텍스트마이닝을 위한 한국어 불용어 목록 연구

길 호 현
(서원대학교 교수)

〈 차 례 〉

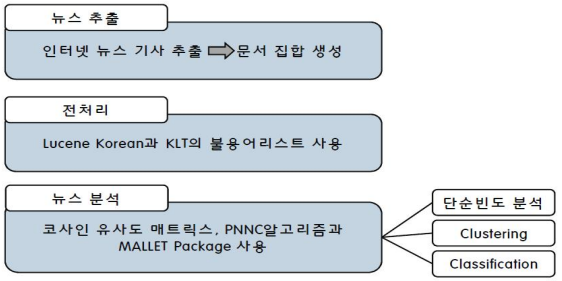
〈국문 요약〉

본 연구의 목적은 텍스트마이닝 방법을 활용하여 한국어 텍스트를 분석할 때 필요한 불용어 목록을 제시하는 것이다. 텍스트마이닝의 전처리 과정에서 불용어를 제거하는 작업이 수행되는데 이를 위한 불용어 목록이 필요하기 때문이다. 이를 위해 국립국어원에서 제시하는 대규모 말뭉치에서 최다 빈도로 출현하는 형태소를 추출하였다. 그리고 이 중에서 중요한 의미를 가지는 형태소와 의미가 없는 형식 형태소를 제외하였다. 결과적으로 실질 형태소이자 자립 형태소이면서 의미적으로는 유용하지 않은 293개의 단어가 불용어 목록으로 선정되었다. 이와 같은 불용어 목록은 다양한 분야에서 한국어 텍스트를 분석할 때 유용하게 활용될 수 있을 것으로 예상된다.

◎ 주제어: 텍스트마이닝, 불용어, 전처리, 말뭉치, 어휘

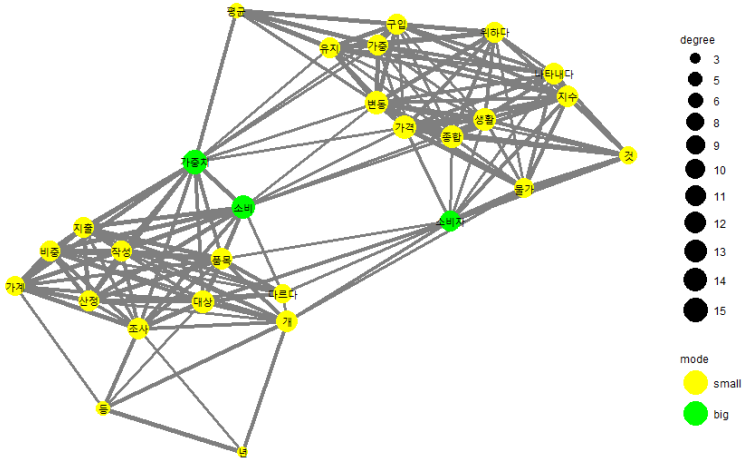
1. 연구의 목적 및 필요성

이 연구는 텍스트의 의미 구조를 텍스트마이닝 기법을 활용하여 분석하는 과정에서 필요한 한국어 불용어의 목록을 제시하는 것을 목적으로 한다. 텍스트마이닝은 대량의 자료에서 유용한 정보를 찾아내는 데이터마이닝의 일종으로, 수치 자료가 아니라 다양한 텍스트를 분석의 대상으로 한다. 그런데 직접적으로 분석이 가능한 숫자로 표시된 일반적인 정형 데이터와는 달리 텍스트는 다양한 형태의 변이형 자료를 포함하는 비정형 데이터이다. 텍스트마이닝은 이러한 텍스트 자료를 분석이 가능한 형태로 가공한 후 이를 바탕으로 정보를 추출하고 해석하게 된다. 이와 같은 텍스트마이닝을 수행하기 위해서는 텍스트를 분석이 가능한 형태로 가공하고 정리하는 작업을 반드시 수행해야 하는데 이를 텍스트 전처리 과정이라고 한다. 전처리를 마친 텍스트라고 해도 텍스트의 정보가 모두 정형화 된 수치 자료로 변경되는 것은 아니다. 그러나 언어가 갖는 비정형성이 어느 정도 감소되기 때문에 통계 처리 등 다양한 분석 방법을 적용할 수 있게 된다. <그림 1>은 텍스트마이닝을 활용하여 뉴스 기사를 분석하는 과정을 도식적으로 표시한 것이다. 분석의 목적과 방법에 따라 다양한 변용이 가능하지만 기본적으로 텍스트마이닝을 적용하는 과정은 이와 유사한 절차에 따라 이루어진다.



<그림 1> 텍스트마이닝 분석의 일반적인 흐름(감미아 외, 2012:56)

텍스트마이닝 분석을 위해서는 정확한 전처리 과정이 필수적이다. 텍스트마이닝 분석을 위한 전처리 과정은 숫자나 문장 부호 제거, 각종 약물이나 밑줄 제거, 누락/제거된 표현 복원, 오탈자 교정, 불용어 제거 등의 작업이 수행된다(이삼형 외, 2018:190). 그런데 다양한 전처리 과정 중 숫자/문자를 제거하거나 대/소문자를 통일하는 등 대부분의 작업은 특정한 함수 등을 이용해서 일괄적·기계적으로 처리할 수 있는 반면 불용어를 제거하는 작업은 자동화 수행이 불가능하다. 불용어란 매우 빈번하게 출현하지만 텍스트의 의미 구성에는 영향을 미치지 않는 불필요한 어휘들로, 이런 어휘들을 제거해야 비로소 텍스트의 의미를 구성하는 중요한 어휘들을 유효하게 산출할 수 있다. 그런데 불용어를 제거하기 위해서는 사전에 제거할 어휘의 목록이 미리 선정되어야 한다.



<그림 2> 전처리를 수행하지 않은 단일 텍스트의 핵심어 네트워크 구조

이러한 불용어 제거의 필요성은 <그림 2>에서 확인할 수 있다. <그림 2>는 텍스트마이닝 분석을 활용하여 한 편의 완결된 텍스트에 사용

된 어휘들의 연결 관계를 시각적으로 제시한 것이다. 그런데 전처리 과정에서 불용어 제거를 수행하지 않았다. 그 결과 '등', '년', '것', '개' 등 다수의 의존명사가 중요한 어휘로 산출되었으며, 다른 어휘들과도 긴밀한 연결 관계를 맺고 있는 것으로 나타났다. 그런데 실질적으로 이러한 의존명사들은 글의 의미 구조에 영향을 미치지 않기 때문에 중요한 어휘라고 볼 수 없으며 분석 대상으로서도 부적절하다. 즉, 불용어 제거가 적절하게 이루어지지 않으면 분석 결과를 왜곡할 우려가 있는 것이다.

그런데 영어 텍스트 분석에는 일반적으로 사용되는 불용어 목록이 존재하는 반면, 한국어 텍스트 분석에 사용되는 불용어 목록은 아직 연구가 충실하게 진행되지 않았다[1]. 최근 국어교육에서도 텍스트마이닝을 활용한 연구가 활발하게 이루어지고 있는데(노형남, 2014; 이수상, 2016; 이슬기·박영민, 2017; 이슬기, 2017 등) 이러한 텍스트마이닝 분석에서 불용어 제거는 반드시 수행되어야 하는 필수적인 과정임에도 불구하고 보편적으로 사용되고 있는 한국어 불용어 목록을 확인하기는 어렵다. 이에 본 연구에서는 텍스트마이닝 분석 방법을 사용하여 개별적인 한국어 텍스트의 의미 구조를 분석할 때 사용할 수 있는 범용 불용어 목록의 시안을 제시하는 것을 목적으로 한다.

2. 선행 연구 검토

Rajaraman, A. & Ullman, J. D.(2011)이 주장한 것처럼 어느 언어에나 통용되는 불용어 리스트는 존재하지 않을 것이다. 그러나 그렇다고 해서 각 언어별 텍스트를 분석할 때 불용어 목록을 활용하지 않을 수도 없다.

1) 아래 주소의 블로그를 방문하면 임의로 구성한 한국어 불용어 목록을 확인할 수 있다. 삭제해서는 안 될 어휘들이 불용어 목록에 포함되어 있음을 확인할 수 있다.
http://bab2min.tistory.com/544

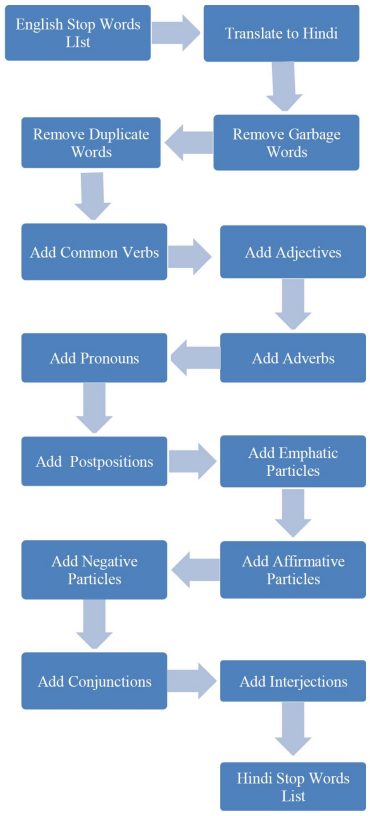


Figure 1 Details of stop word list construction for Hindi

<그림 3> 힌디어 불용어 목록 산출 과정 (Sifatullah Siddiqi & Aditi Sharan, 2018)

영어에서 일반적으로 많이 사용되는 불용어 목록은 1990년에 Fox에 의해 제시된 'A Stop List'이다. Fox는 영어에서 가장 널리 사용되는 말뭉치 중 하나인 Brown 코퍼스를 기반으로 고빈도어를 추출하고 이 목록을 조정하는 방식으로 표준 불용어 목록을 구성하였다. 이러한 영어 불용어 목록을 바탕으로 다양한 언어를 기반으로 불용어 리스트를 선정하려는 연구가 수행되었다. Sifatullah Siddiqi & Aditi Sharan(2018)은 <그림 3>과 같이 영어의 불용어 리스트를 힌디어로 번역한 후 힌디어에 사용되는지를 점검하여 맞지 않는 어휘를 제거하고 힌디어 분석에 필요한 어휘를 추가하는 방법으로 힌디어 불용어 목록을 제작하였다. Raulji & Saini(2017)은 유사한 방법으로 산스크리트 언어의 일반적 불용어 목록을 제시하였다. 그런데 영어와 한국어는 형태론적 특징이 다르기 때문에 이러한 번역 방식으로는 유의미한 결과를 얻기가 어렵다.

한국어의 처리를 위한 불용어 연구도 수행되었다. 불용어의 목록은 주로 검색 엔진에서 무의미한 색인어를 삭제하여 검색의 질을 높이기 위한 목적으로 연구되었다. 김판구 · 조유근(1993)에서는 자동 색인 구성을 위한 불용어 리스트를 구성하는 방안을 제시하였는데, 시스템에서 구문 분석을 시행한 후에 일반 불용어 리스트와 전문 영역 불용어 리스트의 이

중 불용어 리스트를 적용하면 색인어의 적합률이 상승하고 부적합률이 하락함을 보여 주었다. 이를 위해 일반 불용어 812개, 시사 및 경제 분야의 전문 불용어 1826개를 구성하여 제시하였다. 다만 이 연구에서는 <표 1>과 같이 불용어 목록의 예만 제시하고 있을 뿐 구체적인 불용어 목록의 선정 과정과 결과를 제시하지는 않고 있다. 또한 <표 1>에서도 알 수 있듯이 불용어 목록에 의미가 있는 어휘들이 다수 포함되어 있기 때문에 단일 텍스트의 의미 구조 분석을 위한 불용어 목록으로 활용하는 데에 한계가 있다.

<표 1> 일반 및 전문영역 불용어 리스트의 예(김판구·조유근, 1993:811)

일반 불용어 리스트의 예	전문 영역 불용어 리스트의 예(시사 분야)
가득, 가량, 가려, 가령, 가서, 가운, 가장, 가하, 각기, 각자, 각종, 간간, 간의, 갑자, 갑절, 강한, 거기, 거대, 거두, 거듭, 거의, 거치, 결국, 결코, 계속, 고로, 고루, 고사이, 고작, 관하, 관해 등등	가능, 가능성, 가세, 가시, 가중, 각국, 각급, 각당, 각도, 각지, 간격, 간단, 간소, 간접, 간주, 감소, 감수, 감안, 감퇴, 갑을, 강력, 강세, 강제, 강조, 개인, 거듭, 거래, 거북 등등

이 외에도 곽용진(2003)은 말뭉치의 자동 구축을 위해서 포함되어야 할 필수 어휘 목록과 배제해야 할 불용 어휘 목록을 제시하였다. 그중 불용 어휘 목록의 예는 <표 2>와 같다. 그런데 이 불용 어휘는 말뭉치에서 가급적 포함하지 말아야 할 어휘들의 목록이어서 저빈도어와 전문 용어, 외래어 등이 폭넓게 포함될 수밖에 없으므로 이 목록을 본 연구에서 활용하기는 어렵다. 다만 이 연구에서는 다수의 오탈자 등을 고려해야 함을 주장하였는데, 이는 향후 불용어 목록을 적용할 때 개별적으로 반드시 고려해야 할 사항으로 판단된다.

<표 2> 불용 어휘의 예(곽용진, 2003:50)

ID	Entry	Type(형태태그)
1	만만디	외국어
2	류쭝이	외국어
3	프랑클	외국어(고유명사)
4	헤스케	외국어(고유명사)
5	팔대상굉	외국어(고유명사)
6	라잇	외국어(한글전사)
7	오케이	외국어(한글전사)
8	DJ	외국어(이니셜)
9	YS	외국어(이니셜)
10	디릴람미더	방언
11	괘안을지	방언
12	지	방언
13	성새임	방언
14	사람이름	고유명사
15	나라이름	고유명사
16	지역이름	고유명사
17	고금군국지	고유명사
....	....	.....
	(영문)	패턴
	(한자)	패턴

또한 감미아·송민(2012)에서는 Java로 구축된 Lucene Korean 모듈에 KLT(Korean Language Technology)의 불용어 목록을 더하고, 분석 중에 발견된 불용어를 추가하는 방식으로 전처리를 수행하였다. 다만, 정확한 목록과 특히 추가한 불용어 목록은 구체적으로 제시하지 않았다. 한편 안정국·김희웅(2015)에서는 감정어 사전 구축을 위해서 한국어 어휘를 26개의 품사로 나누고 품사별 빈도를 산출한 후 이에 대한 감정 표현 여부를 판단하여 <표 3>과 같은 분류표를 제시하였다. 이는 품사의 분류이므로 불용어 목록으로 활용할 수는 없지만 여기에 사용된 품사의 분류는 불용어 목록 작성을 위한 어휘 분류 기준을 작성하는 데 참조할 수 있다.

<표 3> 단어의 분류 및 감정 표현 여부(안정국·김희웅, 2015: 55)

〈Table 3〉 Classification of Words

Word type	Total	Sentiment	Word type	Total	Sentiment
interjection	682	N	assistant verb	14	N
interjection·noun	85	N	assistant adjective	17	N
determiner	207	N	adverb	17,425	Y
determiner·interjection	11	N	adverb·interjection	3	N
determiner·noun	1,267	Y	numeral	60	N
pronoun	382	N	numeral·determiner	195	N
pronoun·interjection	5	N	numeral·determiner·noun	3	N
pronoun·determiner	3	N	ending	6	N
pronoun·adverb	1	N	bound noun	913	N
verb	68,370	Y	bound noun·postposition	13	N
verb·adjective	2	Y	affix	209	N
noun	337,659	Y	postposition	300	N
noun·adverb	109	N	adjective	16,562	Y
Total words	517,178		Total sentiment words	441,283	

3. 불용어의 선정 기준과 분석 방법

1) 텍스트 마이닝을 위한 전처리 과정

텍스트 전처리 과정은 글의 분석에 직접적으로 영향을 미치지 않는 요소들을 정리하는 작업으로, 다음과 같은 일련의 단계가 수행된다[2].

① 공란 처리

일반적으로 글에서 빈칸은 단어와 단어를 구분하며 한국어에서는 어절 구분의 단위가 된다. 그런데 오타나 기타의 이유로 공란이 2개 이상 연달아 발견되거나 탭 등으로 공란이 생기면 단어나 문장의 입력에 문제가 생기게 된다. 이를 방지하기 위해 글에 사용된 공란을 일괄적으로 1

2) 일반적인 전처리 과정에 대한 설명은 백영민(2017:97~113)의 내용을 요약하여 정리하면서 연구자의 견해를 추가하였다.

칸의 스페이스 공란으로 치환하는 작업을 할 필요가 있다.

② 대·소문자 통일

영문의 경우 문장의 첫 단어의 첫 문자, 고유명사의 첫 문자, 축약어 등에 대문자를 사용한다. 그런데 텍스트 처리를 할 때 대·소문자는 각기 다른 문자로 인식되므로 문장의 첫 머리에 대문자가 사용된 단어는 문장 중간의 소문자 단어와 각기 다른 단어로 처리된다. 이를 방지하기 위해 대문자로 표기된 것을 일괄적으로 소문자로 변환하는 작업이 필요하다. 물론 기계적으로 이러한 과정을 수행할 경우 고유명사 등을 삭제하게 될 수도 있으므로 작업 예외 대상을 세심하게 조정할 필요가 있다.

③ 숫자 표현 제거

글에는 문자 외에도 숫자로 표현된 자료도 포함될 수 있다. 일반적으로 숫자로 표현된 자료는 삭제하거나 모든 숫자를 하나로 통합하는 방식으로 전처리를 한다. 숫자로 나타낸 정보 자체가 글의 내용을 이해하는 데 큰 영향을 미치지는 않기 때문이다. 그러나 순위나 수량 등 수치 정보 자체가 중요할 경우 숫자를 수사의 형태로 변환하여 처리하는 것이 가능하다.

④ 문장부호 및 특수문자 제거

글에는 다양한 문장부호 및 특수문자가 사용되며, 각 문장부호는 문법적 또는 의미론적으로 중요한 기능을 수행한다. 예를 들어, 마침표는 문장을 구분하는 역할을 하는데 문장 단위로 글을 구분하는 것은 글의 내용을 데이터로 입력하거나 의미 구조를 분석할 때 매우 중요한 요인이 된다. 또한 약어나 'et al.' 등에 사용된 마침표는 의미적으로 전혀 다른 기능을 수행하게 된다. 따라서 문장부호 및 특수문자 역시 일반적으로 제거하는 작업을 수행하지만 이 작업의 수행 순서를 면밀하게 조정할 필요가 있다.

⑤ 불용어 제거

불용어(stopwords)는 영어의 a, an, the 등과 같이 빈번하게 사용되나 구체적인 의미를 찾기 어려운 단어들을 의미한다. 이러한 단어들은 글에서 사용되는 빈도가 높은 반면 글의 의미 구조 분석에는 큰 영향을 미치지 않기 때문에 사전에 삭제할 필요가 있다. 불용단어의 목록은 분석자가 직접 구축할 수도 있고 이미 작성되어 있는 목록을 활용할 수도 있다.

⑥ 어근 동일화 처리

동일한 단어라고 해도 문법적 기능에 따라 표현이 바뀌는 경우가 있다. 어근 동일화(stemming)는 파생된 형태의 단어를 동일하게 처리할 수 있도록 체계적인 방식으로 표현을 변환시키는 과정을 의미한다. 영어의 경우 마틴 포터(Martin Porter)의 어근 동일화 알고리즘[3]을 텍스트마이닝 분석에 활용되는 R프로그램의 tm 패키지에서 제공하고 있다.

⑦ 엔그램 적용

엔그램(*n*-gram)이란 'Republic of Korea'처럼 *n*번 연이어 등장하는 단어들의 연쇄를 의미한다. 이러한 단어들은 하나의 단어로 처리하는 것이 자연스럽지만 언제나 그런 것은 아니므로 역시 상황에 따라 다르게 적용할 필요가 있다.

이 외에도 영어 텍스트 분석의 경우 한 글자 단어를 제외하는 등의 작업을 추가로 진행하기도 한다. 다음으로는 이러한 전처리 과정 중 불용어의 목록을 선정하는 방법에 대해서 살펴보겠다.

3) 흔히 '포터의 스테머(Porter's Stemmer)'라고 불린다.

2) 영어 불용어의 선정 방법 및 불용어 목록

Fox(1990)에서는 불용어 리스트를 작성하는 일반적인 작업 과정을 제시하고 있는데 간략히 정리하면 다음과 같다.

① 대량의 문서들로부터 어휘 추출: Brown 코퍼스에서 300회 이상 출현한 단어 278개 선정

② 중요 어휘 제거: 앞의 목록에서 불용처리를 하면 안 되는 중요 단어 32개 제외

③ 중·저빈도 어휘 중 불용어 추가: 앞의 목록에 전통적인 불용어휘 26개 추가

④ 기타 어휘 추가: 한 글자, 특정 접두사와 결합된 단어 등 기타 단어 149개 추가

이와 같은 과정을 거쳐 Fox(1990)에서는 총 421개의 불용어 목록을 제시하였다. 이 외에도 영어 텍스트 분석을 위해 다양한 불용어 목록이 사용되고 있다. Java로 구현된 정보 검색 오픈 소스 라이브러리인 Lucene에서는 색인을 위한 불용어 목록을 제공하고 있는데 <표 4>와 같이 34개밖에 되지 않는다. 또한 이 목록은 영어에서 매우 자주 사용되는 어휘들이기는 하지만 34개의 한정된 어휘 내에 be 동사의 다양한 형태가 포함되어 있는 등 충분히 정련되어 있는 목록은 아니라고 판단된다.

<표 4> Lucene의 불용어 리스트

'a', 'and', 'are', 'as', 'at', 'be', 'but', 'by', 'for', 'if', 'in', 'into', 'is', 'it', 'no', 'not', 'of', 'on', 'or', 's', 'such', 't', 'that', 'the', 'their', 'then', 'there', 'these', 'they', 'this', 'to', 'was', 'will', 'with'

최근 텍스트마이닝 분석을 위해 빈번하게 활용되는 R의 경우 텍스트 분석을 위한 tm 패키지에 <그림 4>와 같이 불용어 리스트를 제공하고 있다. <그림 4>에서도 확인할 수 있듯이 불용어 목록에는 대명사와 전치사, 관사, 부사 등이 포함되어 있으며, be 동사는 원형이 아니라 변이형이 모두 포함되어 있음을 알 수 있다. 또한 다수의 축약어 형태도 제시되어 있다. 이 목록은 Catalan stopwords와 Romanian stopwords를 활용했다고 하는데 해당 목록이 어떤 기준으로 어떻게 선정되었는지에 대해서는 추가적인 설명이 되어 있지 않다[4].

```
> stopwords("en")
  [1] "i"          "me"         "my"         "myself"     "we"         "our"        "ours"       "ourselves"  "you"        "your"
 [11] "yours"      "yourself"   "yourselves" "he"         "him"        "his"        "himself"    "she"        "her"        "hers"
 [21] "herself"    "it"         "its"        "itself"     "they"       "them"       "their"      "theirs"     "themselves" "what"
 [31] "which"      "who"        "whom"       "this"       "that"       "these"      "those"      "am"         "is"         "are"
 [41] "was"        "were"       "be"         "been"       "being"      "have"       "has"        "had"        "having"     "do"
 [51] "does"       "did"        "doing"      "would"      "should"     "could"      "ought"      "i'm"        "you're"     "he's"
 [61] "she's"      "it's"       "we're"      "they're"    "i've"       "you've"     "we've"      "they've"    "i'd"        "you'd"
 [71] "he'd"       "she'd"      "we'd"       "they'd"     "i'll"       "you'll"     "he'll"      "she'll"     "we'll"      "they'll"
 [81] "isn't"      "aren't"     "wasn't"     "weren't"    "hasn't"     "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
 [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"      "cannot"     "couldn't"   "mustn't"    "let's"      "that's"
[101] "who's"      "what's"     "here's"     "there's"    "when's"     "where's"    "why's"      "how's"      "a"          "an"
[111] "the"        "and"        "but"        "if"         "or"         "because"    "as"         "until"      "while"      "of"
[121] "at"         "by"         "for"        "with"       "about"      "against"    "between"    "into"       "through"    "during"
[131] "before"     "after"      "above"      "below"      "to"         "from"       "up"         "down"       "in"         "out"
[141] "on"         "off"        "over"       "under"      "again"      "further"    "then"       "once"       "here"       "there"
[151] "when"       "where"      "why"        "how"        "all"        "any"        "both"       "each"       "few"        "more"
[161] "most"       "other"      "some"       "such"       "no"         "nor"        "not"        "only"       "own"        "same"
[171] "so"         "than"       "too"        "very"
```

<그림 4> tm 패키지의 stopwords list

이 외에도 오픈소스 관계형 데이터베이스 프로그램인 MySQL에서도 불용어 목록을 제공하고 있다[5]. 이처럼 영어에서 불용어 목록은 다양하

4) tm 패키지에 대한 보다 자세한 설명은 아래 주소의 자료에서 참조가 가능하다.
https://cran.r-project.org/web/packages/tm/tm.pdf

5) 구체적인 목록은 아래 주소에서 확인할 수 있다.
http://xpo6.com/list-of-english-stop-words/

게 제시되어 있으며 상황에나 목적에 따라 다르게 적용이 가능하다.

3) 한국어 불용어의 선정 기준 및 분석 방법

한국어 텍스트 분석을 위한 불용어 목록은 아직까지 일반적으로 사용되고 있는 것이 없다. 특정 사이트[6]에서 한국어 불용어 목록을 포함하여 세계 각국 언어의 불용어 목록을 제시하고 있기는 하지만 해당 목록의 타당성을 검증하기는 어려운 상태이다. 물론 불용어 목록은 구체적인 분석의 과정 중에 활용이 되기 때문에 어느 목적에나 범용적으로 사용할 수 있는 일반적인 목록을 구성한다는 것은 불가능할 수도 있다. 효율적인 검색을 위해서 제거해야 할 어휘의 목록과 글의 내용 구조를 파악하기 위해서 제거해야 할 어휘의 목록은 다를 수밖에 없다. 이에 본 연구에서는 한 편의 단일한 글의 내용 구조를 도식화하는 텍스트마이닝 분석 중에 사용할 수 있는 불용어 목록을 제시하는 것을 목표로 한다.

이를 위해 본 연구에서는 Fox(1990)에서 사용한 불용어 선정 과정을 한국어에 맞게 변형하여 적용하고자 한다. 연구의 절차는 다음과 같다.

<그림 5> 한국어 불용어 목록 선정 절차

일반적으로 색인을 위해 언어학적 분석 기법을 활용하여 색인어를 추출하는지 어휘의 출현 빈도에 따라 색인어를 추출하는지에 따라 불용어 목록을 구성하는 방식이 달라진다(김판구·조유근, 1993). 이는 본 연구에서 텍스트마이닝을 적용하는 방법에도 동일하게 적용이 가능하다. 분석 대상이 되는 글의 형태소 분석을 실시하여 글을 형태소 단위로 분해하고 이 중 의미 구성에 기여하지 않는 형태소의 종류를 지정하여 일괄적으로 제거할 수 있는 반면, 글에 사용된 모든 어휘에 대해 어근 동일화 처리(stemming)를 먼저 수행하고 불용어 목록은 기본형으로 처리하는 것도 가능하다. R 프로그램의 KoNLP(Korean Natural Language Processing)[7] 패키지에는 품사를 분류하는 방식에 따라 SimplePos09 함수와 SimplePos22 함수를 사용할 수 있는데 SimplePos09에서는 KAIST 품사 태그 집합의 54개 품사 분류 중 대분류 9개 품사로, SimplePos22에서는 중분류 22개 품사로 나누어 각 형태소를 분석한다. KoNLP 패키지를 이용하여 형태소 분석을 먼저 시행하고 불용어를 제거한 후 어근 동일화 작업을 수행하는 순서로 텍스트마이닝을 진행한다면 불용어 목록이 독립된 단어가 아니라 형태소 단위로 설정되어야 한다. 이럴 경우 불용어의 목록이 지나치게 많아질 뿐 아니라, 이어지는 어근 동일화 작업을 수행함에 있어서도 추가로 많은 작업을 수행해야 한다. 이러한 문제를 해결하기 위해서는 어근 동일화 작업을 불용어 삭제 처리 이전에 수행할 수 있다. 예를 들어, R의 NLP4kec 패키지[8]는 텍스트를 입력하는 단계에서 모든 어휘를 기본형으로 바꾸어 인식한다. 즉 어근 동일화 작업을 텍스트 입력 단계에서 수행하는 것이다. 따라서 불용어의 목록 역시 어근 동등화 작업을 한 이후에 적용할 것을 고려하여 자립 형태소의 형태로 한정하였다.

일반적으로 색인어 선별을 위해 불용어 목록을 작성하는 경우 고빈도 어휘를 중심으로 추출하게 된다. 이는 영어에서 빈도가 높은 단어들이

6) 각 언어별 불용어 목록은 아래 사이트에서 확인할 수 있다.
https://www.ranks.nl/stopwords

7) Jeon, H.(2013). KoNLP: Korean NLP Package. R Package Version 0.80.1. http://CRAN.R-project.org/package=KoNLP

8) NLP4kec는 한글 형태소분석기인 '은전한닢'을 기반으로 개발되었으며 엑셀로 입력된 다량의 문서를 분석하기에 용이하다. 패키지 및 사용법은 아래 주소에서 받을 수 있다.
https://github.com/NamyounKim/NLP4kec

대부분 관사나 전치사, 접속사, be 동사 등 색인어로서는 큰 가치가 없는 것들이기 때문이다. 이처럼 고빈도 어휘를 불용어로 간주하는 방법은 색인 DB의 크기를 줄임으로써 검색 속도를 개선하는 데 매우 효율적이다. 그런데 영어는 굴절어인데 비해 한국어는 교착어이기 때문에 한글 문서에서 색인어와 불용어를 선별하는 방법 또한 한국어의 형태론적 특성이 충분히 고려되어야 한다(강승식, 2004).

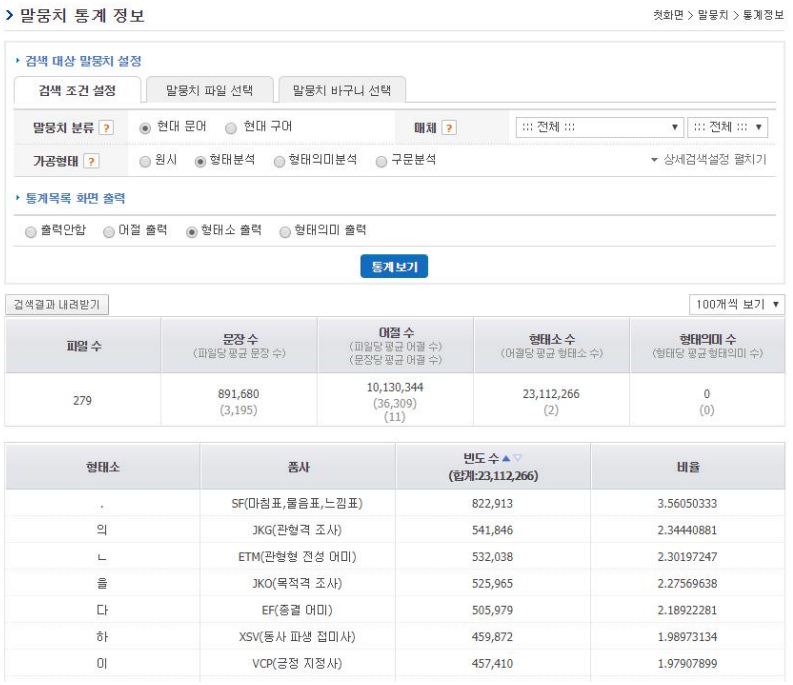
4. 한국어 텍스트마이닝을 위한 표준 불용어 목록

1) 말뭉치를 통한 고빈도 어휘 목록 추출

불용어 선정을 위해 먼저 대규모의 말뭉치에서 빈도수가 높은 어휘를 선정할 필요가 있다. 이를 위해 국립국어원에서 제공하는 말뭉치 통계 정보를 활용하여 빈도수가 높은 어휘들을 추출하였다. 국립국어원에서 제공하는 말뭉치는 21세기 세종계획 구문분석 말뭉치로 출발하는데, 이는 2002년부터 2006년까지 약 80만 어절, 7만 3천 문장 규모로 구축되어 구축 당시 한국어 구문분석 말뭉치 중 가장 큰 규모였으며(홍정하 외, 2008:286) 그 이후로도 국립국어원에서 공식적으로 제공하는 대규모 말뭉치이므로 표준 불용어 선정을 위한 분석 대상으로 적합하다.

분석을 위한 말뭉치의 분류는 현대 문어로 한정하였다. 이는 본 연구에서 주 연구 대상으로 삼는 텍스트 내용 구조 분이 주로 문어 텍스트를 대상으로 한다는 점을 고려하였다. 대신 매체는 특정 매체로 한정하지 않고 전체를 분석 대상으로 하였다. 가공형태는 '형태분석'을 선택하였다. 불용어 목록은 최종적으로 어휘의 목록 형태로 정리될 것이므로 원시 형태인 어절 단위가 아니라 형태소 분석을 통해 단어 수준으로 대상을 정리할 필요가 있기 때문이다. 다만 여기에서 실제로 사용된 의미 구조를 파악할 필요까지는 없으므로 '형태의미분석'이나 '구문분석'을 진행할 필

요는 없을 것으로 판단된다. 결과 목록도 품사를 확인할 수 있는 '형태소 출력'을 선택하였다.



<그림 6> 말뭉치 통계 정보 화면

이러한 조건으로 분석한 결과 말뭉치 파일 중 형태소 분석이 되어 있는 총 279개의 파일에서 빈도에 따른 형태소 목록을 산출하였다. 분석 대상이 되는 문장 수는 891,680개, 어절 수는 10,130,344개, 형태소 수는 23,112,266개였다. 이중 빈도수를 기준으로 상위 10,000개의 형태소를 엑셀 형태로 출력하여 1차 분석 대상 어휘 목록으로 설정하였다. 총 279개의 파일에서 선정된 10,000개의 형태소 중 최소 빈도는 115회이다. Fox에서 최소 빈도를 300으로 설정한 것을 고려한다면 그보다 약간 넓은 범

위를 취했다고 할 수 있으므로, 향후 이 부분을 조정하여 적정한 어휘의 수를 도출할 필요가 있을 것으로 판단된다.

2) 문법적 기능을 수행하는 품사 삭제

<그림 6>에서는 말뭉치를 통한 검색 결과의 일부를 확인할 수 있는데 빈도 수가 상위인 형태소들은 대부분 문법적인 관계를 나타내는 기능을 하고 있음을 발견할 수 있다. 이러한 형태소들은 텍스트의 의미를 구성하는 데에는 중요한 역할을 수행하지 못한다. 다만 이러한 형태소 가운데 의존 형태소들은 독립적으로 사용되지 못하므로 독립된 단어로 보기 어렵고, 실제로 추후 어근 동일화 작업에서 대부분 제거되므로 불용어 목록에 등재할 필요가 없다. 따라서 최빈 형태소 가운데 형식형태소에 해당하는 품사들을 목록에서 제거하여 목록을 간소화 할 필요가 있다. 실제로 최빈 어휘 10,000개의 목록 중에 품사 태그 분류는 총 42개가 등장하였으며, 이중 문법적 기능을 수행하는 형태소로 <표 5>의 목록에 해당하는 637개를 제거하였다.

<표 5> 문법적 기능을 수행하는 품사 태그와 형태소 수

품사 태그	형태소 수
EC(연결 어미)	182
EF(종결 어미)	119
EP(선어말 어미)	8
ETM(관형형 전성 어미)	25
ETN(명사형 전성 어미)	5
JC(접속 조사)	7
JKB(부사격조사)	34
JKC(보격 조사)	2
JKG(관형격 조사)	1
JKO(목적격 조사)	3
JKQ(인용격 조사)	4
JKS(주격 조사)	5
JKV(호격 조사)	4
JX(보조사)	34
SE(줄임표)	2
SF(마침표, 물음표, 느낌표)	3
SO(붙임표)	5
SP(쉼표, 가운뎃점, 콜론, 빗금)	8
SS(따옴표, 괄호표, 줄표)	31
SW(기타 기호)	41
VX(보조 용언)	29
XPN(체언 접두사)	30
XSA(형용사 파생 접미사)	6
XSN(명사 파생 접미사)	44
XSV(동사 파생 접미사)	5
25개	637

3) 의미를 구성하는 핵심 품사 삭제

이 연구에서 대상으로 하고자 하는 텍스트마이닝은 개별 텍스트에서 핵심적인 어휘들을 추출하여 텍스트의 의미 구조를 시각적으로 제시하는

과정을 수행한다. 이 과정은 기본적으로 텍스트에 사용된 어휘들의 빈도가 매우 중요할 수밖에 없으므로, 텍스트의 의미 구조를 형성하는 데 기여하는 핵심 어휘들의 빈도가 중요하게 고려되며 이 외에 문법적 기능을 수행하는 어휘는 분석에서 제외되도록 해야 한다. 따라서 불용어 리스트에는 실질적으로 의미를 가지면서 자립적으로 사용될 수 있는 형태소들은 포함되어서는 안 된다. 의미적으로 중요한 단어가 불용어로 처리되어 제거되어서는 안 되기 때문이다. 불용어 목록에서 제외되는 품사 태그와 형태소 수는 <표 6>과 같다.

<표 6> 의미를 구성하는 핵심 품사 태그와 형태소 수

품사 태그	형태소 수
MAG(일반 부사)	526
NNG(일반 명사)	6,089
NNP(고유 명사)	495
NR(수사)	46
SH(한자)	15
SL(외국어)	81
SN(숫자)	174
VA(형용사)	245
VV(동사)	1,109
XR(어근)	259
10개	9,039

4) 목록 조정 및 정교화

최빈 10,000개 형태소 중 <표 5>와 <표 6>에서 제거된 형태소를 제외하고 남은 형태소는 <표 7>과 같다. 결국 남아있는 어휘는 의미적으로 중요하지 않으면서도 실질 형태소이자 자립 형태소의 성격을 동시에 갖추고 있다는 특징이 있다.

<표 7> 불용어 대상 어휘

품사 태그	형태소 수	형태소 목록
IC (감탄사)	33	그래, 아니, 아, 뭐, 응, 네, 예, 자, 야, 글쎄, 참, 어디, 그럼, 아아, 애, 임마, 아이고, 여보, 어, 저, 원, 아이구, 음, 글쎄요, 아냐, 어머, 오, 흥, 아이, 에이, 허허, 아니야, 여보세요
MAJ (접속 부사)	37	그러나, 그리고, 그런데, 그래서, 따라서, 하지만, 또, 또는, 그러므로, 및, 또한, 그러면, 즉, 그럼, 그러니까, 그렇지만, 오히려, 역시, 그리하여, 다만, 혹은, 그래도, 한편, 이른바, 더구나, 왜냐하면, 근데, 그러자, 더욱이, 하긴, 하기야, 그러면서, 하물며, 그러니, 그러다가, 단, 이리하여
MM (관형사)	64	그, 이, 한, 두, 다른, 그런, 이런, 어떤, 모든, 어느, 몇, 여러, 무슨, 세, 전, 저, 각, 첫, 새, 아무, 약, 네, 아무런, 총, 제, 온, 옛, 오랜, 단, 올, 온갖, 별, 현, 한두, 맨, 양, 몇몇, 수, 딴, 서너, 저런, 두어, 모, 주, 석, 스무, 여느, 이런저런, 본, 동, 웬, 헌, 순, 왠, 요, 두세, 지난, 근, 타, 그까짓, 매, 고, 갖은, 일대
NNB (의존 명사)	136	것, 수, 등, 년, 때문, 일, 중, 월, 씨, 데, 번, 명, 원, 개, 거, 가지, 뿐, 듯, 간, 쪽, 분, 시, 채, 만, 대, 년대, 줄, 놈, 적, 터, 만큼, 바, 측, 내, 편, 차, 자, 세, 대로, 점, 달러, 살, 초, 식, 외, 셈, 듯이, 지, 회, 장, 호, 개월, 말, 따위, 리, 마리, 위, 녀석, 평, 무렵, 나름, 마련, 권, 척, 양, 법, 이, 바람, 쯤, 건, 판, 지경, 도, 달, 군, 조, 주일, 뻔, 시간, 부, 이래, 등등, 주, 개국, 푼, 군데, 체, 거리, 년도, 따름, 주년, 격, 등지, 통, 겸, 설, 참, 둥, 퍼센트, 남짓, 미터, 기, 나위, 즈음, 마당, 바퀴, 여지, 치, 벌, 그루, 박, 투, 승, 김, 가량, 큰술, 엔, 모금, 발짝, 석, 겨를, 전, 개소, 동, 톤, 개년, 한, 짝, 킬로미터, 집, 세기, 섬, 냥, 턱, 되, 발
NP (대명사)	52	나, 그, 우리, 이, 그것, 그녀, 내, 자기, 무엇, 이것, 누구, 저, 여기, 어디, 너, 뭐, 당신, 거기, 이곳, 그곳, 제, 아무, 네, 자네, 언제, 여러분, 이거, 너희, 니, 저희, 아무개, 이놈, 그놈, 저기, 그분, 그대, 그거, 모, 저쪽, 뭣, 저것, 그이, 이쪽, 그쪽, 지, 애, 걔, 저편, 저놈, 네놈, 그네, 쇤네
VCN (부정 지정사)	1	아니
VCP (긍정 지정사)	1	이
7개	324	

<표 7>의 목록 가운데 중복이 되는 어휘들을 제거하여 <표 8>과 같이 텍스트 의미 구조 분석을 위한 불용어 목록 시안을 정리하였다. 중복으로 인해 제거된 어휘는 ‘그’, ‘그럼’, ‘내’, ‘네(2)’, ‘단’, ‘동’, ‘모’, ‘뭐’, ‘석’, ‘세’, ‘수’, ‘아니’, ‘아무’, ‘양’, ‘애’, ‘어’, ‘어디’, ‘원’, ‘이(3)’, ‘자’, ‘저(2)’, ‘전’, ‘제’, ‘주’, ‘지’, ‘참’, ‘한’ 등 31개이며, 시안으로 제시된 불용어 목록은 총 293개이다.

<표 8> 의미 구조 분석을 위한 표준 불용어 목록 시안

가량, 가지, 각, 간, 갖은, 개, 개국, 개년, 개소, 개월, 걔, 거, 거기, 거리, 건, 것, 겨를, 격, 겸, 고, 군, 군데, 권, 그, 그거, 그것, 그곳, 그까짓, 그네, 그녀, 그놈, 그대, 그래, 그래도, 그서, 그러나, 그러니, 그러니까, 그러다가, 그러면, 그러면서, 그러므로, 그러자, 그런, 그런데, 그럼, 그렇지만, 그루, 그리고, 그리하여, 그분, 그이, 그쪽, 근, 근데, 글쎄, 글쎄요, 기, 김, 나, 나름, 나위, 남짓, 내, 냥, 너, 너희, 네, 네놈, 녀석, 년, 년대, 년도, 놈, 누구, 니, 다른, 다만, 단, 달, 달러, 당신, 대, 대로, 더구나, 더욱이, 데, 도, 동, 되, 두, 두세, 두어, 둥, 듯, 듯이, 등, 등등, 등지, 따라서, 따름, 따위, 딴, 때문, 또, 또는, 또한, 리, 마당, 마련, 마리, 만, 만큼, 말, 매, 맨, 명, 몇, 몇몇, 모, 모금, 모든, 무렵, 무슨, 무엇, 뭐, 뭣, 미터, 및, 바, 바람, 바퀴, 박, 발, 발짝, 번, 벌, 법, 별, 본, 부, 분, 뻔, 뿐, 살, 새, 서너, 석, 설, 섬, 세, 세기, 셈, 쇤네, 수, 순, 스무, 승, 시, 시간, 식, 씨, 아, 아냐, 아니, 아니야, 아무, 아무개, 아무런, 아아, 아이, 아이고, 아이구, 야, 약, 양, 애, 어, 어느, 어디, 어머, 언제, 에이, 엔, 여기, 여느, 여러, 여러분, 여보, 여보세요, 여지, 역시, 예, 옛, 오, 오랜, 오히려, 온, 온갖, 올, 왜냐하면, 왠, 외, 요, 우리, 원, 월, 웬, 위, 음, 응, 이, 이거, 이것, 이곳, 이놈, 이래, 이런, 이런저런, 이른바, 이리하여, 이쪽, 일, 일대, 임마, 자, 자기, 자네, 장, 저, 저것, 저기, 저놈, 저런, 저쪽, 저편, 저희, 적, 전, 점, 제, 조, 주, 주년, 주일, 줄, 중, 즈음, 즉, 지, 지경, 지난, 집, 짝, 쪽, 쯤, 차, 참, 채, 척, 첫, 체, 초, 총, 측, 치, 큰, 킬로미터, 타, 터, 턱, 톤, 통, 투, 판, 퍼센트, 편, 평, 푼, 하기야, 하긴, 하물며, 하지만, 한, 한두, 한편, 허허, 헌, 현, 호, 혹은, 회, 흥

5. 논의 및 결론

이 연구에서는 텍스트마이닝 기법을 활용하여 글의 의미 구조를 분석하는 과정에서 텍스트 전처리에 필요한 불용어 목록을 제시하는 것을 목적으로 하였다. 이를 위해 대규모 말뭉치 자료에서 빈도순으로 형태소 10,000개를 추출하고, 텍스트의 의미 구성에 중요한 역할을 하는 독립된 단어들과 의미 구성에 기여하지 않는 형식 형태소 등을 제거하였다. 그 결과 빈도는 높지만 텍스트의 의미 구성에는 기여하지 않는 단어들을 추출할 수 있었는데 이를 텍스트마이닝을 위한 불용어 목록 시안으로 제시하였다.

연구 결과로 제시된 불용어 목록 시안은 기존의 검색 서비스를 위한 불용어와는 목적이 다르다. 뉴스나 색인을 위한 불용어는 영역에 따라 널리 사용되는 용어가 다르기 때문에 영역별로 불용어 목록을 선정할 필요가 있다. 그러나 국어교육 등의 연구 분야에서 텍스트의 연결 관계나 의미 구조를 분석하기 위한 텍스트마이닝 분석에서는 이러한 용어들이 중요한 의미를 가질 수 있기 때문에 함부로 제거되어서는 안 된다. 따라서 실질 형태소이자 자립 형태소이면서 빈도수가 높지만 의미상으로는 중요성이 떨어지는 단어들의 목록이 불용어 목록으로 제시되면 다양한 텍스트마이닝 분석 연구에 유용하게 활용할 수 있을 것으로 예상된다.

다만 이 연구에서 제시하는 불용어 목록은 1차적인 시안이기 때문에 후속 연구를 통해 정교화 해야 한다. 예를 들어, 대명사의 경우 글에서 중요한 내용 구성 기능을 수행하는 경우가 적지 않으므로 이를 불용어 목록에 수록해야 하는지는 더 깊은 논의가 필요하다. 또한, 단위어의 경우 한국어 표기와 외국어 표기가 모두 가능한데, 품사 태그는 각기 다르게 부여된다. 이처럼 각 품사 태그별로 선정되거나 제거되는 형태소들을 세세하게 점검할 필요가 있다. 말뭉치의 목록도 점검 및 조정이 필요하다. <표 8>에 보면 ‘쇤네’ 등의 단어가 보이는데 이는 상위 빈도어로 보기 어려운 측면이 있다. 이는 선정된 말뭉치 중 아동 도서나 문학 작품 등이 포함되어 있기 때문으로, 말뭉치의 목록 또한 보다 실제성을 확보할 수 있는 방향으로 점검할 필요가 있을 것이다.

이처럼 아직 보완해야 할 점이 많지만 이 연구는 텍스트마이닝을 위한 범용 불용어 목록의 시안을 시론적으로 제시하였다는 점에서 연구의 의의를 찾을 수 있다.

참고 문헌

감미아·송민(2012), 텍스트마이닝을 활용한 신문사에 따른 내용 및 논조 차이점 분석,『지능정보연구』18권 3호, 한국지능정보시스템학회, 53~77쪽.

강승식(2004), 한글 문서의 색인어와 색인 기법,『정보과학회지』22권 4호, 한국정보과학회, 72~77쪽.

곽용진(2003), 합목적적 말뭉치(Corpus) 자동 구축,『언어사실과 관점』13권, 연세대 언어정보연구원, 31~68쪽.

김판구·조유근(1993), 한국어 정보 검색을 위한 불용어의 구성 및 적용,『한국정보과학회 학술발표논문집』20권 1호, 한국정보과학회.

노형남(2014), 빅 데이터 텍스트 마이닝 - 정치 연설을 중심으로 -,『화법연구』26, 한국화법학회, 289~325쪽.

백영민(2017),『R를 이용한 텍스트 마이닝』, 한울아카데미.

안정국·김희웅(2015), 집단지성을 이용한 한글 감성어 사전 구축,『지능정보연구』21권 2호, 한국지능정보시스템학회, 809~812쪽.

이삼형·길호현(2018), 텍스트 난이도에 따른 핵심 어휘와 그 관계망 변화에 대한 연구,『우리말연구』53집, 우리말학회, 179~221쪽.

이수상(2016), 독후감 텍스트의 언어 네트워크 분석에 관한 기초연구,『한국도서관정보학회지』47(3), 한국도서관 정보학회, 95~114쪽.

이슬기(2017), 작문 평가에서 텍스트 마이닝의 활용 가능성 탐색,『작문연구』35, 한국작문학회, 99~131쪽.

이슬기·박영민(2017), 쓰기 수행 수준에 따른 중학생 논설문의 텍스트 시각화 분석,『학습자중심 교과교육 연구』17(15), 학습자중심 교과교육학회, 401~422쪽.

홍정하·김주영·강범모(2008), 세종 구문분석 말뭉치의 구축과 통사 범주 및 기능의 통계적 분포,『민족문화연구』49권, 고려대학교 민족문화연구원, 285~331쪽.

Fox, C.(1990), *A Stop List for General Text*, SIGER FORUM 24(1-2), pp.19~35.

Rajaraman, A. & Ullman, J. D. (2011). *"Data Mining". Mining of Massive Datasets*(PDF). pp. 1 - 17.

Raulji, J. K. & Saini, J. R.(2017), *Generating Stopword List for Sanskrit Language*, 2017 IEEE 7th International Advance Computing Conference, pp.799~802.

Siddiqi, S. & Sharan, A.(2018), *Constructive of a generic stopword list for Hindi language without corpus statistics*, International Journal of Advanced Computer Research, Vol 8(34), pp.35~40.

◻ 길호현
❀ 소 속: 서원대학교 사범대학 국어교육과과
❀ 주 소: [28674] 충북 청주시 서원구 무심서로 377-3 서원대학교
❀ 전자우편: roadway@empal.com

◎ 논문접수: 2018년 8월 31일
◎ 논문심사: 2018년 9월 13일 ~ 9월 18일
◎ 게재결정: 2018년 9월 28일

<Abstract>

The Study of Korean Stopwords list for Text mining

Kil, Ho-hyun
(Seowon University)

The purpose of this study is to present a Korean stopwords list needed to analyze Korean text using text mining method. In the preprocessing process of text mining, a task of eliminating stopwords is performed. For this purpose, we extract morphemes with the highest frequency in the large corpus suggested by the National Institute of Korean Language. The morphemes that have important meaning and the morphemes without meaning are excluded. As a result, 293 words, which are substantial morpheme, independent morpheme, and not semantically useful, were selected as an Korean stopwords list. This list is expected to be useful for analyzing Korean texts in various fields.

Keywords: Text mining, stopwords, preprocessing, corpus, vocabulary