DOI : 10.5392/JKCA.2011.11.3.467

구문 다양성 해소를 위한 복합명사구 색인 방법

A Method Of Compound Noun Phrase Indexing for Resolving Syntactic Diversity

조민희, 정도헌
한국과학기술정보연구원

Min-Hee Cho(mini@kisti.re.kr), Do-Heon Jeong(heon@kisti.re.kr)

요약

복합명사구는 단일어보다 명확한 의미를 갖기 때문에 의미적 정보처리에서 중요한 요소로 사용된다. 하지만 명사구의 표현형태의 다양성 때문에 같은 의미를 갖고 있다 할지라도 그 동일성을 판단하기 어렵다. 이에 본 연구에서는 이러한 구문 다양성 해소를 위해 복합명사구 색인 방법을 제안한다. 본 연구의 최종 목적은 다양한 형태로 표현된 동일한 의미의 명사구를 동일한 형태의 색인어로 표현하는 것이며, 이를 위해 다음과 같은 과정을 따른다. 먼저 복합명사구 인식을 위한 규칙 템플릿을 생성하고, 국내학술논문 집합에 적용하여 복합명사구들을 추출한다. 일반적으로 복합명사구는 특정성이 크다. 이에 이를 고려한 색인어 합성규칙을 제안하고, 추출된 명사구에 적용한다. 본 연구의 성능을 객관적으로 평가하기 위해 HANTEC 2.0 테스트셋을 이용하였으며, 그 결과를 기준모델과 비교하였다. 실험과 비교를 통해 본 논문에서 제안하는 색인방법이 검색 정확률 향상에 긍정적으로 영향을 미치며, 정보검색의 성능을 향상시킬 수 있음을 확인하였다.

■ 중심어 : | 복합명사구 | 구문 다양성 | 정보검색 |

Abstract

Compound noun phrase (CNP) is important factor for semantic information process because the meaning of the CNP is more disambiguous than that of single word. However, the CNP can be expressed in various types even though it expresses same meaning. It is called syntactic diversity. It makes information system difficult to grasp sense identity. In order to resolve the syntactic diversity in this research, we propose an indexing method for compound noun phrase. The main purpose is to make identical index term for various types of CNPs which has same meaning. To do so, the research follows next steps. For the first, we make rule template and utilize the template to extract CNPs from set of domestic research papers. In general, the CNP has a unique meaning. Considering the characteristic, we suggest synthesis rules of index terms and apply the rule to CNPs extracted in previous step. For the objective performance evaluation of the research, a test set, HANTEC 2.0, was utilized and the result was compared to baseline model. Through the experiment and the evaluation, we have confirmed that the indexing method suggested in this paper could positively affect retrieval precision and improve performance of the information retrieval.

■ keyword : | Compound Noun Phrase | Syntactic Diversity | CNP | Information Retrieval |

접수번호 : #110128-004
접수일자 : 2011년 01월 28일
심사완료일 : 2011년 03월 16일
교신저자 : 정도헌, e-mail : heon@kisti.re.kr

I. 서 론

인터넷의 발달과 더불어 매일 대용량의 데이터가 쏟아지면서, 정확하고 의미있는 정보를 빠르게 찾기 위한 연구에 대한 중요성이 커지고 있다. 정확한 데이터를 검색하기 위하여 단일어인 키워드를 사용하는 검색어에 부가적인 정보를 제공하는 연구는 다양한 방법으로 진행되고 있지만[1][2], 문서를 대표하는 색인어와 질의어간의 구문적 용어 불일치로 검색 성능을 저하하는 문제를 해결하기 위한 연구는 부족한 실정이다[3]. 이러한 구문적 용어불일치는 띄어쓰기가 자유로운 복합명사와 다양한 형태의 명사구로 인하여 발생한다.

KISTI에서 서비스하고 있는 학술정보 검색 사이트 NDSL(National Discovery for Science Leaders)[1]의 국내 학술 논문 서비스의 경우, 단일어절의 명사 기반 색인시스템으로 구성되어 있다. 대부분의 사용자들은 단일어를 사용하여 검색하고 있으므로[4], 단일어로 표현된 질의어가 문서에서 여러 단어로 구성되어 나타날 때 일치할 수 있어야 한다.

표 1. NDSL 검색 결과 건수(Boolean 검색 모델)

질의어	검색결과건수
스마트그리드기술	6
스마트그리드 기술	40
스마트 그리드 기술	40
스마트그리드의 기술	5
스마트그리드 관련기술	3

예를 들어 [표 1]은 FAST 검색엔진을 사용하는 NDSL의 학술논문 검색 결과 예이다. 질의어는 띄어쓰기, 조사 및 수식어를 제거하면 모두 동일한 의미의 단어이지만 표현 형태가 다르므로 색인어가 다르게 구성되어 검색결과 건수가 모두 다르게 나타난다. 이러한 질의어는 구문적으로 복합명사 혹은 명사구(noun phrase)로 문서에서 발생할 수 있으므로 이들을 동일한 것으로 인식하는 과정은 필요하다. 따라서 데이터의 양이 많아지면서 검색의 정확률이 중요해지고 있기 때문에 기존의 단일어절의 명사 기반 색인 시스템을 고도화하기 위한 인접 어절의 구조적인 관계 및 의미관계까지 고려하는 색인어를 추출하는 방법이 요구된다.

문서를 대표하는 색인어의 대부분은 복합명사, 인접한 두개 이상의 명사, 명사구 등에서 많이 발생한다[5]. 이러한 복합명사 혹은 명사구는 일반적으로 단일어보다 식별력이 커서 문서를 구체적으로 표현할 수 있으므로 색인어로 생성되면 검색의 정확률 개선에 도움이 된다[3]. 본 연구에서 제안하는 색인기술은 문서를 대표하는 복합명사 혹은 복합명사구를 추출해서 이형태로 표현된 명사들을 정규화된 표현으로 변환하고, 동일한 형태의 색인어를 구성하도록 하는 것이다. 또한 4단계 합성규칙을 이용한 색인어 재생성 등의 과정을 통하여 문서에서 발생하는 복합명사구의 불규칙한 형태를 통합하고자 한다. 이러한 방법은 문서에서 자주 등장하는 용어들의 다양한 다른 형태적 표현을 동일한 형태의 색인어들로 일반화함으로써, 색인어와 사용자가 입력한 질의어의 형태상 불일치 문제를 최소화하여 검색 성능을 향상시키고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 명사구 및 복합명사 색인과 관련된 기존의 연구들을 살펴보고, 3장에서는 본 논문에서 제안하는 명사구 인식 및 합성규칙에 의한 복합명사 재생성 방법을 기술한다. 4장에서는 실험을 통하여 본 논문의 연구가 정보검색에 끼치는 영향을 평가한다. 마지막으로 5장에서 결론을 맺는다.

II. 관련 연구

본 논문에서 제안하는 색인기술에 대한 연구는 기존의 복합명사 및 명사구단위 색인 연구와 관련이 있다.

영어권의 경우에는 구를 검색에 활용하는 것이 상당히 일반적이고, TREC[2](Text Retrieval Conference)에 참여한 많은 시스템들이 구 추출방법을 색인에 반영하고 있다. 한국어와는 달리 구를 구성하는 단어들이 모두 떨어져 있으므로 복합명사 분해 문제를 고려하지 않아도 되는 이점이 있다[6][7]. Zhai(1997)는 대규모의 비정형 문서를 처리하는데 완전한 파서(full parser)대신에 명사구 파서(noun phrase parer)를 이용하여 검색

1 http://www.ndsl.kr

2 http://trec.nist.gov/

성능을 개선하는데 상당한 효과를 보여줬다[8].

한글 문서에 대한 명사구와 관련된 색인 문제는 복합명사 분해 문제와 함께 여러 방법으로 연구되어 왔다. 두 단어 이상으로 구성된 명사구 처리에 대한 문제는 실험적으로 연구는 많이 되었지만 기반기술로 많이 이용되는 구문 분석기의 품질 저하 문제 및 속도, 복잡성 등의 이유로 서비스에 활용된 예가 적다.

한국어는 특성상 복합명사가 띄어쓰기 및 조사 생략 등의 형태적 표현방법에 따라 명사구가 될 수 있다. 명사구에 대한 색인은 구문분석을 수행하는 방법, 의존관계를 이용하는 방법, 언어적 휴리스틱 통계정보를 이용하는 방법 등이 있다.

강승식(1997)은 복합명사를 형태소 단위로 분해하는 문제는 단일명사가 어떤 복합명사의 분해로부터 발생하였는지를 구별 할 수 없으므로 검색의 재현율은 향상되지만, 정확률은 떨어진다고 했다[9]. 윤보현(1998)은 의존관계를 이용하여 명사구를 인식하고 색인어구로 정규화하는 과정을 수행하여 검색실험을 통하여 재현율은 저하되지 않고 약 19% 정확률이 향상됨을 보였다[10]. 원형석(2000)의 연구에서는 복합명사 분할 및 명사구 합성시 제한된 자연어처리기법을 이용하여 구문적인 구를 합성해내고 이를 조합한 실험에서 성능향상을 보였다[11]. 하지만 구를 구성하는 각 어절의 복합명사의 부분 명사들간의 관계에 대한 고려는 하지 않았다. 양재형(2000)은 학습말뭉치를 이용하여 규칙기반학습을 통한 기반명사구 인식을 하였으나[12], 초기 말뭉치 구축이 쉽지 않아 다양한 도메인 적용이 힘들다는 한계를 지닌다. 이충희(2002)는 구문분석결과를 통하여 구 단위 색인어 추출을 하였다[13]. 구문분석기를 이용하면, 문장의 의미를 잘 반영하는 색인어를 추출할 수 있지만, 강건한 구문분석기가 없는 실정에서는 성공적으로 수행되기 힘든 문제점이 있다. 또한 구문분석속도는 문장이 길어질 경우 현저히 저하되어 실제상용시스템에서 적용하기 힘들다.

명사들의 결합은 무한한 단어 생성력을 가지고 있으므로, 검색 대상 문서 내의 어휘 또는 색인어와 사용자가 입력한 검색어 간에 형태상 불일치를 일으켜 검색 시 문서를 제대로 검색할 수 없는 문제를 발생시킨다[14]. 따라서 본 논문에서는 비중이 높고 정보 전달의 핵심이 되는 어휘인 연속된 명사열의 모든 결합 관계를 고려한 색인어 생성을 통하여 검색의 효과를 높이려고 한다.

Ⅲ. 한글 복합명사구 색인

복합명사구(compound noun phrase)는 두 개 이상의 명사가 하나의 의미를 가지는 단어 또는 구를 이루는 경우를 말한다. 복합명사(compound noun)는 하나의 어절에서 표현되거나 여러 개의 어절에서 연속된 명사열 형태, 조사나 동사를 사이에 두고 두 개의 명사가 복합명사의 의미를 가지는 경우가 있다[15]. 따라서 본 연구에서는 복합명사구 혹은 명사구(noun phrase)를 “여러 개의 어절에서 표현된 하나의 명사처럼 쓰일 수 있는 명사들의 묶음”으로 정의한다. 이러한 복합 명사구는 문서를 대표하는 색인어이거나 사용자의 검색어와 일치하는 형태일 수 있다. 예를 들면 “정보검색시스템”의 의미를 표현하는 형태는 아래와 같이 다양하다.

(문서 1) 정보검색시스템
(문서 2) 정보검색 시스템
(문서 3) 정보검색을 위한 실용적 시스템
(문서 4) 정보 검색 시스템
(문서 5) 정보를 검색하는 시스템은
(문서 6) 특허정보 검색시스템은
(문서 7) 개인정보 검색서비스 시스템은
(문서 8) 정보검색용 시스템은
(문서 9) 정보검색을 위한 ETLARS 시스템의 개발
(문서 10) 원문정보를 검색하는 시스템은

(문서 11) 시스템 평가 연구를 완료하고 정보 검색에
(문서 12) 시스템회사에서 취업정보 검색
(문서 13) 지리정보시스템, 정보검색
(문서 14) 구글에는 사람이 직접 정보검색결과를 삭제하는 시스템이 없습니다

‘정보검색시스템’으로 검색을 할 경우 복합명사 분해가 적용된 시스템이라면 (문서 1)에서 (문서 14)의 모든 문서가 검색 결과로 나올 것이다. 하지만 (문서 1)에서 (문서 10)은 같은 의미를 뜻하는 문서이지만 (문서 11)에서 (문서 14)는 전혀 다른 의미의 문서이다.

문서들을 살펴보면, (문서 1)에서 (문서 10)은 동일한 의미의 문서이지만 검색어와 일치하는 색인어의 발생 위치 및 출현 빈도수 등에 따라 유사도가 달리 계산되어 검색 순위가 결정된다. 또한 (문서 9)는 '정보검색시스템'을 뜻하는 구문이지만 (문서 13)보다 색인어 간의 거리가 멀어서 검색랭킹순위가 낮다. 이러한 문제점을 해결하기 위하여 본 논문에서는 다양한 형태의 복합 명사구를 정규화하여 색인시스템에서 모두 동일한 색인어로 표현함으로써 검색어와의 불일치 문제를 최소화하려고 한다.

본 연구에서 제안하는 시스템은 크게 [그림 1]과 같다. 형태소 분석 모듈, 명사구 인식 모듈, 명사구 합성 모듈로 나누어진다. 이번 장에서는 각 모듈에 대하여 기술한다.

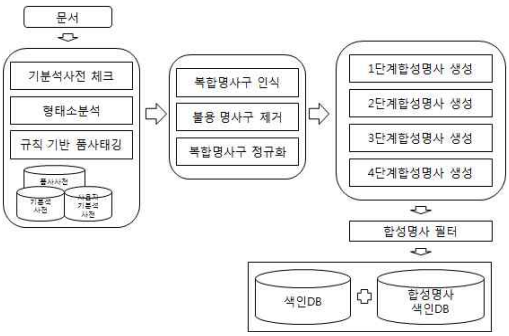

그림 1. 복합명사구 색인시스템 구성도

1. 복합 명사구 인식

본 논문에서는 "여러 개의 어절에서 표현된 하나의 명사처럼 쓰일 수 있는 명사들의 묶음"으로 정의한 복합명사구 인식을 다음과 같이 수행한다.

복합 명사구 인식 모듈은 이전 단계의 품사 태깅된 정보를 이용한다. 본 연구에서는 품사태깅 단계에서 속도를 고려하여 학술논문의 고빈도 단어 및 오분석이 자주 발생하는 단어들에 대하여 사용자기분석사전[3]으로 구축하고, 형태소분석 전에 우선적으로 활용한다. 기분석 사전의 예는 [표 2]와 같다.

표 2. 사용자기분석사전 예

기분석 단어	기분석 결과
효율적인	효율적/91+이/jp+ㄴ/etm
구조물의	구조물/91+의/jcm
추출물의	추출물/91+의/jcm
효과적인	효율적/91+이/jp+ㄴ/etm
전동기의	추출물/91+의/jcm
생물학적	생물학/91+적/xsn
방사선학적	방사선학/ncn+적/xsn
자동차용	자동차/91+용/xsn
여대생의	여대생/91+의/jcm
비실시간	비/xp+실시간/ncn
생물학적인	생물학/91+적/xsn+이/jp+ㄴ/etm
관한연구	관한/pvg+ㄴ/etm+연구/ncn

[표 2]에 포함된 '효율적인' 형태소분석결과는 다음과 같다.

(1) 효율/ncn+적/xsn+이/jp+ㄴ/etm
(2) 효율/ncn+적인/ncn

(2)의 '효율'과 '적인'은 실제로 복합명사로 발생할 수 없으므로, 기분석 사전에 (1)결과만 저장함으로서 형태소분석의 불필요한 분석을 제한한다. '관한연구'는 띄어쓰기의 잘못된 예이나, 실제로 논문에서 자주 발생하는 오류이므로 기분석 사전을 통해 분석할 수 있도록 했다.

앞서 설명한 품사태깅된 정보를 활용하여 복합명사구를 인식한다. 본 연구에서는 대용량의 데이터를 처리하기 위한 학습/통계 기반의 명사구 인식은 속도 문제 때문에 상용 서비스에서는 제약이 있으므로 단순한 규칙 기반의 명사구 인식을 선택하였다.

복합명사구의 인식을 위한 규칙 템플릿 생성을 위하여 KISTI NDSL에서 서비스하고 있는 국내 학술 논문의 제목/초록 정보 10만건을 분석하여 고빈도의 복합명사구 추출을 위한 템플릿을 [표 3]과 같이 생성하였다. '_'는 어절을 구분하는 기호이며 '명사*'는 1개 이상의 명사를 뜻하고 '()'의 품사 어휘정보는 문맥에서 발견되었으나 인식된 명사구에서는 사용하지 않는 정보를 뜻한다.

3 사용자기분석사전은 KISTI에서 개발한 정보검색엔진(KRISTAL)의 색인용 사전으로서, 사용자가 분석어절에 대한 1개 이상의 형태소 분석결과를 미리 저장할 수 있다. 91:접미사를 포함하는 3음절명사, xsn:접사, ncn:명사, jcm:소유격조사, jp:서술격조사, etm:관형형어미, pvg:동사

표 3. 템플릿을 구성하는 패턴 및 예시

패턴	예시
명사명사*	의미검색
명사_명사*	의미 검색
명사(접사)_명사*	의미적 유사성검색
명사*(소유격조사)_명사*	의미의 검색
명사*(소유격조사)_(관형사)_명사*	의미기반 새 검색서비스
명사*(명사구불용어)_명사* 명사*(*~에 의한*)_명사* 명사*(*~에 관한*)_명사* 명사*(*~를 위한*)_명사*	검사에 의한 평가 검사에 관한 평가 검사를 위한 평가

[표 3]과 같은 특정 구문 패턴을 통하여 복합명사구를 인식한다. 기존 연구에 의하면 분석된 학술 논문 데이터의 복합명사구의 96%를 상위 5개의 패턴으로 추출했다[16].

마지막 패턴의 경우 '에 의한', '에 관한' 등은 복합명사구를 구성하고 있지만, 명사구 내에서 중요한 의미를 나타내지 않는 구문이다. 따라서 명사구를 인식한 다음에 정규화 단계에서 ()의 어휘는 제거된다. 국내논문에서 자주 발생하는 수식 구조를 가지는 패턴을 조사하여 233건의 어휘에 대하여 명사구 불용어 사전을 구축하였다. 명사구 불용어 사전은 명사구로 추출되었으나, 명사구에 포함되는 어휘로 부적절하다고 판단되는 어휘 패턴을 구축해놓은 사전이다. 사전의 표제어 예는 다음과 같다. 첫 번째 단어는 독립된 어절의 단어이고, 두 번째 단어는 이전 어절의 조사 정보이다.

```
.....
대한        에
대해        에
..
매우        에
비해        에
의해        에
이          에
이어        에
한          에
.....
위한        을
위해        을
이용한      을
잘          을
주고        을
줄          을
.....
통한        을
통해        을
.....
```

복합명사구가 인식이 되면 구문적으로 다르게 표현된 형태를 하나의 일관된 형태로 표현하기 위한 명사구 정규화 작업을 수행한다. 명사구 정규화는 조사, 접사, 수식어 등을 생략하는 것이다. [표 3]의 패턴에서 '()'로 표현된 품사의 어휘를 제거하는 과정이다.

이와 같은 방법으로 [그림 2]는 형태소 분석결과를 이용하여 [표 3]에서 정의한 템플릿에 의해 인식된 복합명사구를 보여준다. 인식된 복합명사구는 합성명사 생성을 위하여 조사나 접사 등을 생략한 정규화된 복합명사구 형태로 변형되어 출력된다.

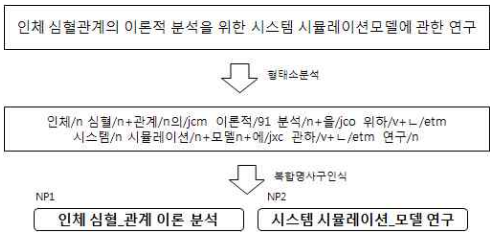

그림 2. 복합명사구 정규화 예

2. 복합명사구 합성

앞서 설명한 특정성이 높은 복합명사구가 색인과정에서 한 개의 색인어(index term)로 추출된다면 검색 단계에서 일치하는 색인어를 찾기가 쉽지 않다. 따라서 복합명사구를 구성하는 단일 명사들을 조합하여 새로운 합성명사를 생성하는 방법을 제안한다.

복합명사구가 추출이 되면 [표 4]와 같은 4가지 합성명사 생성 규칙을 이용하여 합성명사를 만든다. 'N_1N_2 N_3N_4 N_5N_6'은 정규화된 복합명사구이고 'N_1', ... ,'N_6'은 어절내에서 분해된 명사들이다. 생성된 합성명사는 두 개 이상의 명사로 구성된 복합명사구의 내포된 의미를 모두 포함한 경우 혹은 원래의 의미보다 넓은 의미를 지닌 경우가 된다. 복합명사구를 포함하는 명사들의 조합은 문장 곳곳에서 발견되므로, 모든 조합을 고려하여 합성규칙을 [표 4]와 같이 만들었다. 합성규칙을 통해 생성된 색인어는 중복을 제거하여 합성명사 색인 DB에 저장된다.

표 4. 합성명사 생성 규칙

패턴	규칙	
N_1N_2 N_3N_4 N_5N_6	합성규칙1	$N_1N_2N_3N_4N_5N_6$
	합성규칙2	$N_1N_2N_3N_4$ $N_3N_4N_5N_6$
	합성규칙3	$N_1N_2N_3$ $N_1N_2N_4$ $N_1N_3N_4$ $N_2N_3N_4$ $N_3N_4N_5$ $N_3N_4N_6$ $N_3N_5N_6$ $N_4N_5N_6$
	합성규칙4	$N_1N_2N_3N_4N_5N_6$ $N_1N_2N_3N_4$ $N_3N_4N_5N_6$ $N_1N_2N_5N_6$ $N_1N_2N_3$ $N_1N_2N_4$ $N_1N_3N_4$ $N_2N_3N_4$ $N_3N_4N_5$ $N_3N_4N_6$ $N_3N_5N_6$ $N_4N_5N_6$ $N_1N_2N_5$ $N_1N_2N_6$ $N_1N_5N_6$ $N_2N_5N_6$ .....

· 합성규칙 1은 복합명사구를 포함하는 모든 명사들을 합친 것을 말한다. 합성된 명사는 문서 전체에서 색인된 명사들과 비교해 볼 때 특정성이 있어 검색 정확률 측면에서 보면 핵심적인 색인어가 될 수 있다. (1)과 (2)는 서로 다른 형태의 복합명사구이지만 합성을 통하여 동일한 색인어를 생성하고 있다.

(1) '정보 검색 시스템' → '정보검색시스템'

(2) '정보의 검색을 위한 시스템은' → '정보검색시스템'

· 합성규칙 2는 복합명사구가 조사를 생략하거나 띄어 쓰여서 자주 표현되는 점을 고려하여 인접한 두 어절의 명사를 합친 경우이다. 복합명사구 전체가 내포하는 핵심적인 의미보다 포괄적인 의미를 갖는 복합명사를 검색하기 위함이다.

(3) '콘텍트 렌즈 가격' → '콘텍트렌즈', '렌즈가격'

· 합성규칙 3은 복합명사구를 구성하는 인접한 어절의 형태소와 합성한다.

(4) '지적재산권 보호' → '재산권 보호', '지적 보호'

· 합성규칙 4는 복합명사구를 구성하는 모든 형태소의 n-gram정보이다. 많은 합성명사를 만들어 낼 수 있지만, 잘못된 합성명사를 많이 생성하므로 정보량(information gain)이 낮은 합성명사를 제거하는 합성명사 필터를 이용한다[16].

(5) '지리 정보 시스템' → '지리정보', '지리시스템', '정보시스템', '지리정보시스템'

[그림 3]과 [그림 4]는 합성규칙을 적용하여 합성명사를 생성한 예를 보여준다. '_'는 한 어절에서 발생한 복합명사를 분해한 기호이다. 서로 다른 형태의 구문이지만 동일한 색인어 '유비쿼터스컴퓨팅보안'이 합성규칙 4와 합성규칙1을 통하여 생성됨을 보여준다.

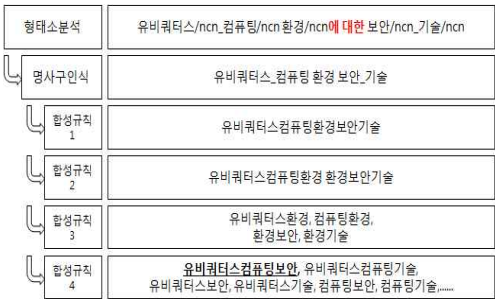

그림 3. 합성규칙 4단계 적용 예 1

그림 4. 합성규칙 4단계 적용 예 2

VI. 실험 결과 및 평가

1. 실험 결과

본 실험은 한글 과학기술 정보를 포함한 학술정보데이터에 대한 복합 명사구 색인기술을 적용하여 검색 정확률(precision) 및 재현율(recall)을 측정하기 위함이다. 하지만 도메인에 적합한 테스트 컬렉션이 마련되어 있지 않으므로 과학기술분야의 정보를 포함한 HANTEC

2.0[4]을 사용하여 시스템의 전반적인 평가를 수행하였다[17]. 테스트컬렉션에 포함된 과학기술분야 4만건 문서에 대한 정보는 [표 5]와 같고, 검색평가를 위한 적합문서집합은 과학기술 분야 집합 L2.rel[5]를 선택했다.

표 5. HANTEC 2.0 실험 문서 집합의 특성

주제	과학기술분야
출처	과기처지원 연구보고서(KRIST) 10,000건 해외과학기술동향(TREND) 18,000건 논문서지사항(SATURN) 12,000건
문서수	40,000건
질의수	과학기술분야 30건
질의 형태	자연어 질의어

HANTEC 2.0 과학기술분야 데이터를 통하여 검색 재현율과 정확률을 측정하였다. 질의어는 과학기술 분야 질의어집합을 사용하였고 질의어를 설명한 <desc> 필드의 자연어 문장을 그대로 사용하였다. 질의어 형태는 아래와 같다.

(1) "산업 폐기물 처리와 관계된 문제점과 처리기술"
(2) "컴퓨터기법이 사용된 영화의 제작과정이나 흥행에 관한 정보"

실험에 사용된 검색엔진은 KISTI에서 개발한 오픈소스 검색엔진 KRISTAL 3.1.29[6]이고, 검색모델은 검색문서의 순위화를 위하여 벡터공간모델(Vector Space Model)[7]을 선택하였다[18]. 실험방법은 다음과 같은 6가지 방법을 수행하였다. Baseline에 덧붙여 단계별 색인 모델을 적용한 후 검색 재현율 및 정확률이 향상됨을 확인한다.

Base1, Base2는 실험의 비교 평가를 위한 단일어절의 명사를 색인하는 기존의 색인시스템[8]이다. Case1, Case2, Case3, Case4는 본 연구에서 제안하는 복합명사구 인식 및 정규화, 합성규칙을 적용한 모델이다.

- Base1 : 복합명사 분해를 하지 않음(1어절 색인)
- Base2 : 복합명사 분해함(1어절 색인)
- Case 1[합성규칙 1] : 복합명사구 합성(2어절이상 색인)
- Case2 [합성규칙 2] : 인접한 어절 바이그램으로 합성한 경우(2어절이상 색인)
- Case3 [합성규칙 3] : 인접한 어절의 형태소 바이그램으로 합성된 경우(2어절이상 색인)
- Case4 [합성규칙 4] : 구를 구성하는 모든 형태소의 n-gram으로 합성된 경우(2어절이상 색인)

표 6. HANTEC 2.0 실험 문서 집합에 대한 재현율

모델	재현율(%)	재현율(%) (rank≤1000)
Base1	86.47%	51.08%
Base2	92.85%	58.20%
Case1	92.85%	60.22%
Case2	92.85%	61.61%
Case3	92.85%	61.61%
Case4	92.85%	62.69%

[표 6]은 적합 문서 집합에 대한 재현율이다. 복합명사 분해된 Base2모델을 통하여 검색어와 일치하는 적합문서를 모두 찾을 수 있으므로, 본 연구에서 제안하는 기술에 따른 검색 재현율의 변화가 없음을 확인할 수 있다. 하지만 상위 1,000개의 문서 검색 결과에 대한 재현율을 살펴보면, 복합명사구를 합성하는 4가지 규칙을 적용한 경우 약 4%정도 재현율이 상승됨을 확인할 수 있다. 이를 통하여 복합명사구 합성 색인과정이 적합 문서를 더 많이 포함할 수 있음을 보여준다.

4 1998년부터 2003년까지 한국과학기술정보연구원과 충남대가 공동으로 개발한 정보검색시스템평가를 위한 한글 테스트컬렉션
5 2인의 평가자가 부여한 점수중 5점 척도에 의해 낮은 점수가 2이상인 문서를 정답으로 한 집합 셋
6 KRISTAL은 과학기술문헌 정보서비스를 목적으로 개발을 시작한 정보검색관리시스템(IRMS; Information Retrieval Management System)
7 문서 색인어 가중치 산출에는 lac, 질의어 색인어 가중치 산출에는 ltc 기법을 적용함.
8 KRISTAL 3.1.29가 포함하고 있는 형태소 분석기를 활용한 색인 시스템. KRISTAL 형태소 분석기는 단일어절의 복합명사를 분해하고, 합성하는 기능을 가지고 있음

표 7. HANTEC 2.0 실험 문서 집합에 대한 정확률

	base1	base2	case1	case2	case3	case4
적합문헌 평균정확률	0.0996	0.1278	0.1456	0.1521	0.1514	0.1628
% change	−18.98	−	+13.93	+19.01	+18.47	+27.39
R 정확률	0.1332	0.1644	0.1995	0.2100	0.2080	0.2121
% change	−18.98	−	+21.35	+27.74	+26.52	+29.01

[표 7][9]은 기준모델(base2)에 덧붙여 합성규칙을 적용한 실험을 통하여 적합문헌평균정확률(Mean Average Precision), R-정확률(R-Precision)을 측정한 결과이다. case3은 case2와 비교하여 볼 경우 재현율 및 정확률의 변화가 미미하지만 제안한 4가지 유형의 복합명사구 합성규칙이 검색의 기준모델에 비교하여 향상됨을 확인할 수 있다.

표 8. 다양한 문서 순위에 따른 정확률

	base1	base2	case1	case2	case3	case4
5	0.2526 (−22.59)	0.3263 −	0.3579 (9.68)	0.4000 (22.59)	0.3895 (9.37)	0.4421 (35.49)
10	0.1789 (−37.05)	0.2842 −	0.3053 (7.42)	0.3053 (7.42)	0.3158 (11.12)	0.3158 (11.12)
15	0.1579 (−34.78)	0.2421 −	0.2667 (10.16)	0.2737 (13.05)	0.2632 (8.72)	0.2702 (11.61)
20	0.1474 (−29.10)	0.2079 −	0.2316 (11.40)	0.2342 (12.65)	0.2342 (12.65)	0.2395 (15.20)
30	0.1316 (−27.89)	0.1825 −	0.1982 (8.60)	0.2035 (11.51)	0.1947 (6.68)	0.2053 (12.49)

30개의 검색 순위에 대한 문서 수준의 정확률 [표 8]을 보면 합성규칙을 적용한 경우 문서 순위에 대한 검색 정확률이 기준모델보다 상승됨을 확인할 수 있다. 이러한 결과는 [표 8]을 나타낸 [그림 5]에서 알 수 있듯이 본 연구에서 제안한 합성규칙을 통한 색인이 검색의 정확률 향상에 유용한 방법임을 제시한다.

9 실험에 대한 평가는 HANTEC 2.0에 포함된 SMART 시스템의 평가 프로그램(trec_eval)을 사용함

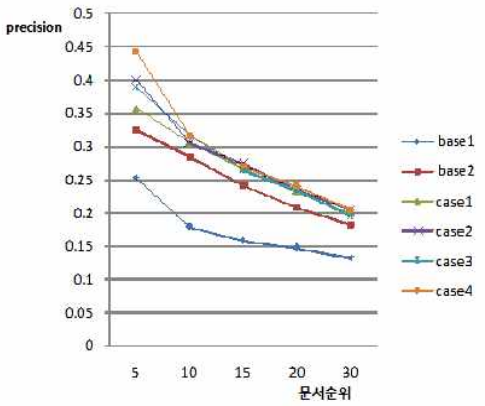

그림 5. 문서 순위에 따른 정확률 그래프

표 9. HANTEC 2.0 실험 문서 집합에 대한 R-정확률

질의어	문서	R-정확률(%change)
단일어절 색인	단일어절 색인	0.1644(−)
복합명사구 색인	단일어절 색인	0.1694(3.04%)
단일어절 색인	복합명사구 색인	0.1809(10.04%)
복합명사구 색인	복합명사구 색인	0.2121(29.01%)

[표 9]는 본 연구에서 제안한 복합명사 합성규칙을 질의어 색인 및 문서 색인에 각각 적용하여 비교 실험한 결과이다. 첫 번째 행의 검색 정확률은 질의어와 문서에 대해 단일어절을 복합명사 분해하는 base2 모델의 결과이다. 네 번째 행은 본 논문에서 제안한 복합명사구 합성 색인을 질의어 및 문서에 모두 적용하였을 경우의 결과이다. 질의어만 복합명사구 색인을 한 경우와 문서만 복합명사구 색인을 한 경우의 실험보다 질의어와 문서 색인에서 모두 복합명사구 색인을 한 경우가 기준모델보다 R-정확률이 29% 향상됨을 확인할 수 있다.

VI. 결론 및 향후연구

국내 학술정보 검색 서비스는 사용자에 따른 다양한 형태의 복합 명사구에 대한 일관성 있는 검색 결과를 내주지 못하고 있다. 본 연구에서는 이러한 문제점을

개선하기 위하여 복합명사구를 효율적으로 처리할 수 있도록 국내학술정보 논문에서 검색어와 색인어간의 불일치 문제를 유발하는 패턴을 분석하여 규칙 템플릿을 만들고, 이를 이용한 복합 명사구의 인식 및 정규화, 복합명사구에 대한 색인어의 특정성 문제를 줄이기 위한 4단계의 복합명사구 합성방법을 제안하였다. 이러한 방법은 실험을 통하여 기존의 색인 모델보다 검색 정확률 개선에 효율적임을 확인하였다.

NDSL에서 서비스하고 있는 국내 학술 논문 데이터의 복합명사구 어절 길이를 살펴보면 7어절[10] 이상으로 구성된 명사열이 상당히 많은 것으로 확인되었다. 이것은 한 개의 복합명사구가 여러 개의 의미를 가진 복합명사구를 포함할 수 있으므로, 서로 다른 의미의 경계 인식 및 정교한 복합명사구의 생성을 위한 템플릿에 대한 추가적 연구가 마련되어야 하고, 4단계 합성명사 생성 과정에서 과다 생성되는 색인어를 필터링하기 위한 기술 연구가 필요하다.

참 고 문 헌

[1] 임지희, 최호섭, 옥철영, “U-WIN 기반의 의미적 정보검색 기술”, 한국콘텐츠학회, pp.547-550, 2006.

[2] 조봉현, 이창기, 안주희, 이근배, “확률적 정보 검색 모델에서의 유사 적합성 피드백 실험”, 한국정보과학회언어공학연구회, pp.183-190, 2001.

[3] 최종희, 최동시, 박세영, “다중단어를 사용한 정보검색 시스템에서의 재현정확도 향상방법”, 한국정보과학회 학술발표 논문집, pp.150-152, 1998.

[4] 강남규, 조민희, 권오석, “NDSL 검색 질의어와 기술용어간의 관계에 대한 분석적 연구”, 정보관리연구, 제39권, 제3호, pp.163-177, 2008.

[5] 박찬이, 김상복, “거리 제한을 이용한 색인 시스템”, 한국컴퓨터정보학회 논문지, 제11권, 제1호, pp.273-282, 2006.

[6] K. Frantzi, S. Ananiadou, and H. Mima, “Automatic Recognition of Multi-Word Terms : the C-value/NC-value Method,” International Journal on Digital Libraries, Vol.3, No.2, pp.115-130, 2000.

[7] W. Zhang, S. Liu, C. Yu, C. Sun, F. Liu ,and W. Meng, “Recognition and classification of noun phrases in queries for effective retrieval,” CIKM, pp.711-720, 2006.

[8] C. Zhai, “Fast statistical parsing of noun phrases for document indexing,” pp.312-319, 1997.

[9] 강승식, “한국어 복합명사 분해 알고리즘”, 정보과학회논문지, 제25권, 제1호, pp.172-182, 1998.

[10] 윤보현, 김상범, 임해창, “한국어정보검색에서 구문적 용어불일치 완화방안”, 제10회 한글 및 한국어 정보처리 학술대회, pp.143-149, 1998.

[11] 원형석, 박미화, 이근배, “복합명사 분할과 명사구 합성을 이용한 통합 색인 기법”, 정보과학회논문지, 제27권, 제1호, pp.84-95, 2000.

[12] 양재형, 서영훈, “규칙 기반 학습에 의한 한국어의 기반 명사구 인식”, 정보과학회논문지, 제27권, 제10호, pp.1062-1071, 2000.

[13] 이충희, 김현진, 장명길, “구 분할을 이용한 명사구기반 색인의 성능향상”,한국정보처리학회 추계학술발표대회 논문집, 제9권, 제2호, pp.585-588, 2002.

[14] 임해창, 윤보현, 강승식, “한국학 서지정보와 전자텍스트를 위한 자동색인 및 검색시스템 개발 연구”, 한국어전산학, 제2권, pp.279-292, 1998.

[15] 최기선, *한국어에서 복합 명사구 인식에 대한 연구*, 한국전자통신연구원, 1993.

[16] 조민희, 정도헌, 홍순찬, 최성필, 최윤수, 전홍우, 정창후, 성원경, *과학기술 지식베이스 시스템의 주요기술개발 및 검증*, 한국과학기술정보연구원, 2010.

[17] http://www.kristalinfo.com/download/#hantec

[18] http://www.kristalinfo.com/download/#kristal

10 2009년까지 수집된 NDSL 국내학술 논문 DB에서 제목, 초록이 모두 포함된 21만여건 메타데이터로부터 추출한 복합명사구의 길이에 대한 통계를 보면 7어절 이상의 복합명사구를 가진 제목 49%, 초록 66%가 발견됨

저 자 소 개

조 민 희(Min-Hee Cho) 정회원

- 2003년 2월 : 연세대학교 전산학과(이학사)
- 2005년 2월 : 연세대학교 전산학과(이학석사)
- 2005년 4월 ~ 현재 : 한국과학기술정보연구원 연구원

<관심분야> : 자연언어처리, 시맨틱웹, 정보검색

정 도 헌(Do-Heon Jeong) 정회원

- 2011년 2월 : 연세대학교 문헌정보학과(박사과정수료)
- 2003년 6월 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 텍스트마이닝, 시맨틱웹, 정보검색