삼성전기 AI전문가 양성과정 - 프로젝트 실습 (비영상)

자연어처리를 위한 Token Classification

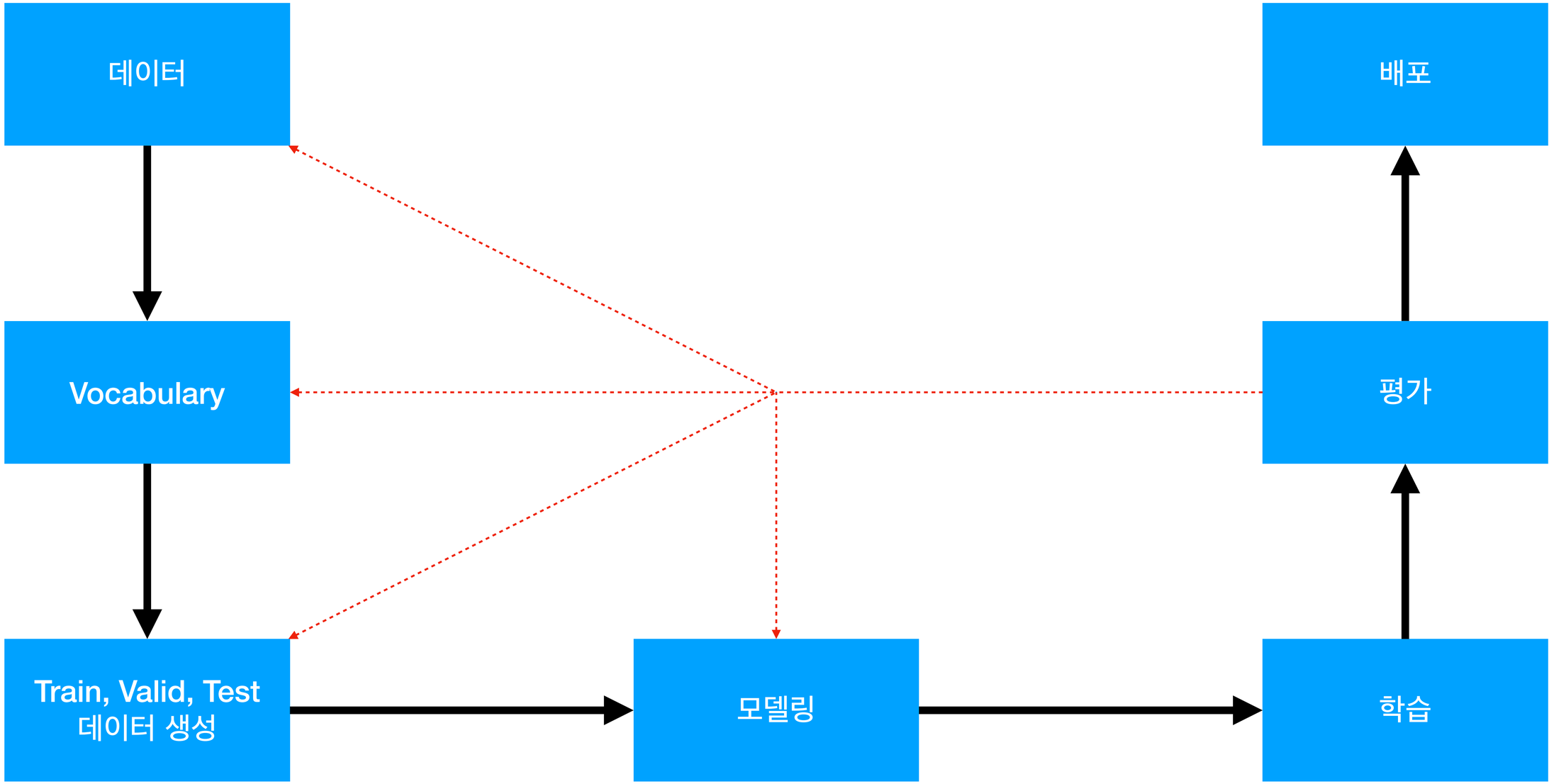
현청천

2022.02.28

Input Data

각 단어를 명사(1), 기타(0)으로 분류하는 Task

Workflow

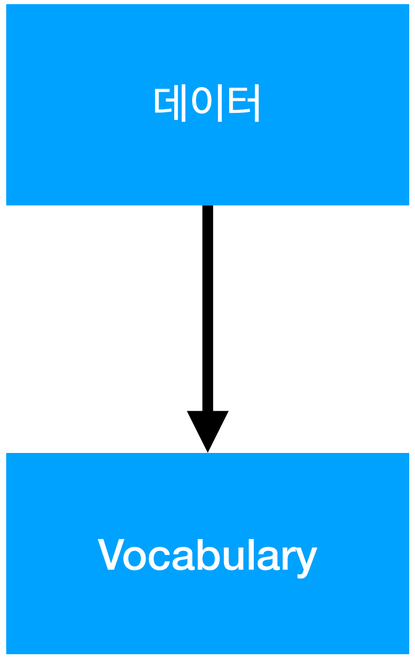


Workflow

데이터

잘 정돈된 데이터는 많을수록 좋음

Workflow

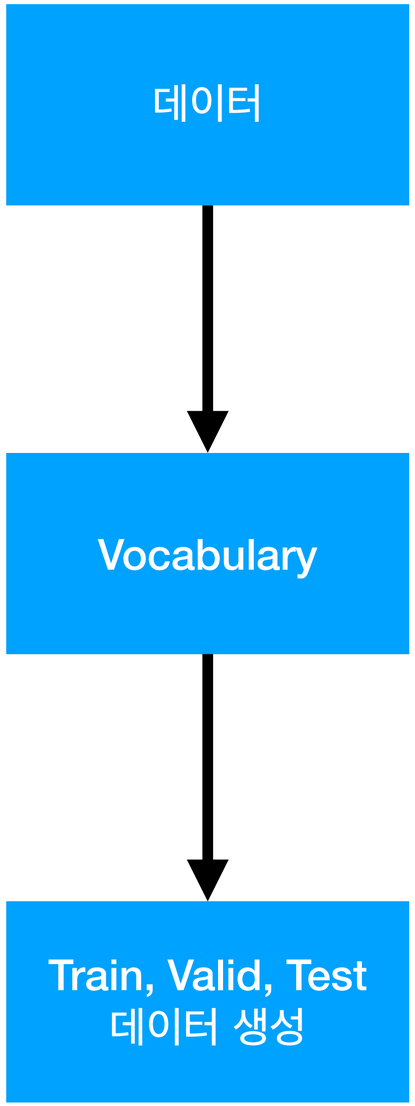


[PAD]	0
[UNK]	1
나는	2
학생	3
입니다	4
당신은	5
수학	6
선생님	7

미리 만들어진
Vocabulary를 사용
하기도 함

각 단어에 고유한 번호 부여

Workflow



- 문장들을 숫자의 배열로 변경
- 숫자의 배열을 하나의 행렬로 변경
  - 배열의 길이가 같도록 변경해 줘야 함

Train

2 (나는)	3 (학생)	4 (입니다)	0 ([PAD])
5 (당신은)	6 (수학)	7 (선생님)	4 (입니다)

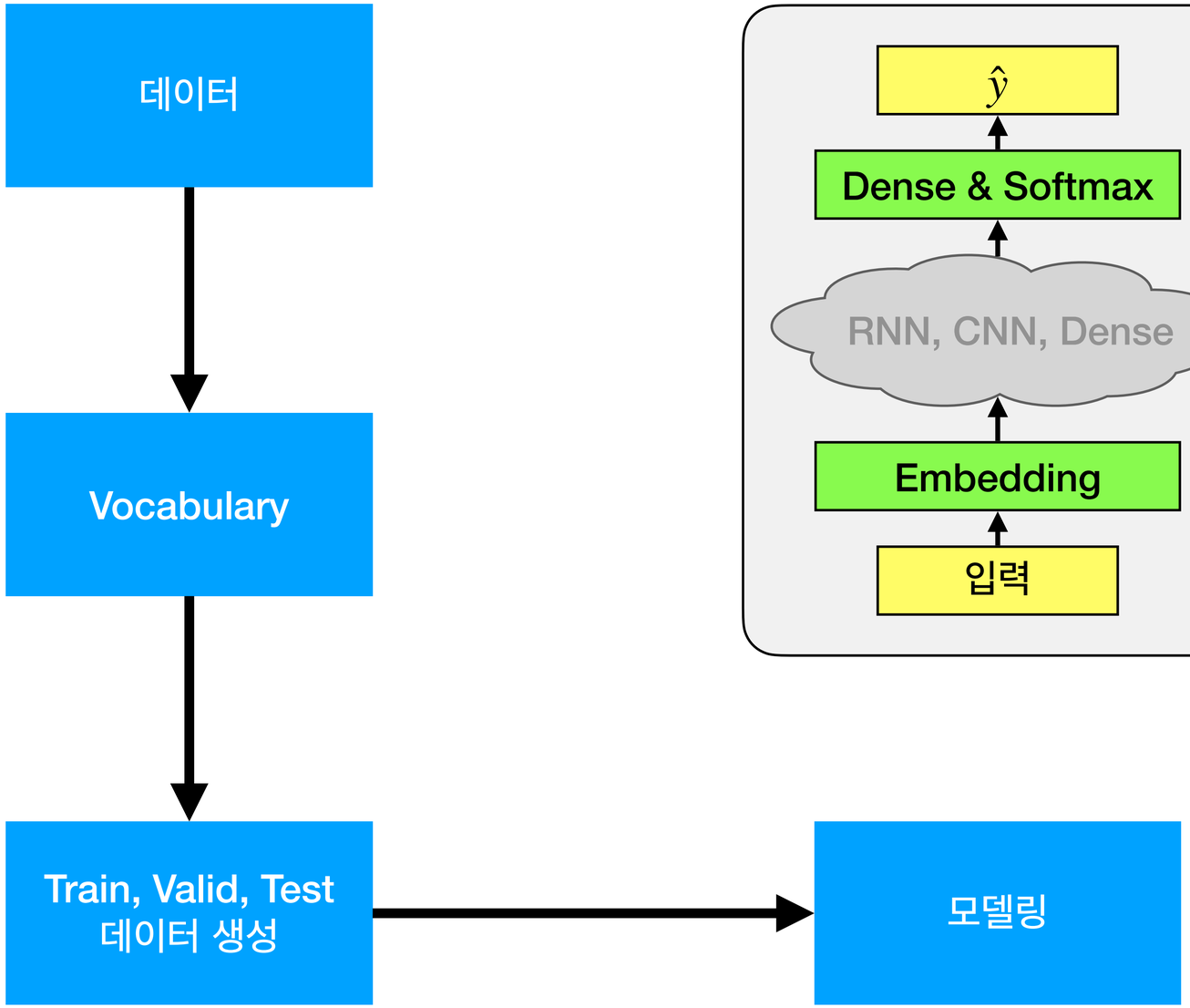
Valid

2 (나는)	7 (선생님)	4 (입니다)	0 ([PAD])

Test

5 (당신은)	6 (수학)	3 (학생)	4 (입니다)

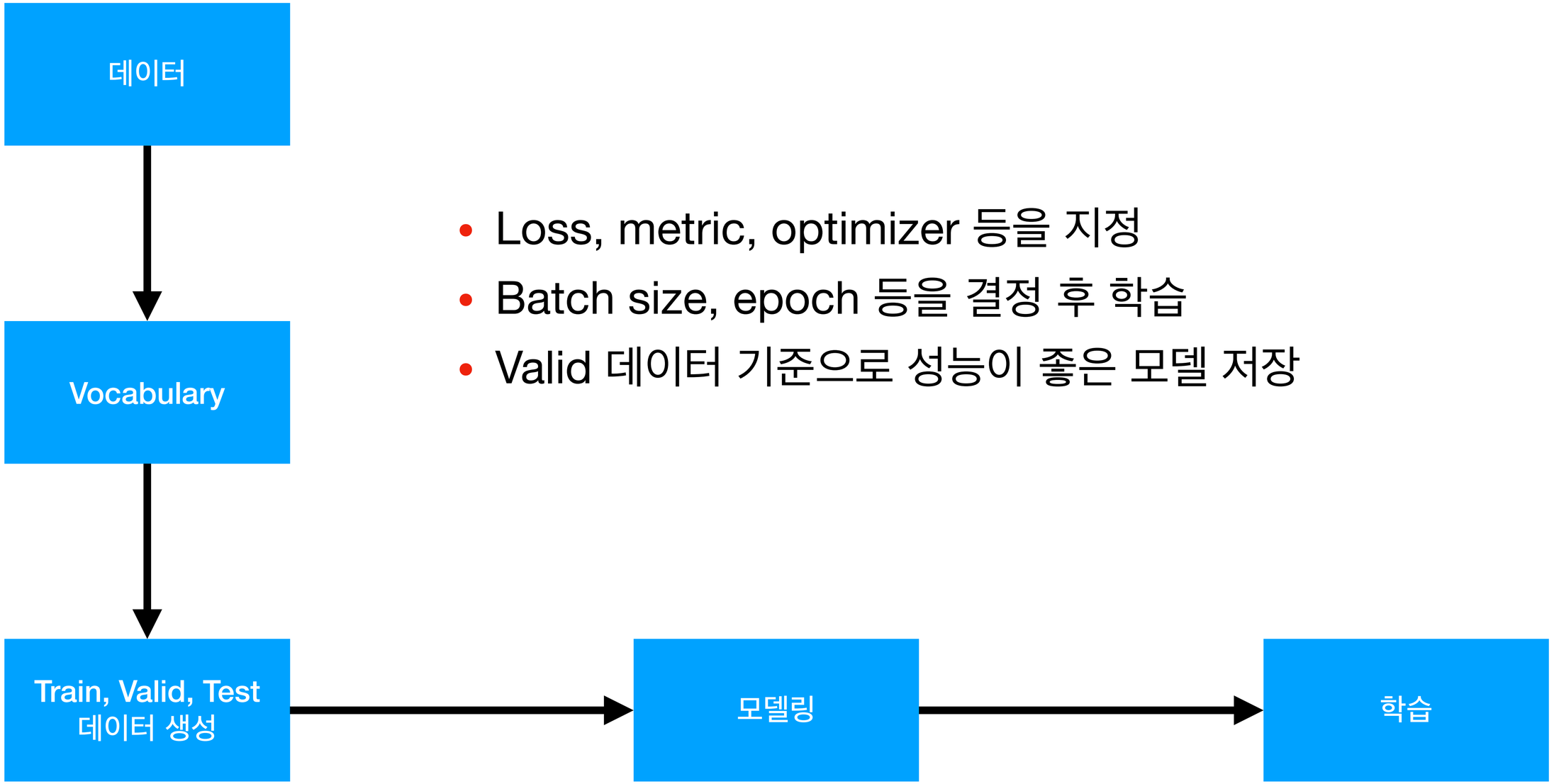
Workflow



$$\hat{y}_i = \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}$$

학습 목적에 맞게 모델링

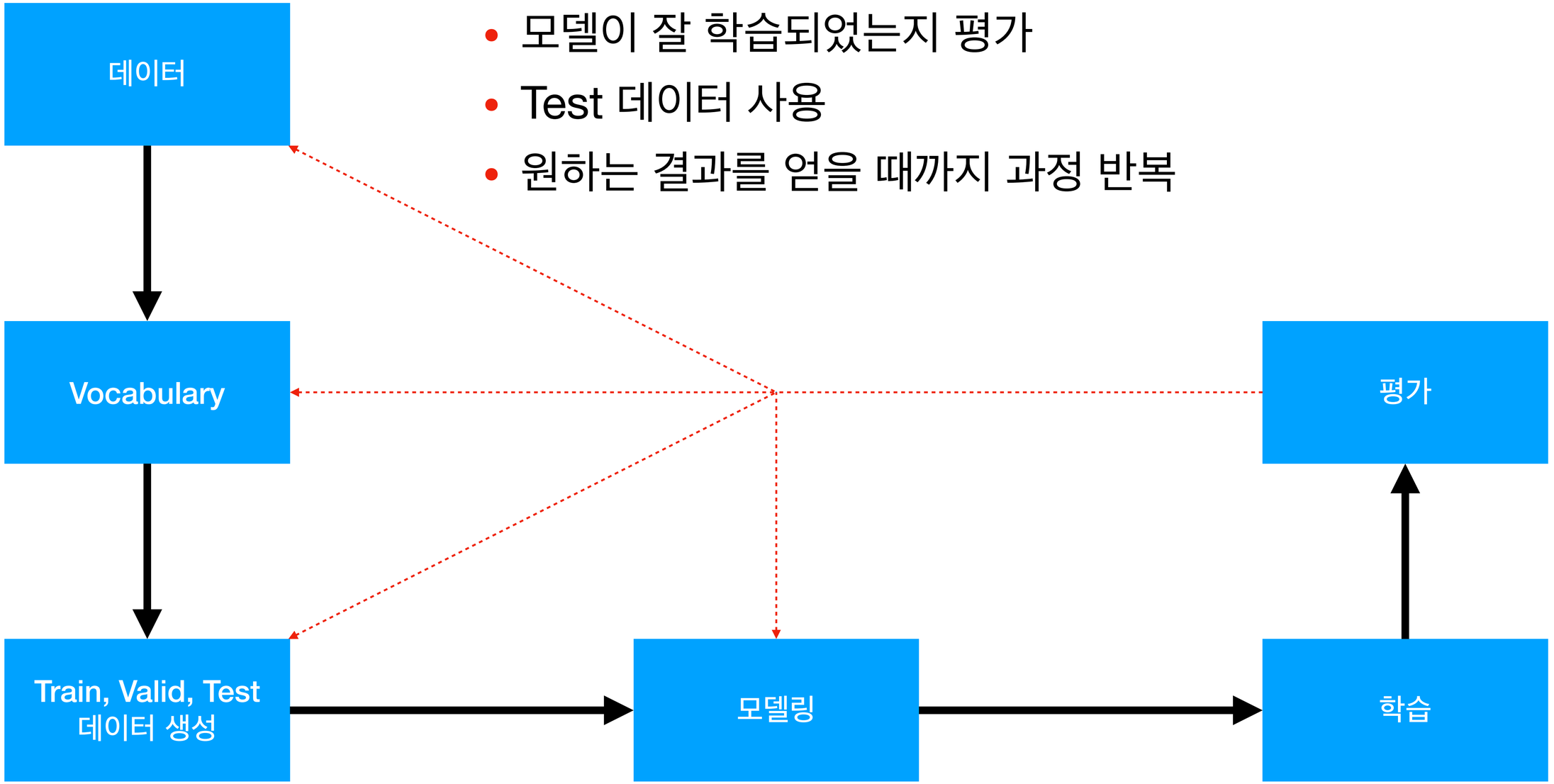
Workflow



- Loss, metric, optimizer 등을 지정
- Batch size, epoch 등을 결정 후 학습
- Valid 데이터 기준으로 성능이 좋은 모델 저장

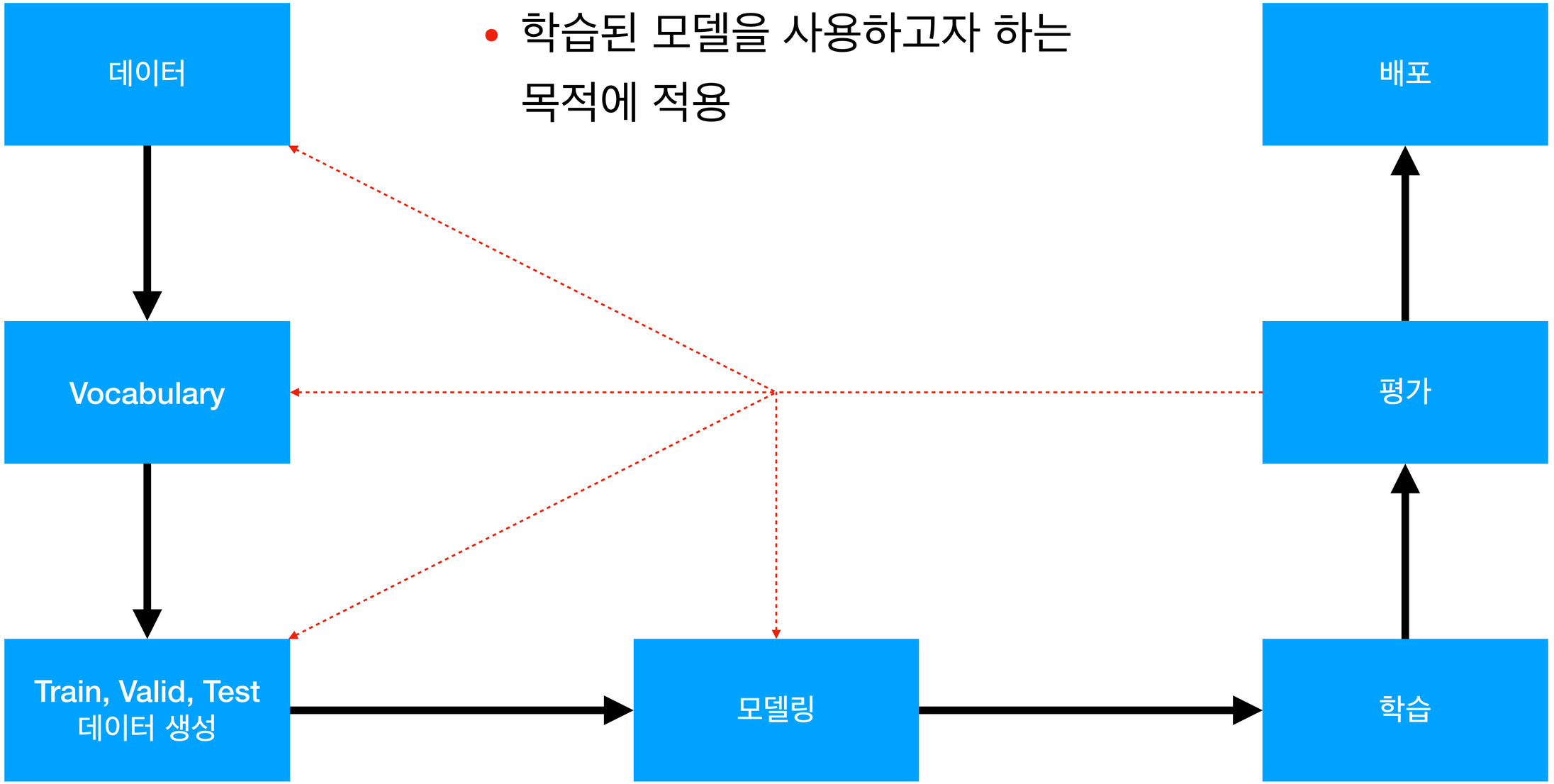
Workflow

- 모델이 잘 학습되었는지 평가
- Test 데이터 사용
- 원하는 결과를 얻을 때까지 과정 반복



Workflow

- 학습된 모델을 사용하고자 하는 목적에 적용



감사합니다.