고빈도어를 이용한 복합명사 색인어 추출 방안

o
김미진*, 박미성*, 장혁창*, 최재혁**, 이상조*
*: 경북대학교 컴퓨터공학과, **: 신라대학교 컴퓨터교육과

The Generation Methods of Composition Noun For Efficient Index Term Extraction

*: Department of Computer Engineering, Kyungpook National University
**: Dept. of Computer Education, Silla University
mjkim@comeng.ce.kyungpook.ac.kr

요약

정보검색이나 자동색인 시스템에서는 정확한 색인어의 추출이 시스템의 성능을 좌우하게 된다. 따라서 정확한 색인어의 추출이 매우 중요하다.

본 논문에서는 정보 검색시에 보다 정확한 문서를 찾아줄 수 있도록, 출현 고빈도어를 이용하여 효율적인 색인어 추출을 위한 합성 명사 생성 방안을 제시한다. 이를 위하여 문서 내에서 출현빈도가 높은 명사, 즉 상위 30%~40%의 고빈도 명사에 합성 및 분해 규칙을 적용하여 합성명사 색인어를 추출한다. 또한 본 논문에서 제시한 상위 30%~40% 고빈도 명사합성에 대한 타당성을 검증하기 위하여 적절한 명사합성 빈도를 구한다. 제안한 방법을 적용한 결과 300어절 이하의 짧은 문서는 출현빈도 상위 30%까지의 명사를 합성했을 경우 저빈도 누락이 작았고 300어절 이상의 문서는 출현빈도 40%까지 합성하면 저빈도 누락이 상당히 줄어듦을 알 수 있었다. 그리하여 전체 색인어의 개수를 줄였고 색인어의 정확률을 높였다.

1 서론

정보검색이나 기계번역과 같은 자연언어 처리 응용 시스템의 성능 향상을 위해 여러 각도에서 다양한 연구가 진행되고 있다. 사용자가 원하는 색인어가 포함된 문서를 검색해 주는 정보 검색 시스템에서나 문서에 나타나는 주요 어휘를 자동으로 색인하는 자동색인 시스템의 경우 색인어의 선정 방법에 대한 연구가 중점적으로 수행되어 왔다. 이는 시스템의 성능을 평가하는 방법 중 하나가 추출된 색인어의 정확도에 의해서 평가되기 때문이다. 즉, 추출된 색인어가 가지는 중요도에 의해 사용자가 찾고자 원하는 문서가 후보로 제시된 문서에 포함될 확률이 좌우되기 때문이다. 한국어의 경우 추출된 색인어의 대부분이 명사나 명사구에 포함되어 있으므로 이들의 비중이 크다. 자립형태소인 명사는 형식 형태소나 파생 접사와 결합하여 단어를 만들 뿐 아니라 명사끼리도 상호 결합하여 새로운 단어인 복합명사를 형성하기도 한다. 그러므로 색인어의 가치는 단일어보다는 복합명사가 더 크고 이에 대한 분석이 중요한 부분을 차지한다[1].

특히 정보검색에서 복합명사를 적절하게 처리하게 되면 시스템의 검색 효율을 향상시킬 수 있는데 이것은 문서 내에서 명사가 차지하는 개념적 중요도가 다른 품사보다 크며 대부분 색인어

로서 사용되기 때문이다[2]. 기존의 한국어 정보 검색 시스템은 대부분 형태소 분석기술을 바탕으로 문서에 있는 단일 명사나 일부 복합명사를 추출하여 검색에 이용한다. 특히 단일 명사에 비해 어휘특정성(specificity)이 큰 복합 명사를 색인어로 이용하면 검색의 정확도가 높아진다. 복합명사는 연속되어 나열된 명사를 연결하거나 간단한 구문 패턴이나 공기(Co-occurence) 정보를 이용하여 생성한다[1,2,3,4,5,6].

복합명사 색인어 추출에 관련된 기존 연구로는 김민정[1], 김판구[3], 정영미[6] 등의 연구가 있다. 논문 [1]에서는 복합어를 추출하기 위해 형태소적 언어 정보뿐만 아니라 간단한 구문 분석을 수행한 후 복합명사 형성조건에 따라 복합어를 추출하는 기법을 제안하고 있다. 이 연구에서는 결과치를 제시하고 있지 않다. 논문 [3]에서는 부분 구문분석만을 수행하여 한국어 문서에 대한 복합어 구성 조건 및 복합어 분해 기법을 제시하였다. 그리고 논문 [6]에서는 격조사의 형태를 설정하여 특정어구의 형태를 만족시키는 체언을 분리해서 후보색인어로 선정하는 방법을 택하여 색인어를 추출하고 있다. 그러나 이상의 연구들에서 [1]은 5개의 휴리스틱한 규칙만을 제시했고, [3]은 복합어 구성시에 병렬명사구를 인식하지 못하고, [6]은 색인어 후보수가 너무 많이 생성되는 문제가 있다.

따라서 본 논문에서는 문서 내의 상위 30%~40%의 출현 고빈도 명사에 본 논문에서 제안한 명사 합성규칙과 분해규칙을 적용하여 색인어를 생성하는, 출현 고빈도어를 이용한 합성명사 색인어 추출 방안을 제시한다. 이 방법을 적용한 결과 전체 색인어의 개수를 줄이고 정보 검색시에 보다 정확한 문서를 찾아줄 수 있었다.

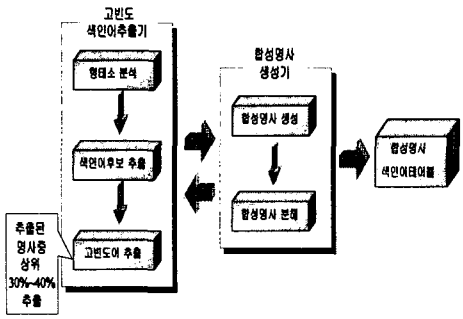
본 논문의 구성은 다음과 같다. 제 2장에서는 효율적으로 색인어를 추출하기 위한 합성명사 생성방안의 전체구성에 대해 설명하였다. 제 3장에서는 상위 30%~40%의 출현 고빈도 명사에 제안된 합성규칙을 적용하여 정확한 색인어 추출 실험과 결과를 분석 평가하였고 제 4장에서는 결론 및 향후 연구방향을 기술하였다.

2 고빈도어를 이용한 합성명사 생성기

효율적인 색인어 추출을 위하여 상위 30%~40%의 출현 고빈도 명사에 제안된 합성규칙과 분해규칙을 적용하여 합성명사를 생성하였다.

2.1 전체 구성도

고빈도어를 이용한 합성명사 생성의 전체 구성도는 [그림 1]과 같다.



[그림 1] 합성명사 생성의 전체 구성도

고빈도어를 이용한 효율적인 합성명사를 생성하기 위하여 자동 키워드 추출기는 먼저 색인 대상이 되는 원문서를 형태소 분석한 후, 색인어 후보들을 추출한다. 여기서 추출된 색인어 후보 중 불용어들은 불용어 사전을 통하여 제거된다. 다음으로 본 논문에서 제안한 적절한 합성빈도에 따라 상위 30%, 40%의 출현 고빈도어를 추출하여, 복합명사 합성 규칙에 따라 합성을 수행한다. 그리고 이미 구성된 복합어, 즉 합성규칙을 적용하지 않고 생성된 복합명사 색인어와 합성규칙 적용으로 생성된 긴 합성명사 색인어를 색인어 후보로 추출해 주기 위하여 복합어를 여러 개념으로 분해한다. 그리고 단일명사 색인어를 위해서는 합성결과에 고빈도어를 추가한다.

2.2 고빈도 색인어 추출기

고빈도 색인어 추출기는 원문서를 형태소 분석한 후 색인어 후보 추출 단계에서 후보를 추출하

여 그 중 상위 30%~40%의 고빈도어를 추출하는 고빈도어 추출 단계를 거친다.

2.2.1 형태소 분석

색인어 추출을 위해서는 형태소 분석을 통한 방법이 가장 일반적으로 행해진다. 색인어 추출 시스템은 문장 내에서 명사만 추출하면 되므로 일반적인 형태소 분석 방법을 그대로 사용할 필요가 없이, 명사 추출을 위한 간단한 형태소 분석만을 행하면 된다[7].

2.2.2 색인어 후보 추출

입력된 문서로 부터 색인어 후보를 추출하기 위해서는 자동 키워드 추출기를 통해 간단한 형태소 분석을 수행한 후 색인어 후보가 될 수 있는 모든 단일 명사과 복합명사들을 후보로 추출해 낸다.

2.2.3 고빈도어 추출

기존의 색인어 추출 시스템은 문서 내에 출현한 모든 단일 명사를 합성 명사 구성조건에 의해 복합 명사 색인어로 추출하므로 너무 많은 색인어가 생성되어 검색 시스템의 정확률을 저하시키는 원인이 된다. 그러므로 본 논문에서는 자동 키워드 추출기를 통해 추출된 색인어 후보들 중에서 출현빈도가 높은 상위 30%~40%의 고빈도어를 선택하여 명사합성을 수행하였다. 출현 고빈도 30%~40% 명사 합성에 대한 타당성을 제시하기 위하여 출현빈도 30%, 40%, 50%에 대해 각 문서당 저빈도 누락과 어절당 저빈도 누락을 조사하였다. 먼저 빈도별 합성명사를 합성규칙에 의해 생성하고 합성명사의 중요도를 판단하기 위하여 수작업 색인어와 비교를 행하였다. 이 비교에서 일치되지 않는 합성명사 색인어 개수를 전체 수작업 색인어 개수로 나눈 값을 저빈도 누락이라 하고 이 저빈도 누락이 출현빈도가 몇 %일 때 가장 작은지를 조사하였다. 수작업 색인어는 국어교육과 학부생 4학년 2명, 1학년 2명, 총 4명이 추출했다. 수작업 색인어의 추출 개수는 전체 어절수 * 5%로 했다. 조사한 결과 300어절 이하의 짧은 문서는 출현빈도 상위 30%까지의 명사를 합성했을 경우 저빈도 누락이 작았고 300어절 이상의 문서는 출현빈도 40%까지 합성하면 저빈도 누락이 상당히 줄어듦을 알 수 있었다. 그리고 출현빈도 상위 30%~40%의 고빈도어를 합성하여 명사를 생성했을 때, 수작업 색인어로 추출한 중요한 색인어를 빠짐없이 추출하려면 상위 몇 %까지의 고빈도어를 합성해야 하는지, 그 적절한 합성빈도도 실험을 통하여 알아보았다. 그 결과 300어절 이하의 짧은 문서는 출현빈도 상위 30%가 적절한 합성빈도임을 알 수 있었고, 300어절 이상의 문서는 출현빈도 40%가 적절한 합성빈도임을 알 수 있었다. 그러므로 본 논문에서는 추출된 색인어 후보들 중 출현빈도 상위 30%~40%의 고빈도어만을 합성규칙과 분해규칙에 따라 합성명사로 생성시킨다.

2.3 합성명사 생성기

합성명사 생성기는 제안한 규칙에 따라 고빈도 명사를 합성하는 합성명사 생성 단계와 전 단계에서 생성된 긴 합성명사를 분해하는 합성명사 분해 단계를 거친다.

2.3.1 합성명사 생성

단일어는 광의의 단어가 색인됨으로써 검색시스템의 정확률을 떨어뜨리고 사용자의 탐색시간을 늘린다. 그러나 복합어는 어휘특정성이 높아 시스템의 정확률을 높여주고 유용한 정보를 포함한 문서들만을 검색하게 한다. 그러므로 색인 어휘의 특정성을 만족하는 색인어휘로서 복합어를 추출하기 위해서는 여러 가지 형태의 구문적 표현을 하나의 복합어로 표현할 수 있는 합성규칙이 필요하다. 본 논문에서는 추출된 상위 30%~40%의 고빈도 명사를 대상으로 기존의 합성규칙[1][3]과 본 시스템에서 제시한 새로운 합성규칙을 사용하여 합성명사를 생성한다.

명사 합성 규칙

색인 어휘의 특정성을 만족하는 색인어휘로서

복합어를 추출하기 위한 합성규칙은 다음과 같다.

1. N-(에)의/에서의/으로(서)의 N --> NN
2. **N-을/를 N-하/시키** --> NN
3. **(N-을/를) N-한/하는/할 N** --> (N)NN
4. **N-을/를 N-받아(고)** --> NN
5. **N-에 N-되/하/시키** --> NN
6. N-에 관한/대한 N --> NN
7. N-적인/적 N --> N-적 N
8. **N-중인/중 N --> N-중 N**
9. **N-에/와 관련된 N --> N 관련 N**
10. N-이/가 N-하/되 --> NN
11. **N-적으로 N-하/되/시키--> N-적 N**
12. N N N
13. N-(으)로 된/만든/인한 N -->NN
14. N_1-**및** N_2 --> N_1-**및** N_2,
    --> N_1
    --> N_2
15. N_1-**와/과** N_2 --> N_1-**와/과** N_2
    --> N_1
    --> N_2

기존 연구에서 제시하고 있는 규칙들 이외에, 기존의 규칙에서는 적용되지 않는 복합명사구를 수용하기 위해 9가지의 새로운 규칙들을 추가하였다. 본 논문에서 새롭게 추가한 규칙들은 다음과 같다.

규칙1) N-중인/중 N --> N-중 N

규칙2) N-을/를N-한/하는/할/하기위한 N
예) 양을 복제하기 위한 작업
-->양복제작업

규칙3) N-을/를 N-시키
예) 위기감을 고조시키
-->위기감고조

규칙4) N-을/를 N-받아/받고
예) 사회보장 진단서를 발급받아
-->사회보장진단서발급

규칙5) N-에 N-되/하/시키
예) 작업에 착수했
--> 작업착수

규칙6) N-적으로 N-되/하/시키
--> N-적 N
예)유전공학적으로 조작함으로써
--> 유전공학적조작

규칙7) N-에(게)/와/과 관련된 N
--> N 관련 N
예) 북한과 관련된 사건
--> 북한관련사건

규칙8) N_1 -및- N_2 --> N_1 및 N_2
--> N_1,
--> N_2
예) 연구의 내용 및 범위
--> 연구내용및범위
--> 연구내용
--> 연구범위

규칙9) N_1-와/과 -N_2 --> N_1와/과 N_2
--> N_1
--> N_2
예)형식언어의 문법과 인식기를 학습
--> 형식언어문법과인식기학습
--> 형식언어문법학습
--> 형식언어인식기학습

기존의 연구에서는 "연구의 내용 및 범위"라는 명사구에서 "연구내용"만을 합성해주고 "및" 이하는 중요한 명사이어도 추출해 주지 못했다. 그러나 규칙8과 규칙9에서 병렬명사구 "와/과", "및"에 대한 처리를 함으로써 위의 문장에서 합성해 내지 못하는 복합명사, 즉 "연구내용및범위", "연구범위"를 합성해 낼 수가 있다.

명사 합성 방법

입력 문서에서 추출된 30%, 40% 고빈도어 명사를 명사 합성규칙에 따라 합성하여 합성 복합명사를 생성한다. 그 자세한 방법은 아래와 같다.

① 고빈도 단어의 앞 어절과 뒷 어절이 명사 합성 규칙에 해당되는지를 확인하여 명사를 합성시킨다.

예를 들면 "조지 워싱턴대의 생태학자팀은 인간의 태아를 복제했다"라는 문장에서 "태아"가 고빈도어인 경우 "태아를"의 앞 어절인 "인간의"와 "태아를"은 "N-의 N"이라는 규칙에 해당하므로, "인간태아"라는 합성명사가 생성된다. 그리고 "태아를"의 그 뒷 어절과의 관계에서도 "태아를 복제했다"는 "N-를 N-하다"라는 규칙에 해당하기 때문에 "태아복제"라는 합성명사도 만들어진다. 그리고 이 세 어절의 명사가 모두 합성된 "인간태아복제"라는 합성명사도 만들어져서 결국 "인간태아", "태아복제", "인간태아복제"라는 세 개의

합성명사가 만들어진다.

② 고빈도어가 문서내에서 복합명사의 일부분인 경우는 그 복합명사를 출력해 주고 위와 같은 방법으로 그 복합명사의 앞 뒤 어절과의 관계에서 규칙이 적용되는지를 살펴 합성시킨다.

예를 들면 "복제인간의 탄생은 꿈인가 악몽인가"라는 문장에서 "복제"가 고빈도어인 경우에 "복제"는 "복제인간"이라는 복합명사의 일부분이다. 이 때에는 "복제인간" 자체를 하나의 복합명사로 출력해 주고, 그 뒤 어절의 관계에서 "복제인간의 탄생"이 "N-의 N"규칙에 해당하므로 "복제인간 탄생"이라는 합성명사가 더 만들어지게 된다.

③ 고빈도어가 문서내에서 "와/과/및"으로 연결되는 병렬명사구를 이룰때 연결된 명사구를 합성규칙에 따라 합성명사로 생성해주고, 그 다음 병렬명사구를 여러 개의 문장으로 나누어서 각각의 문장에 대해 합성규칙을 적용하여 합성 복합명사를 생성한다.

예를 들어 "연구의 내용 및 범위"라는 문장은 "연구내용 및 범위"를 출력해주고, 또 "연구의 내용"과 "연구의 범위"라는 두 개의 문장으로 나누어 생각해서 "연구내용"과 "연구범위"를 생성해준다. 그리고 "형식언어의 문법과 인식기를 학습하고"라는 문장에서 "형식언어"가 고빈도이면 "형식언어문법", "형식언어 문법 학습", "형식언어 인식기 학습"과 "형식언어 문법과 인식기 학습"이 생성된다.

2.3.2 합성명사 분해

추출된 색인어 후보들 중 고빈도 명사들에 합성규칙을 적용하여 합성명사를 생성하기 때문에 합성규칙을 적용하지 않고 생성된 색인어를 찾아주지 못하는 경우가 있다. 그리고 합성규칙 적용으로 생성된 긴 합성명사의 경우에도 색인어로 사용되기 어려우므로 이러한 경우에 합성명사를 분해하여 단일명사 색인어를 만들게 되는데 분리의 기준을 두 가지로 주었다. 자세한 방법은 다음과 같다.

① 조사를 기준으로 분리한다.

문서에서 합성규칙으로 받아들인 규칙들 중 "N-의 N", "N-을/를 N-한/하"라는 규칙에서 "N-의 N"은 조사 "의"를 기준으로 분해되고, "N-을/를 N-한/하" 규칙에서는 조사 "을/를"을 기준으로 분리를 수행한다

예를 들면 "돌리 보고서를 발표한"은 합성규칙 "N-을/를 N-한/하" 에 의해 "돌리보고서발표"가 합성된다. 그러나 긴 합성명사의 경우 색인어로 사용되기 어려우므로 분해가 필요하게 된다. 여기에서는 조사 "를"에서 분리하여 "돌리 보고서"와 "발표"가 색인어가 된다. 그리고 "외국단체의 복제기술 남용이 불가피하므로"는 조사 "의"와 "이"에서 분리가 이루어져 "외국단체"와 "복제기술 남용"과 "불가피"로 분해가 이루어진다.

② 고빈도어가 문서내에서 합성명사의 일부분인 경우는 그 고빈도어를 중심으로 좌측인접명사 앞에서 분해를 하고, 또 우측인접 명사 뒤에서 분해를 해준다.

예를 들면 "**복제** 양 돌리를"은 "복제"가 고빈도어이므로 "복제양"과 "돌리"로 분해된다. 그리고 "전이**유전자** 동물"은 "유전자"가 고빈도어이므로 "전이유전자"와 "동물"로 분해된다.

위의 두 기준을 다 포함한 예인 "형태소 **분석기** 구현이 간편할"을 분해해보면 먼저 조사 기준 분리가 적용되어 "형태소 분석기 구현"과 "간편"으로 분해되고, 두번째로 고빈도어 좌우 인접명사 기준 분리가 적용되어 "형태소 **분석기** 구현"은 분석기가 고빈도어이므로 좌인접 명사와 결합된 "형태소 분석기"와 우인접 명사와 결합된 "분석기 구현"으로 분해된다.

3 실험 및 결과분석

본 논문은 IBM PC상에서 C Language로 구현하였다. 본 논문에서는 효율적인 색인어 추출을 위하여 추출된 색인어 후보들 중에서 출현빈도 상위 30%~40%의 고빈도어를 선택하여 명사합성을 수행하였다. 이 30%~40%의 고빈도어 추출에 대한 타당성을 검증하기 위하여 먼저 문서당 출현빈도에 따른 저빈도 누락과 어절당 출현빈도에 따른 저빈도 누락을 구하고 적절한 명사 합성 빈도도 구하였다.

첫 실험대상 문서는 심마니에서 임의의

keyword로 추출한 6개의 문서를 선택하여 수작업으로 수행하였다.

표 1은 각 문서당 출현 빈도에 따른 저빈도 누락 상태를 보여준다.

표 1. 저빈도 누락과 출현빈도

문서 / 출현빈도	저빈도 누락					
	text 001 (2169 어절)	text 002 (147 어절)	text 003 (533 어절)	text 004 (923 어절)	text 005 (229 어절)	text 006 (273 어절)
30%	21%	23%	26.7%	16.7%	11%	6.7%
40%	15%	23%	14.4%	12%	8.3%	3.3%
50%	15%	23%	14.4%	10%	5%	3.3%

표 1에서 출현빈도 30%는 가장 높은 고빈도로부터, 저빈도로 가면서 <출현 단어빈도/전체 명사갯수>를 계산하여 누적한 값이 30%에 가까운 값을 나타낸다. 예를 들면 아래 그림에서 **북한**이 단어 빈도 7로 가장 높은 고빈도이고 출현빈도는 (7/86=)8.1%, 그 다음 출현 고빈도는 **주일미군**으로서 6.9% ...이렇게 계산된 빈도들을 더하여 30% 내외 까지를 구하면 다음 [그림 2]와 같다.

```
1. 대표   =      1   -->    0.011628
.
.
51. 철수   =     5   -->    0.058140
52. 주일미군 =   6   -->    0.069767
53. 북한   =     7   -->    0.081395

      전체 단어의 갯수는 86 개이다.
----------30%----------
1   북한          -->    0.081395
2   주일미군      -->    0.069767
3   철수          -->    0.058140
4   차석대사      -->    0.046512
5   의제          -->    0.034884
6   회담          -->    0.034884
-------------------------------
    퍼센트        -->    0.325581
```

[그림 2] 명사의 출현빈도 계산

위의 표에서 저빈도 누락은 출현 고빈도 명사를 합성규칙에 따라 합성하여 수작업 색인어와 비교했을 때 일치하지 않는 색인어로서, 저빈도 명사이기 때문에 합성되지 못하고 색인어에서 누락되는 것을 말한다. text001의 경우 이 문서는 2169어절로 출현빈도 상위 30%까지의 고빈도 명사를 합성규칙에 따라 합성했을 경우, 저빈도 누락이 전체 수작업 색인어중 21%에 해당되고, 상위 40%까지의 저빈도 누락은 15%, 50%일 경우에도 15%임을 보여주고 있다. text001은 출현빈도 30%까지를 합성하는 것보다 40%까지를 합성하게 될 경우 6%정도 저빈도 누락이 줄어듦을 알 수 있다. 그리고 어절 수가 273어절인 text006의 경우는 출현빈도 상위 30%까지의 고빈도 명사를 합성규칙에 따라 합성했을 경우, 저빈도 누락이 전체 수작업 색인어중 6.7%에 해당되고, 상위 40%, 50%의 저빈도 누락은 3.3%임을 보여주고 있다. 이 문서의 경우는 출현빈도 30%까지만 합성하여도 저빈도 누락이 매우 작다. 이 실험의 결과로 어절 수에 따라 출현빈도에 따른 저빈도 누락이 달라짐을 알 수 있다. 300어절 이하의 짧은 문서는 출현빈도 상위 30%까지의 명사를 합성했을 경우 저빈도 누락이 작았고 300어절 이상의 문서는 출현빈도 40%까지 합성하면 저빈도 누락이 상당히 줄어듦을 알 수 있다.

표 2에서는 표 1의 결과를 좀더 확실히 하기 위하여 어절 수에 따른 저빈도 누락을 조사했는데 이 실험은 22만 어절의 문서 100개로 수행하였다.

표 2. 어절당 저빈도 누락

어절 / 출현빈도	저빈도 누락			
	100어절 이하	300어절 이하	600어절 이하	600어절 이상
30%	20%	14.7%	24%	31%
40%	17%	10.5%	14%	18%
50%	17%	10%	14%	15%

위의 표에서 보면 어절수에 따라, 출현빈도에 따른 저빈도 누락이 달라짐을 알 수 있다. 표 1에서의 결과와 마찬가지로 300어절 이하의 짧은 문서는 출현빈도 상위 30%까지의 명사를 합성했을

경우 저빈도 누락이 작았고, 300어절 이상의 문서는 출현빈도 40%까지 합성하면 저빈도 누락이 작음을 보여주고 있다. 또한 각 문서의 출현빈도 상위 40%와 50%의 저빈도 누락이 비슷한 결과를 보임도 알 수 있다. 표에서 알 수 있듯이 출현 고빈도 30%~40%의 명사를 합성했을 때 거의 모든 문서에서 수작업 색인어와 일치하는 색인어를 추출할 수 있음을 알 수 있다.

다음 실험은 출현 고빈도어를 이용하여 명사를 합성함으로써 색인어 수가 다른 시스템의 색인어 수보다 월등히 적음을 보여주고 있다. 기존의 색인어 추출 시스템은 문서 내에 출현한 모든 단일 명사를 합성 명사 구성조건에 의해 복합 명사 색인어로 추출하는데 비하여 본 논문은 적절한 명사 합성 빈도에 따라 합성된 명사수에 고빈도 명사를 합한 수가 색인어로 생성된다. 따라서 색인어 수의 비교를 위하여 기존의 색인어를 총 색인어 후보수로 보고 본 논문의 색인어 수는 합성명사수로 나타내어 비교하였다.

표 3. 색인어후보수와 합성명사수 비교

문서	총어절수	총 색인어 후보수	총 색인어 후보수 (중복제외)	고빈도	합성명사수 (고빈도 명사포함)
text1	423	252	172	30%	67
				40%	112
				50%	112
text2	99	78	52	30%	30
				40%	30
				50%	30
text3	1047	656	220	30%	130
				40%	157
				50%	182
:					
text 100	149	103	51	30%	22
				40%	24
				50%	34

표 3에서 총어절 수가 423어절인 text1의 경우, 총 색인어 후보수가 252개이고, 중복을 제외한 후보수가 172개이므로 기존 시스템에서는 그 이상에 상당하는 색인어를 추출하지만 본 논문에서는 출현 고빈도 상위 30%까지의 합성명사 수로 67개의 색인어가 추출되었고 40%, 50%까지는 각각 112개가 색인어로 추출되었다. 본 논문의 출현 고빈도 상위 30%까지의 합성명사 수는 기존 시스템의 색인어 수의 39%정도에 해당되고 40%, 50%까지의 합성명사 수는 약 65%에 해당된다. 이 비교를 통하여, 본 논문에서 제안한 방법으로 추출된 색인어 수가 기존 시스템보다 월등히 적음을 알 수 있다.

표 4는 추출된 색인어 후보들 중 출현빈도 상위 30%~40%를 합성규칙에 의해 합성하게 된 근거로 적절한 명사 합성빈도를 구하여 그 타당함을 보여주고 있다.

표 4. 적절한 명사합성 빈도

	어절	저빈도 누락	TI 누락	FI 누락	출현 빈도	OP
text001	2169	21%	4%	10%	30% (6)	**40% (5)**
text002	147	23%	0%	10%	28% (4)	**15%**
text003	533	26.7%	4.4%	10%	27% (3)	**42% (2)**
text004	923	16.7%	4%	8%	28% (4)	**38% (3)**
text005	229	11%	6.7%	0%	30% (3)	**30%**
text006	273	6.7	0	0%	30% (3)	**30% (3)**
text007	99	25%	0%	0%	30% (2)	**30% (2)**
text008	276	44%	0%	26%	30% (2)	**30% (2)**
text009	80	25%	0%	0%	25% (3)	**7% (6)**
text010	99	20%	20%	0%	30% (3)	**30% (3)**
text011	272	8.5%	0%	0%	30% (3)	**30% (3)**
text012	423	36%	5%	25%	28% (3)	**41% (2)**

각 어절의 길이가 다른 문서 12개에서 적절한 명사 합성빈도를 측정하였다. 여기에서 TI는 네명이 수작업으로 추출한 색인어 중 세명이 공통으로 추출한 색인어를 나타내며, FI는 네명이 수작업으로 추출한 색인어 중 네명 다 공통으로 추출한 색인어를 나타낸다. 그리고 출현빈도의 값인 30%(6)에서 30%는 30%까지의 고빈도를 나타내며 (6)은 출현빈도 수를 말한다. OP는 적절한 합성빈도로서 각 문서의 저빈도 누락을 계산한 뒤에 어느 정도 빈도를 낮추면 또는 높이면 TI, FI 누락이 가장 적은지 OP를 측정하였다. 그 결과를 보면 300어절 이하의 문서는 출현빈도 상위 30%까지의 명사를 합성했을 경우가, 그리고 300어절 이상의 문서는 40%정도가 적절한 합성빈도임을 알 수 있다. 그리하여 본 논문에서는 적절한 명사

합성 빈도인, 상위 30%~40% 고빈도 명사를 합성하여 색인어로 추출하는 방안을 제시하게 되었다.

다음의 표 5에서는 본 논문에서 제시한 방안으로 추출된 색인어가, 색인어휘로 사용될 가능성을 평가하기 위하여 색인어 적합률(relevance)과 부적합률(irrelevance)을 구했다. 일반적으로 자동색인에 대한 평가는 색인어 재현율과 색인어 정확률을 가지고 평가한다[6]. 그러나 색인어 재현율을 구할 때 적합한 색인어를 수동 추출하는 사항은 수동 색인자에 따라 값이 크게 유동적이고 정확한 값을 구하기 어렵기 때문에, 색인어 재현율과 비슷한 색인어 적합률을 구하여 평가할 수 있다[8,9]. 실험을 위하여 심마니에서 임의의 keyword로 추출한 24개의 문서를 선택한 뒤, 표 5와 같이 4개의 부류로 나누어 실험하였다. 적합률과 부적합률은 아래의 식에 의해 구한다[9].

$$\text{적합률} = \frac{\text{수작업색인어} \cap \text{합성색인어}}{\text{수작업색인어}}$$

$$\text{부적합률} = \frac{|\text{합성색인어} - (\text{수작업색인어} \cap \text{합성색인어})|}{\text{수작업색인어}}$$

표 5. 본 논문의 색인어 적합률

	수작업 색인어	합성된 색인어	일치 색인어	적합률	부 적합률
Text(6)	648	676	568	87.6%	16.7%
Text(6)	636	631	572	90%	9.3%
Text(6)	684	689	585	85.5%	16%
Text(6)	696	721	553	79.5%	24.1%

표 5는 본 논문의 색인어 적합률과 부적합률을 보여주고 있다. 결과로 얻은 색인어 적합률은 약 85.65%이고 부적합률은 약 16.5%이다. 타 시스템 [8]의 적합률과 부적합률은 약 80%, 50%이고 시스템[3]은 약 90%와 15%이므로 시스템 [8]과 비교해보면 상당히 높은 적합률과 낮은 부적합률을 가진다. 그러나 시스템 [3]에서는 띄어쓰기 오류로 인한 문장 분석 실패를 없애기 위해 실험데이타의 모든 문장을 한글 맞춤법에 맞게 고치고, 수치데이타는 색인어로 추출하지 않는 등 많은 제약을 가한 후의 실험결과이므로 아무런 제약을 가하지 않은 본 논문의 적합률은 시스템 [3]과 비교할 만 함을 알 수 있다.

4 결론 및 연구방향

본 논문에서는 정보 검색시에 보다 정확한 문서를 찾아줄 수 있도록, 출현 고빈도어를 이용하여 효율적인 색인어 추출을 위한 합성 명사 생성 방안을 제시하였다. 이를 위하여 문서 내에서 출현 빈도가 높은 상위 30%~40% 명사에 명사 합성규칙과 분해규칙을 적용하여 합성명사 색인어를 추출하였다. 또한 적절한 명사 합성 빈도값을 구하여 상위 30%~40% 고빈도 명사합성에 대한 타당성을 증명하였다. 본 논문에서 제시한 합성명사 생성 방안을 적용해 본 결과 300어절 이하의 짧은 문서는 출현빈도 상위 30%까지의 명사를 합성했을 경우 저빈도 누락이 작았고, 300어절 이상의 문서는 출현빈도 40%까지 합성하면 저빈도 누락이 상당히 줄어듦을 알 수 있었다. 그리하여 전체 색인어의 개수를 줄였고 보다 정확한 색인어 추출이 가능해졌다.

향후 연구과제로는 각 어절당 optimal 합성 빈도를 찾는 실험이 이루어져야 하며, 큰 문서를 일정 크기로 분리시켜 색인어를 추출하면 시스템 수행시에 memory와 속도의 효율이 높아질 수 있으므로 *optimal cutting rate*에 대한 연구가 수행되어야 한다.

참고문헌

[1] 김민정, "한글 특성을 고려한 자동 색인기법", 부산대학교 석사학위 논문, 1993.

[2] 신동욱, "복합명사의 통계적 처리에 대한 평가", 한글 및 한국어 정보 처리 학술발표논문집, pp36-41, 1996.

[3] 김판구, 조유근 "상호 정보 기반한 한국어 태스트의 복합어 자동 색인", 한국정보 과학회 논문지, 1994.

[4] 이현아, 이종혁, 이근배 "구문분석과 공기정보를 이용한 개념 기반 명사구 색인방법", 한글 및 한국어 정보 처리 학술발표 논문집, 1995.

[5] 박영찬, 최기선 "통계적 명사 패턴 분류를 이

용한 복합 명사 검색 모델", 한글 및 한국어 정보처리 학술 발표 논문집, pp21-31, 1996.
[6] 정영미, 정보검색론, 정음사, 1988.
[7] 최재혁, "형태소 분석을 통한 한·영 자동 색인어 추출 시스템 ", 정보과학회논문지(B) 제 23권 제 12호, 1996.
[8] Keysun Choi, Young S. Han, "Syntactic Analysis Based Automatic Indexing for Korean Texts", Proceedings of The Korea-US Bilateral Workshop on Computers, Artificial Intelligence and Cognitive Science, pp199-206, Aug. 1991.
[9] 김판구, 조유근 "한국어 정보 검색을 위한 불용어의 구성 및 적용", 한국 정보과학회 '93 봄 학술발표논문집 제20권1호, pp809-812, 1993.