한국어 임베딩

3-2~3-5단원

전찬웅

지도 학습 기반 VS 비지도 학습 기반

- 지도 학습 기반의 형태소 분석
- 언어학 전문가들이 태깅한 분석 말뭉치로부터 학습된 기법
- Ex) KoNLPy, Khaiii

- 비지도 학습 기반 형태소 분석
- 데이터 패턴을 모델 스스로 학습하게 함으로써 형태소를 나누는 기법
- Ex) soynlp, 구글 센텐스피스

1. 지도 학습 기반 형태소 분석

KoNLPy 란?

표 3-3 **형태소 분석 품질 비교**(출처: KoNLPy)

Hannanum	Kkma	Komoran	Mecab	Okt
아버지가방에들어가/N	아버지/NNG	아버지가방에들어가신다/NNP	아버지/NNG	아버지/Noun
이/J	가방/NNG		가/JKS	가방/Noun
시ㄴ다/E	에/JKM		방/NNG	에/Josa
	들어가/VV		에/JKB	들어가신/Verb
	시/EPH		들어가/VV	다/Eomi
	ㄴ다/EFN		신다/EP+EC	

- KoNLPy란 은전한닢, 꼬꼬마, 한나눔 Okt, 코모란 등 5개 오픈소스 형태소 분석기를 파이썬에서 사용할수 있도록 한 한국어 자연어 처리 패키지다.
- 자세한 내용 https://heung-bae-lee.github.io/2020/01/19/NLP_02/

KoNLPy 내 성능 분석

표 3-2 **KoNLPy 내 분석기별 속도**(단위: 초)

분석기명	로딩 시간	실행 시간
Kkma	5.6988	35.7163
Komoran	5.4866	25.6008
Hannanum	0.6591	8.8251
Okt(Twitter)	1.4870	2.4714
Mecab	0.0007	0.2838

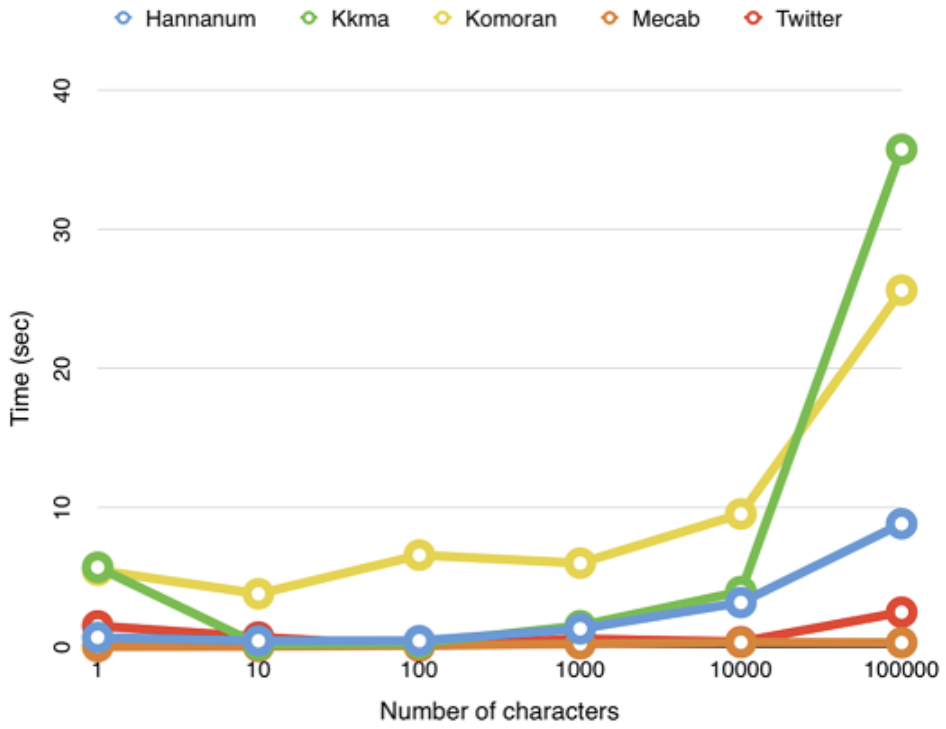


그림 3-11 문자 개수 대비 실행 시간(출처: KoNLPy)

Khaiii 란?

- 카카오가 공개한 오프소스 한국어 형태소 분석기이다.
- CNN 모델 적용해 학습했으며, 입력문장을 문자 단위로 읽어 컨볼루션 필터가 이 문자들을 슬라이딩해 가면서 정보를 추출한다.

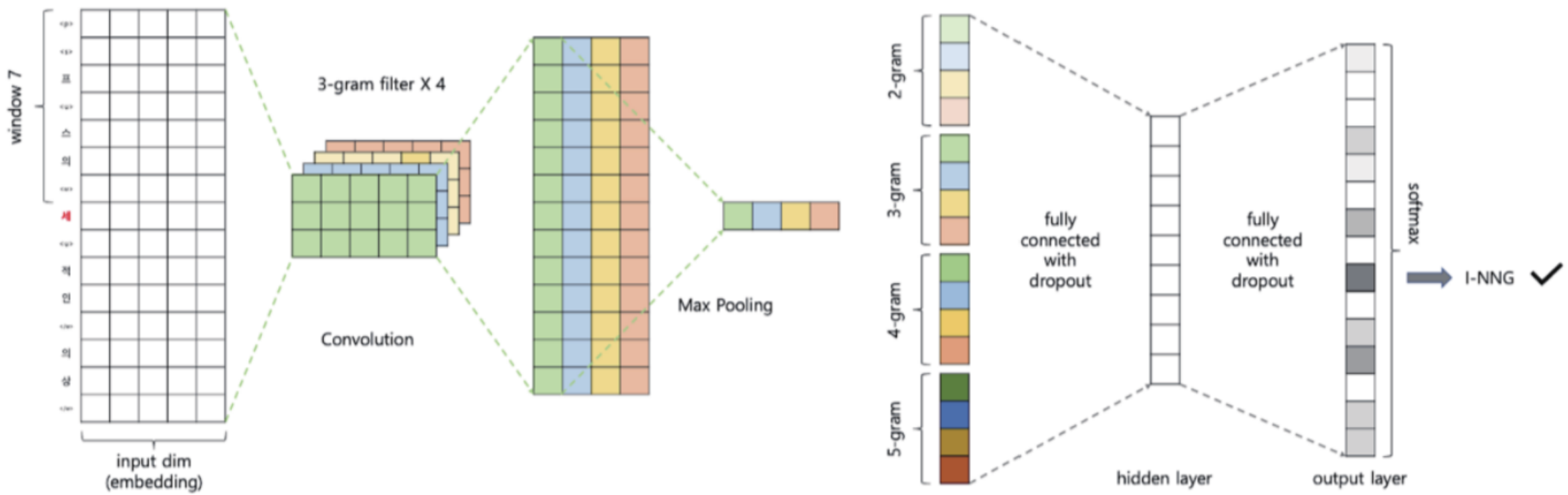


그림 3-12 Khaiii 아키텍처

사전 추가를 해야 하는 이유

“가우스전자 " 사전 추가 전

'가우스전자 텔레비전 정말 좋네요 '

['가우스', '전자', '텔레비전', '정말', '좋', '네요']

“가우스전자 " 사전 추가 후

'가우스전자 텔레비전 정말 좋네요'

['가우스전자', '텔레비전', '정말', '좋', '네요']

사전에 “가우스전자”를 추가함으로써 분석의 정확도를 더 상승시킬 수 있다.

2. 비지도 학습 기반 형태소 분석

soynlp 형태소 분석기

- Soynlp는 형태소 분석, 품사 판별 등을 지원하는 한국어 자연어처리 패키지다.
- 하나의 문서 또는 문장(X)
- 어느정도 규모가 있으면서 동질적인 문서 집합(O)
- 데이터의 통계량을 확인해 만든 단어 점수 표로 작동한다.
- Ltokenizer 클래스를 통하여 입력문장의 왼쪽 부터 문자 단위로 슬라이딩해 가면서 단어 점수가 높은 문자열을 우선으로 형태소로 취급해 분리한다.
- 띄어쓰기 교정 모듈 제공

단어 점수란?

- 단어 점수는 크게 응집 확률과 브랜칭 엔트로피를 활용한다.

- 응집확률 : 주어진 문자열이 유기적으로 연결되어 얼마나 자주 나타나는지에 대한 정도
- 브랜칭 엔트로피 : 특정 단어 앞뒤로 다양한 조사, 어미 혹은 다른 단어가 등장하는 정도

단어점수가 클 때(응집 확률이 높고, 브랜칭 엔트로피가 높을때),
해당 문자열을 형태소로 취급한다

구글 센텐스피스

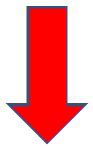
- 구글에서 공개한 비지도 학습 기반 형태소
- 바이트 페어 인코딩(BPE) 기법을 지원한다.

바이트 페어 인코딩(BPE)-학습

aaabdaaabac

aa를 Z로 치환

ZabdZabac



ab를 Y로 치환

ZYdZYac

- 원하는 어휘 집합 크기가 될 때까지 반복적으로 고빈도 문자열들을 병합해 어휘 집합에 추가한다.

바이트 페어 인코딩(BPE)-예측

학교에서 밥을 먹었다

_학교에서, _밥을, _먹었다

_학교, 에서, _밥, 을, _먹었, 다

1. 문장 내 각 어절에 어휘 집합 에 있는 서브워드가 포함돼 있을 경우 해당 서브워드를 어절에서 분리한다
2. 어절의 나머지에서 어휘집합에 있는 서브워드를 다시 찾고,또 분리한다
3. 어절 끝까지 찾았는데 어휘 집합에 없으면 미등록 단어로 취급한다

띄어쓰기 교정

- 말뭉치에서 띄어쓰기 패턴을 학습한 뒤 해당 패턴대로 교정을 수행한다.
- 학습데이터에서 특정 단어 앞뒤로 공백이 있는 것을 다수 발견 시 예측 단어에 그 특정 단어가 나오면 앞뒤를 띄어서 교정하는 방식
- ex) 학습 데이터에 '때' 뒤가 다수 띄어쓰기 되어 있다면,

"어릴때보고 처음 봅니다"

"어릴때 보고 처음 봅니다"