데이터 집합을 이용한 고유명사 추출

김태현[0] 이현숙 하유선 이만호 맹성현
충남대학교 컴퓨터과학과
{heemang, hslee, ysha, mhlee, shmyaeng}@cs.cnu.ac.kr

Proper Noun Extraction Using Data Sets

Tae-Hyun Kim[0] Hyun-Suk Lee You-Sun Ha Mann-Ho Lee Sung-Hyon Myaeng
Dept. of Computer Science, Chungnam National University

요 약

본 논문에서는 한국어 고유명사의 특징에 대해 살펴보고, 이를 기반으로 문서로부터 고유명사를 추출하기 위한 기본 모델을 제안한다. 고유명사는 문서의 내용을 대표하는데 주도적인 역할을 하기 때문에, 이를 효과적으로 추출해내는 것은 문서의 의미를 보다 정확하게 표현할 수 있는 방법이 될 수 있다. 문서에서 고유명사를 효과적으로 추출할 수 있도록 하기 위해, 본 연구에서는 이름집합, 접사집합, 단서집합을 구성할 수 있는 데이터 수집기 모델과 데이터 집합을 기본으로 이용하여 고유명사를 추출하는 고유명사 추출기 모델을 제안하였다. 그리고, 실제로 이 모델을 적용하여, 회사명과 관련된 데이터를 수집하고, 이를 이용해 문서로부터 회사명을 추출할 수 있도록 하는 시스템을 구현하였다. 구현된 회사명 추출 시스템을 이용해 고유명사 추출 실험을 수행한 결과, 외래어로 이루어진 회사명으로 인한 문제를 제외할 경우 만족할 만한 정확율과 재현율을 얻을 수 있었다.

1. 서론

일반적으로 정보검색에 있어서 가장 중요시 되는 것은 검색의 정확율과 재현율 측면에서의 성능이다. 특히 최근의 웹 환경 중심의 정보검색에 있어서는 정확율의 측면이 더욱 강조되고 있는데, 이를 위해서는 문서내의 중요한 정보들을 효과적으로 추출해낼 필요가 있다. 문서의 정보가 정확하게 표현됨으로 인해 검색의 정확도가 향상될 수 있기 때문이다.

하나의 문서를 효과적으로 표현하는데 있어 주로 사용되는 요소는 그 문서를 구성하고 있는 단어들이다. 그 중에서도 문서 내에서 중심이 되는 의미를 지니는 것은 명사인데, 명사는 문장내의 다른 요소들과는 달리 그 자체만으로도 충분한 의미를 지닐 수 있기 때문이다. 따라서, 일반적인 정보검색 시스템에서는 문서를 대표하는 정보를 추출하기 위해 그 문서내의 명사들을 추출하는 방식을 이용하고 있다.

명사는 의미적으로 볼 때 일반적인 의미를 갖는 보통명사와 고유한 대상을 가리키는 고유명사로 나뉜다. 문서 내에서 보통명사는 주로 그 문서의 내용을 유연하게 이끌어가는 역할을 하고, 고유명사는 문서의 주체가 되는 경향이 높다. 즉, 같은 명사라 할지라도 고유명사는 문서 내에서 좀 더 유용한 의미를 내포하고 있다고 볼 수 있는 것이다. 따라서, 문서 내에서 고유명사를 추출하는 것은 문서 전체의 의미를 정확하게 표현하는데 크게 기여할 수 있는 방법이 될 수 있다. 본 논문에서는 문서내의 고유명사 추출을 위한 방법과 관련한 공통 기본 모델을 제시하고, 이를 특정 고유명사 집합에 적용하여 본 결과를 보이고자 한다.

고유명사는 그것이 지칭하는 대상에 따라 인명, 회사명, 지명, 도서명, 음반명 등으로 크게 나누어 볼 수 있다. 이러한 구분은 특정 고유명사 집합의 특성을 유출해내기 쉽게 한다. 즉, 지칭하는 대상에 따라 고유명사 집합을 나누고, 각 집합에 대한 특징을 분석하여 이를 이용할 경우 고유명사 전체를 대상으로 하는 경우 보다 문서로부터 해당되는 고유명사를 추출해내는 것이 쉬워질 수 있다. 이러한 관점에서 각 고유명사

집합을 하나의 객체로 보고, 각 고유명사 집합이 갖는 특성을 객체의 속성으로 본다면, 고유명사 집합들에 공통적인 모델을 적용할 수 있을 것이라는 아이디어를 기본으로 하여 본 연구를 수행하였다.

본 연구에서는 고유명사 추출을 위해 제시한 공통 모델을 검증하기 위해, 고유명사 집합 중에서도 회사명과 관련된 고유명사를 추출할 수 있는 시스템을 고유명사 공통 모델을 기본으로 하여 구현하고 이에 대한 실험을 수행하였다.

2. 고유명사

2.1 고유명사의 특징

정보검색의 견지에서 볼 때, 문서 내의 고유명사는 색인어로서의 가치가 높다. 고유명사는 문서 내에서 해당 문서의 내용을 이끄는 역할을 하기 때문에, 문서의 내용을 효과적으로 대표할 수 있기 때문이다. 실제로 기존 연구에 따르면, 100개의 문서로부터 수작업으로 뽑은 의미있는 색인어 중 약 18%이상이 고유명사에 속하였다.[1] 따라서, 문서로부터 고유명사를 효과적이고 정확하게 추출해낼 필요가 있다.

고유명사는 인명, 회사명, 지명, 도서명, 음반명 등으로 그 대상이 세분화될 수 있다. 일반적인 견지에서 볼 때 이들 각각이 의미하는 바가 다르므로, 각 고유명사 또한 그러한 차이를 인식할 수 있게 하는 나름대로의 특성을 지니고 있다. 이러한 고유명사 집합 중 인명, 회사명, 지명과 관련된 특징을 살펴보면 다음과 같다.

인명의 경우는 일반적으로 성씨와 이름으로 나누어질 수 있는 요소를 가지며, 이에 부가적으로 호칭에 해당하는 것이 인명에 대한 정보로써 더해질 수 있다. 그리고 성씨로 쓰일 수 있는 집합은 한정되어 있으며, 이름은 일반적으로 2~3 문자로 이루어진다. 이름의 경우도 실제로 사용되고 있는 이름에 대한 집합을 이용하여 구분될 수 있지만, 이는 성씨에 대한 집합보다 더 유연한 집합이다. 따라서, 문서로부터 인명을 추출해내는 작업은 이들 기본적인 정보를 이용하여 이루어질 수 있다. 이에 대한 연구는 상당부분 이루어진 것으로 알려져 있다.[1]

회사명의 경우는 인명의 경우보다 다양한 형태를 갖는다. 회사명은 일반적으로 그 회사가 중점적으로 하는 사업을 반영하고 있어, 단일 명사로 대표되기 보다는 그 의미를 반영할 수 있는 여러 명사가 복합적으로 사용되는 경우가 많다. 예를 들어, " 현대 화재 해상 보험" 과 같은 경우 네 개의 명사가 결합되어 하나의 회사명을 구성하고, 각각은 " 현대" 의 경우를 제외하고는 모두 보통명사의 형태를 띄고 있다. 따라서, 인명의 경우에 비해 회사명을 구분해내기란 쉽지 않다. 회사명을 문서로부터 추출하기 위해서는 회사명에 포함되는 단어들의 특성과 그 주변에 나타나는 단어들을 적절히 이용해야 한다. 이에 대한 사항은 2.2절에서 자세히 다루기로 한다.

지명의 경우는 행정구역이나 지리적 특성을 나타내는 특수한 접미사가 동반되는 경우가 많다. 예를 들어, " 성남시" , " 충청남도" , " 지리산" , " 남해안" 과 같은 지명은 각각 " 시" , " 도" , " 산" , " 해안" 같은 접미사를 포함한다. 이러한 경우를 제외하고 접미사 없이도 사용되는 경우도 있다. 예를 들어 " 서울" 이나 " 충북" 과 같은 경우가 그러하다. 따라서, 문서로부터 그에 포함된 지명을 추출하기 위해서는 지명에 사용되는 접미사 집합 뿐만 아니라 지명자체에 대한 집합도 있어야 한다.

위에서 살펴본 바와 같이, 고유명사는 그 자체에 고유명사의 종류를 구분지을 수 있을 만한 정보를 포함하는 경우가 많다. 인명에서의 성씨나 회사명에서의 회사주력 분야(ex.화재 해상 보험), 또는 지명에서의 행정구역이나 지리적 특성을 나타내는 접사(ex. 시, 도, 산 등) 등이 이에 해당된다고 볼 수 있다. 이러한 특징을 이용하면 고유명사 추출을 위한 공통 모델을 제안할 수 있다. 이에 대해서는 다음 절에서 회사명 집합을 중심으로 고유명사 집합이 갖는 특징을 살펴보고 이를 지명 집합과 비교해 보면서 알아보기로 한다.

2.2 회사명과 지명의 특징 및 접근방법

문서 내에서 회사명은 매우 다양한 형태로 나타난다. 우선, 단일 명사로 표현되는 회사명이 있을 수 있다. 예를 들어, " 삼성" , " 현대" , " 기아" 와 같은 경우가 그러하다. 문서에서 이러한 형태의 회사명을 추출하기 위해서는 회사명사전을 이용하여야 한다. 여기에서 회사명 사전이란, 실세계에 존재하는 회사명을 갖고 있는 사전으로 이를 이용할 경우 문서로부터 정확한 회사명을 추출할 수 있다는 장점을 갖는다.

단일 명사로 표현되는 회사명 보다는 일반적으로 복합명사로 표현되는 회사명이 많이 사용되고 있다. 이러한 경우에 해당되는 회사명으로는 " 중앙 소프트웨어" , " 현대 그룹" , " 삼성 생명" , " 한국 가스 공사" 등이 있다. 복합 명사 형태로 이루어지는 회사명은 회사명접사를 포함하는 경우가 많다. 회사명 접사란, 회사명에 함께 쓰여 주로 해당 회사의 주력분야를 표현해주는 요소로써, 회사명의 뒷부분에 많이 사용된다. 위의 예에서 " 소프트웨어" , " 그룹" , " 생명" , " 공사" 등이 이에 속한다. 복합 명사로 이루어진 회사명의 이러한 특징을 이용하여 회사명접사사전을 독

립적으로 만들 수 있고, 이를 회사명 추출에 사용하면 좋은 결과를 얻을 수 있다.

예를 들어, 현재 회사명사전에 " A 건설" 이라는 회사명은 존재하는데, " B 건설" 이라는 회사명은 존재하지 않는다고 생각해 보자. 이 때 회사명사전만을 이용하여 문서로부터 회사명을 추출할 경우에는, 문서에 " B 건설" 이라는 회사명이 있을 지라도 이 회사명을 추출해낼 수 없다. 이러한 경우, 회사명접사사전을 이용하여 " 건설" 이라는 회사명접사를 인식하는 방법을 이용한다면, " B 건설" 이라는 회사명을 효과적으로 추출할 수 있다.

한 문장 내에서 회사명은 주로 문장 전체를 이끄는 역할을 한다. 그리고, 회사명 주위에는 의미적으로 회사명을 뒷받침해주는 단어들이 함께 나타나는 경우가 많다. 이러한 단어를 단서라고 하는데, 우선 다음 예제를 살펴보자.

현대전자가 자사주를 대량 매각해 자금을 마련하고 있다.
회사명 단서 단서 단서

위의 예에서 " 현대전자" 라는 회사명은 문장의 주체로써, " 자사주" , " 매각" , " 자금" 과 같은 단서를 이끈다. 즉, 단서란 문장 내에서 나타난 어떠한 대상의 특징적인 행동이나 상태를 나타내는 단어를 의미한다. 이러한 단서를 회사명 추출에 이용할 수 있다.

예를 들어, 회사명사전에 포함되어 있지는 않지만 회사명일 가능성이 높은 고유명사 후보가 문서로부터 추출되었을 경우, 그 주변의 단어들을 수집하고 단서집합에 있는 통계적인 데이터를 이용한다면, 해당 후보가 회사명일 확률을 추정해낼 수 있다. 따라서, 회사명을 문서로부터 추출함에 있어서 회사명사전, 회사명접사사전 뿐만 아니라 단서집합까지도 이용하는 것이 좀 더 정확한 결과를 얻는데 도움을 줄 수 있다.

즉, 문서로부터 회사명을 추출해내기 위해서는 회사명사전, 회사명접사사전, 단서집합의 세 가지 정보가 있어야 한다. 이러한 세 가지 정보를 적절히 사용해야 좋은 결과를 얻을 수 있는 것이다.

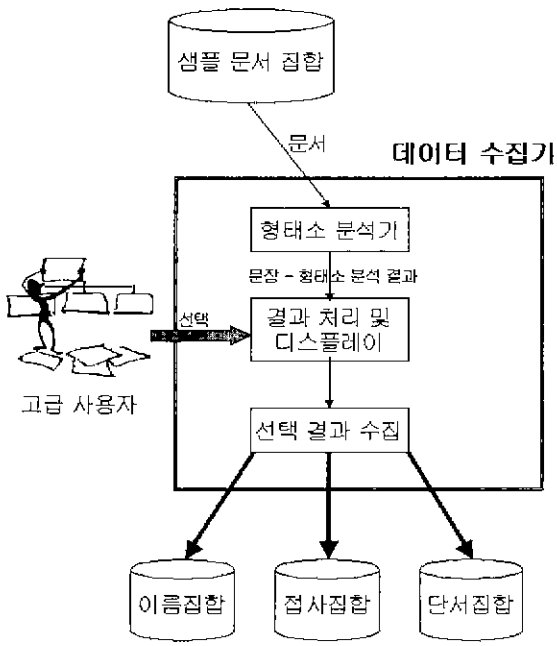
지명의 경우는 회사명과 달리 일반적으로 단일 명사의 형태를 갖는다. 그러나, 이를 회사명과 비교하여 보면, 많은 유사점을 갖는다는 것을 알 수 있다. 예를 들어, " 충남" , " 대전" 과 같이 단순히 이름만으로 지명을 나타내는 경우가 있는가 하면, " 경기도" , " 서울시" 와 같이 " 도" , " 시" 등의 접사를 포함하는 경우가 있다. 이는 각각 회사명 집합의 특성에서 살펴본, 단일명사로 표현되는 회사명과 복합명사 형태로 표현되면서 의미있는 접사를 포함하는 회사명의 경우에 해당된다고 볼 수 있다. 또한, 지명과 함께 나타나는 단어들을 단서로 보아 이를 이용할 수도 있을 것이다.

이와 같은 관점에서 보면, 지명의 경우도 회사명의 경우처럼 지명추출을 위해 지명사전, 지명접사사전, 단서집합의 세 가지 정보를 이용할 수 있다. 이러한 사실은 회사명추출에서 적용한 모델을 지명의 경우에도 적용할 수 있다는 것을 의미한다. 즉, 지명에 대한 간단한 전처리(preprocessing) 부분만 추가해 준다면 회사명추출에서 사용했던 고유명사 추출 모델을 거의 그대로 사용할 수 있다는 것이다. 다음 장에서는 이러한 아이디어를 기반으로 고유명사 추출을 위한 기본 모델을 제안하고, 이를 회사명 고유명사 추출의 경우에 적용한 예를 구체적으로 설명하기로 한다.

3. 고유명사 추출 모델

3.1 데이터 수집기

고유명사를 추출하기 위해서는, 앞서 회사명과 지명의 경우를 예를 들어 설명한 바와 같이 이름집합, 접사집합, 단서집합의 세 가지 데이터 집합이 만들어져야 한다. 이러한 데이터 집합을 구성하기 위해 우선 샘플 문서집합을 이용해 데이터를 수집할 수 있는 " 데이터 수집기" 를 위한 모델을 제안한다. [그림 1] 은 데이터 수집기의 흐름을 간단히 표현한 것이다.



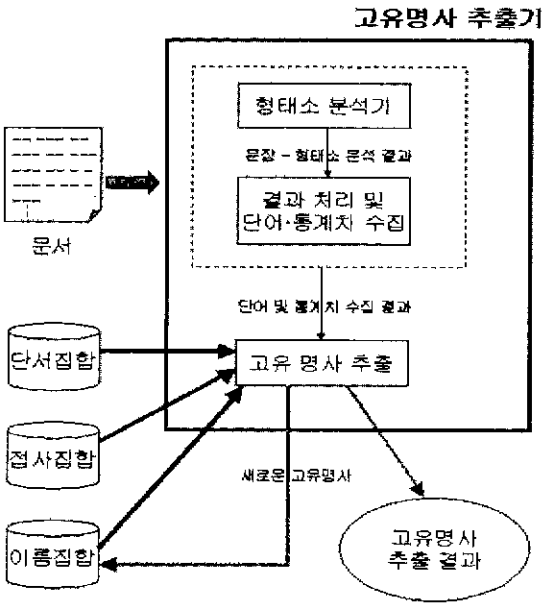
[그림 1] 데이터 수집기

데이터 수집기는 샘플 문서집합으로부터 선택된 문서를 입력으로 하여 사용자가 손쉽게 이름집합, 접사집합, 단서집합을 구성할 수 있도록 하는 역할을 한다. 이를 위해 데이터 수집기는 입력 문서에 대한 형태소 분석 결과를 내부적으로 처리하여, 고유명사일 가능성이 높은 부분을 자동으로 골라내 사용자에게 제시한다. 사용자는 데이터 수집기에 의해 제시된 결과를 보고, 고유명사인지 아닌지, 접사가 포함되어 있는지 아닌지를 결정하고, 접사가 포함되어 있을 경우 이에 해당되는 부분을 선택하게 된다. 또한 최종적으로는 문장 내에서 고유명사를 뒷받침해주는 역할을 하는 단서들을 선택하게 된다.

데이터 수집기를 이용할 경우, 형태소 분석 결과를 이용할 수 있으므로, 문장 내에 있는 보통명사들을 자동으로 고유명사 후보에서 제외시킬 수 있다. 또한 이미 이름집합에 포함된 고유명사들만이 있는 문장의 경우는 단서수집만을 수행하게 할 수 있고, 고유명사 후보가 없는 문장의 경우는 자동으로 데이터 수집 과정 전체를 건너뛰게 할 수 있다. 따라서, 데이터 수집기를 이용하면, 단순히 문서만을 보고 사람이 수작업으로 필요한 데이터들을 골라내는 것보다 훨씬 빠르게 데이터를 수집할 수 있다.

3.2 고유명사 추출기

고유명사 추출기는 데이터 수집기에 의해 수집된 데이터 집합을 기본 데이터 집합으로 하여 문서로부터 고유명사들을 추출해낸다. 다음의 [그림 2]는 고유명사 추출기의 기본적인 흐름을 나타내고 있다.



[그림 2] 고유명사 추출기

고유명사 추출기는 문서의 형태소 분석 결과로부터 해당 문서 내의 단어 및 단어 출현빈도와 동일 문장 내에 존재하는 단어의 리스트를 수집하고, 이 과정에서 수집된 결과와 데이터 집합을 이용해 해당 문서로부터 고유명사를 추출해낸다. 고유명사 추출기를 이용하여 고유명사를 추출해 내는 과정에서 다음의 세 가지 규칙이 기본적으로 적용된다.

1. 이름집합에 있는 고유명사와 동일한 고유명사가 문서처리 결과에 있을 경우, 이를 고유명사 추출 결과로 내어 놓는다. 이는 가장 단순하면서도 정확한 방법으로써, 고유명사추출에서 가장 우선적으로 적용되는 규칙이다.

2. 고유명사 후보가 접사집합에 있는 접사를 포함하고 있을 경우, 이와 같은 문장에 나타난 단어들이 단서집합에 얼마나 포함되어 있는지에 따라 고유명사를 추출한다. 이 방법은 이름집합에 있는 고유명사를 이용하는 것보다 다소 신뢰도가 떨어지지만 접사정보와 단서를 이용하여 이름집합에 포함되어 있지 않은 미등록 고유명사를 추출해낼 수 있다는 장점을 갖는다.

3. 위의 두 가지 규칙에 해당되지 않는 고유명사 후보의 경우에는 단서집합만을 이용한다. 즉, 고유명사 후보와 함께 나타나는 단어들이 어느 정도 단서집합에 포함되어 있는지 만을 고려하는 것이다. 이 경우 위의 두 규칙을 적용한 경우보다 신뢰도가 떨어지지만, 이름집합이나 접사집합을 이용하여서도 추출해낼 수 없는 미등록 고유명사를 추출해낼 수 있다.

위의 세 가지 기본 규칙이외에 고유명사 추출 과정에서, 잘못된 결과를 얻지 않게 하기 위해 부가적으로 사용되는 두 가지 규칙이 있다. 접사배제어와 단서집합을 이용하는 방법인데, 첫째는 일반적으로 접사와 함께 자주 사용되지만, 고유명사가 아닌 단어들을 접사배제어 리스트에 두어 이러한 단어들이 함께 나타나는 고유명사 후보는 해당 접사가 있다고 하더라도 고유명사 후보에서 제외시키는 방법이다. 둘째는 단서집합에 포함된 단어들로만 이루어진 고유명사 후보의 경우, 보통명사로 쓰일 때가 많으므로 이러한 것들을 고유명사 후보에서 제외시키는 것이다.

고유명사 추출 과정에서 부가적으로 할 수 있는 일로써, 이름집합 확장이 있다. 즉, 앞서 설명한 고유명사 추출을 위한 세 가지 기본 규칙 중 두 번째와 세 번째 규칙을 적용하여 새로이 얻은 고유명사를 이름집합에 포함시킴으로써, 이름집합을 확장할 수 있는 것이다.

이상으로 고유명사 추출을 위해 필요한 데이터 수집기 및 고유명사 추출기에 대해 살펴보았다. 다음 장에서는 본 장에서 설명한 모델을 실제로 회사명 추출에 적용한 예를 보이기로 한다.

4. 회사명 추출에의 적용

본 연구에서는 고유명사 추출 모델을 기초로 하여 회사명을 추출하기 위한 시스템을 설계 및 구현하였다. 현재 구현은 한성대학교 한글공학 연구소의 한국어 분석 모듈 HAM version 4.70c의 형태소 분석과 복합명사 분해 기능을 이용하여 Linux version 2.2 상에서 이루어졌다. [4]

4.1 데이터 수집기

샘플 문서를 데이터 수집기에 입력으로 주었다고 하자. 입력된 문서는 형태소 분석기에 의해 형태소 분석이 이루어진다. 예를 들어, 다음과 같은 문장에 대해 형태소 분석을 수행하면,

> 현대전자가 연이어 자사주를 대량으로 매각해 자금을 마련하고 있다.

각 단어의 품사와 조사 정보가 포함되어 있는 다음과 같은 형태소 분석 결과를 얻을 수 있다.

```
(N "현대전자")< :50> + (j "가")<1>
(V "연잇")<Ts:20> + (e "어")<2>
(N "자사주")< :60> + (j "를")<1>
(N "대량")<N:20> + (j "으로")
(N "매각")<N:29> + (t "하") + (e "어")<9>
(N "자금")<N:20> + (j "을")
(N "마련")<N:25> + (t "하") + (e "고")
(V "있")<KIgVJ:20> + (e "다")
(P ".")< :0>
```

데이터 수집기는 이러한 형태소 분석 결과 내에, 복합명사나 형태소 분석기 자체 내의 사전에 보통명사로써 정의되어 있지 않은 명사가 있을 경우, 이를 고유명사 후보로 선택한다. 선택된 고유명사 후보가 이름집합에 존재하지 않는 것일 경우, 데이터 수집기는 이를 사용자에게 제시한다. 그러면, 사용자는 제시된 고유명사 후보를 보고, 그것이 회사명인지 아닌지를 선택하게 된다. 위의 예에서 고유명사 후보로는 " 현대전자" 와 " 자사주" 가 있다. 이들 모두가 이름집합에 없다고 가정할 경우, 사용자는 이 두 고유명사 후보에 대해 차례로 고유명사인지 아닌지를 결정하게 된다. 이 경우 회사명인 고유명사 후보는 " 현대전자" 이다.

일단, 고유명사가 선택되면, 접사수집 단계를 거치게 된다. 이 단계에서는 회사명에 포함되어 공통적으로 자주 쓰이는 단어를 사용자 선택에 의해 수집하게 된다. 위의 예에서 회사명인 " 현대전자" 를 분해하면 " 현대" 와 " 전자" 로 나누어진다. 분해된 결과 단어들이 접사집합에 포함되어 있지 않다면, 사용자는 이들 중에서 접사에 해당되는 단어를 선택하게 된다. 여기에서는 " 전자" 가 접사에 해당된다.

문장 내의 모든 고유명사 후보에 대해 고유명사 선택 및 접사선택의 과정을 거친 후에, 데이터 수집기는 고유명사로 선택되지 않은 명사들을 사용자에게 제시하게 된다. 사용자는 이러한 명사들 중에서 단서가 되는 단어들을 선택하게 되고, 선택된 단어는 단서집합에 저장된다.

데이터 수집 단계에서 접사배제어도 수집된다. 이는 접사를 이용해 고유명사를 추출하는 과정에서 일어나는 문제를 해결하기 위해 사용된다. 예를 들어, 접사 " 은행" 이 접사집합에 포함되어 있다고 가정하자. 그러면, " 하나은행" , " 서울은행" 등과 같은 회사명이 이름집합에 없을 경우, 이들 회사명을 접사 " 은행" 을 이용하여 추출해낼 수 있다. 그러나, 단순히 이러한 방법으로 회사명을 추출할 경우, " 부실은행" , " 국내은행" , " 시중은행" 등과 같은 복합명사의 경우도 회사명으로 잘못 추출될 수 있다. 따라서, 이러한 경우 " 은행" 이라는 접사의 배제어 목록에 " 부실" , " 국내" , " 시중" 등의 단어를 두어, 고유명사 후보 추출 과정에서 이를 참고하면, 단순 복합명사를 추출해내는 것을 막을 수 있는 것이다.

4.2 고유명사 추출기

고유명사 추출기는 입력문서에 대한 형태소 분석결과와 데이터 집합을 이용하여 고유명사를 추출한다. 예를 들어, 다음과 같은 문장에서 고유명사를 추출한다고 할 경우,

> 소고 백화점은 12일 공적 자금 지원에 대한 여론의 질타를 우려해 법정관리를 신청했다.

고유명사 추출기는 다음과 같은 형태소 분석 결과를 얻게 된다.

```
(N "소고")<N:20>
(N "백화점")<N:20> + (j "은")
(N "12일")< :60>
(Z "공적")<E:10>
(N "자금")<N:20>
(N "지원")<N:20> + (j "에")
(V "대하")<IT:24> + (e "ㄴ")<13>
(N "여론")<N:20> + (j "의")
```

(N "질타")<N:20> + (j "를")<1>
(N "우려")<N:29> + (t "하") + (e "어")<9>
(N "법정관리")< :50> + (j "를")<1>
(N "신청")<N:15>+(t"하")+(f"었")<8>+(e"다")
(P ".")< :0>

위의 형태소 분석결과에서는 " 소고 백화점" , " 자금 지원" , " 법정관리" 가 고유명사 후보로 선택된다. 고유명사 추출기는 이런 후보 단어들을 대상으로 하여, 이들을 고유명사로서 추출해 낼 것인지 아닌지를 결정하게 된다. 앞선 예제들에서 고유명사 후보로 선택된 단어들에 대해 고유명사 선택 규칙들을 적용하는 경우, 어떠한 결과가 나오는지 살펴보면 다음과 같다.

첫 번째 예문에서, " 현대전자" 는 이름집합에 있는 단어이므로 첫번째 규칙에 의해 고유명사라는 것이 확인된다. " 자사주" 의 경우는 이름집합에 없으므로 접사를 찾아보게 된다. 그러나 " 자사주" 에는 접사가 없고, 이 단어는 복합명사도 아니므로 다음 단계로 넘어간다. 이 문장에는 명사가 5개(현대전자, 대량, 매각, 자금, 마련) 있는데, 그 중 단서 집합에 있는 것은 2개(매각, 자금)이므로 문장에서 단서가 나오는 확률이 50%가 안 된다. 따라서, " 자사주" 는 고유명사에서 제외된다.

두 번째 예문의 경우, " 소고 백화점" 은 이름집합에 없고, 접사 " 백화점" 이 있는 경우이다. " 소고" 는 접사배제어 집합에 포함되어 있지 않으므로 고유명사가 될 수 있다. 이 문장에는 명사가 6개(자금, 지원, 여론, 질타, 법정관리, 신청) 있고, 그 중 단서 집합에 있는 것은 3개(자금, 지원, 법정관리)이므로 문장전체에서 단서가 나올 확률이 40% 이상이 되어 " 소고 백화점" 은 고유명사로 판정된다. " 자금 지원" 은 이름집합에 없고 접사도 없지만 주위 단어들 중 단서 집합에 있는 단어들의 비율이 50% 이상이 된다. 그러나, 이는 단서 집합에 있는 단어들로만 이루어져 있으므로 고유명사가 될 수 없다. " 법정관리" 의 경우도 " 자금지원" 과 같은 이유로 인해 고유명사가 되지 못한다. 따라서, 첫번째 예문에서는 " 현대전자" 가, 두 번째 예문에서는 " 소고 백화점" 이 고유명사로 추출된다.

이러한 방식으로 고유명사 추출기는 문서로부터 고유명사들을 추출해 내고, 그 결과 미등록 고유명사가 추출되었을 경우엔 이를 이름집합에 추가하여 이름집합을 확장한다.

5. 실험 및 결과분석

5.1 데이터 수집 및 결과

회사명에 관련한 데이터를 수집하기 위하여, 웹 사이트 ' 야후 코리아' 의 경제부분에 있는 ' 로이터' 50개 기사, ' 매일경제' 150개 기사, ' 연합뉴스' 200개 기사, ' 한국경제' 200개 기사의 총 600개 기사를 샘플문서로 이용하였다. 샘플문서집합의 총 크기는 대략 1.2MB이고, 문서 당 평균 크기는 대략 2 KB이다. 이러한 샘플문서집합을 입력으로 하여 데이터를 수집한 결과, 회사명 949개, 회사명접사 130개, 단서 1405개가 각각 수집되었다.

5.2 고유명사 추출 시험 및 결과

회사명 추출기의 성능을 테스트해 보기 위해 ' 로이터' 100개 기사를 대상으로 하여 회사명을 추출하였다. 그 결과는 다음과 같다.

	문서에 있는 회사명	추출한 회사명	제대로 추출한 회사명	추출하지 못한 회사명
			잘못 추출한 회사명	
총 개수	302	214	191	111
			23	

[표 1] 회사명 추출 결과

위의 표를 보면, 문서 당 평균 3개 정도의 회사명이 존재한다는 것을 알 수 있다. 이러한 문서들을 대상으로 하여 회사명을 추출한 결과, 약 89.25%의 정확율과 63.25%의 재현율을 보이는 결과를 얻을 수 있었다. 각 데이터 집합이 결과에 어느 정도의 영향을 미치는지 확인하기 위해, 각 데이터 집합을 제외한 후 회사명을 추출해보았다. 다음의 [표 2]가 그 결과이다.

제외집합	추출 총개수	바른 결과수	잘못된 결과수	미추출 개수
가) 없음	214	191	23	111
나) 이름집합	126	104	22	198
다) 접사집합	166	159	7	143
라) 단서집합	208	191	17	111
마) 배제어정보	223	191	32	111

[표 2] 데이터 집합 사용여부에 따른 결과

위 표의 결과를 보면, 가)에 비해 이름집합을 이용하지 않는 나)의 경우 97개의 회사명이 더 추출되지 않았다. 또한 접사집합을 이용하지 않고 회사명을 추출한 다)의 경우는 159개의 회사명을 추출할 수 있었다. 여기에서 다)의 결과는, 가)와 라)의 결과에서 올바르게 추출된 회사명의 수가 같은 것으로 보아 이름집합을 이용하여 추출된 것임을 알 수 있다. 따라서,

회사명 추출에 있어 이름집합이 큰 비중을 차지하고 있음을 알 수 있다.

접사집합의 경우도, 나)에서 이름집합을 이용하지 않았음에도 불구하고 104개의 회사명이 추출되었고, 다)에서 미추출된 회사명의 수가 가)에 비해 32개가 많은 것으로 보아, 회사명 추출에 큰 몫을 차지하고 있다는 것을 알 수 있다.

그러나, 단서집합의 경우는 회사명 추출에 있어 유용한 집합이 아님을 알 수 있다. 이는 단서집합을 이용한 가)의 경우에 비해, 이용하지 않은 라)의 경우 오히려 잘못 추출된 회사명의 수가 줄어든 것을 보면 알 수 있다.

접사배제어 정보는, 가)의 결과에 비해 접사배제어 정보를 이용하지 않은 마)의 결과에서 잘못 추출된 회사명의 수가 9개 늘어난 것으로 보아, 접사정보만을 이용하여 생기는 회사명 추출 오류를 줄이는데 기여하는 바가 크다고 볼 수 있다.

회사명 추출에서 각 데이터 집합이 기여하는 바는 앞선 결과들에서 알 수 있었지만, 전체적인 회사명 추출의 재현율이 낮은 원인은 위의 결과들만으로는 알 수 없었다. 따라서, 테스트 문서 집합에 포함되어 있는 실제 회사명과 고유명사 추출기에 의해 추출된 회사명들을 분석하여 본 결과, 회사명 추출 실패의 원인이 대부분 외래어에 의한 것임을 알 수 있었다. 따라서, 회사명 추출 결과들에서 외래어만으로 이루어진 회사명들을 제외하고, 재분석해 본 결과 다음과 같은 수치를 얻을 수 있었다.

	문서에 있는 회사명	추출한 회사명	제대로 추출한 회사명	추출하지 못한 회사명
			잘못 추출한 회사명	
총 개수	148	154	138	10
			16	

[표 3] 외래어를 제외한 회사명 추출 결과

즉, 외래어를 제외하였을 경우 고유명사 추출결과는 약 89.6%의 정확율과 93.24% 의 재현율을 보이고 있다. 이는 [표 1]과는 상당한 차이를 보이는 것으로, 이를 통해, 외래어로 표기된 회사명이 문서에서 많이 사용되고 있고, 이들을 문서에서 인식하는데 있어 많은 문제점을 안고 있다는 것을 알 수 있다.

외래어로 표기된 회사명이 문제가 되는 원인은 다음과 같다.

1. 형태소 분석의 실패
   ex) " 아이와" → (아이 + 와)

2. 복합명사분해의 실패로 인한 접사 인식의 실패
   ex) " 넥스텔커뮤니케이션즈"
   (복합명사분해기를 이용하여 분해된 결과를 접사정보와 비교하여 사용하는데, 이 경우 " 넥스텔" 과 " 커뮤니케이션즈" 로 분해되지 않는다.)

3. 회사명 표기의 일관성 문제
   ex) " 코메르츠뱅크" 또는 " 코메르츠방크"

외래어로 표기된 회사명의 경우 이외에, 이름집합이나 접사집합의 데이터 부족과 문서 자체의 오류도 회사명이 추출되지 않거나 잘못 추출되는 원인이 된다. 데이터집합의 부족은 충분한 샘플 데이터를 이용해 데이터를 수집하지 않았기 때문에 생긴 문제이고, 문서 자체의 오류는 띄어쓰기 오류에 의한 경우가 상당부분을 차지하고 있다.

6. 결론 및 향후연구

본 연구에서는 문서 내 고유명사의 효과적인 추출을 위한 고유명사 추출 모델을 제안하였다. 고유명사 추출 모델은 고유명사 추출에 이용할 이름집합, 접사집합, 단서집합을 구성하기 위해 사용되는 데이터 수집기 부분과 이를 사용해 수집된 데이터 집합을 이용하여 실제로 문서로부터 고유명사를 추출해내는 고유명사 추출기 부분으로 크게 나뉘어 진다. 이 모델은 일반적인 고유명사 집합에 적용시킬 수 있는 모델로써, 이를 검증하기 위해 회사명 집합을 대상으로 구현 및 실험을 수행하였다. 실험 결과 외래어로 이루어진 회사명으로 인한 문제를 제외한 경우, 89.6%의 정확율과 93.24%의 재현율을 얻을 수 있었다.

향후 연구에서는 본 논문에서 제시한 고유명사 추출 모델을 다른 고유명사 집합에 적용해 보아 좀 더 일반화된 결과를 얻어야 하고, 단서집합을 고유명사 추출에 효과적으로 이용할 수 있는 방법을 모색해야 할 것이다. 또한 외래어로 표기되는 고유명사들에 대한 연구를 통해 이들에 대한 추출확률을 높여야 할 것이다.

참고 문헌

[1] 정래정, 김준태. " 통계 정보와 어의 정보를 이용한 미등록 고유 명사의 자동 색인" . 1996년도 한국정보과학회 가을 학술발표논문집 Vol.23. No. 2.

[2] Inderjeet Mani, T. Richard MacMillan. Corpus Processing for Lexical Acquisition. MIT Press,

1996.

[3] Paul Thompson, Christopher C. Dozier. Name Recognition and Retrieval Performance. Natural Language Information Retrieval, Kluwer Academic Publishers, 1999

[4] 강승식. "한국어의 형태론적 특성과 형태소 분석 기법", 정보과학회지, 12권 8호, pp.47-59, 1994.