creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

연세대학교
YONSEI UNIVERSITY

[UCI]I804:11046-000000519495

국내 온라인 커뮤니티 게시글에 기반한 신조어 추출 방법 및 형태소 분석 적용에 관한 실증적 연구

연세대학교 정보대학원
빅데이터 석사과정
김한준

연세대학교
YONSEI UNIVERSITY

국내 온라인 커뮤니티 게시글에 기반한 신조어 추출 방법 및 형태소 분석 적용에 관한 실증적 연구

지도교수 이상우

이 논문을 석사 학위논문으로 제출함

2018 년 12 월

연세대학교 정보대학원
빅데이터 석사과정
김한준

연세대학교
YONSEI UNIVERSITY

김한준의 석사 학위논문으로 인준함

심사위원 이 상 우 인

심사위원 최 준 호 인

심사위원 송 민 인

연세대학교 정보대학원

2018 년 12 월

연세대학교
YONSEI UNIVERSITY

감사의 글

우선 부족한 저에게 배움의 기회를 제공해준 삼성 SDS 의 인사 관계자분들과 임원 및 그룹장께 감사드립니다. 그리고 금요일마다 일찍 회사를 나서는 바람에 제 업무를 나누어 맡아주고, 논문 작성을 위해 때때로 자리에서 사라져도 이해해준 동료 선후배 여러분들께도 깊이 감사드립니다. 또한 옆에서, 그리고 멀리서 언제나 한결같은 응원을 보내준 입사 동기 친구들과 멘토에게도 고마움을 표합니다.

2 년 동안 주말을 함께한 동기 여러분, 여러분을 만나 정말 반가웠습니다. 주말마다 같이 식사하고 수업을 들었던 추억을 잘 간직하겠습니다. 작은 일도 서로 도와가며 지낼 수 있어서 참 감사했습니다. 다만 제가 좀 더 먼저 다가가지 못하고, 좀 더 많은 자리에 함께하지 못해서 아쉬웠습니다. 그러나 앞으로 남아 있는 날들이 더 많기에, 이제 다시 시작할 더 뜻깊은 인연을 놓지 않고 이어가겠습니다.

논문을 쓸 수 있도록 용기를 주신 송민 교수님께 감사드립니다. 주제에 대한 확신이 없을 때 교수님의 한 마디가 저에게 큰 힘이 되었습니다. 그리고 심사 때마다 날카로운 의견을 주신 최준호 교수님께도 감사드립니다. 그리고 논문 원고 한 줄 한 줄마다 객관적이고 정성 어린 의견을 주신 이상우 교수님께 큰 감사드립니다. 부족한 제가 이 논문을 마무리할 수 있었던 것은 교수님의 지속적인 관심과 열정적인 지도 덕분입니다. 교수님께서 보여주신 열정을 잊지 않겠습니다.

학기 때마다 주말이면 사라지고 밤이 되어야 나타나는 저를 이해해주고 챙겨준 아내 이현정에게 무한한 감사와 사랑을 보냅니다. 특히 논문 작성 기간에 보여준 아내의 특급 배려 덕분에 논문을 작성할 수 있었습니다. 그리고 충분히 놀아 주지 못하고 공부하던 아빠를 기다려 준(때로는 어서 가서 논문 쓰라고 채근 대던) 우리 딸 다연이에게 고마움을 전합니다. 마지막으로 언제나 저에 대한 끝없는 사랑과 믿음을 보여 주시는 양가 부모님께 감사드립니다.

연세대학교
YONSEI UNIVERSITY

저에게 있어 논문 작성은 새로운 분야를 경험하고, 글쓰기에 대해 고민하며, 제 한계를 느끼며 저 자신에 대해 생각해볼 수 있었던 소중한 기회였습니다. 밤을 지새우며 떠올린 수많은 생각을 잊지 않고 간직하겠습니다. 감사합니다.

2018 년 12 월

김한준

연세대학교
YONSEI UNIVERSITY

차례

표 차례 ii
그림 차례 iii
국문요약 iv
제1장 서론 1
1.1 연구 배경 및 목적 1
1.2 논문의 구성 6
제2장 이론적 고찰 7
2.1 신조어 7
2.2 형태소 분석과 사용자 사전 9
2.3 명사 추출과 미등록어 처리 11
2.4 신조어 추출 12
제3장 연구 방법 14
3.1 데이터 선정 15
3.2 데이터 수집 및 전처리 16
3.3 신조어 후보 추출 18
3.4 변수 선택 22
제4장 연구 결과 26
4.1 신조어 판별 모델 26
4.2 실증적 적용 28
4.2.1 데이터 수집 및 전처리 28
4.2.2 신조어 판별 29
4.2.3 형태소 분석과 워드 클라우드 29
제5장 결론 33
참고 문헌 38

연세대학교
YONSEI UNIVERSITY

표 차례

<표 1> 형태소 분석의 예 2
<표 2> 신조어가 포함된 문장의 잘못된 형태소 분석 결과 3
<표 3> 사용자 사전 추가 이후의 올바른 형태소 분석 결과 4
<표 4> 한글 형태소 분석 프로그램 목록 9
<표 5> 디시인사이드 국내 야구 갤러리 수집 데이터 기간별 통계 17
<표 6> 전처리 전후의 텍스트 예시 17
<표 7> 디시인사이드 국내 야구 갤러리 수집 데이터 음절수별 통계 18
<표 8> 문장으로부터 추출한 후보 단어 예시 19
<표 9> 후보 단어별 빈도수 및 빈도비율과 최종 선정 단어 예시 19
<표 10> 우리말샘으로부터 내려받은 파일 내용 예시 20
<표 11> 우리말샘 데이터 추출 정규식 21
<표 12> WordExtractor 제공 통계지표 22
<표 13> 후보단어의 특징을 더 설명하기 위해 추가한 통계지표 23
<표 14> 수작업을 통해 신조어로 판별된 단어 목록 24
<표 15> 모델 생성시 사용 데이터 분류 26
<표 16> 신조어 판별 모델의 추정 결과 26
<표 17> 신조어 판별 모델의 오분류표 27
<표 18> 디시인사이드 2018년 10월 수집 데이터 갤러리별 통계 29
<표 19> 국내야구 갤러리 2018년 10월 게시물 제목으로 부터 추출된 신조어 목록 29
<표 20> 형태소 분석 프로그램의 사용자 사전 추가 방법 30
<표 21> 자주 사용된 명사 20개 30
<표 22> 신조어 사전 추가 전후, 각각의 경우에만 추출된 명사 목록과 순위 31

연세대학교
YONSEI UNIVERSITY

그림 차례

<그림 1> 형태소 분석 프로그램의 분석 과정 10
<그림 2> 연구 과정 및 방법 14
<그림 3> 신조어 사전을 사용하지 않은 분석 결과로 생성한 워드 클라우드 32
<그림 4> 신조어 사전을 사용한 분석 결과로 생성한 워드 클라우드 32

연세대학교
YONSEI UNIVERSITY

국문요약

신조어는 생성 당시의 사회상을 반영하는 단어라는 특징이 있기 때문에 텍스트 분석 시 무시할 수 없는 중요 단어라고 할 수 있다. 사회의 변화 속도가 점점 빨라짐에 따라 신조어의 생성과 소멸도 빨라지는 경향이 있어서, 신조어가 사전으로 구축되는 시점이 생성된 시점보다 늦기 마련이다. 그리고 한글은 영어와 달리 하나의 어절이 여러 형태소를 가지고 있기 때문에 자연어 처리 시 형태소 분석 과정이 필요한데, 형태소 분석 중 신조어로 인해 미등록어 처리에 대한 문제가 나타난다. 형태소 분석 결과에 신조어가 하나의 형태소로 유지되지 못 하고 더 작은 형태소로 잘 못 분해되어 신조어의 의미를 잃어버리는 경우가 발생하는 것이다.

신조어의 양이 급격하게 증가하고 있는 환경에서 신조어 사전을 미리 구축하여 텍스트 분석 시 활용하기에는 무리가 있다. 만약 분석하고자 하는 텍스트 데이터에 처음으로 등장한 신조어가 있다면 해당 신조어는 하나의 형태소로 추출되지 않을 것이다. 따라서 텍스트 분석을 진행할 때마다 분석하고자 하는 텍스트 데이터로부터 신조어를 추출하는 과정을 먼저 진행하여, 해당 텍스트 데이터만을 위한 신조어 사전을 먼저 구축하고, 이를 형태소 분석 시 다시 활용하는 방법이 텍스트 분석할 때 텍스트의 의미를 더 정확하게 파악하는 방법이 될 것이다.

본 연구에서는 텍스트 데이터로부터 신조어를 추출하는 방법으로 신조어 판별 모델을 제시하였다. 모델을 만들기 위해 국내 최대 온라인 커뮤니티인 디시인사이드의 국내 야구 갤러리의 2018년 7월부터 9월까지 3개월간의 게시물 제목을 수집하였다. 수집된 데이터를 전처리한 후 각 어절에서 조합이 가능한 모든 부분 글자들을 생성하여 빈도수가 전체 어절 수의 0.01%를 넘는 단어를 신조어 후보 단어로 선정하였다. 후보 단어가 전체 텍스트 데이터에서 가지는 통계적 특징을 독립변수로 사용하기 위하여, 각 후보 단어의 길이와 빈도수와 어절에서 시작부분에 위치하는 비율과 끝부분에 위치하는 비율을 계산하고, Python 패키지인 soynlp의 WordExtractor 클래스를 사

연세대학교
YONSEI UNIVERSITY

용하여 각 후보 단어의 글자가 함께 등장하는 정도와 후보 단어의 왼쪽과 오른쪽에 등장하는 글자의 다양성을 수치화하였다. 모델의 종속변수를 확보하기 위하여 신조어 후보 단어가 실제로 신조어인지를 파악해야 하는데, 이는 본 연구자가 각 후보 단어를 인터넷에 검색해 보거나 후보 단어가 사용된 데이터 내의 문장을 검토하여 판단하였다.

확보된 분석 데이터에 로지스틱 회귀분석을 수행하여 7개의 변수로 구성된 신조어 판별 모델을 생성하였다. 한글을 형태소 분석할 때에 신조어가 더 분해되거나 빠뜨리지 않고 포함되도록 하는 것이 본 연구의 궁극적인 목적이기 때문에 신조어가 아닌 단어를 신조어라고 판단하는 경우보다, 신조어인데 신조어로 판별하지 못하는 경우를 줄이는 것이 중요하다. 따라서 모델의 정확도만큼 민감도도 중요하다고 볼 수 있다. 본 연구에서 생성한 신조어 판별 모델의 정확도는 81.94%, 민감도는 81.2%이다.

본 연구에서 생성한 신조어 판별 모델을 실증적으로 적용해보기 위해 디시인사이드 국내 야구 갤러리의 2018년 10월 게시물 제목으로부터 신조어 후보 단어를 추출하고 추출된 후보 단어마다 모델에 필요한 7개 독립변수를 계산하여 127개의 신조어를 추출하였다. 추출한 신조어를 형태소 분석 프로그램의 시스템 사전 또는 사용자 사전에 추가하여 2018년 10월 게시물 중 무작위로 뽑은 5,000건을 형태소 분석하여 데이터에서 자주 언급되는 명사 단어로 워드 클라우드를 생성하였다. 그리고 신조어를 추가하기 전의 형태소 분석 결과와 비교하여, 신조어 사전이 추가됨으로 인해 추출되지 않던 단어가 빈도수 상위 단어로 추출되고, 하나의 신조어가 나누어 추출되던 오류가 수정되는 것을 확인하였다.

핵심 되는 말 : 텍스트마이닝, 형태소분석, 신조어, 온라인커뮤니티, 사용자사전

연세대학교
YONSEI UNIVERSITY

제1장서론

1.1 연구 배경 및 목적

1990년대부터 월드와이드웹으로 대표되는 인터넷의 급속한 발달로 수많은 사용자가 다양한 종류의 데이터를 만들어서 인터넷 공간 속에 그 데이터를 저장하고 있다. 2000년대 이후, 빅데이터라는 키워드로 표현될 정도로 무수히 생성되는 데이터는 기존 데이터베이스로는 처리하기 어려울 정도의 크기를 가지는 것이 보통이고(Özköse, Arı & Gencer, 2015;Brown, Chui & Manyika, 2011), 정형 데이터뿐만이 아니라 이미지와 텍스트와 같은 비정형의 데이터까지 그 형태가 다양하다. 인터넷에서 생성되는 데이터 중에서 비정형 데이터가 80% 정도를 차지할 정도로 비정형 데이터의 양은 정형 데이터의 양을 훨씬 상회하는 것으로 추정된다(Chakraborty & Pagolu, 2014).

이에 따라 텍스트 데이터의 원문을 이해하고 유용한 정보를 추출하는 과정인 텍스트 마이닝(Text Mining)의 중요성이 높아지고 있다. 텍스트 마이닝 기법은 최근 다양한 분석 방법론의 개발, 정보처리 속도의 향상, 저장 장치의 발달 등을 기반으로 급속히 발전하고 있으며 이에 대한 관심도 높아지고 있다(원중호, 이한별, 문혜정, 손원, 2017). 텍스트 마이닝의 분석 기법은 크게 정보 추출(Information Extraction), 정보 검색(Information Retrieval), 자연어 처리(Natural Language Processing), 군집화(Clustering), 문서 요약(Text Summarization)으로 나눌 수 있다(Talib, Hanif, Ayesha, & Fatima, 2016). 자연어 처리를 포함한 대부분의 분석 기법은 분석 대상이 되는 텍스트에 대한 전처리를 필요로 하는데, 전처리 과정에서 문장의 각 단어가 어떤 역할을 하는지를 파악하여 품사를 결정하게 된다. 하나의 단어가 여러 품사가 될 수 있는 모호성을 가지기 때문에 이를 해소하는 과정을 품사 태깅(part-of-speech tagging)이라고 한다(Kroeger, 2005).

연세대학교
YONSEI UNIVERSITY

한글 자연어 처리의 경우에는 영어와 달리 하나의 어절이 여러 형태소[1]를 가지고 있기 때문에 형태소 분석이 자연어 처리에 있어서 가장 중요하고 기초적인 역할을 수행한다(송민, 2017). 형태소 분석이란 단어를 구성하는 각각의 형태소들을 인식하고 형태소의 원형을 복원하는 과정을 말한다(강승식, 2002). 이 과정에서 형태소마다 품사가 결정되는데, 예를 들면 **<오류! 참조 원본을 찾을 수 없습니다.>**과 같이 "나는 학생입니다"라는 문장은 형태소 분석하게 되면 총 5개의 형태소를 가지는 것으로 분석된다.

<표 1> 형태소 분석의 예

문장	나는 학생입니다
형태소분석 결과	나#대명사 + 는#보조사 + 학생#일반명사 + 이#긍정지정사 + ㅂ니다#종결어미

형태소 분석은 일반적으로 분석 후보 형태소를 여러 가지 경우의 수로 생성한 이후에 그중에 옳은 후보를 선택하기 위해서 시스템 사전을 이용하게 된다(송민, 2017). 이 과정에서 이용되는 시스템 사전으로 21세기 세종계획에 의해 구축된 세종 말뭉치 사전[2]이 주로 사용된다.

하지만 문장에 신조어가 포함된 경우의 형태소 분석 결과에서는 올바른 품사 태그 결과를 기대하기가 어렵다. 신조어는 이전에는 알려지지 않은 단어이기 때문에 소프트웨어 형태의 형태소 분석기에서는 더 작은 형태소로 분해하려는 경향이 있고 <<표 2>와 같이 하나의 형태소로 인식되지 않고 더 작은 형태소로 나누어지는 현상이

[1] 형태소(形態素) : 뜻을 가진 가장 작은 말의 단위로서, 더 분석하면 뜻이 없어지는 가장 작은 의미 요소

[2] 세종 말뭉치 사전 : 1998년부터 2007년까지 진행된 21세기 세종계획의 일활으로 발표한 1,000만 어절 규모의 현대 문어 형태의 품사 부착 말뭉치

연세대학교
YONSEI UNIVERSITY

나타난다.

<표 2> 신조어가 포함된 문장의 잘못된 형태소 분석 결과

문장	나 요즘 아싸에서 인싸됨
형태소분석 결과	나#대명사 요즘#일반명사 **아#감탄사** **싸#동사** **아#연결어미** 에서#부사격조사 **인#일반명사** **싸#동사** **아#연결어미** 되#동사 ㅁ#명사형전성어미

"자신이 소속된 무리 내에서 적극적으로 어울려 지내는 사람"을 일컫는 말[3]인 "인싸"를 1개의 형태소가 아닌 "인+싸+아"의 3개의 형태소로 분리하여 원래의 의미를 잃어버리게 된다. "아웃사이더"에서 파생된 줄임말로 "혼자 노는 사람"의 의미[4]로 쓰이는 "아싸"라는 단어 역시 "아+싸+아"로 분석되어 문장에서 쓰인 원래의 의미를 잃어버리게 된다.

신조어는 당시의 사회적 화젯거리나 전반적인 시대상을 반영하여 생성되는 경우가 많기 때문에, 신조어를 올바르게 식별하게 되면 분석 대상 텍스트의 핵심 주제를 파악할 확률이 높아질 것으로 기대할 수 있다(김환, 2017). 또한 분석 대상 텍스트가 온라인 커뮤니티의 자유게시판이나 SNS(Social Network Service)와 같이 비교적 짧은 문장으로 구성된 경우에는 신조어를 식별하는 방법만이 텍스트의 의미를 파악할 수 있는 유일한 방법일 가능성이 높기 때문에 신조어가 식별되지 않고서는 문장의 의미를 파악하지 못하게 된다.

신조어를 식별하지 못하는 한계를 해결하기 위해 몇몇 형태소 분석 프로그램은 내부적으로 사용하는 시스템 사전 외에 사용자 사전을 추가할 수 있는 기능을 제공한다. 사용자 사전은 형태소 분석 프로그램을 이용하는 사용자가 직접 특정 형태소와

[3] 우리말샘 "인싸" : https://opendict.korean.go.kr/dictionary/view?sense_no=1379205

[4] 우리말샘 "아싸" : https://opendict.korean.go.kr/dictionary/view?sense_no=1368957

연세대학교
YONSEI UNIVERSITY

품사를 등록하고자 할 때 이용될 수 있다. 텍스트에서 사용자 사전에 등록한 형태소가 등장하는 경우, 해당 형태소로 분석하고 더 작은 형태소로 나누지 않게 된다. 앞에서 언급된 "인싸"와 "아싸"라는 단어를 일반 명사로 추가하게 되면 <<표 3>과 같이 형태소 분석이 올바르게 이루어진다.

<표 3> 사용자 사전 추가 이후의 올바른 형태소 분석 결과

문장	나 요즘 아싸에서 인싸됨
사용자 사전	아싸 NNG[5] 인싸 NNG
형태소분석 결과	나#대명사 요즘#일반명사 **아싸#일반명사** 에서#부사격조사 **인싸#일반명사** 되#동사파생접미사 ㅁ#명사형전성어미

위의 예에서 보았듯이, 신조어가 포함된 텍스트 데이터를 형태소 분석을 할 경우에는 미리 구축된 신조어 사전이 필수적이다. 그래야만 신조어가 하나의 형태소로 식별되어 본래의 의미를 유지할 수 있기 때문이다. 신조어 대부분은 그 품사가 명사이기 때문에 신조어를 추출하는 것을 명사를 추출하는 문제로 일반화하여 생각할 수 있다. 명사를 추출하고 기존에 알고 있는 단어를 제외하면 그것이 바로 신조어이기 때문이다.

하지만 신조어는 새로 만들어진 말이라는 그 본연의 특성상 형태소 분석 전에 미리 파악하기 어렵다. 만약 분석하고자 하는 텍스트 데이터에 처음으로 등장한 신조어가 있다면 해당 신조어는 명사로 추출되기 어렵다. 그러므로 신조어 사전을 미리 구축하는 것만큼이나 분석하고자 하는 텍스트로부터 신조어를 미리 추출하는 방법이 필요하다.

이를 위해 본 연구에서는 텍스트가 가지고 있는 통계적인 특징을 수치화하여 이

[5] NNG : 품사 태깅시 일반 명사를 나타내는 영문 약어

연세대학교
YONSEI UNIVERSITY

를 기반으로 신조어 판별 모델을 생성하고자 한다. 통계적인 특징을 수치화하기 위해 문장을 공백 기준으로 어절 단위로 나누고, 각 어절에서 조합이 가능한 모든 부분 글자를 생성하여 각 부분 글자마다 빈도수와 부분 글자가 어절의 시작부분인지 끝부분인지, 부분 글자 좌우로 어떤 음절이 나타나는지 등의 통계지표 계산이 필요하다. 그리고 각 부분 글자가 실제로 단어로 쓰이고 있는지에 대한 조사도 필요하다. 앞에서 언급된 문장을 예를 들어 보면, "인싸됨"이라는 어절은 총 6개의 음절 조합을 만들어 낼 수 있다. "인", "인싸", "인싸됨", "싸", "싸됨", "됨"이 그 조합이다. 이 중에서 "인싸"가 실제 단어로 쓰이고 있는 부분 글자이고 이를 "인싸"가 가지고 있는 통계지표로부터 판단하는 방법이 있다면 바로 그 방법이 텍스트로부터 신조어를 미리 추출하는 방법이 될 것이다.

위와 같은 방법으로 신조어 판별 모델을 구축한다면, 신조어 판별 모델을 구축한 데이터와 유사한 텍스트 패턴을 가지는 텍스트에 대해서는 이미 구축한 모델을 적용하여 새로운 신조어를 쉽게 추출할 수 있을 것이다. 예를 들어, 특정 온라인 커뮤니티의 게시글은 해당 커뮤니티에 글을 자주 남기는 사람들이 비슷한 단어 사용 패턴을 가지는 글이 작성되기 마련이다. 그러면 일정 기간의 데이터로부터 부분 글자(신조어 후보 단어)가 가지는 패턴을 계산하고, 부분 글자마다 신조어 여부를 확보한다면 미래에 해당 데이터로부터 신조어 여부를 자동으로 추출할 수 있을 것이다.

본 연구의 목적은 다음과 같다.

첫째, 텍스트 데이터를 형태소 분석할 때 사용자 사전으로 활용할 수 있는 신조어를 추출하는 일련의 분석 절차를 설계하고자 한다. 절차 중 텍스트로부터 추출한 신조어 후보 단어가 실제로 신조어인지를 판별하는 신조어 판별 모델도 생성하였다. 분석 절차와 신조어 판별 모델을 활용하여 신조어가 형태소 분석 시 빠지는 것을 방지할 수 있을 것이고, 이는 곧 텍스트 데이터가 어떤 이야기를 하는지에 대한 이해를 도울 수 있을 것이라 기대한다.

둘째, 앞에서 설계한 분석 절차와 신조어 판별 모델의 실증적인 적용을 위해서 신조어 판별 모델을 국내 온라인 커뮤니티의 텍스트 데이터에 적용하여 신조어를 추

연세대학교
YONSEI UNIVERSITY

출하였다. 그리고 추출한 신조어를 사용자 사전으로 구축하여 사용자 사전을 적용하기 전과 후의 형태소 분석 결과를 비교하여 본 연구의 유용성을 제시하였다.

1.2 논문의 구성

본 연구의 구성은 다음과 같다.

제1장은 서론으로 연구의 배경과 목적에 관해 기술하였다.

제2장은 이론적 고찰로서 문헌연구를 통해 신조어에 대한 기존 조사 연구를 정리해보고 형태소 분석 프로그램과 각 프로그램이 제공하는 사용자 사전 기능을 확인하였다. 그리고 신조어 추출의 기본 원리라고 할 수 있는 미등록어 처리에 대한 기존 연구를 조사하였다. 이를 통해 본 연구가 가지는 의의를 설명하고 연구 문제를 제시하였다.

제3장은 연구 방법으로서 2장에서 도출된 연구 문제인 신조어 판별 모델을 생성하기 위한 연구 방법을 제시하고 수집데이터 선정, 데이터 전처리, 모델 생성을 위한 변수 선택에 대해 구체적으로 설명하였다.

제4장은 연구 결과로서 신조어 판별 모델의 추정 결과를 기술하고 모델의 실증적 적용을 위해 새로운 수집데이터에 모델을 적용하여 신조어를 추출하였다. 그리고 신조어로 구축한 사용자 사전을 사용하기 전과 후의 형태소 분석 결과를 비교하였다.

제5장은 결론으로 본 연구가 가지는 시사점과 한계점에 대해 논의하고 후속 연구에서는 어떤 연구들이 이루어질 필요가 있는지에 대한 방향성을 제시하였다.

연세대학교
YONSEI UNIVERSITY

제2장이론적 고찰

2.1 신조어

새로 생겨난 어휘는 '신어(新語)' 또는 '신조어(新造語)'라 부른다. 표준국어대사전[6]에서도 '신어', '신조어', '새말'을 동의어로 표시하고 있다. 문금현(1999)은 기존 언어와 유연성 없이 새롭게 창조된 말은 '신생어(新生語)', 기존 언어재[7]를 바탕으로 생성된 2차 어휘는 '신조어', 그리고 이 2가지를 아우르는 용어를 '신어'로 정의하였다. 본 연구에서는 이런 세부적인 분류와 정의를 사용하지는 않고 표준국어대사전을 따라 새로 생겨난 어휘를 '신조어'라는 용어로 통일하여 사용하도록 한다.

신조어에 대한 조사 연구는 1994년에 국립국어원에서 신문과 잡지로부터 사전이 등재된 적이 없는 단어를 찾아 자료집 형태의 '신어의 조사 연구'를 발간한 것부터 시작되었다(문금현, 1999). 같은 해에 국립국어원에서 신문, 방송, 잡지에서 사용되는 약어를 정리하여 '현대 국어의 약어 목록'을 발간하였는데 오늘날 많은 신조어가 약어의 형태인 것을 생각하면 이 또한 신조어에 대한 조사 연구라고 볼 수 있을 것이다. 이후 국립국어원은 1995년에 '95년 신어'를 다시 발간하였고, 1996년에는 현대시에 사용된 신조어를 조사하여 '신어의 조사 연구(현대시의 신어 연구)'를 발간하였다(국립국어연구원, 2000). 국립국어원은 이후 2000년부터 다시 매년 신조어 조사를 이어나갔고, 2007년에는 2002년 이후 생겨난 새 말을 사전으로 정리한 '사전에 없는 말 신조어'를 발간했다(국립국어원, 2011).

신조어에 대한 조사 연구와 더불어 신조어의 생성, 정착, 관리방법 등의 연구도 진행되었다. 김태훈과 박상진(2011)의 연구에서는 1920~1930년대의 대중 잡지에 실린

[6] 국립국어원 표준국어대사전 : http://stdweb2.korean.go.kr

[7] 언어재 : 언어적 재료의 의미로 신조어를 만들 때 사용되는 기존 어휘

연세대학교
YONSEI UNIVERSITY

신조어 현황을 정리하고 신조어의 정착을 판단하는 기준을 마련하였다. 김동의와 이상곤(2013)의 연구에서는 네이버의 언론사별 뉴스 기사를 수집하고 국어 전공자로 하여금 신조어를 찾아내어 사전을 구축할 수 있게 하는 관리도구를 제작하였다. 김일환(2014)의 연구에서는 신문 기사로 구성된 '물결21' 코퍼스[8]로부터 신어 명사들을 추출하고 특성을 정리하여 이들의 시기별 사용 양상을 통해 신어의 생성과 정착 과정을 규명하였다. 남길임, 이수진, 그리고 최준(2017)의 연구에서는 네이버 뉴스 기사를 수집하여 구축한 대규모 말뭉치[9]를 통한 신어 사용 추이 조사의 방법론과 의의를 생존 신어의 대표적인 사례를 중심으로 논의하였다. 윤경선(2013)의 연구에서는 트위터에서 수집한 신어를 통해 SNS에 사용된 신어의 양상과 기존 통신 언어와 단어 형성법에 어떤 차이가 있는지를 비교하였다.

연구자에 의한 신조어 조사 외에도 신조어를 사용하는 사람이 직접 신조어를 정리하려는 움직임도 있다. 국립국어원은 2016년에 '우리말샘'이라는 웹사이트 형태의 개방형 한국어 사전을 구축하여 사용자가 직접 사전 작성에 참여할 수 있게 하였다. 2018년 10월 현재 약 100만 개의 단어를 보유하고 있고, 사이트 개설 이후 약 8천여 개의 단어가 추가로 등록되어 '가즈아', '레알'과 같이 비교적 최근에 인터넷에서 유행한 신조어를 포함하고 있는 편이다.

위키피디아[10]와 나무위키[11]와 같은 위키 형식의 온라인 백과사전도 신조어가 정리된 페이지를 가지고 있다. 위키의 특성상 여러 사용자의 노력으로 일정량의 신조어가 등록되어 있고, 새로운 신조어가 추가될 때 비교적 빠르게 등록되는 편이다.

---

[8] 코퍼스(corpus) : 언어 연구를 위해 텍스트를 컴퓨터가 읽을 수 있는 형태로 모아 놓은 언어 자료

[9] 말뭉치 : 코퍼스(corpus)의 순우리말

[10] 위키피디아 대한민국의 인터넷 신조어 목록 : https://ko.wikipedia.org/wiki/대한민국의_인터넷_신조어_목록

[11] 나무위키의 신조어 페이지 : https://namu.wiki/w/신조어

연세대학교
YONSEI UNIVERSITY

2.2 형태소 분석과 사용자 사전

형태소 분석이란 띄어쓰기 단위의 어절인 단어를 구성하는 각각의 형태소들을 인식하고 형태소의 원형을 복원하는 과정을 말한다(강승식, 2002). 형태소 분석에 사용되는 알고리즘에 대한 노력은 1980년대부터 시작되었고(김성용, 1987), 형태소 분석을 자동화하기 위해 소프트웨어 형태의 다양한 형태소 분석 프로그램이 개발되어 대부분 오픈 소스[12]형태로 제공되고 있다. <<표 4>는 Python 한글 형태소 분석 패키지인 KoNLPy[13]에서 사용되는 형태소 분석 프로그램의 목록이다.

<표 4> 한글 형태소 분석 프로그램 목록

형태소 분석 프로그램	URL
한나눔	http://semanticweb.kaist.ac.kr/home/index.php/HanNanum
꼬꼬마	http://kkma.snu.ac.kr/
KOMORAN	https://github.com/shin285/KOMORAN
MeCab-ko	http://eunjeon.blogspot.com/
open-korean-text	https://github.com/open-korean-text/open-korean-text

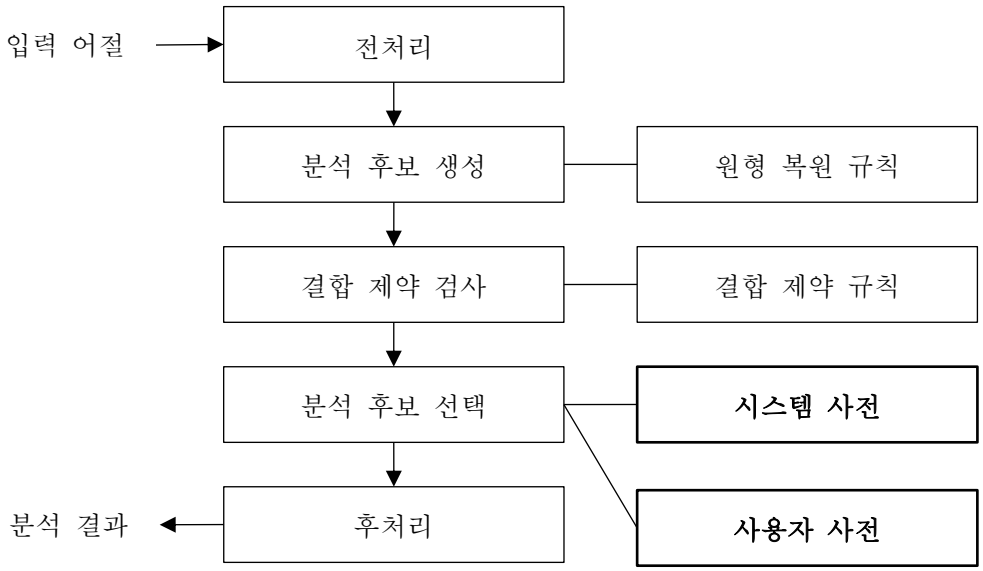
형태소 분석 프로그램마다 텍스트 데이터를 처리하고 형태소를 식별하는 알고리즘은 다르지만, 공통으로 <<그림 1>의 5단계의 절차를 거쳐서 분석 결과를 만들어 낸다.

[12] 오픈 소스 : 소프트웨어 혹은 하드웨어의 제작자의 권리를 지키면서 원시 코드를 누구나 열람할 수 있도록 한 소프트웨어

[13] 파이썬 한국어 NLP : http://konlpy.org/ko/latest/api/konlpy.tag/

연세대학교
YONSEI UNIVERSITY

<그림 1> 형태소 분석 프로그램의 분석 과정



첫 번째 전처리 단계에서는 입력 어절에서 문장부호를 분리하고 숫자나 특수문자를 처리한다. 두 번째 분석 후보 생성 단계에서는 형태소를 여러 가지 조합으로 분리하여 분석 후보를 생성하고, 각 형태소의 원형 복원 규칙을 참고하여 불규칙원형을 복원한다. 세 번째 결합 제약 검사 단계에서는 모음조화, 형태소 결합 제약, 음운 형상 등에 따른 제약을 검사하여 분석 후보를 제거한다. 네 번째 분석 후보 선택 단계에서는 시스템 사전을 참고하고 단어 형성 규칙을 적용해서 옳은 후보를 선택한다. 마지막 다섯 번째 단계에서는 복합명사를 추정하여 분석 결과를 만든다(송민, 2017).

네 번째 단계에서 사용되는 시스템 사전은 각 형태소 분석 프로그램마다 형태소와 품사의 목록으로 구성되어 있다. 새로 생겨난 신조어는 시스템 사전에 포함되어 있지 않기 때문에 형태소 분석 프로그램은 이를 보완하기 위해 사용자 사전이라는 기능을 제공한다. 형태소 분석 프로그램을 이용하는 사용자가 신조어 목록을 별도로 구성하여, 형태소 분석 프로그램이 수행될 때 이 목록에 접근할 수 있게 함으로써 형태소 분석 프로그램이 시스템 사전에 사용자 사전을 더하여 형태소 분석에 활용하게 된

연세대학교
YONSEI UNIVERSITY

다. 형태소 분석 프로그램이 사용자 사전 기능을 제공하지 않더라도, 시스템 사전을 직접 수정하는 방법으로 사용자가 원하는 신조어를 포함하여 형태소 분석을 수행할 수 있다.

2.3 명사 추출과 미등록어 처리

신조어가 포함된 텍스트에 대한 형태소 분석을 올바르게 하기 위해서는 신조어 사전 구축이 제대로 이루어질 필요가 있다. 앞에서 살펴보았듯이 여러 연구를 통해 신조어가 조사되고 있고, 우리말샘이나 위키피디아 등의 웹사이트에 신조어가 등록되고 있으므로 이렇게 구축된 신조어를 사용자 사전으로 사용하는 것을 먼저 생각해볼 수 있다. 하지만 신조어 본연의 특성상 형태소 분석을 수행하는 시점에 텍스트에 포함된 신조어가 모두 사전으로 구축되지 못한다는 것은 분명하다. 텍스트에서 실제로 사용되는 신조어를 미리 알지 못한다면, 문장의 의미 파악이라는 형태소 분석의 목적을 달성하기 어려운 것이다. 따라서 텍스트로부터 신조어를 추출하여 이를 형태소 분석 시 사용자 사전으로 다시 추가하는 방법을 사용함으로써, 미리 구축된 신조어 사전에는 포함되지 않았지만, 텍스트에서는 사용되고 있는 신조어를 최대한 찾을 수 있을 것이다.

우리말샘에 등록된 전체 단어 중 74%가 명사로 구성[14]되어 있다. 그리고 2014년 조사된 신조어의 97%가 명사 또는 명사구로 이루어져 있다(남길임, 2014). 따라서 텍스트로부터 신조어를 찾아내는 과정은 텍스트에서 사용된 명사를 찾아내어 그 중 이미 알고 있는 단어를 제외하는 과정으로 일반화할 수 있다. 이미 알고 있는 단어는 사전을 활용하여 쉽게 제외할 수 있지만, 그 존재를 미리 알 수 없는 명사(즉, 신조어)를 텍스트에서 찾아내는 것은 이미 한글 형태소 분석에서 해결해야 하는 난제로

[14] 우리말샘 사전통계 : https://opendict.korean.go.kr/service/dicStat#static_menu1_3

연세대학교
YONSEI UNIVERSITY

제시된 바가 있을 정도로 쉽지 않다(송민, 2017).

박봉래, 황영숙, 그리고 임해창(1998)의 연구에서는 기존의 미등록어를 처리하는 방법의 문제점을 제시하고, 용례 분석에 근거한 새로운 미등록어 인식 방법을 제안하였다. 이는 사람이 미지의 단어에 대한 인식 과정을 모델화한 것인데, 사람들은 처음 보는 단어를 그 단어가 쓰이는 문장을 접하면서 점점 명확하게 인식하게 된다는 것에 착안한 연구이다.

방진우(2017)의 연구에서는 역사 자료 말뭉치에 대한 분석을 진행하면서 발생한 미등록어를 처리하기 위해 어간부 사전, 문법부 사전, 호칭 사전을 구축하여 활용하였다. 먼저 미등록어절의 뒷부분에서부터 최장일치법으로 문법부 사전에서 일치한 항목을 찾고, 미등록어절의 나머지 부분이 어간부 사전에 매칭되는 경우가 있는지를 확인한다. 만약 조건을 만족시키면 어간부와 문법부 각각을 분석 결과에 추가한다. 예를 들어 "로슉/NNP[15]"이 어간부 사전에 있고, "이/JK[16]"가 문법부 사전에 있는 상황에서, "로슉이"라는 미등록어는 "로슉/NNP"와 "이/JK"로 나누어 분석 결과에 포함되는 것이다. 이 방법으로도 찾을 수 없다면 호칭 사전과 비교하는 과정을 거치게 하였다. 미등록어의 뒷부분 조사를 제거하고 남은 어절의 뒷부분에 호칭 사전의 호칭이 발견되면 호칭을 제외한 나머지를 고유명사로 판단하였다. 예를 들어 "밀양읍에서"가 미등록어라고 할 때, 조사 "에서"를 제거하고 남은 "밀양읍"이 호칭 사전의 "~읍"을 뒷부분으로 가지고 있으므로 "밀양"을 고유명사로 판단한다.

2.4 신조어 추출

앞에서 살펴본 바와 같이 신조어는 새로 생긴 말이라는 특성상 사전에 등록되기

[15] NNP : 품사 태깅시 고유 명사를 나타내는 영문 약어

[16] JK : 품사 태깅시 조사를 나타내는 영문 약어

연세대학교
YONSEI UNIVERSITY

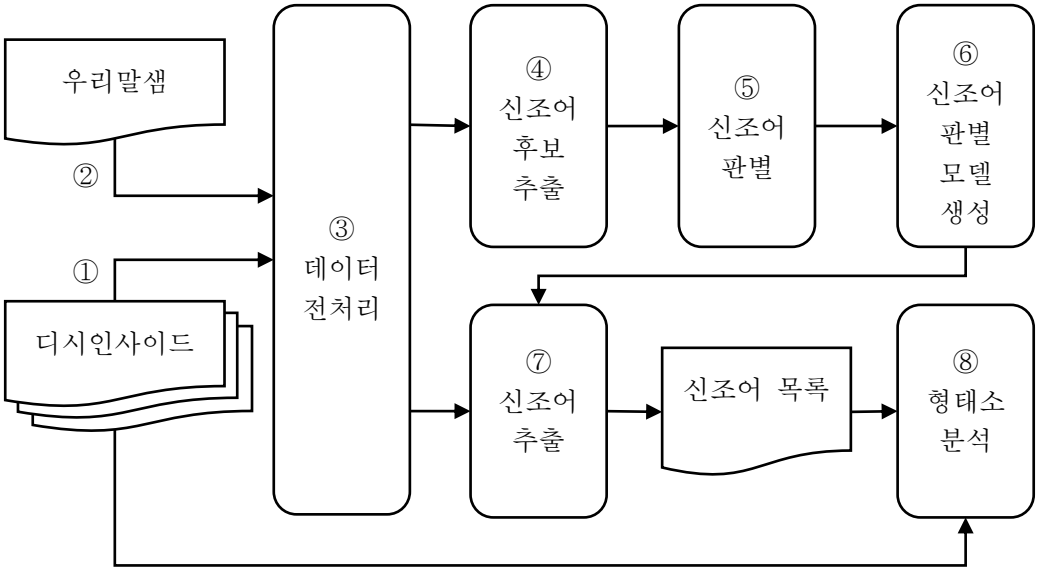
전까지는 하나의 단어로 인식하는 것이 어렵다. 하지만 신조어가 사용된 문장은 신조어가 핵심적인 의미를 지니고 있는 경우가 많기 때문에 이를 단어로 인식하는 것이 중요하다. 한글 데이터를 자연어 처리 방식으로 텍스트 마이닝하기 위해서는 형태소 분석 과정이 필요한데, 형태소 분석 프로그램이 신조어를 하나의 단어로 미리 알고 있지 않다면 의미 없는 형태소로 분해하는 오류를 일으킨다. 이로 인해 신조어가 가지는 의미를 잃어버리게 되고 올바른 텍스트 마이닝을 수행하기가 어려워진다. 따라서 텍스트 분석 이전에 텍스트로부터 신조어를 미리 추출하여 신조어 사전을 구축하고 구축된 사전을 형태소 분석에 이용하는 것이 중요하다고 볼 수 있다.

연세대학교
YONSEI UNIVERSITY

제3장연구 방법

본 연구는 다음 **<오류! 참조 원본을 찾을 수 없습니다.>**와 같은 순서로 진행되었다.

<그림 2> 연구 과정 및 방법



본 연구의 첫 번째 목적인 신조어 판별 모델을 생성하기 위해서는 모델을 구성하는 데이터로 독립변수와 종속변수가 필요하다. 독립변수로는 텍스트 데이터로부터 추출한 신조어 후보 단어와 각 후보 단어가 전체 텍스트 데이터에서 가지는 특징을 선정하였고, 선정된 특징에 대해서는 3.4에서 자세히 설명하였다. 종속변수로는 각 후보 단어가 신조어인지의 여부로 결정하였다.

신조어 판별 모델을 생성하고 검증할 데이터로 디시인사이드의 게시물을 선정하였다. 데이터를 확보하기 위해서 ① Python 프로그램을 작성하여 디시인사이드의

연세대학교
YONSEI UNIVERSITY

2018년 7월부터 9월까지 3개월간의 국내 야구 갤러리의 게시물 제목 데이터를 수집하였다. ② 그리고 신조어 여부를 판별하기 위해 우리말샘의 전체 데이터를 내려받았다. ③ 디시인사이드의 게시물 제목에서 특수기호와 'ㅋ'음절 등의 분석에 불필요한 부분을 제거하도록 전처리하였고 ④ 전처리된 데이터에서 어절별로 부분 글자를 추출하여 신조어 후보 단어를 구성하고, 각 후보 단어가 우리말샘 사전에 이미 등록된 단어라면 제거하는 방법으로 최종 신조어 후보 단어를 선정하였다. ⑤ 추출된 신조어 후보 단어에 대해 인터넷 검색과 신조어 후보 단어가 실제로 사용된 문장 검토를 통해 신조어 판별 모델을 위한 종속변수 데이터를 확보하였다. 총 719개의 후보 단어 중 약 40%에 해당하는 292개가 신조어로 확인되었고, 나머지 60%에 해당하는 427개는 신조어가 아닌 단어로 확인되었다. ⑥ 신조어 여부가 확인된 719건의 후보 단어들이 전체 텍스트 데이터에서 가지는 특징을 12가지 통계지표로 계산하여 독립변수로 사용하였다. 또한 719개 후보 단어들의 신조어 여부를 종속변수로 사용하여 로지스틱 회귀분석을 수행하여 신조어 판별 모델을 생성하였다.

⑦ 앞에서 생성된 신조어 판별 모델을 새로운 텍스트 데이터에 적용하여 그 텍스트로부터 신조어를 추출하여 사용자 사전을 구축하였다. ⑧ 사용자 사전을 사용하여 형태소 분석을 하고, 사용자 사전을 사용하기 전의 형태소 분석의 결과와 비교하였다.

3.1 데이터 선정

신조어를 탐지해낼 텍스트 데이터는 국내 온라인 커뮤니티 디시인사이드의 게시물을 대상으로 선정했다. 디시인사이드를 선정한 첫 번째 이유는 김효원(2016)에 의해 사용된 바 있는 SimilarWeb[17]의 국내 웹사이트 순위에서 2018년 11월 현재 디시인

[17] SimilarWeb(https://www.similarweb.com/top-websites/korea,-republic-of) : 국가별 웹사이

연세대학교
YONSEI UNIVERSITY

사이드가 전체 6위, 커뮤니티로는 1위이기 때문에 그만큼 많은 사용자가 사용한다고 볼 수 있기 때문이다. 이러한 온라인 커뮤니티를 분석대상으로 삼은 두 번째 이유는 이미 '헬조선'[18], '가즈아'[19] 등의 사회적으로 화제가 된 신조어가 이 사이트들에서 탄생했을 정도로 신조어 사용이 활발하게 이루어지기 때문이다.

디시인사이드는 갤러리라는 이름으로 주제별로 게시판을 나누고 있고, 갤러리마다 특정 소재를 가지고 관련된 게시물이 작성된다. 이 중에 국내 야구 갤러리[20]를 분석 대상으로 선정하였다. 국내 야구 갤러리는 디시인사이드 내에서 게시글이 가장 많이 작성되기 때문에 많은 사람이 글을 작성한다고 볼 수 있기 때문이다.

디시인사이드에 등록되는 게시물은 제목과 내용을 사용자가 작성하고, 이미지 파일을 첨부할 수 있다. 커뮤니티의 특성상 사용자가 이미지 파일을 첨부하기 위해 게시물을 작성하는 경우가 자주 있는데, 이럴 때는 내용이 없거나 매우 짧다. 그리고 많은 게시물이 내용이 없거나 2개 이내의 매우 짧은 단어로 이루어져 있거나 제목과 같은 텍스트를 가진 경우가 많다. 이런 이유로 게시물의 내용은 데이터 수집에서 제외하고 제목만 수집하여 데이터 수집의 효율성을 확보하였다.

3.2 데이터 수집 및 전처리

디시인사이드의 게시물은 Python 프로그램을 작성하여 국내 야구 갤러리의 목록 웹페이지의 html 문서를 http 프로토콜을 사용하여 수집하였다. 수집한 html 문서로부터 게시글 번호, 제목, 작성일시를 추출하여 csv 형태의 파일로 저장하였다. 이와

트의 방문자수, 트래픽 소스 등의 통계 및 랭킹 자료 제공

[18] https://namu.wiki/w/%ED%97%AC%EC%A1%B0%EC%84%A0

[19] https://namu.wiki/w/%EA%B0%80%EC%A6%88%EC%95%84

[20] 국내야구 갤러리 : http://gall.dcinside.com/board/lists?id=baseball_new7

연세대학교
YONSEI UNIVERSITY

같은 방법으로 국내 야구 갤러리에 2018년 7월부터 9월까지의 기간에 등록된 게시글을 수집하였다 (<<표 5> 참고).

<표 5> 디시인사이드 국내 야구 갤러리 수집 데이터 기간별 통계

	게시물수	어절수	음절수
2018 년 7 월	1,406,156	3,685,618	23,695,590
2018 년 8 월	1,435,144	3,816,141	23,411,540
2018 년 9 월	1,181,538	3,044,759	19,071,851
합계	4,022,838	10,546,518	66,178,981

디시인사이드의 게시물은 'ㅋ'음절이 전체 음절 중 38%를 넘게 차지할 정도로 매우 많이 등장한다. 그리고 단어와 단어 사이에 'ㅋ'를 넣는 유형이 발견되어 텍스트에서 'ㅋ'를 공백으로 치환하여 'ㅋ'를 기준으로 앞뒤 단어를 나눌 수 있게 <<표 6>과 같이 전처리하였다. 그리고 형태소 분석 시 특수기호는 별도의 형태소로 분리되므로 특수기호는 모두 공백으로 치환하였다.

<표 6> 전처리 전후의 텍스트 예시

전처리 전	ㅋㅋ 분위기 갑분싸 만들지 말고 ㅋㅋㅋㅋ 인싸특)카페앞에서 사진찍음
전처리 후	분위기 갑분싸 만들지 말고 인싸특 카페앞에서 사진찍음

전처리 후 공백을 기준으로 나눈 어절의 음절수 비율은 아래 <<표 7>과 같다. 2음절에서 6음절 사이의 어절이 89.4%를 차지하는 것으로 집계되었다.

연세대학교
YONSEI UNIVERSITY

<표 7> 디시인사이드 국내 야구 갤러리 수집 데이터 음절수별 통계

음절수	어절수	비율	비율
1	797,192	7.6%	7.6%
2	3,079,771	29.3%	89.4%
3	3,338,944	31.7%	
4	1,824,578	17.3%	
5	823,645	7.8%	
6	338,670	3.2%	
7 이상	314,101	3.0%	3.0%
합계	10,516,901	100.0%	100.0%

3.3 신조어 후보 추출

제2장에서 언급한 바와 같이 신조어 추출은 문장으로부터 단어를 추출하고, 추출한 단어 중 사전으로 등록되지 않은 단어를 찾는 과정이다. 따라서 신조어를 추출하기 위해서는 먼저 문장에서 단어를 추출하여야 한다. 문장의 어느 부분이 단어인지를 미리 알 수 없기 때문에 문장을 어절 단위로 분리하고 어절에서 추출 가능한 모든 부분 글자를 신조어 후보 단어로 추출한다. 이때 1글자로 이루어진 후보 단어는 김현중(2018)이 언급한 바와 같이 그 뜻을 판별하기가 어렵기 때문에 제외하였다. 또한 우리말샘의 통계지표[21]를 참고하면 7음절 이상의 단어의 명사는 전체 등록 어휘의 0.7%에 불과하기 때문에 계산의 효율성을 위해 후보 단어에서 제외하였다. <<표 8>는 문

[21] 우리말샘 사전통계 : https://opendict.korean.go.kr/service/dicStat#static_menu1_4

연세대학교
YONSEI UNIVERSITY

장에서 후보 단어를 추출한 예시이다.

<표 8> 문장으로부터 추출한 후보 단어 예시

문장	나 요즘 아싸에서 인싸됨
후보 단어	요즘, 아싸, 아싸에, 싸에, 아싸에서, 싸에서, 에서, 인싸, 싸됨, 인싸됨

후보 단어 중 너무 낮은 빈도로 등장하는 단어는 전체 텍스트에서도 중요하지 않은 단어일 확률이 높기 때문에 분석의 계산량을 줄이기 위해 특정 빈도수 이하의 단어는 후보 단어에서 제외하였다. 분석하고자 하는 텍스트 데이터가 큰 경우에는 큰 빈도수가 기준이 될 것이고, 데이터가 상대적으로 작은 경우에는 작은 빈도수가 기준이 될 것이다. 이렇게 텍스트 데이터의 크기에 맞추어 후보 단어에서 제외하는 기준을 결정하기 위해서, 기준 빈도수를 절댓값이 아닌 전체 어절수의 0.01%의 비율로 선정하였다. 수집한 데이터의 전체 어절수가 10,516,901개이므로 0.01%에 해당하는 1,051개가 후보 선정의 기준이 되는 빈도수가 된다. <표 9>은 위에서 추출한 예시 후보 단어에 대해서 디시인사이드에서 수집한 데이터에서의 빈도수와 빈도비율 및 그에 따른 최종 후보 단어 선정 여부이다.

<표 9> 후보 단어별 빈도수 및 빈도비율과 최종 선정 단어 예시

후보 단어	빈도수	빈도 비율(%)	선정 여부
요즘	**17,640**	**0.1677**	**선정**
아싸	**1,451**	**0.0138**	**선정**
아싸에	5	0.0000	제외
싸에	119	0.0011	제외
아싸에서	2	0.0000	제외

연세대학교
YONSEI UNIVERSITY

싸에서	47	0.0004	제외
에서	**44,723**	**0.4252**	**선정**
인싸	**8,085**	**0.0769**	**선정**
싸됨	37	0.0004	제외
인싸됨	10	0.0001	제외

게시글로부터 추출한 신조어 후보 단어가 이미 알고 있는 단어라면 신조어라 할 수 없을 것이다. 따라서 추출된 단어 중에서 사전에 등록된 단어를 제거하는 작업이 필요하다. 이때 사전에 등록된 단어의 목록이 필요한데, 이 단어 목록은 국립국어원의 우리말샘에 등록된 단어를 사용하였다. 우리말샘에는 표준국어대사전의 모든 단어가 포함되어 있기 때문에 이미 알고 있는 단어를 판별하기에 적합하다고 할 수 있다. 우리말샘에 등록된 사전 단어를 활용하기 위해서 우리말샘 웹사이트에서 제공하는 사전 내려받기 기능으로 2018년 10월 13일에 전체 등록 단어를 텍스트 형태로 내려받았다(<<표 10> 참고).

<표 10> 우리말샘으로부터 내려받은 파일 내용 예시

```
#00 표제어 시작
삼^유간「001」
#01 구분
어휘
#02 품사
명사
#04 고유어 여부
한자어
#05 원어
한자 三遊間
#09 검색용 이형태
3유간
```

연세대학교
YONSEI UNIVERSITY

```
#50 뜻풀이
야구에서, 3루수와 유격수 사이를 이르는 말.
#54 전문 분야
『체육』
#58 용례
000은 삼유간 깊숙한 타구를 날렸다.≪마이데일리 2016년 5월≫
000의 깊숙한 삼유간 땅볼 때에도 유격수 000이 타구를 걷어내 1루에서 아웃시켰다.≪오센 2016년 10월≫
#99 표제어 종료
```

우리말샘 데이터의 여러 항목 중에서 표제어와 검색용 이형태 항목을 사용하였다. 데이터의 형태가 '#번호 항목명'으로 어떤 항목에 대한 데이터인지를 먼저 표시하고 그 다음 줄에 실제 데이터를 표시하기 때문에 아래 <<표 11>와 같은 정규식을 사용하여 실제 사용할 데이터를 추출할 수 있다.

<표 11> 우리말샘 데이터 추출 정규식

구분	추출 정규식	추가 작업	추출 예시
표제어	#00 표제어 시작₩n(.*?)「	-, ^, 공백 제거	삼유간
검색용 이형태	#09 검색용 이형태₩n(.*?)₩n#		3유간

텍스트 데이터로부터 추출한 신조어 후보 단어는 총 2,933개이고, 이 중 1음절 단어 719개와 7음절 이상의 후보 단어 5개를 제외하면 2,209개의 후보 단어가 남는다. 이 중에 우리말샘에 등록된 단어 1,490개를 더 제외하면 719개의 단어가 최종 신조어 후보 단어가 된다.

연세대학교
YONSEI UNIVERSITY

3.4 변수 선택

본 연구에서는 추출된 후보 단어들이 가지는 특징으로부터 해당 후보 단어가 실제 단어로 쓰이고 있는지를 예측하는 단어 판별 모델을 생성하려고 한다. 따라서 텍스트 데이터로부터 각 후보 단어들의 특징을 독립변수로 추출하고, 각 후보 단어가 실제로 단어로 쓰이고 있는지를 종속변수로 하는 모델을 생성하였다.

텍스트 데이터로부터 각 후보 단어가 가지는 특징을 추출하기 위해 Python의 soynlp 패키지[22]의 WordExtractor[23]를 활용하였다. WordExtractor는 입력된 텍스트에서 음절 단위로 부분 글자를 생성하여 각 부분 글자마다 4가지 통계지표를 단어의 왼쪽 방향과 오른쪽 방향으로 각각 계산하여 총 8가지 통계량을 제공한다(김현중, 2018). WordExtractor가 생성하는 부분 글자가 본 연구에서 추출한 후보 단어와 같은 의미이므로, WordExtractor의 통계지표를 사용하여 후보 단어의 특징을 설명하는 것이 적절하다고 판단하였다(<<표 12> 참고).

<표 12> WordExtractor 제공 통계지표

특징	의미
left_frequency	단어가 어절의 왼쪽 부분에 등장한 횟수
right_frequency	단어가 어절의 오른쪽 부분에 등장한 횟수
cohesion_forward	어절의 왼쪽에서부터 단어가 함께 등장하는 정도
cohesion_backward	어절의 오른쪽에서부터 단어가 함께 등장하는 정도
left_accessor_variety	단어의 왼쪽에 등장하는 글자들의 종류

[22] Unsupervised Korean Natural Language Processing Toolkits : https://pypi.org/project/soynlp/

[23] WordExtractor : https://github.com/lovit/soynlp/blob/master/tutorials/wordextractor_lecture.ipynb

연세대학교
YONSEI UNIVERSITY

right_accessor_variety	단어의 오른쪽에 등장하는 글자들의 종류
left_branching_entropy	단어의 왼쪽에 등장하는 글자의 불확실성
right_branching_entropy	단어의 오른쪽에 등장하는 글자의 불확실성

종속변수의 설명력을 높이기 위해 **<오류! 책갈피가 자신을 참조하고 있습니다.>**과 같이 후보 단어들의 특징을 나타내는 통계지표를 추가로 계산하여 독립변수에 포함하였다. 우선 많은 신조어가 줄임말로 구성되기 때문에 2-3음절로 구성된 신조어가 4음절 이상의 신조어보다 더 많으리라 생각하여 후보 단어의 음절수를 독립변수에 추가하였다. 그리고 커뮤니티에서 자주 언급되는 신조어일수록 게시물 제목으로 많이 사용되리라 추측하여, 후보 단어가 전체 텍스트에서 등장한 횟수를 추가하였다. 또한 우리말 어절의 특성상 명사가 어절 왼쪽에 위치하므로 후보 단어가 등장한 어절 중에서 어절 왼쪽에 위치한 경우의 비율을 독립변수에 추가하였다. 이와 반대로 후보 단어가 등장한 어절 중에서 어절 오른쪽에 위치한 경우의 비율도 독립변수에 추가하였다. 어절의 오른쪽은 조사가 등장하는 위치이지만, 수집 데이터가 게시물의 본문이 아니고 제목이기 때문에 작성자가 게시글의 가독성을 높이기 위해 조사를 생략하는 경우가 많다고 판단하였기 때문이다.

<표 13> 후보단어의 특징을 설명하는 추가적 독립변수들

특징	계산방법
length	후보 단어의 길이
frequency	후보 단어가 전체 텍스트에서 등장한 횟수
starts_frequency	후보 단어가 어절의 왼쪽에 등장한 횟수
ends_frequency	후보 단어가 어절의 오른쪽에 등장한 횟수
starts_ratio	후보 단어가 어절의 왼쪽에 등장한 비율 start_frequency / frequency

연세대학교
YONSEI UNIVERSITY

ends_ratio	후보 단어가 어절의 오른쪽에 등장한 비율 end_frequency / frequency

최종 독립변수는 cohesion_forward, cohesion_backward, left_accessor_variety, right_accessor_variety, left_branching_entropy, right_branching_entropy, length, frequency, starts_frequency, ends_frequency, starts_ratio, ends_ratio의 총 12개로 선정하였다.

후보 단어들이 실제로 단어로 쓰이고 있는지에 대한 여부가 데이터로 존재하지 않기 때문에 이를 판단하는 과정을 수작업으로 진행하였다. 후보 단어를 인터넷에 검색하여 그 뜻이나 유래가 정리되어 있는 웹페이지가 있는 경우에는 신조어로 판별하였다. 그리고 해당 단어가 실제로 사용된 예시문을 보고 하나의 단어로 반복적으로 사용되고 있다면 단어로 판별하였다(<<표 14> 참고). 단어인 경우 '1', 단어가 아닌 경우 '0'으로 데이터를 생성하여 기존 데이터에 통합하였다.

<표 14> 수작업을 통해 신조어로 판별된 단어 목록

word, gisa, jyp, ppap, ㄱㄱ, ㄱㄲ, ㄷㄷ, ㄹㅇ, ㅁㅊ, ㅁㅌㅊ, ㅅㅌㅊ, ㅆㅅㅌㅊ, ㅇㄷ, ㅇㅇㄹ, ㅇㅈ, ㅇㅈㄹ, ㅉㅉ, ㅌㅅㅌ, ㅌㅌ, ㅍㅌ, ㅍㅌㅊ, ㅎㅎ, ㅓㅜㅑ, ㅗㅜㅑ, ㅜㅜ, ㅠㅠ, ㅡㅡ, 강혜원, 갤주, 고닉, 국거박, 국뽕, 권은비, 그알, 기어갑, 기즙, 김봉길, 김채원, 까비, 꼴데, 꿀쬠, 끄ㅏ울, 나봉쓰, 나야나, 나연, 넘글, 노꼴, 노사모, 노지선, 노짱, 노회찬, 느그흥, 다현, 담탐, 대깨문, 댕채, 더민주, 더쇼, 데뷔조, 도카콜라, 두니아, 든든이, 딸근, 떡락, 떡상, 뚜두뚜두, 라붐, 러버이, 러블, 러블리즈, 레드벨벳, 렛미, 루다링, 루순경, 류딸, 리멤버미, 막타, 망블, 매켜, 메갈, 모모랜드, 모쏠, 목런, 문재앙, 문재인, 미나야, 미루룽, 반전갑, 배그, 버0, 벅지, 번보, 번즈, 병모닝, 보나가스나, 부산대, 브이앱, 블랙핑크, 블핑, 블핑지츄, 빵린이, 뿌가, 뿌노스, 상받말, 샤샤샤, 설아, 설현, 세이콘, 섹꼴말, 섹록이, 섹리, 섹시안, 섹하영, 섹해인, 손아섭, 손흥민, 쇼미, 스브스, 스시녀, 스트, 스포, 슨판, 슬각, 슴골, 아니정, 아린이, 아육대, 아이린, 아이즈원, 아폭갑, 안양일시, 앰뒤, 앰흑, 야갤, 야갤러, 야고누나, 야구갑, 야념글, 야놀자, 야달소, 야드랑이, 야떼리, 야라니, 야로듀스,

연세대학교
YONSEI UNIVERSITY

야로미스, 야리둡, 야마이걸, 야무치, 야벅지, 야부차기, 야붕이, 야블리즈, 야와이스, 야요미, 야주소녀, 야지안, 야짤, 야쿠라, 야통령, 야프닝, 어우야, 에리이, 엠카, 여찐특, 예넨이, 예은, 오마이걸, 오우야, 오지배, 오지환, 와꾸, 우주소녀, 워너원, 워마드, 원영이, 위스플, 유능갑, 유시아씨, 윤서인, 이가은, 이달소, 인스타, 일베, 임나연, 입갤, 장원영, 정우람, 조보아, 조센징, 조유리, 조현우, 존나, 존못커플특, 존예, 졸커, 주멘, 주은아씨, 쥬리, 지루게이, 짓뚜, 짤녀, 짱개, 쮸쮸, 쯔위, 찐미나, 찐특, 채영, 채원이, 청하앙, 초졸아씨, 최예나, 카톡, 케록이, 케이다케이, 케이야, 코구, 코블린, 코승기, 코용수, 코창력, 쿠자, 퀸사, 퀸친, 퀸친이, 킹사, 킹탄, 타란티노, 탈코, 탑성, 통베, 퇴갤, 트와, 트와이스, 트퀴, 튼튼이, 틀가은, 틀딱, 틀은비, 파오후, 판람차, 팡수, 팩디, 퍄퍄, 페미, 프듀, 프듀48, 프로듀스, 프로미스, 프리파라, 핑까, 한곡갑, 한남충, 한녀, 할카스, 현명호, 혐지안, 호날두, 황의조, 황희찬, 후전드, 훈트, 훈틀, 흔드르라, 흙시, 흙특, 흥민, 희진, 히토미

연세대학교
YONSEI UNIVERSITY

제4장연구 결과

4.1 신조어 판별 모델

3장에서 확보한 719건의 데이터를 R 프로그램을 이용하여 로지스틱 회귀분석을 수행하였다. 신조어 단어 292개와 신조어 단어가 아닌 오류 단어 417개를 각각 무작위으로 7:3의 비율로 나누고, 70%의 데이터를 활용하여 신조어 판별 모델을 생성하고, 나머지 30% 데이터로 신조어 판별 모델의 정확도와 민감도를 검증하였다.

<표 15> 모델 생성시 사용 데이터 분류

신조어		오류 단어	
292		427	
분석용	검증용	분석용	검증용
204	88	299	128

로지스틱 회귀분석을 수행 시에는 12개의 모든 독립변수를 포함하여 초기 분석 모델을 만들고 유의성이 낮은 독립변수를 제거해 나가는 후진제거법을 실시하였다. 후진제거법은 기계적으로 최적화된 모델을 선정하는 방법이라는 한계점이 있다. 12개 독립변수 중 5개가 제거되고 7개의 독립변수가 선정되었다(<표 16> 참고).

<표 16> 신조어 판별 모델의 추정 결과

변수	Estimate	Std. Error	z value	Pr(>|z|)	

연세대학교
YONSEI UNIVERSITY

(Intercept)	-8.7316	1.4903	-5.8590	< .0001	***
length	-0.9497	0.1879	-5.0540	< .0001	***
cohesion_forward	3.6060	0.7216	4.9970	< .0001	***
left_branching_entropy	0.5473	0.2704	2.0240	0.0430	*
right_branching_entropy	0.8755	0.0888	9.8560	< .0001	***
left_accessor_variety	-0.0088	0.0033	-2.6510	0.0080	**
frequency	0.0002	0.0001	2.1730	0.0298	*
starts_ratio	6.7594	1.2481	5.4160	< .0001	***

본 연구에서는 신조어가 아닌 단어를 신조어라고 판단하는 경우보다, 신조어인데 신조어로 판별하지 못하는 경우에 잘못된 형태소 분석 결과가 만들어지므로 모델의 정확도만큼 모델의 민감도[24]가 중요하다고 볼 수 있다(<표 17> 참고).

<표 17> 신조어 판별 모델의 오분류표

분석용				검증용			
예측 / 실제	오류	신조어	오분류율	예측 / 실제	오류	신조어	오분류율
오류	235	61	20.6%	오류	108	23	17.6%
신조어	47	160	22.7%	신조어	16	69	18.8%
정확도 78.53%				정확도 81.94%			
민감도 77.3%				민감도 81.2%			

[24] Sensitivity : https://en.wikipedia.org/wiki/Sensitivity_and_specificity

연세대학교
YONSEI UNIVERSITY

본 연구에서 생성한 신조어 판별 모델은 이론적 논거에서 도출된 모형이라기보다는, 기존 연구들에서 신조어 판별 방식으로 사용할 수 있다고 주장되어 온 독립변수들을 모아서 후진제거법이라는 기계적 방법으로 최적화된 모형을 만든 것이다. 이 모델은 이론적 논거에 의해 만들어진 것은 아니지만, 기존 연구에서 신조어 판별 모델로 사용할 수 있다고 언급되어 왔던 변수들을 사용해서 최적 모형을 만들었다는 점에서 의의가 있다고 볼 수 있다.

4.2 실증적 적용

4.2.1 데이터 수집 및 전처리

앞에서 생성한 신조어 판별 모델을 실증적으로 적용하여 모델의 유용성을 확인하기 위해, 자연어 처리가 필요한 간단한 수준의 텍스트 마이닝 기법인 워드 클라우드를 생성하였다. 워드 클라우드는 텍스트에서 자주 언급되는 단어를 구름 형태의 이미지로 시각화하여 텍스트의 주제를 한눈에 알아보기 쉽다는 장점이 있다. 그리고 자주 언급되는 단어를 찾기 위해 한글 형태소 분석이 필요하므로 본 연구의 주제인 신조어 추출을 검증하기에 알맞다고 할 수 있다.

신조어 판별 모델 생성을 위해 사용했던 데이터 수집 프로그램을 동일하게 사용하여 디시인사이드 국내 야구 갤러리의 2018년 10월 게시물 제목을 수집하였다. 수집된 건수는 <표 18>에서 보듯이 총 1,246,404건이다. 수집한 데이터의 특수기호와 'ㅋ'음절을 제거하고, 전체 어절수의 0.01%에 해당하는 빈도수를 넘는 신조어 후보 단어를 추출하고, 우리말샘에 등록된 단어를 제거하여 최종 신조어 후보 단어를 722개 선정하였다.

연세대학교
YONSEI UNIVERSITY

<표 18> 디시인사이드 2018년 10월 수집 데이터 갤러리별 통계

	게시물수	어절수	음절수
국내야구 갤러리	1,246,404	3,339,603	19,447,573

4.2.2 신조어 판별

신조어 후보 단어들의 독립변수로 쓰일 7개의 통계지표를 계산하고, 모델을 적용해서 신조어와 신조어가 아닌 단어로 분류하였다. <표 19>는 신조어로 분류된 175개의 단어 중 신조어 판별 모델의 추정치가 높은 10개의 단어이다. 본 연구에서 생성한 신조어 판별 모델은 추정치가 0.5 이상일 때 신조어로 판별하므로, 추정치가 높을수록 신조어일 확률이 높다는 것을 의미한다.

<표 19> 국내야구 갤러리 2018년 10월 게시물 제목으로부터 추출된 신조어 목록

순위	신조어	추정치	순위	신조어	추정치
1	ㄹㅇ	7.38	6	ㅅㅂ	3.48
2	센세	4.11	7	킬건	3.44
3	쯔위	3.98	8	류딸	3.36
4	존나	3.87	9	ㅓㅜㅑ	3.21
5	씨발	3.79	10	롬쌔	3.18

4.2.3 형태소 분석과 워드 클라우드

앞에서 추출된 신조어를 활용하여 형태소 분석을 하기 위해, 신조어 목록을 꼬꼬마 형태소 분석 프로그램의 시스템 사전과 KOMORAN 형태소 분석 프로그램의

연세대학교
YONSEI UNIVERSITY

사용자 사전에 추가하였다. 각 형태소 분석 프로그램마다 신조어를 시스템 사전이나 사용자 사전에 추가하는 방법은 아래 <표 20>와 같다.

<표 20> 형태소 분석 프로그램의 사용자 사전 추가 방법

형태소 분석 프로그램	사용자 사전 추가 방법
꼬꼬마	kkma.jar 파일의 압축을 풀고 dic 디렉토리에서 .dic 확장자를 가진 파일에 사용자 사전을 추가
KOMORAN	Komoran 클래스의 setUserDic 메서드에 사용자 사전 파일을 지정

4.2.1에서 수집한 1,246,404개의 게시물 중 5,000건을 무작위으로 샘플링하여 제목을 형태소 분석하여 명사만 추출하여 많이 언급된 단어의 순위를 집계하였다. 그리고 4.2.2에서 추출한 신조어 175개를 형태소 분석 프로그램의 사용자 사전에 추가하여 같은 5,000건에 대해서 Java 프로그램으로 형태소 분석을 진행하여 명사만 추출하여 많이 언급된 단어의 순위를 집계하였다. 아래 <표 21>에서 이렇게 집계된 2세트의 순위표를 비교하였다.

<표 21> 자주 사용된 명사 20개

순위	신조어 사전 미사용	신조어 사전 사용	순위	신조어 사전 미사용	신조어 사전 사용
1	아씨	ㄷㄷ	11	이거	이것
2	케이	ㄹㅇ	12	ㅜㅑ	하노
3	오늘	케이	13	멸망	이거
4	노래	입갤	14	사람	아씨
5	여자	오늘	15	야로	ㅗㅜㅑ
6	이것	노래	16	전드	멸망

연세대학교
YONSEI UNIVERSITY

7	아이	**야붕이**	17	파울	사람
8	하노	여자	18	이유	**야로미스**
9	누나	누나	19	시간	전드
10	**미스**	**야갤**	20	나꼬	파울

신조어 사전을 사용하지 않았을 때는 포함되지 않은 'ㄷㄷ', 'ㄹㅇ', '입갤', '야붕이', '야갤'이라는 신조어가 신조어 사전을 사용한 분석 결과에서는 상위권에 포함되었다. 신조어 사전을 사용하지 않을 때는 12위 'ㅜㅑ'와 같이 'ㅓㅜㅑ'에서 반복되는 부분 글자가 포함되었으나 신조어를 사용할 때는 포함되지 않았다. 그리고 "야로"와 "미스"라는 단어는 신조어 사전을 사용하지 않을 때에 각각 15위, 10위로 포함되었지만, 신조어를 사용한 형태소 분석 결과에서는 제외되고, 대신 2개의 단어를 붙인 신조어 “야로미스”가 18위로 포함되었다.

신조어 사전을 사용하지 않았을 때에만 명사로 추출된 단어와 사용했을 때에 명사로 추출된 단어를 <표 22>와 같이 비교하였다.

<표 22> 신조어 사전 추가 전후, 각각의 경우에만 추출된 명사 목록과 순위

신조어 사전 미사용	(ㅜㅑ, 13), (야로, 16), (붕이, 32), (러블, 62), (이스, 64), (버츠, 103), (ㄷㄷㄷ, 104), (ㅁㅌ, 145), (러버, 146), (저스, 189), (좆크, 191), (여고, 205), (윤서, 211), (문재, 221), (이들, 333), (흔드르, 336), (ㄷㄷㄷㄷ, 355), (백종, 492), (ㄷㄷㄷㄷㄷㄷㄷㄷㄷㄷ, 495)
신조어 사전 사용	(야붕이, 7), (야갤, 10), (야로미스, 18), (아이즈원, 26), (센세, 37), (존나, 39), (할카스, 57), (애미, 67), (원영이, 89), (돌버츠, 92), (롬쌔, 96), (프로미스, 112), (엑시, 119), (노짱, 130), (상왕아씨, 137), (국저스, 152), (문재앙, 157), (킬건, 168), (뭐하노, 195), (킬기준, 204), (윤서인, 218), (꼴데, 223), (문재인, 228), (개죽이, 232), (하빵이, 235), (정공아씨, 238), (류현진, 243), (언럭키, 252), (겐세, 255), (넘글, 263), (븅신, 267), (번즈, 290), (일베, 302),

연세대학교
YONSEI UNIVERSITY

	(함하자, 308), (국뽕, 329), (러버이, 332), (페미, 333), (흔드르라, 352), (ㄱㄲ, 353), (킹탄, 378), (워너원, 410), (국거박, 447), (채원이, 449), (야죽이, 454), (성수좌, 463)

신조어 사전 추가 전에는 'ㅗㅜㅏ'의 'ㅜㅑ'가 잘못 명사로 추출되었지만 추가 후에는 'ㅗㅜㅑ'가 하나의 형태소로 분석되었다. '문재인'의 '문재'와 '백종원'의 '백종'에 대해서도 신조어 사전 추가 전에는 같은 오류가 있었다. '넘글', 'ㄱㄲ' 등의 단어는 신조어 추출 이후에만 명사로 추출되었다.

<그림 3> 신조어 사전을 사용하지 않은 분석 결과로 생성한 워드 클라우드

<그림 4> 신조어 사전을 사용한 분석 결과로 생성한 워드 클라우드

연세대학교
YONSEI UNIVERSITY

제5장결론

현대 사회에서는 급격히 증가하고 있는 비정형 데이터에 다양한 통계분석 방법을 적용해 우리 사회에 나타나고 있는 여러 가지 현상들을 의미 있게 설명하려는 연구들이 나타나고 있다. 문제는 비정형 데이터를 처리하는 과정에서 여러 가지 문제점들이 나타나고 있다는 것이다. 예를 들어, 텍스트 데이터에 대한 자연어 처리의 문제 및 한글 형태소 분석 과정 중에 급격히 증가하는 신조어로 인해 나타나는 미등록어 처리에 대한 문제가 존재한다. 이에 대한 해결책으로 본 연구는 신조어 판별 모델을 생성하여 텍스트 분석 시마다 본 연구에서 개발한 모델을 적용하여 신조어를 먼저 추출하고, 이를 다시 형태소 분석 시 활용하는 방안을 제시하였다.

신조어는 생성 당시의 사회상을 반영하는 단어라는 특징이 있기 때문에 텍스트 분석 시 무시할 수 없는 중요 단어이다. 또한 신조어는 새로 생겨난 말이라는 특성상 사전으로 구축되는 시점이 신조어가 생성된 시점과 차이가 크다. 또한 사회의 변화 속도가 빨라짐에 따라 신조어의 생성과 소멸도 빨라지는 경향이 있다. 한글은 영어와 달리 하나의 어절이 여러 형태소를 가지고 있기 때문에 자연어 처리 시 형태소 분석 과정이 필요한데, 문제는 수시로 등장하는 신조어가 사전에 추가되는 시점이 늦어질 수밖에 없다는 것이다. 따라서 형태소 분석 결과에 신조어가 하나의 형태소로 유지되지 못 하고 더 작은 형태소로 잘 못 분해되어 신조어의 의미를 잃어버리는 경우가 발생한다. 또한 신조어의 양이 급격하게 증가하고 있는 환경에서 신조어 사전을 미리 구축하여 텍스트 분석 시 활용하기에는 무리가 있다. 따라서 텍스트 분석을 진행할 때마다 분석하고자 하는 텍스트 데이터로부터 신조어를 추출하는 과정을 먼저 진행하여, 해당 텍스트 데이터만을 위한 신조어 사전을 먼저 구축하고, 이를 형태소 분석 시 다시 활용하는 방법이 텍스트 분석할 때 텍스트의 의미를 더 정확하게 파악하는 방법이 될 것이다.

본 연구에서는 텍스트 데이터로부터 신조어를 추출하는 방법으로 신조어 판별

연세대학교
YONSEI UNIVERSITY

모델을 제시하였다. 모델을 만들기 위해 국내 최대 온라인 커뮤니티인 디시인사이드의 국내 야구 갤러리의 2018년 7월부터 9월까지 3개월간의 게시물 제목 4,022,838건의 텍스트 데이터를 분석데이터로 선정하였다. Python 프로그램을 통해 사이트로부터 데이터를 수집하여 특수기호와 'ㅋ'음절을 제거하는 등의 전처리를 수행하였다. 그리고 각 어절의 부분 글자를 생성하여 그 빈도수가 전체 어절수의 0.01%를 넘는 단어를 신조어 후보 단어로 선정했다. 신조어 판별 모델의 독립변수는 각 후보 단어가 전체 텍스트 데이터에서 가지는 통계적 특징을 활용하였다. 간단하게는 후보 단어의 길이(length)와 빈도수(frequency)부터, 후보 단어가 각 어절에서 시작 부분에 위치하는 비율(starts_ratio), 끝 부분에 위치하는 비율(ends_ratio)을 독립변수로 선택하였다. 그리고 Python 패키지인 soynlp의 WordExtractor 클래스가 계산하는 통계지표 중 cohesion_forward, cohesion_backward, left_accessor_variety, right_accessor_variety, left_branching_entropy, right_branching_entropy을 독립변수로 선택하였다. 종속변수로는 신조어 후보 단어가 실제로 신조어인지를 파악해야 하는데, 이는 본 연구자가 각 후보 단어를 인터넷에 검색해 보거나 후보 단어가 사용된 데이터 내의 문장을 검토하여 판단하였다.

위에서 선택한 변수의 데이터를 R 프로그램을 통해 로지스틱 회귀분석을 수행하여 length, cohesion_forward, left_branching_entropy, right_branching_entropy, left_accessor_variety, frequency, starts_ratio의 7개의 변수로 구성된 신조어 판별 모델을 생성하였다. 형태소 분석 시에 신조어를 더 분해하거나 빠뜨리지 않고 포함하는 것이 본 연구의 궁극적인 목적이기 때문에 신조어가 아닌 단어를 신조어라고 판단하는 경우보다, 신조어인데 신조어로 판별하지 못하는 경우를 줄이는 것이 중요하다. 따라서 모델의 정확도만큼 민감도도 중요하다고 볼 수 있다. 본 연구에서 생성한 신조어 판별 모델의 정확도는 81.94%, 민감도는 81.2%이다.

본 연구에서 생성한 신조어 판별 모델을 실증적으로 적용해보기 위해 디시인사이드 국내 야구 갤러리의 2018년 10월 게시물 제목 1,246,404건으로부터

연세대학교
YONSEI UNIVERSITY

신조어 후보 단어를 추출하고 추출된 후보 단어마다 모델에 필요한 7개 독립변수를 계산하여 신조어를 175개 추출하였다. 추출한 신조어를 형태소 분석 프로그램의 시스템 사전 또는 사용자 사전에 추가하여 2018년 10월 게시물 중 무작위로 샘플링한 5,000건을 형태소 분석하여 데이터에서 자주 언급되는 명사 단어로 워드 클라우드를 생성하였다. 그리고 신조어를 추가하기 전의 형태소 분석 결과와 비교하여, 신조어 사전이 추가됨으로 인해 추출되지 않던 단어가 빈도수 상위 단어로 추출되고, 하나의 신조어가 나누어 추출되던 오류가 수정되는 것을 확인하였다.

본 연구가 가지는 학술적 시사점은 다음과 같다 첫 번째로는 연구를 위해 수집한 텍스트 데이터가 신문 기사가 아닌 국내 온라인 커뮤니티의 게시글이라는 점이다. 신조어를 조사하는 기존의 연구는 대부분 신문기사를 분석하여 신조어를 조사해왔다. 신문기사는 기자에 의해 본문이 작성되고 편집자의 감수를 거쳐서 공개되기 때문에, 누구나 익명으로 작성이 가능한 온라인 커뮤니티의 게시물에 비해 신조어 사용에 있어서 제한이 있는 것이 사실이다. 따라서 온라인 커뮤니티의 게시물에서 신조어를 추출하는 것이 신문기사에서 추출하는 것보다 더 풍부한 신조어 사전을 구축할 수 있을 것이다.

두 번째로는 한글의 자연어 처리 시 큰 문제 중의 하나인 미등록어 처리(즉, 신조어 처리)를 위한 명사 추출 방법을 신조어 판별 모델을 생성하고 적용하는 것으로 제시하였다는 점이다. 그간의 연구에서는 각 연구에서 제시하는 알고리즘을 이용하여 신조어를 판별하는 데에 비해, 본 연구에서는 분석하고자 하는 텍스트로부터 신조어 후보 단어를 추출하고 이를 기반으로 생성한 신조어 판별 모델로 신조어를 판별하였다. 텍스트 데이터로부터 추출한 신조어 후보 단어의 통계지표를 가장 적합한 형태의 수식으로 만들어 신조어 여부를 판별하기 때문에 연구자의 직관으로 결정한 판별 수식보다 더 높은 정확도를 가질 것으로 추정할 수 있다.

본 연구의 실무적 시사점으로는 첫째, 신조어 여부를 판별할 수 있는 신조어 판별 모델을 생성하였다는 것이다. 한글 텍스트를 자연어 처리하기 위해 형태소

연세대학교
YONSEI UNIVERSITY

분석을 수행하는 다른 연구에서도 본 신조어 판별 모델을 활용할 수 있다. 형태소 분석 전에 본 연구의 3.3에서 제시한 기준으로 신조어 후보 단어를 추출하고 각 후보 단어마다 7가지 통계지표를 계산하여 본 연구의 4.1에서 생성한 신조어 판별 모델을 적용하면 분석하고자 하는 텍스트 데이터만의 신조어 사전이 생성된다. 이를 연구자가 검토하여 신조어로 잘못 판별된 단어를 제외하면 신조어 사전을 효율적으로 구축할 수 있게 된다.

둘째로는 2018년 10월에 디시인사이드 국내 야구 갤러리에서 자주 언급된 주제를 추출하여 워드 클라우드 형태로 시각화했다는 점이다. 이때 신조어 판별 모델을 활용하여 신조어 사전을 구축함으로써 신조어가 빠지지 않고 포함되고, 이로 인해 신조어 사전을 사용하기 전에 비해 커뮤니티에서 언급된 주제를 보다 정확하게 파악할 수 있었다.

본 연구의 한계는 다음과 같다.

첫째, 본 연구에서 사용한 데이터가 디시인사이드라는 온라인 커뮤니티의 게시글이기 때문에 커뮤니티에 글을 작성하는 사용자의 언어적인 특성이 반영되어 있다. 디시인사이드는 디지털 카메라로 촬영한 사진을 공유하는 목적으로 시작하였기 때문에 다른 커뮤니티에 비해 게시물 1개의 텍스트 비중이 매우 낮고, 게시물을 문장을 갖추어 사용하는 경우보다 조사를 생략하면서 단어 위주로 게시글을 작성하는 경향이 있다. 따라서 텍스트의 어절 단위의 특성을 통계지표로 계산하여 독립변수로 사용하여 생성한 신조어 판별 모델은 이러한 게시글의 특성이 반영되어 있을 것으로 추정된다. 또한 국내 야구 갤러리에 작성된 게시물을 사용하였기 때문에 해당 갤러리에 자주 글을 작성하는 사용자의 관심사와 단어 선택의 특성이 모델에 반영되어 있을 것이다. 때문에 다른 온라인 커뮤니티의 게시글에 본 연구의 모델을 그대로 적용할 때에는 작은 크기의 텍스트 데이터로 테스트를 거쳐서 그 결과를 검토한 후 적용할 필요가 있다.

둘째, 커뮤니티의 게시글 본문은 수집하지 않고 제목만 수집하여 분석에 사용하였기 때문에 한 줄로 작성해야 하는 제목 텍스트의 특성이 모델에 반영되었을

연세대학교
YONSEI UNIVERSITY

것으로 추정한다. 따라서 커뮤니티의 본문 텍스트 데이터를 분석할 때에도 작은 크기의 데이터로 테스트를 거쳐서 그 결과를 검토 후 적용할 필요가 있을 것이다.

셋째, 신조어 판별 모델을 생성하기 위해서는 후보 단어의 신조어 여부를 판단하는 과정에 연구자의 수작업이 필요하다. 본 연구에서는 719개의 후보 단어를 연구자가 인터넷에 검색해보거나 후보 단어가 쓰인 문장을 확인하는 과정을 거쳐서 신조어 여부를 판단하였다. 문제는 본 연구에서 사용한 데이터보다 더 큰 크기의 텍스트 데이터에 대해 연구자의 수작업 과정을 거쳐 신조어 여부를 판단한다는 것은 상당한 시간이 걸린다는 것이다. 이는 신조어 판별 모델 생성을 자동화하고자 할 때에는 치명적인 단점이 될 수 있다. 하지만 전체에서 일부 데이터만 샘플링하여 적은 수의 신조어 후보 단어에 대해서 최초 1회에만 수작업으로 신조어 여부를 조사하고 나머지 데이터에 대해서는 생성한 신조어 판별 모델을 사용하는 등의 방법으로 수작업량을 최소화할 수 있을 것으로 기대한다.

넷째, 신조어 판별 모델을 로지스틱 회귀분석으로 생성하였는데, 분류 문제를 해결하는 다른 분석 방법을 이용하여 더 높은 정확도와 민감도를 가지는 모델을 생성하는 것도 향후 연구에서 다룰 수 있는 연구 문제가 될 것이다. 특히 독립변수가 비정형 데이터인 텍스트로부터 계산해낸 통계지표이기 때문에 비선형적인 특성이 있을 것으로 추정한다. 때문에 의사결정나무, 인공신경망, 딥러닝 등의 분석방법을 활용하여 신조어 판별 모델을 생성하여 그 결과를 본 연구와 비교해 보는 것도 의미가 있을 것이다.

연세대학교
YONSEI UNIVERSITY

참고 문헌

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

국립국어연구원 (2000). 국립국어연구원 10년사.
URL:https://www.korean.go.kr/front/page/pageView.do?page_id=P000168&mn_id=76

국립국어원 (2011). 국립국어원 20년사.
URL:https://www.korean.go.kr/front/page/pageView.do?page_id=P000215&mn_id=77

김동의, 이상곤 (2013). 신어를 찾아내고 의미를 기술하여 관리하는 신어 조사용 프로그램의 설계 및 구현. <정보과학회논문지 : 소프트웨어 및 응용>, 40권 12호, 882-894.

김성용 (1987). <TABULAR PARSING 방법과 접속 정보를 이용한 한국어 형태소 분석기>. 한국과학기술원 석사학위 논문.

김일환 (2014). 신어의 생성과 정착. <한국사전학>, 24호, 98-125.

김태훈, 박상진 (2011). 신어의 정착 연구. <한국어 의미학>, 35호, 71-98.

김현중 (2018). wordextractor_lecture.
URL:https://github.com/lovit/soynlp/blob/master/tutorials/wordextractor_lecture.ipynb

김환, 임진희 (2017). 신조어를 활용한 사회적 현상 아카이빙 방안 연구. <기록학연구>, 52호, 322-349.

김효원 (2016). <텍스트 마이닝을 활용한 웹사이트 특성 비교 분석>. 광운대학교 대학원 석사학위 논문.

연세대학교
YONSEI UNIVERSITY

남길임 (2014). <2014년 신어>. 서울: 국립국어원

남길임, 이수진, 최준 (2017). 대규모 웹크롤링 말뭉치를 활용한 신어 사용 추이 조사의 현황과 쟁점. <한국사전학>, 29호, 72-106.

문금현 (1999). 현대국어 신어(新語)의 유형 분류 및 생성 원리. <국어학>, 33호, 295-325.

박봉래, 황영숙, 임해창 (1998). 용례 분석에 기반한 미등록어의 인식. <정보과학회 논문지>, 25권 2호, 397-407.

방진우 (2017). <역사 자료 형태분석에서 미등록어 추정과 분석 중의성 해소>. 연세대학교 대학원 석사학위 논문

송민 (2017). <텍스트 마이닝>. 서울: 도서출판청람.

원중호, 이한별, 문혜정, 손원 (2017). 텍스트 마이닝 기법을 이용한 경제심리 관련 문서 분류.
URL:https://www.bok.or.kr/portal/bbs/B0000233/view.do?nttId=233885&menuNo=200707&pageIndex=1

윤경선. (2013). 소셜미디어 트위터(Twitter)로 살펴본 신어 형성. <한국어 의미학>, 42호, 537-555.

Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of ‘big data’. *McKinsey Quarterly, 4(1)*, 24-35.

Chakraborty, G., Pagolu, K. M. (2014). *Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining*.
URL:http://support.sas.com/resources/papers/proceedings14/1288-2014.pdf

Kroeger, P. R. (2005). *Analyzing grammar: An introduction*. Cambridge, England: Cambridge University Press.

연세대학교
YONSEI UNIVERSITY

Özköse, H., Arı, E. S., & Gencer, C. (2015). Yesterday, today and tomorrow of big data. *Procedia-Social and Behavioral Sciences, 195*, 1042-1050.

Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text mining: Techniques, applications and issues. *International Journal of Advanced Computer Science & Applications, 1(7)*, 414-418.

연세대학교
YONSEI UNIVERSITY

Abstract

Extraction Method of New Word from Online Community: The Application of New Method to Morphological Analysis

Kim, Hanjoon

The Graduate School of Information

Yonsei University

A new-word is an important factor when you implement text analysis because new words are produced rapidly as times go by. The process of new words' appearance and disappearance are as fast as the change of society. Unlike English, Hangul has many morphemes in one word, so morphological analysis is essential in natural language processing. Without using dictionary including newly produced words, you cannot analyze an appropriate morphological analysis because newly produced words might be broken down into smaller morphemes, and thus losing its meaning. When you analyze morphological analysis without including newly produced words, you cannot expect to get good results from morphological analysis.

In addition, it is not easy to construct a dictionary that includes rapidly produced new words. If the text data contains a previously unknown new word, the new word will not be extracted into one morpheme. Therefore, it is necessary to construct a dictionary that includes new word from the text data if you analyze text analysis. This dictionary should be used again for morphological analysis

연세대학교
YONSEI UNIVERSITY

process.

In the study, a new-word discrimination model was proposed as a method of extracting new-word from text data. To make a model, 3 month(from July to September 2018) post titles are collected from Domestic Baseball Gallery of www.dcinside.com, the largest online community in Korea. After preprocessing the collected data, all possible partial characters in each word were generated and partial characters with frequency more than 0.01% of total number of words were selected as new-word candidates word.

The statistical characteristics of each candidate word in the whole text data were used as independent variables. And the logistic regression analysis was performed with dependent variables as the actual new-word. As a result of the analysis, the model is constructed with 7 variables and the accuracy of the model is 81.94% and the sensitivity is 81.2%.

In order to apply the model empirically, the title of the post in October 2018 from the site were collected. 175 new-words were extracted from the text by calculating 7 statistical variables for each candidate word. The extracted new-words were added into the morpheme analysis program as a user dictionary. And two word clouds were generated with most frequently mentioned nouns from 5,000 randomly selected post titles. One was with new-word dictionary, and the other was without new-word dictionary. It was confirmed that the nouns that were not appeared in word cloud without new-word dictionary are appeared in word cloud with new-word dictionary.

Key words : text mining, morphological analysis, new word, online community, user dictionary