한국어 형태소 분석기 성능 비교

2018.12.10 00:58

[개발 이야기](https://iostream.tistory.com/category/%EA%B0%9C%EB%B0%9C%20%EC%9D%B4%EC%95%BC%EA%B8%B0)

[형태소 분석기 비교](https://iostream.tistory.com/144#header-n11)[실험 환경](https://iostream.tistory.com/144#header-n1745)[하드웨어 (MacBook Pro Mid 2015)](https://iostream.tistory.com/144#header-n20)[소프트웨어](https://iostream.tistory.com/144#header-n87)[데이터](https://iostream.tistory.com/144#header-n97)[실험 내용](https://iostream.tistory.com/144#header-n7)[실행 시간 비교](https://iostream.tistory.com/144#header-n14)[로딩 시간](https://iostream.tistory.com/144#header-n15)[형태소 분석 시간](https://iostream.tistory.com/144#header-n30)[문장 분석 품질 비교](https://iostream.tistory.com/144#header-n1882)[띄어쓰기가 없는 문장](https://iostream.tistory.com/144#header-n1125)[자소 분리 및 오탈자가 포함된 문장](https://iostream.tistory.com/144#header-n76)[긴 문장](https://iostream.tistory.com/144#header-n78)[결론](https://iostream.tistory.com/144#header-n1479)

형태소 분석기 비교

자연언어처리 모델 설계 전 데이터 전처리는 매우 중요합니다. 특히, 한국어 자연언어처리에서는 문장 분해의 여러 가지 최소 단위를 가질 수 있습니다. 한국어에서 최소 단위는 자소, 음절, 형태소 등이 될 수 있고 자소나 음절 분해보다 형태소 분해는 문장에서의 위치나 문맥에 따라 달라지므로 쉽지 않은 문제가 있는데요. 예를 들어 '하늘을 나는 자동차'와 '나는 밥을 먹는다'에서 '나는'은 문맥에 따라 주어 '나'를 의미할 수도 있고 '날다'를 의미할 수도 있습니다.

최근 카카오에서 딥러닝 기반의 형태소 분석기 [khaiii](https://github.com/kakao/khaiii)를 발표했습니다. 기존에도 여러 형태소 분석기(한나눔, KOMORAN 등)들이 존재했지만, 딥러닝 기반의 형태소 분석기는 제가 아는 한 처음이었기 때문에 관심이 생겨 여러 형태소 분석기의 성능(수행 시간, 분석 품질 등)을 비교해보고자 합니다.

실험 환경

하드웨어 (MacBook Pro Mid 2015)

macOS Mojave (Ver 10.14.1)

Processor: Intel Core i7 2.5GHz

Memory: 16GB 1600MHz DDR3

소프트웨어

Python 3.6.5

KoNLPy 0.5.1

데이터

[Naver sentiment movie corpus](https://github.com/e9t/nsmc)

실험 내용

실행 시간 비교

Python의 [timeit](https://docs.python.org/3.6/library/timeit.html)를 활용해서 측정

로딩 시간

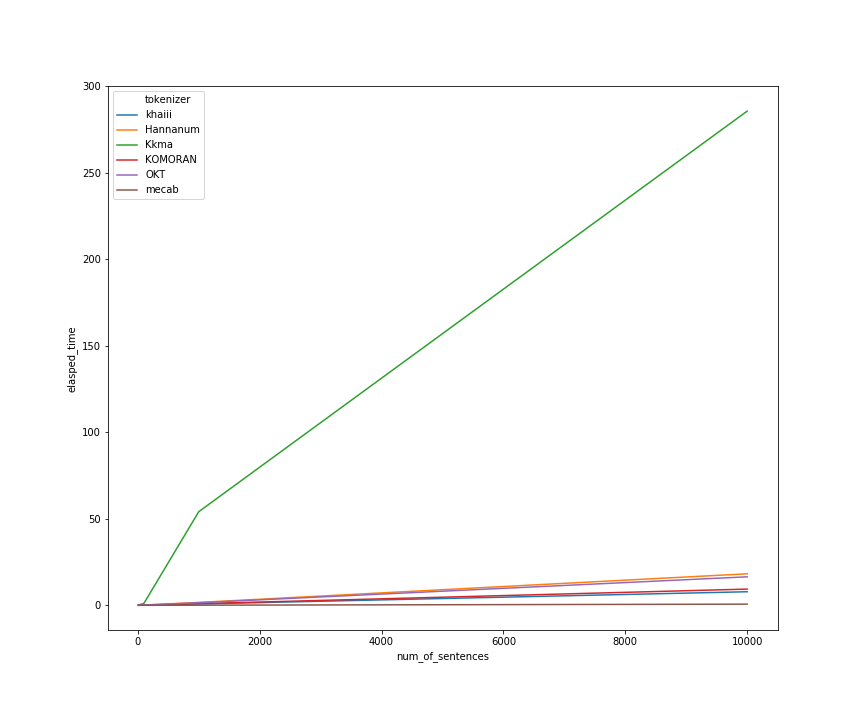
로딩 시간은 100회의 반복 수행 후 최솟값을 로딩 시간으로 간주함

| **분석기** | **로딩 시간 (초)** |
| --- | --- |
| khaiii | 0.0016 |
| 한나눔 | 0.0001 |
| 꼬꼬마 | 0.0002 |
| KOMORAN | 0.9542 |
| Open Korean Text (이하 OKT) | 0.0001 |
| mecab | 0.0004 |

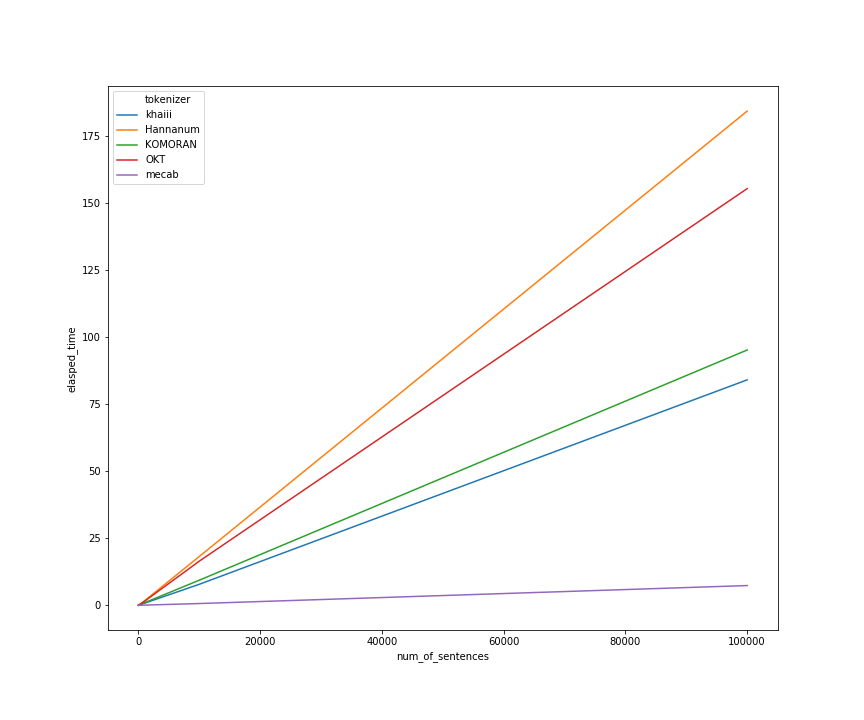
KOMORAN의 로딩 시간이 다른 분석기의 로딩 시간에 비해 눈에 띄네요.

형태소 분석 시간

우선 10, 100, 1000, 10000 문장에 대해서 형태소 분석 시간을 계산해봤습니다. 꼬꼬마 분석기가 갯수가 늘어날 수록 압도적으로 시간이 많이 소요되는 것을 볼 수 있네요.



시간이 오래걸리는 꼬꼬마 분석기를 제외하고 10만 문장까지 늘려서 분석 시간을 체크해봤습니다.



mecab이 압도적으로 빠른 형태소 분석 시간을 보여주고 있네요. 그 뒤를 khaiii와 KOMORAN이 따라가고 있습니다. KOMORAN은 Java 기반 형태소 분석기임에도 khaiii와 큰 차이 없는 수행 시간을 보여주네요.

문장 분석 품질 비교

형태소 분석기의 수행 속도에 대해서는 위에서 비교를 해봤습니다. 수행 속도도 중요하지만, 더 중요한 것은 입력에 대해서 정말 분석을 잘하는 지가 더 중요하다고 생각합니다. 따라서, 분석기가 쉽게 처리하기 어려운 유형 3가지에 대해서 실험을 해봤고 예제 문장에 대해서 결과를 공유합니다. 아래 3개 유형은 쉽진 않지만, 모바일 기기의 사용량이 많은 현재 충분히 입력으로 들어올 수 있는 유형이라고 생각합니다.

* 띄어쓰기가 없는 문장
* 자소 분리 및 오탈자가 포함된 문장
* 긴 문장

띄어쓰기가 없는 문장

실험 문장 1: 너무기대안하고갔나재밌게봤다

| khaiii | 한나눔 | 꼬꼬마 | KOMORAN | OKT | mecab |
| --- | --- | --- | --- | --- | --- |
| 너/MAG | 너무기대안하고갔나재밌게봤다/N | 너무/MAG | 너무/MAG | 너/Modifier | 너무/MAG |
| 무/NNG |  | 기대/NNG | 기대/NNG | 무기/Noun | 기대/NNG |
| 기/MAG |  | 안/NNG | 안/NNG | 대안/Noun | 안/MAG |
| 대안/NNG |  | 하/XSV | 하/XSV | 하고/Josa | 하/VV |
| 하/XSV |  | 고/ECE | 고/EC | 갔나/Verb | 고/EC |
| 고/EC |  | 가/VV | 가/VX | 재밌게/Adjective | 갔/VV+EP |
| 가/VX |  | 었/EPT | 았/EP | 봤다/Verb | 나/EC |
| 았/EP |  | 나/EFQ | 나/EC |  | 재밌/VA |
| 나/VV |  | 재밌/VA | 재밌/VA |  | 게/EC |
| 재밌/VA |  | 게/ECD | 게/EC |  | 봤/VX+EP |
| 게/EC |  | 보/VXV | 보/VX |  | 다/EC |
| 보/VV |  | 았/EPT | 았/EP |  |  |
| 았/EP |  | 다/EFN | 다/EC |  |  |
| 다/EC |  |  |  |  |  |

실험 문장 2: 굉장히잘만든수작지루할틈이없음

| khaiii | 한나눔 | 꼬꼬마 | KOMORAN | OKT | mecab |
| --- | --- | --- | --- | --- | --- |
| 굉/MAG | 굉장히잘만든수작지루할틈이없음/N | 굉장히/MAG | 굉장히/MAG | 굉장히/Adjective | 굉장히/MAG |
| 장히/VV |  | 잘/MAG | 잘/MAG | 잘만/Noun | 잘/MAG |
| 잘/NNP |  | 만들/VV | 만들/VV | 든/Josa | 만든/VV+ETM |
| 만/VV |  | ㄴ/ETD | ㄴ/ETM | 수작/Noun | 수작/NNG |
| 든/XSA |  | 수작/NNG | 수작/NNG | 지루할/Adjective | 지루/XR |
| 수작지루/NNG |  | 지루/XR | 지루/XR | 틈/Noun | 할/XSA+ETM |
| 할틈이/NNG |  | 하/XSA | 하/XSA | 이/Josa | 틈/NNG |
| 없/VA |  | ㄹ/ETD | ㄹ/ETM | 없음/Adjective | 이/JKS |
| 음/ETN |  | 틈/NNG | 틈/NNG |  | 없/VA |
|  |  | 이/JKS | 이/JKS |  | 음/ETN |
|  |  | 없/VA | 없/VA |  |  |
|  |  | 음/ETN | 음/ETN |  |  |

역시나 띄어쓰기가 없는 문장은 분석이 쉽지 않은 것으로 보입니다. khaiii나 한나눔 분석기는 서로 다른 결과(너무 잘게 분석하거나 아예 분석을 못하거나)로 나쁜 분석 품질을 보여준다. 해당 분석기를 사용할 계획이 있다면 앞에서 띄어쓰기 전처리를 해줘야 할 것으로 보입니다. 그래도 꼬꼬마, KOMORAN, mecab 분석기는 생각보다 훌륭한 분석 품질을 보여주며 선전했네요.

자소 분리 및 오탈자가 포함된 문장

실험 문장 1: ㄱㅐOOO같은영화 뭐가무섭다는건지ㅡㅡ

| khaiii | 한나눔 | 꼬꼬마 | KOMORAN | OKT | mecab |
| --- | --- | --- | --- | --- | --- |
| ㄱ/SH | ㄱㅐOOO같은영화/N | ㄱ/NNG | 개/NNB | ㄱㅐ/KoreanParticle | ㄱ/NNG |
| ㅐOOO/SL | 뭐가무섭다는건지ㅡㅡ/N | ㅐ/UN | OOO/NNP | OOO/Alpha | ㅐ/UNKNOWN |
| 같/JKB |  | OOO/OL | 같/VA | 같은/Adjective | OOO/SL |
| 은영화/NNG |  | 같/VA | 은/ETM | 영화/Noun | 같/VA |
| 뭐/XPN |  | 은/ETD | 영화/NNP | 뭐/Noun | 은/ETM |
| 가무/VV |  | 영화/NNG | 뭐가무섭다는건지ㅡㅡ/NA | 가/Josa | 영화/NNG |
| 섭/VA |  | 뭐/NP |  | 무섭다는/Adjective | 뭐/NP |
| 다/EF |  | 가/JKS |  | 건지/Verb | 가/JKS |
| 는/ETM |  | 무섭/VA |  | ㅡㅡ/KoreanParticle | 무섭/VA |
| 건지/NNG |  | 다는/ETD |  |  | 다는/ETM |
| ㅡ/MAG |  | 것/NNB |  |  | 건지/NNB+VCP+EC |
| ㅡ/NNG |  | 이/VCP |  |  | ㅡㅡ/UNKNOWN |
|  |  | ㄴ지/ECS |  |  |  |
|  |  | ㅡㅡ/EMO |  |  |  |

실험 문장 2: ㄴㅓ무합니다이무슨..유치찬란..오글거려못보겠네요

| khaiii | 한나눔 | 꼬꼬마 | KOMORAN | OKT | mecab |
| --- | --- | --- | --- | --- | --- |
| ㄴㅓ무/NNG | ㄴㅓ무합니다이무슨/N | ㄴㅓ무합니다이무슨/UN | 너무/MAG | ㄴㅓ/KoreanParticle | ㄴ/NNG |
| 하/XSA | ../S | ../SW | 하/XSV | 무/Noun | ㅓ무합니다이무슨/UNKNOWN |
| ㅂ/EC | 유치찬란/N | 유치찬란/XR | ㅂ니다/EC | 합니다/Verb | ./SF |
| 니다이무슨/NNP | ../S | ../SW | 이/NNB | 이/Determiner | ./SY |
| ./SF | 오글거려못보/N | 오글거리/VV | 무스/NNP | 무슨/Noun | 유치찬란/XR |
| ./SP | 이/J | 어/ECS | ㄴ/JX | ../Punctuation | ./SF |
| 유치/NNG | 겠네요/E | 못/MAG | ../SE | 유치/Noun | ./SY |
| 찬/NNP |  | 보/VV | 유치/NNP | 찬란/Noun | 오/NR |
| 란/NNG |  | 겠/EPT | 찬란/XR | ../Punctuation | 글/NNG |
| ./SE |  | 네요/EFN | ../SE | 오글거려못/Verb | 거려/XSV+EC |
| ./SP |  |  | 오/NNP | 보겠네요/Verb | 못/MAG |
| 오글거리/VV |  |  | 글/NNG |  | 보/VV |
| 어/EC |  |  | 걸/VV |  | 겠/EP |
| 못보/VV |  |  | 려/EC |  | 네요/EC |
| 겠/EP |  |  | 못/MAG |  |  |
| 네요/EC |  |  | 보/VV |  |  |
|  |  |  | 겠/EP |  |  |
|  |  |  | 네요/EC |  |  |

KOMORAN만 유일하게 자소가 분리된 문장이나 오탈자에 대해서도 어느 정도 강건함을 보여주고 있습니다. 확실하진 않지만 추측컨데, 다른 형태소 분석기는 최소 단위를 음절로 보고 KOMORAN은 자소로 보고 학습을 한 것으로 보이네요.

긴 문장

실험 문장: 개봉했을때부터 지금까지 마음이답답하거나 힘들때 이영화 보고있어요 그때마다 심적인 위로를 받을수있는영화같아요 장면 하나하나가 너무예쁘고 마음에 남아서 진한 여운까지 주는영화 감사합니다

| khaiii | 한나눔 | 꼬꼬마 |  | KOMORAN | OKT | mecab |
| --- | --- | --- | --- | --- | --- | --- |
| 개봉/NNG | 개봉/N | 개봉/NNG |  | 개봉/NNG | 개봉/Noun | 개봉/NNG |
| 하/XSV | 하/X | 하/XSV |  | 하/XSV | 했을/Verb | 했/XSV+EP |
| 였/EP | 었을/E | 었/EPT |  | 았/EP | 때/Noun | 을/ETM |
| 을/ETM | 때/N | 을/ETD |  | 을/ETM | 부터/Josa | 때/NNG |
| 때/NNG | 부터/J | 때/NNG |  | 때/NNG | 지금/Noun | 부터/JX |
| 부터/JX | 지금/M | 부터/JX |  | 부터/JX | 까지/Josa | 지금/NNG |
| 지금/NNG | 까지/J | 지금/NNG |  | 지금/NNG | 마음/Noun | 까지/JX |
| 까지/JX | 마음이답답하거/N | 까지/JX |  | 까지/JX | 이/Josa | 마음/NNG |
| 마음/NNG | 나/J | 마음/NNG |  | 마음/NNG | 답답하거나/Adjective | 이/JKS |
| 이/VCP | 힘들/P | 이/JKS |  | 이/JKS | 힘들/Adjective | 답답/XR |
| 답답/NNG | ㄹ/E | 답답/XR |  | 답답/XR | 때/Noun | 하/XSA |
| 하/XSA | 때/N | 하/XSA |  | 하/XSA | 이영화/Noun | 거나/EC |
| 거나/EC | 이영화/N | 거나/ECE |  | 거나/EC | 보고있어요/Verb | 힘들/VA+ETM |
| 힘들/VA | 보/P | 힘들/VV |  | 힘/NNG | 그때/Noun | 때/NNG |
| ㄹ/ETM | 고/E | ㄹ/ETD |  | 들/XSN | 마다/Josa | 이/MM |
| 때/NNG | 있/P | 때/NNG |  | 때/NNG | 심/Noun | 영화/NNG |
| 이영화/NNP | 어/E | 이/MDT |  | 이영화/NNP | 적/Suffix | 보/VV |
| 보/VV | 요/J | 영화/NNG |  | 보/VV | 인/Josa | 고/EC |
| 고/EC | 그때/N | 보/VV |  | 고/EC | 위로/Noun | 있/VX |
| 있/VX | 마다/J | 고/ECE |  | 있/VX | 를/Josa | 어요/EF |
| 어요/EC | 심/N | 있/VXV |  | 어요/EC | 받을수있는/Verb | 그때/NNG |
| 그때/NNG | 적/X | 어요/EFN |  | 그때/NNG | 영화/Noun | 마다/JX |
| 마다/JX | 이/J | 그때/NNG |  | 마다/JX | 같아요/Adjective | 심/NNG |
| 심/NNG | ㄴ/E | 마다/JX |  | 심/NNG | 장면/Noun | 적/XSN |
| 적/XSN | 위로/N | 심적/NNG |  | 적/XSN | 하나/Noun | 인/VCP+ETM |
| 이/VCP | 를/J | 이/VCP |  | 이/VCP | 하나/Noun | 위로/NNG |
| ㄴ/ETM | 받을수있는영화같아요/N | ㄴ/ETD |  | ㄴ/ETM | 가/Josa | 를/JKO |
| 위로/NNG | 장면/N | 위로/NNG |  | 위로/NNG | 너무/Adverb | 받/VV |
| 를/JKO | 하나하나/N | 를/JKO |  | 를/JKO | 예쁘고/Adjective | 을/ETM |
| 받/VV | 가/J | 받/VV |  | 받/VV | 마음/Noun | 수/NNB |
| 을/ETM | 너무예쁘/N | 을/ETD |  | 을/ETM | 에/Josa | 있/VV |
| 수/NNB | 이/J | 수/NNB |  | 수/NNB | 남아/Noun | 는/ETM |
| 있/VV | 고/E | 있/VV |  | 있/VX | 서/Josa | 영화/NNG |
| 는/ETM | 마음/N | 는/ETD |  | 는/ETM | 진한/Adjective | 같/VA |
| 영화/NNG | 에/J | 영화/NNG |  | 영화/NNP | 여운/Noun | 아요/EF |
| 같/VA | 남/P | 같/VA |  | 같/VA | 까지/Josa | 장면/NNG |
| 아요/EC | 아/E | 아요/EFN |  | 아요/EC | 주는/Verb | 하나하나/NNG |
| 장면/NNG | 지/P | 장면/NNG |  | 장면/NNG | 영화/Noun | 가/JKS |
| 하나하나/NNG | ㄴ/E | 하나하나/NNG |  | 하나하나/NNG | 감사합니다/Verb | 너무/MAG |
| 가/JKS | 한/N | 가/JKS |  | 가/JKS |  | 예쁘/VA |
| 너무/MAG | 여운/N | 너무/MAG |  | 너무/MAG |  | 고/EC |
| 예/VV | 까지/J | 예쁘/VA |  | 예쁘/VA |  | 마음/NNG |
| 쁘/VA | 주는영화/N | 고/ECE |  | 고/EC |  | 에/JKB |
| 고/EC | 감사/N | 마음/NNG |  | 마음/NNG |  | 남/VV |
| 마음/NNG | 하/X | 에/JKM |  | 에/JKB |  | 아서/EC |
| 에/JKB | ㅂ니다/E | 남/VV |  | 남/VV |  | 진한/VA+ETM |
| 남/VV |  | 아서/ECD |  | 아서/EC |  | 여운/NNG |
| 아서/EC |  | 진하/VA |  | 진한/NNP |  | 까지/JX |
| 진하/VA |  | ㄴ/ETD |  | 여운/NNP |  | 주/VX |
| ㄴ/ETM |  | 여운/NNG |  | 까지/JX |  | 는/ETM |
| 여운/NNG |  | 까/VV |  | 주/VX |  | 영화/NNG |
| 까지/JX |  | 지/ECD |  | 는/ETM |  | 감사/NNG |
| 주는영화/NNG |  | 주/VXV |  | 영화/NNP |  | 합니다/XSV+EC |
| 감사/NNG |  | 는/ETD |  | 감사/NNG |  |  |
| 하/XSV |  | 영화/NNG |  | 하/XSV |  |  |
| ㅂ니다/EC |  | 감사/NNG |  | ㅂ니다/EC |  |  |
|  |  | 하/XSV |  |  |  |  |
|  |  | ㅂ니다/EFN |  |  |  |  |

긴 문장이지만 띄어쓰기가 어느 정도 되어있는 문장에 대해서 테스트 결과입니다. 조금씩의 차이는 있지만 대부분의 분석기들이 잘 분석해주고 있는 것을 볼 수 있습니다. 따라서 문장의 길이는 띄어쓰기 유무와 자소/오탈자와 달리 형태소 분석기의 분석 품질에 큰 영향을 주지 않는 것으로 보입니다.

결론

* 사용할 데이터의 특성(띄어쓰기 유무 등)이나 개발 환경(Python, Java)에 따라서 적합한 형태소 분석기를 고려해야함
* 연산 속도가 중요하다면 mecab을 최우선으로 고려해야하며, 심지어 분석 품질도 상위권으로 보여짐
* 자소 분리나 오탈자에 대해서도 어느 정도 분석 품질이 보장되야 한다면 KOMORAN 사용을 고려
* 한나눔과 khaiii는 일부 케이스에 대한 분석 품질, 꼬꼬마는 분석 시간에서 약간 아쉬운 점이 보임