

현대중국어 형태소 분석기의 현황과 활용*

박민준**

<목 차>

1. 서론
2. 각종 분석기의 설치 및 사용 방법
3. 각종 분석기의 특징 및 용도 분석
4. 결론

1. 서론

전통적인 언어학 연구의 출발점은 언어를 그것의 구성 요소로 환원하는 것에서 출발한다. 가령 구조주의 언어학의 연구 방법론은 언어를 구성하는 요소를 음소 혹은 형태소 등의 최소단위로 분절하고 그들 간의 변별적 자질과 상호 제약을 연구하는 것이다. 유사한 관점에서, 형태소 분석¹⁾(词法分析, morphological analysis)은 통사 분석(句法分析, syntactic parsing)과 의미 분석(语义分析, semantic analysis)으로 나아가기 위한 가장 기초적인 작업이므로, 형태소 분석이 올바르게 선행되어야 통사 및 의미 분석이 올바르게 수행될 수 있다.

코퍼스 언어학 또는 자연어처리(Natural Language Processing, NLP)의 분석 과정도 이와 크게 다르지 않다. 통사·의미 분석에 앞서, 먼저 연구 대상인 전체 텍스트를 그 구성 요소, 즉 특정한 문법적 단위로 분절하는 일종의 정규화 작업(text normalization)²⁾이 요구되는데, 이것이 NLP의 여러 가지 하위 작업

* 본 연구는 2021년도 덕성여자대학교 교내연구비 지원으로 이루어졌음.

** 덕성여자대학교 중어중문학과전공 조교수

1) 중국어는 형태변화가 상대적으로 무표적이므로 형태소 분석(morphological analysis)이라는 표현 대신 어휘 분석(lexical analysis)으로 칭하기도 한다.

2) 데이터 과학(data science)의 관점에서는 불필요한 데이터를 제거(혹은 정제 data

(task) 중 하나인 형태소 분석 단계에 해당한다.

영어와 한국어의 경우, 띄어쓰기 규칙이 존재하므로 공백 문자(space)를 기준으로 삼아 텍스트를 그 하위 구성 요소(즉 어휘 혹은 어절) 단위로 분절할 수 있는데, 이를 토큰화(tokenization)³⁾라 한다. 이렇게 분절된 요소는 다양한 어형으로 나타나기 마련인데⁴⁾, 이 때문에 영어와 한국어의 형태소 분석은 다양한 어형을 하나의 표제어로 통합하는 표제어 추출(lemmatization)⁵⁾ 작업이 요구되기도 한다.

개별 토큰(영어/한국어)	일반화된 토큰(원형/어근)
eats / 먹는다	eat / 먹-
ate / 먹었다	
eating / 먹는	
eaten / 먹힌	

[표 1] 표제어 추출의 예

이와 달리, 중국어 텍스트는 띄어쓰기가 없고 고립어적 특성으로 인해 별다른 형태 변화가 없기 때문에 영어나 한국어와 같은 표제어 추출은 생각된다. 따라서 중국어의 형태소 분석은 사실상 주어진 텍스트(문자열)를 어떻게 분절할 것인지, 즉 단어 분리(word segmentation)의 문제로 귀결된다.⁶⁾

cleansing)하고 각각각색의 개별 데이터를 소수의 유형으로 통합하는 정규화(normalization) 작업으로 볼 수 있다.

3) Tokenization is the task of cutting a string into identifiable linguistic units that constitute a piece of language data. (Bird, et al 2009)

4) 이는 굴절(inflexion)과 활용(conjugation) 등 형태론적 어형의 변화가 풍부한 영어와 한국어의 특징에 기인한다.

5) Lemmatization is a process that maps the various forms of a word (such as *appeared*, *appears*) to the canonical or citation form of the word, also known as the lexeme or lemma (e.g. APPEAR). (Bird, et al 2009).

6) 이 때문에 중국어의 단어 분리를 중국어 토큰화(Chinese tokenization)라고 칭하기도 한다.

원문	师兄，一起吃饭吗？ 哦，我吃过了。
단어 분리 (分词)	师兄， 一起 吃 饭 吗 ？ 哦， 我 吃 过 了 。
품사 태깅 (词性标注)	师兄/n， /w 一起/d 吃/v 饭/n 吗/y ？/w 哦/e， /w 我/r 吃/v 过/u 了/y 。 /w

[표 2] 중국어 형태소 분석의 예

대규모 텍스트의 경우, 단어 분리 작업을 사람이 일일이 수행하는 것은 비효율적이다. 이때, 단어 분리기(word segmenter), 혹은 토큰라이저(tokenizer)라 불리는 컴퓨터 프로그램을 사용하면 효율적으로 단어 분리를 수행할 수 있다. 이에 더해 품사 태깅(part-of-speech tagger) 프로그램을 사용하면, 각 단어에 대한 품사 정보도 자동으로 부착할 수 있다(표 2).

일반적으로 중국어 형태소 분석 프로그램은 단어 분리(word segmentation) 뿐만 아니라 분리된 단어에 대한 품사 분류(POS tagging) 기능도 함께 수행한다. 요컨대, ‘중국어 형태소 분석기⁷⁾(分词及词性标注器) = 단어 분리기(分词器) + 품사 태깅(词性标注器)’로 이해할 수 있다. 주요 중국어 형태소 분석기의 현황은 다음 표 3과 같다.

7) 중국어의 형태소 분석은 사실상 단어 분리에 가까우므로 어휘 분석기라 불러도 될 것이다. 다만, 국내 국어정보학 분야에서 이에 상응하는 한국어 분석 프로그램을 ‘형태소 분석기’라 칭하고 있기에 본고 역시 일종의 보통명사로써 ‘중국어 형태소 분석기’라 지칭한다. 실제로 중국어 분석 결과도 형태소(语素, g), 명사성 형태소(ng), 동사성 형태소(vg), 형용사성 형태소(ag) 등을 출력하기 때문에 모순되지 않는다.

명칭	기본 언어 (추가 지원 언어)	GUI 지원여부	지원 품사 tagset
ICTCLAS ⁸⁾	C++ (+Python, Java, Lucene)	지원, Online Demo ⁹⁾	PKU, ICTPOS
LTP ¹⁰⁾	Python (+Java, Ruby)	Online Demo ¹¹⁾	863
Jieba ¹²⁾	Python (+R, .NET, iOS, Android, PHP, Node.js, Go, Erlang, Rust)	미지원	자체 Tagset ¹³⁾ (PKU 기반)
Stanford Word Segmenter ¹⁴⁾ Tagger ¹⁵⁾	Java (+Python, Ruby, PHP, Node.js, Matlab, .NET, GATE)	Online Demo ¹⁶⁾ 미지원(Segmenter) 지원(Tagger)	CTB ¹⁷⁾
FudanNLP ¹⁸⁾	Java	미지원	자체 Tagset (CTB 기반)
THULAC ¹⁹⁾	C++ (+Python, Java)	Online Demo ²⁰⁾	PKU
PKUSEG ²¹⁾	Python	미지원	PKU
HanLP ²²⁾	Python (+Java, Golang)	API 사용 (postman)	PKU, CTB, 863
CorpusWord Parser	Windows MFC	지원 ²³⁾ , Online Demo ²⁴⁾	863

[표 3] 주요 중국어 형태소 분석기 현황

8) <https://github.com/NLPIR-team/NLPIR>9) <http://kgb.lingjoin.com/nlpir/>10) <https://github.com/HIT-SCIR/ltp>11) <http://ltp.ai/demo.html>12) <https://github.com/fxsjy/jieba>13) <https://gist.github.com/hscspring/c985355e0814f01437eaf8fd55fd7998>14) <https://nlp.stanford.edu/software/segmenter.shtml> (단어 분리기)15) <https://nlp.stanford.edu/software/tagger.shtml> (품사 태거)16) <https://corenlp.run/>17) <https://catalog.ldc.upenn.edu/docs/LDC2009T24/treebank/chinese-treebank-postags.pdf>18) <https://code.google.com/archive/p/fudannlp/downloads>19) <https://nlp.csai.tsinghua.edu.cn/project/thulac/>20) thulac.thunlp.org/demo21) <https://github.com/lancopku/pkuseg-python>22) <https://hanlp.hankcs.com/>23) <http://corpus.zhonghuayuwen.org/Resources.aspx>24) <http://corpus.zhonghuayuwen.org/CpsWParser.aspx>

중국어 형태소 분석, 즉 단어 분리와 품사 태깅은 텍스트 분석(text analytics), 코퍼스 언어학(corpus linguistics) 연구에 수반되는 양적 분석을 위한 토대가 되는 가장 기초적이고도 중요한 작업이다. 양적 분석의 목표가 중국어 텍스트라는 데이터에서 유용한 정보나 통찰을 얻어내는 것이라고 할 때, 중국어 단어 분리와 품사 태깅 작업이 미흡하다면 부실한 데이터 전처리(data preprocessing) 결과를 토대로 도출된 양적 분석 결과는 그 신뢰도를 담보하지 못하기 때문이다. 입력 데이터가 잘못되면 엉뚱한 분석 결과를 낼 수 있기에 (Garbage in, garbage out), 중국어 형태소 분석기의 정확한 이해와 활용이 매우 중요하다. 이에 본고는 중국어 형태소 분석 프로그램의 사용 방법을 상세히 설명하고 (2장), 프로그램별 특징과 최적 용도를 분석 (3장)함으로써 말뭉치 기반의 양적 분석을 위한 실용적 지침서의 역할을 하고자 한다.

2. 각종 분석기의 설치 및 사용 방법

본 장에서는 사용자의 관점에서 중국어 형태소 분석기 (단어 분리 및 품사 태깅 프로그램)을 GUI 기반 분석기와 CLI 기반 분석기 두 부류로 나누고, 개별 프로그램의 이용 방법과 특징을 살펴본다. 전자는 코딩에 익숙하지 않은 초심자 및 일반 사용자에게 적합하며, 후자는 중국어 단어 분리 및 품사 태깅의 세부 사항을 조작할 필요성이 있는 고급 사용자에게 적합하다.

2.1 GUI 기반 분석기

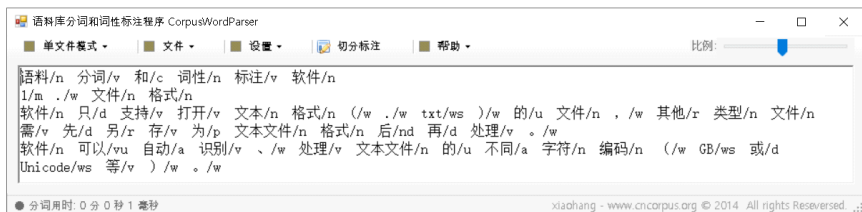
GUI (graphical user interface)는 마우스 클릭만으로 프로그램을 구동, 조작할 수 있는 그래픽 기반의 인터페이스를 일컫는다. 일례로, 윈도우즈나 카카오

특은 GUI 기반 운영체제와 애플리케이션이라 할 수 있다. GUI는 명령어를 입력할 필요 없이 주어진 메뉴를 선택함으로써 프로그램을 조작하기 때문에 사용이 편리하고 직관적이다.

2.1.1 CorpusWordParser

CorpusWordParser는 중국 国家语委(语料库在线)에서 개발한 중국어 형태소 분석기이다. 대표적인 GUI 기반 중국어 어휘 분석 프로그램인 CorpusWordParser를 사용한 중국어 어휘 분석 과정은 다음과 같다.

[그림 1] CorpusWordParser







- (1) 분석 대상 텍스트(코퍼스) 입력
 - [文件-打开文件] 에서 분석 대상 코퍼스(.txt)를 불러오기 혹은 텍스트 창에 직접 입력
- (2) 단어 분리 및 품사 태깅
 - [切分标注] 클릭
- (3) 작업 완료된 텍스트(출력)를 파일로 저장
 - 메뉴 [文件-保存文件] 클릭 후 텍스트 파일(.txt)로 저장

2.1.2 기타 시범용 웹 페이지 (Online Demo)

사용자 친화적인 그래픽 인터페이스(GUI)에서는 코딩을 전혀 모르는 사용자도 별다른 어려움 없이 중국어 형태소 분석이 가능하다. 하지만 위 [표 3]에서 알 수 있듯이, 주요 중국어 어휘 분석 프로그램 대다수는 GUI를 지원하지 않고 있다. 비록 LTP와 THULAC 등 몇몇 프로그램은 시범용 웹 페이지(Online

demo) 형식으로나마 GUI 서비스를 제공하고 있지만(아래 그림 2), 이들 웹 페이지는 입력할 수 있는 텍스트 용량에 상한(통상 1만 자 내외)이 존재하므로 실제 대량의 말뭉치 기반 연구에 활용하기에는 제약이 크다.

[그림 2] 중국어 형태소 분석기 (Online Demo) 예시

	
<p>Stanford CoreNLP (https://corenlp.run/)</p> 	<p>THULAC (http://thulac.thunlp.org/demo)</p> 
<p>LTP (http://ltp.ai/demo.html)</p>	<p>CorpusWordParser (http://corpus.zhonghuayuwen.org/CpsWParser.aspx)</p>

위에서 살펴본 CorpusWordParser 역시 시범용 웹 페이지²⁵⁾의 경우 텍스트 처리 용량에 한계가 있으며 파일 업·출력 및 사용자 사전(customized user dictionary) 기능도 사용이 불가하다. 따라서 빅데이터 기반의 심화 연구를 진행하기 위해서는 부득불 다음의 CLI 기반 프로그램에 익숙해질 필요가 있다.

25) <http://www.aihanyu.org/cncorpus/CpsWParser.aspx>

2.2 CLI 기반 분석기

앞서 살펴본 GUI는 명령어를 입력할 필요 없이 메뉴를 선택함으로써 사용자가 손쉽게 사용할 수 있다는 장점이 있으나, 한편으로 이는 주어진 메뉴만을 사용하도록 함으로써 사용자의 자유도와 확장성을 제약하는 단점으로 작용한다. 또한, 개발자 입장에서도 메뉴바나 버튼 등의 요소를 별도로 설계해야 하는 부담이 있고, 비용적인 측면에서도 이를 구동하고 서비스하기 위한 시스템 및 서버 리소스가 추가로 소모되므로, 대중적이고 범용적인 프로그램이 아니라면 굳이 GUI 기반으로 프로그램을 설계·배포할 필요가 없다. 이 때문에 현재 대다수의 중국어 형태소 분석기는 CLI(Command Line Interface, 명령어 기반 인터페이스)로 구축되어 있다. 2022년 8월 현재 GUI 기반 중국어 형태소 분석기는 CorpusWordParser와 ICTCLAS뿐이며, 기타 분석기와 풍부한 고급 기능을 활용하기 위해서는 CLI 기반 중국어 형태소 분석기를 사용하여야 한다. 또한, CLI 기반 중국어 분석기는 형태소 분석 결과를 감정 분석(sentiment analysis), 구문 분석(parsing) 등의 후속 작업(downstream task)과 연계하여 응용하고 확장할 수 있기 때문에 종합적인 중국어 자연어처리를 위해서는 CLI 기반 분석기의 사용 방법을 익히는 것이 필요하다. 아래에서는 CLI 환경에 익숙하지 않은 사용자의 관점에서 몇 가지 중국어 형태소 분석기의 설치와 실행 방법을 소개하도록 한다.

CLI 기반의 중국어 품사 태거를 구동하기 위해서는 별도의 프로그래밍 언어 환경의 설치가 필요한데, 현재 가장 많은 중국어 어휘 분석기가 공통으로 사용하는 프로그래밍 언어는 파이썬(Python)이다(표 3). 로컬 컴퓨터에 파이썬을 설치하는 방식은 (1) 최소 설치²⁶⁾와 (2)통합개발환경(IDE) 설치²⁷⁾ 두 가지가 있는데, 초심자라면 버전 호환성과 안정성 측면에서 통합개발환경 설치를 권장한다.²⁸⁾

26) <https://www.python.org/downloads/>

27) 다양한 선택지가 있음. 아래 참조.

<https://wiki.python.org/moin/IntegratedDevelopmentEnvironments>

28) 파이썬 최소 설치만으로는 Scipy, Numpy, nltk, BeautifulSoup 등 주요 라이브러리를

[실행 예시] 파이썬 Anaconda 통합개발환경 설치 및 파이썬 구동

- [1] <https://www.anaconda.com/products/individual#Downloads>에서 installer 다운로드
- [2] 설치 완료 후 [윈도우즈❖ > Anaconda3 > Spyder] 클릭

2.2.1 LTP

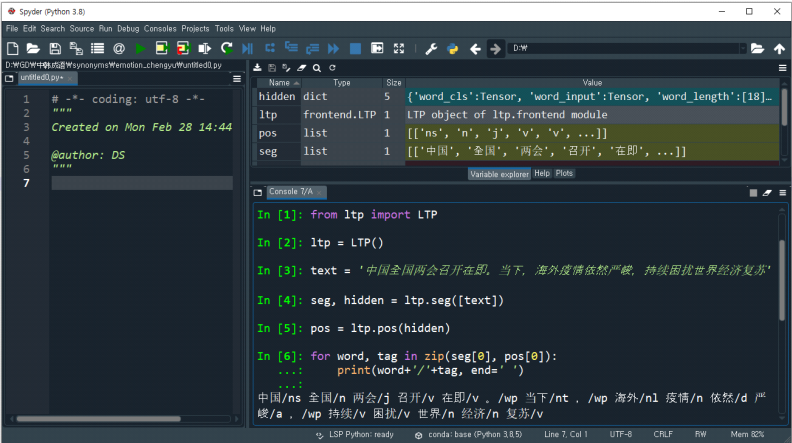
윈도우즈 운영체제, 파이썬 통합개발환경(IDE) 환경에서 하얼빈공대 LTP (Che et al, 2020) 중국어 형태소 분석기의 구동과 사용 방법을 살펴본다.

[실행 예시] 하얼빈공대 LTP

- [1] LTP 설치
[윈도우즈❖ > Anaconda3 > Anaconda Prompt] 클릭
Anaconda Prompt 명령 창에서 pip install ltp 입력 후 실행



- [2] Spyder 명령 창에서 LTP 로드 및 실행
[윈도우즈❖ > Anaconda3 > Spyder] 클릭하여 Spyder 실행
Spyder의 우측 하단 명령 창(console)에 아래 명령행 [line 1] ~ [line 6]을 순차적으로 입력하여 중국어 분석을 수행



[line 1-2] LTP 로드 및 LTP 클래스 객체 ltp 생성
[line 3] 분석 대상 텍스트(말뭉치) 입력 후 text 변수에 저장
[line 4] ltp.seg 로 단어분리 수행 후 seg 변수에 저장
[line 5] ltp.pos 로 품사태깅 수행 후 pos 변수에 저장
[line 6] seg, pos에 저장된 단어(word)와 품사(tag)를 '단어/품사' 형식으로 화면에 출력

위 코드에서 텍스트 입력(line 3)과 출력(line 6) 코드를 수정하면 파일로 말뭉치를 읽고 쓰는 것도 가능하다.

2.2.2 Jieba

‘말더듬이(结巴)’라는 명칭의 Jieba는 Tencent(腾讯) 엔지니어 Sun Junyi가 개발한 중국어 형태소 분석기이다. 윈도우즈 운영체제, 파이썬 통합개발환경 (IDE) 환경 하에서의 구동과 사용 방법은 아래와 같다.

[실행 예시] Jieba

- [1] Jieba 설치
[윈도우즈❖ > Anaconda3 > Anaconda Prompt] 클릭
Anaconda Prompt 명령 창에서 pip install jieba 입력 후 실행

```

Anaconda Prompt (anaconda3)

(base) C:\Users\LD> pip install jieba
Collecting jieba
  Downloading jieba-0.42.1.tar.gz (19.2 MB)
    Building wheels for collected packages: jieba
      Building wheel for jieba (setup.py) ... done
    Created wheel for jieba: filename=jieba-0.42.1-py3-none-any.whl size=19314482 sha256=f4b3ffe19327aea85a734ad05b0f4b191e2dd3ad2b94873f022b34422f45a
    Stored in directory: c:\users\ld\appdata\local\pip\cache\wheels\ca\35\d8\df\dte73bec1d12020b3cb7ce8de0f310ea2cf155ee018ae
Successfully built jieba
Installing collected packages: jieba
Successfully installed jieba-0.42.1

(base) C:\Users\LD>

```

[2] Spyder 명령 창에서 Jieba 로드 및 실행

[윈도우즈❖ > Anaconda3 > Spyder] 클릭, 우측 하단 명령 창 (console)에 명령행 [line 1] ~ [line 4] 입력 및 실행

```

# -*- coding: utf-8 -*-
"""
Created on Mon Feb 28 14:44
@author: DS
"""

1 #
2 """
3 Created on Mon Feb 28 14:44
4
5 @author: DS
6 """
7

Name      Type      Size      Value
seg        generator  1         generator object
tag         str        1         v
text       str        1         中国全国两会召开在即。当下，海外疫情依然严峻，持续困扰世界经济复苏
word       str        1         复苏

In [1]: import jieba.posseg as pseg
In [2]: text = '中国全国两会召开在即。当下，海外疫情依然严峻，持续困扰世界经济复苏'
In [3]: seg = pseg.cut(text)
In [4]: for word, tag in seg:
...:     print(word+'/'+tag, end=' ')
...:
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\DS\AppData\Local\Temp\jieba.cache
Loading model cost 0.768 seconds.
Prefix dict has been built successfully.
中国/ns 全国/n 两会/m 召开/v 在/p 即/v . /x 当下/t . /x 海外/s 疫情/n 依然/d 严
峻/a . /x 持续/vd 困扰/v 世界/n 经济/n 复苏/v

```

[line 1] Jieba 로드 후 pseg 라는 모듈 이름으로 저장

[line 2] 분석 대상 텍스트(말뭉치) 입력 후 text 변수에 저장

[line 3] pseg.cut 으로 단어분리 및 품사태깅 수행²⁹⁾ 후 seg 변수에 저장

[line 4] seg에 지정된 단어분리 및 품사태깅 수행 후 결과를 '단어/품사' 형식으로 화면에 출력

2.3 GUI 래퍼(Wrapper)

앞서 살펴본 LTP와 Jieba 이외에도 대부분의 품사 태거는 CLI 환경에서 동작한다(표 3). 문제는 중국어학 전공 학생들과 연구자들이 컴퓨터 전공자가 아니기

29) 품사태깅 작업 없이 단어분리만을 수행하려면 jieba.cut을 사용

에 CLI 환경에 익숙하지 못하다는 점이다. 2000년 이후 우리의 생활 깊숙이 자리한 인터넷 브라우저와 앱 등의 각종 소프트웨어는 기본적으로 GUI 기반이고 이러한 GUI 환경에 익숙해진 상황에서 마우스 대신 키보드, 아이콘 클릭 대신 명령어를 입력하여야 하는 CLI 기반 프로그램을 사용하는 것은 분명 쉽지 않은 도전이 될 수 있다.

이에 따라, 필자는 위에서 살펴본 CLI 기반 분석기의 실행을 돕는 래퍼(wrapper³⁰⁾) 프로그램을 윈도우즈 GUI 환경으로 구현하였다. 2022년 8월 현재 ICTCLAS, Jieba, THULAC 3개 분석기에 대한 GUI 래퍼를 개발하였고, 추후 요청에 따라 기타 분석기로 확장할 계획이다. 3개 래퍼 모두 동일한 인터페이스로 구현하였기 때문에 실행방법은 모두 동일하다. 아래에서 다운로드 및 사용방법을 설명하도록 한다.

[실행 예시] ICTCLAS / Jieba / THULAC windows GUI wrapper

(1) 파일 다운로드

웹페이지 <https://github.com/karmalet?tab=repositories>에서 'ictclas_GUI', 'jieba_GUI', 'THULAC-Python-GUI' 중 사용할 프로그램의 제목을 클릭하여 이동한다. 그 후, 'Windows에서 실행' 섹션의 '파일 다운로드' 링크를 클릭하여 압축 파일(.zip)을 다운로드한다.

(2) 압축 해제 및 실행

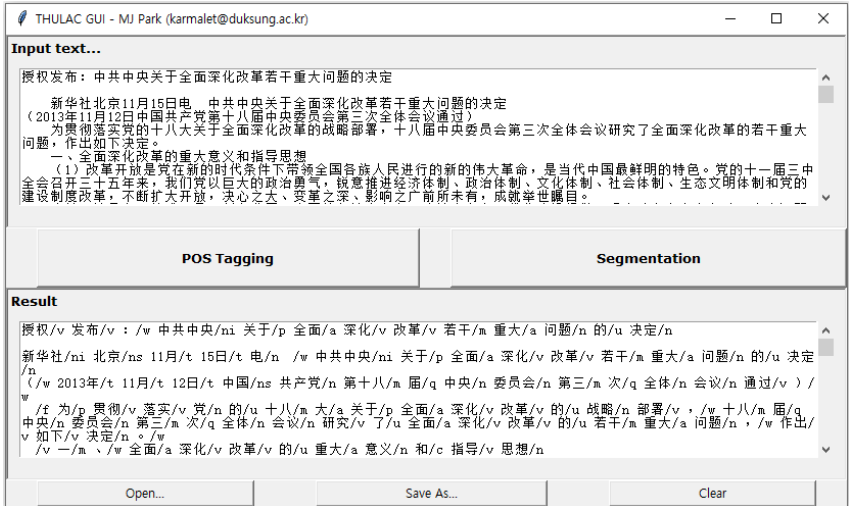
다운로드 받은 .zip 파일을 반드시 압축 해제 후, 해당 폴더 안의 .exe 파일을 실행한다. 최초 실행 시, 시스템에 따라 보안 경고 메시지가 나타나는데, 메시지를 무시하고 '실행' 버튼을 누르면 된다.

본 GUI 래퍼는 Windows 환경에서 동작하며, MacOS 등 기타 환경에서는 windows 실행파일용 에뮬레이터를 사용하여야 한다. CLI 환경에 익숙하지 않은 초보자를 위한 프로그램이므로 ICTCLAS, Jieba, THULAC 각 분석기의 가장 기초적인 기능만 구현하여 단순화하였으며 사용자가 로드한 텍스트 파일의 인

30) 문자 그대로 다른 프로그램을 감싸는 소프트웨어를 말하며 서로 다른 소프트웨어 간의 호환성 혹은 상호 운용성을 돕는 기능을 한다.

코딩(encoding)을 자동 인식하는 기능을 첨가하여 편의성을 높였다.

[그림 3] THULAC GUI 실행 화면



[Open] 실행 파일 로드 (파일 인코딩 자동 인식, 사용자 지정 불필요)

[Segmentation] 단어 분리만 수행

[POS tagging] 단어 분리 + 품사 태깅 수행

[Save As...] 단어 분리 결과창(Result)의 내용을 파일로 저장
(UTF-8 인코딩)

[Clear] 모든 창의 내용을 지움

다만, 이는 초심자의 접근성을 높이기 위한 입문용 프로그램일 뿐이며, 개별 중국어 형태소 분석기에 특화된 다양하고 풍부한 기능을 활용하기 위해서는 §2.2 결과 같이 CLI 환경에서 직접 코딩을 하여야 한다.

3. 각종 분석기의 특징 및 용도 분석

중지하듯이, 중국어의 단어(词)는 독립적으로 운용 가능한 최소 언어단위³¹⁾로 정의된다. 1992년에 제정된 정보처리용 중국어 단어분리규범(信息处理用现代汉语分词规范, GB/T13715)에도 이 정의가 명시되어 있다.³²⁾ 문제는 실제 중국어 정보처리에서 해당 정의의 실효성이 낮다는 점이다. 구문 및 담화적인 의미 맥락에 따라 중국어의 단어 구획은 유동적으로 변화하며 또한 필요에 따라 언어 외적인 사용 빈도를 고려하여야 할 경우도 존재한다. 가령, 鸡蛋을 하나의 단어(word)로 볼 것인가 아니면 구(phrase)로 처리할 것인가? 음식 재료의 관점에서는 鸡蛋을 荷包蛋, 皮蛋, 松花蛋 등과 같이 하나의 단어로 처리하는 것이 좋을 것이다. 그러나 특정 동물의 알(卵)이라는 관점에서는 鸡蛋, 鹅蛋, 蛇蛋 등은 수식어구(偏正词组)로 처리하는 것이 분석에 보다 용이할 것이다. 이처럼 중국어 단어(词)에 대한 학술적 정의가 그것이 실제 적용되는 개별 사례에 일률적으로 적용될 수는 없기 때문에, 중국어 형태소 분석기의 개발 주체와 사용 목적에 따라 단어 구획(分词)과 품사 태깅(词性标注) 기준이 저마다 조금씩 다르다. 일례로 각각의 분석기별로 채택하고 있는 품사 태그셋을 살펴보면 품사 분류에 다소간의 차이가 존재함을 알 수 있다(표 4).

31) 最小的有意义的独立运用的语言单位 (北京大学中文系现代汉语教研室, 2012)

32) 아래의 중국 국가표준(GB) 열람 시스템에서 찾아볼 수 있음

<https://openstd.samr.gov.cn/bzgk/gb/index>

[표 4] 주요 중국어 품사 태그셋(Tagset)

	GB/T20532 (속칭 863)	PKU	ICTPOS	CTB 3.0
주요 분류* (하위 분류)	20 (+29)	26 (+80)	22 (+77)	33 (+0)
규범	靳光瑾等 (2005)	俞士汶等 (2003)	刘群等(2004)	Fei(2000)
대표 분석기	LTP, CorpusWord Parser	ICTCLAS, THULAC	ICTCLAS	Stanford Tagger

* 품사 분류의 세부 사항은 [부록] 참조

현재 중국 경내에서 통용되는 품사 태그셋은 일반적으로 정보처리용 중국어 품사표기규범 (GB/T20532, 속칭 863 태그셋³³⁾)을 기반으로 각 대학·기관·연구소에 따라 자신의 목적에 부합하는 태그셋을 변용하여 사용한다. 대표적인 예로 북경대 태그셋(PKU)과 중국과학원 전산연구소 태그셋(ICTPOS)이 있다. [부록]을 살펴보면 각 기관별 품사 분류 체계의 특징이 드러난다. PKU는 상태사(状态词)를 주요 품사(一级分类)인 'z'로 분류하고 있으며, ICTPOS는 개사(介词, p)의 하위 분류로 '被'(pbei)와 '把'(pba)를, 동사(v) 중에서는 '是'(vshi)와 '有'(vyou)를 별도로 구분하여 태깅한다.

중국 바깥의 해외 학계에서는 Chinese Treebank (CTB, Xia 2000; Xue et al 2005)의 태그셋을 주로 사용한다. 이는 영어 구문분석 말뭉치인 PennTreebank (PTB, Marcus et al 1994)를 바탕으로 한 중국어 품사 태깅 체계로서, 형용사를 동사의 하위 분류로 처리하고(VA), 명사성 수식 성분(定语)을 영어의 형용사에 준하여 처리(JJ)하며³⁴⁾, 접속사를 대등 접속사(CC) 및 종

33) 863 태그셋은 1986년 3월 국가첨단기술발전계획(国家高技术研究发展计划, 속칭 863계획)의 일환으로 실시된 중국어 정보화사업, 그중에서도 품사표기규범 평가의 산물이며, 현재 널리 쓰이고 있는 拼音输入法, 五笔输入法 등의 중국어 입력 방식도 이때 정립되었다.

34) CTB 3.0 태깅 규범(Xia 2000)에 따르면, V+N형 (获奖/JJ 学者/NN), VA+N형 (高速

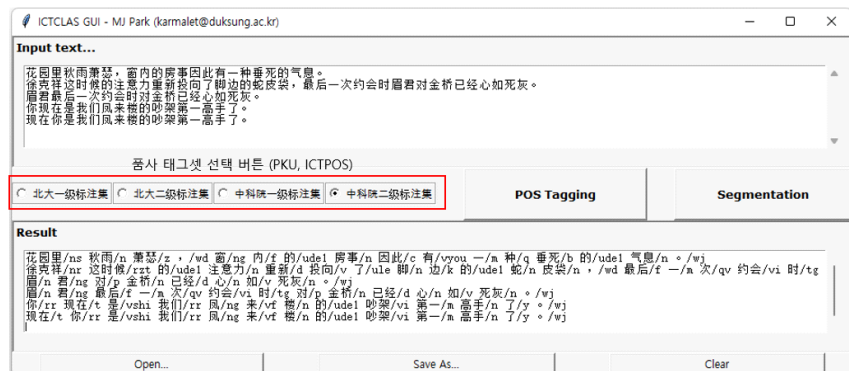
속(SC) 접속사로 세분하는 등 863 태그셋과는 상당히 다른 면모를 보여주고 있다.

품사 태그셋의 다양성만큼이나 이에 기반한 형태소 분석기들의 특징과 기능도 천차만별이다. 따라서 연구자의 활용 용도에 따라 알맞은 품사 태거를 선택하는 것이 매우 중요하다. 이어서 각각의 품사 태거별 고유한 특징을 살펴보도록 한다.

3.1 세밀한 품사 분류 - ICTCLAS

가장 세밀한 품사 태깅이 가능한 프로그램은 ICTCLAS이다. 중국 대륙 내에서 표준 품사표기규범으로 통용되고 있는 863 태그셋(GB/T20532)에 기초한 북경대학 태그셋(PKU)과 중국과학원 전산연구소 태그셋(ICTPOS)을 두루 사용할 수 있다. 부가적으로 인명(nr), 장소명사(ns), 고유명사(nz), 동명사(vn) 및 각종 구조조사(ude1/2/3) 등의 하위 품사 분류(二级分类)에 따른 세밀한 품사 태깅을 수행할 수도 있다. §2.3에서 소개한 ICTCLAS GUI 래퍼에서 하위 품사 태그셋(二级标注集) 버튼을 선택하면 편리하게 해당 기능을 활용할 수 있다.

[그림 4] ICTCLAS GUI 래퍼의 품사 태그셋 선택 (朴敏浚 2016)



/JJ 公路/NN), CD+N형 (两国/JJ 关系/NN) 등의 수식성분을 기타수식구(Other noun-modifier)라는 동일한 유형으로 분류하고 있다. 이는 사실상 중국어의 문장성분(定语)을 영어의 품사 분류(JJ)에 대응시킨 것이다.

ICTCLAS는 현재 북경이공대학의 张华平 교수팀이 개발한 NLPiR이라는 종합 자연어처리 소프트웨어 안에 포함되어 있으나, 본래 그 전신은 90년대 말 중국의 기초연구발전계획(73计划)의 일환으로 중국과학원 전산연구소(中科院计算所, Institute of Computing Technology, ICT)에서 개발한 중국어 어휘 분석 시스템(CT Chinese Lexical Analysis System, ICTCLAS)이다. ICTCLAS는 기계학습(계층적 은닉 마르코프 모델)을 이용하여 만들어졌으며³⁵⁾, 주요 학습 데이터는 중국 최초의 대규모 수동 품사 태깅 데이터인 《人民日报》 말뭉치에 기반하고 있다. 《人民日报》 말뭉치는 PKU 품사 태깅 규범(俞士汶等 2002; 俞士汶等 2003)에 따라 북경대학교 전산언어학연구소(北京大学计算语言学研究所) 연구원들이 3년에 걸쳐 세밀하고 엄격한 가공 및 검토 작업을 통해 획득한 고품질의 인공 주석 데이터(labeled data)이며, 이를 온전히 활용하고 있는 형태소 분석기는 ICTCLAS가 유일하기에³⁶⁾ 품사 태깅의 정밀성과 신뢰도가 높은 편이다.

3.2 기본에 충실한 교육용 분석기 - CorpusWordParser

중국 교육부 언어문자응용연구소(教育部语言文字应用研究所)에서 개발한 CorpusWordParser는 기타 중국어 형태소 분석기들과 달리 사용자 친화적인 그래픽 기반의 인터페이스(GUI)를 갖추고 있다. 별도의 프로그램 설치가 필요 없으며 실행 파일(.exe) 하나만을 다운로드하여 실행하면 동작한다.³⁷⁾ 따라서 GUI 래퍼 없이 마우스 클릭만으로 손쉽게 중국어 단어 분리 및 품사 태깅을 수행할 수 있다. 또한, 기타 분석기에서 CLI 환경에서만 활용할 수 있는 다양한 기능들(사용자 정의 사전, 다중 문서 일괄 처리 등)을 코딩 지식이 없어도 손쉽게 이

35) Zhang et al.(2003) 참조

36) 일반에는 《人民日报》 말뭉치의 8% 비중에 해당하는 200만 자 규모의 일부 데이터만 공개되었다. 즉, 비공개 데이터를 모두 활용한다는 측면에서 ICTCLAS는 우위를 지닌다.

37) 단, MacOS 등 기타 환경에서는 windows 실행파일용 에뮬레이터 필요

용할 수 있다는 장점이 있다.

다만 품사 태깅에 있어 중국어 품사표기규범 (GB/T20532) 단일 태그셋만을 사용³⁸⁾함으로써 PKU, ICTPOS 규범에 따른 품사 태깅은 불가능하다. 또한, 중국어 변체 인코딩 규격인 Big5를 지원하지 않기 때문에, 유니코드 인코딩이 아닌 중국어 변체를 포함한 텍스트 파일은 텍스트 깨짐이 나타날 수 있다.

비록 다양한 품사 태깅과 인코딩을 지원하지 않지만, 프로그램 설치가 필요 없고 구동이 안정적이며, 기본적인 기능에 충실하고 사용자 정의 사전 적용이 가능하다는 점에서 중국어 전처리 작업 및 교육용으로는 충분하다고 여겨진다.

3.3 용도별 맞춤 텍스트 처리 - PKUSEG, Jieba

때로는 구어·문어, 시기·주제별로 어느 특정한 영역(domain)에 속하는 텍스트를 분석하여야 하는 경우가 있다. 이 때 개별 영역의 텍스트를 따로 구분하여 맞춤 기계학습을 진행하면, 각 영역별 어휘, 문체 특성에 최적화된 개별 모델이 생성된다. 이는 일종의 커스터마이징 혹은 파인 튜닝(fine-tuning)으로 볼 수 있는데, 북경대학교 전산언어학실험실(计算语言学教育部重点实验室)의 언어계산·기계학습연구조(语言计算与机器学习研究组)에서 개발한 PKUSEG (Luo et al, 2019) 은 영역별로 최적화된 단어 분리 모드(预训练模型)를 제공한다.

[표 5] PKUSEG의 영역별 모델

영역(domain)	출처(source)	범용 모델
news	MSRA 데이터셋 ³⁹⁾ (인민일보)	mixed.zip (2022년 공개)
web	웨이보(微博) ⁴⁰⁾	
medicine	의료 분야 말뭉치 ⁴¹⁾	
tourism	관광 분야 말뭉치 ⁴²⁾	
art	위키피디아(wikipedia) 예술 분야 텍스트	default_v2.zip (2019년 공개)
entertainment	위키피디아(wikipedia) 연예·스포츠 분야 텍스트	
science	위키피디아(wikipedia) 과학 분야 텍스트	

38) 총 49가지 품사태그 중 35가지를 사용함

PKUSEG은 뉴스, 웹, 의학, 관광 등 영역별 텍스트를 기반으로 훈련된 분석기를 제공한다(표 5).⁴³⁾ 사용자는 작업 대상 텍스트의 특성에 따라 다양한 모델을 자유롭게 선택할 수 있다. 물론 대상 텍스트가 선택한 영역의 어휘·문체 특성에 잘 부합한다면 범용 모델보다 더 정확한 형태소 분석 결과를 기대할 수 있다. (구체적인 사항은 §4 결론 참조)

Jieba 형태소 분석기는 정확도 우선(精准模式), 최대 분리(全模式), 검색엔진 모드(搜索引擎模式)의 3가지 모드를 제공한다.

먼저, 정확도 우선 모드는 Jieba의 기본 모드이며, 여타 분리기와 마찬가지로 가장 가능성이 높은 단어 분리 결과를 반환한다. 이에 더해, Jieba 설명서⁴⁴⁾는 사용자 정의 사전에 없는 단어(新词)들도 인식하여 분리한다고 밝히고 있다.(예: 아래 표의 ‘杭研’⁴⁵⁾)

[표 6] Jieba의 세 가지 단어 분리 모드

모드	단어분리 예시
정확도 우선 精准模式	他来到了网易杭研大厦 [‘他’, ‘来到’, ‘了’, ‘网易’, ‘杭研’, ‘大厦’]
최대 분리 全模式	“乒乓球拍卖完了” [‘乒乓’, ‘乒乓球’, ‘乒乓球拍’, ‘球拍’, ‘拍卖’, ‘卖完’, ‘了’]
검색 엔진 搜索引擎模式	“他来到了中国科学技术大学” [‘他’, ‘来到’, ‘了’, ‘中国’, ‘科学’, ‘技术’, ‘科学技术’, ‘大学’]

둘째로, 최대 분리 모드(全模式)는 해당 문자열 내에서 분리 가능한 단어의 모든 경우를 나열하여 주는데, 이는 중의성(歧义)이 내재되어 있는 문자열을 분석

39) 제 2회 중국어 단어분리 공개경쟁 (SIGHAN 2005 Word Segmentation Bakeoff) 에서 마이크로소프트 아시아 연구센터(MicroSoft Research Asia)가 공개한 중국어 단어분리 데이터. 30만 단어(token) 규모의 인민일보 말뭉치를 수동 태깅한 자료임. [다운로드](<https://www.microsoft.com/en-us/download/details.aspx?id=52531>)

40) <https://github.com/FudanNLP/NLPCC-WordSeg-Weibo>(NLPCC2016 Shared Task)

41) 邱立坤等(2015)

42) 41)과 동일

43) 모델 다운로드: <https://github.com/lancopku/pkuseg-python/releases>

44) <https://github.com/foxsjy/jieba#readme>

45) ‘网易杭州研究院’의 약칭

하기에 적합하여 보인다.

마지막으로, 검색엔진 모드(搜索引擎模式)는 정확도 우선 모드의 분리 결과를 바탕으로, 길이가 긴 단어들을 다시금 분리하여 검색의 재현율(召回率, recall)을 높여주므로 검색엔진에 유용할 것이라고 Jieba 설명서는 제안하고 있다.⁴⁶⁾ 필자가 보기에 검색엔진 모드는 텍스트를 대표하는 단어들, 즉 키워드(关键词)를 추출해내는 데 유용할 것으로 보인다.

3.4 생성문법과의 연계성 - Stanford CoreNLP

앞서 소개한 ICTCLAS, CorpusWordParser, Jieba 등의 형태소 분석기는 모두 중국에서 개발된 것들로, 제각기 조금씩 상이한 부분은 있지만 대체적으로 중국어 단어분리 표기규범(GB/T13715)과 품사표기규범(GB/T20532, 속칭 863 태그셋)을 기반으로 하고 있다.

이와 달리, 중국 밖의 해외 영미권에서는 대개 CTB(Chinese Treebank⁴⁷⁾, Xia 2000: Xue et al 2005)를 표준 규범으로 삼고 있다. CTB는 중국 대륙에서 통용되는 863 태그셋과 다소 차이가 있는데, 아래 예시를 통해 살펴보면 “你/PN”(pronoun), “这/DT”(determiner), “后/LC”(localizaer), “大/JJ”(other noun modifier) 등 품사 표기가 863 태그셋과 상당히 다른 것을 알 수 있다. (보다 상세한 차이점은 [부록]을 참고)

(예) 为了/P 排毒/VV, /PU 你/PN 起床/VV 后/LC 要/VV 喝/VV 这/DT 大/JJ 杯/M 水/NN 或者/CC 那/DT 杯/M 鲜果汁/NN。/PU

CTB (Chinese Treebank)는 명칭 그대로 중국어 트리뱅크, 즉 통사구조 정

46) 搜索引擎模式, 在精确模式的基础上, 对长词再次切分, 提高召回率, 适合用于搜索引擎分词。(https://github.com/fxsjy/jieba#%E7%89%B9%E7%82%B9)

47) https://catalog.ldc.upenn.edu/LDC2013T21

보가 표기된 중국어 구문분석 말뭉치 구축을 위해 마련된 규범이다. 따라서 단어 분리와 품사 태깅은 최종 목표인 통사구조 표기 (syntactic bracketing)를 위한 준비적, 도구적 성격을 지닌다. CTB 규범을 따르는 대표적인 중국어 텍스트 처리 프로그램인 Stanford CoreNLP 시스템의 중국어 통사구조 분석 과정을 살펴 보면 다음과 같다.

[실행 예시] Stanford CoreNLP(ver 4.2.0) 를 활용한 중국어 통사구조 분석

(입력 문장 예시) 前面走着两个女同学，她们交头接耳地谈着话。(input_sample.txt)

[1] Stanford Chinese Word Segmenter (Chang et al, 2008) - CLI (GUI 미지원)
 c:\stanford-segmenter-2020-11-17> segment.bat ctb input_sample.txt UTF-8 0
 (출력 문장 예시) 前面 走 着 两 个 女 同 学 ， 她 们 交 头 接 耳 地 谈 着 话 。
 (주의) input 파일은 반드시 Unicode (UTF-8) 형식이어야 함. GB 불가.

[2] Stanford Chinese Part-of-Speech Tagger (Toutanova et al, 2003) - GUI 지원
 c:\stanford-postagger-full-2020-11-17> java -mx200m -cp "stanford-postagger.jar;" edu.stanford.nlp.tagger.maxent.MaxentTaggerGUI -model models\chinese-distsim.tagger
 Loading POS tagger from models\chinese-distsim.tagger ... done [0.5 sec]



(주의) 반드시 단어 분리가 완료된 문장을 넣어야 함.

[3] Stanford Chinese Parser (Levy et al, 2003) - GUI 지원

1. lexparser-gui.bat 더블클릭
2. Load Parser 클릭 > stanford-corenlp-4.2.0-models-chinese.jar 파일⁴⁸⁾ 선택
3. 팝업 창에서 'edu/stanford/nlp/models/lexparser/' 경로의 모델 중 택일
4. 중국어 파서(parser) 로드 완료

48) <https://nlp.stanford.edu/software/stanford-corenlp-4.2.0-models-chinese.jar> 에서 다운로드

Parser

File

Load File

Load Parser

Save Output

< Prev

Next >

Parse

Parse >

Clear

前面走着两个女同学，她们交头接耳地谈着话。

Parser: edu/stanford/nlp/models/lexparser/chineseFactored.ser.gz

ROOT

IP

IP

NP

NN

前面

VP

VV

走

AS

着

NP

QP

CD

两

CLP

M

个

ADJP

JJ

女

NN

同学

PU

,

IP

NP

PN

她们

VP

DVP

VP

VA

交头接耳

DEV

地

VV

谈

AS

着

NP

NN

话

PU

.

Done

Bracketed Chinese Tree:

(ROOT (IP (IP (NP (NN 前面)) (VP (VV 走) (AS 着) (NP (QP (CD 两) (CLP (M 个))) (ADJP (JJ 女)) (NP (NN 同学))))) (PU ,) (IP (NP (PN 她们)) (VP (DVP (VP (VA 交头接耳)) (DEV 地)) (VP (VV 谈) (AS 着) (NP (NN 话))))) (PU .)))

- [1] Stanford Chinese Segmenter⁴⁹⁾를 CMD에서 실행
- [2] Stanford Chinese Part-of-Speech Tagger⁵⁰⁾를 CMD에서 실행
 - 입력 창에 [1]의 단어 분리 결과를 넣고 'Tag sentence!' 클릭
- [3] Stanford Chinese Parser⁵¹⁾를 실행
 - 입력 창에 [1]의 단어 분리 결과를 넣고 'Parse' 클릭
 - 구문 분석 (Parsing) 결과가 수형도로 나타나며, 'Save Output' 버튼을 클릭하면 분석 결과를 위의 [3] Bracketed Chinese Tree 형식으로 저장 가능함

※ 주의: 실행 환경에 Java가 반드시 설치되어야 함. 유틸 메모리 1G 이상 권장.

49) <https://nlp.stanford.edu/software/segmenter.shtml>
50) <https://nlp.stanford.edu/software/tagger.shtml>
51) <https://nlp.stanford.edu/software/lex-parser.shtml>

위와 같이, Stanford CoreNLP가 제공하는 [1] 중국어 단어 분리 - [2] 품사 태깅 - [3] 구문 분석으로 이어지는 작업 흐름을 활용하면 최종적으로 CTB 형식의 수행도를 얻을 수 있다. Xue et. al.(2005)에 따르면, CTB는 X-bar 이론과 함께 GB 이론의 이론적 토대 위에서 구축되었으므로, Stanford CoreNLP는 생성문법의 관점에서 중국어 통사론을 연구하는 데 유용한 도구가 될 것이다.

다만, 단어 분리, 품사 태깅 및 구문 분석기가 각기 다른 별도의 프로그램으로 실행되기 때문에 Java에 익숙한 사용자가 아니라면 사용에 어려움을 느낄 수 있다.⁵²⁾ 이 경우, 북경대학의 중국어 트리뱅크(PKU Treebank)를 기반으로 하는 Cparser (Zhan et al, 2004)를 대안으로 고려해 볼 수 있다. Cparser는 중국어 텍스트의 단어 분리, 품사 태깅 및 구문 분석을 윈도우즈 실행 파일(.exe) 클릭만으로 수행할 수 있는 일원화된 통합 환경(IDE)을 제공한다.⁵³⁾

4. 결론

지금까지 ICTCLAS 등 6가지 중국어 형태소 분석기의 사용 방법과 주요 특징 및 용도에 관하여 살펴보았다. 혹자는 ‘그래서 어느 분석기의 성능이 가장 좋은 데?’라고 되물을지 모르겠다. 결론부터 말하면 중국어 분석기의 객관적 성능 비교는 어려우며, 2022년 현재 성능은 상향평준화 되어 그 비교 의미가 퇴색되었다.

2005년 중국어 단어분리 콘테스트(SIGHAN CWS bakeoff 2005)⁵⁴⁾ 이후, 세상에 모습을 드러낸 각종 중국어 단어 분리기의 성능(정확도) 비교는 투명하게 이루어지고 있지 못한 실정이다. 이러한 배경에는 중국어가 영어에 비해 상대적으로 소수 언어라는 점, 벤치마크용 공용 실험 데이터가 부족하다는 한계가 존재

52) 사실 사용상의 문제뿐 아니라 전체 NLP 작업 흐름에서 시너지 효과를 기대하기 어렵다. 자세한 내용은 Che et al.(2020) 참조

53) 구체적인 사용방법과 응용예시는 박민준·강병규(2019) 참조.

54) <http://sighan.cs.uchicago.edu/bakeoff2005/>

한다. 여기에 2017년 이후 단어 임베딩(Word Embedding)을 입력으로 하는 딥러닝 모델(Transformer, Vaswani et al 2017)이 주류가 되면서, 점차 더 많은 모델이 단어 분리 및 품사 태깅 단계를 생략하거나 문자열을 그대로 입력받아 처리하는 end-to-end 방식을 채택하는 추세 전환(강병규·박민준 2022)도 한몫을 했다. 최근 문맥 정보를 민감하게 벡터 공간에 표상하는 임베딩 기술이 발전하면서, 극단적인 경우 후행 과제(downstream task)인 감정 분석, 기계번역 등에서 선행 과제(upstream task)로서의 단어 분리 과정을 생략하고도 충분한 성능 향상을 이루기도 한다(Yang et al. 2016, Su et al. 2017, Li et al. 2019).

위와 같은 종합적인 사유로 인해 전통적인 세부 과제로서의 중국어 단어 분리(CWS)의 성능 비교만을 전문적으로 다루는 학술 논문 혹은 공개 경쟁⁵⁵⁾은 찾아보기 힘들어졌다. 하지만 제한적으로나마 PKUSEG, THULAC 홈페이지에서 성능 비교 결과를 공개하고 있기에(표 7, 8) 이를 통해 간접적으로 각 모델의 성능을 유추할 수 있다.

[표 7] 중국어 단어 분리 정확도 비교 (f-score, PKUSEG 제공⁵⁶⁾)

말뭉치 분석기	MSR	CTB8	PKU	WEIBO	macro average
Jieba	81.45	79.58	81.83	83.56	81.61
THULAC	85.55	87.84	92.29	86.65	88.08
PKUSEG	87.29	91.77	92.68	93.43	91.29

PKUSEG은 각 영역별 말뭉치를 대상으로 Jieba, THULAC, PKUSEG 세 모델의 성능 테스트 결과를 게시하고 있다. [표 7]에 따르면, 신문(MSR, PKU), 웨이보(WEIBO) 및 종합 말뭉치(CTB8)에서 PKUSEG의 성능이 가장 뛰어난 것으로 나타난다.

55) <https://paperswithcode.com/task/chinese-word-segmentation#benchmarks>

56) <https://github.com/lancopku/pkuseg-python>

[표 8] 중국어 단어 분리 정확도 비교 (THULAC 측정값⁵⁷⁾으로부터 도출한 f-score)

분석기	말뭉치	MSR ⁵⁸⁾		PKU	
		f-score	time	f-score	time
	LTP-3.2.0	88.13	3.21s	95.35	3.83s
	ICTCLAS(2015版)	89.09	0.55s	94.15	0.53s
	Jieba	81.15	0.26s	81.57	0.23s
	THULAC	88.79	0.62s	92.57	0.51s

청화대학의 THULAC(Li and Sun 2009)은 중국어 단어분리 콘테스트 (SIGHAN CWS bakeoff 2005)의 테스트셋 중에서 마이크로소프트(MSR)와 북경대학(PKU) 말뭉치를 대상으로 LTP(v.3.2), ICTCLAS(v.2015), Jieba, THULAC 네 모델의 성능을 비교하였다. LTP와 ICTCLAS가 정확도 면에서 우수한 성능을 보여주고 있으며, 처리 속도 측면에서는 Jieba가 두드러진다.

[표 7]에서 동일한 PKU 말뭉치를 대상으로 THULAC과 PKUSEG이 거의 동등한 정확도(약 92%)를 보여주고 있으므로, 비록 [표 8]에 PKUSEG의 측정 기록은 없지만 THULAC과 비슷한 수준이라고 가정한다면 PKUSEG의 상대적인 성능을 가늠해 볼 수 있다.

[표 7]과 [표 8]의 정확도 수치를 살펴보면 MSR, PKU의 동일한 데이터셋을 가지고도 실험을 했음에도 차이가 난다. 이는 실험 수행과정에 수반되는 여러 가지 변수들로 인해 성능 측정에 오차가 발생하기 때문이다.⁵⁹⁾ 일반적으로 자연어 처리분야에서 85-90%를 상회하는 F-score는 상당히 높은 정확도이며, 한 자리

57) <https://github.com/thunlp/THULAC-Python>

58) 같은 MSR 데이터셋도 버전에 따라 훈련 데이터와 테스트 데이터 셋이 다름. (v4: <http://sighan.cs.uchicago.edu/bakeoff2005/>, v5: <https://www.microsoft.com/en-us/download/details.aspx?id=52531>) [표 7, 8]의 차이는 이에 따른 성능 측정 편차로 보임.

59) 각 모델별 파라미터 설정, 훈련 및 테스트 데이터의 분할, 실험을 수행하는 컴퓨터의 성능 등 수많은 변수가 수치 측정에 영향을 미친다. 이 때문에 동일 컴퓨팅 환경, 동일 데이터로 수행하는 공개 경쟁 대회가 아니면 객관적인 성능 평가가 현실적으로 어렵다. PKUSEG 랩도 이러한 점을 인식하고 있다. (https://github.com/lancopku/pkuseg-python/wiki/FAQ#4-E6%98AF%E5%A6%82%E4%BD%95%E8%B7%9F%E5%85%B6%E5%AE%83%E5%B7%A5%E5%85%B7%E5%8C%85%E5%9C%A8%E5%A4%9A%E9%A2%86%E5%9F%9F%E6%95%B0%E6%8D%AE%E4%B8%8A%E8%BF%9B%E8%A1%8C%E6%AF%94%E8%BE%83%E7%9A%84))

수 차이는 공학적 측면에서는 커다란 차이로 여겨질 수 있으나 일반 사용자의 관점에서는 모델별 우열을 체감하기 어려운 수준이라고 생각한다.

따라서 [표 7]과 [표 8]의 정확도 수치는 참고용으로 삼되, 정확도보다는 연구자 개인의 연구 목적과 분석 대상 말뭉치의 특징, 그리고 형태소 분석기의 기능 및 활용가능성을 기준으로 본인의 연구에 가장 적합한 분석기를 선택하는 것이 중요하다. 이때, [표 3]의 중국어 형태소 분석기 현황 및 [표 4]의 각종 품사 태그셋에 대한 이해를 바탕으로 3장에서 제시한 분석기별 특징을 참고한다면, 자신의 필요에 맞는 중국어 분석기를 고르는 데 도움이 될 것이다. 또한, [부록]의 중국어 품사 태그셋 비교표를 참고한다면 분석기마다 서로 다른 품사 태그셋에서 기인하는 혼동을 줄일 수 있으리라 기대한다. 이와 같은 중국어 형태소 분석기에 대한 전면적인 현황 분석과 개별 품사 태그셋 간의 세부 비교는 중국 학계를 포함하여 본고에서 최초로 제시되는 것이라고 필자는 생각한다.

무엇보다도 본고의 가장 큰 의의는 중국어 형태소 분석의 입문서 역할에 있다. 이에 따라 최신 형태소 분석기⁶⁰⁾의 훈련(training)과 작동 원리 등의 서술은 과감히 생략하고 GUI, CLI 등 다양한 환경에서의 실행 예시를 풍부히 제시하고자 노력하였다. 특히, 컴퓨터에 익숙하지 않은 언어학 연구자들을 위해 단순화된 윈도우 실행 프로그램을 개발·배포하였다. § 2.3 절의 GUI 래퍼를 활용하면 코딩에 익숙하지 않은 연구자도 부담 없이 마우스 클릭만으로 중국어 단어 분리와 품사 태깅을 할 수 있을 것이다. 부디 본고가 중국어 양적 분석 연구의 유용한 길잡이가 되기를 바라며 글을 마친다.

< 參考文獻 >

강병규·박민준(2022), 「중국의 자연어처리 연구 현황과 발전 추세」, 『언어와 정보 사회』 45, 193-231.

60) 최신 중국어 형태소 분석기의 진전은 다음을 참고할 것.

http://nlpprogress.com/chinese/chinese_word_segmentation.html

- 박민준·강병규(2019), 「중국어 구문분석기의 작동 원리와 응용 사례 - Cparser를 중심으로」, 『중국어언어연구』 82, 233-266.
- 박민준·이창호(2019), 「중국어 트리뱅크의 시각화」, 『언어와 정보사회』 38.
- 이창호·이지현(2019), 「트리뱅크 (TreeBank)와 중국어교육」, 『중국어언어연구』 81, 205-243.
- 北京大学中文系现代汉语教研室(2012), 《现代汉语(增订本)》, 北京: 商务印书馆.
- 刘群·张华平·俞鸿魁·程学旗(2004), <基于层叠隐马模型的汉语词法分析>, 《计算机研究与发展》 08, 1421-1429.
- 朴敏浚(2016), <面向初学者的中文信息处理平台构建及应用>, 《数字化汉语教学》 10, 227-232.
- 邱立坤·史林林·王厚峰(2015), <多领域中文依存树库构建与影响统计句法分析因素之分析>, 《中文信息学报》 05, 69-75.
- 俞士汶·段慧明·朱学锋·孙斌(2002), <北京大学现代汉语语料库基本加工规范>, 《中文信息学报》 05, 49-64.
- 俞士汶·段慧明·朱学锋·孙斌·常宝宝(2003), <北大语料库加工规范: 切分·词性标注·注音>, 《汉语语言与计算学报》 13(2), 121-158.
- 靳光瑾·肖航·富丽. (2005). <信息处理用现代汉语词类标记规范(修订)>, 《第四届全国语言文字应用学术研讨会论文集》.
- Bird, S., Klein, E., & Loper, E. (2009), *Natural language processing with Python: analyzing text with the natural language toolkit*, O'Reilly Media, Inc.
- Che, W., Feng, Y., Qin, L., & Liu, T. (2020), N-ltp: A open-source neural chinese language technology platform with pretrained models, *arXiv:2009.11616*.
- Chang, P. C., Galley, M., & Manning, C. D. (2008), Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the third workshop on statistical machine translation*, 224-232.
- Levy, R., & Manning, C. D. (2003, July), Is it harder to parse Chinese, or the Chinese Treebank?. In *Proceedings of the ACL 2003*, 439-446.
- Li, X., Meng, Y., Sun, X., Han, Q., Yuan, A., and Li, J. (2019), Is word segmentation necessary for deep learning of Chinese representations? *arXiv:1905.05526*.

- Li, Z., & Sun, M. (2009), Punctuation as implicit annotations for Chinese word segmentation, *Computational Linguistics* 35(4), 505-512.
- Luo, R., Xu, J., Zhang, Y., Ren, X., & Sun, X. (2019), Pkuseg: A toolkit for multi-domain chinese word segmentation. arXiv preprint *arXiv:1906.11455*.
- Marcus, Mitchell & Marcinkiewicz, Mary & Santorini, Beatrice (2002), Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics* 19, 313-330.
- Qiu, X., Qian, P., & Shi, Z. (2016), Overview of the NLPCC-ICCPOL 2016 shared task: Chinese word segmentation for micro-blog texts. In *Natural Language Understanding and Intelligent Applications* (pp. 901-906). Springer, Cham.
- Qiu, X., Zhang, Q., & Huang, X. J. (2013), Fudannlp: A toolkit for chinese natural language processing. In *Proceedings of the ACL 2013: system demonstrations* (pp. 49-54).
- Su, J., Tan, Z., Xiong, D., Ji, R., Shi, X., and Liu, Y. (2017), Lattice-Based Recurrent Neural Network Encoders for Neural Machine Translation, In *Proceedings of the AAAI Conference on Artificial Intelligence* 31(1).
- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003), Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the NAACL 2003*, 252-259.
- Tseng, H., Chang, P., Andrew, G., Jurafsky, D., & Manning, C. (2005), A conditional random field word segmenter. In *Fourth SIGHAN Workshop*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017), Attention is all you need, *Advances in neural information processing systems* 30.
- Xia, F. (2000), The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0), *IRCS Technical Reports Series* 38.
- Xue, N., Xia, F., Chiou, F. D., & Palmer, M. (2005), The penn chinese treebank: Phrase structure annotation of a large corpus, *Natural language engineering* 11(2), 207-238.
- Yang, Z., Chen, W., Wang, F., & Xu, B. (2016), A character-aware encoder for

- neural machine translation, In *Proceedings of the 26th International Conference on Computational Linguistics*, 3063-3070.
- Zhan, W., et al. (2004), An Integrated Chinese Grammar Development Environment, *The 4th China-Japan joint conference to promote cooperation in Natural Language Processing (CJNLP-04)*, City University of Hong Kong.
- Zhan, W. (2016), Peking University Treebank([pdf]), In *Encyclopedia of Chinese Language and Linguistics*, Volume 3, 332-336. First published online: 2015, Brill Publishing House.
- Zhang, H., et al.(2003), HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the second SIGHAN workshop on Chinese language processing* (17), 184-187.

[부록] 중국어 품사 태그셋 비교표

품사	예시	GB/T20 532 (863)	PKU	ICTPOS	CTB
		靳光瑾等 (2005)	俞士汶等 (2003)	刘群等(2004)	Fei(2000)
形容词	好 高 危险 漂亮 干净	a	a	a	JJ/VA
区别词	男 共同 大型 西式	f	b	b	JJ
连词	和 与 并 或 虽然 并且 而且 因为	c	c	c	CC, CS
副词	都 就 很 不 却 正在 重新 曾经 居然	d	d	d	AD
叹词	嗯 唉 哼 哦 哎哟	e	e	e	IJ
方位词	上下左右前后里外 前边 左面 里头 中间	nd	f	f	LC
语素	民 究 遥	g	g	ng,vg,ag	-
前接成分 (前缀)	小 老 啊	h	h	h	-
成语	百花齐放 史无前例 平白无故 彬彬有礼	i	i	nl,vl,al	-

품사	예시	GB/T20532(863)	PKU	ICTPOS	CTB
		靳光瑾等(2005)	俞士汶等(2003)	刘群等(2004)	Fei(2000)
简称略语 (缩略语)	公检法	j	j	-	-
后接成分 (后缀)	们 儿 界 率	k	k	k	-
习用语	绘声绘色 木头疙瘩 五大三粗 居高临下	i	l	nl,vl,al	-
数词	一百 第一	m	m	m	CD OD
名词	书 人 马 飞机 桌子 美国 北京 浙江 中关村	n ns	n ns	n ns	NN, NR
拟声词	扑通 咕咚 叮叮当当	o	o	o	ON
介词	从 被 给 把 将	p	p	p pbei pba	P LB,SB BA
量词	个 条 片 辆 斤 次 遍 公里	q	q	q	M
代词	这 那 大家 我 你 他 我们 你们 谁 哪里 怎么	r	r rr(人称代词) ry(疑问代词)	r rr(人称代词) ry(疑问代词)	DT PN
处所词	前院,后街,右侧,地上	nl	s	s	NN
时间词	年 月 日 分 秒 现在 过去 昨天 去年 星期一	nt	t	t	NT
助词	的 (的, 得, 地) 着 所 了 过 等	u	u (ud, ue, ui) uz us ul uo	u (ude1, ude2, ude3) uzhe usuo ule uguo	DEC, (DEG, DER, DEV) AS MSP AS/SP AS ETC
动词	吃 打 洗 喜欢 告诉 是 有	v	v	v vshi vyou	VV VC(是) VE(有)

품사	예시	GB/T20 532 (863)	PKU	ICTPOS	CTB
		靳光瑾等 (2005)	俞士汶等 (2003)	刘群等(2004)	Fei(2000)
标点符号	, 。 、 : ? ! “ ” ……	w	w	w	PU
非语素字 (字符串)	鸱, 鸱, 葡, 萄, 窈, 窕 =) karmalet@duksung.ac.kr	x	x	x	FW
语气词	了, 吗	u	y	y	SP
状态词	雪白 通红 冰凉 绿油油 冷冰冰	as	z	z	VA

* 로마자 1자리는 주요 분류(一级分类), 2자리는 하위 분류(二级分类)임 (CTB는 하위분류 없음).

< Abstract >

A Practical Guide to Chinese Word Segmenters and POS Taggers

Park, Minjun

This paper introduces and explains in detail the overall information and tutorial about the commonly used Chinese morphological analyzers (e.g. ICTCLAS, Jieba, Stanford CoreNLP) which are employed in Chinese preprocessing tasks of Chinese Word Segmentation (CWS) and Part-of-speech tagging. In particular, the usability of the tools was enhanced by developing simple executables distributed to linguistic researchers unfamiliar with coding, along with rich execution examples in GUI and CLI environments. Plus, by introducing the unique features and functions of each morphological analyzer, it was recommended the most suitable analyzer tailored to the needs of individual researchers. As a guide for Chinese morphological analysis, which is inevitably accompanied by data-driven quantitative research, this study presents practical

tools and useful guidelines for Chinese text preprocessing to researchers who want to expand their research interests to corpus linguistics, computational linguistics, and natural language processing.

Key Words: Chinese Word Segmentation, POS tagging, ICTCLAS, Jieba, LTP, Stanford tagger, CorpusWordParser, THULAC, PKUSEG

투고일: 2022.08.29. / 심사일: 2022.09.20. ~ 2022.09.26. / 게재확정일: 2022.09.27.
--