Code Paper https://doi.org/10.23287/KJDH.2024.1.2.5

근대 국한문혼용체 자료 서브워드 기반 형태소 분석기의 설계와 적용

Design and Implementation of a Subword-based Morphological Analyzer for Modern Sino-Korean Mixed Texts

김병준(Kim, Byungjun)*
0000-0001-9925-9343

목차

초록

본 연구는 근대 국한문혼용체 자료의 자동화된 분석을 위한 서브워드 기반 형태소 분석기를 설계하고 적용하는 방법을 제안한다. 현재 구축된 대규모 근대 문헌 데이터베이스는 한자어와 옛한글이 혼재된 특성으로 인해 현존하는 형태소 분석기로는 효과적인 처리가 어렵다. 이러한 문제를 해결하기 위해 본 연구에서는 kiwipiepy 라이브러리의 sw_tokenizer 를 활용하여 서브워드 토큰화 기반의 새로운 접근법을 제시한다. 1890-1940 년대의 신문 및 잡지 자료 약 230 만 건(약 7 억 7 천만 음절)을 학습 데이터로 활용하여 세 가지 다른 vocab_size(32000, 48000, 64000)를 적용한 모델을 구현하고 그 성능을 비교하였다. 실험 결과, vocab_size 가 커질수록 복합 한자어의 의미 단위가 더 잘 보존되는 것을 확인하였으며, 연구 목적에 따라 적절한 분석 단위를 선택할 수 있음을 보였다. 본 연구는 근대 국한문혼용체 자료의 자동화된 분석을 위한 실용적인 도구를 제시함으로써 디지털 인문학 연구의 새로운 방향을 제시했다는 의의가 있다.

주제어: 근대 국한문혼용체, 서브워드 토큰화, 형태소 분석, 말뭉치 구축

*단독저자: 한국학중앙연구원 한국학대학원 인문정보학, 조교수, bjkim@byungjunkim.com

Abstract

This study proposes the design and implementation of a subword-based morphological analyzer for automated analysis of modern Sino-Korean mixed texts. Current large-scale modern literature databases are difficult to process effectively with existing morphological analyzers due to their characteristics of mixed Sino-Korean characters and archaic Korean. To address this issue, we present a new approach using the sw_tokenizer of the kiwipiepy library based on subword tokenization. We implemented three models with different vocab sizes (32000, 48000, 64000) using approximately 2.3 million newspaper and magazine articles (about 771.5 million syllables) from 1890-1940 as training data. The experimental results show that larger vocab sizes better preserve the semantic units of compound Sino-Korean characters, and researchers can select appropriate analysis units according to their research purposes. This study contributes to digital humanities research by providing a practical tool for automated analysis of modern Sino-Korean mixed texts and suggests new directions for future research in this field.
Keywords: modern Sino-Korean mixed text, subword tokenization, morphological analysis, corpus construction.

1. 서론

1.1 연구 배경 및 목적

근대 시기 한국의 지적 흐름을 이해하기 위한 핵심 자료인 신문과 잡지들이 국가적 차원의 디지털화 사업을 통해 전산화되면서, 연구자들의 자료 접근성이 크게 향상되었다. 국립중앙도서관의 <대한민국 신문 아카이브>와 한국사데이터베이스의 <한국근현대잡지자료>와 같은 데이터베이스 구축으로 연구자들은 방대한 양의 근대 신문 기사 텍스트를 손쉽게 접할 수 있게 되었다. 그러나 이렇게 축적된 디지털 자료를 대상으로 디지털인문학 연구방법론을 적용하여 분석하는 데에는 여전히 기술적 제약이 따른다[1].

현재 한국어 자연어 처리 분야에서는 BERT, GPT와 같은 사전학습 언어 모델(Pretrained Language Model, PLM)이 주류를 이루고 있으나, 이러한 모델들은 대부분 현대 한국어를 대상으로 학습되었기에 근대 국한문 혼용체 자료를 처리하는 데 한계가 있다. Park et al. (2020)[2]의 연구에서는 한국어 NLP 작업에서 형태소를 고려한 토큰화 방식이 더 효과적임을 실험적으로 입증했으며, 이는 형태소 분석이 PLM 시대에도 여전히 중요한 역할을 할 수 있음을 시사한다. 나아가 이용태 외(2019)[3]는 한국어 기계 독해를 위한 다양한 토큰화 방식을 비교 분석하여, 서브워드 기반의 접근이 효과적일 수 있음을 보였다.

이러한 배경에서 본 연구는 서브워드(Subword) 기반의 토큰화 방식을 활용하여 근대 국한문 혼용체 자료를 효과적으로 분석할 수 있는 방법을 제안하고자 한다. 서브워드 토큰화는 단어보다 작은 단위로 텍스트를 분할하여 처리하는 방식으로, 미등록어(Out-of-Vocabulary) 문제를 해결하고 어휘 사전의 크기를 효율적으로 관리할 수 있다는 장점이 있다. 이와 관련하여 어수경 외(2021)[4]는 한국어의 교착어적 특성을 고려한 음절 기반 토큰화 방식을 제안함으로써, 한국어의 언어적 특성에 맞는 서브워드 분절의 가능성을 제시한 바 있다.

1.2 근대 국한문혼용체 자료의 특성과 분석의 난점

근대 국한문혼용체 자료는 현존하는 어떤 형태소 분석기로도 적절한 처리가 어려운 독특한 언어적 특성을 가지고 있다. 이러한 난점은 크게 세 가지 측면에서 발생한다.

첫째, 한자어와 한글이 복잡하게 결합되어 있다는 점이다. 예를 들어 "今般 皇城新聞社에셔 論說을 草하야 我邦의 敎育事業이 日進月步라 니"라는 문장을 살펴보면, '에셔', '하야' 등의 옛한글 문법 요소가 한자어와 결합되어 있으며, '草하야'처럼 한자와 한글이 불규칙하게 혼용되어 있다. 또한 '論說', '敎育事業'과 같은 복합 한자어와 '日進月步'와 같은 사자성어까지 포함되어 있어, 현대 한국어 형태소 분석기로는 이러한 복잡한 결합 구조를 제대로 파악하기 어렵다. 현대 한국어로 이 문장을 옮기면 "이번에 황성신문사에서 논설을 작성하여 우리나라의 교육사업이 나날이 발전한다고 하니"가 되는데, 이처럼 한 문장 안에서도 다양한 층위의 언어적 특성이 혼재되어 있다.

둘째, 동일한 개념이 한자어와 한글로 혼용되어 표기되는 특징이 있다. 앞서 제시한 예문에서도 '論說'은 '논설'로, '敎育事業'은 '교육사업'으로도 표기될 수 있다. 이러한 표기의 유동성은 당시 신문들이 한글 보급과 독자층 확대를 위해 한자어를 점진적으로 한글로 바꾸어가는 과정에서 발생했다. 예를 들어 "學界의 소식을 新聞으로 傳하다"라는 문장에서는 '學界'와 '新聞'은 한자로, '소식'은 한글로 표기되어 있다. 이처럼 동일한 텍스트 내에서도 표기 방식이 일관되지 않아, 자동 분석 시 의미적으로 동일한 단위를 포착하기 어렵다는 문제가 있다.

셋째, 당시의 어문 규범이 현재와 달라 문법적 일관성이 떨어진다. 앞선 예문의 '에셔'(에서), '하야'(하여)와 같은 옛한글 표기는 현대 한국어 문법 체계와는 다른 방식으로 활용된다. 특히 한자어 용언의 활용에서 이러한 특징이 두드러지는데, '草하야'(초하여), '傳하다'(전하다)와 같이 한자어 어근에 한글 어미가 결합하는 방식이 현대 한국어의 규칙과는 다르다. 더구나 '是以로'(이로써), '爲야'(위하여)처럼 한문 허사가 한글 조사와 결합하는 경우도 있어, 기존의 형태소 분석 방식으로는 이러한 비정형적 결합을 효과적으로 처리하기 어렵다.

이러한 특성들로 인해 근대 국한문혼용체 자료는 현존하는 어떤 형태소 분석 도구로도 적절한 처리가 어려운 상황이다. 현대 한국어 형태소 분석기는 한자어와 옛한글을 제대로 인식하지 못하고, 한문이나 현대 중국어 형태소 분석기는 한글로 표기된 조사와 어미를 처리하지 못한다. 특히 정유경, 반재유(2019)[5]의 연구에 따르면, 현대 중국어 형태소 분석기의 경우 한자어가 문법적 기능어로 쓰이는 경우를 제대로 인식하지 못해 불필요한 색인어를 과도하게 추출하는 문제가 있다. 예를 들어 '是以로'와 같이 한문 허사가 조사와 결합된 경우, '是'와 '以'를 각각 독립된 의미 단위로 인식하여 색인어로 추출하는 오류를 범한다. 이처럼 규칙 기반의 접근법은 표기와 문법의 비일관성으로 인해 근본적인 한계를 보이며, 따라서 이러한 자료의 특수성을 고려한 새로운 분석 방법론이 필요한 상황이다.

1.3 국한문 혼용체 자료의 형태소 분석 시도

지금까지 국한문 혼용체 자료의 형태소 분석을 위해 다양한 시도가 있어왔다[6]. 초기에는 주로 사전과 규칙에 기반한 형태소 인식 모듈이 개발되었다. 이러한 접근법은 연구자가 수작업으로 선별한 규칙에 기반해 형태소를 분석하는데, 규칙에 일치하는 입력에 대해서는 높은 정확도를 달성할 수 있었다. 그러나 이러한 방식은 크게 두 가지 한계점을 가진다. 첫째, 데이터 구축에 많은 시간과 인력이 소요된다. 연구자가 넓은 범위의 규칙을 작성해야 하므로 작업량이 많고, 특히 근대 시기의 다양한 표기법과 어휘를 모두 포괄하기 위해서는 대규모 인력이 필요하다. 둘째, 모델이 새로운 형태소나 신조어 등 처음 접하는 단어(Out-of-Vocabulary)에 취약하다는 점이다.

이러한 기존 접근법의 한계를 극복하기 위해 본 연구에서는 서브워드 토큰화 방식을 활용하여

kiwipiepy 라이브러리의 sw_tokenizer를 구현하고자 한다. 특히 최근 이용태 외(2024)[7]의 연구에서 토큰화 방식이 대규모 언어모델의 성능에 미치는 영향이 실험적으로 입증된 만큼, 적절한 토큰화 방식의 선택이 중요하다. 본 연구의 목적은 크게 세 가지로 요약할 수 있다. 첫째, 근대 국한문 혼용체 자료의 특성을 고려한 서브워드 토큰화 방식을 설계하고 구현한다. 둘째, 하이퍼파라미터를 다르게 적용한 복수의 토크나이저를 동일한 텍스트에 적용해 그 차이를 평가하고 개선점을 도출한다. 셋째, 실제 연구 현장에서 활용할 수 있는 구체적인 가이드라인을 제시한다. 이를 통해 근대 한국학 연구에서 디지털 방법론의 적용 가능성을 확장하고, 대규모 텍스트 자료의 계량적 분석을 위한 기반을 마련하고자 한다.

2. 서브워드 기반 형태소 분석기 설계 [8]

2.1. 서브워드 토크나이저의 이론적 배경

앞서 살펴본 근대 국한문혼용체 자료의 분석 난점을 해결하기 위해, 본 연구에서는 서브워드 토큰화 방식에 주목한다. 서브워드 토크나이저는 기존의 단어나 형태소 단위 토큰화가 가진 한계, 특히 미등록어 문제와 어휘 사전 크기 문제를 해결하기 위해 등장했다[9][10]. 현재 널리 활용되는 서브워드 토큰화 알고리즘은 크게 네 가지로 구분할 수 있다. BPE(Byte Pair Encoding)는 빈도수 기반으로 문자열 쌍을 병합하는 방식으로, 구현이 단순하면서도 효과적이다. WordPiece는 BPE의 변형으로, 병합 기준을 빈도수 대신 우도(likelihood)로 삼아 더 의미있는 서브워드 단위를 포착하고자 한다. Unigram Language Model은 확률 모델을 기반으로 하여 문맥을 고려한 다양한 분절 가능성을 제공한다. Byte-Level BPE는 문자 대신 바이트 단위로 접근하여 다국어 처리에 강점을 보인다.

이러한 알고리즘들은 각각의 장단점을 가지고 있는데, 특히 한자어와 옛한글이 혼재된 근대 국한문혼용체의 경우 기존 알고리즘의 직접적인 적용은 한계가 있다. 이에 본 연구에서는 kiwi 형태소 분석기의 새로운 버전(0.15.1)에서 도입된 서브워드 토크나이저를 활용하여 이러한 문제를 해결하고자 한다[11].

2.2. kiwipiepy 의 sw_tokenizer 구조와 특징

kiwipiepy의 sw_tokenizer는 기존 서브워드 토크나이저와 차별화되는 독특한 구조를 가진다. 특히 주목할 점은 sw_tokenizer가 kiwipiepy의 tokenize 함수처럼 바로 불러와 쓸 수 있는 것이 아니라, 사용자가 직접 학습할 말뭉치를 가져와 자신만의 서브워드 토크나이저를 만들거나 다른 사람이 만든 토크나이저를 따로 불러와 사용해야 한다는 점이다. 가장 큰 특징은 한국어의 형태소 정보를 토큰화 과정에 통합했다는 점이다. 유니그램 언어모델을 기반으로 하되, 형태소 분석 결과를 함께 고려함으로써 언어학적으로 더 타당한 분절이 가능하도록 설계되었다. 이러한 구현의 세부 사항과 사용 방법은 2024년 11월 현재 최신 버전인 0.20.1의 공식 문서[12]에서 확인할 수 있다.

특히 한자어 처리에 있어 주목할 만한 특징을 가지고 있다. SwTokenizerConfig의

split_chinese 옵션을 False로 설정함으로써, 한자어가 의미 단위로 보존될 수 있도록 한다. 이는 앞서 살펴본 중국어 형태소 분석기의 한계를 극복할 수 있는 중요한 설계적 특징이다. 또한 vocab_size를 통해 토크나이저의 어휘 사전 크기를 조절할 수 있어, 자료의 특성에 따라 유연한 조정이 가능하다. 예를 들어 '新聞(신문)'이라는 2음절 한자어의 경우, '新'과 '聞'이 자주 같이 등장할 때 하나의 서브워드로 등재될 수 있는데, vocab_size 설정에 따라 이러한 등재 여부가 달라질 수 있다.

2.3. 학습 데이터 구성과 전처리

제안하는 토크나이저의 성능을 검증하기 위해, 1900-1940년대의 대규모 근대 문헌 자료를 학습 데이터로 구성했다. 구체적으로 세 가지 주요 자료를 활용했다.

먼저, 한국사데이터베이스의 한국근현대잡지자료 83건 중 1940년까지 게재된 잡지 65종을 선별했다[13]. 총 557,481개의 단락이 포함되었으며, 연대별 분포는 1890년대 1,664건(173,903음절), 1900년대 47,261건(5,803,690음절), 1910년대 8,198건(323,459음절), 1920년대 171,235건(18,194,659음절), 1930년대 302,766건(22,495,046음절), 1940년대 26,357건(1,720,566음절)이다. 이는 약 4,870만 음절에 달하는 규모로, 특히 1920-30년대의 자료가 풍부하게 포함되어 있다는 특징이 있다.

둘째, 조선일보 뉴스 라이브러리에서 1920년부터 일제 강제 폐간 전인 1940년까지의 기사를 수집했다. 정규 기사뿐만 아니라 각종 광고 및 연재 소설까지 모두 포함하여 총 900,476건의 기사를 확보했다. 시대별로는 1920년대 340,320건(115,917,400음절), 1930년대 525,253건(219,571,285음절), 1940년대 34,903건(14,498,992음절)의 분포를 보인다. 약 3억 5천만 음절에 달하는 대규모 데이터로, 특히 1930년대 자료가 전체의 63%를 차지하여 이 시기의 언어 사용을 풍부하게 반영하고 있다.

셋째, 동아일보 디지털아카이브에서 역시 1920년부터 1940년까지의 기사를 수집했다. 총 922,472건의 기사가 포함되었으며, 시대별로는 1920년대 372,290건(132,352,172음절), 1930년대 511,036건(223,379,850음절), 1940년대 39,146건(17,528,979음절)이다. 총 3억 7천만 음절 규모로, 조선일보와 마찬가지로 1930년대 자료가 전체의 60%를 차지한다.

이렇게 세 가지 자료를 합산하면 총 7억 7천만 음절에 달하는 대규모 말뭉치가 구성되며, 특히 1920-30년대 자료가 전체의 80% 이상을 차지하여 일제강점기 국한문혼용체의 핵심 시기를 충실히 포괄하고 있다.

수집된 자료의 품질 확보를 위해 체계적인 정제 과정을 거쳤다. 먼저 자료의 특성을 고려하여 잡지는 문단 단위로, 신문은 기사 단위로 추출하여 개별 리스트를 구성했다. 이후 세 가지 자료를 하나의 문자열 리스트로 통합하는 과정에서 중복 텍스트를 제거하여 데이터의 중복성을 최소화했다. 이러한 정제 과정을 거쳐 최종적으로 총 2,287,985건(771,519,502 음절)의 문단/기사를 학습 데이터로 확보했다. 이는 기존 연구들에서 사용된 학습 데이터의 규모를 크게 상회하는 것으로, 보다 신뢰성 있는 토크나이저 학습이 가능할 것으로 기대된다.

3. 형태소 분석기 구현과 성능 평가

3.1. 실험 설계와 구현

앞서 2장에서 논의한 kiwipiepy의 sw_tokenizer의 구조적 특징을 바탕으로, 본 연구에서는 서로 다른 vocab_size를 적용한 세 가지 모델을 구현하여 그 성능을 비교 평가하였다. 구현 과정에서는 1900-1940년대의 근대 문헌 자료로 구축한 약 230만건의 학습 데이터를 활용하였다.

토크나이저의 구현은 kiwipiepy의 SwTokenizer.train 함수를 통해 이루어졌다. 설정에 있어 가장 중요한 두 가지 요소는 SwTokenizerConfig와 vocab_size이다. SwTokenizerConfig에서는 앞서 설명했듯이 한자어의 무분별한 분할을 방지하기 위해 split_chinese 옵션만을 False로 설정하고 나머지는 기본값을 유지하였다. vocab_size는 32000, 48000, 64000 세 가지로 설정하여 각각 다른 크기의 토크나이저를 생성하였다. 구현을 위한 기본 코드는 다음과 같다:

```
config = SwTokenizerConfig(split_chinese=False)
tokenizer = SwTokenizer.train(
  save_path='./241112_vo32000_tokenizer.json',
  texts=paragraph_list,
  vocab_size=32000,
  config=config)
```

3.2. 모델 성능 평가

구현된 세 가지 모델의 성능 차이를 분석하기 위해 실제 조선일보 기사에서 발췌한 예시 문장 "國民新報時代에政合邦說을提唱하야日韓聯邦하기로主論하기도經하얏고"를 대상으로 토큰화를 수행하였다. 각 모델의 토큰화 결과는 다음과 같다:

- 모델 A (Vocab 32000): ['國民', '新', '報', '時代', '에/J', '政', '合', '邦', '說', '을/J', '提唱', '하/V', '야/E', '日', '韓', '聯', '邦', '하/V', '기/E', '로/J', '主', '論', '하/V', '기/E', '도/J', '經', '하얏고']
- 모델 B (Vocab 48000): ['國民', '新', '報', '時代', '에/J', '政', '合', '邦', '說', '을/J', '提唱', '하/V', '야/E', '日', '韓', '聯邦', '하/V', '기/E', '로/J', '主', '論', '하/V', '기/E', '도/J', '經', '하얏고']
- 모델 C (Vocab 64000): ['國民', '新報', '時代', '에/J', '政', '合', '邦', '說', '을/J', '提唱', '하/V', '야/E', '日', '韓', '聯邦', '하/V', '기/E', '로/J', '主', '論', '하/V', '기/E', '도/J', '經', '하얏고']

이 결과를 분석해보면 매우 흥미로운 차이점들이 발견된다. 모델 A와 모델 B의 가장 두드러진

차이는 '聯邦' 한자어의 처리 방식이다. 모델 A는 이를 '聯'과 '邦' 두 개의 개별 토큰으로 분리한 반면, 모델 B는 '聯邦'이라는 하나의 토큰으로 처리하였다. 이는 vocab_size가 커짐에 따라 더 많은 복합 한자어를 하나의 단위로 처리할 수 있게 되었음을 보여준다. 모델 B와 모델 C를 비교해 보면, '新報'의 처리에서 차이가 나타난다. 모델 B가 '新'과 '報' 두 개의 토큰으로 분리한 것에 비해, 가장 큰 어휘 사전을 가진 모델 C는 이를 '新報' 하나의 토큰으로 처리하였다. 이는 vocab_size가 더욱 커짐에 따라 더 많은 의미 단위를 보존할 수 있게 되었음을 시사한다.

이러한 결과는 연구자들이 자신의 연구 목적에 맞는 토크나이저를 선택해야 할 필요성을 보여준다. 예를 들어, 개별 한자의 의미 분석이 중요한 연구라면 모델 A와 같은 작은 vocab_size의 토크나이저가 적합할 수 있다. 반면, 복합 한자어의 의미 단위를 보존하는 것이 중요한 연구라면 모델 C와 같은 큰 vocab_size의 토크나이저가 더 적합할 것이다.

본 연구에서는 서로 다른 vocab_size에 따른 기본적인 토큰화 성능을 비교하는 데 초점을 맞추었으나, 향후 더욱 체계적인 성능 평가를 위해서는 정량적/정성적 평가 방법의 개발과 한자-한글 혼용 비율에 따른 세부적인 성능 비교가 필요하다. 이러한 추가 연구와 실험 결과는 본 연구의 깃허브 레포지토리를 통해 지속적으로 보완하여 공개할 예정이다.

4. 의의와 전망

4.1 근대 자료 분석의 새로운 가능성과 한계

그동안 한국 근대 텍스트 자료 분석에서는 국한문혼용체 전용 형태소 분석기의 부재로 인해 여러 가지 제약이 있었다. 연구자들은 어쩔 수 없이 현대어 번역본을 활용하거나, 한자어를 강제로 1음절 단위로 분절하거나, 때로는 고전/현대 중국어 형태소 분석기를 사용해야 했다. 이러한 방법들은 모두 원문의 의미를 온전히 보존하지 못하거나, 부적절한 분석 결과를 초래할 위험이 있었다.

본 연구에서 제안한 서브워드 기반 형태소 분석기는 이러한 한계를 극복하고 원문 자료를 직접 토큰화할 수 있게 했다는 점에서 큰 의의가 있다. 특히 3장의 실험에서 확인했듯이, vocab_size 조정을 통해 한자어의 의미 단위를 연구 목적에 맞게 보존할 수 있다는 점은 텍스트 마이닝 결과의 품질을 크게 제고할 수 있는 가능성을 보여준다.

그러나 현재 구현에는 몇 가지 중요한 한계점이 존재한다. 첫째, 정량적/정성적 성능 평가를 위한 표준화된 방법론이 부재하다는 점이다. 특히 근대 국한문혼용체의 특성상 현대 한국어 형태소 분석기의 성능 평가 방식을 그대로 적용하기 어렵다. 둘째, 한자와 한글의 혼용 비율에 따른 세부적인 성능 차이를 아직 체계적으로 검증하지 못했다. 이는 시대별, 매체별로 상이한 국한문혼용체의 특성을 고려할 때 반드시 보완되어야 할 부분이다. 또한 학습이 완료된 토크나이저에 사용자가 새로운 서브워드를 추가할 수 없다는 기술적 제약도 존재한다. 이는 특히 특정 분야나 시기의 고유한 용어를 추가로 처리해야 하는 전문 연구자들에게 제약이 될 수 있다.

이러한 한계점들은 본 연구의 깃허브 레포지토리를 통해 지속적으로 개선해 나갈 예정이며, 이를 통해 연구 결과의 재현성과 확장성을 보장하고자 한다. 특히 다양한 시기와 매체의 국한문혼용체 자료에 대한 벤치마크 데이터셋을 구축하고, 이를 통해 보다 객관적인 성능 평가가 가능하도록 할 계획이다. 또한 토크나이저의 확장성을 높이기 위한 기술적 개선도 지속적으로 진행할

예정이다.

4.2 향후 연구 방향과 전망

본 연구의 성과를 바탕으로, 향후 연구는 크게 두 가지 방향으로 발전될 수 있다. 첫째, 옛한글 서브워드 토크나이저로의 확장이다. 현재 개발된 방법론은 근대 국한문혼용체에 초점을 맞추고 있지만, 이를 중세 국어와 같은 옛한글 텍스트에도 적용할 수 있는 가능성이 있다. 앞서 살펴본 한자어 처리 방식과 유사하게, 옛한글의 특성을 고려한 토큰화 방식을 개발할 수 있을 것이다. 둘째, 다양한 시대를 아우르는 한국어 서브워드 토크나이저의 개발이다. 현재는 특정 시기의 텍스트만을 대상으로 하고 있지만, 향후에는 고대부터 현대에 이르는 다양한 시기의 한국어 텍스트를 모두 처리할 수 있는 통합적인 토크나이저의 개발이 필요하다. 이는 한국어의 통시적 연구를 위한 강력한 도구가 될 수 있을 것이다. 이러한 발전 방향은 단순히 기술적 개선을 넘어, 디지털 인문학 연구의 새로운 지평을 열 수 있을 것으로 기대된다. 특히 한국학 연구에서 시대와 언어의 경계를 넘어서는 통합적 연구를 가능하게 함으로써, 보다 풍부하고 심도 있는 학문적 성과를 이끌어낼 수 있을 것이다.

---

[1] 김선영 (2023). "교양교육에서의 근대 국한문 혼용 신문자료 활용을 위한 텍스트 처리(text processing) 자동화 방식의 고안과 추가적인 과제". <교양교육연구>. 41-52.

[2] Park, Kyubyong; Lee, Joohong; Jang, Sangkeun; Jung, Dongwoo (2020). "An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks." Proceedings of the AACL-IJCNLP. 133-142. https://aclanthology.org/2020.aacl-main.17

[3] 이용태, 김수민, 이홍노 (2024). "한국어 학습 시 토큰화 방식에 따른 대규모언어모델의 언어 이해 수준에 관한 연구". <한국통신학회 학술대회논문집>. 1442-1443. http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11737673

[4] 어수경, 박찬준, 문현석, 임희석 (2021). "한국어 인공신경망 기계번역의 서브 워드 분절 연구 및 음절 기반 종성 분리 토큰화 제안". <한국융합학회논문지>. 1-7.

[5] 정유경, 반재유 (2019). "국한문 혼용 텍스트 색인어 추출기법 연구". <정보관리학회지>. 7-19.

[6] 정유경 (2021). "근대 한국학 자료의 국한문 혼용 텍스트 처리와 적용". <한국정보관리학회 학술대회 논문집>. 51-64.

[7] 이용태, 김수민, 이홍노 (2024). "한국어 학습 시 토큰화 방식에 따른 대규모언어모델의 언어 이해 수준에 관한 연구". <한국통신학회 학술대회논문집>. 1442-1443. http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11737673

[8] 이 논문에서 개발한 국한문혼용체 자료 서브워드 기반 형태소 분석기는 아래 깃허브 링크에서 학습 과정과 테스트 결과를 모두 코드로 확인할 수 있다.
https://github.com/ByungjunKim/ModernKoreanSubword

[9] Sennrich, Rico; Haddow, Barry; Birch, Alexandra (2016). "Neural Machine Translation of Rare Words with Subword Units." Proceedings of the ACL. 1715-1725. https://doi.org/10.18653/v1/P16-1162

[10] 송현제 (2021). "서브워드 토큰화와 한국어 형태소 분석기". <정보과학회지> 39-4. 15-20.

[11] 이민철 (2024). "Kiwi: 통계적 언어 모델과 Skip-Bigram 을 이용한 한국어 형태소 분석기 구현". <디지털인문학> 1-1. 109-136. https://doi.org/10.23287/KJDH.2024.1.1.6

[12] https://bab2min.github.io/kiwipiepy/v0.20.1/kr/sw_tokenizer.html

[13] 김바로 (2024). "국사편찬위원회 한국근현대잡지자료 데이터(2024.03.27.)". <디지털인문학> 1-1. 143-156. https://doi.org/10.23287/KJDH.2024.1.1.8

참고문헌

강이경, 이해준, 김재원, 윤희원, 류원호 (2019). "한국어 기계 독해를 위한 언어 모델의 효과적 토큰화 방법 탐구". <한국어 언어처리 학술대회 논문집>. 197-202.

김바로 (2024). "국사편찬위원회 한국근현대잡지자료 데이터(2024.03.27.)". <디지털인문학> 1-1. 143-156. https://doi.org/10.23287/KJDH.2024.1.1.8

김선영 (2023). "교양교육에서의 근대 국한문 혼용 신문자료 활용을 위한 텍스트 처리(text processing) 자동화 방식의 고안과 추가적인 과제". <교양교육연구>. 41-52.

송현제 (2021). "서브워드 토큰화와 한국어 형태소 분석기". <정보과학회지> 39-4. 15-20.

어수경, 박찬준, 문현석, 임희석 (2021). "한국어 인공신경망 기계번역의 서브 워드 분절 연구 및 음절 기반 종성 분리 토큰화 제안". <한국융합학회논문지>. 1-7.

이민철 (2024). "Kiwi: 통계적 언어 모델과 Skip-Bigram 을 이용한 한국어 형태소 분석기 구현". <디지털인문학> 1-1. 109-136. https://doi.org/10.23287/KJDH.2024.1.1.6

이용태, 김수민, 이흥노 (2024). "한국어 학습 시 토큰화 방식에 따른 대규모언어모델의 언어 이해 수준에 관한 연구". <한국통신학회 학술대회논문집>. 1442-1443. http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11737673

정유경 (2021). "근대 한국학 자료의 국한문 혼용 텍스트 처리와 적용". <한국정보관리학회 학술대회 논문집>. 51-64.

정유경, 반재유 (2019). "국한문 혼용 텍스트 색인어 추출기법 연구". <정보관리학회지>. 7-19.

Park, Kyubyong; Lee, Joohong; Jang, Sangkeun; Jung, Dongwoo (2020). "An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks." *Proceedings of the AACL-IJCNLP*. 133-142. https://aclanthology.org/2020.aacl-main.17

Sennrich, Rico; Haddow, Barry; Birch, Alexandra (2016). "Neural Machine Translation of Rare Words with Subword Units." *Proceedings of the ACL*. 1715-1725. https://doi.org/10.18653/v1/P16-1162