

서 론

오늘날 많은 사람들이 웹(Web)으로부터 많은 정보를 얻고 있다. 이들 웹페이지에는 여러 가지 정보를 포함 있지만, 그 중에서도 인명, 지명, 조직명, 상품명 등의 고유 명사가 해당 웹페이지를 구분하는데 중요한 정보를 제공하는데 이러한 고유 명사를 개체명(Named Entity)이라고 한다. 하지만 이렇게 중요한 정보를 제공하는 개체명은 그 수가 한정되어 있지 않고, 계속해서 생성되기 때문에 모든 개체명을 사전에 등록하는 것은 현실적으로 불가능하다[1]. 본 논문에서는 이러한 약점을 보완하기 위해서 영역별로 개체명 사전을 구축하기 위한 도구인 '집합 확장 시스템(Set Expansion System)'을 이용하여 개체명을 추출하는 시스템을 제안한다. 이 집합 확장 시스템을 이용하여 영역별로 개체명 사전을 구축한다면, 만들어진 개체명 사전을 이용하여 많은 분야에서 빠르고 정확한 개체명을 인식할 수 있을 것이다[2, 3].

집합 확장 시스템은 해당 개체명의 집합을 웹을 통해서 자동으로 수집 해주는 도구를 말한다[4]. 이것은 알고 싶은 집합의 원소의 두 개 내지 세 개를 씨앗단어(seed word)로 사용하여 대상(target)이 되는 집합을 찾아내고, 그 집합의 원소를 리스트(lists) 형식으로 보여주는 시스템을 말한다[4, 5]. 이 시스템을 이용하여 영역별로 개체명 사전을 쉽게 구축할 수 있을 것이다[4]. 현재 이 집합 확장 시스템을 상용화하고 있는 대표적인 시스템은 구글에서 제공하고 있는 구글셋(Google SetsTM : <http://labs.google.com/sets>)이 있다. 이것은 집합이 있는 개체명의 원소를 1개에서 5개까지 입력받아 그 원소의 집합에 속해있는 다른 원소들을 일정 수준의 개수만큼 보여준다.

집합 확장 시스템을 이용해서 구축된 개체명 사전은 의견 추출(Opinion Mining) 분야에서 유용하게 사용될 수 있다[6]. 의견 추출이란, 기사평, 상품평, 블로그 댓글 등의 평가나 의견을 추출 후, 분석하여 자동으로 전반적인 평가를 도출해 내는 기술이다. 의견의 대상이 되는 상품명이나, 영화명, 정치인, 관광지 등은 모두 고유 명사이다. 이러한 영역들의 개체명 추출은 의견 추출시 의견을 제시한 사람(opinion holder)이 누구인지, 무엇에 대한 의견(opinion object)인지를 분석하기 위해서 꼭 필요한 작업이지만, 의견 추출의 영역에 따라 다른 개체명 사전의 구축이 필요하다

는 어려움이 있다. 이러한 어려움을 본 논문에서 제안하는 집합 확장 시스템을 이용하여 영역별로 따로 개체명 사전을 자동으로 구축하게 된다면, 새로운 영역의 의견추출 시 필요한 개체명 사전의 구축을 보다 손쉽게 정확하게 수행할 수 있을 것이다.

집합 확장 시스템에 관한 연구로는 상용화 된 구글셋 외에 카네기멜론 대학의 SEAL(Set Expander for Any Language)이 있다. 이 시스템은 웹을 이용하여 개체명 대상들을 추출하고 "Laze Work Process"라는 기법을 통해 개체명 대상들의 중요도를 계산하였다[4].

본 논문에서는 효과적인 집합 확장 시스템을 구축하기 위해서 씨앗단어 정보와 패턴규칙(pattern rule)을 이용해서 해당 영역의 개체명 후보들을 추출하고, 씨앗단어 정보와 웹페이지(Web page), 그리고 개체명 후보 간의 관계를 그래프로 표현하여 웹페이지와 개체명 후보 간의 상호 중요도를 계산함으로써 최종 개체명 후보들의 중요도를 계산하고 순위화 한다. 실험으로 관찰한 제안된 시스템의 성능은 영어 개체명을 위해서는 구글셋의 결과보다 더 좋은 성능을 보였으며, 한국어 개체명을 위한 실험에서는 그래프 기반의 상호중요도 측정 기법을 사용했을 때가 더 좋은 성능을 보이고 있다. 그러므로, 제안된 상호중요도 측정 알고리즘의 효과적임을 알 수 있었다. 한국어 개체명의 경우 구글셋이 아직 서비스를 제공하지 않아 구글셋과는 비교하지 못했다.

본 논문의 구성은 다음과 같다. 2절에서는 제안하는 집합 확장 시스템을 자세히 설명하고, 3절에서는 실험 방법과 실험결과에 대해서 토의하고. 마지막으로 4절에서는 본 논문의 결론에 대해서 설명하고 향후 연구에 대해서 기술한다.

시스템

제안하는 시스템은 [그림 1]과 같이 구성되어 있다. 먼저 연구자 하는 분야의 씨앗단어 세 개를 구글 API(Google Application Program Interface)의 검색 단어로 사용하여 세 개의 씨앗단어를 모두 포함하고 있는 웹페이지들을 추출한다[7]. 이렇게 검색된 웹페이지에서 개체명 추출을 위한 패턴(pattern)을 추출하고, 그 패턴에 맞는

개체명 후보들을 선정한다. 선정된 후보들에 그래프 기반의 상호 중요도 측정 기법을 이용해서 중요도를 계산하여 순위화 된 개체명 후보 리스트(list)를 출력한다.

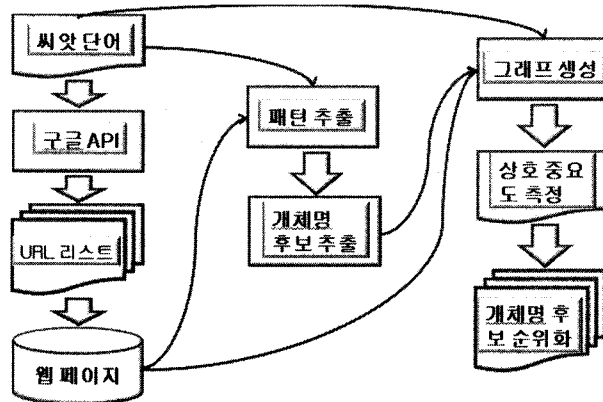


그림 1. 전체적인 시스템 구성도

웹페이지 검색

세계적인 검색 사이트인 구글은 그들이 웹 검색 시에 사용하고 있는 방식과 보유하고 있는 데이터들을 바탕으로 질의(query)에 맞는 웹페이지를 찾아주는 API를 제공하고 있다. 본 논문에서는 이 구글 API를 사용하여 웹페이지를 검색한다. 세계의 질의를 사용하여 구글 API를 통해 검색된 웹페이지는 상위 100개를 사용하고 있다. 이 상위 100개의 웹페이지는 구글에서 선정한 순위를 사용한다.

패턴을 이용한 개체명 후보 추출

오늘날의 웹페이지는 반구조적(semi-structure)인 형식이다. 각각의 웹페이지는 각자 다른 구조를 가지고 있지만, 하나의 웹페이지 안에서는 같은 구조를 가지게 된다[4]. 본 논문은 이러한 웹페이지의 정보와 키워드의 정보를 이용한다.

[그림 2]는 한국에 있는 대학교를 추출하기 위해 키워드 '동아대학교, 서울대

단계 1: 씨앗단어 포함 문장 추출

구글 API에서 검색된 웹페이지에서 씨앗단어가 하나라도 포함된 문장을 모두 추출한다.

단계 2: 패턴규칙 추출

단계 1에서 추출된 문장들 중 입력으로 사용된 세 개의 씨앗단어를 중심으로 왼쪽 문맥과 오른쪽 문맥을 추출하여 그 쌍을 패턴으로 사용하는데, 이때 문맥 추출 크기로는 문자단위로 2이상, 5이하를 사용한다. 씨앗단어를 포함하고 있는 문장에서 왼쪽 문맥과 오른쪽 문맥을 쌍으로 하는 패턴을 씨앗단어별로 각각 추출하고, 세 개의 씨앗단어에서 모두 추출된 패턴들이 해당 웹페이지의 최종 패턴규칙으로 사용된다.

단계 3: 개체명 추출

단계 2에서 추출한 왼쪽과 오른쪽 문맥을 사용하여 그 웹페이지에서 두 문맥이 모두 출현한 패턴을 찾는다. 예를 들어, 왼쪽 문맥이 L, 오른쪽 문맥이 R이라고 한다면 두 문맥이 모두 출현한 문장은 "L개체명R"이란 구조를 가지고 있을 것이다. 이런 패턴에서 왼쪽 문맥과 오른쪽 문맥을 제거하여 개체명 후보들을 얻는다.

그래프 기반의 상호중요도 계산 기법

2.2절에서의 기술한 방법으로 추출된 개체명들은 씨앗단어와 비슷한 단어들도 있지만 그렇지 않은 단어들도 다수 포함되어 있다. 본 논문에서는 추출된 개체명 후보들의 중요도를 부여하여 순위화하기 위하여 그래프(graph)를 생성하여 상호중요도를 계산하는 기법을 고안하였다.

[그림 3]은 씨앗단어와 그에 의해서 검색되어진 웹페이지, 그리고 추출된 개체명 후보들을 노드로 하는 그래프의 예시이다. '동아대학교', '서울대학교', '부산대학교'의 씨앗단어를 통해 검색되어진 임의의 웹페이지 A, B, C가 있고, 그 웹페이지들에서 개체명 후보로 추출된 '경희대학교', '동국대학교', '고려대학교', '고려대학교구로병원'이 있다. 이 중 '고려대학교구로병원'은 우리가 얻고 싶은 대학교의 이

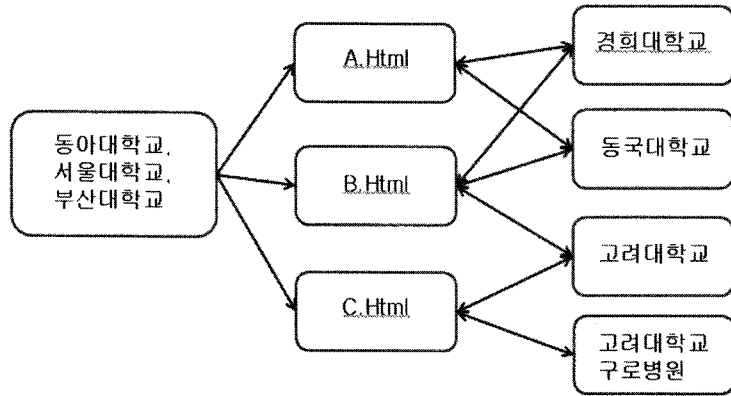


그림 3. 상호 중요도 측정 기법을 위한 그래프

름이 아니기 때문에 그것의 순위를 낮추어야 한다. 그래서 본 논문에서는 다음의 가정을 기반으로 하는 상호 중요도 계산 기법을 제안한다.

1. 중요한 개체명은 중요한 웹페이지로부터 출현한다.
2. 중요한 웹페이지는 중요한 개체명을 많이 포함하고 있다.

이 가정은 재귀적으로 정의되어 있으며, 재귀적 알고리즘(algorithm)을 통해 구현 된다. [그림 3]에서 보는 바와 같이 웹페이지와 개체명 후보 간의 중요도 계산을 위해서 씨앗단어로부터 추출된 웹페이지들의 중요도를 계산하여야 하는데, 본 연구에서는 구글로부터의 각 웹페이지의 랭킹(ranking) 정보를 이용하여 다음의 식으로 계산이 된다.

$$Initial\ Importance\ of\ Webpage = (1 - \alpha) + \alpha \left(\frac{n + 1 - i}{n} \right) \quad (1)$$

위 식에서 n 은 구글에서 검색된 전체 웹페이지 중에서 씨앗단어가 모두 포함된 웹페이지의 전체 개수이고, i 는 구글에서 제공된 웹페이지의 랭킹 정보이다. 그리고 α 는 랭킹된 웹페이지 사이의 가중치를 안정화하기 위한 변수로 본 논문에서는

0.6의 값을 사용하였다. 이들 초기 웹페이지 중요도 값을 이용하여 다음 과정의 재귀적 알고리즘을 통해 개체명 후보들의 중요도를 계산한다.

IW: Importance of Web Page

IN: Importance of Named Entity

NIW: Normalized Importance of Web Page

m_i: the number of named entities of *i*-th Web Page

for(개체명 후보 중요도 값의 총합이 수렴할 때까지)

{

1. 웹페이지 중요도 정규화: 측정된 웹페이지들의 중요도는 최대값으로 나누어 0~1사이 값으로 정규화.

$$NIW_i = \frac{IW_i}{\text{Maximum}(IW_j)}$$

2. 개체명 후보 중요도 계산: 출현한 웹페이지들의 중요도의 합으로 개체명 후보들의 중요도 계산.

$$IN_{ne} = \sum_{ne \in \text{WebPage}(i)} NIW_i$$

3. 웹페이지 중요도 계산: 포함하고 있는 개체명 후보들의 중요도의 평균값

$$IW_i = \frac{\sum_{ne \in \text{WebPage}(i)} IN_{ne}}{m_i}$$

}

실험 결과

각 영역별 씨앗 단어와 성능 평가 방법

본 논문에서는 구축된 시스템을 사용하여 총 11가지의 영역에 대해서 실험을 하였다. 먼저 본 논문에서 제안하는 상호 중요도 측정 기법의 성능을 입증하기 위해서 [표 1]과 같이 3가지 한국어 개체명 영역에 대해 실험을 수행하였고, 구글셋의 결과와 비교하기 위하여 [표 2]와 같이 8가지의 영어개체명 영역을 실험 하였다.

표 1. 한글 개체명 영역과 씨앗단어

영역(한글)	씨앗 단어
대학교	동아대학교, 서울대학교, 부산대학교
영화 & 드라마 제목	쉬리, 박하사탕, 친구
산 이름	소백산, 설악산, 지리산

표 2. 영어 개체명 영역과 씨앗단어

영역(영어)	씨앗 단어
classic-disney	Mary Poppins, Cinderella, Toy Story
mlb-teams	Red sox, Tigers, Marlins
nba-teams	Celtics, Nuggets, Mavericks
nfl-teams	Browns, Vikings, Colts
popular-car-makers	Ford, Nissan, Toyota
us-presidents	George Washington, John Adams, Abraham Lincoln
us-states	California, Indiana, Kansas
watch-brand	Seiko, Omega, Cartier

본 논문에서는 실험 결과들을 비교하기 위해서 문서수준 정확률과 MAP(mean average precision)을 사용한다. 정확률은 전체 검색 문서에 대한 검색된 적합 문서의 비율로서 정의되며, 따라서 정확률을 계산하기 위해서는 검색된 문서의 수가 결정되어야 한다. 여기서 문서수준 n 에서의 정확률이란 상위 n 개의 문서들에 포함된 적합 문서들의 비율로서 정의되며 다음과 같은 식으로 표현된다.

$$\text{문서수준정확률} = \frac{\text{상위 } n \text{개 문서들에 포함된 적합문서수}}{\text{문서수준 } n} \quad (2)$$

MAP는 정보 검색 영역에서 순위가 있는 리스트(lists)를 평가할 때 사용하는 기법이다. 이는 재현율과 정확률을 모두 포함하고 있으며, 전체의 순위에 민감하다. 다양한 영역에서 순위화 된 리스트가 구성되어 있는 경우를 평가하기 위한 방법인 MAP는 간단하게 순위화 된 각 영역에서의 평균정확률(average precision)의 값의 평균이다. 각 영역에 대한 평균정확률은 다음 식과 같이 표현된다.

$$\text{평균정확률}(L) = \frac{\sum_{r=1}^{|L|} \text{정확률}(r)}{\text{실제 정답수}} \quad (3)$$

위 식에서 L 은 개체명 후보들의 순위화 된 리스트이고, r 은 순위이며, 정확률(r)은 순위 r 에서의 정확률이다. 본 논문에서는 만약 순위 r 보다 아래에 같은 개체명이 존재하면, 아래에 있는 개체명을 제외시켰다. 그리고, 실제 정확한 정답집합을 얻기가 어렵기 때문에 영어 개체명 영역에서는 구글셋에서 나온 개체명들 중 정답만의 개수를 실제정답으로 가정하였고, 한글 개체명 영역의 경우는 본 논문의 기법을 사용하여 추출된 개체명 후보들 중에서 정답의 수를 실제 정답수로 계산하였다.

실험 결과 및 토의

본 논문에서는 한국어 개체명 추출을 위한 실험과 영어 개체명 추출을 위한 실

험을 평가하기 위하여, 한국어 개체명 추출의 결과는 각 영역의 평균정확률을 계산하여 전체 MAP를 구하였고, 영어 개체명 추출의 결과는 구글셋과의 정확한 비교를 위하여 문서 수준 정확률과 MAP로 모두 비교하였다. 그리고, 후보들의 재현율을 알아보기 위해서 유한한 영역인 한국어 영역의 ‘대학교’ 영역을 선택하였다.

한국어 개체명 추출 실험 결과

유한한 영역의 재현율

[표 3]은 한국어 영역에서 정답의 수가 유한한 영역인 ‘대학교’ 영역에서의 재현율을 나타낸 것이다. 실제 대한민국의 대학교 수는 410개에 해당하고, 그 중 후보로 추출된 것은 297개로 재현율은 약 72% 정도에 해당한다.

표 3. 유한한 영역의 재현율

	실제 정답수	추출된 정답수	재현율
대학교	410	297	72.44%

구글 랭킹을 이용한 웹페이지 중요도 적용

[표 4]에서 보는 바와 같이 ‘대학교’ 영역에서는 웹페이지의 중요도를 미적용한 것 보다 적용한 것이 8%가량 상승하였고, ‘영화&드라마’ 제목과 ‘산이름’ 영역의 경우는 각각 약 2%정도의 하락이 있었다.

표 4. 웹페이지 중요도 적용 비교

	웹페이지 중요도 미적용	웹페이지 중요도 적용
대학교(297/644)	76.57%	84.61%
영화&드라마(165/183)	81.84%	80.19%
산이름(100/160)	58.19%	56.02%
평균	72.20%	73.61%

하지만, 전체적인 평균정확률인 MAP의 경우는 웹페이지의 중요도를 적용한 경우가 약 1.4%정도 상승한 것을 볼 수 있다.

그래프 기반의 상호 중요도 측정 기법 적용

[표 5]에서 보는 바와 같이 ‘대학교’ 영역은 상호 중요도 측정 기법을 미적용한 경우보다 약 4.7%가량 상승하였고, ‘영화&드라마’ 영역은 약 4.8%, ‘산이름’ 영역은 약 5.8%가량이 상승한 것을 볼 수 있다. 또한, 전체적인 MAP는 약 5.11%정도 상승하였다.

표 5. 상호 중요도 측정 기법 적용 비교

	상호 중요도 미적용	상호 중요도 적용
대학교(297/644)	84.61%	89.34%
영화&드라마(165/183)	80.19%	85.01%
산이름(100/160)	56.02%	61.82%
평균	73.61%	78.72%

[표 4]과 [표 5]의 영역 이름 옆 괄호안의 수치는 (정답수/후보수)를 의미한다.

영어 개체명 추출 실험 결과

[표 6]는 구글셋과 본 논문에서 제안하는 시스템의 문서 수준 정확률을 비교해 놓은 것이다. 대부분의 영역에서 본 논문에서 제안하는 방식의 문서 수준 정확률이 높게 나왔으며, 그 결과 전체적인 평균에서도 본 논문에서 제안하는 방식이 더 좋은 성능을 보였다.

[표 7]은 MAP 방식을 사용하여 구글셋과 비교해 놓은 것이다. MAP 역시 구글셋보다 본 논문에서 제안하고 있는 상호 중요도를 측정하여 적용하는 방식이 약 4%가량 더 높은 성능을 보였다.

[그림 4]에서는 전체적으로 한국어 영역에서의 MAP와 영어 영역에서의 문서수준 정확률, MAP의 값을 그래프 형태로 나타내었다. 두 분야 모두 본 논문에서 제

안된 방식의 시스템에서 더 높은 성능을 가진다는 것을 알 수 있다.

표 6. 영어 개체명 영역 문서 수준 정확률 비교

영역(영어)	구글셋	상호 중요도 측정 적용
classic-disney	82.05%	85%
mlb-teams	72.97%	83.78%
nba-teams	71.79%	74.36%
nfl-teams	71.11%	71.11%
popular-car-makers	86.67%	97.78%
us-presidents	97.87%	93.62%
us-states	100%	100%
watch-brand	87.80%	100%
평균	83.78%	88.21%

표 7. 영어 개체명 영역 MAP 비교

영역(영어)	구글셋	상호 중요도 측정 적용
classic-disney	80.11%	78.78%
mlb-teams	91.49%	100%
nba-teams	96.3%	100%
nfl-teams	100%	100%
popular-car-makers	82.35%	97.44%
us-presidents	97.47%	95.65%
us-states	100%	100%
watch-brand	93.45%	100%
평균	92.65%	96.48%

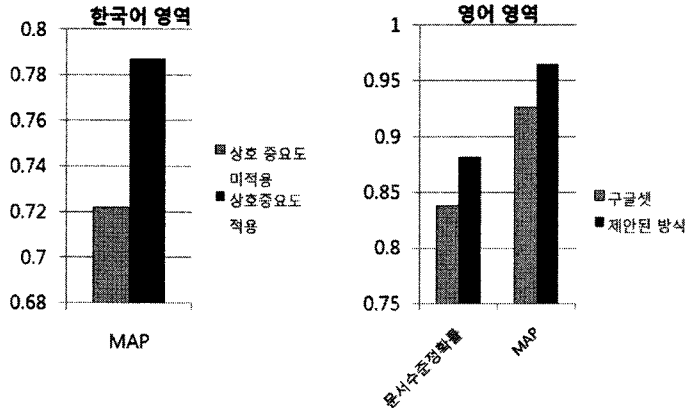


그림 4. 한국어 영역과 영어 영역의 성능 비교

결론 및 향후 연구

본 논문에서는 영역별로 사전을 구축하기 위해서 자동으로 개체명을 추출할 수 있도록 개체명 후보들에게 상호 중요도를 계산한 가중치를 부여하는 방식을 제안하였다. 그 결과 현재 상용화 되고 있는 구글셋과의 성능 평가를 통하여 본 논문에서 제안하고 있는 방식이 구글셋보다 더 높은 성능 향상을 가지고 왔다는 것을 알 수 있었다. 또한, 현재 구글셋에서 서비스하고 있지 않은 한국어 개체명 영역에서도 80%에 가까운 성능을 보이고 있다는 것을 확인하였다.

앞으로 씨앗단어와의 관계가 정확하고 많은 개체명을 추출할 수 있는 웹페이지에 높은 가중치를 부여할 수 있는 기법을 개발한다면 개체명을 더 많고 정확하게 추출할 수 있을 것이다.

참고문헌

[1] 이경희, 이주호, 최명석, 김길창, “한국어 문서에서 개체명 인식에 관한 연구,”

한글 및 한국어 정보처리 학술발표논문집, pp. 292-299, 2000.

- [2] W. J. Black, F. Rinaldi, and D. Mowatt, "Facile: description of the ne system used for muc-7," In *Proceedings of MUC-7*, 1998.
- [3] H. Chen, Y. Ding, S. Tsai, and G. Bian, "Description of the ntu system used for met2," In *Proceedings of MUC-7*, 1998.
- [4] R. C. Wang and W.W. Cohen, "Language -Independent Set Expansion of Named Entities using the Web," In *Proceedings of ICDM*, pp. 342-350, 2007.
- [5] W. W. Cohen, "Automatically Extracting Features for Concept Learning from the Web," In *Proceedings of ICML*, pp. 159-166, 2000.
- [6] 김묘실, 강승식, "SVM을 이용한 악성 댓글 판별 시스템의 설계 및 구현," 한글 및 한국어 정보처리 학술 발표논문집, pp. 285-289, 2006.
- [7] H. Qu, A.L. Pietra, and S. Poon, "Automated Blog Classification: Challenges and Pitfalls," In *Proceedings of The AAAI Spring Symposia on Computational Approaches to Analysing Weblogs*, pp. 184-186, 2006.
- [8] B. Settles, "Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets," In *Proceedings of NLPBA/BioNLP*, pp. 107-110, 2004.

1 차원고접수 : 2008. 11. 4

2 차원고접수 : 2008. 11. 19

최종게재승인 : 2008. 12. 1

(Abstract)

The Set Expansion System Using the Mutual Importance Measurement Method to Automatically Build up Named Entity Domain Dictionaries

Sangjoon Bae

Youngjoong Ko

Dept. of Computer Engineering, Dong-A University

Since Web pages contain a lot of information today, the Web becomes an important resource to extract some information. In this paper, we propose a set expansion system which can automatically extract named entities from the Web. Overall, the proposed method consists of three steps. First of all, Web pages, which may include many named entities of a domain, are collected by using several seed words of the domain. Then some pattern rules are extracted by using seed words and the collected Web pages, and the named entity candidates are selected through applying the extracted pattern rules into Web pages. To distinguish real named entities, we develop the new mutual importance measurement method which estimates the importance of named entity candidates. We conducted experiments for 3 domains for Korean and for 8 domains for English. As a result, the proposed method obtained 78.72% MAP in Korean and 96.48% MAP in English. In particular, the performances of English domains are better than the results of the Google set.

Keywords : Mutual Importance Measurement Method, Domain Named Entity, Named Entity Recognition