텍스트 전처리를 위한 불용어 목록의 구축

왕인서[O1], 변석우[2], 우균[1]
[1]부산대학교 정보컴퓨터공학과
[2]경성대학교 소프트웨어학과
inseowang@pusan.ac.kr, swbyun@ks.ac.kr, woogyun@pusan.ac.kr

Building a List of Stopwords for Text Preprocessing

Inseo Wang[O1], Seokwoo Byun[2], Gyun Woo[1]
[1]School of Information Computer Engineering Pusan National University,
[2]School of Software Engineering Kyngseong University

요 약

한국어 자연어 처리에서 표준화된 불용어 목록은 미비한 상태다. 본격적인 자연어 처리를 위해 텍스트 데이터를 전처리하는 단계는 필수적인데, 유의미한 토큰만을 남겨두기 위해 불용어의 제거는 중요하다. 본 논문에서는 특정 도메인의 텍스트 데이터 집합에서 불용어 목록을 구축하여 단어 집합을 생성하고, 구축한 불용어 목록을 사용하여 세 개의 다른 도메인에 존재하는 텍스트 데이터 집합의 불용어를 정제한다. 구축한 불용어 목록을 네이버 영화, 세종 코퍼스, 청와대 국민청원 데이터 집합에 적용한 결과, 각각 88.50%, 84.65%, 77.34%의 정제율을 보였다.

1. 서론

텍스트 전처리 단계에서 불용어의 제거는 중요하다[1]. 의미 분석에 기여하지 않는 불용어들을 제거하고 실질적 의미가 있는 어휘만을 선정해서 기계에 학습시켜야 실제 텍스트의 의미에 접근할 수 있기 때문이다. 텍스트 데이터는 정형화된 수치 자료와 달리, 그 자체로 변이성을 가지는 비정형 데이터이다. 텍스트 전처리는, 텍스트 데이터를 기계에 학습시키기 위해 정형화된 데이터로 변환하는 과정이라고 할 수 있다. 이 과정에서 의미 분석에 유의미하게 작용하지 않는 어휘들은 불용어 목록을 통해 정제되고, 정제되지 않은 나머지 어휘들로 단어 집합을 구성하여 텍스트 데이터를 정수로 대응시키는 과정을 거친다.

Christopher Fox가 발표한 A Stop List는 영어의 대표적인 불용어 목록이다[2]. 이는 다수의 문학 자료에서 추출한 텍스트 데이터 집합 Brown Corpus를 활용하여 의미론적으로 필요하지 않은 어휘 421개를 선정한 것이다. 하지만 한국어 자연어 처리에서, 이처럼 일반적으로 사용할 수 있는 불용어목록에 대한 연구는 부족한 상태이다[1]. 영어 자연어 처리를 위한 범용 불용어 목록인 'A Stop List'가 1990년에 등장한 것과는 뚜렷하게 대조된다. 다만 다른 분야에서는 일반적으로 사용할 수 있는 한국어 불용어에 관한 연구가 여전히 진행되고 있다.

본 논문에서는 문장 분석 시 의미 분석에 크게 유효하지 않은 품사를 개별적으로 선정하여 다수의 문장으로 이루어진 텍스트 데이터 집합에서 불용어 품사에 일치하는 어휘들을 추출하여 불용어 목록을 구성한다. 이후 동일 도메인의 데이터 집합에서 얼마만큼 원하는 불용어를 정제하는지 실험한다. 2절에서는 불용어 목록 구현을 위한 프로그램과 구성 기법에 관해 기술하고, 3절에서는 불용어 목록 구현 및 목록 적용에 관해 기술하고 4절에서는 적용 결과를, 5절에서는 적용 결과를 바탕으로 한 토의를, 6절에서는 결론을 기술한다.

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2014-3-00035, 매니코어 기반 초고성능 스케일러블 OS 기초연구 (차세대 OS 기초연구센터)).

*교신 저자: 우균(부산대학교, woogyun@pusan.ac.kr).

2. 불용어 목록 구현을 위한 준비

2.1 한국어 형태소 처리 패키지

불용어 목록 구현에 앞서 문장의 형태소를 분석하기 위해 KoNLPy[3]를 사용하였다. KoNLPy는 2014년 서울대학교 산업공학부의 박은정과 조성준에 의해 개발된 Python에서 사용할 수 있는 한국어 정보 처리를 위한 패키지이다. 해당 패키지는 특정 문구의 용례 검색을 지원하며, 꼬꼬마, 한나눔, MeCab-ko, Komoran, Okt를 활용하여 형태소 분석, 품사 태깅 기능 또한 존재한다.

본 논문에서 실시할 불용어 목록 구현을 위해서, Okt 분석기를 사용하였다. Okt는 KoNLPy에 내장된 다른 형태소 처리기보다 형태소 분석 속도가 빠르며, 다른 처리기에 비해 형태소 분석 시 태깅하는 품사의 범위가 포괄적이어서, 사용자가 직관적으로 확인할 수 있는 분석 결과를 제시해주었기에 Okt 처리기를 사용하였다.

2.2 불용어 목록 구성 기법

불용어 목록의 구성 방법은 크게 두 가지로 나뉜다. 첫 번째는 언어학적 분석 기법을 통하여 추출하는 방법이며, 두 번째는 어휘의 출현 빈도를 기준으로 추출하는 방법이다. 전자의 분석 방법은, 문장을 분석하는 단계에 따라 불

용어 구성이 다를 수가 있다. 후자의 분석 방법은 한 언어에서 공통으로 사용되면서 고빈도로 등장하는 어휘를 불용어로 선정할 수 있다[4].

본 논문에서 구성할 불용어 목록은 후자의 방법을 채택한다. 후자의 방법을 선택하여 불용어 목록을 구성했던 선행 연구로는, 권호경의 논문[5]이 있다.

권호경의 논문에서는 불용어의 가중치를 산정하는 통계적인 기법을 사용하여 불용어 목록을 구성하였다. 불용어 어휘의 선정 방법은, 대량의 문헌으로부터 추출한 어휘에서 고빈도로 출현하는 어휘를 선정하여, 중요 어휘를 제거하고 중, 저빈도의 단어 중 불필요한 어휘를 첨가하는 방식으로 불용어를 선정하였다.

본 논문의 불용어 목록 구현 역시 소개했던 논문에서 채택했던 방법과 유사하게, 문장에서 고빈도로 출현하면서, 문법적 기능을 담당하는 품사들을 선정하여 일반적으로 사용할 수 있는 불용어 목록을 구성한다.

3. 불용어 목록 구축

3.1 불용어 목록 구현을 위한 품사 선정

불용어 선정에 앞서 어떤 품사의 어휘를 불용어로서 선정할 것인지에 대한 논의가 필요하다. 본 실험에서는 의미 분석에 불필요하다고 판단되는 형식 형태소들을 중심으로 불용어 품사를 선정하였다. 부사와 감탄사, 조사, 접속사를 불용어 선정 품사로 선택하였다. 선정된 품사는 특정 품사를 수식(부사)하거나 문장에서 큰 의미 없이 독립적으로 존재(감탄사)하며, 혹은 실질적인 의미를 담당하는 어휘에 부착(조사)되어 사용되는 품사들이다. 해당 품사들을 먼저 제거할 수 있다면 향후 텍스트 의미 분석을 수행하는데 크게 도움이 될 것으로 기대된다.

3.2 불용어 목록과 단어 집합 생성

불용어 목록 구성을 위해 문장 단위의 말뭉치들을 모은 데이터 집합을 분석한다. 실험에는 Python을 활용하였다. 또한 형태소 분석을 위해 KoNLPy의 Okt 형태소 처리기를 활용하였다.

실험을 위한 데이터 집합은 도서 한국어 임베딩[6]에서 지원하는 튜토리얼 페이지의 네이버 영화 말뭉치 테스트 셋을 사용하였다. 불용어 목록 구성을 위한 실험 절차는 그림 1과 같다.

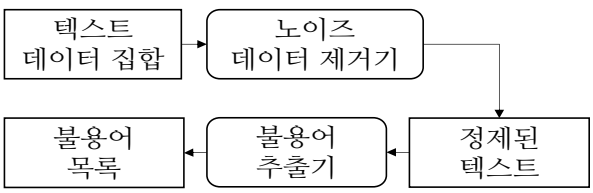

그림 1. 실험 절차

불용어 목록 구현을 위해 사용할 텍스트 데이터 집합은 신문 기사, 공문서, 논설문과 같은 공적인 성격이 옅기 때문에 띄어쓰기와 맞춤법이 정확하지 않은 문장들이 포진해있을 가능성이 매우 크다. 뿐만 아니라 이모티콘과 영어, “ㅋㅋㅋㅋ”, “ㅎㅎㅎㅎ”와 같이 의성어를 나타내기 위한 요소가 많다. 따라서 노이즈 데이터들을 제거하기 위한 정제 단계를 먼저 실시한다. 노이즈 데이터 제거가 끝난 문장은 형태소 분석을 진행한다. 그리고 앞서 선정한 불용어 품사에 해당되는 어휘가 등장하면 불용어 목록에 추가한다.

불용어 목록의 구성은, 문장과 사전에 결정해 놓은 품사 이름을 모은 리스트와 실제 불용어를 구성할 리스트를 입력으로 받는 함수를 통해 생성한다. 입력받은 문장을 형태소 분석하여 분석된 어휘들의 품사가 불용어로 선정한 품사와 일치하는지를 확인한다. 이후 불용어 리스트에 넣을 어휘가 이전에 분석했던 어휘인지를 확인하여 중복되지 않았다면 불용어 리스트에 추가한다. 이를 구현한 알고리즘은 그림 2와 같다.

```
입력: 문장, 품사 리스트, 불용어 리스트
출력: 갱신된 불용어 리스트

1. 입력받은 문장을 형태소 분석
2. 불용어 저장을 위한 불용어 리스트 생성.
3. 분석된 문장의 어휘 수만큼 아래 단계 반복
    3-1. 어휘가 품사 리스트에 포함 여부 확인
    3-2. 어휘가 불용어 리스트에 있는지 확인
    3-3. 위 조건을 만족하면 불용어 리스트에 추가
4. 갱신된 불용어 리스트 반환
```

그림 2. 불용어 목록 생성을 위한 알고리즘

불용어 목록이 구성되면, 생성된 불용어 목록을 활용하여 새로운 데이터 집합의 말뭉치들을 형태소 분석한다. 불용어 목록에 일치하는 어휘들은 제거되고, 제거되지 않은 어휘들은 단어 집합에 추가된다.

단어 집합은 Python의 기본 자료 구조인 사전으로 구현하였다. 어휘가 단어 집합의 키가 되며, 해당 어휘의 등장 횟수가 단어 집합의 값이 된다. 또한, 불용어로 선정되지는 않았지만, 길이가 짧아 의미 분석에 크게 유효하지 않은 어휘들도 단어 집합 생성 단계에서 정제한다.

3.3 불용어 목록의 적용

네이버 영화 댓글 데이터 집합에서 추출한 불용어 목록은 총 530개다. 이 중, 접속사가 18개, 감탄사가 23개, 조사가 170개, 부사가 319개로 가장 많았다. 공적인 성격이 옅은 문어체의 특성상 개인의 감정과 가치판단을 표현하기 위해 다양한 어휘들이 등장했을 가능성이 크다. 그 중, 의태어나 의성어를 비롯하여 문장 생성에 직접적으로 참여하지 않으면서, 다른 문장 성분을 수식하는 품사, 즉 부사가 불용어로 가장 많이 추출되었음을 알 수 있다.

이전 단계에서 생성한 불용어 목록을 활용하여 다른 도메인의 데이터 집합에서 단어 집합을 생성한다. 신규

로 생성된 단어 집합에서 앞서 불용어로 선정했던 품사들이 얼마만큼 정제되었는지를 확인할 수 있다면 이전 단계에서 구성한 불용어 목록의 유효성을 어느 정도 측정할 수 있을 것이다.

분석을 진행할 신규 데이터 집합은, 네이버 영화 말뭉치 데이터 집합[6]과 청와대 국민청원[7], 세종 코퍼스 데이터 집합[8]을 사용하였다. 신규로 분석할 데이터 집합은 10만 개 이상의 문장이 포함되어있다. 단어 집합을 생성하기 위한 절차는 기존의 실험과 동일하다.

4. 적용 결과

본 논문에서는 자체적으로 생성한 불용어 목록을 서로 다른 도메인의 텍스트 데이터 집합에 적용하여 단어 집합을 생성하였다. 신규 데이터에 불용어 목록을 적용하여 각 도메인에서의 불용어 정제율을 아래의 표 1로 나타내었다.

표 1. 불용어 목록 적용 결과(단위:%)

불용어 품사	영화 리뷰	세종코퍼스	국민청원
데이터 집합1	100.00	84.31	77.35
데이터 집합2	100.00	91.52	76.62
데이터 집합3	74.12	77.44	76.70
데이터 집합4	79.92	85.32	78.70
평균 정제율	88.50	84.65	77.34

총 세 개의 다른 도메인을 선정하여 도메인별 네 개의 데이터 집합마다 이전 단계에 생성했던 불용어 목록을 적용하였다. 그 결과, 불용어 목록을 생성했던 동일 도메인의 데이터 집합에서는 높은 정제율을 보였으며, 세종 코퍼스의 정제율은 84.65%의 정제율을 보였다. 세종 코퍼스 도메인에서는 이전 도메인들과 비교했을 때 상대적으로 낮은 정제율을 보였으나 70%가 넘는 정제율을 보였다.

5. 토의

본 논문에서 구성한 불용어 목록에서 불용어로서 예상된 어휘와 불용어로 추출되어선 안 되는 어휘 몇 개를 추려 아래의 표 2로 나타내었다.

표 2. 예상 불용어와 부적합 불용어

예상 불용어		부적합 불용어	
어휘	품사	어휘	품사
그러다가	접속사	쓸데없이	부사
그래서	부사	왜냐하면	부사
밖에도	조사	난데없이	부사
스럽게	조사	어영부영	부사
그렇다고	접속사	때때로	부사

예상되는 불용어로 선정한 어휘들은 단순히 문장 사이의 연결 관계를 의미할 뿐 문장의 의미 분석에 크게 기여하지 못한다. 해당 어휘들은 불용어로서 적합하다.

불용어로서 부적합한 어휘들은 문장 사이의 인과관계(왜냐하면), 빈도(때때로), 어떤 현상의 상태나 글쓴이의 가치판단(어영부영, 난데없이, 쓸데없이)을 나타내었다. 해당 어휘들은 의미 분석에도 유효하게 작용할 수 있으므로 불용어로 선택되는 것은 부적합하다.

본 논문에서 구현한 불용어 목록을 통해 향후 자연어 처리 과정에서 얼마만큼 실질적인 의미에 접근할 수 있는지를 측정할 수 있다면 범용 불용어 목록으로서의 유효성을 확인할 수 있을 것으로 기대된다. 다만, 본 논문에서는 서로 다른 도메인의 데이터 집합에서 통용되는 불용어의 유무와 그의 정제에 집중하고자 했기에 추가 실험은 진행하지 않았다.

6. 결론

본 논문에서는 문법적 기능을 담당하는 품사들을 선정하여 네이버 영화 말뭉치 데이터 집합에서 불용어를 추출하였다. 구성한 불용어 목록을 활용하여 다양한 데이터 집합에서 불용어 정제율을 확인하여 다양한 상황에서 적용 가능한 불용어 목록을 구성하고자 했다. 그 결과 모두 평균 83.50%의 유의미한 정제율을 보였다.

대부분 상황에서 사용할 수 있는 범용 불용어 목록의 미비는 자연어 처리 과정에서 번거로움으로 작용한다. 본 논문의 실험을 통해, 문법적 기능을 담당하는 품사에 한하여 여러 도메인에 적용 가능한 불용어 목록을 어느 정도 구성할 수 있을 것으로 기대된다.

참 고 문 헌

[1] 길호현, “텍스트마이닝을 위한 한국어 불용어 목록 연구, ” *우리말글*, Vol. 78, pp. 1-26, 2018.

[2] Fox Christoper, “A Stop List for General Text,” *Acm sigir forum*, Vol. 24, No. 1-2, pp. 19-21, 1989

[3] 박은정, 조성준, “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지,” *제26회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 133-136, 2014.

[4] 김판구, 조유근, “한국어 정보 검색을 위한 불용어의 구성 및 적용,” *한국정보과학회 학술발표논문집*, 제20권 제1호, pp. 809-812, 1993.

[5] 권호경, 이상훈, 박미영, 이승우, 한기태, 서창덕, 임인칠, “통계정보를 이용한 가중치 부여 불용어 사전의 구성,” *한국정보과학회 학술발표논문집*, 제23권 제1호(A), pp. 903-906, 1996.

[6] 이기창, *한국어 임베딩*, 3판, pp. 347, 에이콘, 2019.

[7] 김현중, (2019, 10, 7) [Online]. Available:https://github.com/lovit/petitions_dataset

[8] 정국재, (2019, 10, 12) [Online]. Available:https://github.com/jeongukjae/sejong-downloader