명사의 연어 정보와 서술성 명사의 공기 정보를 활용한 복합명사 분석 및 자동 색인 *

양성현°, 정의석, 윤준태, 송만석
연세대학교 컴퓨터과학과 한글 정보처리 연구실

Analysis of Compound Noun and Automatic Indexing Using Collocation Information of Nouns and Co-occurrence Information of Predicative Nouns

Seong-hyeon Yang, Eui-sok Chung, Jun-tae Yoon, Mansuk Song
Natural Language Processing Lab., Department of Computer Science, Yonsei University

요약

복합명사로부터 적절한 색인어를 추출하는 것은 한국어 정보검색 시스템의 성능 향상에 중요한 역할을 한다. 본 논문에서는 복합명사로부터 색인어 추출을 하기 위해 복합명사 구문 구조 분석 결과를 활용한다. 단일명사가 3개 이상 결합된 복합명사의 경우 각 단일명사의 구문적 관계를 파악하여 적절한 괄호치기를 한 후 색인어를 추출하면 보다 좋은 결과를 얻을 수 있다. 이러한 복합명사 구문 구조 분석을 위해 말뭉치로부터 구조적 중의성이 없는 연어 관계의 완전 복합명사와, 서술성 명사와 공기하는 명사쌍을 추출한 결과를 이용한다. 또한 서술성 명사는 이와 공기하는 명사와 결합되어 복합명사를 이룰 가능성이 많고, 복합명사의 형태로 인식되어야만 정확한 의미 파악이 가능하다. 서술성 명사와 공기하는 명사를 파악하여 복합명사를 추출하기 위해서 부분 파서로 공기쌍을 찾아 복합명사 후보를 생성한 후, 이 후보 가운데 적합한 복합명사만을 선택하기 위해 말뭉치에서 추출한 완전 복합명사 사전을 통해 검증한다. 이러한 방법으로 서술성 명사에서 복합명사 형태의 색인어를 추출한다.

1. 서론

정보검색 시스템에서 색인어 선정을 하는 방법으로는 크게 통계학적 방법과 언어학적 방법이 있으며[1], 현재 한국어 정보검색 분야에서는 언어학적 방법인 형태소 분석과 복합명사 분석을 거쳐 색인어 추출을 한 후, 문헌 내의 색인어 출현 빈도에 따른 색인어 중요도를 평가하는 통계학적 방법으로 검색 시스템을 구현하는 방법이 주류를 이루고 있다.

문헌 중에 색인어 대상이 될 수 있는 단어는 대체로 명사에 국한되는데 이는 명사가 문헌의 중요한 의미를 나타낼 수 있는 언어 성분이기 때문이다. 한국어에서 명사 중심의 색인어 선정에서 중요한 문제가 되는 것은 복합명사이다. 복합명사는 띄어쓰기가 자유롭고, 단일명사가 복합명사로 결합되면서 구문적 관계를 내포한다. 복합명사 분석에는 일반적으로 두 가지 방법이 시도되고 있는데, 첫째는 직접적으로 복합명사의 형태로 등장하지는 않지만 복합명사로 받아들일 수 있는 유형을 판별하여 복합명사 형태로 색인을 하는 것이고, 둘째는 띄어쓰기 없이 표기된 복합명사를 단일명사로 분리하는 것이다.

본 논문에서는 말뭉치로부터 추출한 완전 복합명사 정보와 서술성 명사가 포함된 술어의 공기 정보를 이용하여, 복합명사 분석 및 이를 활용한 자동색인을 시도한다.

용언화 접미사 '하', '되'와 결합되어 서술어로 사용될 수 있는 서술성 명사는 그 자체로서 명사 색인어로 선정될 수도 있지만 이러한 형태의 색인은 문장 내에서 그 명사의 쓰임을 제대로 반영하지 못한다. 서술성 명사는 공기하는 다른 명사와 결합된 의미로서 문장 속에서의 명확한 의미가 드러난다. 이러한 공기쌍은 복합명사의 형태로 인식될 수 있으며, 서술성 명사로부터 복합명사 색인어를 추출하는 것은 각각을 단일명사로 색인하는 것에 비해 문헌 내에서 명사의 쓰임새를 정확히 반영하는 색인기법이 된다. 또한 3개 이상의 단일명사가 연속되는 명사열의 경우 복합명사로서의 구조 해석에 중의성이 발생하고, 이러한 중의성을 해소한 상태에서 올바른 명사 결합만을 색인하는 것은 각각의 단일명사를 색인하는 것에 비해 좋은 색인어를 도출해 낼 수 있다. 본 논문에서는 이러한 두 가지 방법의 색인기법을 시도하고 있다. 이러한 색인기법을 위한 학습 데이터로는 말뭉치로부터 추출한 연어관계의 완전 복합명사 및 서술성 명사의 공기정보를 활용하고 있다.

* 본 연구는 정보통신부 '97 국책연구개발사업[과제 번호 AB-97-B-0182] 연구비를 지원 받아 수행된 것임.

2. 서술성 명사로부터 복합 명사 생성

2.1 서술성 명사 색인의 필요성

서술어와 명사 간의 공기(co-occurrence)정보는 자연어처리 여러 분야에서 중요한 정보원이 될 수 있다. 또한 자동색인에 있어서는 서술성 명사가 포함된 술어와 공기하는 명사를 찾아 복합명사 형태로 색인하는데 있어 공기 관계의 파악은 매우 중요하다. 서술성 명사는 공기 관계를 파악함으로써 해당 명사의 쓰임새가 명확하게 드러나기 때문이다. 예를 들어, '문자를 정확히 인식하는'과 같은 구문에서, 형태소 분석 결과 명사를 색인어로 선택하는 일반적인 방법에 의해서 '문자'와 '인식'이 색인어로 선택될 수 있다. 이러한 색인어는 같은 문헌 내의 각각 다른 문장 속에서, '문자'와 '인식'이 사용된 것과 다름 없는 색인어를 추출함으로써, 예문의 의미를 정확히 반영하는 색인어가 될 수 없다. 위의 예문에서 '인식'이라는 서술성 명사가 문장에서 쓰이고 있는 의미를 반영하기 위해서는 '문자인식'이라는 형태의 복합명사로 인식되어야 한다. '서술성 명사+하다' 유형과 '서술성 명사+되다' 유형의 술어는, 연세 말뭉치 약 3000만 어절의 자동 태깅된 결과에 의하면 전체 서술어의 약 18%를 차지하고 있다. 또한 여러 가지 유형의 복합명사 중 약 20%를 차지하는 것이 서술성 명사에 의해 생성되는 복합명사이다[10]. 즉 이러한 유형의 복합명사를 올바르게 인식하여 색인어로 선정하는 것은 정보검색 시스템의 성능 향상에 기여한다.

2.2 서술성 명사 공기 관계로부터 색인어 추출

한국어에서의 다양한 접사들의 쓰임새를 파악하여 색인을 하려는 시도 중 하나로 이러한 서술성 명사로부터 적합한 색인어를 추출하려는 시도[7]가 있었으나, 올바른 공기쌍을 찾아내는 것과 더불어 공기 관계에 의해 파악된 복합명사 후보의 타당성을 점검하는 작업이 이루어지지 않으면 적합하지 않은 색인어를 생성할 가능성이 많아진다.

공기 관계 파악을 위해서는 문장의 구문 구조를 제한적으로나마 파악하여야 한다. 본 논문에서는 몇 가지 휴리스틱을 사용하여 술어와 결합되는 명사쌍을 찾아주는 부분 파서를 이용한다[6]. 부분파서의 공기쌍 추출의 정확도는 약 85-90%이다. 이러한 부분 파서에서 추출된 공기쌍의 정보는 다음과 같은 형태이다.

Head (서술어)	POS (서술어의 품사)	Noun (공기하는 명사)	PP (명사에 결합된 조사)

예) <인식하>-<동사>-<문자>-<를>

'문자를 정확히 인식하는'과 같은 구문에서는 '문자인식'이 추출될 수 있는데, 이러한 방법으로 복합명사 형태를 구성했을 때 올바른 복합명사만이 생성되는 것은 아니다. 올바른 복합명사의 형태는 각 서술성 명사가 요구하는 격관계에 의해서 파악될 수 있으나[7], 이러한 격 관계 정보를 구축하는 것은 어려운 작업이다. 또한 휴리스틱에 의한 공기쌍 추출기의 오류로 인해 잘못된 색인어가 추출될 수도 있으며, 직관적으로 올바른 복합명사인지 애매하거나 일반적으로 사용되지 않는 복합명사 후보도 생성된다.

공기 관계에 의한 복합명사 후보 중에서 올바른 쌍만을 추출하기 위해 말뭉치에서 연어로 등장하는 명사쌍과 형태소 분석기 사전을 검증 집합으로 선정한다. 서술성 명사는 2음절의 한자어 명사가 대부분이다. 또한 이와 공기하는 명사도 대부분 2음절 이상이다. 본 논문에서는 4음절 이상의 복합명사만을 고려 대상으로 한다.

색인어를 추출할 때, 문헌에서 위와 같은 공기쌍을 찾아 서술성 명사가 포함된 술어 중 복합명사 형태로 타당한 후보만을 색인어로 선정한다. 이를 위해서 적합한 색인어 후보 집합을 가지고 있어야 한다.

2.3 복합명사 후보 사전의 구성

1) 형태소 분석기를 위한 명사 사전에서 4음절 이상의 명사를 추출한다. 이를 CN_1이라고 한다. CN_1은 다수의 복합명사를 포함하고 있다.

2) 연세 말뭉치 약 3000만 어절과 KT SET 95로부터 자동 태깅에 의해 생성된 형태소열 중에서

$w_1, w_2, \ldots, w_{n-1}, w_n (w_1, w_n \notin N, w_2, \ldots, w_{n-1} \in N, n \geq 4,$ N은 명사의 집합)

이 등장할 경우, $(w_2 w_3), \ldots, (w_{n-2} w_{n-1})$과 같이 2개 이상 연속된 명사열이 있을 경우 두 명사를 묶어 하나의 복합명사 후보를 생성한다. 이를 CN_2라 한다.

이 두 집합을 합하여 복합명사 후보를 검증할 수 있는 집합을 CN을 구축한다.

즉 CN을 $CN_1 \cup CN_2$라 한다.

공기 관계는 2진 관계만을 추출하기 때문에 2)의 방법에서 2개씩의 쌍만을 추출하였으며, 대부분이 2음절의 한자어인 서술성 명사에 의한 복합명사 검증 집합을 설정하면서, 1음절의 단일명사를 고려하여 발생하는 오류를 없애기 위해 1)의 방법에서 4음절 이상의 명사만을 추출하였다.

이렇게 구축된 사전의 내용은 대량의 말뭉치에서 실제 사용되고 있는 복합명사집합(CN_2)이거나 수작업에 의한 신뢰성 있는 형태소 분석기 사전의 내용(CN_1)이므로 신뢰성이 높다고 판단되며, 약 3000만 어절의 균형 있는 말뭉치에서 추출한 데이터이므로 실제 사용되는 대부분의 복합명사를 포함하고 있다고 판단된다.

2.4 서술성 명사 공기 관계에 의한 올바른 복합명사 집합 생성

이와 같이 구축된 복합명사 집합 *CN*을 바탕으로 서술성 명사의 공기쌍에 의한 복합명사 후보를 검증한 집합은 다음과 같이 구축된다.

1) 부분 파서를 이용한 공기쌍 추출기[6]를 약 3000만 어절의 연세 말뭉치와 KT SET 95에 적용하여 얻어낸 쌍 중에서, 서술어의 어간이 '서술성 명사+하' 또는 '서술성 명사+되'인 집합만 선택한다.

2) 1)에 의해 생성된 집합에서 조사 및 술어정보 중 접사 '하'와 '되'를 제외하고, 서술어에서 서술성 명사만 추출하여, 이와 공기하는 명사와 결합시켜 복합명사 후보 형태를 생성한다. 이 집합을 *CN_CANDIDATE* 이라고 하자.

또한 *CN_COOC*는 $CN \cap CN\_CANDIDATE$ 라 한다.

*CN_COOC*는 2)에 의해 생성된 복합명사 후보 중 말뭉치와 형태소 분석기 사전을 통해 검증된 올바른 복합명사 집합이 된다.

색인어를 추출하는 과정에서는 색인되는 문헌에서 추출된 서술성 명사의 공기 관계에 의한 복합명사 후보를 *CN_COOC*에서 검색하여 사전에 등록된 복합명사만을 색인어로 선정하게 된다. 물론 서술성 명사 자체와 이와 공기하는 명사도 각각 단일명사로 색인된다. 이는 복합명사 분리에 의한 재현률을 높이기 위한 기법과 같은 맥락에서 수행된다. *CN_COOC*에 포함되어 있지 않은 복합명사 후보는 색인어로 선정되지 않고, 각각의 단일 명사만 색인어로 선정된다. *CN_COOC*집합을 미리 구축하는 이유는 색인어 추출시 매번 대량의 *CN*에서 검증하는 것은 사전 검색 시간을 매우 많이 요구하기 때문이다.

서술성 명사의 의미를 명확하게 하기 위한 방법으로 공기관계 파악에 의한 복합명사 형태로의 색인 기법을 도입하면 정확률을 높일 수 있다. 실제로 '자동번역', '문자인식', '위성방송' 등의 질의어에서 '번역', '인식', '방송'은 서술성 명사로서, 이러한 단어들이 문헌 중에서 용언화 접미사 '하', '되'와 결합하여 서술어로 쓰일 때, 단일명사로 색인되는 것보다 복합명사로 색인되었을 때 명확한 의미가 드러나게 된다.

3. 명사 구문 분석에 의한 색인어 선정[4]

복합명사는 단일명사들로 구성되어 있고, 간단한 CFG규칙 N->N N으로 분석될 수 있다. 이러한 규칙은 3개 이상의 단일명사로 구성된 복합명사에 대해 구조적 모호성을 발생시킨다. 이러한 구조적 모호성이 있는 복합명사의 구조를 밝혀, 올바른 복합명사만을 색인어 후보로 선정할 경우 정보검색 시스템의 성능이 향상된다[5]. 복합명사 구조 분석을 위해 분석 대상이 되는 복합 명사 내부 구조는 다음 두 가지로 가정한다.

첫째, 보어와 술어 관계
[예 1] 수소 이온 교환 / 정보 검색

위의 예에서 '교환'과 '검색'은 명사구 내에서 서술어로서의 역할을 하고 있고, '이온'이나 '정보'는 목적어로서의 역할을 하고 있다. 이러한 복합명사 구조는 서술성 명사인 '교환'과 '검색'이 목적어를 취하는 술어-논항 구조로 파악될 수 있다.

둘째, 술어적 성격이 없는 명사들 간의 관계
[예 2] 복합 명사 / 고층 건물 옥상

위의 예는 앞의 명사가 관형사이거나 두 명사가 일정한 관계를 가지고 합쳐진 복합 명사이다. 이러한 복합 명사는 말뭉치에서 두 개의 명사 간의 공기 강도를 측정함으로써 구조를 분석할 수 있다.

3.1 연어 관계의 완전 복합명사로부터의 학습

구조적 모호성을 해결하기 위한 학습 데이터로 말뭉치로부터 추출한 완전 복합명사(complete compound noun)를 사용한다. 완전 복합명사는 구조적 중의성이 없는 복합명사로, 말뭉치에서 다른 품사 형태소를 사이에 두지 않고 연속해서 나타나는 2개의 명사의 결합이다.

이렇게 추출된 완전 복합명사는 3개의 이상의 단일명사로 이루어진 복합명사의 구조 분석을 위한 학습데이터가 된다. 예를 들면 n_1, n_2, n_3에 대해 (n_1 n_2)가 (n_2 n_3)보다 말뭉치에서 출현 빈도가 높으면 이를 먼저 결합하여((n_1 n_2) n_3)로 분석한다. 즉, 다음 수식에서 R_{adj} 가 1보다 크면 (w_1 w_2), (w_2 w_3)에, 1보다 작으면 (w_1 w_3) (w_2 w_3)에 연관도가 주어진다.

$$R_{adj} = \frac{\Pr(w_1 \rightarrow w_2)}{\Pr(w_2 \rightarrow w_3)}$$

명사 간의 연어 관계 학습은 다음과 같은 방법으로 이루어진다.

1) 말뭉치로부터 w_1, w_2, w_3(단, $w_1 \in N, w_2 \in N, w_3 \in N_p$)를 만족하는 어절열로부터 쌍 ($w_2, n_3$)를 추출한다. 여기서 N은 조사를 취하지 않는 순수 명사이며, N_P는 조사를 취하고 있는 명사이며 n_3는 어절 w_3에 포함되어 있는 명사이다.

2) 위와 같이 추출된 (w_2 w_3)을 모아 복합 명사 데이터베이스를 구축하였다.

3.2 서술성 명사에 의한 학습

말뭉치로부터 추출한 동사와 명사간의 공기 관계 <동사, 명사, 격정보>는 구문 분석 시 명사의 미지격 해결에 유용한 정보가 된다[6]. 보어 명사와 서술 명사 간의 관계에 주어지는 복합 명사는 조사가 생략된 미지격 명사구와 동사의 관계로 볼 수 있다. 학습 데이터는 [6]에서 추출한 공기쌍으로, 약 3000만 어절의 연세 말뭉치에서 추출되었다.

이 중에서 용언화 접미사 '-하'가 붙은 경우로 서술성 명사를 한정하고, 또한 체언과의 관계가 주격과 목적격으로 나타

나는 쌍에 대해서만, '서술어'에서 접미사 '하'를 제거한 '서술성 명사'의 형태로 <서술성 명사, 품사, 체언, 품사, 조사, 빈도수>의 데이터를 공기 사전에 추가하여 학습 데이터로 이용한다.

3.3 복합명사 분석

복합명사 내의 각 명사들은 다음 [표 1]의 관계를 가진다.

$(n_1\ (n_2\ n_3))$로 분석되는 경우	(1) n_3가 서술성인 경우	n_2가 n_3의 목적어 n_1은 n_3의 주어
	(2) n_3가 서술성이 아닌 경우	$(n_1\ n_3)$의 연관도가 $(n_2\ n_3)$의 연관도보다 높다
$((n_1\ n_2)\ n_3)$로 분석되는 경우	(3) n_3가 서술성인 경우	n_2가 n_3의 목적어 n_1은 n_3의 주어
	(4) n_3가 서술성이 아닌 경우	$(n_1\ n_3)$의 연관도가 $(n_2\ n_3)$의 연관도보다 높다

[표 1] 한국어 복합명사의 분석 모델

3.4 분석 과정

3개의 명사로 이루어진 복합명사 $(n_1\ n_2\ n_3)$에 대해 서술성 명사의 공기 관계에 의해 구해진 데이터에 의해 보어와 술어 관계가 포착되면 [표 1]의 (1)과 (3)에 의해 분석된다. 보어와 술어 관계에 의해 분석이 되지 않으면, 연어 관계에 의해 분석을 시도한다. 연어 관계 데이터가 존재하면 의존 모델에 입각하여 연관도를 주고, 그렇지 않다면 연관도 선호도에 근거하여 왼쪽 우선 연관을 시킨다.

즉, 다음과 같은 세 가지 방법을 순차적으로 적용시킨다.

1. method 1 (Association of Verb-Noun; AssocVN)

먼저, 연관도를 구하고자 하는 명사-명사 쌍에 대해 서술성 명사-명사 데이터베이스로부터 주격, 혹은 목적격 데이터가 있는지 검사한다. 이들에 주어지는 연관도는 주어진 문법 관계, 즉 조사에 대해 서술명사와 일반명사가 얼마나 자주 공기하는가에 의해 측정되었다.

$$Assoc_p(n_1, n_2) = P_p(n_1, n_2)$$

(단, $p \in$ 주격조사, 목적격조사, $n_1, n_2 \in N$)

2. method 2 (Association of Noun-Noun; AssocNN)

이는 완전 복합명사, 즉 명사의 연어 관계에 의해 구축된 데이터베이스로부터 추출된 데이터로부터 연관도를 구하는 방법이다. 이의 연관도는 빈도수에 의한 확률에 의거하여 측정된다.

$$Assoc(n_1, n_2) = P(n_1, n_2)$$

3. method 3 (Associative Preference; AssocPREF)

1과 2의 방법에 의해 결정을 하지 못할 경우, 연관의 선호도에 의해 결합을 결정한다. 한국어의 경우 왼쪽 우선 연관(left association)과 오른쪽 우선 연관(right association)의 비율이 7:3 정도로 파악된다. 1과 2의 방법에 의해 결정하지 못한 명사구는 왼쪽 우선 연관에 의해 분석한다.

3.5 실험 및 성능

학습 말뭉치와 분리된 실험 말뭉치에서 추출한 274개의 복합 명사를 임의로 선정한 후, 인간의 언어 직관에 의해서도 판단 불가능한 37개를 제거, 237개의 복합 명사에 대한 실험 결과 83.5%의 성공률을 보여, 무조건 좌결합을 시키는 시스템의 69.6%나 연어 관계 파악에 의한 80.2%에 비해 높은 성능을 나타내고 있다. 술어-논항 관계로 파악되는 명사열이 연어 관계로도 많이 나타나기 때문에 술어-논항 관계 파악을 시도한 결과에 의한 성능 향상은 크게 나타나지는 않는다.

$n_1\ n_2\ n_3$	구문 구조	$n_1\ n_2$ 관계	$n_1\ n_3$ 관계	$n_2\ n_3$ 관계
무속 신앙 전통	((무속 신앙) 전통)	연어	-	왼쪽우선
경제 성장 정책	((경제 성장) 정책)	주어	-	연어
프랑스 근대 문학	(프랑스 (근대 문학))	-	연어	연어

[표 2] 복합 명사의 분석 예

[표 2]는 복합명사 구문 분석의 예이다. 이러한 복합 명사 구분 분석 결과에 의해 파악된 Head-Modifier 구조를 색인어 선정에 활용하면 정보 검색 시스템의 성능 향상을 꾀할 수 있다[5]. 위의 예에서 '프랑스 근대 문학'의 경우, '프랑스문학'과 '근대문학'의 두 가지 색인어를 추출하게 되며, '프랑스근대'와 같은 부적절한 색인어는 추출되지 않는다.

4. 복합명사 분해를 위한 형태소 분석 후처리기

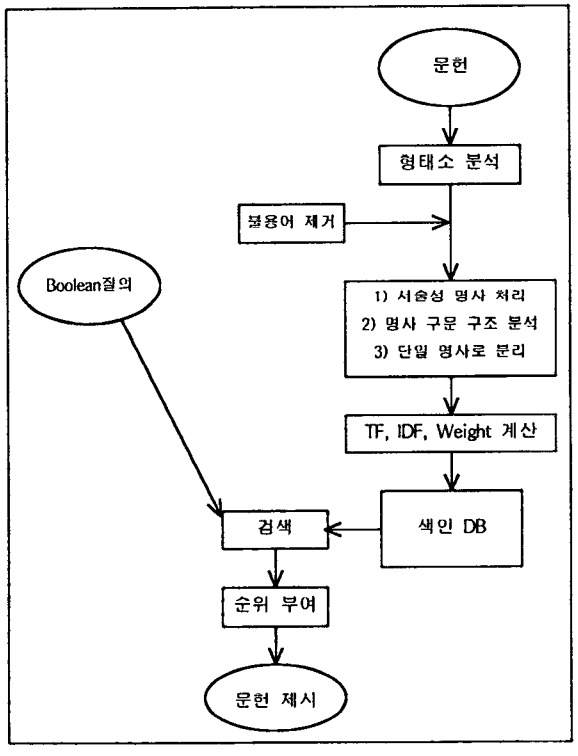
복합명사 처리는 한국어의 정보검색 시스템에 중요한 영향을 미친다. 띄어쓰기가 자유로운 복합명사를 띄어쓴 형태와 붙여쓴 형태를 같은 명사로 인식하는 문제라든가 붙여쓴 복합명사를 단일명사로 분리해 내는 문제 등이 주요한 문제이다. 복합 명사를 적절히 단일명사로 분리하여 색인하는 것은 재현률을 높이는데 기여한다. 본 연구에서 사용되는 형태소 분석기는 그 자체로 복합명사 분리 및 미등록어 추정 기능을 갖고 있다. 그러나 이러한 기능이 정보검색 분야에서 요구되는 세밀한 복합명사 분석에 미흡하여, 복합명사 분리가 이루어지지 않은 형태소 분석 결과는 복합명사 후처리를 통해 단일명사로 분리한다.

복합명사 분리 후처리기는 단일명사 사전 기반으로 처리된다. 3음절의 복합명사는 대부분 사전에 등록시켜 단일명사로서 처리를 하며, 4음절 이상의 복합명사를 구성하는 단일명사는 대부분 2,3음절의 명사이다[9]. 형태소 분석기가 4음절 이상의 명사로 출력해준 단어에 대해 최장일치 우선으로 2,3음절의 단일명사들의 조합으로 분리 가능한지를 검사한다. 만일 단일명사들의 조합으로 분리 가능하면 분리된 형태를

결과로 내 주게 되고, 그렇지 않을 경우 분리시키지 않고 단일명사로 판단한다. 복합명사 분리를 위해 복합명사의 일부가 사전 등록 단일명사이고 나머지가 그렇지 않을 경우 미등록 명사 추정을 하는 경우도 있으나, 본 연구에 사용된 형태소 분석의 결과에 대한 재차 미등록 명사 추정은 단일명사를 과도하게 분석하는 오류를 내고 있기 때문에, 완전한 사전 등록 명사에 의한 복합명사 분리만을 시도하였다.

이러한 방법으로 복합명사를 단일명사로 분리하여 색인하는 방법은 재현률을 높이는데 기여한다.

5. 시스템 구현



[그림 1] 시스템 구성도

전체적인 시스템의 구성은 [그림 1]과 같다.

검색 대상이 되는 문헌을 형태소 분석한 후, 불용어를 제거한다. 색인어로서의 가치가 없는 불용어 제거를 위해서 [11]에서 제시한 불용어 목록을 이용하였다. 형태소 분석기의 결과 중 불용어에 속하지 않은 색인어 후보들은 3가지 방법에 의하여 분석이 된 후, 색인어로서 선정이 된다. 3가지 방법은 위에서 소개한 서술성 명사로부터 복합명사의 형태로 색인어를 추출하는 방법과, 복합명사 구문 구조 분석에 의해 올바른 복합명사 형태로 색인 하는 방법, 그리고 복합명사를 단일명사로 분리하는 방법이다.

이렇게 선정된 색인어들은 통계적 방법에 의해 가중치를 부여 받아 색인 DB로 구축되는데, 여기에 사용되는 방법으로는 일반적으로 통계적 정보검색 시스템에서 사용하고 있는 TF와 IDF를 이용한 방법이다.

TF는 각 문헌별로 상대적인 길이에 따라 정규화가 필요하므로, 해당 문헌의 최대 빈도로 나누어준 후, 비례 상수 K로 TF의 적용 강도를 조절하는, Croft가 제안한 식 (1)을 사용하고, IDF는 Sparck Jones 가 제안한 식 (2)를 사용한다[2].

TF와 IDF로부터 최종적으로 정규화된 색인어의 weight가 계산되어야 p-norm 모델에 적용한 유사도 계산을 할 수 있다. weight를 정규화하는 방법은 [8]에서 제시한, 최대 가중치값으로 나누어 정규화시키는 식 (3)을 사용한다.

$$TF = K + (1-K)\frac{freq}{maxfreq} \quad \text{(Croft, 1983)} \qquad (1)$$

$$IDF = \log_2 \frac{n}{N} + 1 \quad (n\text{은 } DF,\ N\text{은 문헌의 수})$$

(Sparck Jones, 1972) (2)

$$Weight = \frac{TF * IDF}{C} \quad (C\text{는 문서 내 최대 가중치의 값}) \qquad (3)$$

이렇게 구축된 색인 DB를 이용하여, P-norm 모델에 의해 질의어와 문헌간의 유사도를 계산하여 순위대로 결과를 보여준다. 유사도 측정과 순위 부여를 위해서는 Boolean 모델을 개선시켜, 질의어와 색인어에 모두 가중치를 둘 수 있으며 유사도 순서로 순위 매김을 할 수 있는 Extended boolean 모델 중의 하나인 P-norm 모델을 사용하였다[2][3]. 이 모델은 정보 검색 모델 중 가장 성능이 우수한 것으로 일반적으로 알려져 있다. 이 모델은 다음과 같은 수식에 의해 유사도 계산을 한다.

$$SIM(Q_{or\ p}, D) = \sqrt[p]{\frac{a_1^p d_{A1}^p + a_2^p d_{A2}^p + \cdots + a_n^p d_{An}^p}{a_1^p + a_2^p + \cdots + a_n^p}}$$

$$SIM(Q_{and\ p}, D) = 1 - \sqrt[p]{\frac{a_1^p (1-d)_{A1}^p + a_2^p (1-d)_{A2}^p + \cdots + a_n^p d(1-d)_{An}^p}{a_1^p + a_2^p + \cdots + a_n^p}}$$

$$SIM(Q_{not}, D) = 1 - SIM(Q, D)$$

6. 실험 및 평가

4414개의 문헌과 50개의 질의어가 마련된 KT SET 95에서의 실험 결과는 다음과 같다. 복합명사 분리에 의한 실험은 질의어에서도 복합명사 분리를 수행하였으며, 서술명 명사 처리 및 구분 분석의 기법을 적용할 때에는 문헌에서 띄어쓰기 없는 형태의 색인어가 추출되므로 질의어의 복합명사를 띄어쓰기 없는 형태로도 주어진다.

[표 3]은 재현률에 따른 정확률의 결과를 나타낸다. 평균 정확률은 46.5%이다.

재현률	정확률
20%	58.3%
40%	49.8%
60%	44.2%
80%	33.7%

[표 3] 실험 결과

7. 결론 및 향후 연구 과제

본 논문에서는 복합 명사의 구조 분석 및 서술성 명사로부터 복합명사를 추출하는 색인 기법을 시도하고 있다. 이러한 시도는, 일반적으로 사용자의 질의 특성상 구체적인 정보를 원할 때 복합명사의 형태로 질의를 주는 경우가 많으며 이에 적합한 색인 방법은 문헌 내에서 복합명사로 파악 가능한 색인어들을 정확하면서도 빠짐없이 추출하기 위함이다. 또한 복합명사를 단일명사로 분리함으로써 재현률을 높이는 방법도 채택하여야 한다.

한국어 정보검색 시스템의 일반적인 형태는 언어학적 방법들을 사용하여 추출된 색인어에 적절한 가중치를 부여하여 통계적 방법의 검색 시스템으로 구현하는 것이다. 자연어처리 기법을 사용하여 언어학적 분석에 통해 양질의 색인을 만들어 내는 연구와 함께, 추출된 색인어들이 통계적 가중치로서만이 아니라 문헌 내의 의미적 쓰임새로서도 표현될 수 있는 지능적 정보 검색 시스템과 함께 이러한 시스템에 사용자가 쉽게 접근할 수 있는 자연어 인터페이스의 개발이 필요하겠다.

< 참고문헌 >

[1] 정영미, "정보검색론", 구미무역 출판부, 1988.

[2] William B. Frakes, Ricardo Baeza-Yates, "Information Retrieval : Data structure and Algorithm", Prentice hall, 1992.

[3] Gerard Salton, "Automatic Text Processing", Addison Wesley, 1989.

[4] Juntae Yoon, Mansuk Song, "Yet Another Compound Noun Analysis Using Co-occurrence Relation", To appear in NLPRS 97, 1997.

[5] Chengxiang Zhai, "Fast Statistical Parsing of Noun Phrases for Document Indexing", Fifth conference on Applied Natural language Processing, April 1997.

[6] 김선호, "통계 정보를 기반으로 한 어휘 관계 예측", 연세대학교 컴퓨터과학과 석사학위 논문, 1996.

[7] 김민정, 권혁철, "한국어 특성을 이용한 자동 색인 기법", 한국정보과학회 가을 학술발표논문집 Vol. 19, No. 2, 1992.

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

[9] 강승식, "한국어 형태소 분석을 위한 복합 명사의 인식 방법", 한국인지과학회 춘계학술대회 논문집, 1993.

[10] 김동일, 이신원, 윤후병, 장재우, 정성종, "한국어 텍스트에서의 복합 명사구 출현 빈도에 관한 연구", 한국정보과학회 봄 학술발표논문집 Vol. 21, No. 1, 1994.

[11] 남영준, "색인어형태분석에 의한 한국어 자동색인기법연구", 중앙대학교 문헌정보학과 박사학위 논문, 1994.