ICT이노베이션스퀘어 AI복합교육 고급 언어과정

자연어처리를 위한 한국어 전처리 패키기

현청천

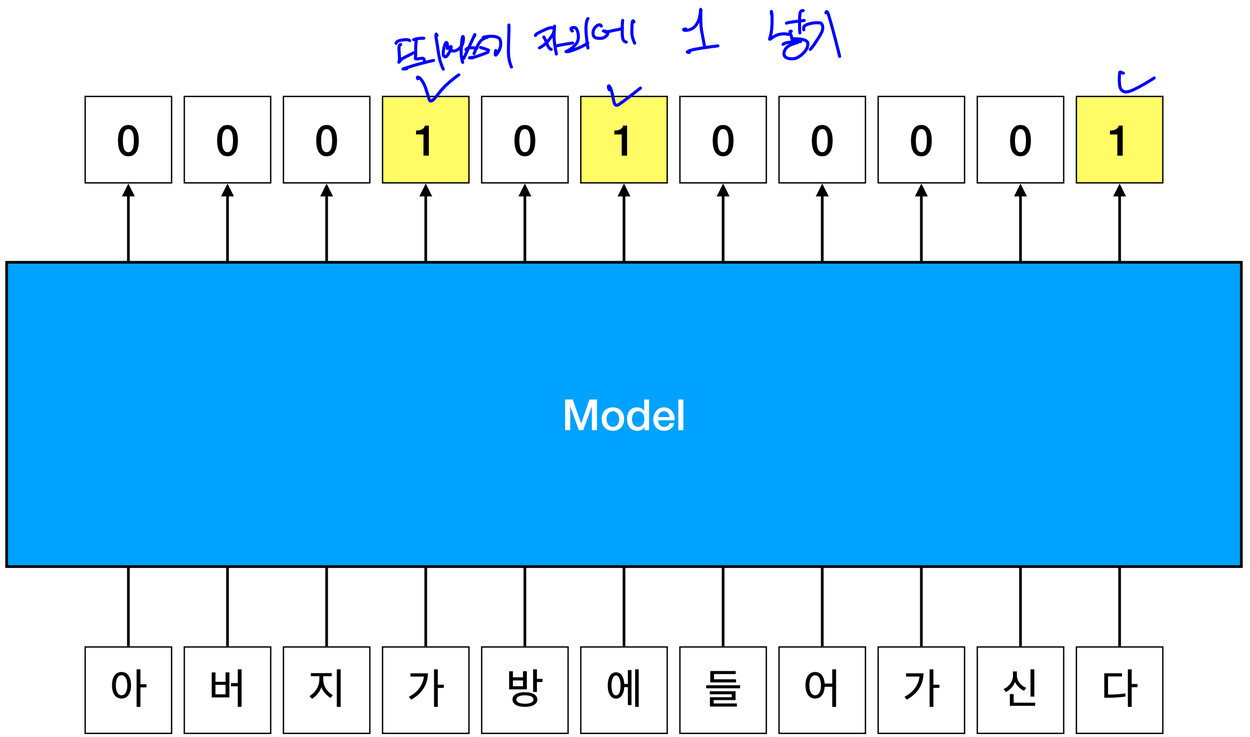
2021.04.19

띄어쓰기

- 띄어쓰기는 적당한 의미 단위로 구분하는데 큰 영향을 줌
  - 아버지가 방에 들어가신다 : [‘아버지가’, ‘방에’, ‘들어가신다’]
  - 아버지 가방에 들어가신다 : [‘아버지’, ‘가방에’ ‘들어가신다’]

- 인터넷에서 수집된 데이터에는 많은 오류가 있음
- 띄어쓰기에 따라서 딥러닝 모델이 다른 뜻으로 이해할 수 있음

띄어쓰기



띄어쓰기

Soyspacing

https://github.com/lovit/soyspacing

띄어쓰기 문제를 해결하기 위한 휴리스틱 알고리즘

띄어쓰기

Pycrfsuite

https://github.com/lovit/pycrfsuite_spacing

CRF를 이용한 띄어쓰기 알고리즘

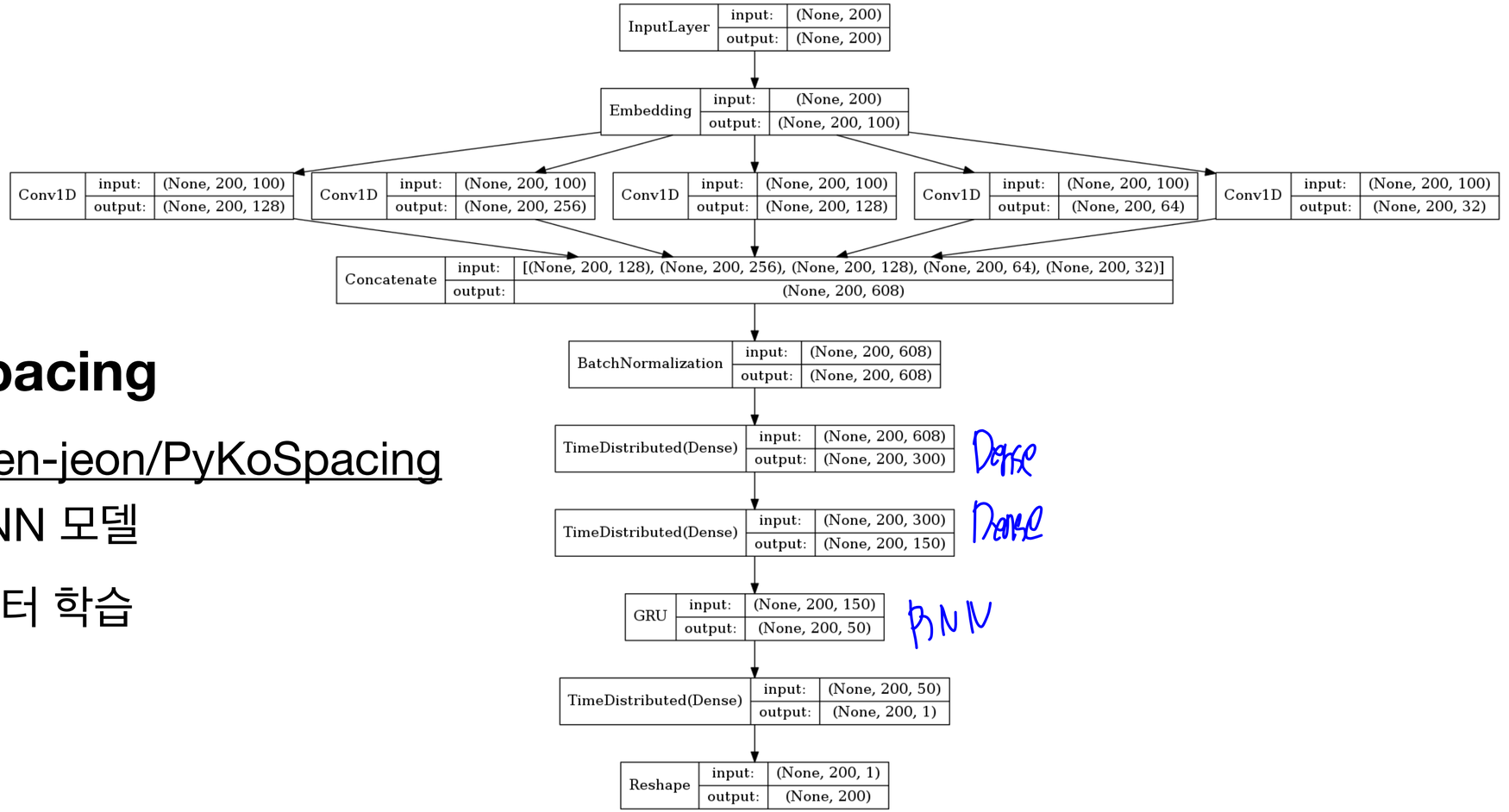
띄어쓰기

PyKoSpacing

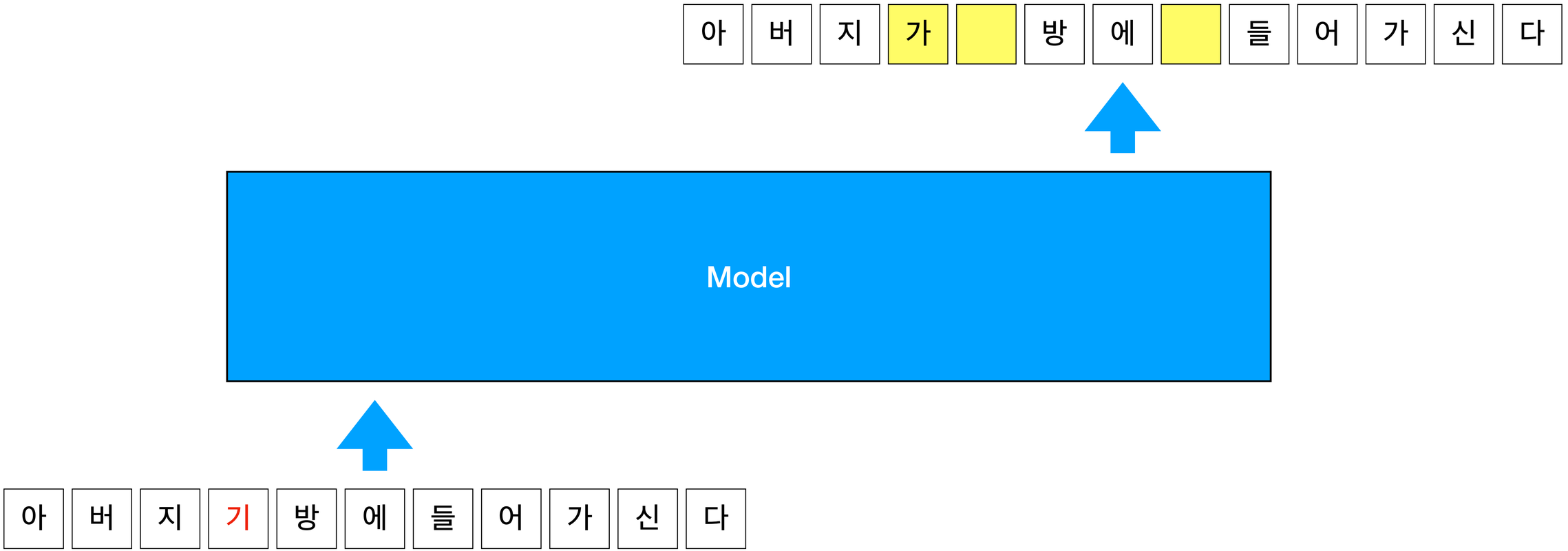
https://github.com/haven-jeon/PyKoSpacing

CNN, RNN 모델

뉴스 데이터 학습



맞춤법 검사



맞춤법 검사

Py-hanspell

https://github.com/ssut/py-hanspell.git + 네이버연동

네이버 맞춤법 검사기를 이용한 파이썬용 한글 맞춤법 검사 라이브러리

형태소분석기

KoNLPy

https://github.com/konlpy/konlpy

Hannanum, Kkma, Komoran, Mecab, Okt 등 다양한 형태소분석기 제공

Soynlp

SoyNLP

https://github.com/lovit/soynlp

한국어 처리를 위한 패키지

SoynIp (LRNounExtractor)

L+[R] : 조사를 보고 명사여부를 판단

내서	-0.530702
있게	1.000000
있는	0.327824
쓰는	0.079298
었다면	-1.000000
였다면	0.437399

- 재미있게 3번
- 재미있는 2번

$$\frac{(3 \times 1.0 + 2 \times 0.33)}{5} = 0.732$$

Threshold 보다 크면 명사

Soynlp (WordExtractor)

학습을 통해 단어추출 (미등록 단어문제)

Soynlp (WordExtractor-cohesion)

노 (200)
노래 (100)
노란 (100)
노래가 (50)
노래는 (30)
노래를 (20)
노란색 (100)
노란색을 (10)

빈도수

P(노래|노) = 0.5
P(노란|노) = 0.5
P(노란색|노란) = 1
P(노란색을|노란색) = 0.1

확률분포

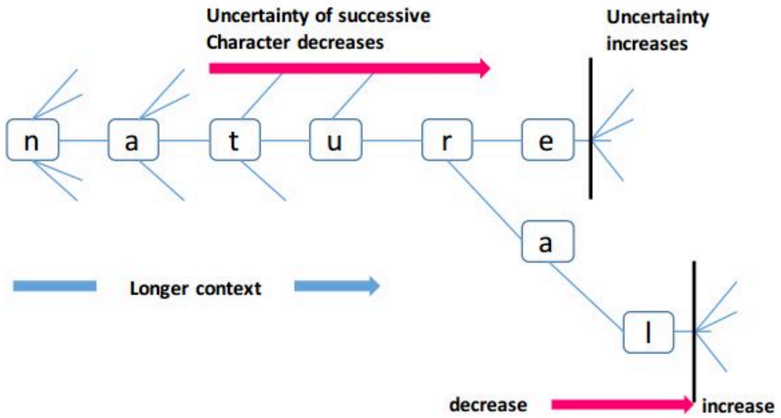
o 노란에서 노란색이 나올 확률

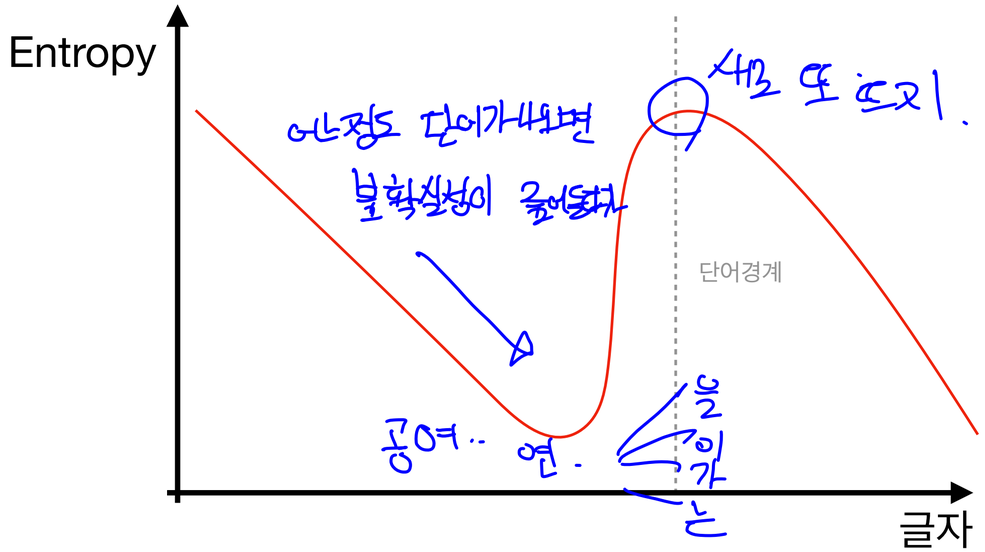
$$cohesion(c_{0:n}) = \left(\prod P(c_{0:i+1} \mid c_{0:i}) \right)^{\frac{1}{n-1}}$$

$$cohesion(\text{노란색}) = \left(P(\text{노란} \mid \text{노}) \times P(\text{노란색} \mid \text{노란}) \right)^{\frac{1}{2}} = (0.5 \times 1)^{\frac{1}{2}} = 0.707$$

SoynIp (WordExtractor-branching entropy)

$$H(X\,|X_n) = \sum_{x \in X} P(x\,|x_n) * \log(\,P(x\,|x_n)\,)$$





공연은 : 30
공연을 : 20
공연이 : 50

$$entropy(\text{공연}) = -\,(0.3 * \log(0.3) + 0.2 * \log(0.2) + 0.5 * \log(0.5)) = 1.03$$

손나은 : 98
손나응 : 1
손나으 : 1

$$entropy(\text{손나}) = -\,(0.98 * \log(0.98) + 0.01 * \log(0.01) + 0.01 * \log(0.01)) = 0.11$$

Soynlp (WordExtractor-Accessor Variety)

공연이 왼쪽 3개

공연을 : 30
공연을 : 20
공연이 : 50

$av_l(\text{공연}) = 3$

공연 오른쪽

이번공연 : 30
저번공연 : 20
올해공연 : 50

$av_r(\text{공연}) = 2$

$$AV(\textbf{공연}) = min\left(av_r(\textbf{공연}), av_l(\textbf{공연})\right)$$

불확실성을 단어 경계 다음에 등장한 글자의 종류

Soynlp (Tokenizer)

문장을 단어의 경계에 따라 단어단위로 분해

Soynlp (Tokenizer-LTokenizer)

L(명사/동사/형용사/부사)과 R(기타)로 분해

띄어쓰기가 잘 되어 있는 문장

Soynlp (Tokenizer-MaxScoreTokenizer)

Score를 이용해 분해

띄어쓰기가 안 되어 있는 문장

Soynlp (Tokenizer-RegexTokenizer)

규칙을 이용해 분해 (정규식)

단어의 형태가 바뀌는 경우

Soynlp (Part of Speech Tagger)

사전기반 품사 판별기

```
pos_dict = {
    'Adverb': {'너무', '매우'},
    'Noun': {'너무너무너무', '아이오아이', '아이', '노래', '오', '이', '고양'},
    'Josa': {'는', '의', '이다', '입니다', '이', '이는', '를', '라', '라는'},
    'Verb': {'하는', '하다', '하고'},
    'Adjective': {'예쁜', '예쁘다'},
    'Exclamation': {'우와'}
}
```

감사합니다.