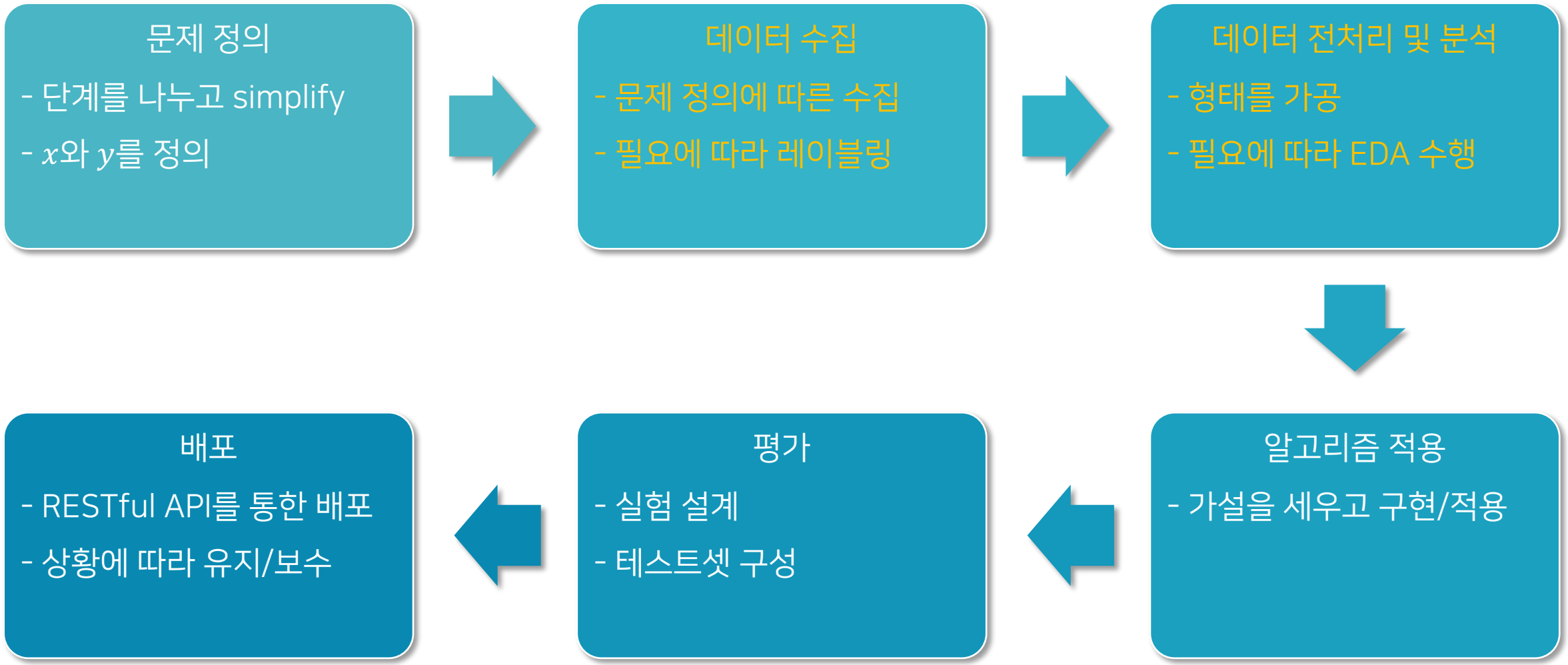
Preprocessing Summary

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

NLP Project Workflow



Preprocessing Workflow

데이터(코퍼스) 수집
- 구입, 외주
- 크롤링을 통한 수집

정제
- Task에 따른 노이즈 제거
- 인코딩 변환

레이블링(Optional)
- 문장마다 또는 단어마다 labeling 수행

Tokenization
- 형태소 분석기를 활용하여 분절 수행

Subword Segmentation (Optional)
- 단어보다 더 작은 의미 단위 추가 분절 수행

Batchify
- 사전 생성 및 word2index 맵핑 수행
- 효율화를 위한 전/후처리

Service Pipeline

정제

- 학습 데이터와 같은 방식의 정제 수행

Tokenization

- 학습 데이터와 같은 방식의 분절 수행

Subword Segmentation

- 학습 데이터로부터 얻은 '모델'을 활용하여 똑같은 분절 수행

Batchify

- 학습 데이터로부터 얻은 사전에 따른 word2index 맵핑 수행

Prediction

- 모델에 넣고 추론 수행
- 필요에 따라 search 수행 (자연어 생성)

Detokenization (Optional)

- 사람이 읽을 수 있는 형태로 변환(index2word)
- 분절 복원

Summary

- 정제
  - Task와 언어 및 도메인에 따른 특성
    - 풀고자 하는 문제의 특성에 따라 전처리 전략이 다름
  - 끝이 없는 과정
    - 노력과 품질 사이의 trade-off
    - Sweet spot을 찾아야함

- 분절
  - 한국어의 경우 띄어쓰기 normalization을 위해 형태소 분석기 활용이 필요
  - Subword segmentation을 통해 좀 더 잘게 분절 할 수 있음

- 모두 비슷한 알고리즘을 사용하고 있으므로, 결국 데이터의 양과 품질이 좌우함
  - 따라서 전처리 과정을 경시해서는 안됨