한국어 위키피디아를 이용한 분류체계 생성과 개체명 사전 자동 구축

(Automatic Construction of Class Hierarchies and Named Entity Dictionaries using Korean Wikipedia)

배 상 준 † 고 영 중 ††
(Sangjoon Bae) (Youngjoong Ko)

요 약 위키피디아는 개방형 백과사전으로서 수많은 편집자들에 의해 작성되기 때문에 빠른 시간에 방대한 양의 정보가 축적되고 있으며, 축적되는 정보의 신뢰성 또한 매우 높다. 본 논문에서는 이러한 장점을 가진 위키피디아의 여러 가지 세부정보를 이용하여 한국어 개체명 사전을 자동으로 구축하는 방법을 제안한다. 먼저 위키피디아의 각 엔트리(entry)의 분류정보를 사용하여 분류체계(class hierarchy)를 생성한다. 생성된 분류체계에 위키피디아 엔트리를 자동으로 매핑(mapping)시킨 다음, 분류체계에서 최상위 계층의 불확실성(entropy)을 계산한다. 마지막으로, 임계값 이상의 불확실성을 가지는 분류체계를 제거함으로써 정확률이 높은 개체명 사전을 구축한다. 본 논문에서 제안하는 방법으로 실험을 한 결과 최고 81.12%(83.94%:정확률, 78.48%:재현율)의 F1-measure의 성능을 보였다.

키워드 : 위키피디아, 분류체계, 개체명 사전, 텍스트 마이닝

Abstract Wikipedia as an open encyclopedia contains immense human knowledge written by thousands of volunteer editors and its reliability is also high. In this paper, we propose to automatically construct a Korean named entity dictionary using the several features of the Wikipedia. Firstly, we generate class hierarchies using the class information from each article of Wikipedia. Secondly, the titles of each article are mapped to our class hierarchies, and then we calculate the entropy value of the root node in each class hierarchy. Finally, we construct named entity dictionary with high performance by removing the class hierarchies which have a higher entropy value than threshold. Our experiment results achieved overall F1-measure of 81.12% (precision : 83.94%, recall : 78.48%).

Key words : Wikipedia, Class hierarchy, Named entity dictionary, Text mining

1. 서 론

오늘날 정보가 기하급수적으로 증가함에 따라 정보 집합으로부터 의미 있는 지식을 찾아내는 작업은 많은 분야에서 요구되고 있다. 정보검색이나 정보추출과 같은 자연어 처리의 응용 분야에서는 의미 있는 지식을 찾아내는 작업을 위하여 문서 내의 핵심어를 추출하고 있다. 핵심어는 정보 검색에서 주요 검색대상이 되며, 정보추출 시에는 추출할 정보를 구성하는 요소가 될 수 있다. 이러한 핵심어의 대부분은 인명, 조직명, 지명 등의 개체명(named entity)이다[1]. 개체명은 대부분 문서에서 중요한 역할을 하지만 한정된 것이 아니라 계속 생성되고, 그 수 또한 방대하기 때문에 생성될 때마다 사전에 등록시키는 것은 현실적으로 불가능하다. 따라서, 본 논문에서는 개체명을 효과적이고 지속적으로 추출하여 자동으로 사전을 구축하기 위해 현재 널리 이용되고 있는 위키피디아(wikipedia.org)를 이용하고자 한다.

위키피디아는 현재 가장 큰 온라인 백과사전으로 수많은 편집자에 의해서 다양한 언어로 작성되고 있다[2]. 수많은 편집자들은 이미 위키피디아에 등록된 개체에 대해서는 자신이 얻은 최신의 정보를 업데이트(update)하거나 등록되어 있지 않은 개체의 경우에는 새로운 문서를 작성한다. 또한 편집자들은 그 시대에 잘 알려진 개체에 대해서 문서를 작성하는 경우가 많으므로, 다른 웹사이트(Web site)에서 개체명을 추출하는 것보다 양질의 개체명 사전을 구축할 수 있다. 위키피디아의 큰 장점은 지속적으로 개체명을 모을 수 있다는 것이다. 전 세계의 편집자들은 여러 가지 언어로 빠른 시간에 방대한 양의 문서를 작성하고 있다. 현재까지 한국어 문서에 대한 양은 24만개 정도이지만 영어 문서의 경우에는 100만개가 넘어가는 양의 문서가 작성되어 있다. 그 이유는 위키피디아가 미국에서 시작된 웹사이트이고 영어를 사용하는 국가가 많기 때문이다. 이처럼 한국어 문서

· 이 논문은 동아대학교 학술연구비 지원에 의하여 연구되었음
· 이 논문은 제36회 추계학술발표회에서 '위키피디아로부터의 한국어 개체명 사전 자동 구축'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 동아대학교 컴퓨터공학과
sjbae0525@gmail.com
†† 종신회원 : 동아대학교 컴퓨터공학과 교수
yjko@dau.ac.kr

논문접수 : 2009년 12월 15일
심사완료 : 2010년 2월 2일

의 양도 시간이 갈수록 증가하게 될 것이다. 따라서, 본 논문에서 제안하는 방법을 사용하여 자동으로 개체명 사전을 구축한다면 지속적으로 많은 양의 양질의 개체명을 추가해 나갈 수 있다.

본 논문에서는 다음과 같은 방법에 의해서 개체명 사전을 구축하는 방법을 제안한다. 먼저 위키피디아 문서 내에 있는 분류정보를 추출하여 분류정보의 빈도수를 중심으로 분류체계를 구성한다. 다음으로 분류체계에 엔트리(entry)를 매핑(mapping)시키고, 매핑된 엔트리에 대한 노이즈(noise)를 줄이기 위하여 불확실성(entropy)을 측정한다. 마지막으로 임계값 이상의 불확실성을 가지는 분류체계를 제거함으로써 최종적인 개체명 사전을 구축한다. 그 결과 불확실성 값 1.0에서 최고의 성능을 얻었으며, 83.94%의 정확률과 78.48%의 재현율, 81.12%의 F1-measure 성능을 얻었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 기술하고, 3장에서는 전체적인 제안 방법과 분류체계를 생성하는 방법, 불확실성 측정기법을 이용하여 분류체계 제거하는 방법에 대해서 자세히 기술한다. 4장에서는 본 논문에서 제안하는 방법에 대한 성능을 평가하고, 마지막 5장에서는 향후 연구 방안에 대해서 기술한다.

2. 관련 연구

개체명 사전을 구축하는 연구는 오래전부터 연구되어져 왔다. Riloff & Jones(1999), Agichtein & Gravano (2000), Thelen & Riloff(2002) 등의 연구가 본 연구와 밀접한 관련이 있다[2-4]. 위 연구의 중점 사항은 웹페이지에서 개체명 추출을 위한 높은 정확도를 가진 패턴을 만들기 위해 문법적인 정보와 통계적인 기법을 결합한 방법에 대해서 연구하고 있다. 하지만 위의 연구들은 영어권에서만 속하는 언어 의존적인 정보들을 사용하므로 한국어에 대한 개체명을 추출하기 위해서는 불합리적이다. 또한 본 논문에서 제안하는 방식은 위키피디아라는 신뢰도가 높은 웹페이지로부터 개체명을 추출하는 것이기 때문에 높은 성능을 기대할 수 있다.

위키피디아에 대한 연구는 현재 국외에서 활발히 진행중이다[5,6]. 그 중에서 Richman & Schone(2008)의 연구는 본 논문과 밀접한 관련이 있다[7]. 이 연구는 위키피디아 문서에서 개체명 인식을 위한 데이터들의 활용 방안에 대해서 제안하고 있다. 위키피디아 문서에는 여러 가지의 링크(Link) 정보를 포함하고 있다. 이러한 링크 정보를 사용하여 언어학적인 패턴을 생성하고, 패턴에 일치하는 개체명들에 대한 인식이 이루어진다. 이에 비해 본 논문은 여러 가지 링크 정보 중에서 분류정보 링크를 이용하여 체계를 생성하고, 생성된 분류체계를 통한 개체명 사전을 구축하는 방법을 제안한다.

3. 제안 방법

제안하는 방법은 그림 1과 같이 구성되어 있다. 먼저 위키피디아 문서 내의 여러 가지 자원으로부터 분류정보를 추출하여 분류체계를 생성한다. 그리고 생성된 분류체계에 개체명에 해당하는 위키피디아 문서의 엔트리를 매핑시킨다. 매핑된 개체명들은 노이즈를 포함하고 있기 때문에 불확실성 측정을 통하여 임계값 이상의 엔트로피 값을 가지는 분류체계를 제거하고, 최종적인 개체명 사전을 구축한다.

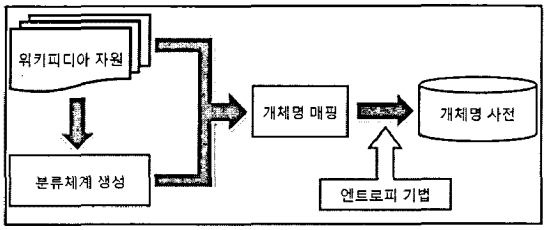

그림 1 전체적인 제안 방법 구성도

3.1 위키피디아 자원

위키피디아 문서 내에는 유용하게 사용될 수 있는 많은 정보들이 있다. 개체명에 해당하는 엔트리와 개체명을 설명하고 있는 본문정보, 그리고 비슷한 엔트리들을 분류 해놓은 분류정보 등이 있다. 이들 중에서 본 논문에서는 분류정보를 사용하여 분류체계를 생성한다.

3.2 분류체계 생성

위키피디아의 각 엔트리는 분류정보를 가지고 있다. 이러한 분류정보는 엔트리에 대한 특징을 설명하고 있기 때문에 같은 분류정보 내의 엔트리들은 같은 타입의 개체명이라고 생각할 수 있다. 하지만, 분류정보의 수가 너무 많기 때문에 각각의 분류정보에 대해서 개체명 타입을 결정하는 일은 시간과 비용이 많이 든다. 따라서 본 논문에서는 이러한 분류정보들을 사용해서 자동으로 분류체계를 생성하고 각 엔트리를 분류체계에 매핑한다.

분류정보의 체계를 생성하기 위하여 다음과 같이 가정한다.

1. 분류정보 내에서 가장 우측에 있는 단어가 중요하다.
2. 출현 빈도수가 높은 분류정보 내의 단어가 상위 계층에 위치한다.

분류체계를 생성하기 위하여 위키피디아 문서에 있는 분류정보를 추출하고, 위의 가정을 바탕으로 분류체계를 생성한다. 예를 들어, 그림 2와 같이 분류정보가 있을 때 각 분류정보들을 오른쪽부터 띄어쓰기를 기준으로 구분을 하면, '선수', '야구 선수', '미국의 야구 선수', '선수', '골프 선수', '여자 골프 선수', '선수', '축구 선수', '리버풀의 축구 선수'와 같이 나누어질 수 있고, 각 단어

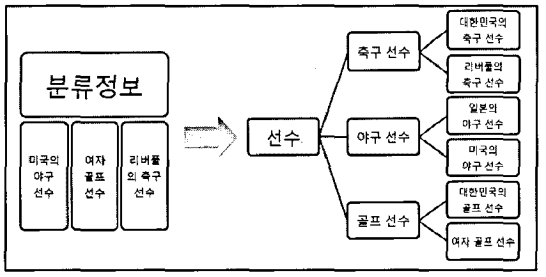

그림 2 분류체계 생성의 예

들의 출현 빈도수를 조사하여 빈도수가 높은 단어가 상위 계층이 된다. 본 논문에서는 최대 3계층까지의 분류체계만을 생성한다. 따라서 그림 2의 오른쪽과 같은 계층이 생성된다.

3.3 개체명 매핑

개체명을 추출하기 위해서는 생성된 분류체계에 위키피디아 엔트리들을 매핑하는 작업이 필요하다. 만약 위키피디아 문서 내의 엔트리가 생성된 분류체계의 최하위 계층의 분류정보를 가지고 있다면 최하위 계층에 엔트리를 매핑시킨다. 예를 들어, '이운재'라는 엔트리가 '대한민국의 축구 선수'라는 분류정보를 가지고 있으면 '[선수]-[축구 선수]-[대한민국의 축구 선수]-이운재'처럼 매핑된다.

3.4 불확실성 측정기법

이렇게 생성된 분류체계에는 양질이 아니어서 제거해야 하는 분류체계가 있다. 이러한 분류체계의 특징은 여러 가지 형태(type)의 개체명을 포함하고 있다는 것이다. 이를 자동으로 측정하고 불량의 분류체계를 자동으로 제거하기 위해서 본 논문에서는 이미 구축되어 있는 ETRI 개체명 사전과 엔트로피(entropy)개념을 이용한 불확실성 측정기법을 사용한다.

엔트로피는 정보를 내보내는 근원의 불확실성을 나타내는 양을 뜻한다[8]. 정보의 양이 많을수록 불확실성이 높아지게 되고 엔트로피의 값은 증가하게 된다. 본 논문의 경우에는 하나의 분류체계에 여러 종류의 개체명 형태가 매핑되어 있는 경우에 불확실성이 높아진다. 그리고, 이러한 엔트로피를 계산하려면 이미 분류되어 있는 개체명 사전이 필요하다. 본 논문에서는 ETRI 개체명 사전을 이용하여 ETRI 개체명 사전과 위키피디아에 공통으로 출현한 엔트리들을 정답집합으로 고려하여 엔트로피도 구하고 최종 성능도 측정하였다. 만약 어느 분류체계에 ETRI 개체명 사전에 의해서 "조직명", "인공물", "지역명" 등과 같은 여러 가지 형태를 가지는 개체명들이 포함되어 있다면, 이 분류체계는 정확한 정보를 제공하고 있는 것이 아니기 때문에 엔트로피 값이 높게 측정되고, 높은 엔트로피 값을 가지는 분류체계를 제거함으로써 정확한 분류체계를 생성할 수 있다.

최종 엔트로피는 다음과 같은 수식에 의해서 표현된다.

$$H(x)=\sum_{i=1}^{n}p(i)\log_2\left(\frac{1}{p(i)}\right)=-\sum_{i=1}^{n}p(i)\log_2 p(i) \quad (1)$$

위 식에서 x는 분류체계를 표시하고, i는 매핑되어 있는 개체명 타입을 말하며, $p(i)$는 i가 일어날 확률을 말한다.

본 논문에서는 최상위 계층에 대해서 엔트로피 값을 계산하고, 실험을 통해서 엔트로피의 임계값을 설정하였다. 그 실험에 대한 결과는 4절 성능 평가 부분에서 설명할 것이다.

4. 실험 성능 평가

4.1 실험 데이터

위키피디아는 위키피디아 내에서 작성된 문서를 xml 형식의 파일로 제공하고 있다(http://download.wikipedia.org). 본 논문에서 제안하는 방법의 실험을 위해서 2009년 1월 19일에 제공된 파일을 사용하였다.

표 1 위키피디아 엔트리 수

내용	개수
위키피디아 전체 엔트리 수	235,093
불용어를 제외한 엔트리 수	77,656
위키피디아 엔트리∩ETRI 개체명 사전	26,090

표 1과 같이 위키피디아의 전체 엔트리 수는 235,093개이다. 하지만 이러한 엔트리에는 토론이나 틀과 같은 엔트리를 설명하기 위한 문서가 아닌 편집자들을 위한 것이나 위키피디아 자체적으로 사용하는 엔트리도 포함되어 있고, 분류정보가 없는 엔트리도 있다. 본 논문에서는 개체명 사전을 만들기 위해서 위키피디아의 분류정보를 사용하므로 위와 같은 불용어가 들어있는 엔트리나 분류정보가 없는 엔트리는 사용하지 않는다.

본 논문에서 제안하는 방법을 실험하기 위하여 ETRI에서 제공하는 개체명 사전을 이용하였다. 이 개체명 사전은 17개의 대분류와 139개의 소분류로 나뉘어져 있다. 이 개체명 사전을 이용하여 ETRI 개체명 사전의 개체명과 분류계층에 매핑된 엔트리가 공통으로 속해있는 개체명만을 가지고 실험을 하였다. 여기에서 분류계층에 매핑된 엔트리의 정답 태그(tag)를 ETRI 개체명 사전에 17개의 대분류 타입으로 태깅(tagging)하고 이것을 전체 정답셋으로 정하였다. 그 수는 표 1과 같이 26,090개이다.

4.2 전체 분류정보 사용

본 논문에서 제안하는 방법의 순서에 따라서 전체 분류정보를 추출하여 분류체계를 만들고 엔트리를 매핑하는 과정을 거친다. 매핑된 엔트리 중에 ETRI 개체명

사전의 등록이 되어 있는 엔트리는 ETRI 개체명 태그를 부여하여 성능을 측정하였다.

표 2는 위키피디아 문서에서 출현하는 모든 분류정보에 대해서 3계층까지의 분류체계를 만들고 실험을 한 결과이다. 분류계층에 매핑된 위키피디아 엔트리는 최상위 계층의 태그를 부여받기 때문에 다중으로 개체명 타입을 가질 수 있다. 예를 들어, '경복궁'은 생성된 분류체계의 최상위 계층에 따라서 '인공물'과 '문명/문화'의 태그를 부여받을 수 있다. 본 논문에서는 이와 같은 경우 구분을 하지 않고 전부 추출한 다음, 성능 평가 시에는 ETRI 개체명 사전에 태깅된 분류 태그가 나왔을 시에만 맞는 정답으로 간주하였다. 위 예에서 '경복궁'의 경우 ETRI 개체명 사전에서 '인공물'로 태깅되어 있다면 '문명/문화'가 태깅된 '경복궁'은 틀린 정답으로 처리된다.

표 2 전체 분류정보를 사용한 실험 결과

내용	개수
분류계층에 매핑되는 단일 타입의 엔트리	71,344
위키피디아 엔트리∩ETRI 개체명 사전 (단일 타입)	26,090
위키피디아 엔트리∩ETRI 개체명 사전 (다중 타입 포함)	32,138
맞는 정답 개수	22,016
정확률	**68.505%**
재현율	**84.385%**
F1-measure	**75.62%**

위 표에서 정확률과 재현율은 다음과 같이 평가된다.

$$\text{정확률} = \frac{\text{맞는 정답 개수}}{\text{위키피디아 엔트리} \cap \textit{ETRI} \text{개체명 사전(다중 타입 포함)}}$$

$$\text{재현률} = \frac{\text{맞는 정답 개수}}{\text{위키피디아 엔트리} \cap \textit{ETRI} \text{개체명 사전(단일 타입)}}$$

본 논문에서는 다중 타입의 엔트리에 대해서 구분하고 있지 않기 때문에 정확률을 구하기 위해서 다중 타입을 포함하는 위키피디아 엔트리와 ETRI 개체명 사전의 공통 개체명을 사용하였다. 재현율은 4.1절에서 설명한 전체 정답셋을 사용하였다. 그 결과 75.62%의 F1-measure 성능을 얻을 수 있었다.

4.3 단어 수가 하나인 분류정보 제외

오류분석을 통해 살펴본 결과 분류정보가 하나의 단어로 이루어진 경우가 오류를 많이 일으키는 것으로 관찰되었다. 이를 해소하기 위해서 단어수가 하나인 분류정보를 제외하고 분류체계를 생성하여 위와 같은 과정의 실험을 해보았다. 그 결과는 표 3과 같다.

표 3에서 보는 것처럼 단어 수가 하나인 분류정보를 제외했을 경우 그렇지 않은 경우보다 2%가량이 오른 77.721%의 F1-measure 성능을 보였다.

표 3 단어 수가 하나인 분류정보를 제외한 실험 결과

내용	개수
분류계층에 매핑되는 단일 타입의 엔트리	57,078
위키피디아 엔트리∩ETRI 개체명 사전 (단일 타입)	20,993
위키피디아 엔트리∩ETRI 개체명 사전 (다중 타입 포함)	24,722
맞는 정답 개수	17,765
정확률	**71.859%**
재현율	**84.623%**
F1-measure	**77.721%**

4.4 불확실성 측정을 이용한 분류체계 제거

다중 타입의 엔트리를 제외하기 위해서 엔트로피 기법을 이용하여 분류체계의 최상위 계층에 엔트로피 값을 부여하고 임계값을 찾기 위한 실험을 하였다. 그 결과는 표 4와 같다.

표 4 엔트로피 값의 임계값을 찾기 위한 실험 결과

엔트로피 값	정확률	재현율	F1-measure
0.8	88.519%	74.115%	80.679%
0.9	86.474%	75.463%	80.594%
1.0	**83.937%**	**78.483%**	**81.119%**
1.1	82.362%	79.765%	81.042%
1.2	81.055%	80.174%	80.612%
1.3	80.003%	80.765%	80.382%

위 표에서와 같이 임계값이 1.0일때 F1-measure의 성능이 가장 좋았다. 이 실험 결과를 바탕으로 1.0 이하의 엔트로피 값을 가지는 분류체계에 매핑된 엔트리에 대해서 81.119%는 정답 태그가 부여된다는 것을 유추할 수 있다. 표 5는 각 임계값에서의 분류체계에 매핑되고 ETRI 개체명 사전에 포함되지 않는 엔트리를 포함하는 모든 엔트리에 대한 개수이다.

표 5 엔트로피 값에 따른 분류체계에 매핑된 엔트리 개수

엔트로피 값	분류체계에 매핑된 모든 엔트리 개수
0.8	39,961
0.9	42,079
1.0	**44,886**
1.1	46,183
1.2	46,818
1.3	47,618

5. 결론 및 향후 연구

본 논문에서는 위키피디아의 자원을 이용하여 분류체계를 생성하고 개체명 사전을 구축하는 방법을 제안하고 있다. 분류체계를 생성하고 엔트리를 매핑하는 과정

에서 나타나는 오류를 분석해본 결과 분류정보의 단어 수가 영향을 미친다는 것을 알았다. 또한 다중으로 개체명 타입이 태깅되는 엔트리에 대해서 분류계층의 최상위 계층의 엔트로피 값을 계산하고 분류체계를 제거하기 위한 임계값을 실험을 통해서 찾아내었다.

향후에는 아직까지 다중으로 개체명 타입이 태깅되는 엔트리를 줄일 수 있는 방법을 연구하고, 최하위 계층의 엔트로피 값을 계산해서 최하위 계층의 분류체계를 제거 할 수 있는 방법을 연구하겠다. 또한 위키피디아에서 제공하는 다른 자원 중에서 개체명 사전을 구축하는 데 영향을 줄 수 있는 자원의 모색도 필요하겠다.

참 고 문 헌

[1] K. Lee, J. Lee, M. Chol, G. Kim, "Study on Named Entity Recognition in Korean Text," *Proc. of the Annual Conference on Human Cognitive Language Technology*, vol.21, no.1(C), pp.292-299, 2000. (in Korean)

[2] E. Riloff And R. Jones, "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping," *Proc. of the Sixteenth National Conference on Artificial Intelligence*, pp.474-479, 1999.

[3] E. Agichtein And L. Gravano, "Snowball: extracting relations from large plain-text collections," *Comm. ACM*, pp.85-94, 2000.

[4] M. Thelen And E. Riloff, "A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts," *Proc. of the Conference on EMNLP*, pp.214-221, 2002.

[5] W. Dakka And S. Cucerzan, "Augmenting Wikipedia with Named Entity Tags," *Proc. of the IJCNLP*, pp.545-552, 2008.

[6] S. Ye , T. Seng, J. Iu, "Summarizing Definition from Wikipedia," *Proc. of the ACL-IJCNLP*, pp.199-207, 2009.

[7] A. Richman And P. Schone, "Mining Wiki Resources for Multilingual Named Entity Recognition," *Proc. of the ACL*, pp.1-9, 2008.

[8] A. L. Berger, S. A. Della Pietra, S. A. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Proc. of the Computational Linguistics*, pp.39-71, 1996.