# Tokenization

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

# Two Steps

1. Sentence Segmentation

2. Tokenization

# Sentence Segmentation

- 보통 훈련 시 우리가 원하는 입력 데이터는
  - 1 sentence/line

- 우리가 수집한 corpus는
  - 한 라인에 여러 문장이 들어있거나,
  - 한 문장이 여러 라인에 들어있음

- Sentence Segmentation을 통해 원하는 형태로 변환
  - 마침표 등을 단순히 문장의 끝으로 처리하면 안됨!
    - e.g. 3.141592, U.S.

- NLTK를 활용하여 변환 가능
  - `from nltk.tokenize import sent_tokenize`

# Multiple sentence/line

현재 TED 웹사이트에는 1,000개가 넘는 TED강연들이 있습니다. 여기 계신 여러분의 대다수는 정말 대단한 일이라고 생각하시겠죠 -- 전 다릅니다. 전 그렇게 생각하지 않아요. 저는 여기 한 가지 문제점이 있다고 생각합니다. 왜냐하면 강연이 1,000개라는 것은, 공유할 만한 아이디어들이 1,000개 이상이라는 뜻이 되기 때문이죠. 도대체 무슨 수로 1,000개나 되는 아이디어를 널리 알릴 건가요?

```
현재 TED 웹사이트에는 1,000개가 넘는 TED강연들이 있습니다.
여기 계신 여러분의 대다수는 정말 대단한 일이라고 생각하시겠죠 -- 전 다릅니다.
전 그렇게 생각하지 않아요.
저는 여기 한 가지 문제점이 있다고 생각합니다.
왜냐하면 강연이 1,000개라는 것은, 공유할 만한 아이디어들이 1,000개 이상이라는 뜻이 되기 때문이죠.
도대체 무슨 수로 1,000개나 되는 아이디어를 널리 알릴 건가요?
```

# Multiple line/sentence

```
현재 TED 웹사이트에는 1,000개가 넘는 TED강연들이 있습니다.
여기 계신 여러분의 대다수는
정말 대단한 일이라고 생각하시겠죠 --
전 다릅니다. 전 그렇게 생각하지 않아요.
저는 여기 한 가지 문제점이 있다고 생각합니다.
왜냐하면 강연이 1,000개라는 것은,
공유할 만한 아이디어들이 1,000개 이상이라는 뜻이 되기 때문이죠.
도대체 무슨 수로
1,000개나 되는 아이디어를 널리 알릴 건가요?
1,000개의 TED 영상 전부를 보면서
```

```
현재 TED 웹사이트에는 1,000개가 넘는 TED강연들이 있습니다.
여기 계신 여러분의 대다수는 정말 대단한 일이라고 생각하시겠죠 -- 전 다릅니다.
전 그렇게 생각하지 않아요.
저는 여기 한 가지 문제점이 있다고 생각합니다.
왜냐하면 강연이 1,000개라는 것은, 공유할 만한 아이디어들이 1,000개 이상이라는 뜻이 되기 때문이죠.
도대체 무슨 수로 1,000개나 되는 아이디어를 널리 알릴 건가요?
1,000개의 TED 영상 전부를 보면서
```

# Tokenization

- Why?
  - 두 개 이상의 다른 token들의 결합으로 이루어진 단어를 쪼개어, vocabulary 숫자를 줄이고, 희소성(sparseness)을 낮추기 위함

- Example

before:

> North Korea's state mouthpiece, the Rodong Sinmun, is also keeping mum on Kim's summit with Trump while denouncing ever-tougher U.S. sanctions on the rogue state.

after:

> North Korea 's state mouthpiece , the Rodong Sinmun , is also keeping mum on Kim 's summit with Trump while denouncing ever-tougher U.S. sanctions on the rogue state .

# Korean Tokenization

- Why?
  - 교착어: 어근에 접사가 붙어 다양한 단어가 파생됨
  - 띄어쓰기 통일의 필요성

# Tokenization for Other Languages

- 영어: 띄어쓰기가 이미 잘 되어 있음. NLTK를 사용하여 comma 등 후처리
- 중국어: 기본적인 띄어쓰기가 없음. Character 단위로 사용해도 무방
- 일본어: 기본적인 띄어쓰기가 없음.

# 형태소 분석 및 품사 태깅 (Part of Speech Tagging)

- 형태소 분석: 형태소를 비롯하여, 어근, 접두사/접미사, 품사(POS, part-of-speech) 등 다양한 언어적 속성의 구조를 파악하는 것

- 품사 태깅: 형태소의 뜻과 문맥을 고려하여 그것에 마크업을 하는 일

출처: https://konlpy-ko.readthedocs.io/ko/v0.4.3/morph/

# POS Tagger for Other Languages

| 언어 | 프로그램명 | 제작언어 | 특징 |
|---|---|---|---|
| 한국어 | Mecab | C++ | 일본어 Mecab을 wrapping. 속도가 가장 빠름 |
| 한국어 | KoNLPy | 복합 | 설치와 사용이 편리하나, 일부 tagger의 경우 속도가 느림 |
| 일본어 | Mecab | C++ | 속도가 가장 빠름 |
| 중국어 | Stanford Parser | Java | 미국 스탠포드에서 개발 |
| 중국어 | PKU Parser | Java | 북경대학교에서 개발 |
| 중국어 | Jieba | Python | 가장 최근에 개발. Python으로 제작되어 시스템 구성에 용이 |

# 품사 태깅 예제 (feat. Mecab)

- 아버지가 방에 들어가신다.
  - 아버지 NNG
  - 가 JKS
  - 방 NNG
  - 에 JKB
  - 들어가 VV
  - 신다 EP+EF
  - . SF
  - <EOS>
- 아버지 가방에 들어가신다.
  - 아버지 NNG
  - 가방 NNG
  - 에 JKB
  - 들어가 VV
  - 신다 EP+EF
  - . SF
  - <EOS>

| 태그 | 설명 | 태그 | 설명 |
|---|---|---|---|
| NNG | 일반 명사 | EP | 선어말 어미 |
| NNP | 고유 명사 | EF | 종결 어미 |
| NNB | 의존 명사 | EC | 연결 어미 |
| NNBC | 단위를 나타내는 명사 | ETN | 명사형 전성 어미 |
| NR | 수사 | ETM | 관형형 전성 어미 |
| NP | 대명사 | XPN | 체언 접두사 |
| VV | 동사 | XSN | 명사 파생 접미사 |
| VA | 형용사 | XSV | 동사 파생 접미사 |
| VX | 보조 용언 | XSA | 형용사 파생 접미사 |
| VCP | 긍정 지정사 | XR | 어근 |
| VCN | 부정 지정사 | SF | 마침표, 물음표, 느낌표 |
| MM | 관형사 | SE | 줄임표 … |
| MAG | 일반 부사 | SSO | 여는 괄호 (, [ |
| MAJ | 접속 부사 | SSC | 닫는 괄호 ), ] |
| IC | 감탄사 | SC | 구분자 , · / : |
| JKS | 주격 조사 | SY | |
| JKC | 보격 조사 | SL | 외국어 |
| JKG | 관형격 조사 | SH | 한자 |
| JKO | 목적격 조사 | SN | 숫자 |
| JKB | 부사격 조사 | | |
| JKV | 호격 조사 | | |
| JKQ | 인용격 조사 | | |
| JX | 보조사 | | |
| JC | 접속 조사 | | |

# 분절 예제 (feat. Mecab with –O wakati)

$ echo '아버지가 방에 들어가신다.' | mecab -O wakati
아버지 가 방 에 들어가 신다 .

$ echo '아버지 가방에 들어가신다.' | mecab -O wakati
아버지 가방 에 들어가 신다 .

# Summary

- 한국어의 경우
  1) 접사를 분리하여 희소성을 낮추고,
  2) 띄어쓰기를 통일하기 위해 tokenization을 수행

- 굉장히 많은 POS Tagger가 존재하는데,
  - 전형적인 쉬운 문장(표준 문법을 따르며, 구조가 명확한 문장)의 경우, 성능이 비슷함
  - 하지만 신조어나 고유명사를 처리하는 능력이 다름
  - 따라서, 주어진 문제에 맞는 정책을 가진 tagger를 선택하여 사용해야 함