특허 문서를 위한 형태소 분석기 비교 평가

Comparison and Evaluation of Morphological Analyzers for Patent Documents

저자 (Authors)
이유진, 김세빈, 홍현석, 김장원
Yujin Lee, Sebin Kim, Hyunseok Hong, Jangwon Gim

출처 (Source)
Proceedings of KIIT Conference , 2019.6, 264-265(2 pages)

발행처 (Publisher)
한국정보기술학회
Korean Institute of Information Technology

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08750007

APA Style
이유진, 김세빈, 홍현석, 김장원 (2019). 특허 문서를 위한 형태소 분석기 비교 평가. Proceedings of KIIT Conference, 264-265

특허 문서를 위한 형태소 분석기 비교 평가

이유진*, 김세빈*, 홍현석*, 김장원*†

Comparison and Evaluation of Morphological Analyzers for Patent Documents

Yujin Lee*, Sebin Kim*, Hyunseok Hong*, and Jangwon Gim*†

요 약

최근 비정형 텍스트에 대한 분석이 증가하면서 한국어에 대한 자연어 처리가 더욱 중요하게 되었다. 또한 분석대상에 포함된 문서의 형태(문장 길이, 전문 용어 등)에 따라 다양한 형태소 분석기가 사용 되고 있다. 특히 특허 문서는 일반 뉴스나 웹 상의 글과는 다르게 전문적인 용어가 다수 출현하고 있어 이러한 형태를 고려한 형태소 분석기가 필요하다. 따라서 본 연구에서는 기존에 개발된 형태소 분석기(OKT, MeCab, Komoran, Khaiii)에 대한 성능을 비교 평가하였다.

Abstract

As recent research on unstructured text analysis increases, natural language processing of Korean has become more important. In addition, various morphological analyzers are used depending on the targeted type(sentence length, terminology, etc.) of document in the analysis. Patent documents, in particular, has many technical terms compared to general documents like news and web articles. Therefore, selecting a proper morphological analyzer is necessary. In this study, we evaluated the performance of various morpheme analyzers such as OKT, MeCab, Komoran, and Khaiii.

Key words
Patent claim, MeCab, OKT, Komoran, Khaiii

Ⅰ. 서 론

최근 비정형 텍스트에 대한 분석이 증가하면서 자연어 처리 기술은 더욱 중요하게 되었다[1]. 대표적인 지식재산권인 특허에 포함된 청구항에는 전문적인 용어, 복합명사 및 신규어의 출현 빈도가 높다. 한글로 작성된 특허 문서에서의 기술 개체 식별을 위해서는 교착어의 특징을 갖는 한국어의 특징을 고려할 필요가 있다. 따라서 본 논문에서는 특허 문헌에 적용 가능한 형태소 분석기(OKT, MeCab, Komoran, Khaiii)의 성능을 비교 평가한다[2,3].

* 국립군산대학교
† 교신저자

※ 이 성과는 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2018R1C1B6008624).

II. 관련 연구

자연어 처리에 있어서 전처리 과정은 매우 중요하며 그 과정에서 유의미한 단어를 추출하기 위해 기본적으로 형태소 분석을 한다. 최근에는 전문용어를 분석하기 위해 형태소 분석기가 중요한 역할을 하고 있다[4]. 또한 딥 러닝 기술을 기반으로 형태소 분석기의 정확도를 높이는 연구 또한 진행되고 있다[5]. 그 결과 다양한 형태소 분석기들이 텍스트 분석에 활용되고 있으며 전문 용어를 포함할 경우 기존의 형태소 분석기의 한계가 있다. 따라서 본 연구는 전문용어 출현이 많은 특허 문서 분석을 위해 형태소 분석기에 대한 성능을 비교 평가한다.

III. 실 험

한국특허정보원(KIPRIS)에서 제공되는 특허 데이터(100,000건)로부터 청구항(1,120,000개) 및 고유 단어(1,580,000개)로 실험데이터를 구축하였다.

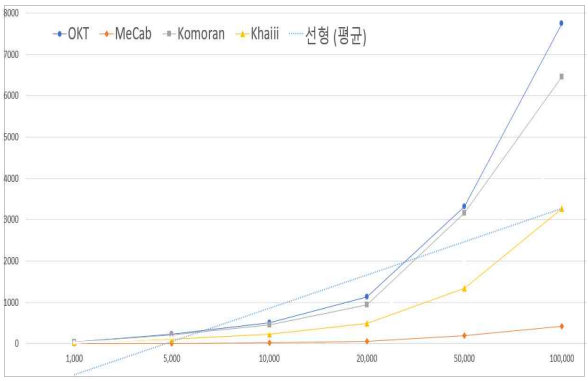

그림 1. 특허 건수에 따른 품사 태깅 시간
Fig. 1. POS tagging time per patent count

그림 1에서 x축은 특허 건수이고, y축은 품사 태깅 시간을 나타낸다. 50,000건 이하의 데이터에서는 태깅 시간이 선형으로 증가하지만, 이후 데이터에서는 그래프가 급격하게 변화하는 것을 볼 수 있다. 또한 청구항에서 동사 추출에 대한 비교를 수행한 결과인 그림 2는 모든 형태소 분석기에서 나온 단어를 색인화를 하여 빈도수를 나타낸 그래프이다.

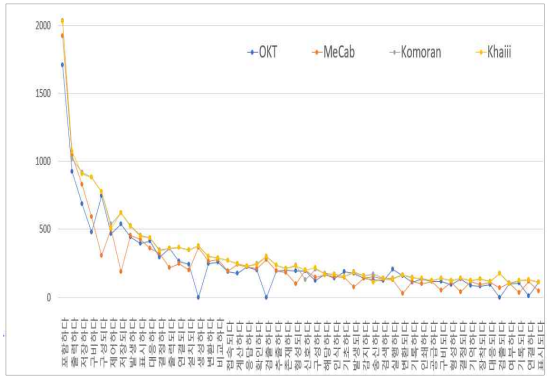

그림 2. 형태소 분석기별 색인어 추출 빈도
Fig. 2. Unique word frequency per morphological analyzer

IV. 결 론

본 논문에서는 전문용어들을 포함한 특허 문서로부터 유의미한 단어 추출에 사용하는 분석기들의 성능을 비교 평가 하였다. Khaiii와 MeCab은 데이터가 증가함에도 처리 시간이 크게 늘어나지 않았음을 확인할 수 있었으며 추출된 색인어를 비교한 결과 분석기들 간에 비슷한 결과를 도출함을 확인하였다. 향후에는 고전적인 방법의 형태소 분석기뿐만 아니라 딥 러닝을 이용한 형태소 분석을 진행할 예정이다.

참 고 문 헌

[1] 남기훈. "한국어 비정형 데이터 처리를 위한 효율적인 오피니언 마이닝 기법." 예술인문 사회융합 멀티미디어 논문지. Vol.7, pp.759-766, June 2017.

[2] Konlpy https://konlpy-ko.readthedocs.io May 2019.

[3] Khaiii https://github.com/kakao/khaiii May 2019.

[4] 이상백, 손윤희, 장현철, 이규철. "전문용어 정제를 위한 형태소 분석을 이용한 한의학 증상 진단 시스템 개발." 정보과학회 컴퓨팅의 실제 논문지 Vol.22, no2, pp.77-82, February 2016.

[5]이동준, 임유빈, 권태경. "형태소 기반 효율적인 한국어 단어 임베딩." 정보과학회논문지 Vol.45, no5, pp.444-450, May 2018.