2000년 한국정보처리학회 추계 학술발표논문집 제 7 권 제 2 호

한국어 구문 분석기를 이용한 지명 추정 시스템 설계 및 구현

이현숙*, 하유선*, 김태현*, 이만호*, 맹성현*
*충남대학교 컴퓨터과학과
e-mail : {hslee, ysha, heemang, mhlee, shmyaeng}@cs.cnu.ac.kr

A Method for Unknown-Word Extraction from Korean Text

Hyun-Suk Lee*, You-Sun Ha*, Tae-Hyun Kim*, Mann-ho Lee*, SungHyon Myaeng*
*Dept. of Computer Science, Chungnam National University

요 약

본 논문에서는 학습데이터를 이용하여 텍스트로부터 미등록 고유명사를 추정하는 방법을 제안한다. 고유명사 추정을 위해 먼저 형태소 분석기를 이용하여 품사가 명사인 단어들을 후보단어로 선택한다. 이렇게 선택된 후보단어가 고유명사인지 추정해 보기 위해 학습데이터를 이용하여 구성한 정보집합을 사용한다. 이러한 정보집합으로는 이름집합, 접미사집합, 단서집합, 배제어집합이 있다. 본 논문에서는 이런 정보를 이용하여 한국어 지명을 추정하는 시스템을 구현하여 실험한 결과 77.2%의 정확도와 84.9%의 재현율을 보였다.

1. 서론

정보 검색은 색인어 추출 과정을 통해 이루어진다. 색인어는 문서들을 구별짓거나 문서의 전체적인 내용을 나타내기 때문에 정확하고 의미있는 색인이 중요하다.[2,3]

한국어 문서에서는 주로 명사가 문서의 내용을 나타낸다. 그러므로 명사 색인어는 문서의 특성을 잘 표현해 준다. 특히 고유명사는 사람이나 사물 등의 고유한 의미를 나타내 주기 때문에 문서에서 변별력을 높여 주는 역할을 할 수 있다. 이러한 미등록 고유 명사의 처리는 색인의 정확도를 높이는데 기여한다. 그러나 대부분의 고유명사는 사전에 등록되어 있지 않기 때문에 자동적으로 추출하는 것이 용이하지 않다.[1,2,4]

이처럼 한국어 색인시, 문서에 나타나는 명사들을 인식하고 이를 색인어로 추출하기 위해 형태소 분석과 미등록어 처리가 아직 어려운 문제로 남아 있다.

본 연구에서는 이러한 미등록 고유명사의 추정을 위해 통계 정보와 구문 분석을 이용하는 방식을 제안하였다.

2. 미등록 고유명사 추정기

미등록어에는 고유명사, 신조어, 전문용어, 약어, 의성어, 의태어, 복합동사, 조사, 어미 등이 있다. 명사 이외의 다른 품사의 미등록어는 그 빈도수가 매우 낮기 때문에 미등록어 추정은 일반적으로 입력된 문장에서 미등록 명사를 추정하는 데 중점을 둔다. 미등록 명사 중 고유명사는 새로이 만들어지는 것이 많기 때문에 사전을 만들어 모두 처리할 수는 없다.[2]

본 논문에서는 미등록 고유명사 중에서 지명의 특징을 살펴 보고 지명과 관련된 고유명사를 추정할 수 있는 방법을 기술하며 시스템을 구현하여 평가한 결과를 보인다.

2.1 지명의 특성

한국어의 지명은 대부분 단일 명사 형태이고 크게 행정구역이나 지리적 특성을 나타내는 접미사를 동반하는 경우와 접미사 없이 단독으로 사용되는 경우로 나눌 수 있다. 지명에 포함되는 접미사 중 행정구역을 나타내는 것은 "도", "시", "구", "군" 등이 있고 지리적 특성을 나타내는 것으로는 "산", "강", "해수욕장", "만", "항", "골" 등이 있다. 예를 들면 접미사를 포함한 지명으로는 "대전시", "유성구", "계룡산", "금강", "아산만", "대천해수욕장" 등이 있고 단독으로 사용되는 지명으로는 "서울", "경기", "철원", "강릉", "오죽헌" 등이 있다.

그러므로 어떤 문장에서 지명을 추정하기 위해서는 지명에 사용되는 접미사집합과 실제 이름집합이 있어야 한다. 이 때, 지명과 관련된 접미사를 포함하는 단어들 중에는 지명이 아닌 단어들도 있으므로 이러한 단어들로 배제어집합을 구성하여 이용하면 정확도를 높일 수 있다. 또한 지명이 나오는 문장에서 자주 나타나는 단어들 중 지명일 가능성을 높여 주는 단어들의 집합, 즉 단서 집합을 이용하는 것도 유용하다.

본 연구에서는 학습 데이터를 구문분석하여 접사집합, 이름집합, 단서집합, 배제어집합을 반자동으로 구성하였다. 접미사집합은 지명에 사용되는 접미사들로 이루어진다. 이름집합은 각 접미사별로 구성하였고 접미사가 없는 지명들은 따로 모아 놓았다. 그리고 지명과 함께 나오는 단서가 될 수 있는 명사들로 단서집합을 구성하였다. 배제어집합은 지명접미사를 포함하지만 지명이 아닌 단어들의 집합이다. 배제어집합도 이름집합과 마찬가지로 각 접미사별로 구성하였는데 예를 들면 "산" 접미사의 배제어로는 "명산", "등산", "뒷동산" 등이 있고, "리" 접미사의 배제어로는 "다리", "머리", "소리" 등이 있다.

2.2 접근 방법

이 논문에서 제안한 고유명사 추정 방법은 형태소 분석과 지명 추정의 두 단계로 이루어진다. [그림 1]은 본 시스템에서 지명을 추정하는 과정을 도식화한 그림이다.

문장에서 지명을 식별하려면 먼저 지명이 될 수 있는 품사인 명사를 찾기 위해 형태소 분석을 한다. 형태소 분석에는 한성 대학교의 강승식 교수가 개발한 HAM(Hangul Analysis Module)시스템을 사용하였다. 이 분석을 통해 지명후보를 선택한 다음 이 후보단어가 실제 지명인지 아닌지를 추정하기 위해 다음의 단계를 거친다.

첫째, 접미사사전을 이용하여 지명후보에 지명을 나타내는 접미사가 있는지를 찾아 본다. 접미사가 있으면 그 접미사에 해당하는 이름집합에 지명후보가 존재하면 지명으로 확정한다. 접미사가 없는 경우에는 그에 해당하는 사전에서 찾아본다.

둘째, 위에서 지명으로 선택되지 않은 경우, 즉 접미사는 있으나 지명사전에 없는 경우에는 그 문장에서 단서가 되는 단어들의 통계적인 정보를 이용하여

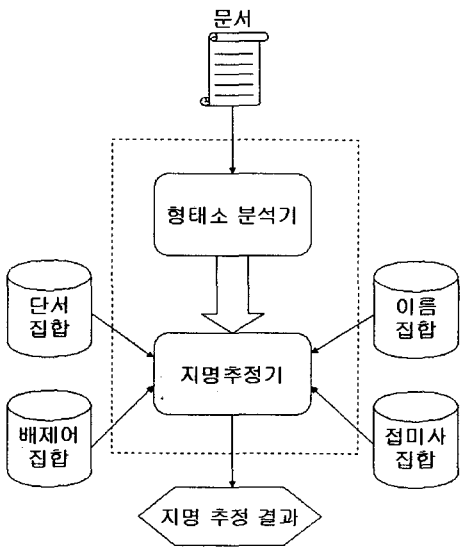

그림 1 고유명사 추정기

지명인지 아닌지의 여부를 결정한다. 문장의 단어들 중 단서가 되는 단어의 비율이 특정 한계치 이상일 경우에 지명으로 추정한다. 이 때, 그 후보가 배제어 집합에 들어 있다면 지명이 될 수 없다.

셋째, 위의 두 단계에서 지명으로 선택되지 않은 경우, 즉 접미사도 없고 지명사전에도 없는 경우에는 단서집합만을 이용한다. 단서집합을 이용하는 방법은 두번째 단계와 같으나 접미사가 없기 때문에 다른 한계치를 사용하여 지명을 추정한다.

단서집합을 이용할 때의 한계치는 접미사가 있는 경우는 40%, 접미사가 없는 경우는 60%로 정하였다. 이 한계치들은 시스템에서 여러 수치를 적용해 본 결과 가장 적절한 값으로 정한 것이다. 다음 2.3 절에서는 실제로 지명을 추정하는 예를 살펴본다.

2.3 지명 추정

예를 들어 다음 문장에서 지명을 추정하는 과정을 살펴 보자.

"강원도 태백산맥 자락에 위치한 불국사에는 석굴암이 자리잡고 있다."

먼저 형태소 분석을 하면 다음과 같다.

```
강원도  (N "강원")<N:20> + (j "도")
        (N "강원도")<N:20>
태백산맥 (N "태백산맥")< :60>
자락에 (N "자락")<N:20> + (j "에")
위치한 (N "위치")<N:29> + (t "하") + (e "ㄴ")<13>
불국사에는 (N "불국사")<N:20> + (j "에는")
석굴암이 (N "석굴암")<N:20> + (j "이")
```

자리잡고 (V "자리잡")<I:20> + (e "고")
있다 (V "있")<KIgVJ:20> + (e "다")
. (P ".")< :0>

위의 분석에서 지명후보는 명사인 "강원도", "태백산맥", "자락", "위치", "불국사", "석굴암"이다. 이들 중 "강원도"는 명사 "강원"과 조사 "도"로 분석되지만 조사 "도"가 지명접미사가 될 수 있으므로 "강원도"를 후보로 선택한다. 이제 이 후보들을 앞에서 언급한 세 단계에 적용시킨다.

먼저, "강원도"를 보면 "도"가 접미사가 될 수 있으므로 "도"에 해당하는 이름집합에서 "강원"을 찾아 본다. "강원"이 있으면 "강원도"는 지명으로 추정되고 없으면 다음 단계로 넘어간다. 다음 단계에서는 "강원"이 "도"에 해당하는 배제어집합에 있으면 지명이 아니라고 결정된다. 그렇지 않으면 강원도 주변의 단어들 중 단서인 것의 백분율을 구한다. 이 값이 40% 이상이면 "강원도"는 지명이라고 추정되고 그 이하이면 그 다음 단계를 적용한다. 이 때는 주변의 단어들 중 단서인 것의 백분율만을 이용하는데 이 경우는 접미사가 없는 경우이므로 한계값을 60%로 한다.

다음으로 "태백산맥", "불국사", "석굴암"은 각각 "산맥", "사", "암"을 접미사로 갖는다. 이 후보들도 접미사를 갖기 때문에 위의 "강원도"와 같은 방식으로 추정한다.

접미사를 갖지 않는 후보인 "자락"과 "위치"는 첫 번째와 두번째 단계를 거치지 않고 단서만을 이용하여 지명 여부를 결정한다.

3. 실험 및 결과

3.1 실험

본 시스템은 LINUX 상에서 C 언어와 mSQL 을 이용하여 구현하였다.

지명과 관련된 학습 데이터로는 국내여행정보 관련 사이트에서 수집한 600 개의 문서와 우편번호부에 나오는 정보를 사용하였다. 우선 이 학습 데이터에서 반자동으로 지명, 접미사, 단서, 배제어가 될 단어를 뽑아서 데이터베이스를 구성하였다. 학습 데이터의 총 크기는 641KB 이고 문서 당 평균 크기는 대략 1.1KB 이다. 학습 데이터를 통해 수집된 지명의 수는 15510 개, 접미사는 70 개, 단서는 282 개, 배제어는 742 개이다.

테스트데이터는 스포츠투데이의 여행 관련 기사 20 개와 스포츠조선의 여행 관련 기사 80 개를 대상으로 하였다. 이 기사들을 이용하여 본 논문에서 구현한 시스템이 어느 정도의 정확도와 재현율을 보이는지 실험해 보았다.

3.2 정확도와 재현율

본 논문에서 제시한 방법을 사용한 실험 결과는 [표 1]과 같다.

	문서에 있는 지명	추정한 지명	제대로 추정한 지명	추정하지 못한 지명
			잘못 추정한 지명	
총 개수	338	372	287	51
			85	

표 1 지명 추정 결과

위의 결과에서 구한 정확도와 재현률은 각각 77.2%, 84.9%였다.

위의 실험에 추가하여 본 시스템에서 사용한 각 집합이 지명 추정에 어느 정도 기여하는지를 알아 보기 위해 이름집합을 제외한 경우, 배제어집합을 제외한 경우, 단서집합을 제외한 경우의 성능을 비교해 보았다. 각각의 결과는 아래의 [표 2]와 같다.

	정확도 (precision rate)	재현율 (recall rate)
모두 사용	77.2%	84.9%
이름집합 제외	69.6 (%)	75.4 (%)
배제어집합 제외	66.2 (%)	84.9 (%)
단서집합 제외	72.8 (%)	85.8 (%)

표 2 지명 추정 결과의 정확도와 재현율

[표 2]에 나타난 바와 같이 이름집합을 제외하고 지명을 추정하였을 경우 정확도와 재현율이 모두 낮아졌다. 이름집합에는 실제 지명이 들어있는데 이 정보를 이용하지 못하기 때문이다. 특히 접사가 없는 지명의 경우, 이름집합을 사용하지 않으면 단서 정보에만 의존하여 추정하므로 단서가 적을 경우 기각될 확률이 높아진다. 배제어 집합은 재현율에는 별로 영향을 주지 않지만 사용하지 않으면 정확도가 떨어진다. 그 이유는 지명과 관련된 접미사를 가지고 있지만 지명이 아닌 단어들, 즉 배제어들을 제외할 수 없기 때문이다. 그리고 단서집합을 사용하지 않고 지명을 추정하였을 때에는 정확도는 낮아지지만 재현율은 약간 높아지는 경향을 보인다. 이것은 문장에 단서로 쓰이는 단어가 많지 않아 한계값을 넘지 못해 기각되었던 지명후보들이 지명이라고 추정되기 때문이다. 단서 집합을 사용하지 않으면 본래 시스템보다 재현율이 조금 좋아질 수는 있지만 정확도가 감소하므로 단서집합을 사용하는 것이 더 좋다고 할 수 있다.

3.3 잘못 추정한 경우

문장에 철자법 오류가 있다든지 띄어쓰기가 잘못된 경우와 형태소 분석기 자체의 오류 때문에 분석이 제대로 되지 않을 때 잘못된 지명을 추정할 수 있다. 이런 문제는 철자 교정과 띄어쓰기를 제대로 해주고, 형태소 분석기의 오류를 수정해 주면 해결될 수 있다. 그리고 이름집합에 없는 지명후보를 접미사와 단서를

이용하여 추정할 때에도 잘못된 지명을 추정할 수 있다. 이런 경우 잘못 추정하게 되는 가장 큰 요인은 실제 지명이 아니더라도 그 문장에 단서집합에 있는 단어가 많이 나타나기 때문이다. 그러나 궁극적으로는 지명 추정에 이용할 정보의 부족에서 기인한다고 볼 수 있다. 이를 해결하기 위해서는 더 많은 학습데이터를 이용하여 정보집합을 늘려주어야 한다.

4. 결론

본 연구에서는 형태소 분석과 학습데이터에서 수집한 정보를 이용하여 자동으로 한국어 미등록 고유명사를 추정하는 방법을 제안하였다. 그리고 이를 지명 추정에 적용해 보았다.

학습데이터에서 구문분석을 통해 수집한 정보는 이름집합, 접미사집합, 단서집합, 배제어집합이라는 4 개의 정보집합을 구성하는데 사용되었다. 지명 추정은 이름집합을 이용하는 단계, 접미사집합을 이용하는 단계, 단서집합만을 이용하는 단계가 있다. 배제어집합은 지명과 관련된 접미사를 가지지만 지명이 아닌 단어들이 지명으로 추정될 확률을 낮춰 준다.

이런 방식으로 각 단계에서 문서에 있는 지명을 추정하여 보고 각 집합이 지명 추정에서 어떠한 역할을 하는지 알아보았다. 정확도를 더 높이기 위해서는 정보집합의 확장과 고유명사 후보 선택을 위해 형태소 분석기의 기능의 강화가 필요할 것으로 보여진다.

향후 과제로는 이 시스템을 다른 고유명사 집합에도 적용할 수 있도록 일반화하는 것과 정보집합의 확장을 자동화하는 연구가 필요하다.

참고문헌

[1] 양장모, 김민정, 권혁철, "언어 정보를 이용한 한국어 미등록어 추정", 정보과학회 봄 학술발표논문지, 1996

[2] 정래정, 김준태, "통계 정보와 어의 정보를 이용한 미등록 고유 명사의 자동 색인"

[3] Paul Thompson, Christopher C. Dozier, "Name Recognition and Retrieval Performance"

[4] 강승식, "한국어의 형태론적 특성과 형태소 분석 기법", 정보과학회지, 12 권 8 호, 1994