Journal of the Korea Industrial Information Systems Research Vol. 21 No.3, June. 2016 :35-46
http://dx.doi.org./10.9723/jksiis.2016.21.3.035 ISSN:1229-3741

소셜네트워크서비스에 활용할 비표준어 한글 처리 방법 연구

(Research on Methods for Processing Nonstandard Korean Words on Social Network Services)

이 종 화[1], 레 환 수[2], 이 현 규[3]*
(Jong-Hwa Lee, Hoanh Su Le, and Hyun-Kyu Lee)

요 약 특정한 관심이나 활동을 공유하는 관계망을 구축해주는 온라인 서비스인 소셜네트워크서비스(SNS), 자신의 관심사에 따라 자유롭게 글, 사진, 동영상 등을 올릴 수 있는 공간인 블로그(Blog) 등은 자신을 알리고 표현하는 사회현상으로 자리 매김하고 있다. 이러한 SNS나 블로그를 통해 사용자들이 자유롭게 표현한 글들을 분석하여 의미있는 정보와 가치, 그리고 패턴을 찾기 위한 텍스트 마이닝(Text Mining), 오피니언 마이닝(Opinion Mining), 의미 분석(Semantic Analysis) 등의 연구가 활발히 이루어지고 있다. 또한, 연구자들의 연구 효율을 보다 높이기 위하여 키워드 기반 연구들도 이루어져있다. 하지만 대부분의 연구들은 한글의 맞춤법에 많은 한계점을 나타내고 있다. 본 연구는 어근을 찾기 힘든 이상한 외계 언어, 무분별하게 표현되는 속어, 알기 힘든 한글 이모티콘 인터넷 언어, 마이닝 처리 과정에서 파악하기 어려운 단어들을 데이터베이스에 구축하여 데이터 사전 기반 마이닝 처리 기법의 한계를 극복하고자 한다. 특정 주제에 대한 주관적 견해로 구성된 블로그를 사례 분석 대상으로 연구를 진행하였으며 유니코드를 활용한 비표준어 추출은 텍스트 마이닝 처리에 유용함을 발견할 수 있었다.

핵심주제어 : Text Mining, Non-standard, Stemming Korean, Unicode

Abstract Social network services (SNS) that help to build relationship network and share a particular interest or activity freely according to their interests by posting comments, photos, videos,... on online communities such as blogs have adopted and developed widely as a social phenomenon. Several researches have been done to explore the pattern and valuable information in social networks data via text mining such as opinion mining and semantic analysis. For improving the efficiency of text mining, keyword-based approach have been applied but most of researchers argued the limitations of the rules of Korean orthography. This research aims to construct a database of non-standard Korean words which are difficulty in data mining such abbreviations, slangs, strange expressions, emoticons in order to improve the limitations in keyword-based text mining techniques. Based on the study of subjective opinions about specific topics on blogs, this research extracted non-standard words that were found useful in text mining process.

Key Words : Text Mining, Non-standard, Stemming Korean, Unicode

* Corresponding Author : hyunqlee@pknu.ac.kr
† 본 논문은 부경대학교 경영대학 간접연구경비 2016년도 우수논문 지원 사업으로 수행된 연구임.
Manuscript received June 2, 2016 / accepted June 27, 2016

1) 부경대학교 일반대학원, 주저자
2) 부경대학교 일반대학원, 공동저자
3) 부경대학교 경영대학, 교신저자

1. 서 론

다양한 사회 분야의 편의를 제공하며 많은 변화를 주도하는 사회적 배경에는 인터넷의 등장이 있다. 디지털 정보 사회를 맞아 다양한 데이터가 축적되고 자료의 가공 처리에 대한 활용이 증가하였을 뿐 아니라, 데이터 형태 또한 다양하게 변화하였다. 사용자가 적극적으로 콘텐츠 제작에 참여하고 생산 주체로 자리매김할 수 있는 웹2.0의 환경이 이 모든 변화들을 제공하였고 사이버 공간의 네트워크 확장으로 인하여 현실사회에서의 변화 또한 증폭되고 있다는 것을 알 수 있다. 또한, 구조화된 정형적 데이터보다 문자, 사진, 동영상과 같은 비정형적 데이터가 훨씬 많은 양을 차지하고 있다[1, 2].

이러한 네트워크를 사용하는 온라인 이용자는 빠르고 광범위하게 증가하였고 콘텐츠를 생산하는 주체로서 소셜 네트워크 서비스(Social Network Service)의 중심에 자리 잡고 있다. SNS는 인터넷이 연결되지 않은 오프라인(Offline) 공간을 인터넷이 연결된 온라인(Online) 공간으로 확장시키면서 온라인 환경의 특성이 형성되었다. SNS 환경은 사용자들의 콘텐츠 생성과 다수의 의견, 경험, 관점 등을 공유함으로써 집단지성(Collective Knowledge)화 되는 소셜미디어(Social Media)로 확산되었다. 이는 SNS의 기본적 네트워크 특성과 함께 콘텐츠의 생성과 사용자들의 공유 측면이 부각되어 온 것이다[3, 4].

SNS는 사용자의 개인 네트워크 범위를 넘어서 이젠 기업이 고객의 정보를 활용하는 기준으로 작용되고 있는 사회 현상도 나타나고 있다. 요즘 금융계의 새로 등장한 금융 기술인 핀테크(Fin Tech)는 다양한 소셜미디어 분석 툴을 개발하여 고객의 SNS를 반영해 개인 신용도를 평가하는 금융서비스를 실시하고 있다. 이제는 마케팅 전략에 SNS에서의 소비자들의 행동 패턴과 소비 형태, 감정표현들을 분석하는 기업들이 대부분이며 그 범위는 보험, 은행, 증권을 비롯한 금융, 유통, 통신 등 서비스업 뿐만 아니라 제조업(원자재 재고, 수요와 공급예측, 고객의 니즈에 의한 연구개발 등)까지 빅데이터의 활용은 없어서는 안 되는 전략적 파수꾼이 되었다. 다양한 소셜망들의 방대한 데이터들은 많은 연구자들의 연구 대상으로 자리 잡고 있으며 텍스트 마이닝 처리기법을 활용한 연구도 활발히 진행되고 있다[5, 6, 7, 8, 9, 10, 11].

SNS의 사회적 팽창 보급과 확산은 세대 간 소통의 많은 차이를 낳고 있다. 국어의 정확한 표준어와 닷컴(Dot com) 세대의 자유분방한 표현 언어들과의 괴리의 연결고리를 연구하자고 한다.

2. 선행연구

데이터마이닝은 다양한 미디어의 텍스트를 수집하여 자연어 처리 과정을 거쳐 텍스트 요약과 그 빈도를 활용한 분류과정까지의 일련의 분석과정이 필요한 분석 기법이다. 텍스트 분석은 정보수집과정, 단순어의 어근, 접두·접미사, 조사 등 형태소 분석의 정보 처리 과정을 거쳐 나온 키워드들의 연관성과 중요도를 확인하기 위한 2차원 매트릭스 과정을 통한 정보 추출 과정, 연구 목적에 부합한 다양한 분석 결과를 도출하는 정보분석 과정을 통하여 이루어진다[6, 16].

그러나 자신의 의견을 자유롭게 사용하는 소셜미디어들의 사용자 표현들을 분석하는 과정에는 많은 한계점을 보이고 있다[3]. 표준어 사용을 기본으로 하는 언론매체들의 글들과는 다르게 어근을 찾기 힘든 이상한 외계 언어를 비롯하여 무분별하게 표현하는 속어, 알기 힘든 한글 이모티콘 인터넷 언어 등 한글의 말 줄임 현상들이 가세하여 기존의 형태소를 분석하기에 어려운 점이 연구자들의 한계점으로 남아 있다[5]. 이러한 부분을 조금이라도 해결하기 위한 방법으로 연구자의 연구 방향에 맞는 관련 단어들을 사전으로 구축하여 기존 마이닝 처리과정에 데이터 사전을 비교하여, 연구자의 연구 방향에 더 근접할 수 있는 방법으로 연구되고 있다. 하지만 비교 원천인 사용자들의 글이 비교되는 사전과의 대조작업이 어렵다보니 이 또한 비표준어를 분석하기 어려운 현실이다[12].

한글의 형태소 분석 연구 또한 다양한 영역에 활발히 진행되고 있다. 음절 단위의 한국어 품사 태깅 연구[13], 기존 사전들의 오류 수정 연구

[14], 형태소 분석에 필요한 품사 구분과 조사, 감탄사, 관형사를 제외한 감성의 표현을 연구하기 위한 명사, 형용사, 동사의 연구[15] 등이 연구되고 있다. 사회과학 및 인문학(사회 현상, 트렌드 분석, 텍스트마이닝 등), 자연과학(자연어 처리, 시맨틱 웹, 기계학습 등), 경영(SNS 댓글분석, 기업 평판 리스크, 마케팅 효과 측정 등)영역 등에 빅데이터 연구가 이루어져 있다[15].

본 연구는 선행연구를 통한 한글의 텍스트마이닝 처리 기법(데이터 사전 기반 마이닝 처리 기법 포함)의 한계를 극복하고자 한다. 인터넷에서 사용하는 인터넷어, 신조어 들의 한글 분석에 기존 연구자들의 키워드 기반 텍스트마이닝 처리 기법을 많이 사용하였다[6, 7, 9]. 하지만, 한글 처리에 한계점이 지적되어왔고 소비자의 행동 패턴, 감정 표현 정보 분석이 늘어가는 SNS 전성시대에 비표준어 연구의 필요가 절실하다고 판단되어 본 연구의 주제로 정하였다. 본 연구는 비표준어 사전을 통한 텍스트마이닝 처리 기법을 연구하고자 한다. 한국복제전송저작권협회의 승인을 받아 국립국어원에서 제공하는 표준국어대사전의 표제어 리스트를 근거로 비표준어 사전을 구축하였다.

본 연구의 진행 과정은 다음의 단계로 진행하였다. 먼저, 표준국어대사전의 모든 품사를 말뭉치(Corpus)로 조합하여 명사 추출 과정을 거쳐 표준어 단어사전을 작성하였다. 특정 주제에 대한 주관적 견해로 구성된 소셜 미디어의 비정형 데이터 중 텍스트 자료들을 1차 수집한다. 수집된 자료들 중 연구에 필요한 한글 자음, 모음을 비롯한 한글을 추출하기 위하여 유니코드를 활용하여 정제된 연구 대상 말뭉치(Corpus)를 작성한다. 실제 사례 분석 대상으로 자신의 의견을 자유롭게 표현하는 소셜미디어인 블로그를 대상으로 진행하였으며 표준어 단어사전과 비교하여 그 외의 단어들을 비표준어 단어로 구분하여 사전을 구축, 구현하였다.

3. 연구 방법과 프레임워크

본 연구는 자신의 관심에 따라 자유롭게 글과 사진을 올릴 수 있는 블로그를 대상으로 하였다. 사용자들의 성별, 연령, 성향에 따라 표현하는 글들은 각기 다른 표현 방법으로 작성되었으며 심각한 한글 파괴의 현상은 이미 오래된 이야기이다. 다양한 소셜미디어들의 등장만큼이나 다양하게 한글이 표현되고 있으며 인터넷어, 신조어, 채팅어 등 새로운 언어를 양상하고 있다. 단어의 축약, 함축, 비약, 합성 등이 혼재된 온갖 신조어가 난무하는 네티즌들의 대화는 언뜻 들으면 한국어라고 인식하지 못하는 경우마저 있다. 세대간 소통 또한 어렵게 하는 현재 청소년의 언어문화를 얼마나 이해할 수 있을까라는 의문이 본 연구 취지이기도 하다.

국어의 발전과 국민의 언어 생활 향상을 슬로우건으로 표준어 규정, 한글 맞춤법 등의 어문규정을 준수한 “표준국어대사전”이라는 국어사전이 있다. 계층과의 원활한 의사소통 증대를 위하여 국어 사용 환경을 개선하고 한국어 교육의 질적 향상을 위한 기반을 조성하며 국립국어원에서 관리하고 있다[17]. 또한, 우리 글자 한글의 우수성을 기리기 위한 국경일인 “한글날”을 통하여 선조들의 지혜와 후손들의 한글 바로 쓰기를 고취시키고자 함이다.

분석의 대상으로 네이버 블로그 페이지를 선정하였다. 국내 최대 검색 포털 사이트로 검색은 물론 이메일, 카페, 블로그, 지식iN, 사전, 지도, 동영상, 이미지검색 등 다양한 서비스를 제공하며 가입자 및 사용자가 국내 최대를 자랑하고 있다. 네이버 블로그는 인터넷 1인 미디어를 제공하며 성향별 블로그, 밋더 블로거, 이웃 커넥트 위젯, 포토앨범 등을 제공하고 있다[18].

분석 대상인 네이버 블로그 내 “신조어” 검색어로 13만여개의 블로그 리스트가 확인 되었다. 먼저 자료 수집의 용이성을 위하여 ‘특정출처만 검색’이란 출처 메뉴에서 네이버 블로그(blog.naver.com)로 제한을 하였다. 검색 포털 사이트는 검색어에 대한 정확도를 고려하여 네이버 블로그에서는 한 검색 조건 당 1,000개의 블로그만을 제공한다는 것을 확인하였다. 본 연구자는 2012년, 2013년, 2014년, 2015년 등 각 기간 설정을 활용하여 처리 조건을 확대하여 4000여개의 텍스트를 수집하였다. 네이버측은 사용자의 컴퓨

터 고유 숫자 주소인 IP를 분석하여 짧은 시간에 같은 IP 주소에서 블로그를 접속한다면 무차별 로봇들의 공격으로 판단하여 블로그 서버에서 같은 IP의 모든 접촉을 차단하고 있는 것 또한 확인하였다. 이러한 보안 정책을 경험한 본 연구자는 연구 목적 로봇의 접속 시간을 1분으로 지연처리하여 분당 하나의 블로그를 수집하는 것으로 만족해야만 하였다. 1,000개의 블로그를 수집하는데 17시간에 가까운 시간을 사용한 것으로 나타났다. 물론, 블로그 URL 수집기 또한 별도의 프로그램 과정을 통하여 먼저 수집 후 진행하였다.

Table 1 Number of blogs collected by year

Year	Number of blogs	Number of words	Number of pages
2012	949	364,646	4,030
2013	977	367,604	4,706
2014	958	266,084	3,913
2015	966	252,677	3,873
Total	3,850	1,251,011	16,522

이미지, 동영상을 제외한 텍스트 기준으로 수집하였고 이렇게 추출된 자료는 <표 1>과 같이 3,850개의 블로그 수와 범용 워드프로세서를 이용하여 수집된 블로그의 단어 수는 1,251,011개의 낱말 수를 대상으로 표준 단어 이외의 단어들을 찾고자 한다.

비표준어 단어 추출을 위한 우리말 표준국어대사전의 표제부 리스트의 자료 수집이 추가로 이루어졌다. 국립국어원 언어정보나눔터에서 전자사전의 파일자료를 수집하였다. 체언, 용언, 수식언, 독립언, 관계언, 어미, 접사 등 품사별 단어들을 수집하였고 연어과 같은 두 단어들의 결합된 문장도 함께 수집하였다. 이렇게 추출된 표준어 품사 단어 수는 <표 2>과 같이 553,103개의 단어로 블로그 문서들에서 제외될 표준 단어들을 수집하였다.

본 연구는 표준 단어 중심의 분석 과정의 이전 단계로써 보다 의미 있는 비표준어 단어를 포함시키는 비표준어 표제어 리스트를 구축하는 과정을 연구하였다. 원진영·김대곤(2014), 이종화·이현

Table 2 Classification of word forms in standard Korean dictionary

Order	Word form	Number of words	Remarks
1	감탄사	675	
2	고유명사	168,123	
3	관용표현	10,037	
4	관형사	727	
5	대명사	385	
6	복합명사구	20,931	
7	부사	15,959	
8	분류사	1,411	
9	수사	364	
10	어근	3,956	
11	어미	2,682	
12	연어	17,970	
13	용언	67,573	
14	의존명사	285	
15	접사	256	
16	조사	1,092	
17	체언	146,084	
18	특수어	94,593	
Total		553,103	

규(2015), 장청윤 외(2013) 연구 대상 문서의 1차 텍스트마이닝 결과와 연구 관점의 데이터 사전을 비교함으로써 보다 연구자의 연구 방향에 집중할 수 있는 텍스트마이닝 처리 기법을 기반으로 텍스트마이닝을 넘어 스마트한 텍스트마이닝 처리를 제안한다.

이와 같은 배경으로 연구 프레임워크을 제시하였고 다음과 같이 설명하였다.

<그림 1>은 소셜 데이터에서 우리말인 한글만을 추출하기 위하여 유니코드(Uni code)의 각 음절 코드 값을 이용하여 숫자, 특수문자, 외국어 등을 제외한 순수 한글만을 포함시켰다. 연구 대상 문서의 명사 처리와 명사 처리의 비정상 명사를 정확히 판단하기 위하여 형태소 분석의 품사 태그를 부착하여 명사 여과망을 보다 견고하게 설계하였다. 한편, 국립국어원 표준국어대사전 표제어 데이터는 모든 품사별 자료를 확보하고 '복

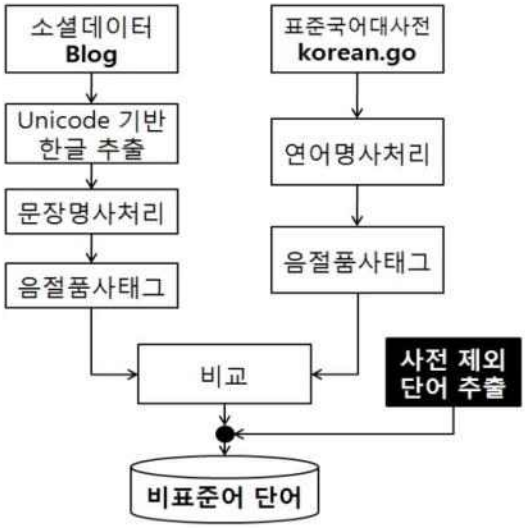

Fig 1. Research framework in this study

합명사구', '연어' 등은 명사 처리가 불가피하므로 연구 대상 문서와 같은 방법으로 명사 처리, 품사 태그 부착과정을 진행하면서 텍스트마이닝을 처리하였다. 단어별 빈도 분석 기법이 아닌 해당 단어의 유무가 중요한 연구이다 보니 비교하는 과정에서 연구 문서의 단어가 표준국어대사전 리스트에 없으면 연구 문서의 단어는 비표준어 단어로 선발된다.

이런 표준어와의 비교를 통하여 발굴한 비표준어 단어를 몇 가지 살펴보면 "갑툭튀", "금사빠", "버카충", "쓸고퀄", "호갱님", "돌취생" 등 단어들의 축약과 함축적 표현이 많은 것으로 나타났다.

4. 연구 분석 결과

빅데이터 분석 기법인 텍스트마이닝, 오피니언 마이닝, 의미분석 등 다양한 연구가 진행되어 왔다[1, 5, 7, 8, 9, 11, 12]. 본 연구는 2012년부터 2015년까지 4,000개의 블로그 소셜 데이터를 이용하여 마이닝 처리의 신뢰성을 높여 비표준어 단어 리스트를 작성하였다. 연구 대상 문서의 단어 수 1,251,011개, 표준국어대사전 단어 수 316,838개를 비교 과정을 거치며 표준어 이 외의 단어를 추출하였다.

본 연구를 위한 오픈 소스 통계분석용 SW인 R 프로그램을 사용하였다. 비정형 텍스트마이닝 처리의 필수 패키지인 tm()를 활용하여 같은 폴더의 텍스트 모두를 컴퓨터가 읽을 수 있는 형태로 모아 놓아 단어들의 뭉치로 변환되는 Corpus 기능과 한글 명사 처리를 위한 KoNLP() 패키지를 활용하여 명사를 추출하는 extractNoun 명령과 단어들의 형태소를 9개의 품사로 분리 가능한 명령어 SimplePos09 등을 이용하여 명사 처리의 신뢰도를 높였다. 수집된 사전들의 명사화는 extractNoun 명령을 사용하여 진행되었다. 하지만 <그림 2> 같이 예를 보면 일부 명사 추출에 조사가 편입된 것을 확인할 수 있었다.

<그림 2>과 같은 문장 중에서 정상적인 명사 처리가 되는 어휘가 있는 반면에 비정상적인 명사 처리로 인하여 비교 대상에서 제외되는 경우가 발생한다. 예를 들면 "한글날을", "한글에" 와 같은 어휘들이 대표적인 예이다. 정상적인 명사 처리가 되려면 "한글날을"은 "한글날"로 "한글에"는 "한글"로 처리되어야 한다.

```
> extractNoun("한글날을 하루 앞두고 개최된 이 행사는 외국인 어린이들의 한글에 대한 흥미를 높이고 마련되었다.")
 [1] "한글날을" "하루"     "개최"     "행사"     "외국"     "어린이"   "들"       "한글에"   "흥미"     "마련"
>
```

Fig 2. 1st work of noun classification(extractNoun)

```
> SimplePos09("한글날을")
$`한글날을`
[1] "한글날/N+을/J"
```

```
> SimplePos09("한글에")
$`한글에`
[1] "한글/N+에/J"
```

Fig 3. 2nd work of noun classification(SimplePos09)

```
> extractNoun("핵노잼은 정말 재미가 핵폭탄급으로 없다는 뜻이다.")
[1] "핵노잼은" "재미"     "핵폭탄"   "급"        "뜻"
>
```

Fig 4. 1st work of noun classification on research document

```
> SimplePos09("핵노잼은")
$`핵노잼은`
[1] "핵노잼/N+은/J"
```

Fig 5. 2nd work of noun classification on research document

이러한 문제는 “KoNLP”의 패키지의 단점이기도 하지만 한글 처리의 어려움을 보여주는 예이기도 하다. 본 연구자는 이러한 문제를 해결하고자 단어의 형태소 분석을 통한 품사 분리 작업을 시도해 보았다. <그림 3> 는 SimplePos09명령을 통한 형태소 분리 작업한 예이다. 비정상적 명사 처리인 “한글날을”이 “한글날”, “을”로 분리된 것을 볼 수 있다.

본 연구에 사용된 문장들은 <표 1>과 같이 1,251,011개의 수집된 단어들 전체에 형태소 분리 작업 처리가 이루어졌으며 “/N”에 해당하는 명사 단어만 분리 작업을 하였다.

연구 대상인 블로그 수집 자료 또한 같은 방법으로 처리하였다. 더욱이 수거된 자료들내에 비표준어 단어들이 포함되어 있어서 더욱 그 효과가 있었다. <그림 4>와 같이 비표준어 단어가 포함된 문장의 명사 처리 과정이다.

<그림 5>와 같이 단어에 품사 태그를 부착한 것을 확인할 수 있으며 모든 연구 문서 전체를 2차 처리까지 완료하였다.

또한, 웹 환경에서의 수집과정에 “ ”, 영문, 숫자 등을 제외한 순수 한글 음절의 구분을 위하여 유니코드 체계를 이용하였다. 문자 체계 코드의 각 문자들을 아스키코드 값으로 변환한 결과 ‘-21504’에서 ‘-10333’사이에 해당하는 코드 값이 Hangul Jamo,Hangul Moeum, Hangul Compatibility Jamo 등 11,172 글자를 포함하고 있었다[19]. 유니코드에서 한글은 AC00부터 D7A3까지 총 11,172자(초성 19 × 중성 21 × 종성 28 = 11,172)를 표현하는데, 이런 글자 수는 초성 19개, 중성 21개, 종성 28개를 곱한 값이다. 종성은 원래 27개지만, 종성이 없는 경우를 포함해 28개가 된 것이다. 서버측 웹 스크립트(ASP)의 ChrW함수는 유니코드 문자 코드 인수를 사용하고 ANSI에서 유니코드로 변환할 필요가 없으며 본 연구의 한글 아스키코드 값 추출 과정에 활용하였다<그림 6>.

```
Do
    flag ← 1
    Input st
    For( i = 1, I < length(st), ++i) {
        st1 ← Ascii(Mid(st, i, 1)
        if (st1 >= -21504 and
                st1 <= -10333))
            Print Mid(st, i, 1);
    }
Loop
```

Fig 6. Unicode Research Function

국립국어원의 언어 정보 중 다양한 품사나 어구들이 형용사 역할을 하며 명사구를 만드는 과정에 ‘복합명사구’는 “~가게”, “~공연”, “~정당성” 등의 형태로 모든 단어 앞에 “~”가 표시되어 있었다. 또한, 다른 어휘적 관계성을 보이는 단어들 사이의 결합 양상으로 구성된 ‘연어’는 “가벼운 농담”, “얼굴이 환하다”, “포스터를 떼다” 등의 단어의 결합으로 이루어져 있었다. 표준어 전체를 명사 처리하여 조사 및 특수 문자 등을 제외하고 명사의 수를 확보하였다. 그렇게 처리된 명사 단어의 수는 중복을 제외하고

316,838개를 확보하였다.

Table 3 Classification of Standard Syllables

Number of syllables	Number of words	Start number	End number
1	967	1	967
2	48,403	968	49,370
3	97,169	49,371	146,539
4	79,920	146,540	226,459
5	33,625	226,460	260,084
6	29,131	260,084	289,214
7	12,498	289,215	301,712
8	7,721	301,713	309,433
9	3,637	309,434	313,070
10	1,954	313,071	315,024
11이상	1,814	315,025	316,838

소셜데이터와 표준국어대사전의 비교 과정은 다음과 같이 진행되었다.

먼저 비교 대상이 빅데이터이며 연구 문서에서 표준화 사전의 단어를 최대 추출하여 후처리 과정이나 수작업을 줄일 수 있기 때문에 연구 대상 단어 하나를 표준 사전 전체와 비교하는 방식으로 구축하였다. 또한 효율성을 높이기 위해 31만

```
Hash-Search(T, k)
i ← 0
repeat j ← h(k, I)
    if T[j] = k
        then return j
    i ← i +1

    until T[j]=NIL or i = m
return NIL
```

Fig 7. Hash Function

단어의 음절수 순으로 정렬하여 <표 3>과 같은 해시 테이블(Hash Table) 세트를 만들었다. 연구 대상 단어의 길이가 3자리이면 49,371 ~ 146,539 사이의 단어를 검색, 5자리이면 226,460 ~ 260,084 사이의 단어를 검색하여 유휴시간(Idle time)으로 느껴지는 반복을 최소화 시켰다. 또한 11자리 이상 되는 단어는 사전 끝까지 검색되도록 설정하였다. 표준단어들을 살펴보면 3음절 단어가 97,169개로 가장 많았고 4음절 79,920개, 2음절 48,403개 순의 분포를 보였다. <그림 7>의 Hash-Search(T, k)는 해시 테이블 T와 키 k를 입력 인자로 받아, 위치 j가 키 k를 저장하고 있으면 위치 j를 리턴하고 키 k가 테이블 T에 존재하지 않은 경우는 NILL값을 리턴한다.

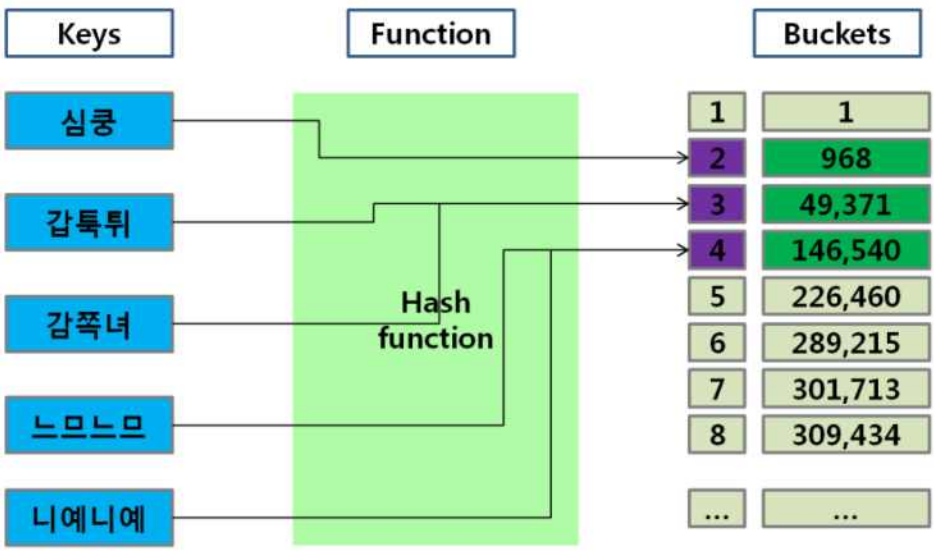

Fig 8. Structure of Hash Function

<그림 8>은 해시 테이블 구조로 연구 대상 단어 리스트에서 한 단어를 받아 주어진 단어의 길이 정보와 동일 길이의 표준어 리스트 영역 정보를 확보하는 과정을 거쳐 해싱함수에 의해 계산된 주소(버킷 주소)에 레코드 키를 저장하는 과정을 나타낸 것이며, 비교 처리 과정의 반복 횟

Table 4 List of headwords in non-standard word dictionary

Order	Words	Order	Words	Order	Words	Order	Words	Order	Words	Order	Words
1	끌	54	은탱	107	길냥이	160	셀카족	213	훈대딩	266	디질랜드
2	듄	55	저퀄	108	김치녀	161	셀피족	214	훈선수	267	똥꼬발랄
3	쩐	56	존맛	109	까도남	162	쇼루머	215	힐링녀	268	리즈시절
4	가놀	57	졸팅	110	까도녀	163	스빈다	216	ㄸㄹㄹ	269	리터루족
5	감솨	58	즐감	111	깔맞춤	164	씹덕사	217	ㅍㅎㅎ	270	멘탈붕괴
6	갬튀	59	직캠	112	꼬돌남	165	악화녀	218	ㅎㄷㄷ	271	모루밍족
7	갠소	60	출첵	113	꼬돌남	166	엄친딸	219	돌취생	272	모루밍족
8	갠톡	61	치맥	114	꿀벅지	167	엔디족	220	글설리	273	몰고어족
9	게삭	62	칭구	115	나굉족	168	엔지족	221	넘사벽	274	미담진족
10	겜방	63	칼답	116	낫닝겐	169	여미족	222	극혐오	275	번달번줌
11	겜종	64	컴겜	117	내칭구	170	완벽녀	223	쓸고퀄	276	베이글남
12	겟판	65	쿡방	118	내팅구	171	완판녀	224	피꺼솟	277	베이글녀
13	겨냄	66	쿨녀	119	노노족	172	요섹남	225	솔까말	278	부끄부끄
14	구챗	67	폰겜	120	노모족	173	우쭈쭈	226	사바사	279	불펌금지
15	구플	68	피닝	121	뇌색남	174	운도남	227	고고씽	280	브런치족
16	극혐	69	핸폰	122	뇌섹남	175	운도녀	228	김여사	281	브로맨스
17	길막	70	헐ㅋ	123	뇌섹녀	176	울아들	229	품절남	282	블링블링
18	길빵	71	홧팅	124	눔프족	177	울아빠	230	가격딴지	283	빼박캔트
19	꿀잼	72	훈남	125	뉴빠족	178	울언니	231	감성포텐	284	사토리족
20	노답	73	훈녀	126	능욕짤	179	웃겨염	232	건어물녀	285	삼포세대
21	노잼	74	힐빙	127	니트족	180	유느님	233	걸크러쉬	286	세테크족
22	노켐	75	광클	128	답정녀	181	유트족	234	고럼고럼	287	쇼루밍족
23	놀족	76	엄크	129	당글녀	182	응믐응	235	고퀄리티	288	쓰담쓰담
24	뇌섹	77	깜놀	130	대세녀	183	의느님	236	고퀄병맛	289	아리얀족
25	닉추	78	짝남	131	듣보잡	184	자출족	237	골로갈뻔	290	알바추노
26	님선	79	짝여	132	떡실신	185	좋슴당	238	군대드립	291	야지디족
27	닝겐	80	ㅋㅋ	133	똥남자	186	즐설리	239	군데렐라	292	오랜지족
28	덕질	81	ㅇㅋ	134	똥따떠	187	짜짜시	240	굿쩐문화	293	으랏차차
29	덧답	82	ㅎㅎ	135	띠껍다	188	짱짱맨	241	그래니룩	294	치렐루야
30	득템	83	겨털	136	런피스	189	쩍벌남	242	그루밍족	295	캥거루족
31	딘치	84	팀킬	137	레어템	190	차도녀	243	꺄르르르	296	코드갱어
32	렉방	85	멘붕	138	로엘족	191	출장족	244	꽁냥꽁냥	297	프리타족
33	막방	86	득템	139	로엘족	192	츤데레	245	꿀잼꿀팁	298	프티타족
34	맞팔	87	ㅇㅇ	140	로코남	193	치느님	246	끄어어억	299	힙스터족
35	먹튀	88	호갱님	141	먹튀꾼	194	치맥족	247	낄끼빠빠	300	앵그리맘
36	멍미	89	가싶남	142	몰링족	195	칭구덜	248	나포츠족	301	금사빠녀
37	모객	90	간장녀	143	몰캉족	196	컴터끄	249	나홀로족	302	남성눔프족
38	병맛	91	감쪽녀	144	물계급	197	컴터앞	250	날르가즘	303	매스티지족
39	블그	92	갑툭튀	145	뭐하삼	198	퇴장족	251	낮져밤이	304	베재댜겨쇼
40	생얼	93	개교주	146	반갑남	199	퉁퉁녀	252	내일러들	305	부키싱글족
41	소푸	94	개미족	147	반갑녀	200	펫팸족	253	노마드족	306	썸썸남썸녀
42	스샷	95	개소름	148	방콕남	201	포미족	254	뉘예뉘예	307	역쇼루밍족
43	스압	96	개이득	149	방콕녀	202	포미족	255	뉴시니어	308	찌아찌아족
44	스포	97	검색질	150	방통력	203	푸어족	256	느므느므	309	커우커우족
45	심남	98	꼴좁이	151	버카충	204	푸하핫	257	니예니예	310	코드커터족
46	심쿵	99	구글링	152	빌빌족	205	피딩족	258	달관세대	311	하이타오족
47	썸남	100	굿굿굿	153	뽀샤시	206	하빈다	259	달빛동맹	312	흐규규규뀨
48	썸녀	101	귀요미	154	뽀통령	207	핫핑크	260	대프리카	313	
49	썸맥	102	글램핑	155	삑사리	208	핫핫핫	261	덕페이스	314	
50	썸친	103	금벅지	156	사빈다	209	호갱님	262	도로로로	315	
51	안농	104	금사빠	157	상오빠	210	호빗족	263	도찐개찐	316	
52	안습	105	긔요미	158	세피족	211	후덜덜	264	등쉰색히	317	
53	왤케	106	기승전	159	셀카봉	212	훅와닿	265	디스하다	318	

수를 줄이는 코딩 작업을 하였다.

<표 1>과 같이 3,850여개의 블로그에서 1,251,011개 낱말의 단어들 중 표준 사전 316,838개의 단어를 제외한 결과 40,447개의 단어가 비표준어 단어로 나타났다. “가격할인간접할인”, “완료함으로명실상부한” 등 띄어쓰기 문제의 단어와 “오픈캐스트”와 같은 영문 한글 표기, “올라왔답니다짜자잔”과 같은 마침표(.) 오류 표기 등으로 인하여 영문 한글 표시는 삭제, 띄어쓰기 오류는 줄 바꿈, 기호 오류는 기호 삽입 및 줄 바꿈으로 교정하여 전 과정을 다시 진행하였다.

본 연구의 모형으로 4000여개의 블로그를 대상으로 비표준어 단어를 추출한 결과는 <표 4>에서 312여개의 비표준어 단어를 제시하였다. 비표준어 단어를 살펴보면 3음절 단어가 142개로 가장 많았고 2음절 84개, 4음절 72개 순의 분포를 보였다. 두 단어의 조합을 2음절 또는 3음절로 표현한 것들이 대부분이며 한 가지 재미있는 예를 들면 다음과 같다.

“삼촌, 생선으로 가방 사주세요.”

‘생일 선물’의 줄임말을 사용하여 ‘생선’이란 새로운 단어가 등장하여 청소년들 사이에 통용되는 단어 중 하나이다. 하지만 이미 국어에는 “먹기 위해 잡은 물고기”의 뜻으로 정의된 명사 단어이다 보니 형태소 분석에서 제외된 경우도 있었다. 연구 결과로 300여개의 비표준어 단어 리스트를 <표 4>에서 제시한다.

5. 결론 및 향후 과제

어근을 찾기 힘든 이상한 외계 언어, 무분별하게 표현되는 속어, 알기 힘든 한글 이모티콘 인터넷 언어, 마이닝 처리 과정에서 파악하기 어려운 단어들을 데이터베이스에 구축함으로써 보다 마이닝 처리의 속도와 신뢰도를 높일 수 있었고, 특정 주제에 대한 주관적 견해로 구성된 블로그를 실제 사례 분석 대상으로 연구를 진행하였으며 비표준어 사전을 통한 텍스트 마이닝 처리의 유용한 점을 발견할 수 있었다.

본 연구는 앞선 예문과 같이 “삼촌, 생선으로 가방 사주세요.”와 같이 소셜 미디어 간 통용되는 단어들의 차이와 그 문화가 세대 간의 소통 장벽으로 작용되는 것을 조금이나마 해소하고 건전한 언어 문화에 기여하고자 비표준어 사전을 연구하였고 그에 앞서 표제부 리스트를 기존의 소셜 데이터에서 찾고자 노력하였다.

웹 기반의 블로그 문서들은 연구 본질이 한국어에 국한적으로 적용되어 있어서 먼저 유니코드 중 한글의 음절 단위 모음, 자음 등의 자료의 범위 코드 값을 발견하여 연구하는 주제에 보다 집중할 수 있는 기준을 마련하였다.

또한 R 프로그램을 통한 한글 명사 처리 신뢰성을 높이기 위한 실험으로 명사 추출용 extractNoun명령의 결과를 형태소 태그 부착을 통한 여과망 역할의 SimplePos09을 통해 보다 깔끔한 명사 처리에 기여한 알고리즘을 제시하였다. extractNoun명령을 사용하여 아래와 같은 문장의 명사 처리 결과는 다음과 같다.

문장 : “핵노잼은 정말 재미가 핵폭탄급으로 없다는 뜻이다.”
명사 : “핵노잼은”, “재미”, “핵폭탄”, “급”, “뜻”

조사 역할의 “은/는”이 “핵노잼은”이란 명사에 포함되어 명사 처리에 미비함을 보이고 있다. 이에 연구자는 SimplePos09를 사용하여 명사의 형태소 분석과 품사별 태그를 부착하여 명사 확인 과정을 거쳤다.

명령 : SimplePos09(“핵노잼은”)
결과 : “핵노잼/N+은/J”

“핵노잼은”은 N과 J의 품사 태그가 부착되었으며 N은 체언인 명사, 대명사, 수사 등에 부착된 태그이고 J는 관계언에 속하는 조사인 격조사, 보조사, 서술격조사 등에 부착되는 태그이다. 이러한 과정을 통하여 정확한 명사 추출은 본 연구 설계 중 중요한 부분을 차지함을 물론 표준 사전과의 비교 과정에 속도면에서 직접적 영향을 주었다.

국립국어원에서 제공하는 표준국어대사전의 말뭉치 중 '복합명사구'나 '연어' 관련 단어들을 명사 처리 과정과 모든 품사 단어들을 하나로 뭉쳐 중복된 단어를 필터링 하는 과정을 거쳐 316,838개의 방대한 표준어 명사 리스트를 추출하였다. 또한, 해시테이블(Hash Table) 원리를 이용한 단어의 음절 수 기준으로 각각의 인덱스(Index) 값을 기억하여 보다 효율적인 검색 알고리즘을 구현하였다.

연구 대상 문서를 유니코드의 한글 영역 코드 값을 이용한 한글 추출과 1, 2차 명사 처리 과정과 표준어 사전 명사 리스트, 그리고 해시테이블을 이용한 검색 알고리즘을 활용하여 소셜 빅데이터를 활용한 비표준어 단어 추출 과정을 설계함으로써 보다 가치 있고 의미 있는 단어의 추출을 위한 과정을 설계하였다.

이 결과로 312개의 비표준어 단어 리스트를 <표 4>에서 제시하였으며 몇 가지 특징을 살펴보면 다음과 같다.

인터넷에서 쓰는 채팅어는 "안뇽", "칭구", "귀요미", "긔요미", "내칭구", "내팅구", "칭구덜" 등의 단어들을 사용하고 있으며 오타에서 파생된 파생어 "스빈다"-(습니다), "사빈다"-(삽니다), "하빈다"-(합니다) 등을 사용하기도 하였다.

국어와 외국어를 조합하는 합성 단어는 "코드갱어", "낫닝겐", "노답", "노잼", "딘치" 등을 확인하였고 초성만을 사용하는 이모티콘 언어는 "ㅋㅋ", "ㅇㅋ", "ㅎㅎ", "ㅇㅇ" 등이 나타났으며, 단어들의 첫 글자 혹은 첫 음절을 줄이고 결합하는 형태의 단어들이 포진되어 있었다.

SNS는 이미 우리 생활에 엄청난 부분을 함께 공유하고 있다. 세대별 언어장애의 빌미를 제공하는 부작용도 있었지만 젊은 사용자들에게 더 이상 표준어를 강요하는 것은 이미 한계점을 넘어선 것으로 보인다. 기성세대와 소통의 장이 될 수 있는 자연어 처리를 비표준어 사전 구축으로 시작해 본다. 또한, 수집된 비표준어를 표준어와 매칭하여 SNS 환경에 비표준어와 표준어의 변환 자동화를 기대해 본다. 포털사이트의 한글 영문 자동 변환 기능처럼 비표준어 연구는 세대 간 소통에 새로운 기준이 마련될 것으로 본다.

또한, 비정형화된 데이터인 자연어를 대상으로 하는 연구는 신조어가 꾸준하게 발생하는 동적인 성질이 있다. 또한 "생일선물"을 "생선"으로 표현하듯 같은 언어로 다른 뜻을 가지는 모호한 면도 있다. 물론 같은 뜻이지만 다른 표현 방법도 자연어에서는 가능한 일들이다. 이러한 한계점을 극복하기 위한 노력은 계속 진행되어야 하며 소셜 네트워크 내 소통 언어의 형태 분석, 의미 분석, 대화 분석 등을 연구하는 표준화 처리 과정의 자연어 처리 연구는 지속될 것으로 기대된다.

References

[1] Lee, J. H., "Big Data, Data Mining and Temporary Reproduction," The Journal of Intellectual Property, Vol. 8, No. 4, 2013, pp. 93-125.

[2] Kang, S. J., "Constructing a Large Interlinked Ontology Network for the Web of Data," Journal of Korean Industrial Information Systems Society, Vol. 15, No. 1, 2010, pp. 15-23.

[3] Park, C. S., Hong, Y. J. and Cho, I. H., "An Analysis on Journalism Characteristics of SNS based on Issued Cases : With Twitter as the Center," Proceedings in 2012 Fall Conference of The Korean Entertainment Industry Association, 2012, pp. 36-40.

[4] Boyd, D. M. and Ellison, N. B., "Social Network Sites: Definition, History, and Scholarship," Journal of Computer-Mediated Communication, Vol. 13, No. 4, 2007, pp. 210-230.

[5] Kim, W. S., Lee, J. H., Park, j. W. and Choi, j. H.,"A Technique of the Approval Rating Analysis for Political Party Using Opinion Mining,", Journal of Korean Institute of Information Technology, Vol. 12, No. 10, 2014, pp. 133-141.

[6] Won, J. Y. and Kim, D. G., "Deduction of Social Risk Issues Using Text Mining,"

Journal of safety and crisis management, Vol. 10, No. 7, 2014, pp. 33-52.
[7] Lee, J. H. and Lee, H. K., "A Study on Unstructured Text Mining Algorithm through R Programming based on Data Dictionary," Journal of the Korea Society Industrial Information System, Vol. 20, No. 2, 2015, pp. 113-124.
[8] Chang, J. Y., Lee, s. Y. and Han, J. B., "Machine-Learned Classification Technique for Opinion Documents Retrieval in Social Network Services," Proceedings in 2013 Conference of Korean Institute of Information Scientists and Engineers, 2013, pp. 245-247.
[9] Chang, C. Y., Jang, J. H., Kim, S, H., Lee, H. K. and Lee, C. H., "A Study on the Efficient Patent Search Process using Big Data Analysis Tool R," Journal of Korea Safety Management & Science, Vol. 15, No. 4, 2013, pp. 289-294.
[10] Le, H., and Lee, H. K., "Exploring Relationship Between Social ICT Issues And Academic Research Interests Through Text Mining Analysis," The Journal of Internet Electronic Commerce Research, Vol. 14, No. 5, 2014, pp. 161-180.
[11] Le, H., Lee, J. H. and Lee, H. K., "Purchase Process Aspect-based Opinion Mining : An Application for Online Shopping Mall," The Journal of Internet Electronic Commerce Research, Vol. 15, No. 2, 2015, pp. 15-28.
[12] Yun, B. H., "Natural Language Processing based Information Extraction for Newspapers," Journal of Korean Institute of Information Technology, Vol. 6, No. 4, 2008, pp. 188-195.
[13] Hong, J. P. and Cha, J. W., "Error Correction of Sejong Morphological Annotation Corpora using Part-of-Speech Tagger andFrequency Information," Journal of KISS : Software and Applications, 2013, Vol. 40, No. 7, pp. 417-428.
[14] Sim, K. S.,"Syllable-based POS Tagging without Korean Morphological Analysis," Korean Journal of Cognitive Science, Vol. 22, No. 3, 2011, pp. 327-345.
[15] An, J. K. and Kim, H. U., "Building a Korean Sentiment Dictionary and Applications of Natural Language Processing," Proceedings in 2014 Summer Conference of Korea Intelligent Information Systems Society, 2014, pp. 177-182.
[16] Kwon H. R., Na J. H., Yoo J. S. and Cho W. S., "Text-mining Techniques for Metabolic Pathway Reconstruction," Journal of Korean Industrial Information Systems Society, Vol. 12, No. 4, pp. 138-147, 2007.
[17] URL http://www.korean.go.kr/
[18] URL http://www.naver.com/
[19] URL http://www.unicode.org/

이 종 화 (Jong-Hwa Lee)

• 정회원
• 부경대학교 경영학 석사
• 부경대학교 경영학 박사수료
• 관심분야 : Big Data, Mining, Content Analysis

레 환 수 (Hoanh Su Le)

• 비회원
• 호치민국립대학교 경영학 석사
• 부경대학교 경영학 박사
• 관심분야 : Big Data, Data-Mining & Analytics

이 현 규 (Hyun-Kyu Lee)

- 정회원
- 연세대학교 경영학 박사
- 부경대학교 경영학부 교수
- 관심분야 : 정보시스템 전략, Data-Mining & Analytics