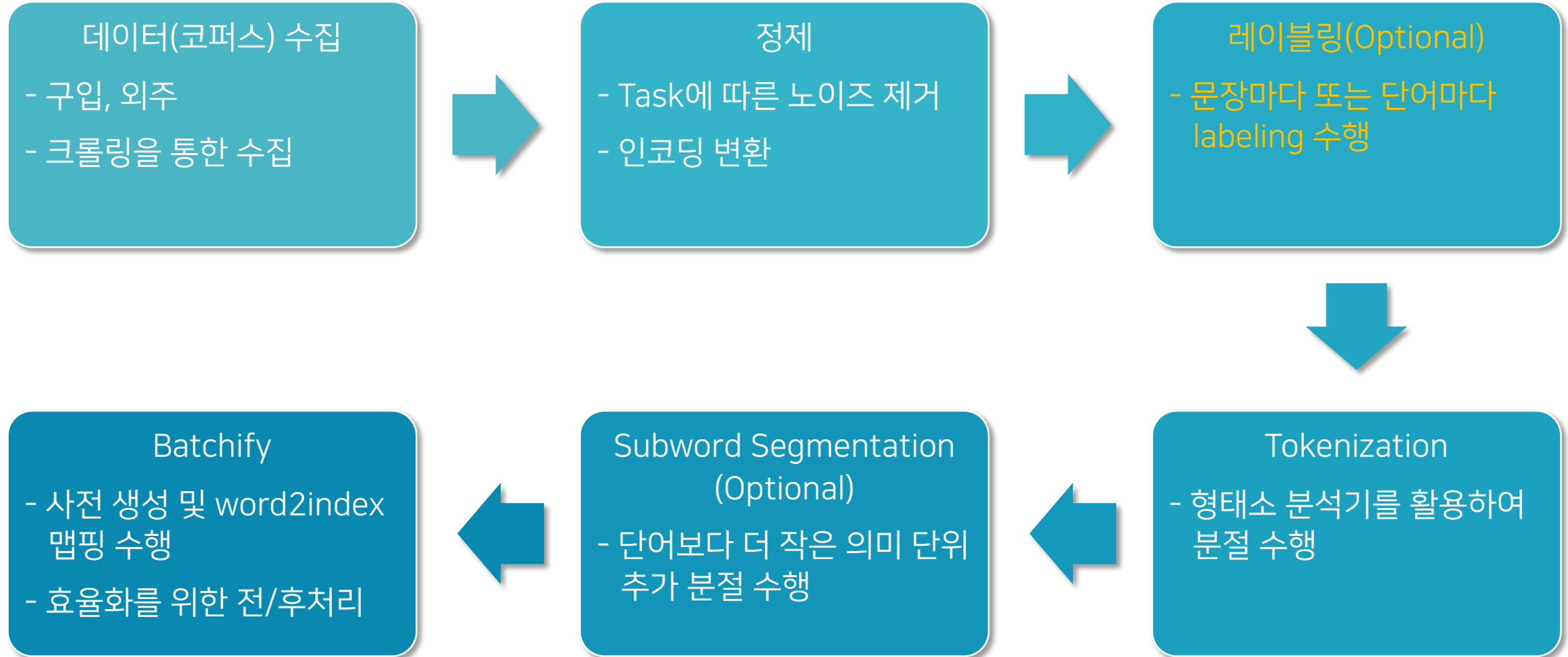


Labeling

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Preprocessing Workflow



Label

- Text Classification
 - INPUT: sentence
 - OUTPUT: class
- Token Classification
 - INPUT: sentence
 - OUTPUT: tag for each token → sequence
- Sequence-to-Sequence
 - INPUT: sentence
 - OUTPUT: sentence

Label Example

Sentence → Class

- TSV 형태의 하나의 파일
 - 각 row가 문장과 대응되는 레이블
 - 문장 column과 레이블 column 구성

Sentence → Sentence (Sequence)

- TSV 형태의 하나의 파일
 - 각 row가 대응되는 문장 쌍
 - 각 문장 별로 column을 구성
- 두 개 이상의 파일로 구성
 - 같은 순서의 row가 대응되는 문장 쌍
 - 한 문장당 여러 레이블이 존재 할 경우
 - e.g. 한국어 ↔ 영어 ↔ 중국어

Tip: 레이블링 직접 진행하기

- Human Labeling은 prototyping 시, 굉장히 강력한 도구 (두려워 하지 말자)
- 효율적인 레이블링 도구를 구성하자 (e.g. 엑셀)