복합명사 분할과 명사구 합성을 이용한 통합 색인 기법

(Integrated Indexing Method using Compound Noun Segmentation and Noun Phrase Synthesis)

원 형 석 † 박 미 화 † 이 근 배 ††

(Hyungsuk Won) (Mihwa Park) (Geunbae Lee)

요 약 본 논문에서는 명사구 색인과 복합명사 분할을 포함한 복합명사 처리를 위해 통계 정보와 자연언어 처리를 제한적으로 이용 가능하게 하는 통합적 색인 기법을 제안한다. 먼저 색인과 검색에서 복합명사 분할 및 합성 모두를 고려한 통합 기법을 제시하고, 이를 위해 통계 정보와 제한적인 자연언어 처리를 모두 이용하는 통합 색인 기법을 제안한다. 먼저 형태소 분석 및 태깅 과정에서 단일어를 색인어로 추출하고 구문분석의 결과에서 명사구를 합성해 낸다. 구문 분석 실패 시에는 형태소 분석 및 태깅의 결과만을 사용하게 된다. 또한 태깅의 결과에서 복합명사를 골라 통계 정보를 이용하여 단일 명사로 분할하고 재합성한다. 분할된 단일 명사와 합성된 명사구는 기존의 단일어로만 이루어진 색인어를 보완하기 위해 색인어로 사용된다. 실험은 한국어 정보검색의 실험 집합인 KTSET 2.0과 KRIST SET을 사용하여 통합 색인 기법이 복합명사 처리에 효율적임을 보였다

Abstract In this paper, we propose an integrated indexing method with compound noun segmentation and noun phrase synthesis. Statistical information is used in the compound noun segmentation and natural language processing techniques are carefully utilized in the noun phrase synthesis. Firstly, we choose index terms from simple words through morphological analysis and part-of-speech tagging results. Secondly, noun phrases are automatically synthesized from the syntactic analysis results. If syntactic analysis fails, only morphological analysis and tagging results are applied. Thirdly, we select compound nouns from the tagging results and then segment and re-synthesize them using statistical information. In this way, segmented and synthesized terms are used together as index terms to supplement the single terms. We demonstrate the effectiveness of the proposed integrated indexing method for Korean compound noun processing using KTSET2.0 and KRIST SET which are a standard test collection for Korean information retrieval.

1. 서 론

한국어 정보검색에서는 성능을 높이기 위한 시도로 복합명사 처리가 연구되고 있다. 여기서 복합명사란 두 개 이상의 단일명사가 서로 결합하여 새로운 의미를 갖게 되는 단어를 말하고 구문적으로 한 단어로 다루어지고, 그렇지 않은 경우는 구(phrase)로 다룬다. 특히 한국어 문서의 경우, 문장의 내용을 주로 명사가 나타내고 있기 때문에 복합명사에 관한 연구가 색인에서 활발히 이루어지고 있다. 이것은 복합명사가 단일 명사보다 특정성(specificity)이 커서 문서를 더 잘 표현할 수 있어 색인어로서의 가치가 높기 때문이다.

이러한 복합명사를 복합명사를 구성하고 있는 단일명사로 정확하게 분할하게 되면 분할된 단일명사들은 복합명사에 비해 특정성은 떨어지게 되지만 검색에서 재

· 본 연구는 과기처 특정기초 소프트 사이언스(96-99, stepi 관리) 지원에 의한 것임.

† 비 회 원 : 포항공과대학교 학술정보원 연구원
moho@postech.ac.kr
bfpark@postech.ac.kr
†† 종신회원 : 포항공과대학교 컴퓨터공학과 교수
gblee@postech.ac.kr

논문접수 : 1999년 1월 4일
심사완료 : 1999년 11월 15일

현율을 올릴 수 있고, 문장상에 나타나는 단어들로 명사구[1]를 합성하게 되면 특정성이 큰 색인어를 만드는 것이 되어 정확도를 올릴 수 있다. 즉, 복합명사의 분할 및 합성이 모두 정보 검색 시스템의 성능 향상에 도움이 된다.

이외에도 복합명사 처리에는 위에서 예로 든 복합명사 분할이나 명사구 색인 이외에도 색인과 검색과정의 일치, 분할된 단일명사와 합성된 복합명사의 가중치 문제 등 여러 문제들이 서로 얽혀 있다.

본 논문에서는 이와 같은 복합명사 처리와 관련된 여러 문제들을 해결하기 위해 명사구 색인과 복합명사 분할을 통한 통합적 색인 모델을 제시한다. 그리고 제시한 통합적 색인 모델을 자연언어 처리와 통계 정보를 이용하는 통합적 색인 시스템을 구현하여 검증한다. 명사구 색인을 위해 SKOPE(Standard KOrean Processing Engine) 시스템[1, 2]의 형태소 분석, 품사 태깅, 구문 분석을 이용하여 색인어를 단계적으로 추출하는 색인 시스템을 개발하였다. 형태소 분석 및 태깅 과정을 거쳐 구문 분석이 이루어지면 구문 분석 결과를 이용하여 복합명사를 합성해낸다. 만약 구문 분석이 실패하면 형태소 분석 및 태깅의 결과만을 이용하여 명사류를 색인어로 추출한다. 그리고 태깅의 결과에서 추출된 명사들 중 복합명사를 통계 정보를 사용하여 단일명사로 분할하고 재합성 한다. 합성된 명사구와 분할된 단일 명사도 모두 색인어로 사용한다. 복합명사에서 분할된 단일명사와 문장에 나타난 명사들로 합성된 명사구의 가중치는 문장에서 추출된 단어들을 기반으로 재결정한다.

2장에서는 복합명사와 관련된 기존의 연구들을 살펴보고, 3장에서는 복합명사 처리를 위한 통합적 색인 모델을 제시하고, 4장에서는 복합명사 처리를 위한 통계 정보와 자연언어 처리를 이용한 통합적 색인 시스템을 제안, 설명하고, 5장에서는 실험 및 실험 결과를 검토하고, 6장에서 결론을 맺는다.

2. 기존 연구

복합명사에 대한 연구는 크게 통계를 이용하는 방법과 자연언어 처리를 이용하는 방법으로 나눌 수 있다. 통계적인 방법론이 대규모의 문서를 효율적으로 처리하기 위한 방법에 치중하는 반면, 자연언어 처리를 이용하는 방법은 문서의 보다 정교한 표현을 통한 성능 향상을 꾀하고 있다. 하지만 무제한으로 쓰여진 대규모의 문서를 다루어야 하는 정보검색에 사용될 수 있는 강건한 자연언어 처리기의 부재로 최근에는 제한적으로 자연언어 처리 기법을 이용하는 방법이 주로 연구되고 있다. 특히, 영어의 경우, 한국어와는 달리 복합어를 구성하는 성분단어가 문장상에 떨어져서 존재하므로 분할의 필요가 없다. 따라서 복합명사에 대한 국내의 연구가 복합명사 분할과 구 색인으로 나누어진 반면에 영어권의 경우 구 색인에 대한 연구만이 행해지고 있다.

2.1 영어권의 구 색인 연구

구 색인에 대한 연구 초기에 통계적 방법만으로 기존의 단어 기반 색인 방식에 비해 큰 성능 향상을 얻을 수 있었다. 그 후 자연언어 처리를 이용한 구 색인이 많이 연구되었으나, 앞서 행해졌던 통계적 방법에 비해 성능 향상을 보이지 못하였다. 최근에는 자연언어 처리를 제한적으로 색인에 이용하거나 자연언어 처리기의 사용으로 인한 과부하를 피하기 위해 간단한 질의 문장만을 처리하면 되는 검색에만 사용하는 주로 연구가 이루어졌다[3].

대표적인 구 색인에 관한 연구들을 살펴보면 다음과 같다.

통계에 기반한 연구를 살펴보면, Salton & Buckley [4]은 구 생성을 다룬 비교 연구에서, 통계적인 방법과 구문분석 결과를 이용한 방법의 실험 결과가 비슷함을 보이면서, 복잡한 자연언어 처리를 이용한 방법보다는 간단한 통계적인 방법이 선호되어야 한다고 주장했다. 하지만 논문에서 통계적 방법과 비교된 구문구조를 이용하는 방법은 구문분석 결과의 아주 제한된 패턴만을 받아 들인 뒤 가중치에 의해 여과하는 과정을 거치는 단순한 방법이었다.

Fagan[5]은 구를 색인어로 합성해 내기 위해, 단일어들의 출현 빈도를 기반으로, 단어의 출현 위치, 단일어간의 거리, 수식어로의 출현 빈도 등 여러 가지 파라미터들을 합성에 이용하였다. 구를 합성하여 색인어로 사용함으로써 다섯 개의 실험 집합에 대해 단어 기반 색인에 비해 평균적으로 10%정도의 정확도를 향상 시켰다. 하지만, 실험 집합에 따라 2.2%에서 22.7%까지 성능의 차이가 일률적이지 못했고, 합성할 때 이용되는 파라미터들이 대상 도메인에 따라 반복적으로 구해져야 한다는 문제점을 가지고 있었다. 또한 구문관계를 고려하지 않음으로써 부적절한 구가 생성되거나 적절한 구를 생성하지 못하는 경우가 많았다.

Fagan[6]은 앞서 자신이 행했던 통계적 방법에 기반한 연구의 단점을 보완하기 위해 구문분석기를 이용하

1) 복합명사와 구(phrase) 의 구별은 여러 방법이 제시되고 있으나 주관성을 배제하기 힘들다. 본 논문에서는 문장상에 존재할 경우 복합명사로 보고 그렇지 않은 경우는 보다 넓은 의미로 구라는 용어를 사용한다.

여 구문구조에서 구를 생성하여 부적절한 구 생성을 방지할 수 있었다. 하지만 너무 적은 수의 구를 생성하는데 그쳐 검색 성능에 거의 영향을 주지 못해 오히려 실험 결과는 앞서 행해졌던 통계적 방법에 의한 구 생성보다 나쁜 결과를 보였다.

최근의 연구에서 Zhai[7]는 대규모의 무제한 문장을 처리할 수 있는 강건한 자연언어 처리기를 사용하기 위해 구문분석기 대신 명사구 구문분석기를 명사구 색인에 이용하였다. 250MB의 대규모 문서 집합을 실험하여 단일어 만을 색인했을 때보다 합성한 명사구를 색인어에 추가했을 때 정확도에서 최고 18%, 재현율에서 13%의 성능 향상을 얻었다.

2.2 한국어에 대한 연구

복합명사에 대한 국내의 연구의 특징은 외국의 연구들과 달리 구 색인만이 아니라 복합명사의 분할을 다룬다는 것이다. 이것은 띄어쓰기가 자유로운 한국어의 특성을 반영한 것이다. 복합명사 분할은 여러 연구들에서 대규모의 말뭉치에서 얻은 통계 정보와 규칙들을 사용하여 95%이상의 좋은 분할 성능과 그로 인한 검색 성능 향상에 많이 기여하고 있음을 보이고 있다. 한편, 구 색인에 대한 연구는 외국의 경우와 마찬가지로 강건한 자연언어 처리기가 없는 이유로 통계에 기반하여 기초적인 자연언어 처리 방법인 형태소 분석 결과를 이용하는 방법이 주로 연구되고 있고, 구문관계를 고려한 구 색인에 대한 연구는 거의 이루어지지 않고 있다.

국내의 연구를 구 색인과 복합명사 분할에 대한 연구를 구분하여 살펴보기로 한다.

먼저 자연언어 처리를 이용하지 않는 방법으로, 남세진[8]은 자연언어 처리 방법을 사용하지 않고, 명사사전과 복합명사 구성 패턴의 통계적인 정보만을 이용하여 복합명사를 합성하였다. 그러나 복합명사 합성 시 사전에 등록되어 있는 명사에 대해서만 복합명사 합성이 이루어지므로, 사전에 없는 의미 있는 미등록 명사에 대해 자연언어 처리 방법을 통한 명사 추정을 할 수가 없어, 복합명사가 합성이 이루어지지 않았다.

다음은 자연언어 처리를 이용한 연구로, 이현아[9]는 문장을 의존 문법에 기반한 구문분석 기법을 이용하여 하나의 개념 단위인 단문으로 분할한 뒤, 5가지 패턴에 의한 명사를 합성하여 명사구 후보들을 생성한다. 명사구 후보들 중 잘못된 구문 관계를 가진 명사구를 삭제하고 연결 가능한 명사구를 통합하고 불용어 리스트를 이용하여 필터링을 하는 과정을 거친다. KTSET 276문서의 1,328 문장에 대한 구 색인어 추출에 대해서만 실험을 하였다. 대상 실험 집합이 너무 작고 색인어의 생성에 대한 실험만 하였으므로 실제 시스템의 성능 향상 여부를 보일 수 없었다.

윤보현[10]은 문장을 형태소 분석 후, 대등적 연결어미나 종속적 연결어미의 위치에서 문장을 단문으로 분할한 뒤, 패턴들에 해당하는 명사들을 명사구로 색인하였다. KTSET2.0의 2,600 문서들에 대해 검색실험을 하여 약 11~26.5%의 정확도 향상을 보였다.

김미진[11]은 형태소 분석 후 색인어 후보들 중 상위 30~40%의 고빈도어 단어의 앞 어절과 뒤 어절을 15개의 패턴에 적용하여 복합명사를 합성하고, 합성된 복합명사 중 길이가 긴 것은 색인어의 가치가 없다고 보고 분해하였다. 고빈도어 주위에만 패턴을 적용함으로써 생성되는 명사구 수는 줄였으나 사용되는 패턴 수가 다른 실험에 비해 너무 많고, 명사구 생성에 대한 실험만이 이루어져 실제 검색 성능 결과가 없었다.

윤준태[12]는 명사구내 성분명사간의 관계를 술어-보어 관계와 한정 관계로 파악하고, 이러한 두 관계를 이루는 명사간의 공기 정보를 대규모 말뭉치로부터 추출하여 명사구의 구문분석에 적용하였다. 이 방법은 명사구에 대해 83.8%의 분석 성공률을 보였으나, 실제 검색실험에 이용되지는 않았다.

복합명사 분할에 대한 연구는 활발히 이루어지고 있는데 대부분 대규모의 말뭉치에서 얻은 통계 정보를 기반으로 추가의 우선 적용 규칙들을 적용하거나 관용어 사전 등을 이용하여 분할하고 있다.

윤보현[13]은 통계 정보와 우선 적용 규칙을 사용하고, 미등록어를 포함한 복합명사는 휴리스틱을 이용하여 분할하여 약 96%의 분할 성능을 보였다.

심광섭[14]은 복합명사 분리 문제를 문장 띄어쓰기의 일부분으로 보고, 단어 레벨이 아닌 음절 레벨에서 문제를 해결하려고 시도하였다. 110만 어절의 말뭉치로부터 학습된 상호정보를 이용하여, 4,322 어절의 복합명사를 실험하여 미등록어가 없는 상태에서 분리도가 최대 98.6%까지 나타났다.

장동현[15]은 형태소 분석시에 사용되는 명사사전과 말뭉치로부터 추출한 복합명사의 분해결과를 이용하여 정방향 및 역방향 트라이 사전을 구성하고, 복합명사 분해시 정방향, 역방향 트라이를 탐색하여 정방향 일치와 역방향 일치를 행하는 과정과 문법형태소 일치 과정을 거쳐, 일치 결과로 나타나는 8가지 현상에 대하여 경험적 규칙을 이용하여 복합명사를 분해하였다.

강승식[16]은 형태소 분석결과로 추정되는 복합명사를 네 개의 분해규칙과 두 가지 예외규칙을 사용하여 가능한 분해 후보를 생성하고, 후보들에 대하여 가중치

를 부여함으로써 최적 후보를 선택하는 알고리즘을 이용해 복합명사를 분할하였다. 이 방법은 미등록 단위명사가 포함되어 있는 복합명사뿐만 아니라, 복합명사의 길이에 관계없이 동일하게 적용되었다.

하지만 국내에서 아직 복합명사 분할과 구 색인(복합명사 합성)을 통합적으로 다룬 연구는 없었으며 복합명사의 분할, 합성 및 가중치 설정의 상호관계를 체계적으로 규명할 필요가 있다. 본 논문은 이러한 노력의 첫 시도가 될 것이다.

3. 복합명사 처리를 위한 통합 색인 모델

일반적으로 색인과 검색은 같은 처리 과정을 거치는데 이것은 검색할 때 색인어와 검색어의 불일치 문제를 피하기 위해서이다. 만약 색인과 검색의 처리 과정이 다르다면 처리 효과가 달라져서 검색 성능에 차이가 나게 될 것이다. 복합명사 처리에는 복합명사 합성과 복합명사 분할이 있다. 색인과 검색에서 복합명사 처리를 하면, 아래와 같이 각각 4가지씩 처리하는 방식이 존재한다. 첫째, 문장에 나타나는 형태 그대로 색인하는 경우 즉 복합명사 처리를 전혀 하지 않는 경우(A) 둘째, 복합명사 분할을 하는 경우(B) 셋째, 복합명사 합성을 하는 경우(C) 넷째, 복합명사 분할과 복합명사 합성을 모두 하는 경우(D)가 있다. 따라서 색인과 검색을 동시에 생각하면 모두 16가지의 조합이 나온다.

(A) : 문장에 나타난 단어만 사용
(B) : (A) + 복합명사 분할
(C) : (A) + 복합명사 합성
(D) : (A) + (B) + (C)

16가지의 모든 조합을 열거하면 다음과 같다.

Case 1 : 색인(D), 검색(D)	Case 9 : 색인(B), 검색(D)
Case 2 : 색인(D), 검색(C)	Case10 : 색인(B), 검색(C)
Case 3 : 색인(D), 검색(B)	Case11 : 색인(B), 검색(B)
Case 4 : 색인(D), 검색(A)	Case12 : 색인(B), 검색(A)
Case 5 : 색인(C), 검색(D)	Case13 : 색인(A), 검색(D)
Case 6 : 색인(C), 검색(C)	Case14 : 색인(A), 검색(C)
Case 7 : 색인(C), 검색(B)	Case15 : 색인(A), 검색(B)
Case 8 : 색인(C), 검색(A)	Case16 : 색인(A), 검색(A)

이러한 16가지 경우에 대해서 색인어와 검색어의 일치 문제를 살펴보도록 한다. 이것을 살펴보는 이유는 색인과 검색 과정 각각에서 행해지는 복합명사 처리 방법들이 검색 성능에 어떠한 영향을 미치는 지 검증하기 위한 것이다. 검색과 관련되는 척도로 정보검색에서 많이 사용되는 재현율과 정확도 측면에서 위의 16가지 경우를 고려한다.

먼저 재현율만을 고려한다. 재현율이란 검색에서 관련 문서를 추출해 낸 비율이다. 따라서 같은 내용을 담고 있는 문서들에 대해 그에 관련된 질의어로 모든 문서를 검색할 수 있는지를 알아본다. 아래의 문서와 질의어는 복합명사 합성과 복합명사 분할이 필요한 경우를 포함하고 있다.

(문서3) 정보를 검색하는 시스템은
(문서4) 정보검색을 도와주는 시스템에 대한
(문서5) 정보검색시스템은

(질의어3) 정보 검색에 관한 문서
(질의어4) 정보검색에 관한 문서
(질의어5) 정보검색시스템에 관한 문서

위의 문서와 질의어를 16가지 경우에 따라 색인어와 검색어로 바꾸어 보면 [표1], [표2]와 같다. [표1]에서 색인에서 분할한 경우, 일반적인 복합명사 분할 외에 본 논문에서 사용할 분할 후 구성 단일 명사가 3이상 일 때 재합성을 한 것을 반영한 것이다. [표3]은 [표2]의 검색어로 [표1]의 색인어를 검색할 때, 16가지 경우 모두에 대한 검색 여부를 보여준다.

표 1 문서 3,4,5에 대한 색인어

	색인D	색인 C	색인 B	색인 A
문서 3	정보 검색 시스템 정보검색 정보시스템 검색시스템 정보검색시스템	정보 검색 시스템 정보검색 정보시스템 검색시스템 정보검색시스템	정보 검색 시스템	정보 검색 시스템
문서 4	정보검색 시스템 정보 검색 정보검색시스템	정보검색 시스템 정보검색시스템	정보검색 시스템 정보 검색	정보검색 시스템
문서 5	정보검색시스템 정보 검색 시스템 정보검색 검색시스템 정보시스템	정보검색시스템	정보검색시스템 정보 검색 시스템 정보검색 정보시스템 검색시스템	정보검색시스템

표 2 질의어 3,4,5에 대한 검색어

	검색 D	검색 C	검색 B	검색 A
질의어3	정보 검색 정보검색	정보 검색 정보검색	정보 검색	정보 검색
질의어4	정보검색 정보 검색	정보검색	정보검색 정보 검색	정보검색
질의어5	정보검색시스템 정보 검색 시스템 정보검색 정보시스템 검색시스템	정보검색시스템	정보검색시스템 정보 검색 시스템 정보검색 정보시스템 검색시스템	정보검색시스템

표 3 모든 경우에 대한 검색 결과

	질의어 3			질의어4			질의어5		
	문서 3	문서 4	문서 5	문서 3	문서 4	문서 5	문서 3	문서 4	문서 5
Case1	O	O	O	O	O	O	O	O	O
Case2	O	O	O	O	O	O	O	O	O
Case3	O	O	O	O	O	O	O	O	O
Case4	O	O	O	O	O	O	O	O	O
Case5	O	O	X	O	O	X	O	O	O
Case6	O	O	X	O	O	X	O	O	O
Case7	O	X	X	O	O	X	O	O	O
Case8	O	X	X	O	O	X	O	O	O
Case9	O	O	O	O	O	O	O	O	O
Case10	O	O	O	O	O	O	X	X	O
Case11	O	O	O	O	O	O	O	O	O
Case12	O	O	O	X	O	O	X	X	O
Case13	O	O	X	O	O	X	O	O	O
Case14	O	O	X	X	O	X	X	X	O
Case15	O	X	X	O	O	X	O	O	O
Case16	O	X	X	X	O	X	X	X	O

Case1,2,3,4,9,11의 경우 모든 문서를 검색해 낼 수 있었다[2]. Case11의 경우, 색인과 검색 양 쪽에서 분할만을 이용하여서 모든 문서를 검색해 냈다. 그리고 Case1,2,3,4의 경우 즉, 색인에서 복합명사 합성과 복합명사 분할을 모두 하였을 때, 질의어에 대한 처리가 어떠한 경우라도 모든 문서를 검색할 수 있었다.

다음으로 정확도 측면에서 색인과 검색 과정의 일치를 살펴본다. 정확도란 검색에서 추출한 문서 중 관련된 문서의 비율을 말한다. 비슷하지만 조금씩 다른 내용을 담고 있는 문서들을 같은 내용의 다른 형태를 가진 질의어로 검색을 해서 추출하였을 때 관련 문서가 어떤 비율을 차지하는지를 알아본다. 여기서는 앞서 재현율만을 고려한 경우 좋은 성능을 보인 Case 1,2,3,4,9,11을 대상으로 살펴본다.

(문서6) 컴퓨터 공학의 세부 연구 분야에는
(문서7) 공학과 경제 분야에서 컴퓨터의 사용은
(문서8) 컴퓨터공학에서 배우는 과목들은

(질의어6) 컴퓨터공학과 관련된 문서
(질의어7) 컴퓨터 공학과 관련된 문서

문서 6,7,8의 색인 결과와 질의어 6,7의 검색어로 사용되는 단어들은 [표4]와 [표5]에 각각 나타나 있다. [표5]에서 검색과 관련되는 단어들만을 색인어로 보이고 또한 정확도와 관련 짓기 위해 가중치를 표시하였다.

표 4 문서 6,7,8 의 색인어

	색인 D	색인 C	색인 B	색인 A
문서 6	컴퓨터(1) 공학(1) 컴퓨터공학(1)	컴퓨터(1) 공학(1) 컴퓨터공학(1)	컴퓨터(1) 공학(1)	컴퓨터(1) 공학(1)
문서 7	컴퓨터(1) 공학(1)	컴퓨터(1) 공학(1)	컴퓨터(1) 공학(1)	컴퓨터(1) 공학(1)
문서 8	컴퓨터공학(1) 컴퓨터(0.5) 공학(0.5)	컴퓨터공학(1)	컴퓨터공학(1) 컴퓨터(0.5) 공학(0.5)	컴퓨터공학(1)

표 5 질의어 6,7의 검색어

	검색 D	검색 C	검색 B	검색 A
질의어 6	컴퓨터공학(1) 컴퓨터(0.5) 공학(0.5)	컴퓨터공학(1)	컴퓨터공학(1) 컴퓨터(0.5) 공학(0.5)	컴퓨터공학(1)
질의어 7	컴퓨터(1) 공학(1) 컴퓨터공학(2)	컴퓨터(1) 공학(1) 컴퓨터공학(2)	컴퓨터(1) 공학(1)	컴퓨터(1) 공학(1)

문서 6,7,8의 내용을 보면, 문서 6이 질의의 내용에

2) 물론 검색에서 부분 일치(partial match) 기법을 사용하면 대부분의 경우에 문서를 검색해 낼 수 있다. 하지만 그런 방법은 단어의 구성 형태를 고려하지 않는 방법이어서 무분별한 검색이 행해지게 되므로 본 논문에서는 고려하지 않았다. 그리고 분할 후 재합성을 고려하지 않을 경우 6가지 경우 중 2가지(Case2,Case4)가 제외된다.

가장 부합되고 그 다음으로 문서8이 관련이 있고 문서 7은 거의 관련이 없는 문서이다. 또한 문서 6은 합성이 필요한 경우이고 문서 8은 분할이 필요한 경우이다. 그리고 질의어 6,7은 각각 분할과 합성이 필요한 동일한 내용을 담고 있다. 색인어의 가중치는 문장상에 나타나는 단어를 1로 보고 명사구는 구성 단어들의 가중치의 평균을 가중치 값으로 가지고 분할된 명사는 복합명사의 가중치를 분할된 단어의 수로 나눈 값을 가진다고 가정한다. 또한 검색에서는 질의어에 나타난 단어의 형태를 1로 보았을 때, 합성된 단어의 가중치는 구성 단어들의 합으로 하고, 분할된 명사의 가중치는 색인어와 마찬가지로 분할된 단어 수로 나눈 값을 가진다고 가정하였다. 이상의 사항을 고려한 결과가 [표6]에 나타나 있다.

표 6 질의어 6,7에 대한 검색 결과

		Case 1	Case 2	Case 3	Case 4	Case 9	Case 11
질의어 6	문서6	2	1	2	1	1	1
	문서7	1	X	1	X	1	1
	문서8	1.5	1	1.5	1	1.5	1.5
질의어 7	문서6	4	4	2	2	2	2
	문서7	2	2	2	2	2	2
	문서8	3	3	1	1	3	1

[표6]에서 보듯이, 질의어 6에 대해 Case1, 3이 좋은 검색 결과를 보이고 있고, 질의어 7에 대해서는 Case1, 2가 좋은 검색 성능을 보이고 있다. 질의어 6,7의 결과를 다 같이 고려해 볼 때, 색인에서 분할과 합성을 모두 하고 검색에서 분할과 합성을 모두 하는 Case1과 검색에서 분할만 하는 Case3이 비교적 좋은 검색 결과를 보이는데 반해 검색에서 각각 같은 방식을 사용하지만 색인에서 분할만을 한 경우인 Case9와 Case11의 검색 결과는 그보다 떨어졌다. 즉 색인에서 분할만을 하였을 때, 색인에서 분할과 합성을 모두 한 경우에 비해 재현율은 비슷하지만 정확도는 떨어진다는 것을 보여주고 있다. 특히 색인과 검색에서 분할과 합성을 모두 처리한 Case1이 가장 좋은 성능을 보이고 있다. 즉 색인에서 분할과 합성을 모두 처리하고 검색에서도 분할과 합성을 하여 가중치를 고려한 검색을 한다면 가장 좋은 성능을 얻을 수 있을 것으로 예상된다.

4. 복합명사 통합 색인 모델의 구현

4.1 구문분석 시스템을 이용한 통합적 색인 시스템

본 논문에서 복합명사 합성에 이용하는 SKOPE 시스템은 [그림1]과 같다[1]. 전체 SKOPE시스템 중 형태소 분석기, 태거, 구문 분석기를 사용한다. 그리고 복합명사 합성과 복합명사 분할을 위한 통계 정보와 SKOPE 시스템을 이용한 통합적 색인 시스템의 구성도는 [그림2]와 같다.

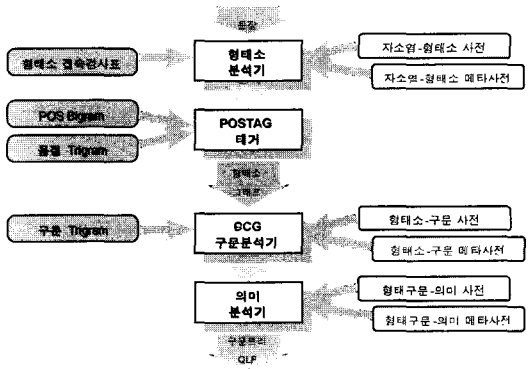

그림 1 SKOPE 시스템

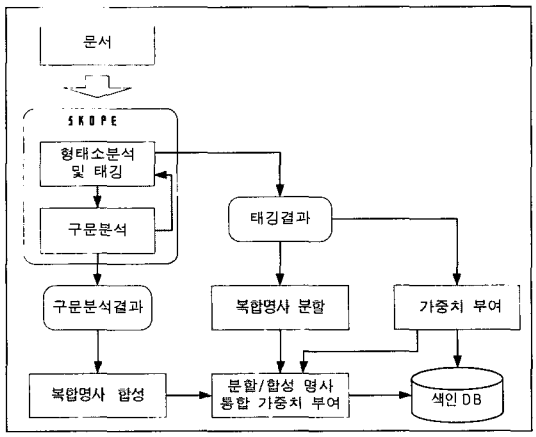

그림 2 SKOPE 시스템을 이용한 통합적 색인 시스템

4.2 명사구 색인 시스템

4.2.1 구문구조를 이용한 명사구 색인

구를 색인하는 것은 문서상에 실제로 나타나지 않은 복합명사를 만들어 내어 문서를 더 잘 표현할 수 있게 하여 정확도를 높이기 위한 방법이다.

복합명사 합성에 사용되는 SKOPE 시스템의 구문분석기는 범주문법을 사용하여 구문분석을 하며, 구문분석기의 성능은 134개의 한국어 표준문장에 대해 89%의 구문분석 정확도를 보였고, 501개의 호텔예약 말뭉치에

대해서는 94%의 정확도를 보였다[2]. 문장의 구문구조를 이용하여 복합명사를 합성하면 통계에 기반한 방법이나 패턴에 기반한 방법보다 문장상에 존재하는 다양한 명사구 표현을 찾아 복합명사를 합성할 수 있는 장점을 가진다.

예를 들어 아래와 같은 예들에서 구문구조를 이용함으로써 보다 나은 문서 표현을 위한 적합한 색인어를 합성해 낼 수 있다.

(예1) **공학**과 경제 분야에서의 **컴퓨터**의 응용
(예2) **컴퓨터 공학**과 관련된 여러 응용 분야들
(예3) **영어**와 **일본어**의 **해석**

만약 통계 정보만을 이용한다면 (예1)과 (예2)는 "공학"과 "컴퓨터"라는 단어가 동시에 나타난 문장으로 동일하게 취급되겠지만 구문구조를 이용하여 명사구 색인을 하게 되면 (예1)과 달리 (예2)만 "컴퓨터공학" 이라는 명사구를 색인 해 내어 부적절한 구 생성을 줄이고 정확한 구 생성이 가능하다. 또한 (예3)의 경우도 패턴을 이용하면 "일본어해석" 이라는 것만 생성할 수 있지만 구문 구조를 이용하면 명사들 간의 거리에 무관하게 구문구조에 의해 "영어해석" "일본어해석" 이라는 정확한 명사구를 더 만들어 낸다.

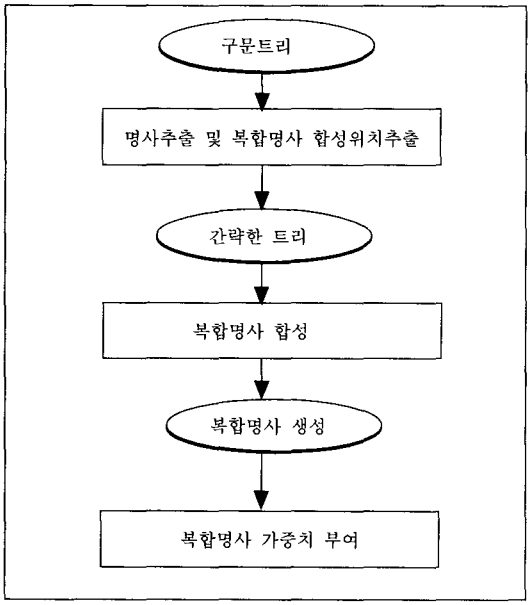

그림 3 복합명사 합성 시스템

즉, 명사 N개로 구성된 문장의 경우, 기존의 패턴을 이용하면 길이가 2인 N-1개의 명사구만을 추출할 수 있고, 통계에 의한 방법은 길이가 2인 N(N-1)/2개의 모든 조합을 생성할 가능성이 있지만 너무 많은 가치 없는 명사구들이 생성되게 된다. 하지만 구문구조를 이용하면, 구문 제약에 의해 문장 내에서 명사들간의 수식관계가 정해져 불필요한 합성을 피할 수 있고 또한 구문구조를 이용한 점진적 합성으로 명사들 간의 거리나 합성되는 명사 개수에 상관없는 정확한 명사구를 생성할 수 있다.

SKOPE 시스템의 구문 분석기를 이용한 복합명사 합성 과정은 [그림 3]과 같다.

4.2.2 복합명사 합성 과정

구문분석기의 강건성을 높이고 분석속도를 높이기 위해 먼저 문장을 구문구조에 영향을 미치지 않는 대등절로 분할한다. 문장 분할은 태깅의 결과에서 연결어미가 나오면 이루어진다. 원래 문장보다 길이가 짧은 분할된 절을 구문분석하면 분석 속도가 빨라져 구문분석기를 효율적으로 이용할 수 있게 된다.

구문분석 결과로 구문트리가 나오는데 구문트리를 후위순회하면서, 단말노드인 경우 태깅 결과를 검사하여 명사류(보통명사, 고유명사, 수사, 외국어)이면 단말노드에 그대로 두고, 복합명사 합성위치 후보(조사, 어미)에 해당되면 그 노드의 상위노드 중에서 가장 처음 만나는 명사구 범주노드에 복합명사 합성위치 후보를 표시한다. 그리고 해당노드와 그 노드의 부모노드를 삭제하면 간략화 트리가 만들어 진다.

복합명사의 합성은 간략화 트리를 후위순회하면서 단말노드이면 해당 태깅 결과를 복합명사 합성 후보로 등록하고, 단말노드가 아니면 그 노드의 태깅 결과를 바탕으로 양 자식 노드의 복합명사 합성 후보가 합성이 가능한지 패턴[표 7]을 비교하여 합성조건을 만족하면 복합명사를 합성하였고, 또한 상위 노드에서는 복합명사 합성을 위해 하위노드의 복합명사 합성 후보를 가져와 패턴과 비교하여 합성조건을 만족하면, 점진적으로 합성이 계속된다. 본 논문에서는 합성되는 복합명사의 길이[3]는 2와 3로 제한하였다. 이것은 문서 상에 나타나는 복합명사들이 실제로 구성 단일 명사가 4이상인 것은 거의 존재 하지 않는다는 사실에 기반한 것이다. 실제 실험 집합인 KTSET2.0의 경우 복합명사를 구성하고 있는 단일 명사는 3개 이하인 것이 전체 명사의 97%였고,

3) 본 논문에서 명사구, 복합명사의 길이란 복합명사 합성에 사용된 단일 명사의 개수를 말한다.

KRIST SET은 96%였다. 따라서 길이가 4이상인 명사구는 명사구가 과도하게 생성되어 시스템의 부하만을 초래할 뿐 검색 성능 향상에 전혀 도움이 되지 않으므로 생성하지 않는다.

마지막 과정으로, 복합명사 합성 후 생성된 명사구에 가중치를 부여한다.

명사구 색인의 예는 [그림4]과 같다.

표 7 복합명사 합성에 사용되는 패턴들

1. 조사가 생략된 명사/인접한 명사 (예: 형태소 해석)
2. 관형격 조사 결합 명사/ 피수식 명사 (예: 정보의 검색)
3. 목적격, 주격조사 결합 명사/서술형 명사 (예: 정보를 검색하는)
4. 관형화된 서술형 명사/피수식 내포문의 명사 (예: 정보를 처리하는 시스템)
5. 조사 상당 용언이 이끄는 관형화된 내포문의 명사/피수식 명사
   (예: 일기에 대한 예보)

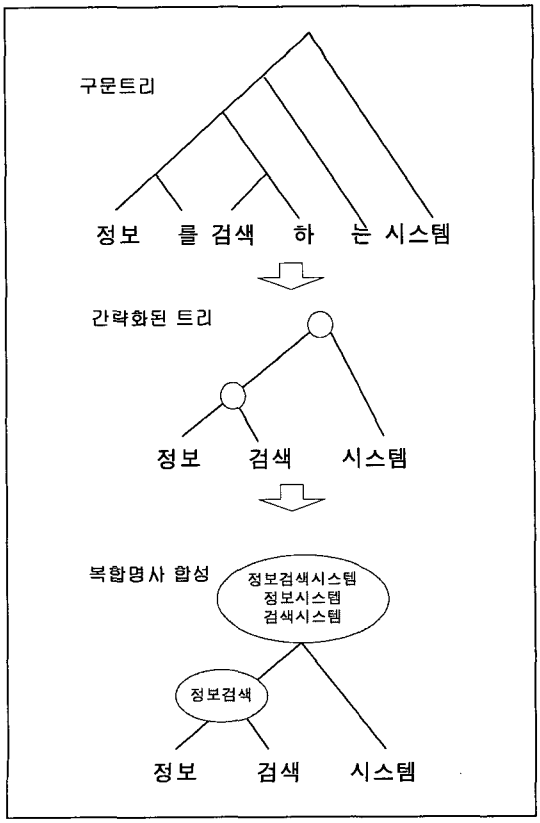

그림 4 복합명사 합성 예

4.3 통계 정보를 이용한 복합명사 분할

4.3.1 복합명사 분할

복합명사 분할은 한국어에서 띄어쓰기가 자유로워 발생하는 불일치의 문제를 해결하므로 재현율을 올릴 수 있는 방법이다.

분할은 복합명사의 구성 패턴, 상호 정보, 관용어 사전을 이용한다. 분할할 형태 결정에 필요한 복합명사의 구성 패턴은 채영숙[17]의 분석 결과에서 얻었고, 관용어 사전은 1,000개의 단어로 이루어져 있으며, 전체의 의미와 상관없는 성분 단어로 나누어지는 복합명사들(예: 데이터베이스, 멀티미디어)을 관용어 사전에 등록하였다. 분할 과정은 [그림 5]과 같다.

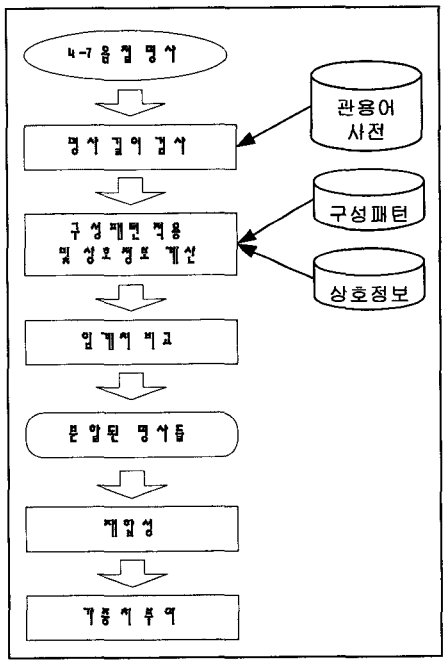

그림 5 복합명사 분할 과정

4.3.2 복합명사 분할 과정

형태소 분석 및 태깅을 거쳐 나온 결과에서 명사류를 추출하여 그 중 음절 수가 4-7인 명사류를 대상으로 복합명사 분할이 시도된다. 음절이 4-7인 명사에 대해서 분할을 시도하는 이유는 한국어에서 4음절 이상의 단일어는 극히 드물기 때문에 4음절 이상의 명사라면 복합어일 확률이 아주 높기 때문이다. 또한 8음절 이상의 명사는 실제로 거의 존재하지 않는다[4].

4) KTSET2.0의 명사들을 분석한 결과 8음절 이상의 명사는 전체

일반적인 상호 정보의 계산식은 아래 식(1) 이지만, 본 논문에서는 데이터 부족 문제를 해결하기 위해 제안된 평탄화 기법 중의 한가지인 식(2)를 사용하였다[18].

$$MI(x,y) \approx \log_2 \frac{N \cdot f(x,y)}{f(x) \cdot f(y)} \quad (1)$$

$$MI(x,y) \approx \log_2 \frac{N \cdot f(x,y)+1}{f(x) \cdot f(y)+|V|} \quad (2)$$

$f(x)$: 단어 x의 빈도수
$f(y)$: 단어 y의 빈도수
$f(x,y)$: 단어 x,y의 공기 빈도수
N : 말뭉치 크기
V : 말뭉치에 사용된 어휘수

분할은 다음의 과정을 거친다. 일단 단어가 분할 대상으로 들어오면 입력된 단어가 관용어 사전에 들어있는지를 검사한다. 들어 있지 않은 경우, 입력단어의 음절 수를 검사한 뒤 음절에 따른 구성 패턴을 적용한다. 4음절 단어의 경우 2+2 형태로 상호정보를 계산하여 임계치를 넘을 경우 분할한다. 5음절 단어의 경우에는 2+3, 3+2형태로, 6음절 단어의 경우는 4+2, 3+3, 2+4 의 형태로 상호정보를 계산하여 임계치를 넘는 가장 큰 값을 갖는 형태로 분할이 된다. 다시 4음절은 2+2로 분할이 시도된다. 이렇게 분할된 명사들은 다시 재합성 된다[그림6]. 분할 및 재합성된 단일 명사와 명사구는 가중치 부여 과정을 거친다.

이런 방법으로 KTSET2.0의 단어들의 10%를 대상으로 실험한 결과 95%의 분할 정확도 성능을 보였다.

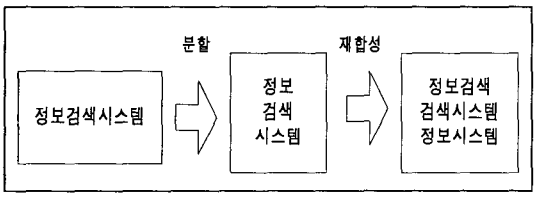

그림 6 분할과 재합성

4.4 통합 가중치 부여

앞서의 과정들이 문서를 표현하는 적절한 색인어를 찾는 과정이었다면 가중치 부여는 선택된 색인어들의 중요한 정도를 결정짓는 과정이다. "정보검색에서 색인어가 검색되는 문서의 범위를 결정하고 색인어의 가중치가 검색된 문서들의 순위를 결정한다[19]."는 것을 고려하면 적절한 색인어의 추출 방법만큼이나 가중치 부여도 중요하게 다루어져야 높은 검색 성능을 기대할 수 있다.

일반적인 정보 검색 시스템에서는 색인어의 가중치 부여시 색인어 빈도와 문서 빈도를 이용하는 통계적 가중치 부여 방법을 이용하지만, 합성된 복합명사 및 분할된 성분명사의 가중치 부여는 실제 문서 집합에 존재하지 않는 새로운 단어를 문서 집합에 추가 후 가중치를 부여하는 과정이므로 기존의 가중치 부여 방법과 구별이 되어야 한다.

합성된 복합명사의 가중치 부여는 2가지 경우를 생각해 볼 수 있는데, 하나는 합성된 복합명사가 이미 문서 집합에 있는 경우이고, 다른 경우는 합성된 복합명사가 문서 집합에 새롭게 나타난 경우이다. 전자의 경우에는 이미 문서상에 존재하는 색인어와 같이 통계적 방법을 사용하여 가중치를 부여하면 되고, 후자의 경우에는 2가지 방법을 생각해 볼 수 있는데, 하나는 기존의 단어와 동일하게 통계적인 방법을 사용하는 것이고, 다른 하나는 복합명사를 구성하고 있는 성분명사의 가중치를 이용하는 방법이다.

위와 같은 방법을 고려하여, 본 실험에서는 색인과 검색에서 합성명사에 대한 적합한 가중치 부여 방법을 결정하기 위해 다음과 같은 3가지 방법을 사용했다.

> 방법 1: 기존 단어와 같이 통계적인 방법을 사용
> 방법 2: 구성 성분 명사의 평균을 사용
> 방법 3: 구성 성분 명사의 합을 사용

색인과 검색에서 각각 3개의 방법을 조합하여, 검색 성능 실험을 한 결과 색인에서는 통계적인 방법, 그리고 검색에서는 구성 성분 명사의 합을 사용하는 조합이 검색 성능이 가장 좋아 검색, 색인 통합 실험에 사용하였다.

그리고 복합명사에서 분할된 단일명사의 가중치 부여 방법은 복합명사가 가지는 가중치를 복합명사를 구성하고 있는 구성 단어의 개수로 나눈 값을 가중치로 사용하였다. 즉 아래의 식과 같이 복합명사 c를 구성하고 있는 단일명사 a의 가중치는 복합명사 c의 가중치를 구성 단일 명사 수로 나눈 값이다.

$$W_a = \frac{W_c}{|C|} \times |\alpha|$$

$|\alpha|$: x를 구성하고 있는 단일 명사의 수

지금까지 자연언어 처리 시스템인 SKOPE 시스템의 형태소 분석기, 태거, 구문 분석기와 통계 정보를 이용

명사 중 2.8%에 불과했다.

한 통합적 색인 방법을 통해 복합명사 합성, 복합명사 분할, 가중치 부여를 통합적으로 다루는 색인 시스템을 설계하고 구현하였다.

5. 실 험

본 논문에서 제안한 복합명사 합성과 복합명사 분할 처리를 통합한 SKOPE 시스템을 이용한 통합적 색인 모델의 타당성과 시스템의 성능을 보인다. 실험은 한국어 정보검색 실험 집합인 KTSET2.0[20]의 4,414건의 문서와 50개의 자연언어 질의어와 KRIST SET[21]의 13,515건의 문서와 25개의 자연언어 질의어를 사용하였다. 실험에 사용된 검색 시스템은 자연언어 질의를 불리언 질의로 변환하여 검색하는 시스템이다[22].

복합명사 처리에 있어 앞서 3장에서 고려한 모든 경우에 대한 재현율은 아래 [표8]과 같다.

표 8 16가지 경우에 대한 재현율

	KTSET2.0	KRIST
CASE 1	89.29	87.15
CASE 2	86.85	78.56
CASE 3	88.88	85.90
CASE 4	86.22	77.76
CASE 5	86.14	82.06
CASE 6	85.03	76.64
CASE 7	86.46	80.45
CASE 8	84.07	75.04
CASE 9	89.29	87.15
CASE 10	82.77	77.72
CASE 11	88.88	85.90
CASE 12	82.14	76.92
CASE 13	86.14	82.06
CASE 14	80.42	75.80
CASE 15	86.46	80.45
CASE 16	79.46	74.20

앞의 [표3]에서 예상했던 대로 KTSET2.0에서는 색인에서 분할을 했을 경우가 다른 방법에 비해 재현율의 향상 폭이 컸고 아래의 여섯 가지 경우에서 다른 경우들에 비해 좋은 재현율을 나타내었다.

(1) 색인에서 분할과 합성을 하고 검색에서도 분할과 합성을 하는 경우 (Case1)
(2) 색인에서 분할과 합성을 하고 검색에서 합성을 하는 경우 (Case2)
(3) 색인에서 분할과 합성을 하고 검색에서 분할을 하는 경우 (Case3)
(4) 색인에서 분할과 합성을 하고 검색에서 복합명사 처리를 하지 않는 경우 (Case4)
(5) 색인에서 분할을 하고 검색에서 분할과 합성을 하는 경우 (Case9)
(6) 색인에서 분할을 하고 검색에서 분할을 하는 경우 (Case11)

하지만 KRIST SET의 경우, 재현율이 좋을 것으로 예상되었던 Case2, Case4의 경우의 재현율이 낮게 나타났다. 이는 KRIST SET이 KTSET보다 복합명사를 붙여 쓰는 경향이 강해 불일치의 문제가 KTSET보다 더 크다고 볼 수 있으므로, 검색에서 불일치 문제를 해결할 수 있는 분할의 역할이 중요하다고 볼 수 있다.

좋은 재현율을 보인 경우들에 대한 KTSET2.0과 KRIST SET의 11point에 대한 정확도는 [표9] [표10]과 같다.

표 9 KTSET2.0의 6가지 경우에 대한 11point 정확도

	PRECISION						
RECALL	Baseline	CASE 1	CASE 2	CASE 3	CASE 4	CASE 9	CASE 11
0	74.00	64.00	62.00	66.00	68.00	64.00	70.00
10	77.99	74.40	73.55	77.03	77.41	71.45	77.94
20	68.74	72.97	70.10	75.18	74.38	66.07	71.05
30	66.05	71.38	68.58	70.94	70.59	65.92	68.14
40	56.32	64.70	63.65	64.63	65.19	60.69	62.16
50	54.10	58.86	58.85	58.60	59.93	57.53	56.07
60	40.36	49.79	48.43	49.51	47.58	49.27	50.19
70	32.90	41.34	41.59	41.30	40.96	42.52	42.41
80	27.47	37.13	34.77	36.70	33.76	36.90	36.06
90	10.20	18.38	14.15	17.10	14.47	18.03	16.15
100	5.69	8.18	6.38	7.04	5.67	8.19	7.05
Average	46.71	51.01	49.28	51.27	50.72	49.14	50.84
%change		9.21	5.49	9.77	8.58	5.20	8.84

표 10 KRIST SET의 4가지 경우에 대한 11 point 정확도

	PRECISION				
RECALL	Baseline	CASE 1	CASE 3	CASE 9	CASE 11
0	40.00	56.00	52.00	56.00	52.00
10	50.90	62.65	58.82	61.02	60.42
20	42.34	58.24	54.24	51.07	49.83
30	35.83	50.80	50.49	48.26	46.90
40	35.23	47.96	47.66	46.35	45.83
50	35.08	43.71	43.74	43.97	42.91
60	23.71	32.25	32.26	32.63	32.58
70	21.14	29.61	29.61	29.33	29.19
80	20.29	28.10	28.24	26.65	26.40
90	14.16	21.37	21.30	19.54	19.53
100	11.24	14.50	14.43	13.83	13.89
Average	29.99	40.47	39.34	38.97	38.13
%change		34.93	31.18	29.93	27.14

[표9] [표10]에서 Baseline은 [표8]의 Case16 즉, 색인과 검색 과정 모두에서 복합명사 처리를 하지 않은 경우이다. [표9] [표10]에서 보듯 모두 Baseline에 비해 정확도가 향상되었으며 특히 Case1과 Case3에서 높은

성능 향상을 보였다. 즉 색인에서 합성과 분할을 다 했을 경우 같은 검색 방식을 취하지만 색인에서 분할만을 한 Case9와 Case11에 비해 높은 정확도를 보였다. 그리고 KTSET2.0의 실험 결과인 [표9]에서, 가장 좋은 성능을 보일 것으로 기대했던 Case1보다 Case3이 더 좋은 검색 성능을 보인 것은 KTSET의 질의어는 복합명사가 주로 붙어서 있으므로, 특정성을 증가 시킬 수 있는 합성의 역할이 적고, 오히려 불필요한 합성명사를 만들 수 있어, 정확도를 감소시킬 수 있는 문서집합의 특성에 기인한 것이다.

실험결과를 분할과 합성 각각의 효과 측면에서 살펴보면, 분할을 통해 불일치 문제를 해결한 경우(Case 11) Baseline과 비교하여, KTSET에서 8.84%, KRIST SET에서 27.14%의 정확도 향상을 보여, 한국어 정보검색에서 분할이 필수적임을 알 수 있고, 불일치 문제를 해결 후, 정확도 향상을 위한 복합명사 합성 시에는 (Case 1) 각각의 SET에서 불일치 문제 해결 시(Case 11)보다 0.37%, 9.79%의 정확도 향상을 보였다. 그러나 합성 시 색인어의 수가 분할 만 했을 때 보다 21.87% 증가하고, 검색 시간도 7.35% 증가했기 때문에 효율성 측면에서 본다면 합성의 효과가 적어졌다.

6. 결 론

본 논문에서는 한국어 정보검색에서 성능 향상을 위해 반드시 해결되어야 하는 복합명사 처리 문제를 효율적으로 처리하기 위한 분할합성가중치의 통합 모델과 그를 위한 통합적 색인 시스템을 제안하였다. 통합 모델 제시로 높은 재현율과 정확도를 얻기 위해서는 색인에서 분할과 합성이 도움이 됨을 보였고 통계 정보와 SKOPE시스템의 구문분석을 이용한 통합적 색인 시스템으로 실험하여 제시한 통합 모델이 복합명사 처리에 있어 효과적임을 보였다.

향후, 효율성 측면에서 색인 시스템의 명사구 색인 과정에서 과다 생성되는 구를 억제하기 위한 방법이 마련되어야 하고, 또한 정확도 향상을 위한 다양한 통합 가중치 부여에 대한 연구가 필요하다.

참 고 문 헌

[1] 이원일, "단일화 기반 범주 문법에 기반한 음성 한국어 처리", 포항공대 박사학위 논문, 1998.

[2] Jeongwon Cha, Wonil Lee, Geunbae Lee and Jong-Hyeok Lee, "Morpho-Syntactic Modeling of Korean with K-CCG," Proceedings of the 18th ICCPOL, pp. 67-74, 1999.

[3] Smith, M.E., "Aspects of the P-Norm model of Information Retrieval: Syntactical query generation, Efficiency and Theoretical properties," Ph.D. Thesis, CS, Cornell Univ., 1990.

[4] Gerard Salton, Chris buckley, "A comparison between statistically and syntactically generated term phrases," Tr89-1027, CS department, Cornell Univ., 1989.

[5] Joel L. Fagan, "The effectiveness of a non-syntactic approach to automatic phrase indexing for document retrieval," JASIS, Vol.40, No.2, pp115-132, 1989.

[6] Joel L. Fagan, "Experiments in automatic phrase indexing for document retrieval: a comparison of syntactic and non-syntactic methods," Ph.D. thesis, Cornell University, 1987.

[7] Chengxiang Zhai, "Fast statistical parsing of noun phrases for document indexing," Fifth conference on applied natural language processing, pp.312-319, 1997.

[8] 남세진, 이지연, 신동욱, 채미옥, "복합명사의 통계적 처리에 대한 평가", 제8회 한글 및 한국어 정보처리 학술발표논문집, pp. 36-41,1996.

[9] 이현아, "구문분석과 공기 정보를 이용한 개념 기반 명사구 색인 방법", 포항공대 전산과 석사 학위 논문, 1996.

[10] 윤보현, 김상범, 임해창, "한국어 정보검색에서 구문적 용어 불일치 완화방안", 제 10회 한글 및 한국어 정보 처리 학술발표 논문집, pp.143-149, 1998.

[11] 김미진, 박미성, 장혁창, 이상조, 최재혁, "고빈도어를 이용한 복합명사 색인어 추출 방안", 제 10회 한글 및 한국어 정보 처리 학술 발표 논문집, pp.121-129, 1998.

[12] 윤준태, 정의석, 송만석, "명사간 어휘 정보를 이용한 한국어 복합 명사 분석", 정보과학회논문지(B), 제 25권, 제 11호, 1998.

[13] 윤보현, 조민정, 임해창, "통계 정보와 선호 규칙을 이용한 한국어 복합 명사의 분해", 정보과학회논문지(B), 제 24권, 제 8호, 1997.

[14] 심광섭, "합성된 상호 정보를 이용한 복합명사 분리", 정보과학회 논문지(B), 제 24권, 제 11호, pp.1307-1317, 1997.

[15] 장동현, 맹성현, "효율적인 색인어 추출을 위한 복합명사 분석 방법", 제8회 한글 및 한국어 정보처리 학술발표논문집, pp.32-35, 1996.

[16] 강승식, "한국어 복합명사 분해 알고리즘", 정보과학회논문지(B), 제 25권, 제 1호, pp172-182, 1998.

[17] 채영숙, 권혁철, "말뭉치로부터 추출된 통계정보를 활용한 한국어 복합명사 분석", 인지과학회 논문지, 제8권, 제2호, pp.101-108, 1997.

[18] Stanley Chen and Joshua Goodman, "An empirical study of smoothing techniques for language modeling," Proceedings of the 34th Annual meeting of the Association for Computational Linguistics, pp.310-318, 1996.

[19] 최대선, "구 색인에서 성분 단어의 가중치 부여 방법에 관한 연구", 포항공대 석사학위 논문, 1997.

[20] 김재군, 김영환, 김성혁, "한국어 정보검색연구를 위한 시험용 데이터 모음(KTSET) 개발", 제6회 한글 및 한국어 정보처리 학술 발표 논문집, pp. 378-385, 1994.
[21] 이준호, 최광남, 한현숙, 김종원, 남성원, " 정보검색 연구를 위한 KRIST 테스트 컬렉션의 개발", 정보관리학회지, 제 12권, 제 2호, pp. 225-232, 1995.
[22] 박미화, 원형석, 이원일, 이근배, "구문분석에 기반한 자연언어 질의로부터의 불리언 질의 생성", 제10회 한글 및 한국어 정보처리 학술 발표 논문집, pp73-80, 1998.

원 형 석

1997년 경북대학교 컴퓨터공학과 졸업. 1999년 포항공대 대학원 전산과 졸업. 1999년 ~ 현재 포항공대 학술정보원 연구원. 관심분야는 정보검색, 자연어처리

박 미 화

1989년 동아대학교 컴퓨터공학과 졸업. 1989년 ~ 1994년 포스데이타 근무. 1999년 포항공대 정보통신대학원 졸업. 1999년 ~ 현재 포항공대 학술정보원 연구원. 관심분야는 정보검색, 자연어처리

이 근 배

1984년 서울대학교 컴퓨터공학과 졸업. 1986년 서울대학교 컴퓨터공학과 석사. 1991년 미국 UCLA 전자계산학과 박사. 1984년 ~ 1986년 서울대학교 연구조교. 1987년 ~ 1991년 UCLA 전자계산과와 생명과학학과에서 연구조교와 연구원으로 근무. 1991년 ~ 현재 포항공대 부교수. 관심분야는 자연어처리, 인공지능/신경망, 음성인식, 정보검색등