

Traditional NLP

- 단어 = symbolic 대상

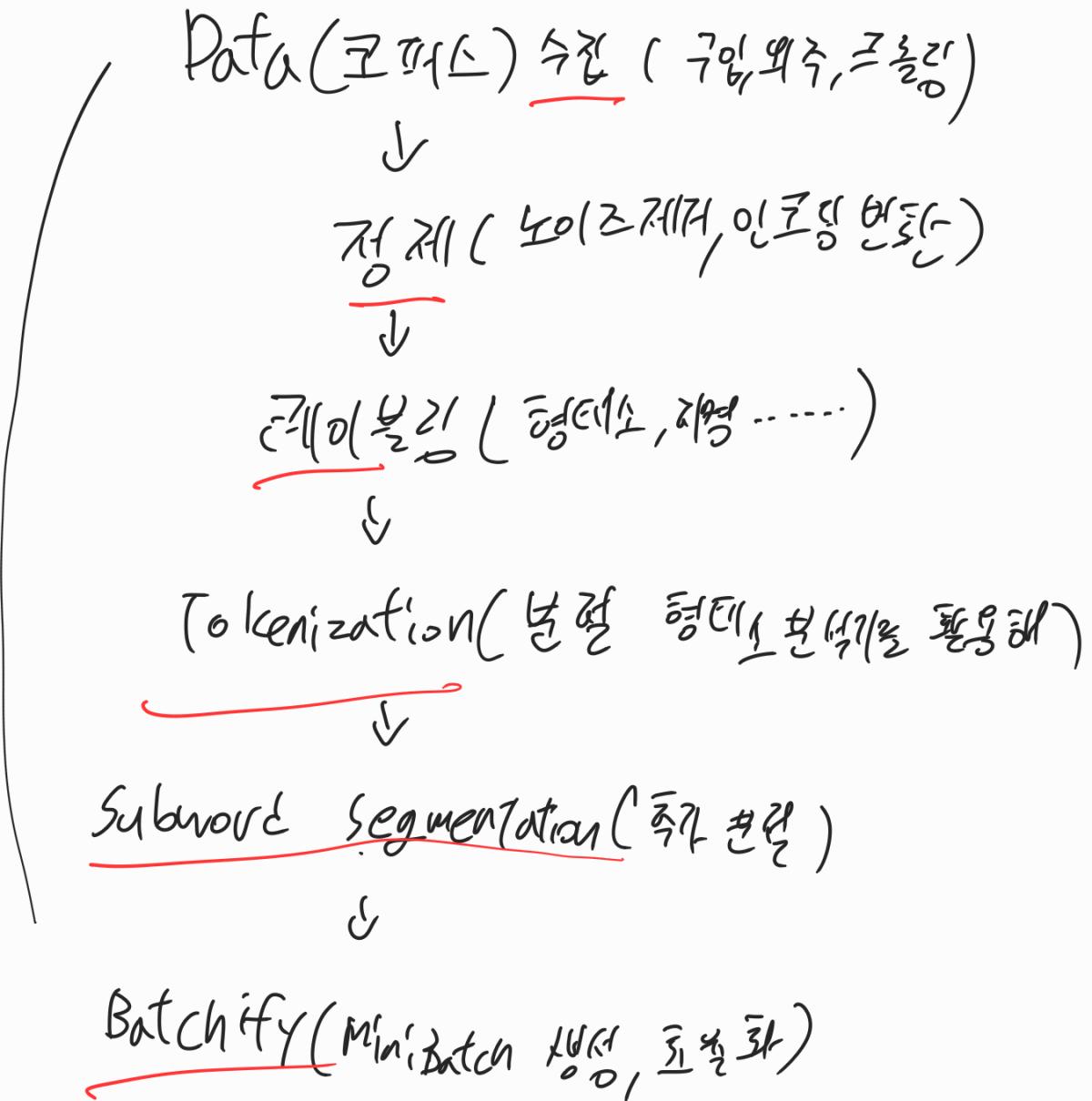
Sympolic

- 의미적
- 사상이 인지하기 쉽다
- C(비)트성이
- 언어는 속도 빠름
- 모든 것에 대해 원-핫 표현 (one-hot representation)

NLP with deep learning

- 연속적
- 사상이 인지하기 어렵다
- C(비)트 형태
- 모든 것에 대해 계속적인 (continuous) 값 (value)

NLP
Preprocess



(corpus)

- NLP에 문장들로 구성된 데이터셋
 - 포함된 언어수로 따라서 Mono / Bi, Multi lingual corpus
- Parallel corpus
 - 대응되는 문장쌍이 같은 디밍 혹은 같은 형태
 - 질문 : 응답

NLP Preprocess 중 Data 수집

- 대(대)터 구조, 외국, 크롤링
 - 국립
 - 양질의 data
 - AI의 매우 제한적
 - 국립처: 대학교, ETRI, 플랫폼...
 - 외국
 - 가장 높은 비중
 - AI의 제한적
 - 품질 관리를 위한 인력 추가로 필요
- 최대 10억 단위

1. 네트워크

- AI-hub
- Kyber
- wmt competition
- OPUS

2. 양의 데이터

3. 한국어 적용

4. Crawling

- INF 양의 CORPUS 수집기능 (영어는 domain 별로 수집 가능)
- 품질이 오락가락, 정제과정 많을 노력 필요

Crawling

- 방적 문제 해결 -> 접두어 및 차이(다른 크롤링)
- robots.txt ex) `http://www.tec.com/robots.txt`
 6
허용범위 (`http://www.robotstxt.org/`)

- 수집자와 나-영한 특성

Domain, 문서, 수집 날짜, 양방향(parrel data), 정제/정오

블로그	Category	대화체	낮음	X	초기화
자신	"/"	"/"	"/"	"	중간
News	News	문자체	"/"	O	낮음
wiki	Q&A	"/"	"/"	O	"/"
구글내비	지도형	대화체	중간	X	높음
Tech	기술형	"/"	낮음	O	낮음
지식	일반	대화체	낮음	O	높음
					6) 목적 주의

Data 2-6-11

전각
P VS P

반각 문자로 표기 가능한 전각문자와
경우 반각문자로 치환

• 7/7(목)인 노년증 제거

- 전각 들지, 번진
 - fas(에) 때는 (전번각인) Noise 제거

Interactive Noise 2/1/21

- 코퍼스의 특성에 따라 노이즈 제거
 - 작동기-기상화물 = 가진 정보 = 학습 데이터

ナミヤ

TaskColl $\approx 1 - \frac{2}{\pi} \approx 0.5$

- 풀고자 하는 문제의 특성에 따라 전처리 단계가 다르다
 - 선증한 진단 패턴

ex) 이모티콘 → 강점분석에 중요 다른 알아는 꼭 알아두고

0.01, 0.001, 0.0001

- 그는 양식, 도시적인 조각-미술을 특성화(特徵化)하는데 다른 전통(傳統) 전략과는

6. 전기(2)

- 전기율자 -> 단위온자 치환

• C(이소용자) 동일(Cpt)

NYC

NYC \Rightarrow new york city

N.Y.C

• 정규식을 활용한 정제

• (복잡한 규칙의 노이즈 제거 / 치환)

RegEx 적용

노이즈 제거

RegEx 구현

반복

217/11

[2345(C)] \rightarrow 2,3,4,5,C 중 하나

[2-5C-C] \rightarrow 2~5 C-C 중 하나

[12-5C-C] \rightarrow 2~5 C~5 예 해당하지 않는가? 1 = Not

(X)(YZ) \rightarrow X를 통해 Y와 Z를 연결하는 것

(e) 양 끝에 알파벳으로 둘러싸인 bcdefg

· abcde

((a-z))bc((a-z))를 통해 wzl \rightarrow ad

(X|Y) \rightarrow X 또는 Y 통해 wzl

X? X가 어떤 or 어떤 경우.

X+ X가 아니면 0이나 1이나 2이나

X* X가 아니거나 많을 때, 단복수로 있을 때

X{n} n번의 반복 ex) {0-9}{4} \rightarrow 4자리 수인 44444444

X{n,3} n번이상 반복

· any character

X \dagger 1=문장의 시작 \dagger =문장의 끝