고유 명사 출현 패턴을 이용한 색인의 성능 향상에 관한 연구*

정래정, 김준태
동국대학교 컴퓨터공학과

Improving Indexing Performance by using Occurrence Pattern Information of Proper Nouns

Raejung Jung and Juntae Kim
Department of Computer Engineering, Dongguk University

요 약

본 논문에서는 고유 명사 출현 패턴 정보와 부가 정보를 이용한 미등록 고유 명사의 색인 방법을 제안한다. 정보 검색 시스템에서 고유 명사의 처리는 정확하고 의미 있는 색인을 위해 매우 중요하다. 본 논문은 형태소 분석 결과에 고유 명사 출현 패턴과 패턴 부가 정보를 사용하여 인명, 기관명, 회사명 등의 고유 명사 추출의 정확도를 높이는 방법을 제시한다. 총 827개의 인명과 기관 및 회사명을 포함하고 있는 조선일보 경제면 기사 100개 7416 어절에 대하여 본 시스템으로 실험한 결과, 인명의 경우 89%의 정확률을 보였다. 본 논문에서 제시한 출현 패턴과 고유 명사의 부가 정보를 적용했을 때 단순한 형태소 분석 결과에 비하여 고유 명사 추출 오류가 크게 개선되었다.

1. 서론

정보검색은 색인어를 통해 이루어진다. 색인이란 어떤 문헌에 대해 그 문헌의 전체적 내용을 나타내거나, 그 문헌을 다른 문서들로부터 구별할 수 있도록 그 문서의 선택 단서가 되는 단어 또는 단어구 등을 추출하는 것을 말한다[3,4]. 신속하고 정확한 검색은 색인의 추출 작업에 의해 좌우된다. 정보검색 시스템의 성공 여부는 사전정보와 정확한 색인에 의존하며 평가되어진다[3]. 한국어 색인 시에는 형태소 분석이 필요한데 그 이유는 형태소 분석을 이용하여 문서에 나타나는 명사들을 인식하고 이를 색인어로 추출하기 위한 것과, 색인어를 선별하는데 있어 기준을 가지기 위해서이다[11].

한국어 처리에는 미등록어와 복합 명사 처리 등이 어려운 문제로 남아있다. 미등록어와 복합 명사 처리는 색인의 정확성을 높이는데 기여한다[2]. 복합어에 대해서는 색인어로서의 복합어를 구성하기 위한 복합어 구성조건과 분해규칙을 만들고 복합어의 유용성을 측정하는 척도로 상호정보 개념을 제시해 구성 단어간의 연관도를 측정하는 연구가 있었고[8], 미등록어에 대해서는 언어정보인 한 어절 내의 미등록어의 결합 관계 정보, 미등록어 뒤에 나오는 형태소의 의미 정보, 좌우어절과의 연관 관계, 전체 문헌에서의 미등록어가 나타난 유형을 이용하여 미등록어를 추정하는 연구가 있었다[6]. 미등록된 단어 중에는 명사 미등록어가 대부분을 차지하므로 미등록어 추정은 입력된 어절로부터 미등록 명사를 추정하는 것이다[8].

미등록어에는 고유 명사, 신조어, 전문 용어, 약어 등 여러 가지가 있는데, 전문 용어와 약어는 분야별로 사전에 추가하여 처리할 수 있으나 고유 명사는 새로이 만들어지는 것이 많기 때문에 사전을 만들어 처리할 수 없다. 특히 신문 기사나 온라인 뉴스 등의 텍스트에는 색인어로서의 가치가 높은 고유 명사들이 다수 포함되어 있어 이들에 대한 정확한 추출이 전체적인 정보 검색의 정확도를 높이는데 매우 중요하다. 예를 들어 조선일보 경제기사 100개로부터 수작업으로 뽑은 의미 있는 색인어 중 약 18% 이상이 고유 명사였는데 고유 명사가 색인으로 가지는 중요성을 고려했을 때 매우 높은 비중을 차지하고 있다.

본 논문에서는 한국어 색인어로서 중요한 비중을 차지하는 고유 명사를 추출하는데 있어서의 문제점을 살펴보고, 출현 패턴과 고유 명사의 부가 정보를 복합적으로 이용한 고유

* 본 연구는 1996년도 한국과학재단 핵심연구과제 연구비 지원을 받고 있음.

명사 색인 방식을 제안하였다. 조선일보 경제면 기사를 대상으로 색인 실험을 수행하여 한국과학기술원에서 개발한 KTS(Korean Tagging System)시스템[2]의 고유 명사 추정오류를 크게 개선한 결과를 설명하였다.

2. 일반적인 고유 명사처리

한국어에서의 색인시 복합 명사와 고유 명사, 외래어 등 사전에 등록되지 않은 어휘가 색인어로 선택되는 경우가 많으므로 미등록어를 정확하게 추정하는 기능이 필요하다[10].

지금까지 연구되어온 미등록 고유 명사 처리 방법은 조사 사전을 이용해서 형태소 분석에 실패한 어절로부터 최장조사를 떼어내는 최장 조사 일치법이 있는데 이 방법으로 미등록 고유 명사를 추출하는데 있어서 다음과 같은 문제점들을 가지고 있다.

첫째, 고유 명사를 분리하지 못하는 경우가 많다. 즉, 잘못된 추출을 하는 경우인데, [예2.1]은 미등록어에 대한 색인이 제대로 되지 않고 있음을 보여주고 있다.

예 2.1) "권혁수기업문화부장도"

권혁수기업문화부/nc+장/xn+도/jx (X)

정확한 색인의 경우는 다음과 같다.

권혁수(이름미등록어)+기업문화부(미등록어)+장(호칭)+도

둘째, 형태소 분석 성공으로 추출을 못하는 경우도 있었다. 즉, 하나의 고유 명사를 다른 형태소들의 조합으로 분석하는 경우를 말하는데 대량의 말뭉치를 이용해 단어의 품사 태그(tag)를 결정하는 KTS 시스템에서는 다음 [예2.2]에서와 같이 "이장한"이라는 미등록 이름 고유 명사를 이(수사), 장(단위성 의존명사), 하(동사), ㄴ(관형사형 전성 어미)로 이름 고유 명사에 대한 처리에 있어서 잘못된 추출을 하는 경우가 있었다. 예는 [예2.2]와 같다.

예2.2) "이장한 회장은"

이/nn+장/nbu+하/pv+ㄴ/exm 회장/nc+은/jc (X)

정확한 색인의 경우는 다음과 같다.

이장한(이름 고유 명사) 회장(호칭) + 은(조사)

[보기1]은 KTS 시스템의 53개의 품사 태그 중에 몇 개를 보기로 들었다. KTS 시스템에서는 통사론적(기능), 형태론적(형태), 의미론적(의미) 기준에 따라 크게 품사 태그를 체언(명사, 대명사, 수사), 용언(동사, 형용사), 숫식언(관형사, 부사), 관계언(조사), 독립언(감탄사) 외에도 기호, 외국어, 접사로 나뉘었다[6].

[보기1] KTS시스템의 품사 태그(Tag) 중 일부

nq : 고유 명사, nc : 보통명사, nn:수사, jc:격 조사, nbu:단위성의존명사, jx:보조사, exm:관형사형 전성 어미, md:지시관형사, mn:수관형사, m:관형사, xn:명사접미사, ncs:상태성 보통명사, xpa:형용사 파생 접미사, pv:동사

* KTS 시스템은 다음의 53개의 태그(Tag)를 가지고 있다

기호(s) : 9개, 외국어(f) : 1개, 체언(n) : 11개, 용언(p) : 4개, 관용어(m) : 3개, 부사어(a) : 5개, 독립어(i) : 1개, 조사(j) : 7개, 어미(E) : 8개, 접사(x) : 4개 => 총 53개 태그(tag)

KTS 시스템은 형태소 배열 규칙을 위한 품사 접속표와 약 8,000개의 표제어를 갖는 사전을 약 55,000어절의 태깅된 말뭉치에서 추출해서 만들어 갖고 있다[6,9]. 이 시스템의 경우 조선일보 경제면 기사 100개를 대상으로 실험한 결과 위와 같은 이유로 고유 명사 추출에 실패한 경우가 인명 고유명사의 경우 45%, 회사명의 경우 55%, 지명의 경우 12% 등으로 매우 높았다. 특히 인명, 회사명 등이 색인어로서 가지는 가치를 감안해 본다면 위의 수치는 매우 높은 수치라고 할 수 있다.

신문 기사, 온라인 뉴스 등 고유 명사가 많이 포함된 경우 고유 명사의 빈도가 높으므로 이들 고유 명사의 추출에 정확도를 높이는 방법이 필요하다.

3. 출현 패턴을 이용한 고유 명사 추정

3.1 일반적인 고유 명사의 출현 패턴

고유 명사의 출현 패턴은 매우 다양하나 일반적으로 자주 출현하는 패턴은 다음 [표1] 및 [표2]과 같다. [표1]은 한 어절에서 고유 명사가 나타날 수 있는 경우를 11개의 패턴으로 나타냈고 이들 11개의 패턴의 조합들을 [표2]에 나타냈다. 여기에서 호칭을 나타내는 단어를 포함하는 어절은 다시 같은 어절이나 앞어절 또는 다음 어절을 고려하는 알고리즘을 사용하였다.

예로 "정주영 현대그룹명예회장은", "종근당 김상조 사장은", "이사장 정낙영"과 같은 자주 나타나는 패턴의 경우 한 어절의 패턴정보와 복합적으로 나타나는 패턴정보들을 형태소 분석 후에 미등록 고유 명사를 추정하고 추출하는데 이러한 패턴들을 사용한다.

2) 본 논문의 실험에서 사용된 KTS 시스템은 실험용 버전으로 현재의 과학기술원 형태소 분석 시스템을 대표하는 것은 아님.

고유 명사 출현 패턴	예
① 이름	정주영
② 이름 + 조사	이건희는
③ 이름 + 호칭	이석재씨
④ 이름 + 호칭 + 조사	김영배이사는
⑤ 이름+ 고유명사 + 호칭 + 조사	정인영한라그룹회장은
⑥ 고유명사 (회사/기관)	포항제철
⑦ 고유명사 + 조사	현대그룹은
⑧ 고유명사 + 호칭	한라시멘트부사장
⑨ 고유명사 + 호칭 + 조사	삼익악기회장은
⑩ 호칭	대표
⑪ 호칭 + 조사	사장은

[표1] 한 어절 내에서의 고유 명사의 출현 패턴

패턴 조합	예
a. ① + ⑥ / ⑦	고판남 세풍그룹(은)
b. ① + ⑩ / ⑪	정명식 회장(은)
c. ⑥ + ⑩ / ⑪	제일은행 사장(은)
d. ⑥ + ③ / ④	대웅제약 이승철사장(이)
e. ⑩ + ⑪	대표이사 부회장에
f. ① + ⑧ / ⑨	정세영 현대그룹회장(은)

[표2] 한 어절 이상에서의 고유 명사 패턴 조합

3.2 고유 명사 부가 정보를 이용한 고유 명사 후보 선정 및 판별

한국어 이름의 경우 특히 형태소의 오분석에 따라 색인으로 채택되지 못하는 경우가 많다. 즉, 형태소 분석에 성공하였더라도 한 어절 내에서나 혹은 앞어절이나 다음 어절에 호칭이 있는 경우 [예2.2]에서 본 것처럼 사람 이름일 가능성이 높다.

우선 인명이 어떤 어절로 이루어져 있는가에 대한 패턴 부가 정보인 통계정보를 이용하여 인명일 가능성이 있는 어절을 후보로 선택한다. 인명인지를 판단하기 위해 사용하는 정보는 다음 3가지로 구분한다.

첫째, 음절의 수를 고려한다. (2 - 4음절)

둘째, 첫 두음절의 종류를 판별한다. (성씨 여부)

셋째, 나머지 음절의 이름에서의 출현 빈도를 고려한다.

이러한 세 가지 조건에 입력 어절이 맞으면 형태소 분석에 성공한 경우라도 이름 후보로 두고 다음 어절에 대한 처리가 끝난 후 인명임을 암시하는 호칭이 있을 때는 이를 이름 고유 명사로 처리한다. 이를 위하여 [예3.1]과 같이 인명을 구성하는 성씨 약 105개와 다음에 올 수 있는 가능한 글자 약405개를 성씨사전과 이름사전으로 구성했다.

[예 3.1] < 성씨사전 >

강, 견, 경, 계, 고, 공, 곽, 구, 국, 궁, 권, 기, 길, 김, 나, 남, 노, 당, 도, 독고, 동, 두, 류, 마, 맹, 명, 모, 목, 문, 민, 박, 방, 배, 백, 변, 복, 봉, 부, ...

< 이름사전 >

가, 각, 간, 갈, 감, 갑, 강, 개, 객, 거, 걱, 건, 걸, 검, 겁, 겨, 격, 견, 겸, 겹, 경, 계, 고, 곡, 곤, 곰, 공, 곽, 관, 광, 교, 구, 국, 군, 궁, 귄, 귈, 규, 균, ...

위에서 선택된 후보들중 실제 인명인가를 경우를 판단하기 위하여 인명 뒤에 올 수 있는 어휘들을 수집해서 [예3.2]와 같이 사전으로 구성했다. 조선일보 경제면 기사로부터 약 42개의 어휘들을 추출하여 호칭사전을 구성했다.

[예 3.2] < 호칭사전 >

강사, 계장, 과장, 교수, 국장, 군, 님, 담당, 대리, 대통령, 대표, 대표이사, 박사, 반장, 법무사, 변호사, 부교수, 부사 장, 부실장, 부장, 부회장, 사원, ...

한 어절내에 호칭에 해당하는 어휘가 발견되면 같은 어절 혹은 바로 전 어절에 인명의 후보가 있을 경우 이를 색인어로 채택한다.

3.3 고유 명사 색인 알고리즘과 판별 우선 순위

형태소 분석기를 통해 분석이 끝낸 어절들에 대해서 분석이 실패한 경우를 가지고 미등록 고유 명사를 판별하는 간략한 색인 알고리즘은 한 어절이 형태소 분석이 되었더라도 다음 어절에 따라 미등록 인명 고유명사로 처리되어야 하는 경우가 있으므로 오류의 소지가 많다. 본 알고리즘은 두 어절을 보면서 한 어절의 처리결과를 다음 어절이 처리된 후에 색인으로 택하게 된다.

또한 한 어절이 여러 형태로 인식될 수 있는 미등록어 중의성 해결을 위해 다음과 같은 우선 순위를 결정하는 규칙을 가지고 그 중 우선 순위가 높은 결과를 선택한다.

첫째, 한 어절에서의 추정 결과보다 앞, 뒤 어절에 관계가 우선 한다.

예) 이상한 사장은

-> 이상/ncs+하/xpa+ㄴ/exm 사장/nc+은/jx (X)

이상한(이름 고유 명사) 사장(호칭)+은(조사) (O)

둘째, 일반적인 형태소 분석 결과보다 고유 명사 추정결과가 우선한다.

예) 현대자동차는

-> 현/nc+대/nc+자동차/nc+는/jx (X)

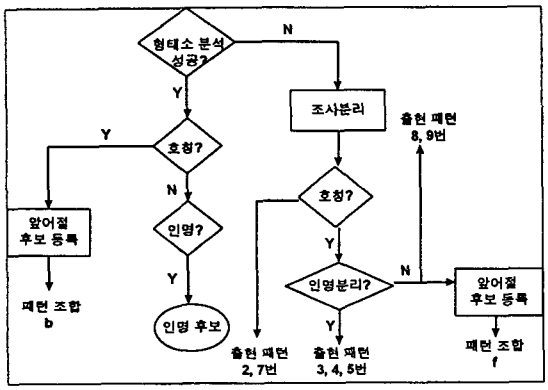
현대자동차(고유 명사)+는(조사) (O)

셋째, 호칭사전에 있는 최장일치 단어가 우선한다.

예) 현대부회장은

-> 현/nc+대부/nc+회장/nc+은/jx (X)

현대(고유 명사)+부회장(호칭)+은(조사) (O)



< 그림1 > 고유 명사 판별 색인 순서도

고유 명사를 판별하는 순서도는 <그림1>과 같다. 미등록어를 추정하는데 있어 한 어절에서는 최장일치 조사를 분리하고 고유 명사를 판별하는 색인기 처리를 거친 후 우선 순위 기준에 따라 미등록어를 고유 명사로 추정해 낸다.

4. 실험 및 결과

4.1 실험

본 시스템은 IBM PC상에서 C언어로 구현하였다. 사전은 명사사전과 조사사전, 호칭사전, 성씨사전, 이름사전 등이 있다. 본 시스템에서는 형태소분석기를 이용해 대상 문헌을 형태소 분석한 후, 그 결과로부터 추출된 명사와 미등록어를 대상으로 패턴 정보와 패턴 부가 정보에 기반한 고유 명사 추출을 위한 판별 알고리즘을 가지고 분석을 해나가는 시스템이다.

실험은 조선일보 경제면 기사 100개를 대상으로 하였으며 우선 수작업을 통해서 인명과 회사 및 기관명을 뽑고, 본 논문에서 구현한 시스템과 KTS시스템에서 뽑은 미등록 고유 명사들을 이들과 비교하여 두 시스템의 고유 명사 색인 재현율과 정확률을 비교 분석을 해 보았다.

4.2 색인의 재현율과 정확률

색인어의 성능을 평가하는 방법에는 다음과 같은 재현율(recall rate)과 정확률(precision rate)을 사용한다[1,4] 재현율은 텍스트에서 찾아내어야 할 총 정보의 양 대 시스템이 찾아낸 정확한 정보의 양의 비율로서, 시스템의 이해력을 나타낸다. 정확률은 시스템이 찾아낸 총 정보의 양 중 정확한 정보의 양의 비율로서 시스템의 정확도를 나타낸다. 본 실험에서는 고유 명사에 대한 재현율과 정확률을 다음과 같이 정의하였다.

◆ 재현율 (Recall rate)

= 찾아낸 올바른 고유 명사 / 총 고유 명사

◆ 정확률 (Precision rate)

= 찾아낸 올바른 고유 명사 / 찾아낸 고유 명사

본 실험에서는 조선일보 경제면 기사 100개로부터 본 논문에서 제안한 시스템과 한국 과학 기술원의 KTS 시스템의 고유 명사에 대한 색인 성능을 위의 재현율과 정확률 공식을 이용하여 비교하여 보았다. 총 어절의 수는 7416개이며 이중 인명은 324개이고 회사 및 기관명은 503개이다.

KTS 대 본 논문에서 제안한 방법에서의 성능 비교는 인명의 경우와 회사 및 기관명의 경우에 다음 [표2] 및 [표3]과 같다. [표2]에서는 인명 고유 명사에 대한 성능 비교 실험인데 324개의 인명 고유 명사중에 KTS시스템은 286개를 추출하여 이중 175개가 올바른 고유 명사고 본 시스템은 301개를 추출하여 이중 267개가 올바른 고유 명사였으며, [표3]은 회사 및 기관명에 대한 성능 평가인데 총 503개의 회사 및 기관 고유 명사 중 KTS시스템은 추출한 412개 중 정확한 고유 명사가 222개였고 본 시스템은 439개를 추출하여 이 중 323개가 정확한 고유 명사였다.

성능 / 시스템	재현율 (recall rate)	정확률 (precision rate)
KTS 시스템	54.1 %	61.9 %
본 시스템	82.4 %	88.7 %

[표2] 인명에 대한 색인어에 대한 정확률과 재현율

성능 / 시스템	재현율 (recall rate)	정확률 (precision rate)
KTS 시스템	44.1 %	53.9 %
본 시스템	64.2 %	73.6 %

[표3] 회사/기관명에 대한 색인어에 대한 정확률과 재현율

[표2]와 [표3]에서의 성능 비교 결과에서 인명에 대한 색인어 추출 성능이 더 우수하게 나타났는데 이는 인명 고유 명사에 대한 처리에 있어서 다른 고유 명사 처리보다는 더 많은 정보를 이용하여 미등록 고유 명사를 추정했기 때문이다.

4.3 추출 실패와 오인식한 경우

실험에서 추출하지 못한 경우와 추출하였어도 오인식한 경우는 다음 [예4.1],[예4.2]와 같고 그외 띄어쓰기 오류로 인하여 추출에 실패한 경우가 있었다.

[예4.1]은 이름 고유 명사와 호칭이 보통 명사와 붙어있는 관계로 조사 앞 전체를 고유 명사로 한 경우로 추출하지 못한 경우이다.

예4.1) "김선홍회장주제로 그룹사장단회의를"

김선홍회장주제(고유명사) + 로(조사) (X)

김선홍(이름 고유 명사) + 회장(호칭) + 주제(보통 명사) + 로(조사) (O)

[예4.2]은 고유명사를 이름 고유 명사와 고유 명사로 추출한 경우인데 "현대종"이 이름 고유 명사 판별 알고리즘 과정을 통해 이름 고유 명사 추출에 성공해서 잘못 추출된 경우이다.

예4.2 "현대종합금속사장을"

현대종(이름 고유 명사) + 합금속(고유 명사) + 사장(호칭) + 을(조사) (X)

현대종합금속(고유 명사)+사장(호칭)+을(조사) (O)

5. 결론

정보검색에서 시스템의 성능을 향상시키는데 있어서 정확한 색인은 매우 중요하다. 특히 신문기사나 온라인 뉴스 등의 텍스트에는 색인어로서의 가치가 높은 고유 명사들이 다수 포함되어 있어 이들에 대한 정확한 추출이 전체적인 정보검색의 정확도를 높이는데 매우 중요하다.

본 논문에서는 이러한 고유 명사 추출의 문제점을 살펴보고, 패턴 정보와 고유 명사의 부가적 정보를 복합적으로 이용한 고유 명사의 자동색인 방식을 제안하였다. 조선일보 경제면 기사를 대상으로한 색인의 재현율과 정확률 실험 결과, 패턴 정보를 이용해서 추출한 경우 재현율은 약 72%이고 정확률은 약 80%를 보였다. 인명 고유 명사의 경우에는 약 89%의 정확도를 보였다. 보다 많은 어휘사전의 구성과 추정 우선순위 판별과정에 관한 연구로 정확률을 보다 높일 수 있을 것으로 판단되어진다.

향후 과제로는 미등록어 처리를 하는데 있어 복합명사 단위로 나타나는 미등록어 처리를 단위 명사로 분할하는 방법과 올바른 색인을 수행할 수 있도록 문장처리시에 나타나는 중의성을 해결하는 문제, 또한 사전의 자동 확장 및 문장 패턴에 관한 연구가 필요하다.

참고문헌

[1] Gerard Salton, "Automatic Text Processing, "Addison Wesley Publishing Co., 1989.

[2] William B.Frakes, Ricardo Baeza-Yates, "Information Retrieval," PTR Prentice Hall, 1992.

[3] 정영미, "정보검색," 구미무역, 1993.

[4] 김영택, "자연언어 처리," 교학사, 1994.

[5] 오길록, 최기선, 박세영, "한글공학," 대영사, 1994.

[6] 김재훈, 서정연, "자연언어처리를 위한 한국어 품사태그," 한국과학기술원 인공지능연구센터, CAIR-TR-94-5 5, 1994.

[7] 최기선, "한국어 정보검색," 정보과학회 논문지, 제12권, 제8호, 1994

[8] 김판구, 조유근, "상호정보에 기반한 한국어 텍스트의 복합어 자동색인," 정보과학회 논문지, 제21권, 제7호, 1994

[9] 이상호, 김재훈, 조정미, 서정연, "부분 분석 결과를 공유하는 한국어 형태소 분석," 제 11회 통신 및 신호처리 워크샵 논문지, 1994.

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

[11] 강승식, 권혁일, 김동렬, "한국어 자동색인을 위한 형태소 분석 기능," 정보과학회 봄 학술발표논문지, 1995.

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