그래프 기반의 상호 중요도 측정 기법을 이용한 영역별 개체명 자동 추출

배상준○
동아대학교
컴퓨터공학과
jooniking@gmail.com

고영중
동아대학교
컴퓨터공학과
yjko@dau.ac.kr

Automatic Named Entities Extraction Using the Graph-based Measurement Technique of the Mutual Importance

Sangjoon Bae○
Computer Engineering,
Dong-A University

Youngjoong Ko
Computer Engineering,
Dong-A University

요 약

본 논문에서는 영역별로 자동으로 개체명을 추출하기 위하여 씨앗단어를 이용하고, 웹페이지와 개체명 후보들 간의 상호 중요도를 측정하여 개체명 후보들의 순위를 정하는 방식을 제안한다. 제안된 방식은 크게 세 단계에 의해서 수행되어 지는데 먼저 씨앗단어 정보를 이용하여 웹페이지를 검색하고, 검색되어진 웹페이지와 씨앗단어 정보를 이용하여 패턴 규칙을 추출한다. 추출된 패턴 규칙을 웹페이지에 적용하여 개체명 후보들을 추출하고 추출된 후보들과 웹페이지 사이의 상호 중요도를 재귀적으로 계산하여 최종적으로 개체명 후보들의 순위가 정해 진다. 한국어와 영어 개체명 영역에 제안된 기법을 적용하여 실험한 결과 한국어에서는 78.72%의 MAP를 얻을 수 있었고, 영어에서는 96.48%의 MAP를 얻었다. 특히 영어 개체명 인식에서의 성능은 구글에서 제공하고 있는 구글셋의 결과보다도 높은 성능을 보였다.*

1. 서 론

오늘날 많은 사람들이 웹(Web)으로부터 많은 정보를 얻고 있다. 스포츠 팀이나 영화 제목과 같은 개체명 집합에 대해서 몇 가지 정보만을 사용해서, 그 집합의 모든 원소 정보를 손쉽게 얻을 수 있다면 정보추출에 매우 유용한 도구가 될 수 있을 것이다. 해당 개체명의 집합을 웹을 통해서 자동으로 수집 해주는 도구를 '집합 확장 시스템(Set Expansion System)'이라고 부르며, 이를 통해 매우 쉽게 개체명 사전을 구축할 수 있을 것이다[1]. 집합 확장이란, 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[1, 2]. 집합 확장의 대표적 시스템은 구글(google)에서 제공하고 있는 구글셋(Google Sets^{TM} : http://labs.google.com/sets)이 있다.

이런 집합 확장 시스템은 주로 개체명(Named Entity)을 인식하거나 추출할 경우에 많이 사용되게 된다[1]. 개체명이란, 인명, 지명, 조직명, 상품명 등의 고유 명사를 말한다. 이러한 고유 명사는 대부분의 문서에서 중요한 정보를 제공하지만, 사전에 등록되지 않은 경우가 많다. 고유 명사는 한정된 것이 아니라 계속 생성되기 때문에 모든 고유 명사를 사전에 등록하는 것은 현실적으로 불가능하다[3]. 하지만, 집합 확장 시스템을 이용해서 자동으로 개체명을 추출하여 사전을 구축할 수 있다면, 만들어진 개체명 사전을 이용하여 많은 분야에서 빠르고 정확한 개체명을 인식할 수 있을 것이다[4, 5].

집합 확장 시스템이 유용하게 사용될 수 있는 분야 중 하나가 의견 추출(Opinion Mining) 분야이다[6]. 의견

* 이 논문 또는 저서는 2008년 정부(교육과학기술부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2008-331-D00553)

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 명사이며 개체명이다. 이러한 영역들의 개체명 추출은 의견 추출 시 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 필요하다는 어려움이 있다. 본 논문에서 제안하는 집합 확장 기술을 이용한다면, 새로운 영역의 의견추출 시 필요한 개체명 사전의 구축을 보다 손쉽고 정확하게 수행할 수 있을 것이다.

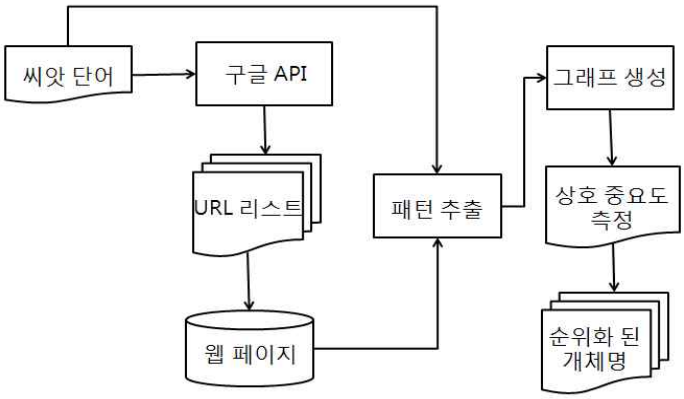
집합 확장 시스템에 관한 연구로는 상용화 된 구글셋 외에 카네기멜론 대학의 SEAL(Set Expander for Any Language)이 있다. 이 시스템은 웹을 이용하여 개체명 대상들을 추출하고 “Laze Work Process“라는 기법을 통해 개체명 대상들의 중요도를 계산하였다[1].

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 실험방법과 실험결과에 대해서 토의하고. 마지막으로 4절에서는 본 논문의 결론에 대해서 설명하고 향후 연구에 대해서 기술한다.

2. 시스템

제안하는 시스템은 [그림1]과 같이 구성되어 있다. 먼저 얻고자 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 48] 전체적인 시스템 구성도

2.1 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. 세 개의 질의를 사용하여 구글 API를 통해 검색된 웹페이지는 상위 100개를 사용하고 있다. 이 상위 100개의 웹페이지는 구글에서 선정한 순위를 사용한다.

2.2 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[1]. 본 논문은 이러한 웹페이지의 정보와 씨앗단어의 정보를 이용한다.

[그림 2]는 한국에 있는 대학교를 추출하기 위해 씨앗단어 ‘동아대학교, 서울대학교, 부산대학교’를 이용하여 검색된 웹페이지 중 하나의 일부분이다. 그림에서 보면 씨앗단어인 ‘동아대학교’와 ‘부산대학교’ 포함된 문장의 구조가 ‘ · 동아대학교’와 ‘ · 부산대학교’로 구성되어 있다. 이처럼 씨앗단어를 중심으로 왼쪽과 오른쪽이 비슷한 구조를 취하고 있다는 것을 알 수

있다. 그리고, 이런 구조를 가진 다른 문장을 보면 추출하고자 하는 대학교의 이름을 가진 문장이 대부분이라는 것을 알 수 있다. 본 논문은 이와 같이 씨앗단어를 가진 문장에서 패턴규칙을 추출하고, 모든 씨앗단어에서 동일하게 추출된 패턴규칙을 그 웹페이지에서 적용시켜 씨앗단어와 유사한 단어들을 추출한다.

```
<b>・동아대학교</b> (산업시스템공학과) 부산 사하구<br>
    <a
href="http://www.donga.ac.kr/"
target=_blank>http://www.donga.ac.kr</a> (<a
href="http://ie.donga.ac.kr/"
target=_blank>http://ie.donga.ac.kr</a>)<br>
<b>・동의대학교</b> (기계・산업시스템공학부) 부산시 부산진구<br>
    <a href="http://www.dongeui.ac.kr/"
target=_blank>http://www.dongeui.ac.kr</a> (<a
href="http://ie.dongeui.ac.kr/"
target=_blank>http://ie.dongeui.ac.kr</a>)<br>
<b>・명지대학교</b> (산업시스템공학부)<br>
    <a href="http://www.mju.ac.kr/"
target=_blank>http://www.mju.ac.kr</a> (<a
href="http://ie.myongji.ac.kr/"
target=_blank>http://ie.myongji.ac.kr</a>)<br>
<b>・부경대학교</b> (산업시스템안전공학부) 부산 남구<br>
    <a
href="http://www.pknu.ac.kr/"
target=_blank>http://www.pknu.ac.kr</a> (<a
href="http://ie1.pknu.ac.kr/"
target=_blank>http://ie1.pknu.ac.kr</a>)<br>
<b>・부산대학교</b> 부산 금정구<br>
    <a href="http://www.pusan.ac.kr/"
target=_blank>http://www.pusan.ac.kr</a> (<a
href="http://www.ie.pusan.ac.kr/"
target=_blank>http://www.ie.pusan.ac.kr</a>)</td>
```

[그림 49] 씨앗단어를 바탕으로 검색된 페이지

다음 3단계의 과정을 거쳐서 개체명 후보를 추출하게 된다. 이 3단계는 검색된 웹페이지 각각에서 적용된다.

2.2.1 단계 1: 씨앗단어 포함 문장 추출

구글 API에서 검색된 웹페이지에서 씨앗단어가 하나라도 포함된 문장을 모두 추출한다.

2.2.2 단계 2: 패턴규칙 추출

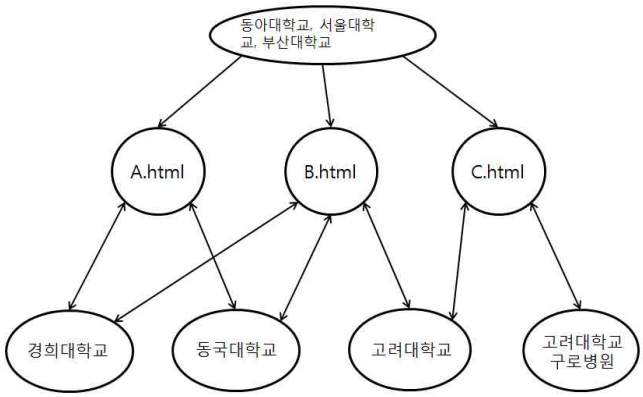
입력으로 사용된 세 개의 씨앗단어를 중심으로 왼쪽 문맥과 오른쪽 문맥을 추출하는데, 이때 문맥 추출 크기로서 문자단위로 2이상, 5이하를 사용한다. 최종 문맥은 세 개의 씨앗단어의 왼쪽 문맥의 교집합과 오른쪽 문맥의 교집합이 각각 선택된다. 이렇게 해서 나온 왼쪽 문맥과 오른쪽 문맥은 해당 웹페이지 안에서의 패턴규칙이 된다.

2.2.3. 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.

2.3 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 50] 상호중요도 계산을 위한 그래프

[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 이름이 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 구현된다. [그림 3]에서 보는 바와 같이 웹페이지와 개체명 후보 간의 중요도 계산을 위

해서 씨앗단어로부터 추출된 웹페이지들의 중요도를 계산하여야 하는데, 본 연구에서는 구글로 부터의 각 웹페이지의 랭킹(ranking) 정보를 이용하여 다음의 식으로 계산이 된다.

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

위 식에서 n는 구글에서 검색된 전체 웹페이지 중에서 씨앗단어가 모두 포함된 웹페이지의 전체 개수이고, i는 구글에서 제공된 웹페이지의 랭킹 정보이다. 그리고 α는 랭킹된 웹페이지 사이의 가중치를 안정화하기 위한 변수로 본 논문에서는 0.6의 값을 사용하였다. 이들 최기 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.

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

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

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

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

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

}

3. 실험 결과

3.1 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 n에서의 정확률이란 상위 n개의 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.

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

위 식에서 L은 개체명 후보들의 순위화 된 리스트이고, r은 순위이며, 정확률(r)은 순위 r에서의 정확률이다. 본 논문에서는 만약 순위 r보다 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.

3.2. 실험 결과 및 토의

본 논문에서는 한국어 개체명 추출을 위한 실험과 영어 개체명 추출을 위한 실험을 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.

3.2.1 한국어 개체명 추출 실험 결과

(1) 구글 랭킹을 이용한 웹페이지 중요도 적용

[표 3]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 3] 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.

(2) 그래프 기반의 상호 중요도 측정 기법 적용

[표 4]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 4] 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 3]과 [표 4]의 영역 이름 옆 괄호안의 수치는 (정답수/후보수)를 의미한다.

3.2.2 영어 개체명 추출 실험 결과

[표 5]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 5] 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%**

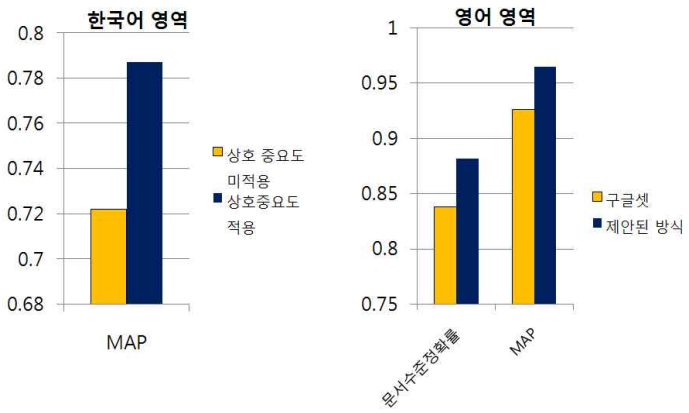
[표 6]은 MAP 방식을 사용하여 구글셋과 비교해 놓은 것이다. MAP 역시 구글셋보다 본 논문에서 제안하고 있

는 상호 중요도를 측정하여 적용하는 방식이 약 4%가량 더 높은 성능을 보였다.

[표 6] 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]에서는 전체적으로 한국어 영역에서의 MAP와 영어 영역에서의 문서수준 정확률, MAP의 값을 그래프 형태로 나타내었다. 두 분야 모두 본 논문에서 제안된 방식의 시스템에서 더 높은 성능을 가진다는 것을 알 수 있다.



[그림 4] 한국어 영역과 영어 영역의 성능 비교

4. 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] 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.

[2] W.W. Cohen, "Automatically Extracting Features for Concept Learning from the Web," In *Proceedings of ICML*, pp. 159-166, 2000.

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

[4] W.J. Black, F. Rinaldi, and D. Mowatt, "Facile: description of the ne system used for muc-7," In *Proceedings of MUC-7*, 1998.

[5] H. Chen, Y. Ding, S. Tsai, and G. Bian, "Description of the ntu system used for met2," In *Proceedings of MUC-7*, 1998.

[6] 김묘실, 강승식, "SVM을 이용한 악성 댓글 판별 시스템의 설계 및 구현," *한글 및 한국어 정보처리 학술발표논문집*, pp. 285-289, 2006.

[7] H. Qu, A.L. Pietra, and S. Poon, "Automated Blog Classification: Challenges and Pitfalls," In *Proceedings of The AAA1 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.