따라서 최근에는 웹을 무한한 언어자원으로 보고 웹에서 필요한 언어자원을 자동으로 뽑는 연구가 활발히 진행되고 있다. 본 연구는 이러한 연구의 첫 시도로 웹으로부터 다량의 원시(raw) 코퍼스를 얻어 개체명 태깅 학습을 위한 태그 부착 코퍼스를 자동으로 생성하고 이렇게 생성된 말뭉치를 개체명 태깅 학습에 적용하는 비교 실험을 통해 수집된 말뭉치의 유효성을 검증하고자 한다. 향후에는 자동으로 웹으로부터 개체명 태깅 규칙과 패턴을 뽑아내어 실제 개체명 태거를 빨리 개발하여 유용하게 사용할 수 있다.

## 1. 서론

개체명은 사람의 이름, 지명, 기관명 등 특정한 개체의 이름을 일컫는다. 이러한 개체명 태깅은 정보 검색이나 기계 번역, 정보 추출(Information Extraction), 질의응답 시스템(Qustion Answering System) 등의 자연어처리 어플리케이션에 있어서 매우 중요한 부분이다. 특히 최근에는 정보 추출이나 질의응답 시스템 같은 고정밀 어플리케이션이 부각되고 있어 개체명 태깅은 그 중요도가 더해지고 있다.

많은 개체명 태깅 시스템은 학습을 통하여 시스템을 구현하게 되는데, 이 때 학습을 하기 위해서는 태그가 부착되어 있는 코퍼스가 필요할 뿐만 아니라 일반적으로 통계적 언어처리 분야에서 시스템의 정확도에 가장 큰 영향을 미치는 요인 중의 하나인 자료희귀문제(data sparseness problem)를 피하기 위해 대용량의 코퍼스가 필요하다. 따라서 개체명이 태깅된 코퍼스의 확보가 매우 중요한 문제가 된다. 하지만 태그가 부착된 코퍼스를 만드는 것은 매우 많은 인력과 시간이 드는 일이다. 실제 사람이 160 텍스트를 태깅하는데 드는 시간은 8시간이라고 한다. 즉, 사람이 1000 텍스트의 태깅된 코퍼스를 만들기 위해서는 일주일의 시간이 걸리게 된다 [3].

최근 인터넷의 발전으로 인해 다양한 정보가 문서화되어 공유되고 있으며, 그 양 또한 매우 빠른 속도로 증가하고 있다. 따라서 최근에는 웹을 무한한 언어자원으로 보고 웹에서 필요한 언어자원을 자동으로 뽑는 연구가 활발히 진행되고 있다. 우리는 이러한 추세에 맞추어 웹에서 Named Entity를 포함하고 있는 텍스트를 검색 엔진을 사용함으로써 많은 양을 자동

으로 얻을 수 있다.

본 연구는 이러한 맥락에서 첫 시도로 좌우 문맥 패턴을 이용한 개체명 태깅 시스템 학습을 위한 개체명 부착 코퍼스를 자동으로 얻기 위한 방법을 제안한다. 그리고 웹으로부터 자동으로 얻은 개체명 부착 말뭉치의 유효성을 검증하기 위해 본 연구실의 개체명 태깅 시스템 POSNER(POStech Named Entity Recognizer)를 이용하여 수작업으로 만들어진 말뭉치와 자동 생성된 말뭉치를 각각 학습하여 성능을 비교 평가하는 실험을 수행한다. 태그가 부착되지 않은 웹 텍스트를 메타 서치 엔진과 웹 텍스트 추출기를 통하여 얻고, 얻어진 텍스트의 정제 과정을 거쳐서 개체명 태깅의 좌우 문맥 패턴 학습을 위한 개체명 태그 부착 코퍼스를 자동생성하고, 생성된 태그 부착 코퍼스의 정확도를 보임으로써 태그 부착 코퍼스 생성의 재료로서 웹 데이터 이용의 가능성을 보인다. 본 논문은 2장에서 본연구의 관련 연구들을 소개하고 3장에서 시스템의 개관을 보이며 4장에서 생성된 코퍼스의 정확도를 보여줄 것이다. 그리고 마지막으로 5장에서 결론을 맺는다.

## 2. 관련연구

MUC6(Message Understanding Conference)에서 처음 등장한 개체명 태깅은 그 시스템을 크게 나누면 규칙 기반의 시스템과 통계적 방법의 시스템 그리고 두 방법을 혼합한 시스템의 3가지로 볼 수 있다. MUC7 참가팀들에게는 학습 데이터로 1997년부터 방송된 뉴스 기사 32000 단어의 개체명 태그부착 코퍼스가 주어졌다. 후에 BBN( MUC7 참가팀)은 자신들의 시스템의 학습을 위해 수작업으로 만든 백만 단어 크기의 개체명 태그 부착 코퍼스를 배포하였다.

태그가 부착된 대량의 코퍼스로부터 통계적 방법 혹은 기계 학습을 통한 시스템의 성능 향상을 위한 연구는 90년대 초반부터 집중적으로 수행되어 왔고, 현재 많은 자연어처리 시스템들은 코퍼스에 기반한 방법을 통하여 시스템을 개발하고 있다. 그러나 이러한 방법들은 태그 부착 코퍼스의 부족이라는 문제에 직면하게 되었다.

코퍼스 기반 자연어처리의 신뢰도를 높이기 위해 요구되는 코퍼스 크기를 예측하고자 하는 연구에서는 실용적인 자연어처리 시스템을 만들려는 목적에서 보면 자연어처리를 위한 코퍼스의 수집은 균형보다는 크기에 우선순위를 두는 것이 바람직하다는 결론을 도출하였다 [6].

이렇듯 대량의 태그부착 말뭉치의 구축은 말뭉치에 기반한 학습에 있어서 매우 중요하다 그러나 수작업에 의한 말뭉치의 구축은 많은 시간과 노력을 요하는 힘든 작업이다. 따라서 본 연구는 웹 검색을 통해 태그가 부착된 말뭉치를 자동으로 획득하고, 구축된 말뭉치의 유효성을 보이고자 한다.

## 3. 시스템 개관

본 연구는 POSNER의 패턴 학습을 위한 개체명 태그 부착 코퍼스를 생성하고, 이를 이용한 개체명의 어휘 패턴의 학습에 적용하고자 하는 것이다. POSNER의 패턴 학습은 개체명의 문맥을 학습하여 개체명 태그를 부착하는 학습으로, POSNER의 패턴은 다음과 같이 이루어져있다.

```
왼쪽 문맥 <Person> 오른쪽 문맥
왼쪽 문맥 <Location> 오른쪽 문맥
왼쪽 문맥 <Organization> 오른쪽 문맥
…
```

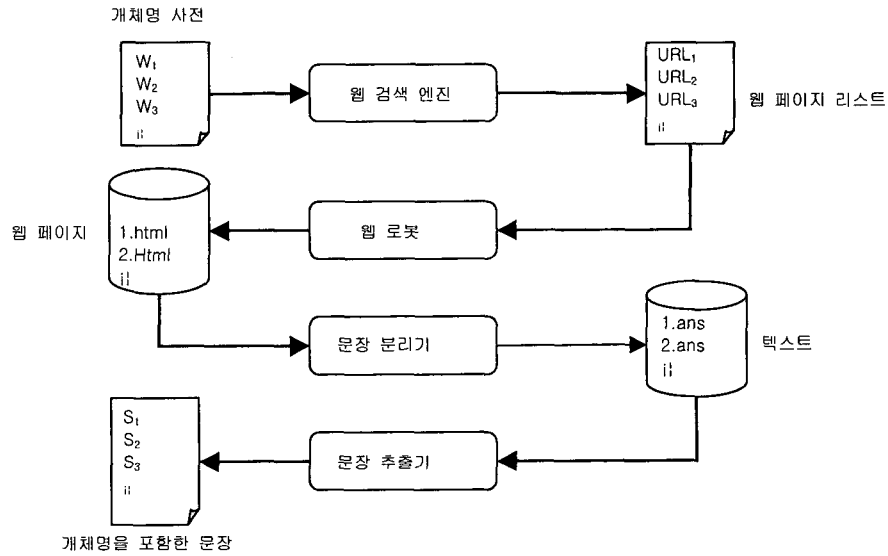
이 장에서는 위와 같은 학습에 필요한 태깅된 개체명과 좌우 문맥을 자동으로 획득 할 수 있는 방법을 제안한다.

개체명 사전을 어느 정도 확보하고 있으면 웹 검색 엔진을 통해 개체명이 나타나는 많은 문서를 찾을 수 있고, 개체명의 유형은 이미 알고 있기 때문에 이들 문서 내에서 나타나는 개체명들에 대해 해당 개체명 태그를 부착 할 수 있다.

웹 검색엔진을 이용한 개체명 태그 부착 코퍼스의 자동 생성은 다음과 같은 순서로 이루어진다.(그림1)

1. 웹 검색을 위한 개체명 리스트를 준비한다.
2. 리스트의 각 엔트리를 웹 검색 엔진에 질의로 사용하여 검색된 문서의 URL을 얻는다.
3. 웹 로봇을 이용하여 각 URL의 텍스트를 다운로드한다.
4. HTML 태그를 제거하고 텍스트를 문장단위로 분할한다.
5. 정확한 개체명을 포함하는 문장을 뽑아온다.
6. 해당 개체명의 NE category를 태깅 후 저장한다.

사용한 개체명 엔트리와 웹 검색 엔진을 통한 텍스트 수집 및 텍스트 정제 과정에 대해 차례로 기술하겠다.




<그림 1: 시스템 개관 >

### 3.1 개체명 엔트리

웹 서치를 위한 개체명 리스트는 본 연구실에서 보유하고 있는 Person, Organization, Location의 세가지 Named Entity 타입으로 이루어진 개체명 사전(49만 엔트리)에서 일부를 추출한 것이다. 각 타입별로 Person : 937 엔트리, Organization : 1050 엔트리, Location : 1000 엔트리를 추출하여 리스트를 구성하였다.

<표1: 개체명 리스트 예>

| | |
|---|---|
| 가남숙 | <P> |
| 김갑동 | <P> |
| 도로교통안전협회 | <O> |
| 63빌딩 | <L> |
| 가나안땅 | <L> |
| ... | |

### 3.2 웹 검색 엔진을 이용한 수집

다음으로 웹 검색 엔진을 사용하여 개체명 엔트리를 포함하는 웹 페이지의 URL을 얻어온다. 본 연구에서는 검색엔진으로 엠파스(http://www.empas.com)를 사용하였다.

개체명 엔트리 각각을 엠파스에 질의로 요구하여 하나의 개체명 표제어마다 최고 10개의 웹 페이지 리스트를 가져오도록 하였다. 각 웹 페이지에는 개체명이 여러 번 발생할 수도 있지만 전혀 나타나지 않을 수도 있다. '핑클' 이 '써핑클럽' 에 잘못 매칭되듯이 전혀 엉뚱한 단어에 매칭된 문서가 검색될 수도 있고, 개체명이 둘 이상의 단위 단어로 분할 될 수 있을 때 각각이 개별적으로 매칭되어 검색되는 경우도 있다. 따라서 검색된 텍스트에서 원하는 개체명이 포함된 문장을 찾기 위해서는 단순한 문자열 매칭 이상의 기술이 필요하다. 이에 대해서는 3.3 장에서 자세히 다루겠다.

웹 로봇은 검색된 서치 엔진 검색 결과로 얻어진 리스트의 각각의 사이트를 방문하여 텍스트를 얻어온다. 이 때 해당 페이지의 텍스트만을 대상으로 다운로드하고 페이지가 링크하고 있는 페이지들은 무시한다. 이 작업은 로봇이 방문해야 할 모든 페이지 리스트에 대하여 이루어지며 얻어오는 페이지에 대하여 중복을 허용하지 않았다.

로봇이 웹의 다큐먼트를 얻어올 때 웹 페이지 형식을 위한 태그들을 제거한 뒤 텍스트 만을 얻어오며 header, label, caption등은 추출대상에서 제외된다.

추출된 텍스트들은 문장 분할 과정을 거쳐 문장 단위로 저장된다. 정규 규칙과 휴리스틱 그리고 C4.5의 학습 데이터를 이용하여 문장을 분할한다[5].

<표 2: 자동 생성된 개체명 태그 부착 코퍼스 예>

<L>가야산국립공원</L> 백운동 집단시설지구 일원에서 1999년부터 "산과 꽃 천연의 만남" 의 주제로…
장소 : <L>가야산국립공원</L> 백운동집단시설지구 일원
<L>가야산국립공원</L>과 인접한 합천은 발표회,전시회등 문화예술활동을 활발히 전개하고 있어…
…
우리 교육청에서는 <O>경상북도교육청</O>과 함께 ;『새 천년 열어갈 정직하고 창의적인 인간 육성』…
<O>경남스틸</O>은 주주들에게 30%의 고율배당을 시행할 계획이다.
<O>경남스틸</O>은 현재 소폭절단이 가능한 미니슬리터 기계를 증설 중이다.
…
중국도서관사 개설, <P>김화자</P>, 2000.
2002 인후초등 1학년 5반 <P>김진영</P>입니다.
<P>김정일</P> 총서기가 "우리가 더 빨리 발전하자면 다른 나라에서…
…

웹에서 페이지 단위로 텍스트를 추출하게 되면 개체명 엔트리를 포함하지 않는 문장들 또한 다량 포함되어 있다. 본 연구에서는 개체명 엔트리 태깅을 위한 좌우 문맥 패턴 학습 코퍼스를 생성하는 것이므로 이러한 문장들은 코퍼스 수집 대상에서 제외되어야 한다. 따라서 웹에서 추출한 텍스트 중에서 개체명 엔트리을 포함하고 있는 문장만을 추출하는 작업을 거치게 된다.

### 3.3 텍스트 정제

검색 엔진을 이용해서 개체명 엔트리를 포함하는 텍스트를 추출하는 과정에서 잘못 매치된 텍스트 또한 가지고 있게 된다. 실제로 엠파스는 n-gram 매칭을 이용하여 검색을 한다. 충분한 자연어처리 기술을 이용해서 정확한 결과를 출력하는 것이 아니기 때문에 오류를 포함할 수 있다. 따라서 잘못 매치된 문장을 정제하는 것이 매우 중요하다. 이러한 문장을 정제하기 위해서는 조사를 분리해내고 복합명사를 검사하여 추출된 문장이 올바른 결과인지를 검사할 필요가 있다.

우선적으로 질의문에 대하여 단어가 분리되어 나오는 문장들이 제외된다. 예를들어 '포항시의회' 를 질의어로 검색한 결과는 다음과 같은 문장을 포함하고 있다.

`포항시중학교 의회에 참석 학생대표들과 함께...'
`민주노총포항시협의회 의장 김병일 위원'

이 문장들은 '포항시의회'를 문장내에 질의로 사용하여 검색엔진으로부터 얻은 결과이지만 '포항시의회'의 개체명 엔트리를 포함하고 있지 않으므로 제외되어야 한다.

개체명은 뒤에 조사가 붙어서 하나의 어절을 이룰 수 있다. 따라서 문장에서 개체명을 포함하는 어절 내에서 조사를 분리해 낸다. 조사의 분리는 포항공대 품사 부착기(POStech TAGger)를 사용하여 분리해낸다[8].

조사를 분리해낸 어절에 대해서 복합 명사인지를 검사해야 한다. 이때에 복합명사를 분리해서 잘못 매치된 것일 경우 문장을 제외시킨다. 예를 들어, '핑클' 이라는 개체명 엔트리가 있어서 이를 웹에서 찾았는데, '써핑클럽' 을 포함하는 문장이 추출되었다고 하자. 그러면 '써핑클럽' 은 복합명사 처리를 통하여 '써핑' +' 클럽' 으로 나뉘어 지고 이는 '핑클' 과는 매치가 되지 않는다. 따라서 '핑클' 이 '써핑클럽' 에 매치되어 추출되어진 문장들은 개체명 태그 부착 코퍼스 생성에서 제외시킬 수 있다.

한국의 복합 명사는 공백없이 붙여서 사용될 수도 있고 띄워서 사용할 수 있다. 또한 중의적 분할이 발생할 수 있기 때문에, 복합 명사를 단위 명사로 분해하는 것은 어려운 작업이다. 본 연구에서의 복합 명사의 분할은 음절 수에 따른 분할 패턴과 상호정보(Mutual Informaion)를 사용한다. 복합 명사 분할 패턴은 다음과 같다[7].

4음절 : 2/2
5음절 : 2/3, 3/2
6음절 : 3/3, 2/4, 4/2, 2/2/2
7음절 : 2/3/2, 3/4, 4/3, 5/2, 2/5 …
8음절 : 2/3/3, 3/3/2, 2/4/2, 3/5, 5/3, 6/2, 2/6

위의 복합 명사 분할 패턴과 함께 분할 패턴의 중의성을 해결하기 위해 상호 정보(Mutual

Information)를 사용한다. 여기에서 사용하는 상호 정보 식은 다음과 같다. 식에서 x, y 는 복합 명사를 구성하는 각 단어이다.

$$MI(x,y) = \frac{P(x,y)}{P(x)P(y)} = \frac{f(x,y)}{f(x,*)f(*,y)}$$

예를 들어 '일본' 을 질의로 던져서 검색을 한 결과로 다음과 같은 문장이 있다.

'1894 8월 일본전권공사로 임명됨.'

'일본전권공사' 는 6음절로 이루어져 있으므로 '일본' + '전권공사' , '일본전' + '권공사' , '일본전권' + '공사' , '일본' + '전권' + '공사' 의 4가지 패턴으로 분할이 가능하다. 따라서 이 4가지의 경우에 대하여 상호정보를 구하여 상호정보 값이 가장 큰 값을 가지는 패턴으로 분리가 된다. 위 예의 경우 '일본' + '전권' + '공사' 로 분리 될 수 있고, 따라서 '일본' 을 포함하는 개체명 코퍼스로 선택될 수 있다.

'1894 8월 <0>일본</0>전권공사로 임명됨.'

또 '김덕' 이라는 사람 이름을 검색 했을 때 다음과 같은 문장들을 얻을 수 있다.

'김덕호홈페이지 - 표지.'

위 문장은 복합 명사 분할을 한 결과가 '김덕호' + '홈페이지' 가 되어 질의어에 매치되지 않는다. 따라서 이런 문장도 복합명사 처리를 통해 제외될 수 있다.

텍스트 정제 과정에서는 또한 충분한 문맥을 포함하지 않고 단어만 매치되서 추출된 것은 학습에 도움이 되지 않기 때문에 이러한 데이터는 코퍼스 생성에서 제외된다.

### 3.4 개체명 태그 부착 코퍼스

개체명 정제 과정을 통해 수집된 문장들은 웹 검색에 사용된 개체명의 유형에 따라 개체명 태그가 부착된다. 웹 데이터를 이용한 개체명 태그 부착 코퍼스의 자동 생성을 통해 사람 이름, 기관명, 지역 이름의 총 엔트리 2987개에 대하여 17000개의 개체명을 포함하는 태그 부착 코퍼스를 얻었다. 본 연구에서 제안하는 방법은 문장 내에 개체명이 2개 이상 나타나는 경우 1개 밖에 태깅하지 못한다. 따라서 코퍼스의 크기에 대한 효율은 떨어질 수 있으나 웹 데이터를 이용한 자동 생성으로 코퍼스의 양을 무한히 늘릴 수 있기 때문에 실제 학습에는 문제가 되지 않는다.

## 4. 코퍼스의 정확도

학습을 효과적으로 하기 위해서는 코퍼스의 양뿐만 아니라 코퍼스의 정확도 또한 매우 중요하다. 그러나 일반적으로 개체명 태그가 부착된 코퍼스를 자동으로 얻는 방법은 손으로 태그를 부착한 코퍼스에 비해 그 정확도가 떨어진다. 따라서 본 연구에서 제안하는 방법에 의해 자동으로 생성된 코퍼스의 정확도를 측정하여 학습에 이용가능한지를 판단하는 것은 매우 중요한 일이다.

자동 생성된 개체명 태그 부착 코퍼스의 정확도를 알아보기 위해, 본 연구실의 기존의 개체명 태깅 시스템 POSNER의 학습 코퍼스로 사용하여 시스템의 성능을 측정하여, 이의 성능이 손으로 태그를 부착한 코퍼스로 동일한 시스템을 학습했을 경우의 성능과 비교하였다. POSNER는 개체명 태그가 부착된 코퍼스로부터 seed pattern을 학습하고 원시 코퍼스를 이용하여 DL-CoTraining을 하여 결정트리를 생성한다[4].

수동으로 개체명 태그를 부착한 코퍼스를 이용하여 학습한 시스템의 실험 결과는 다음과 같다.

<표 3 : 개체명 태그 부착 코퍼스>

| | News domain | | | Non-news domain | | |
|---|---|---|---|---|---|---|
| | P | L | O | P | L | O |
| Training | 337 | 133 | 994 | 677 | 591 | 344 |
| Test | 26 | 44 | 193 | 102 | 72 | 57 |

<표 4 : 원시 코퍼스>

| Corpus | Nes |
|---|---|
| News domain 1(A) | 26,394 |
| News domain 2(B) | 51,318 |
| Non-news domain 1(C) | 50,555 |
| Non-news domain 2(D) | 91,127 |
| Mixed domain (E) | 76,949 |

<표 5 : 뉴스 도메인 코퍼스로 학습>

| | Corpus A | Corpus B | Corpus E |
|---|---|---|---|
| Precision | 81.12 | 83.01 | 79.54 |
| Recall | 86.31 | 81.75 | 78.83 |
| F-measure | 78.91 | 82.38 | 78.93 |

<표 6 : 뉴스가 아닌 도메인으로 학습>

| | Corpus C | Corpus D | Corpus E |
|---|---|---|---|
| Precision | 83.33 | 81.22 | 87.28 |
| Recall | 80.09 | 80.52 | 86.15 |
| F-measure | 81.68 | 80.87 | 86.71 |

위의 표들은 POSNER를 다양한 코퍼스에서 실험한 결과들로서 웹으로부터 구축한 코퍼스와의 비교가 용이하도록 [4]에 실린 표를 옮겨 실은 것이다. POSNER는 개체명 사전을 사용하지 않고 개체명 태그 부착 코퍼스만으로 학습하여 seed rule을 생성하고 이를 대용량의 원시 코퍼스를 이용하여 Co-training 을 통해 학습하여 결정트리를 생성한다[4].
표 5는 태그부착 코퍼스를 뉴스 도메인으로 하고 원시 코퍼스 A,B,E 각각으로 학습하고 실험한 결과이며, 표 6은 태그부착 코퍼스도 뉴스가 아닌 도메인의 코퍼스를 사용하고 원시 코퍼스 C,D,E 각각으로 학습하고 실험한 결과이다.

<표 7 : 웹으로부터 자동구축한 코퍼스로 학습>

| | Precision | Recall | F-measure |
|---|---|---|---|
| web | 88.09 | 72.54 | 80.32 |

표 7의 결과는 POSNER를 웹으로부터 자동으로 생성된 코퍼스로 학습을 하고 개체명 태깅 한 결과이다. 역시 개체명 사전을 사용하지 않고 태깅하였으며, Co_Training을 사용하지 않았다. 그리고 테스트를 위해 POSNER에서 테스트를 위해 사용한 데이터와 동일한 263개의 개체명을 포함한 뉴스 도메인의 데이터를 사용하였다.

위의 결과로 볼 때 수동으로 태그 부착한 코퍼스로 학습한 시스템과 웹에서 자동으로 생성된 코퍼스로 학습한 시스템은 Recall에서 가장 큰 차이를 보이고 있다. 그러나 이는 학습 도메인이 매우 상이하기 때문이라고 볼 수 있으며, 코퍼스를 더 수집해서 학습하면 향상될 것이다. 반면에 Precision은 비슷한 성능을 보이고 있다. 이는 웹 다큐먼트를 통해 자동으로 생성된 코퍼스가 수동으로 생성된 코퍼스와 비슷한 정도의 어휘 자질과 문맥 자질을 제공하기 때문이다. 따라서 웹 데이터를 이용하여 자동으로 구축된 개체명 태그 부착 코퍼스는 수동으로 구축된 개체명 태그 부착 코퍼스를 대체할 수 있으며 대량의 코퍼스 구축이 가능하기 때문에 개체명 인식 시스템의 학습을 위한 코퍼스 구축의 좋은 방법이 될 수 있다.

## 5. 결론

본 논문에서는 개체명 리스트를 이용하여 웹 검색을 통하여 태그가 부착되지 않은 코퍼스를 생성하고 이를 자동으로 태그를 부착하여 패턴을 만들기 위한 태그 부착 코퍼스를 만드는 방법을 제안하였다. 본 연구에서는 3개의 개체명 카테고리(Person, Location, Organization)에 대하여 웹 검색을 통하여 45만 단어의 개체명 태그 부착 코퍼스를 얻을 수 있었고, 이렇게 생성된 코퍼스를 기존의 개체명 태깅 시스템에 적용하여 실험한 결과 수동으로 태그 부착한 코퍼스에 견주어 신뢰할 만한 결과를 얻었다. 이 결과를 볼 때, 본 연구에서 제안하는 개체명 태그 부착 코퍼스의 자동 생성 방법은 대량의 개체명 태그 부착 코퍼스의 생성을 통해 개체명 인식 시스템의 학습을 위한 데이터를 제공할 수 있는 좋은 대안이 될 수 있다.

보다 정확한 코퍼스 구축을 위해서는 3.3 장에서 기술한 텍스트 정제의 보다 높은 정확도와 재현율이 요구된다. 그러나 현재 구현된 시스템에서는 정확한 개체명을 정제하지 못하는 경우도 있다. 예를 들어,

'광동제약측은…'

에서처럼 '광동제약' 이 하나의 개체명이고 뒤에 붙은 '측' 은 한단어 명사이다. 현재는 한단어 명사에 대한 처리는 하지 않고 있기 때문에, 이런 경우 올바른 개체명을 포함하고 있음에도 불구하고 텍스트 정제 과정에서 누락되는 경우가 있었다. 이런 경우 보다 정확한 복합명사 분할과 함께 접두사, 접미사를 인식하여 보다 정확한 개체명을 정제할 필요가 있다. 향후에는 이를 좀더 보강하고 자동으로 생성된 코퍼스를 이용하여 패턴학습을 이용한 개체명 태깅에 대하여 연구할 예정이다.

## 6. 참고 문헌


[1] Gregory W. Lesher, Christian Sanelli, 2000, "A Web_Based System for Autonomous Text Corpus Generation" International Society for Augmentative and Alternative Communation (ISAAC)

[2] Deepak Ravichandran, Eduard Hovy, 2002, "Learning Surface Text Patterns for a Question Answering System", Proceedings of the 40th ACL conference.

[3] Ellen Riloff, 1996, "Automatically Generating Extraction Patterns from Untagged Text", Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), 1044-1049

[4] 곽병관, 2001, "Co-training을 이용한 실용적 개체명 태깅", 포항공과대학교 석사학위 논문

[5] 심준혁, 2001, "웹 문서 처리를 위한 통합된 문서 전처리 시스템", 포항공과대학교 석사학위 논문
[6] 양단희, 임수종, 송만석, 1999, "자료 빈약성을 해소하기 위한 말뭉치 크기의 예측", 한국정보과학회 논문지, 제26권 제4호, pp. 568-583
[7] 윤보현, 조민정, 임해창, 1997, "통계 정보와 선호 규칙을 이용한 한국어 복합 명사의 분해", 한국 정보과학회 논문지(B) ,제24권, 제 8호, 900-909
[8] 이근배, 차정원, 이종혁, 1997, "Hybrid POS tagging with generalized unknown generalized unknown-word handling.", Proceedings of the 2nd international workshop on information retrieval with Asian languages (iral97), Tsukuba-City, Japan, pp43-50