한국어 정보검색에서 복합명사 색인 실험[1]

강병주, 최기선, 윤준태
한국과학기술원 전산학과
대전시 유성구 구성동 373-1, 우:305-701
{bjkang, kschoi, jtyoon}@world.kaist.ac.kr

Compound Noun Indexing Experiments in Korean Information Retrieval

Byung-Ju Kang, Key-Sun Choi, Juntae Yoon
Department of Computer Science
Korea Advanced Institute of Science and Technology
373-1 Kusong-dong, Yusong-gu, Taejon, 305-701

요약

한국어 정보검색에서 복합명사의 불규칙한 표기 형태로 인하여 발생하는 색인과 질의의 불일치 문제는 단순명사 단위로 색인하고 질의함으로써 해결할 수 있지만 원래의 복합명사가 가지고 있던 정보를 상실함으로써 정확도의 하락이 예상된다. 따라서 보다 정교한 문서검색을 위해서는 복합명사를 색인으로 사용하는 것이 필요하다. 본 논문에서는 단순한 패턴을 이용한 복합명사 색인 방법으로부터 정교한 명사구 구문분석을 통한 복합명사 색인 방법까지 그 동안 연구되었던 대표적인 복합명사 색인 방법을 실험을 통하여 비교 평가하여 복합명사 색인의 검색성능에 대한 효과성을 검증한다.

1 서론

단어기반색인 (word-based indexing)은 상용 영어 문서검색시스템에서 가장 널리 사용되는 색인 방식이다. 단어기반색인에서 질의에 대한 문서의 적합성 정도는 그들 사이에 공유되는 단어의 수에 비례한다. 하지만 단어기반색인은 복합명사의 불규칙한 띄어쓰기로 인해 한국어에 대해서는 직접적으로 적용하기가 곤란하다. 한국어에서는 명사들이 비교적 자유롭게 다른 명사들과 결합하여 새로운 복합명사를 이룰 수 있다. 그리고 복합명사를 구성하는 단순명사들의 결합 강도의 정도는 개별 복합명사들에 따라 매우 다양하다. 따라서 복합명사를 하나의 표기단위로 간주하여 붙여 쓸지 아니면 띄어 쓸지는 부분적으로는 글 쓰는 사람의 기분에 부분적으로는 복합명사의 결합강도에 달려있다.

(1) '정보검색시스템'

예를 들어 복합명사 (1)은 다음과 같은 4가지 방식으로 쓰일 수 있다.

- 정보검색시스템
- 정보검색_시스템
- 정보_검색시스템
- 정보_검색_시스템

따라서 공백문자를 분리자로 해서 키워드를 추출한다면 질의와 문서의 색인 사이에 형태상 불일치를 피할 수 없다. 우리는 앞으로 이를 복합명사 띄어쓰기 문제 (spacing problem of compound noun)라고 부르기로 한다.

*복합명사 띄어쓰기 문제*는 단순명사 단위로 색인하고 또한 질의도 단순명사로 하면 해결될 수 있다. 이 방법은 형태소기반색인 (morpheme-based indexing)이라고 불린다. 하지만 불행히도 이 방법에는 문제가 있다. 복합명사는 때때로 단순명사들의 의미들의 단순한 합 (compositional sum)보다는 좀 더 구체적인 의미를 가지는 경우가 많고 또는 완전히 다른 의미를 가지기도 한다. 따라서 단순히 복합명사를 단순명사들로 분리하면 색인시 정보의 손실이 발생하게 되어

[1] 본고는 과학기술부의 지원을 받아 수행중인 '대용량 국어정보 심층처리 및 품질관리 기술개발'의 일환으로 이루어졌다.

패턴	규칙
1. N_1_N_2	N_1/N_2
2. N_1/의_N_2	N_1/N_2
3. N_1_N_2_N_3	N_1/N_2, N_2/N_3, $N_1/N_2/N_3$
4. N_1/의_N_2 _N_3	N_1/N_2, N_2/N_3, $N_1/N_2/N_3$
5. N_1 _N_2/의 _N_3	N_1/N_2, N_2/N_3, $N_1/N_2/N_3$

그림 1. 복합명사 생성 패턴과 규칙
(여기서 '_'는 공백문자를 있음을, '/'는 공백문자가 없음을 의미한다)

검색의 정확도 (precision)를 떨어뜨린다. 예를 들어 '데이타베이스'를 '데이타'와 '베이스'로 분리하여 색인하는 것은 적절치 않다. 왜냐하면 '데이타베이스'라는 개념은 단순히 '데이타'와 '베이스'의 의미의 단순 합으로는 표현될 수 없기 때문이다. 이러한 이유로 복합명사를 색인어에 포함시키기를 원한다면 또 다른 문제에 대한 해결이 필요하다. 먼저 띄어 쓰여진 복합명사의 복원이 필요하다. 이렇게 띄어 쓰여진 복합명사를 복구하는 일도 간단치 않은 일이지만 더욱 어려운 문제가 남아 있다. 한국어에서는 복합명사를 구성하는 단순명사들이 그들 사이에 구문관계를 가지는 경우가 많다. 따라서 복합명사와 같은 의미를 가지는 구나 절의 형태가 존재할 가능성이 많다. 예를 들어 복합명사 (2)는 (3)과 같은 구문구조에 해당한다.

(2) '정보검색'
(3) "정보를 검색하다"

질의가 (2)로 주어지면 형태소기반방법을 사용하여 단순명사 수준에서는 (3)을 포함한 문서와 매치될 수 있지만 복합명사 수준에서는 매치될 수 없다. 따라서 (3)에서 복합명사 (2)를 색인어로 추출하는 것이 필요하다. 우리는 이 문제를 복합명사 정규화문제 (normalization problem of compound noun)라고 부른다.

그 동안 한국어 정보검색에서 *복합명사 띄어쓰기 문제*와 *복합명사 정규화 문제*가 혼재되어 복합명사 문제를 다소 어렵게 보이게 한 것이 사실이다. 하지만 복합명사 문제를 이렇게 둘로 나누어 보면 두 번째 복합명사 정규화 문제는 영어 권의 구색인 (phrase indexing) 문제와 본질적으로 거의 동일하다.

영어 권의 구색인에 대한 연구는 역사가 깊다. 처음으로 체계적으로 그리고 대규모적으로 구문적 구색인 방법과 통계적 구색인 방법을 비교 평가한 Fagan의 연구 [15]로부터 최근 Zhai의 연구 [25]까지 그 동안 많은 연구가 있어 왔다. Zhai는 대규모 문서집합 (250M)에 대해서 구문적 구색인 방법을 실험하여 단어기반의 방법에 비해서 약 18%의 성능개선을 얻었다 [25]. 하지만 이전까지는 소규모의 문서집합에서 구의 좋지않은 통계적 특성 (주로 문서집합 내에서의 구의 낮은 빈도)때문에 구색인이 효과가 있다는 절대적인 실험 결과를 얻는데 실패하였다 [18].

대규모 한국어 실험문서집합이 없는 관계로 Zhai의 실험을 한국어에서 재현하는 것은 현재로서는 불가능하다. 하지만 본 연구에서는 한국어에서 복합명사가 빈번하게 사용된다는 한국어 복합명사의 통계적 특성에 착안하여 소규모 문서집합에서도 복합명사 색인의 효력이 있을 수 있음을 다양한 복합명사 색인 실험을 통해서 검증하고자 한다.

2 한국어 복합명사 색인

한국어 복합명사 색인 방법도 영어 권의 구색인 (phrase indexing)과 마찬가지로 크게 통계적 방법 (statistical method)과 구문적 방법 (syntactic method)으로 나눌 수 있고 구문적 방법은 다시 형판 기반 방법 (template-based method)과 파서 기반 방법 (parser-based method)으로 나뉘어 진다 [15, 18].

구문적 방법은 단어들 사이의 구문적인 관계에 근거하여 단어들을 묶어 복합명사를 구성하는 방법이다. 먼저 가장 단순한 방법은 복잡한 구문 분석 없이 문장 내에서 복합명사를 만들 수 있는 특정 구문 패턴 (또는 형판)을 이용하여 대략적으로 복합명사를 생성하는 것이다 [1, 2, 4, 5, 7, 10]. 기존 연구에 의하면, 대규모의 말뭉치를 사용한 결과는 아니지만, 실험 문서 집합 내의 복합명사의 93.53%가 그림 1의 간단한 5가지의 패턴과 규칙으로부터 추출되었다 [10].

(4) 음성/의 _인식
(5) 음성/인식

명사구 (4)에 그림 1의 패턴 2가 매치될 수 있고 그 결과 복합명사 (5)가 생성된다.

나머지 6~7%의 복합명사를 인식해 내기 위하여 더 정교하고, 많은 패턴과 규칙들이 사용되었다. 적게는 5개부터 많게는 58개까지의 구문 패턴이 사용되었다 [1, 2, 4, 7].

(6) 기계/의_문자_인식

형판기반 방법은 구현이 간단하다는 장점이 있지만 의미 없는 복합명사를 많이 생성할 가능성이 높다. 예를 들어 (6)의 명사구에서 3개의 복합명사 - '기계문자', '문자인식', '기계문자인식' - 가 패턴 4에 의해 생성된다. 하지만 '기계문자'는 의미 없는 명사열이고 그보다는 '기계인식'이 보다 적절한 복합명사 후보이다. 이 경우에 보다 정확한 복합명사 생성을 위해서 명사구 (6)이 (7)과 같은 구문 (또는 의존) 구조를 가진다는 정보가 필요하다.

(7) 기계 문자 인식

완전 구문분석 또는 부분 구분분석을 시도하는 파서기반 방법은 최근에야 적용되기 시작하였다 [8, 11, 12, 22, 23]. 복합명사를 구성하는 명사-명사 쌍 사이에는 중심어-수식어 (head-modifier) 관계 또는 서술어-논항 (predicate-argument) 관계가 성립한다. 따라서 명사구로부터 중심어-수식어 관계나 서술어-논항 관계를 가지는 명사들을 묶어서 복합명사를 추출할 수 있다. 명사구 구문 분석에서 구조적인 중의성을 피할 수 없기 때문에 올바른 분석을 위해서 대규모 구문 트리뱅크 (tree bank)로부터 추출된 명사-명사 간 공기정보가 사용된다 [8, 22, 23]. 이전의 명사구 예 (6)에서 "기계의 문자 인식"은 (기계, 문자)의 결합강도가 (기계, 인식)보다 높으면 [[기계 문자] 인식]으로 분석되고, (기계, 인식)의 결합강도가 (기계, 문자)보다 높으면 [기계 [문자 인식]]으로 분석된다 [22, 23]. 이 경우 [기계 [문자 인식]]이 옳은 분석이다. 분석 결과 2개의 서술어-논항 관계, (문자, 인식)과 (기계, 인식)이 인식되고 그로부터 복합명사, '문자인식'과 '기계인식', 이 생성된다.

(8) "기계가 문자를 인식하다"

또한 문 단위에서도 완전한 구문분석을 통해서 명사들 사이에 서술어-논항 관계를 파악하여 복합명사를 생성할 수 있다. (8)에서 서술어-논항 관계, (기계$_{주어}$, 문자$_{목적어}$, 인식$_{서술어}$), 성립한다. 분석결과, <주어-서술어> 관계와 <목적어-서술어> 관계로부터 복합명사 '기계인식'과 '문자인식'이 생성될 수 있다. 또한 <주어-목적어-서술어> 관계로부터 '기계문자인식'이 가능한 복합명사 후보이다.

하지만 중심어-수식어 관계와 서술어-논항 관계를 만족한다고 해서 반드시 의미 있는 복합명사가 되는 것은 아니다. 따라서 복합명사의 자격을 어떻게 판단할 것인 지가 또 다른 주요 연구 대상이다. 대규모 말뭉치에서 일정한 빈도 이상으로 나오는 복합명사만을 추출할 수 도 있을 것이다.

통계적 방법은 단순명사의 출현빈도와 단순명사들 사이의 공기빈도만을 고려하여 통계적으로 가치 있는 명사열을 복합명사로 생성하는 방법이다 [15]. 다음의 두 연구 [6, 19, 24] 는 순수한 통계적인 방법이라고는 할 수 없으나 통계적인 복합명사 어휘 특성에 기반하고 있다. [24]는 3단어 이상의 복합명사에서 중심어-수식어 관계의 2단어 복합명사를 생성한다. 중심어-수식어 관계의 판별은 단순히 두 명사의 상호정보에 의한다. 상호정보가 일정한 임계치 이상인 단순명사 bigram을 복합명사로 추출한다. 데이타 희소성 (data sparseness) 문제를 피하기 위하여 명사들의 세부 품사를 정의하고 이들 품사 빈도와 품사 bigram 빈도를 앞의 어휘빈도 상호정보 계산에 반영하였다. KT 문서집합의 일부분만 사용해서 실험하였는데 약 형태소기반 방법에 대해 2.7%의 정확도 개선이 있었다.

[6, 19]는 복합명사의, 문서집합 내에서의, 통계적인 특성에 따라, 항상 붙여져 쓰이는 복합명사 (유형 1), 단순 명사로 나누어도 전혀 의미의 변화가 없는 복합명사 (유형 2), 유형1과 유형 2 이외의 복합명사, 등의 3가지 유형으로 구분하고, 질의에 복합명사가 포함된 경우 우선 복합명사의 유형을 검색 대상 문서집합에서 단어의 통계적인 분포를 이용하여 알아내고 그 유형에 따라 검색방법을 달리한다. 이는 검색단계에서 질의 복합명사의 가중치를 조정함으로써 이루어지는데 복합명사와 단순명사 사이의 의존 관계를 색인단계가 아니라 검색단계에서 고려하고자 하는 것이다. KT 문서집합 2.0에서 실험하였고 약 7.7%의 정확도 향상을 보였다.

두 실험 모두 벡터공간모델이 아니라 확장부울린모델에 기반하고 있기 때문에, 그리고 [24]는 KT 문서집합의 일부만 사용하였기 때문에 그들의 실험결과를 본 실험과 직접적으로 비교하기는 곤란하다.

3 복합명사 색인 실험

3.1 실험 환경

이번 실험의 또 다른 주요 목적 중의 하나가 대표적인 한국어 정보검색 방법들을 투명하고 객관적으로 성능을 비교 평가해보고자 하는 것이다. 따라서 표준 실험문서집합, 표준 검색엔진 등을 사용하는 것이 중요하다.

본 실험에서는 한국어 정보검색 실험문서집합으로 가장 대표적인 KT 문서집합 2.0 [14]과 KRIST 문서집합 [9]이 사용된다. 몇 가지 중요한 문서집합의 특성이 표 1에 나와있다. 두 문서집합의 가장 큰 차이점의 하나는 문서집합의 크기이다. KRIST가 KT에 비해 3배 정도 크다. 또 다른 중요한 차이는 질의의 길이이다. KT에서는 질의의 길이가 1 문장 정도로 짧고 KRIST는 2-5 문장정도로 비교적 길다. 실험의 객관성을 확보하기 위해 모든 질의를 하나도 빠짐없이 그리고 변경 없이 사용하였다.

	KT	KRIST
주제	컴퓨터, 정보과학	생명과학, 기계/전기/전자 공학
출처	기술보고서, 신문기사	논문 서지정보
문서 수	4,414 (4.4M)	13,515 (12M)
질의 수	50	30
질의 형태	자연언어, 부울린	자연언어

표 1. KT와 KRIST 실험 문서집합의 특성

검색엔진은 Cornell 대학의 SMART 시스템을 사용하였다. 한국어 처리를 위하여 SMART 시스템의 한국어 버전이 사용되었다 [17]. SMART 시스템은 벡터검색모델에 기반을 두고 있다 [20]. 본 실험에서 색인 가중치 방법 (term weighting scheme)으로 일반적으로 한국어 정보검색에 가장 좋은 성능을 보인다는 *atc.atc* 방법을 사용하였다 [17].

3.2 실험 설계

6가지의 구문적 복합명사 색인 방법을 고안하였다. 각 색인 방법은 KT와 KRIST에 대해서 평가되었다. 각 실험은 다음과 같은 이름붙이기에 따라서 구분된다.

<experiment_name>:=
    <test_collection><indexing_method>
where
    <test_collection>:=KT|KR
<indexing_method>:=W|M|E|B|R|P

3.2.1 단어기반 색인 (W)

붙여 쓰여진 복합명사 전체가 하나의 색인어로 추출된다. 하지만 복합명사 띄어쓰기 문제와 한국어 텍스트에서 매우 빈번한 복합명사의 사용으로 인해서 이 방법은 매우 좋지 않은 성능을 보인다. 현재 이 방법은 실제적으로 한국어 정보검색시스템에서 더 이상 사용되지 않는다.

3.2.2 형태소기반 색인 (M)

형태소기반 색인에서 색인 단위는 단순명사이다. 질의에 있는 복합명사도 단순명사로 분리된다. 따라서 복합명사 띄어쓰기 문제는 해결된다. 이 방법은 실제로 영어의 단어기반의 색인 방법에 해당되므로 본 실험에서 평가 기준선 (baseline)으로 사용된다.

3.2.3 복합명사기반 색인 (E)

이 방법은 형태소기반 방법과 단어기반 방법을 합친 방법이다. 즉, 붙여 쓴 복합명사 자체도 색인하고 복합명사를 구성하는 단순명사도 색인한다. 이 방법은 일반적으로 형태소기반 색인보다 성능이 좋을 거라고 믿어지는 색인 방법이다.

3.2.4 명사 bigram 방법 (B)

복합명사가 띄어 쓰여 있는 경우 원래의 복합명사는 텍스트 상의 연속된 명사들로부터 복구될 수 있다는 휴리스틱을 이용한다. 이 방법에서는 연속된 단순명사들로부터 각 명사 bigram을 복합명사로 추출한다. 이 방법은 가장 간단한 방법이지만 따라서 단점도 많다. 2단어 복합명사만을 생성하고 잘못된 복합명사를 많이 생성한다.

3.2.5 패턴 기반 방법 (R)

명사구가 NP:={N+[의]}*+{N}$^+$ 의 형태를 가진다고 가정하고 그림 1의 5가지의 패턴과 규칙을 사용하여 2단어 혹은 3단어 복합명사만을 생성한다. 이 방법은 분명히 명사 bigram 방법보다는 많은 복합명사를 생성해낼 수 있지만 반대로 잘못된 복합명사도 더 많이 생성할 가능성이 있다.

3.2.6 명사구 구문분석 방법 (P)

마찬가지로 NP:={N+[의]}*+{N}$^+$ 형태의 단순화된 명사구를 가정하고 명사구 구문분석을 하여 복합명사를 생성한다. 명사구 구문분석의 구조적인 중의성 문제를 해결하기 위하여 대량의 말뭉치로부터 추출된 명사-명사 간 공기 정보를 이용한다. 명사구 구문분석을 통해 머리-수식어 모든 명사 쌍들이 복합명사로 추출된다. 대략적인 색인어 추출과정은 다음과 같다:

1. 형태소분석기를 사용하여 명사를 색인으로 추출한다. 형태소 분석 과정에서 모든 붙여 쓴 복합명사는 단순명사로 분리된다. 이때 원래의 텍스트상에서 연속된 명사들은

표시를 해둔다. 그리고 소유격조사 '의'가 붙어 있는 경우도 연속된 것으로 본다.

2. 연속된 명사 열들로부터 구문분석을 통해 중심어-수식어 관계에 있는 명사 쌍들을 복합명사로 추출한다.
3. 연속된 전체 명사열도 복합명사로 추출한다.

예를 들어 명사구 (6)에서 추출되는 색인들은 다음과 같다:

- 단순명사: 기계, 문자, 인식
- 중심어-수식어 명사 쌍: 기계/인식, 문자/인식
- 연속된 전체 명사열: 기계/문자/인식

이 방법은 본 논문의 복합명사 색인 방법들 중 가장 정교한 방법이며, 방법 B와 R에 비해 정교하게 복합명사를 생성하기 때문에 추출되는 복합명사의 수는 방법 R보다 적다. 하지만 3단어 이상의 복합명사를 포함하기 때문에 방법 B보다는 많은 복합명사를 생성한다 (표 2). 그러나 이 방법에서는 절이나 문 단위의 서술어-논항 관계의 복합명사는 생성하지 않았다.

색인방법	총 색인 수	생성된 복합명사의 수
KTW	375,265	-
KTM	406,472	-
KTE	492,758	-
KTB	571,456	78,698
KTR	804,550	311,792
KTP	693,215	200,457

표 2. 총 색인 및 생성된 복합명사의 수 비교 (KT 에서)

4 결과 분석

실험 결과는 표 3, 4와 같다. 방법 P (명사구 구문분석 방법)가 KT, KRIST 모두에서 가장 좋은 결과를 내었지만, 복합명사를 전혀 사용하지 않는 방법 M (형태소기반 색인)에 비해서는 KT에서 3.02% , KRIST에서 1.31%의 미미한 정확도 개선 밖에 얻을 수 없었다. 방법 B와 방법 R은 오히려 방법 M보다도 성능이 많이 떨어지는 것으로 나왔는데 이는 복합명사 생성 방법이 정교하지 않아 많은 부적절한 복합명사가 오히려 단순명사의 가중치를 떨어뜨리는 역작용을 가져와 전체적인 성능이 큰 폭으로 떨어진 것 같다. 하지만 방법 P에서는 조금이지만 정확도가 올라갔다는 사실은 복합명사 생성이 더욱 정교해지면 성능이 더 올라갈 가능성도 있다는 점에서 고무적이다. 그리고 여기에는 자세한 내용을 실을 수는 없었지만 현재 계속되고 있는 결과 분석에서 복합명사 질의에서 복합명사를 구성하는 단순명사를 질의에 포함시키느냐 아니냐에 따라 방법 M에 대해, KT에서, 큰 폭으로 성능개선을 얻을 수 있었다. 분리되면 안 되는 복합명사가 분리되었기 때문에 일어나는 성능 하락의 폭이 큰 것 같다. 또한 개별 질의별로 분석해 본 결과 복합명사 색인 및 질의에 의해 적합문헌의 순위가 큰 폭으로 올라가는 것을 확인할 수 있었다.

일부 KT 질의에 대해 적합성 평가를 조사해 본 결과 적합성 평가가 잘못된 경우가 적지 않았다. 이러한 잘못된 적합성 평가도 복합명사 색인의 효과를 감소시키는데 일조하고 있었다. 또한 일부 질의는 벡터공간모델에 적합하지 않은 질의였기 때문에 복합명사 색인의 효력이 전혀 반영되지 못하는 결과를 가져왔다.

이번 실험에서는 복합명사와 구성 단순명사 사이의 의존 관계를 고려하지 않은 색인 및 검색 모델이었다. 사실 보다 바람직한 복합명사 색인 및 검색 모델의 구축을 위해서는 이 의존관계가 복합명사와 단순명사의 가중치에 반영되어야 할 것이다 [21].

방법 P에서 생성된 복합명사를 조사해본 결과 의미 없는 복합명사들이 아직 많이 있음을 알 수 있었다. 이러한 복합명사들은 결국 문서에 대한 잡음으로 작용해서 전체적인 검색 성능을 떨어뜨리게 될 것이다. 하지만 이러한 잡음이 검색성능에 얼마나 영향을 미치는지 그리고 검색성능의 큰 폭의 저하 없이 얼마만큼의 잡음이 허용될 수 있는지에 대해서는 아직 알 수 없다.

5 결론

우리는 한국어 정보검색에서 복합명사 문제를 2가지 하위 문제, *복합명사 띄어쓰기 문제*와 *복합명사 정규화 문제*, 로 분리하였다. 복합명사 띄어쓰기 문제는 복합명사를 단순명사로 분리함으로써 해결할 수 있고, 그 결과 단순명사만을 포함하는 문서에서 복합명사를 생성하는 문제 (복합명사 정규화 문제)는 영어권의 구색인 문제와 본질적으로 같은 문제라고 볼 수 있다.

영어의 구색인은 검색성능에 대한 효과성을 입증하기 위해 대규모 실험문서집합을 기다려야 했지만 한국어에서는 복합명사가 빈번하게 사용된다는 통계적 특성때문에 비교적 작은 실험문서집합에서도 효과가 있을 수 있다는 점에 착안하여 여러 가지 다양한 구문적 복합명사 색인 방법을 실험하였다. 가장 정교한 명사구 구문분석 방법만이 미미한 성능개선을 가져왔지만 그 가능성은 볼 수 있었다.

Recall	Precision					
	KTW	KTM	KTE	KTB	KTR	KTP
0.00	0.736 9	0.7854	0.7397	0.7189	0.7303	0.7450
0.10	0.6528	0.6581	0.6488	0.6704	0.6692	0.6808
0.20	0.5331	0.5596	0.5523	0.5379	0.5367	0.5633
0.30	0.4442	0.5161	0.4969	0.4542	0.4595	0.5057
0.40	0.3620	0.4681	0.4627	0.4078	0.4075	0.4457
0.50	0.2876	0.3844	0.4111	0.3698	0.3688	0.4092
0.60	0.2163	0.2990	0.3176	0.2985	0.2987	0.3325
0.70	0.1357	0.2360	0.2604	0.2254	0.2269	0.2688
0.80	0.1024	0.1725	0.2200	0.1657	0.1670	0.2308
0.90	0.0459	0.0877	0.1140	0.0709	0.0703	0.0989
1.00	0.0371	0.0553	0.0875	0.0471	0.0459	0.0692
11-pt Avg.	0.3231	0.3838	0.3919	0.3606	0.3619	0.3954
% change	-15.81	-	2.11	-6.04	-5.70	3.02
3-pt Avg.	0.3077	0.3722	0.3945	0.3578	0.3575	0.4011

표 3. KT 문서집합에서의 실험 결과

Recall	Precision					
	KRW	KRM	KRE	KRB	KRR	KRP
0.00	0.6116	0.6674	0.6083	0.6639	0.6603	0.6725
0.10	0.5377	0.6543	0.5810	0.6384	0.6231	0.6671
0.20	0.4780	0.5677	0.5170	0.5536	0.5609	0.5812
0.30	0.3839	0.4982	0.4784	0.4428	0.4279	0.5133
0.40	0.2829	0.3897	0.3767	0.3612	0.3181	0.3870
0.50	0.2493	0.3383	0.3279	0.3086	0.2788	0.3259
0.60	0.1897	0.2829	0.2465	0.2444	0.2284	0.2714
0.70	0.1389	0.2185	0.1937	0.1808	0.1650	0.2100
0.80	0.1100	0.1673	0.1805	0.1437	0.1323	0.1836
0.90	0.0686	0.1359	0.1432	0.1081	0.1046	0.1481
1.00	0.0538	0.1103	0.1159	0.0892	0.0855	0.1232
11-pt Avg.	0.2822	0.3664	0.3426	0.3395	0.3259	0.3712
% change	-22.98	-	-6.49	-7.34	-11.05	1.31
3-pt Avg.	0.2791	0.3578	0.3418	0.3353	0.3240	0.3636

표 4. KRIST 문서집합에서의 실험 결과

명사구 구문분석 방법의 좋지 않은 성능에 대한 계속적인 원인분석에서 정확도 향상을 방해하는 2가지 원인을 생각해 볼 수 있겠다. 첫째, 복합명사와 그 구성 단순명사들을 독립적으로 보고 그들 사이의 의존관계를 색인 및 검색 모델에 제대로 반영하지 못하였기 때문이라고 생각된다. 따라서 앞으로의 연구는 복합명사와 단순명사들 사이의 의존관계를 반영하는 색인 가중치 방법을 개발하는데 우선할 것이다. 두 번째는 보다 정교한 복합명사 생성의 필요성이다. 방법 P에서도 실제로 많은 부적절한 복합명사를 생성하였다. 의미 없는 복합명사의 생성은 최대한 억제하면서 문서의 구체적인 선택 단서가 될 수 있는 복합명사는 반드시 생성하는 보다 정확한 방법에 대한 연구가 필요하다.

참고문헌

[1] 김동일, 이신원, 윤후병, 장재우, 정성종. "한국어 텍스트에서의 복합명사구 출현빈도 분석에 관한 연구". 한국정보과학회 봄 학술발표논문집, 1994.

[2] 김민정, 권혁철. "한국어 특성을 이용한 자동 색인 기법". 한국정보과학회 가을 학술발표논문집, 1992.

[3] 김성혁, 서은경, 이원규, 김명철, 김영환, 김재군. "자동색인기 성능 시험을 위한 Test Set 개발". 정보관리학회지 11(1), 1994.

[4] 김판구, 조유근. "상호 정보에 기반한 한국어 텍스트의 복합어 자동 색인". 한국정보과학회 논문지, 21(17), 1994.

[5] 남세진, 이지연, 신동욱, 채미옥. "복합명사의 통계적 처리에 대한 평가". '96 한글 및 한국어 정보처리 학술대회, 1996.

[6] 박영찬, 최기선. "통계적 명사 패턴 분류를 이용한 복합명사 검색 모델". '96 한글 및 한국어 정보처리 학술대회, 1996.

[7] 서은경. "구문 통계적 기법을 이용한 한국어 자동 색인에 관한 연구". 정보관리학회지 10(1), 1993.

[8] 양성현, 정의석, 윤준태, 송만석. "명사의 연어 정보와 서술성 명사의 공기 정보를 활용한 복합명사의 분석 및 자동색인". '97 한글 및 한국어 정보처리 학술대회, 1997.

[9] 이준호, 최광남, 한현숙, 김종원, 남성원. "정보검색 연구를 위한 KRIST 테스트 컬렉션의 개발". 정보관리학회지 12(2), 1995.

[10] 이창열, 강현규, 장호욱, 박세영. "자동 키워드 제작기 시스템". '93 한글 및 한국어 정보처리 학술대회, 1993.

[11] 이현아, 이종혁, 이근배. "구문분석과 공기정보를 이용한 개념 기반 명사구 색인 방법". '95 한글 및 한국어 정보처리 학술대회, 1995.

[12] 이현아, 이종혁, 이근배. "단문 분할을 통한 명사구 색인 방법". 정보과학회논문지, 24(3), 1997.

[13] 은종진. "효율적인 구문 분석을 위한 전처리기 구현과 복합 명사의 구조 분석". 한국과학기술원, 석사학위논문, 1996.

[14] 최기선. "지능형 정보검색 환경", 보고서, 한국통신, 1995.

[15] Fagan, J. L. "Experiments in Automated Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods". Ph.D. Thesis, Department of Computer Science, Cornell University, Ithaca, NY, 1987.

[16] Lauer, M. "Corpus Statistics Meet the Noun Compound: Some Empirical Results". In *Proceedings of 33rd Annual Meeting of ACL*, 1995.

[17] Lee, J. H. and Ahn, J. S. "Using n-grams for Korean Text Retrieval". In *Proceedings of 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

[18] Lewis, M. "Representation and Learning in Information Retrieval". Ph.D. Thesis, Department of Computer and Information Science, University of Massachusetts, Amherst, MA, 1992.

[19] Park, Y. C. and Choi, K. S. "A Korean Compound Noun Retrieval Model using Statistical Noun-Pattern Categorisation". In *Proceedings of the 17th International Conference on Computer Processing of Oriental Language (ICCPOL97)*, 1997.

[20] Salton, G., Wong, A. and Yang, C. S. "A Vector Space Model for Automatic Indexing". Communications of the ACM, 18(11), 1975.

[21] Strzalkowski, T. "Natural Language Information Retrieval". Information Processing and Management, 31(3), 1995.

[22] Yoon, J. T. and Song, M. S. "Yet Another Compound Noun Analysis Using Co-occurrence Relation". In Proceedings of NLPRS'97, 1997.

[23] Yoon, J. T., Choi, K. S., Yang,. S. and Song, M. S. "Document Indexing Based on Noun Compound Analysis and Term Normalization". accepted In *Proceedings of the 3rd International Workshop on Information Retrieval with Asian Languages (IRAL'98)*, 1998.

[24] Yun, B. H., Cho, M. J., and Rim, H. C. "Korean Compound Noun Indexing based on Lexical Association and Conceptual Association". In *Proceedings of the 2nd International Workshop on Information Retrieval with Asian Languages (IRAL'97)*, 1997.

[25] Zhai, C. "Fast Statistical Parsing of Noun Phrases for Document Indexing. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, 1997.