국한문 혼용 텍스트 색인어 추출기법 연구 『시사총보』를 중심으로*

An Experimental Approach of Keyword Extraction in Korean-Chinese Text

정유경 (Yoo Kyung Jeong)**
반재유 (Jae-yu Ban)***

초 록

본 연구는 국한문 혼용 텍스트를 대상으로 한글 형태소 분석 기법과 한문 어조사를 반영한 색인어 추출기법을 제안하였다. 국한문 혼용체로 작성된 『시사총보』 논설을 대상으로 해당 시기에 사용된 고유명사 및 한자어 사전을 보완하였으며 한자어 불용어 리스트를 고려하여 색인어를 추출하였다. 본 연구에서 제안한 국한문 색인 시스템은 수작업 색인 결과를 기준으로, 중국어형태소 분석기에 비해 재현율과 정확률 측면에서 상대적으로 높은 성능을 보였으며, 어문법이 확립되지 않은 근현대 시기의 국한문 혼용체를 대상으로 한 첫 번째 색인어 추출기법을 제안하였다는 데에서 연구의 차별점이 있다.

ABSTRACT

The aim of this study is to develop a technique for keyword extraction in Korean-Chinese text in the modern period. We considered a Korean morphological analyzer and a particle in classical Chinese as a possible method for this study. We applied our method to the journal "Sisachongbo," employing proper-noun dictionaries and a list of stop words to extract index terms. The results show that our system achieved better performance than a Chinese morphological analyzer in terms of recall and precision. This study is the first research to develop an automatic indexing system in the traditional Korean-Chinese mixed text.

키워드: 말뭉치, 자동 색인, 한자 형태소 분석, 국한문 혼용체, 시사총보
corpus, automatic indexing, traditional Chinese morphological analysis, Korean-Chinese character style, Sisachongbo

---

* 본 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017S1A6A3A01079581).
** 연세대학교 근대한국학연구소 HK연구교수(yk.jeong@yonsei.ac.kr) (제1저자)
*** 연세대학교 근대한국학연구소 HK연구교수(juban@yonsei.ac.kr) (교신저자)
■ 논문접수일자 : 2019년 8월 18일 ■ 최초심사일자 : 2019년 9월 18일 ■ 게재확정일자 : 2019년 12월 13일
■ 정보관리학회지, 36(4), 7-19, 2019. http://dx.doi.org/10.3743/KOSIM.2019.36.4.007

1. 서 론

한국 근대 초기에 해당하는 19세기 말은 독립신문・시사총보・황성신문・제국신문 등 언론 출판이 본격화된 시기로, 당대의 시대상을 이해할 수 있는 다양한 근현대 역사자료들이 생성되었던 때였다. 그 시기의 역사와 문화를 이해하기 위해 개별 연구자들은 해당 시기에 출판된 자료를 직접 발굴하고 그 자료의 의미를 파악하는 작업을 수행해 왔다. 최근 토대사업과 같은 역사자료의 디지털화 사업이 국가 차원에서 진행되면서 디지털화된 역사자료에 대한 접근성이 높아졌으며, 축적된 데이터의 양 또한 방대해졌다. 그럼에도 해당 시기에 축적된 방대한 양의 디지털 자료를 통계학 등의 사회과학적 연구방법론을 적용하여 분석하는 데에는 기술적 제약이 따른다. 특히, 해당 시기의 텍스트는 한글과 한자가 섞인 국한문 혼용체로 되어 있으며, 순한글 자료들도 'ᄋᆞ(아래아)'와 같은 옛 한글을 포함하고 있어 분석에 한계가 있다.

따라서 본 연구에서는 국한문 혼용 텍스트를 대상으로 한글 형태소 분석 기법을 적용한 텍스트 분석과 주제적으로 의미 있는 단어들을 선정하기 위한 색인어 추출 방법에 대해 연구하였다. 근대 초기 국한문 신문은 한자와 옛 한글이 섞여 있는 형태이지만 글이 쓰인 방식과 문법은 현재의 한글 어문법과 상당히 유사한 부분이 존재한다. 이러한 점에 착안하여 현재 통용되고 있는 한글 형태소 분석기에 한문 문법 적용과 어휘사전 보강 등을 통하여, 국한문 혼용 텍스트를 대상으로 한 색인어 추출기법(국한문 색인 시스템)을 제안하고자 한다. 본 연구에 사용된 말뭉치는 국한문 혼용체로 작성된 『시사총보』의 논설을 대상으로 하였으며, 전공 분야의 전문가 6인이 색인한 수작업 색인 결과와 비교하여 본 연구에서 제안한 국한문 색인 시스템의 효용성을 검증하고자 하였다.

2. 이론적 배경

2.1 국한문 형태소 분석의 난점

국한문 혼용체에서 체언의 경우 주로 한자어로 표기되며 격조사는 한글 형태로 구성된다는 특징이 있기 때문에 체언의 구분을 통한 형태소 분석의 가능성이 있다. 그러나 근대 시기의 국한문 혼용체는 문장 내에서 허사(虛辭)로서 문법적 역할만을 담당하는 어조사들이 실사(實辭)인 체언・용언 등과 함께 뒤섞여 표기되고, 격조사 역할을 하는 한글조차 옛 한글의 형태로 표기되기 때문에 형태소 분석기 개발의 난점이 많았으며, 그로인해 지금껏 형태소 분석기 개발의 시도조차 이루어지지 못하였다.

이 같은 문제점은 현행 국문법에서조차 한자 형태소의 성격이 명확히 규정되지 못한 한계점에서 비롯된다. 그동안 이익섭(1969), 김창섭(2001, 2013), 이상복(2012), 주지연(2015, 2017) 등의 꾸준한 논의들을 통해 한국어에 존재하는 한자 형태소를 적극적으로 인정하고 식별해야 한다는 의견이 거론되었다. 그러나 1음절 한자어의 다양한 의미・분포의 문제점에 대한 명확한 해결책을 제시해주지 못한 채, 2음절 한자어(2字語)를 '복합어(한자형태소 인정)로 볼 것인가' 아니면 '단일어(인정하지 않음)로 볼 것

인가'에 대한 논의가 여전히 첨예한 대립각을 세우고 있다.

① 형용사 +명사(小+國):
② 대명사+명사(對+韓)
③ 동사+명사(訪+韓)
④ 명사+명사(詩+集)
⑤ 대명사+명사(對+韓國)
⑥ 동사 +명사(訪+韓國)
⑦ 명사구 +동사구(一筆+揮之)
⑧ 명사(NN)+명사(VN)(身體+檢查)

위의 8가지 예는 한·중·일 언어에서 기본적으로 단어를 구성하는 방법으로, 해당 논의는 2음절 한자어(①~④)와, 관용적으로 사용되는 3음절·4음절 한자어(⑤~⑧)들을 어떠한 기준에서 복합어로 구분해야할 지에 대한 고민들로 이어진다.

이같이 국문 내 한자 형태소의 기준조차 정립되지 못한 채, 국한문 텍스트의 형태소를 분석하고 색인어를 추출한다는 것은 불가능한 작업이다.

또한, 근대의 국한문은 단어이면서 상황에 따라 어조사의 기능을 하는 한자 어휘들이 상당수 출현하기 때문에, 특정 색인어를 추출하기 위하여 일정한 문법적 규칙을 적용하기 매우 어렵다. 결국, 위와 같은 난점들은 국문과 한문이 혼재되어 있는 이중언어 구조에서 비롯한 것이다.

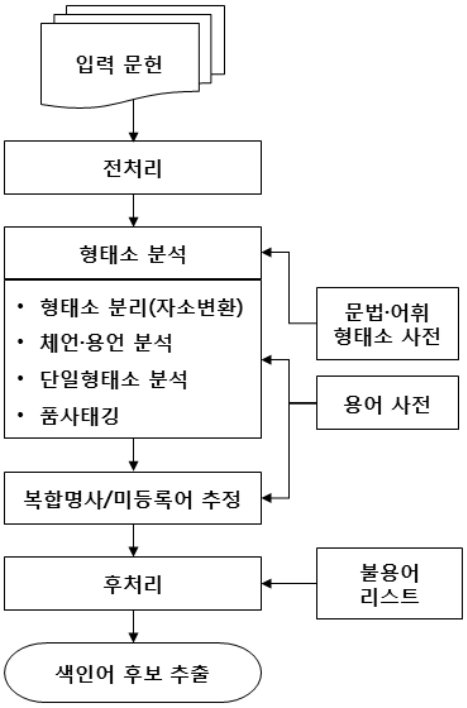
그러나 국한문 혼용체는 근대라는 특정 시기에 사용된 우리말의 글말 형태이기 때문에, 한자(實辭·虛辭)와 옛한글('ㆍ')의 방식을 현 국문 표기에 가까운 형태로 변환할 수 있다면 변환된 텍스트에 대한 색인어 추출이 가능해질 것이다. 본 연구에서 제안한 '국한문 색인 시스템'은 이 같은 고민들이 반영된 결과라 할 수 있다.

2.2 형태소 분석과 색인

자동색인은 색인어휘집이나 기능어휘집 등 미리 정해진 사전을 활용하는 방법과 형태소 분석기를 활용하는 방법이 있다. 사전에 정해진 어휘집을 사용해서 색인어를 추출할 경우, 불필요한 색인어가 추출되는 경우가 많고 기능의 한계가 있기 때문에, 형태소 분석의 결과를 사용한 자동색인 기법이 주로 사용된다(강승식, 권혁일, 김동렬, 1995). 또한, 단어의 유형과 형태소의 분리, 품사 정보를 이용하여 해당 텍스트에서 색인어를 추출하기 때문에 형태소 분석의 결과가 자동색인의 성능에 큰 영향을 미친다.

초기에 개발된 한글 형태소 분석기는 HAM(Hangul Analysis Module)으로 한국어 음절 특성을 이용하여 음절 단위의 형태소 분석 기법을 채택하였다. 그 외에도 KAIST의 한나눔 형태소 분석기, 일본어 형태소 분석기인 MeCab를 변형한 은전한닢 프로젝트, 꼬꼬마 분석기 등의 형태소 분석기가 있으며, 품사태깅을 위해 기분석 사전이나 시스템 사전 등 품사정보를 포함하고 있는 사전과의 매칭을 통해 형태소 분석을 수행한다. 특히 색인어 추출 성능에 중요한 영향을 미치는 복합명사 추출에서, 한나눔 형태소 분석기가 꼬꼬마 분석기에 비해 단일명사의 범주를 잘 잡아내고 복합명사의 인식율이 높은 것으로 나타났다(박주희, 맹성현, 2017). 한글 형태소 분석기를 활용한 색인어 추

출 과정은 다음 〈그림 1〉과 같은 순서로 이루어진다(강승식, 2002).



〈그림 1〉 색인어 후보 추출 개요

입력 문헌을 대상으로 가장 먼저 수행되는 것은 전처리(preprocessing)과정으로, 이 과정에서 형태소 분석을 위해 입력 문자열인 원 텍스트(raw text)로부터 문장 부호, 숫자, 특수문자와 같이 형태소 분석 대상이 되지 않는 문자들을 분리하거나 공백으로 대치하여 문자열을 제거한다. 전처리를 거친 텍스트는 어절(token)단위로 분리한 후, 용언과 체언 등의 품사정보를 포함한 기본사전과 비교하여 문법형태소로 분석하고 한국어 단어 형성규칙에 따라 최소형태소 단위로 분리하는 과정을 거친다. 이 과정에서 문법형태소인 조사와 어미를 분리하고 용언인지 체언인지에 따라 각 품사에 따른 최소형태소 분석이 이루어진다. 색인어는 주로 명사와 같은 체언으로 구성되어 있기 때문에 이 과정에서 이루어지는 형태소 분석의 성능이 색인어 추출에 가장 많은 영향을 미친다. 조사가 분리된 단어는 체언으로 추정하고 어휘 사전에서 단어를 확인하여 체언 여부를 결정한다. 용언으로 추정되는 단어의 경우 용언화 접미사인 '-하다/되다/시키다'와 같은 접미사가 발견되면 체언 여부를 검사한다. 서술격 조사 '이'가 발견되거나 서술격 조사가 생략된 경우에도 체언 분석과정에서 처리한다. 형태소 분리 단계에서 어조사 사전 참조를 통해 조사나 어미가 발견되지 않은 단어는 단일형태소로 이루어진 단어라고 추정하여 이 경우에도 어휘사전을 참조하여 체언 여부를 확인한다(강승식, 2002; 정영미, 2012).

근대시기 국한문 자료는 본격적으로 디지털화된 기간 자체가 짧기 때문에, 해당 텍스트에 대한 형태소 분석 연구는 찾아보기 어렵다.

다만, 국한문 혼용 텍스트 대상의 형태소 분석을 위해 홍성혁, 김철수, 이용석 등(1996)이 기존의 사전 구성 방식을 보완한 한글・한자사전 통합저장 방식을 제안한 것이 주목할 만하다. 형태소 분석에서 성능에 가장 중요한 영향을 미치는 것은 품사사전으로 한글과 한자가 섞인 텍스트의 경우, 단일문장을 처리할 때 한글사전과 한자사전이 동시에 요구되므로 메모리 측면에서 매우 비효율적이며 동음이의어 처리 등의 문제에 한계가 있다는 점을 지적하였다. 이러한 난점을 극복하기 위해 한글 사전만으로 형태소 분석을 수행할 수 있도록 하는 사전 저장방식을 고안하였다.

그러나 이 방법은 현대 국어문장에 체언만 한

자로 표기한 국주한종(國主漢從)의 국한문 텍스트에서 구성할 수 있는 방법으로, 한자어가 대부분을 차지하는 근대 시기의 텍스트 분석에는 적합하지 않다는 한계가 있다. 또한, 메모리의 측면에서 한자어 사전을 효율적으로 구성할 수 있는 방법에 대한 제안으로서, 본격적인 형태소 분석이나 색인어 추출에 대한 연구라고 볼 수 없다.

3. 연구 방법

본 연구에서는 한자의 비중이 높고 옛 한글이 섞여 있는 근현대 시기의 국한문 텍스트에서 색인어를 추출하는 방법을 제안하기 위해 한글형태소 분석기와 해당 시기에 사용되었던 한자어를 보강한 한자 사전, 한자어 어조사 문법을 적용한 색인어 추출방법으로 제안하고자 한다.

자동색인에서는 일반적으로 사전에 등재된 단어 외에 복합명사를 인식하지 못하는 문제와 미등록어 및 불용어 처리의 문제가 색인 성능에 많은 영향을 미친다. 본 연구에서는 한국어 형태소 분석기의 문장 분석기능과 체언 위주의 품사 태깅 결과를 활용하였으며 복합명사의 인식과 색인어 추출의 재현율을 높이기 위해서 표준대국어사전의 한자어와 약 50만 어휘에 달하는 한국한자어사전의 한자어를 보완하였다. 또한, 한문 어문법에서 어조사로 쓰이는 한자어들을 사용하여 어절 분리 및 색인어 추출에 사용하였다. 또한, 미등록어로 표기된 단어와 한문 불용어를 추가하여 색인의 성능을 높이도록 하였다.

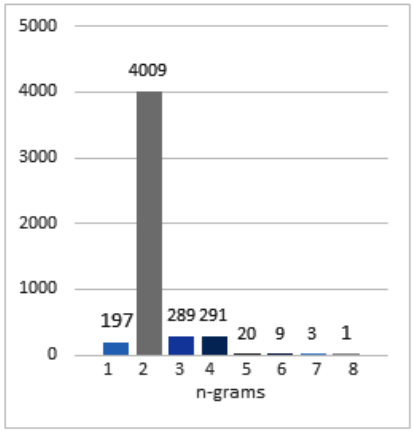
3.1 대상 말뭉치

색인어 추출 실험을 위해 사용된 말뭉치는 한말 잡지인『시사총보』로, 해당 잡지에 수록된 논설 100건을 대상으로 원문입력과 색인어 추출 실험을 진행하였다.『시사총보』는 1899년 1월 22일부터 8월17일까지 격일간(隔日間)으로 발행된 근대초기 대표적 국한문신문이다. 특히「논설」(『시사총보』)은『皇城新聞』과『大韓每日申報』등으로 이어지는 근대초기 한국학의 지적 흐름을 이해할 수 있는 중요한 자료이다.

원문입력 과정에서는 형태소 분석과 문장 분리를 위해 말뭉치에 문단과 문장 분리 태깅을 수행하였으며, 색인어 선정과정에서는 국어국문학과 역사학 전공의 색인자 6명이 색인어 선정과 검증에 참여하였다. 색인어 선정에는 명사 위주의 추출을 기본으로 하였으며, 한자어 대명사(余, 吾, 吾儕 등), 감정 관련 어휘(慘毒, 欣感 등), 불특정 다수를 지칭하는 대명사(人, 民 등)는 색인어에서 배제하였다. 다만, 불특정 다수를 지칭하는 대명사 人, 民 등의 경우에는, 露人(러시아사람), 人民, 公民과 같이 합성어로 사용되어 별도의 의미를 갖는 경우 색인어로 추출하였다. 또한, 賑穀(진곡)을 賑之穀로 사용하는 예와 같이 중간에 어조사가 있지만 전후가 연결되어야 뜻이 성립하는 경우와, '之'가 관용적으로 함께 지칭되는 단어('惠民之署'), 그리고 고유명사('興仁之門') 등은 모두 색인어로 포함하였다.

원문으로 입력된 텍스트는 제목과 공백을 제외하고 총 95,037자로 구성되어 있으며, 전문가의 수작업 색인을 통해 색인어로 선정된 단어는 8,256개(고유단어 4,824개)의 단어였다.

최종 색인어 선정과정에서 불필요한 색인어를 제외하기 위해 수작업 색인어들의 글자 수 분포를 반영하여 불용어로 선정에 사용하였다. 다음 〈그림 2〉는 전문가가 수작업으로 색인한 색인어의 글자 수 분포이다.



〈그림 2〉 수작업 색인어 글자 수의 분포

색인어로 선정된 단어는 4,824단어 중 4,009개로 약 80% 이상의 색인어가 두 글자로 나타났다. 이러한 결과를 바탕으로 형태소 분석에서 미등록어로 판별된 단어들을 대상으로 bi-gram으로 단어를 분할하여 불용어 리스트 선정에 참조하였다. 수작업 색인을 통해 추출된 단어들은 색인 작업의 기준(golden standard)으로 간주하여 본 연구에서 제안하는 색인어 추출기법의 성능 측정에 사용하였다.

3.2 색인어 추출기법

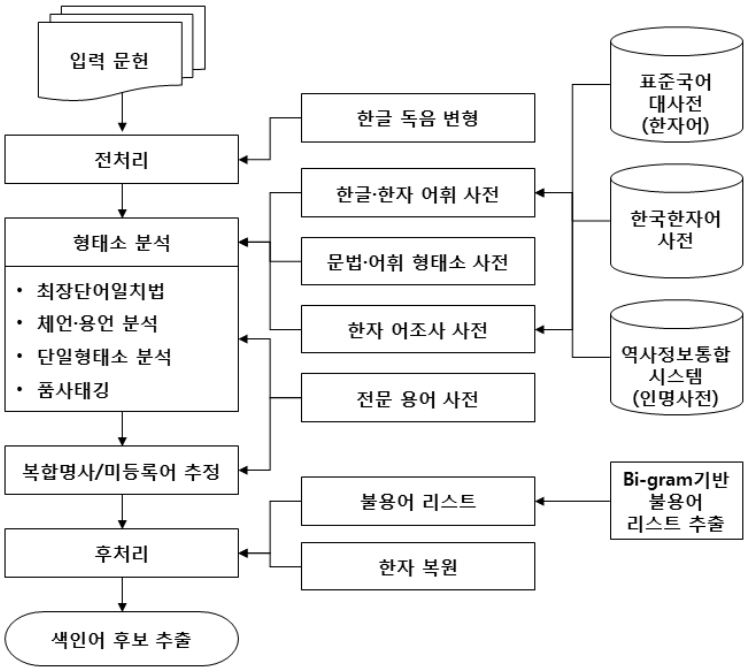
본 연구는 국한문 혼용 텍스트에서 색인어를 추출하기 위해 한글 형태소 분석기와 한자어 사전을 사용하여 색인어 추출하는 방법을 제안한다. 기존 사전기반의 색인어 추출방법은 불필요한 색인어가 과도하게 많이 추출되는 문제가 있기 때문에 형태소 분석기의 품사태깅 결과와 용어가중치 등을 반영한 색인어 추출이 이루어진다. 하지만 국한문 혼용 텍스트를 대상으로 한 형태소 분석기가 존재하지 않고 문법체계 또한 명확하게 정립되어 있지 않기 때문에 국한문 혼용 텍스트에서 색인어를 추출하기 위해서는 전적으로 사전에 의존할 수밖에 없는 실정이다.

한국 근현대 시기의 국한문 혼용 텍스트는 한자어와 옛 한글이 혼합되어 문장을 구성하고 있지만, 한글 독음으로 문장을 해석했을 때 의미의 변화가 크지 않다는 점에 착안하여 한글 형태소 분석기를 활용한 추출기법을 제안하였다. 〈그림 3〉은 본 연구에서 제안하는 색인어 추출의 개요이다.

전처리과정에서는 기존에 존재하는 한글 형태소 분석기를 사용하기 위해 텍스트에 포함되어있는 한자어를 한글 독음으로 변형하는 과정을 거쳤다. 예를 들어 원문이 한자와 옛 한글로 구성된 경우 다음과 같이 한글 독음으로 변형하고 옛 한글인 아래아를 변형하여 분석에 사용하였다.

원문: 八月一日에 皇上陛下계셔 法部로 詔勅을 下ᄒᆞ셧ᄂᆞᆫᄃᆡ 恭錄如左ᄒᆞ노라
변형: 팔월일일에 황상폐하계셔 법부로 조칙을 하하셧는데 공록여좌하노라

이때 한자의 원형과 한글 독음을 쌍으로 저장하여 최종 색인어 후보 리스트를 추출할 때 한자어로 복원하였다.



〈그림 3〉 색인어 추출의 개요

한글 형태소 분석기를 활용한 분석과정에서는 현대 한글 기분석 사전에 등록되지 않은 한자어-한글독음 사전을 추가하여 식별될 단어의 재현율을 높였다. 사용한 사전으로는 국립국어원의 표준국어대사전에 수록된 한자어 434,233단어와 한국한자어 사전에 등재되어있는 318,746단어, 역사정보통합시스템에 있는 한국근현대 인물자료, 독립운동가자료, 독립운동가옥고(獄苦)기록, 독립운동자 공적자료, 한국역대 인물자료, 영남 유학 인물자료, 한국족보 인물자료, 한국역대과거합격자, 동아시아역사인물, 일제감시대상인물카드에 수록된 193,246개 고유명사를 사용하였다. 한글의 경우 사전매칭과정에서 자소 단위의 분리를 통해 단어의 전방일치 여부를 따지지만, 국한문 혼용체의 경우 자소 단위의 분할이 되지 않기 때문에 완전일치법을 사용하여 색인어를 추출하였다. 다음으로 한자어 어조사 사전을 참조하여 한자어 문법을 적용한 어절 분할을 수행하였다. 문장 내에서 문법적 기능만을 담당하는 9개의 품사들(대명사・조동사・부사・접속사・전치사/후치사・종결사・감탄사・양사・기타)을 고려하였으며, 而, 又, 然, 且와 같은 접속사와 以, 之 같은 전치사 및 후치사를 사용하여 어절을 분할하였다.

최종 색인어 후보 리스트를 추출하는 과정에서 주제어로서 의미가 없는 단어들을 제외하기 위해 두 글자 고빈도어와 미등록어로 판별된 단어를 bi-gram으로 분할하여 추출한 총 4,145개의 단어 중 83개 단어를 불용어로 선정하였다. 또한, 한 글자 단어는 색인어로서의 주제적 식별력이 떨어진다고 판단하여 색인어 후보로 추출하지 않았다.

3.3 실험

본 연구에서 제안한 기법의 유용성을 검증하기 위해, 전문가 6인이 색인한 수작업 색인어를 기준으로 하여 평가를 진행하였다. 추가로 다양한 형태소 분석기 중 국한문 혼용체의 텍스트에 알고리즘의 변형이나 수정 없이 적용할 수 있으며, 문법적 유사성이 있어 한자어 대상의 자동 형태소 분석이 가능한 중국어 형태소 분석기를 추가 비교 대상으로 고려하였다. 동등한 결과 비교를 위해 동일한 언어 대상의 형태소 분석기를 사용해야 하지만, 번자체 한자어를 대상으로 한 형태소 분석기가 존재하지 않으며, 중국어 형태소 분석기에 탑재된 사전에는 번자체 정보를 함께 수록하고 있기 때문에 근현대시기의 한자어 텍스트에 적용 가능한 중국어 형태소 분석기를 사용하여 차이를 살펴보았다.

중국어 형태소 분석기로는 품사 태깅에 HMM 모델을 적용한 結巴(jieba)[1] 형태소 분석기를 사용하였으며, 한글 텍스트에는 한나눔 형태소 분석기[2]를 사용하였다. 한자-한글 독음 변형에는 hanja 라이브러리[3]를 사용하였다.

색인어의 성능평가를 위해서는 총 색인어 후보 중에 정확히 추출된 색인어의 비율을 의미하는 재현율(recall)과 추출된 색인어 후보 중 정확히 추출된 색인어의 비율을 의미하는 정확률(precision)을 사용하여 측정하였다. 단, 실험을 수행한 기사에 따라 텍스트의 분량이 길거나 연재형일 경우 다수의 색인어가 중복으로 추출되는 문제점이 발생했다. 그러므로 본 실험에서는 중복된 색인어(동일 색인어)를 제외한 결과로 성능을 평가하였다.

4. 실험 결과

4.1 색인어 추출 결과

『시사총보』의 논설 100건을 대상으로 색인어 후보 추출 실험을 수행한 결과, 각 방법별로 추출된 색인어 후보의 수는 다음 〈표 1〉과 같다.

〈표 1〉 추출 색인어의 수

방법	추출 색인어의 수 (동일 색인어 제외)
수작업 색인어	8,256단어 (4,626)
중국어 형태소 분석기	16,234단어 (12,585)
국한문 색인 시스템	21,861단어 (6,596)

중국어 형태소 분석기를 사용하여 색인어 후보를 추출했을 경우, 국한문 색인 시스템에 비해 색인어로 추출된 한자어가 상대적으로 작았으나, 기사 내 동일 색인어를 제외하고 추출한 고유단어의 수는 12,585단어로 상당히 많았다. 특히, 현대 중국어에서는 의미를 갖는 한자어로 취급하지만, 국한문 혼용 텍스트에서는 어조사로 취급하는 단어들이 포함되어 의미 없는 색인어들을 다수 추출한 것으로 나타났다.

1) https://github.com/fxsjy/jieba
2) http://semanticweb.kaist.ac.kr/hannanum/index.html
3) https://github.com/littmus/hanja

반면에 본 연구에서 제안한 방법의 경우, 중국어 형태소 분석기에서 추출된 색인어 후보 리스트에 비해 상당히 많은 단어들이 추출되었다. 이는 현대 한국어 형태소 분석기에서 명사, 고유명사 등으로 판별된 단어들을 색인어 후보로 모두 추출한 후 불용어와 양사, 허사 등 색인어로서 부적합한 단어가 제외하였기 때문에 주제어로서도 적합한 단어들이 상당히 많았다. 다음 〈표 2〉는 각 방법별로 추출한 색인어의 빈도수 상위 10단어들이다.

〈표 2〉 빈도수 기준 색인어 상위 10단어

순위	수작업 색인어	중국어 형태소 분석기	국한문 색인 시스템
1	天下	天下	天下
2	國家	國家	國家
3	政府	是故	政府
4	人民	以來	我國
5	兩班	我國	議政
6	議政	所謂	人民
7	全國	政府	全國
8	法律	然則	兩班
9	生民	然後	富强
10	富强	而已	天地

〈표 2〉의 결과에서 알 수 있듯이 중국어 형태소 분석기를 사용하여 분석한 경우, 상위에 추출된 색인어 후보 10단어 중 겹치는 단어는 '天下'와 '國家' 두 단어로 나타났다. '是故'와 같이 접속사의 역할을 하거나, 국한문 혼용체에서 어조사의 역할을 하는 단어들을 포함하여 문법적인 기능은 있으나 색인어로서는 의미 없는 단어들이 다수 추출되었다. 반면에 본 연구에서 제안한 방법은 수작업으로 색인한 단어들과 거의 유사한 색인어 추출 결과를 보여주고 있다. 상위 10개의 단어 중 8개의 단어가 일치했으며, 단어의 빈도순위도 유사하였다.

중국어 형태소 분석기 대 본 연구에서 제안한 방법의 성능 비교 결과는 다음 〈표 3〉과 같다.

〈표 3〉 성능 평가

방법	재현율	정확률
중국어 형태소 분석기	64.1%	23.6%
국한문 색인 시스템	72.3%	50.5%

중국어 형태소 분석기의 경우 각 기사별로 올바르게 추출된 색인어로 판별된 단어는 2,967개로 나타났으며, 잘못 추출된 단어들의 경우 네 글자 이상의 단어나, 현대 중국어에서는 문법적 기능을 하지만 색인어로서 의미 없는 품사들이 다수 추출되었다. 상당히 다양한 고유단어가 추출된 것에 반해, 수작업 색인과 일치한 단어의 수가 매우 적었기 때문에 낮은 정확률을 보였다. 본 연구에서 제안한 방법에서는 각 기사 별로 3,332개의 색인어가 수작업 색인어와 일치된 결과를 보였으며, 한 글자로만 구성된 197개의 색인어는 성능평가에서 제외되었다. 또한 歐陽氏(구양씨)와 같이 인명사전에는 등재되지 않은 단어들이 색인어로 추출되지 않았다. 완전 자동색인은 색인전문가의 개입이 없기 때문에 얼마나 많은 색인어를 추출하였는지를 측정하는 재현율에 비해, 정확한 색인어를 추출하였는지를 측정하는 정확률이 중요시되는 경향이 있다(김판준, 2006). 본 연구에서 제안한 방법의 경우, 재현율뿐만 아니라 정확률의 측면에서도 중국어 형태소 분석기에 비해 높은 성능을 보였다.

〈그림 4〉는『시사총보』의 논설 중 "학교설

〈그림 4〉 논설 '학교설립제도'의 색인어 추출 예

립제도" 원문 기사의 일부와 각 방법별로 추출된 색인어를 표시하고 있다. 수작업 색인어는 "Manual Index", 중국어 형태소 분석기로 추출된 결과는 "Ch Index", 본 연구에서 제안한 국한문 색인 시스템은 "Auto Index"로 표시하였다. "學敎"와 같이 각 단어 위에 세 개의 태그가 모두 붙어 있을 경우, 세 가지 방법에서 모두 추출된 단어를 의미한다. 중국어 형태소 분석기에서는 〈그림 4〉의 예에서도 색인어로서 의미 없는 단어들이 많이 추출되었다. 첫 번째 줄의 '敎育'은 추출되지 않았으며, 세 번째 줄의 '我國中'의 경우, '我國'이 추출되어야 하지만 '國中'이 추출되는 등 단어 분절시 오류가 있었다. 국한문 색인 시스템의 경우 수작업 색인어의 대부분을 추출했으며, 용언의 역할을 하는 한자어들까지 색인어로 추정하였다.

중국어 형태소 분석기에서는 사용하는 문법 체계와 어휘가 우리나라 근·현대시기의 한자체와 다르기 때문에, 추출된 색인어의 수가 상대적으로 적게 나타났다. 사전에 등재되지 않은 옛 어휘와 한자 이체자(異體字)가 포함된 단어를 색인어로 잡지 못하는 데 비해, 국한문 색인 시스템의 경우 해당 어휘들을 모두 색인어로 잡아내고 있다.

4.2 논의점

재현율의 측면에서 국한문 색인 시스템을 통

해 수작업 색인어로 추출된 단어의 대부분을 추출했음에도, 정확률의 측면에서 성능향상을 위해 해결해야 할 문제점을 발견할 수 있었다.

먼저, 같은 어절에 어조사가 포함된 어휘가 중복으로 추출되는 경우이다. 〈그림 4〉의 본문 첫 번째 줄의 '夫人才'에서 '夫'는 어조사로서 문법적 기능을 담당하는 단어이다. 그러나 국한문 색인 시스템의 경우, 어조사 불용어 처리의 순서보다 사전에 등재된 색인어 추출이 먼저 수행되기 때문에, '부인(夫人)'과 '인재(人才)' 모두 색인어로 잡아주는 결과가 나타났다. 본 연구에서는 연구자가 문장의 내용을 식별하여 색인어를 임의 선택하는 과정을 거쳤지만, 후속 작업에서는 각 어조사에 따른 개별 규칙을 부기하여 자동으로 색인어를 선별할 수 있도록 개선할 수 있을 것이다.

다음으로는 한자의 동자이음(同字異音)으로 인하여, 국한문의 한글 변환시 동음이의어(同音異議語)의 어휘가 색인어로 추출되는 경우이다. 〈그림 4〉의 6번째 줄 '善則善'을 살펴보면, '則善'이 색인어로 추출된 것을 알 수 있다. '則(법칙 칙, 곧 즉)'은 한글 독음으로 변형했을 때 가장 첫 번째 독음인 '칙'으로 변형된다. '즉'이 아닌, '칙'으로 변환되기 때문에, 기존 한글 사전에 등재된 '칙선(勅宣・勅選)'으로 해당 어휘를 판별하여 색인어로서 '즉선(則善)'을 추출하는 문제점을 발생시켰다. 이러한 동자이음(同字異音)에 대한 문제는 추후 연구를 통해 선별규칙에 대한 기준을 마련하여 개선할 필요성이 있다.

5. 결 론

본 연구에서는 국한문 혼용 텍스트를 대상으로 한글 형태소 분석 기법과 한문어조사를 반영한 색인어 추출기법을 제안하였다. 국한문 혼용체로 작성된『시사총보』논설을 대상으로 하였으며, 해당 시기에 사용된 고유명사 및 한자어 사전 보완과 한자어 불용어 리스트를 고려하여 색인어를 추출하였다. 색인어 추출실험 결과, 약 70%의 재현율로 사전에 수록된 색인어들은 비교적 잘 추출하였으나, 정확률 측면에서 비교적 낮은 성능을 보였다. 어조사 어휘가 포함된 단어를 중복으로 추출하는 문제와 한글 독음변형에 있어 다양한 독음으로 읽히는 문제로 인해 색인어가 잘못 추출되었다. 이를 해결하기 위해 추후 연구를 통해 한자 어조사와 관련된 문법 규칙을 추가하고, 한자독음규칙에 따른 한글변형사전을 추가하여 시스템의 성능을 향상하고자 한다.

한국 근현대시기의 텍스트는 시기와 매체에 따라 어문법의 변화와 국한문 혼용의 비율이 다른 경향이 있다. 따라서 후속연구로『시사총보』에 비해 출판의 기간이 길고 다양한 형태의 글을 포함하고 있는 신문 매체인『황성신문』을 대상으로 추가적인 색인어 추출실험을 진행하여 색인어 추출법의 유용성과 어휘와 문법적 보완을 고려하고자 한다. 본 연구에서 처음 제안한 국한문 혼용체 대상의 색인어 추출법은 추후 한국 근현대 시기의 다양한 텍스트에 적용될 수 있을 것이며, 색인어 추출 이외에 텍스트 자동요약 및 분류 등 다양한 분석에 적용할 수 있을 것이다.

참 고 문 헌

강승식 (2002). 한국어 형태소 분석과 정보 검색. 서울: 홍릉과학출판사.

강승식, 권혁일, 김동렬 (1995). 한국어 자동 색인을 위한 형태소 분석 기능. 한국정보과학회 학술발표논문집, 22(1), 929-932.

김창섭 (2001). 한자어 형성과 고유어 문법의 제약. 국어학, 37, 177-195.

김창섭 (2013). '-的'의 두음 경음화와 2 자어 3 자어론. 국어학, 68, 167-188. https://doi.org/10.15811/jkl.2013..68.006

김판준 (2006). 디스크립터 프로파일을 사용한 통제어휘 자동색인. 한국정보관리학회 학술대회 논문집, 2006, 153-160.

박주희, 맹성현 (2017). 개념 그래프 구축을 위한 사전 및 구문 분할 기반 한국어 복합 명사 개념 경계 탐지 방법론. 한국정보과학회 학술발표논문집, 44(1), 651-653.

이상복 (2012). 국어학: 국어의 형태소 분석에 대한 일고찰 (1) - 고유명사를 중심으로. 배달말, 50, 1-35.

이익섭 (1969). 한자어의 비일음절 단일어에 대하여. 김재원 박사 회갑기념논총, 을유문화사, 837-844.

정영미 (2012). 정보검색연구. 서울: 연세대학교 대학출판문화원.

주지연 (2015). 한국 한자어의 형태소 분포 조사. 국어학, 76, 39-66.

주지연 (2017). 韓國語 漢字 形態素 개념 정립을 위한 시론(1) - 2字語 構成 要素의 形態素 지위를 중심으로. 어문연구, 45(1), 67-98.

홍성혁, 김철수, 이용석 (1996). 국·한문 혼용 문장의 형태소 분석을 위한 사전 구성. 한국정보과학회 학술발표논문집, 23(2A), 541-544.

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Chung, Yeong-Mi (2012). Research in information retrieval. Seoul: Yonsei University Press.

Hong, Seong-Hyeok, Kim, Cheol-Su, & Lee, Yong-Seok (1996). Construction of electronic dictionary for morphological analysis in hangul - Hanja mixed sentences. Proceedings of the 23th KISS Conference, 22(1), 23(2A), 541-544.

Ju, Ji-Yeon (2015). A research on the distribution of sino-korean morphemes. Journal of Korean Linguistics, 76, 39-66.

Ju, Ji-Yeon (2017). To establish the sino-korean morpheme concept (1) - Focusing on the

morphological status of 2syllable-word component. The Society for Korean Language & Literary Research, 45(1), 67-98.

Kang, Seung-Sik (2002). Korean morphological analysis and information retrieval. Seoul: Hongrung Publishing Company.

Kang, Seung-Sik, Kwon, Hyuk-Il, & Kim, Dong-Ryul (1995). The role of morphological analysis for korean automatic indexing. Proceedings of the 22th KISS Conference, 22(1), 929-932.

Kim, Chang-Sop (2001). Word-formation in sino-korean and a constraint of the grammar of native korean. Journal of Korean Linguistics, 37, 177-195.

Kim, Chang-Sop (2013). The tensification of initial consonant of the sino-korean suffix -jok(的) and the two-character word and three-character word theory. Journal of Korean Linguistics, 68, 167-188. https://doi.org/10.15811/jkl.2013..68.006

Kim, Pan-Jun (2006). Automatic indexing with controlled vocabulary using a descriptor profile. Proceedings of Korea Society for Information Management Conference, 2006, 153-160.

Lee, Ik-Seop (1969). Chinese non 1syllable-word, a collection of scholarly papers in celebration of the 60th anniversary of Dr. Jae Won Kim, Eulyoo Publishing, 837-844.

Lee, Sang-Bok (2012). A study on morphological analysis of Korean (1) - For proper nouns. Korean Languages, 50, 1-35.

Park, Ju-Hee, & Myaeng, Seong-Hyeon (2017). A method for establishing korean multi-word concept boundary harnessing dictionaries and sentence segmentation for constructing concept graph. Proceedings of the 44th KISS Conference, 44(1), 651-653.