자연어 처리 기반 텍스트 마이닝을 위한 한글 어간 추출 알고리즘

최기원* · 최성훈* · 조상현* · 김희철**

*인제대학교 디지털 항노화 헬스케어학과 대학원

Hangeul Stem Extraction Algorithm for Text Mining Based on Natural Language Processing

Ki-won Choi* · Seong-hun Choi* · Sang-hyeon Jo* · Hee-cheol Kim**

*Inje-University

Institute of Digital Anti-aging Healthcare

E-mail : kiwon2819@naver.com* · shuhoony@naver.com*

요 약

텍스트 마이닝의 기반이 되는 자연어 처리는 언어의 종류에 따라 처리 방법이 다를 수 있다. 특히 타 언어에 비해 비교적 표현의 자유도가 높은 한글은 어미의 활용에 따라서 여러 가지 단어의 형태가 존재한다. 이처럼 다양한 형태로 굴절하는 단어에서 변화하지 않는 부분을 어간이라고 하며, 효과적인 텍스트 마이닝을 위해선 어간을 추출하여 다양한 형태의 단어들을 단일화하는 과정이 필수적이다. 따라서 본 논문에서는 한글문서의 효과적인 텍스트 마이닝을 위하여 한글 어간 추출 알고리즘을 제시한다.

ABSTRACT

Natural language processing, which is the basis of text mining, differs depending on the type of language. Especially, Hangeul, which has relatively high freedom of expression compared to other languages, has various forms of words depending on the use of ending. The part that does not change in these various forms of words is called the stem. For effective text mining, it is essential to extract words and unify various types of words. Therefore, this paper proposes an extraction algorithm for Hangul word for effective text mining of Hangul document.

키워드

텍스트 마이닝, 자연어 처리, 어간 추출, 스테밍 알고리즘

Ⅰ. 서 론

"빅 데이터"란 단어는 이제 IT업계를 벗어나 비IT종사자들에게도 익숙한 단어가 되었다. 스마트폰의 등장으로 컴퓨터를 통해 고정된 장소에서 이루어지던 작업은 장소를 구애받지 않게 되었고, 소셜 네트워크 서비스의 등장으로 개개인이 정보를 생성하고 전달할 수 있게 되면서 수많은 정보가 범람하며 빅 데이터라는 개념이 생기게 되었다. 빅 데이터는 기존의 정형화되어있는 데이터는 물론 가공되지 않은 비정형데이터까지 포함하는 개념인데, 빅 데이터의 대부분을 차지하는 비정형데이터는 가공되지 않았기 때문에 비정형데이터 자체로는 특별한 정보를 얻어낼 수 없었다. 따라서 방대한 데이터를 처리 및 분석하기 위한 방법이 필요했고 이 방법이 바로 "데이터 마이닝"이다.[1]

데이터 마이닝은 대규모로 저장된 데이터 안에서 숨겨져 있는 상관관계를 발견하여 유의미한 정보를 추출해내는 과정을 뜻한다. 데이터 마이닝

은 다루는 데이터의 대상에 따라 여러 종류로 나뉘는데 본 논문에서는 문자를 대상으로 하는 “텍스트 마이닝”에 대해서 다룬다.[2]

텍스트 마이닝은 구조화되지 않은 대규모의 텍스트 집합으로부터 유의미한 정보를 추출해내는데, 정보를 추출하기 위해선 일련의 전처리 과정이 필요하며 그 과정 중 하나에 해당되는 것이 “스테밍 알고리즘”이다.

스테밍 알고리즘은 어간을 추출해내는 과정이다. 같은 의미를 가진 단어라도 어미의 활용에 따라서 여러 가지 활용어가 존재하는데, 다양한 형태를 가진 활용어에서 변화하지 않는 부분을 어간이라고 한다. 텍스트 마이닝을 위해서는 스테밍 알고리즘을 동반한 전처리 과정을 거쳐 텍스트에서 의미 없는 부분은 제거하고 의미를 가지는 어간만을 추출해야 한다.[3]

하지만 대다수의 어간 추출 알고리즘은 영문 중심으로 연구되고 있으며, 한글 특유의 중의적인 의미를 가진 단어나 용언의 불규칙한 활용 등, 자연어 처리의 어려움으로 인하여 한글을 위한 어간 추출 알고리즘(Stemming Algorithm)은 부족한 실정이다.[4]

따라서 본 논문에서는 한글 텍스트 마이닝을 위한 스테밍 알고리즘을 제시한다. 또한 추출한 어간의 데이터베이스화와 문서화의 두 가지 방법을 비교하여 최적의 방법을 연구한다.

II. 어간 추출 방법

본 논문에서는 올바른 어간 추출을 위하여 국립국어원의 통합자료실에서 제공하는 말뭉치 파일 참조하였다. 말뭉치 파일을 토대로 어절에 대한 어간 정보를 추출하고, 추출한 정보를 다시 데이터베이스화 하는 과정을 거치게 된다.

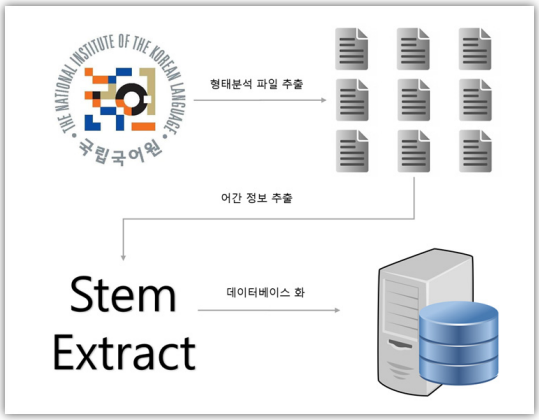

그림 1. 어간 추출 프로세스

2.1 형태분석 파일

말뭉치 데이터베이스는 원시 말뭉치 파일과 어간 추출에 필요한 형태분석 말뭉치 파일로 구성되어 있다. 형태분석 말뭉치는 특정한 문장에 대해서 공백을 기준으로 분리한 어절들에 대한 형태분석 정보를 나열하고 있으며, 그 정보에는 어절에 대한 어간정보를 포함하고 있다.

형태분석 말뭉치 파일은 그림 2와 같이 구성되어 있다. 중간의 긴 공백을 기준으로 좌측은 파일명과 행 번호를 나타내며 우측에 텍스트에 대한 정보가 있다. 우측의 목록을 보면 공백을 기준으로 좌측은 어절, 우측은 어절에 대한 형태가 분석이 되어 있으며 형태분석부분의 첫 번째 마디가 어절의 어간이다. 즉, 그림 2에 보이는 “여행을”이란 어절에서 형태분석의 첫 번째 마디인 “여행/NNG” 의 여행이 어간이 되고 NNG는 여행의 품사를 나타낸다. [5]

```
6CT_0013-0000050        <s n="00002">
6CT_0013-0000060        <vocal desc="웃음,후후후"/>
6CT_0013-0000070        </s>
6CT_0013-0000080        <s n="00003">
6CT_0013-0000090        이번    이번/NNG
6CT_0013-0000100        여름에,  여름/NNG+에/JKB+,/SP
6CT_0013-0000110        </s>
6CT_0013-0000120        <s n="00004">
6CT_0013-0000130        그      그/IC
6CT_0013-0000140        아프가니스탄으로 아프가니스탄/NNP+으로/JKB
6CT_0013-0000150        여행을   여행/NNG+을/JKO
6CT_0013-0000160        갔다    가/VV+ㅏ쓰/EP+다/EC
6CT_0013-0000170        왔거든요,          오/VV+ㅏ쓰/EP+거든요/EF+,/SP
```

그림 2. 형태분석 말뭉치 파일 구조

2.2 어간 데이터베이스 구축

형태분석 말뭉치 파일은 각각 약 500~30000개의 형태분석 정보를 담은 행을 가지고 있으며, 데이터베이스 구축을 위해서 200개의 말뭉치 파일을 사용하였다. 파일에서 가져올 정보는 어절과 품사를 제거한 어간이므로 필요한 정보만 가져오기 위한 전처리 과정을 거친다. 말뭉치 파일에는 중복된 형태분석 정보가 존재하기 때문에, 전처리 과정에서 중복된 형태분석을 모두 제거한다. 그 후, 어절이 들어갈 colum을 Primary Key로 지정하고 어절로부터 추출된 어간을 매칭 시킨다. 이와 같은 과정을 통해서 그림 3과 같은 데이터베이스가 구성이 되었으며, 중복된 형태분석 제거를 통하여 약 11만개의 row가 생성되었다.

word	stem	word	stem	word	stem
가	가	가겠다고	가	가격을	가격
가가	가	가겠다구	가	가격이	가격
가거거든요	가거거든요	가겠다는	가	가격이나	가격
가거나	가	가겠단	가	가격인데	가격
가거든	가	가겠대	가	가격표	가격표
가거든요	가	가겠대는데	가	가격표를	가격표
가거라	가	가겠습니다	가	가계	가계
가게	가	가겠어	가	가계부	가계부
가게가	가게	가겠어요	가	가계부는	가계부
가게끔	가	가겠지	가	가고	가
가게는	가게	가격	가격	가고는	가
가게를	가게	가격대	가격대	가고자	가
가게라면지	가게	가격대가	가격대	가공	가공
가게라든가	가게	가격대도	가격대	가공되거나	가공
가게로	가게	가격대를	가격대	가공되지	가공
가게를	가게	가격대에	가격대	가공된	가공
가게에	가게	가격대와	가격대	가공은	가공
가게에도	가게	가격대의	가격대	가공을	가공
가게에서	가게	가격도	가격	가공의	가공
가게요	가	가격두	가격	가공이	가공
가겠	가게	가격들	가격	가공하기	가공
가겠거니	가	가격만	가격	가공하는	가공
가겠구나	가	가격보다	가격	가공한	가공
가겠군	가	가격에	가격	가공해	가공
가겠네	가	가격으로	가격	가공해서	가공
가겠는데	가	가격으루	가격	가구	가
가겠다	가	가격은	가격	가구가	가구
가겠다고	가	가격을	가격	가구나	가

109031 row(s) returned

그림 3. 어간 데이터베이스

III. 스테밍 알고리즘 구현

3.1 텍스트 어간 추출

스테밍 알고리즘은 어간을 추출할 텍스트 파일의 전처리 과정과 위에서 설계한 어간 데이터베이스의 참조로 이뤄진다.

첫 번째로 텍스트 파일의 전처리 과정은 텍스트 내용들을 데이터베이스와 비교 가능한 어절로 바꾸기 위함이다. 파일의 내용을 공백을 기준으로 Split한 후 특수문자를 제거한다. 그 후 String형 자료구조에 분리된 어절들을 삽입한다.

어절이 담긴 자료구조가 준비되었으면 어간 데이터베이스와 비교를 통해 어간 변환 과정을 수행한다. 어절이 담긴 자료 구조를 탐색하면서 데이터베이스에 같은 어절이 존재한다면 그 어절의 어간형태를 얻어와 어간형태로 변환한다. 그림 4는 특정 텍스트 문서에 대한 어간 처리 결과를 보여준다.

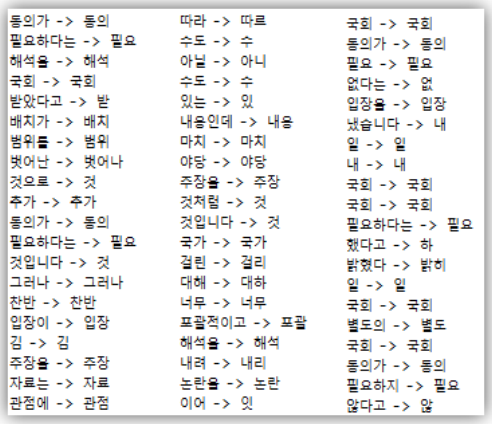

그림 4. 어절의 어간 변환 결과

3.2 어간 데이터베이스 비교

mysql로 작성된 데이터베이스와 달리 텍스트 파일로 작성된 데이터베이스를 구축해서 실행 속도 비교를 수행했다. 텍스트 파일 데이터베이스의 구축방법은 2.2와 동일하며 쿼리문으로 데이터베이스에 INSERT하는 대신 스트림을 통해 어절과 어간 한 행을 출력하는 방식으로 작성하였다. A4 용지 한 장 분량의 실행 속도는 둘 다 비슷한 속도를 보였으나, A4 용지 100장 분량의 문서로 테스트해본 결과, mysql은 약 7초 가량의 속도를 보인데 비해 텍스트 파일 데이터베이스는 0.7초의 약 10배 가량 빠른 속도를 보였다.

그림 5는 200개의 문서를 데이터베이스화 할 때의 INTSERT 확률을 보여주고 있다. 200개 문서를 20개씩 묶어서 문서 20개당 INSERT되는 횟수를 퍼센테이지로 구했는데, x축은 20개씩 묶은 문서의 개수이며, y축은 INSERT 확률이다. 본 연구에서는 20개의 문서까지는 약 25%의 INSERT 확률을 보였으나 문서 수가 180개가 넘어가면서 부터는 확률의 변화가 미미한 것을 확인 할 수 있었다. 이는 180개의 문서를 데이터베이스화 했을 시, 약 10만개의 단어가 생성되었고 이후에는 데이터베이스에 추가되는 단어가 미미해 성능의 변화가 크지 않다는 것을 보여주고 있다.

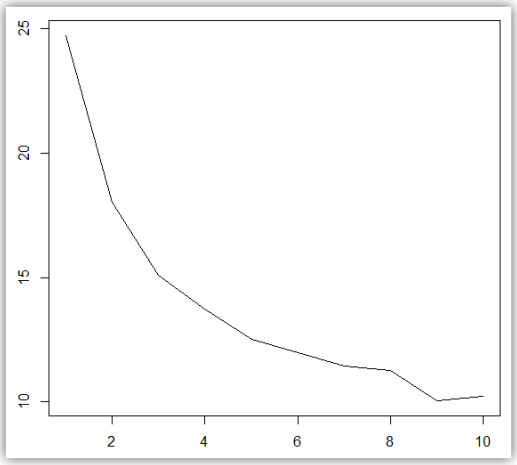

그림 5. 데이터베이스 INSERT 확률 그래프

IV. 결 론

본 논문에서는 자연어 처리 효과를 높이기 위한 한글 스테밍 알고리즘을 설계하고 구현하였다.

국립국어원에서 제공하는 공공데이터를 활용하여 어간 정보에 대한 신뢰성을 높였으며, 처리 속도 비교를 통해 좀 더 신속한 어간 데이터베이스를 구축할 수 있었다.

또한 증가하는 문서 수에 대한 어간삽입확률 그래프를 통해 성능에 관한 유의미한 결과를 도출해 낼 수 있었다.

문제점으로는 일반적인 단어들에 대해서는 어느 정도 적절한 추출 결과를 보였으나 고유명사나 일반적으로 자주 쓰이지 않는 단어에 대해서는 취약점을 보였으므로 차후 개선해 나가야 할 것으로 보인다.

감사의 글

이 논문은 2017년도 산업통상자원부의 '창의산업융합 특성화 인재 양성사업'의 지원을 받아 연구되었음.(과제번호 N0000717)

참고문헌

[1] 김동완 "빅데이터의 분야별 활용사례" 경영논총, 제34집, pp. 39-52, 2013.

[2] 최윤정, 박승수 "웹 컨텐츠의 분류를 위한 텍스트마이닝과 데이터마이닝의 통합 방법 연구" 인지과학 제13권 제3호, pp. 33-46, 2002

[3] 이효숙 “자연어검색시스템을 위한 스태밍알고리즘의 설계 및 구현” 정보관리학회지, 제14권, 제2호 213-234, 1997.

[4] 김근형, 오성열 “온라인 고객리뷰 분석을 통한 시장세분화에 텍스트마이닝 기술을 적용하기 위한 방법론” 한국콘텐츠학회논문지 제9권 제8호, pp. 272-284, 2009.

[5] 이종화, 레환수, 이현규 “소셜네트워크서비스에 활용할 비표준어 한글 처리 방법 연구” 한국산업정보학회논문지 제21권 제3호, pp. 35-46, 2016