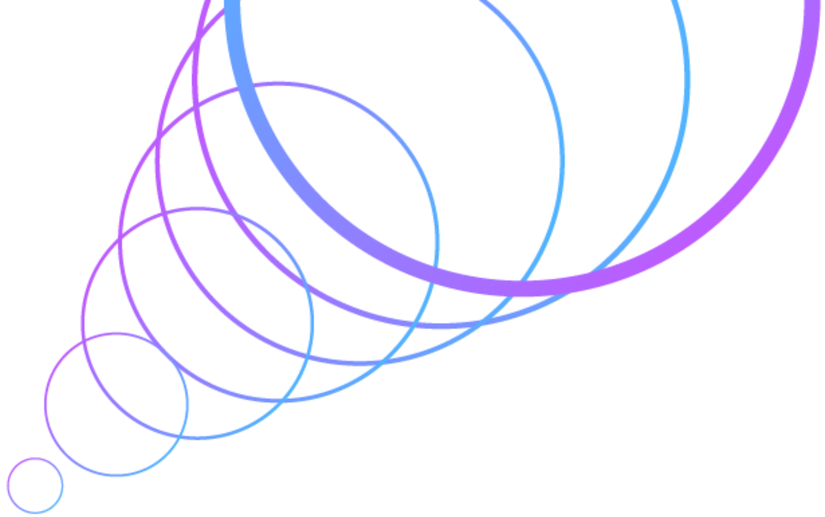


세상에 쓸모없는 데이터는 없다: HyperCLOVA를 이용한 반지도 학습

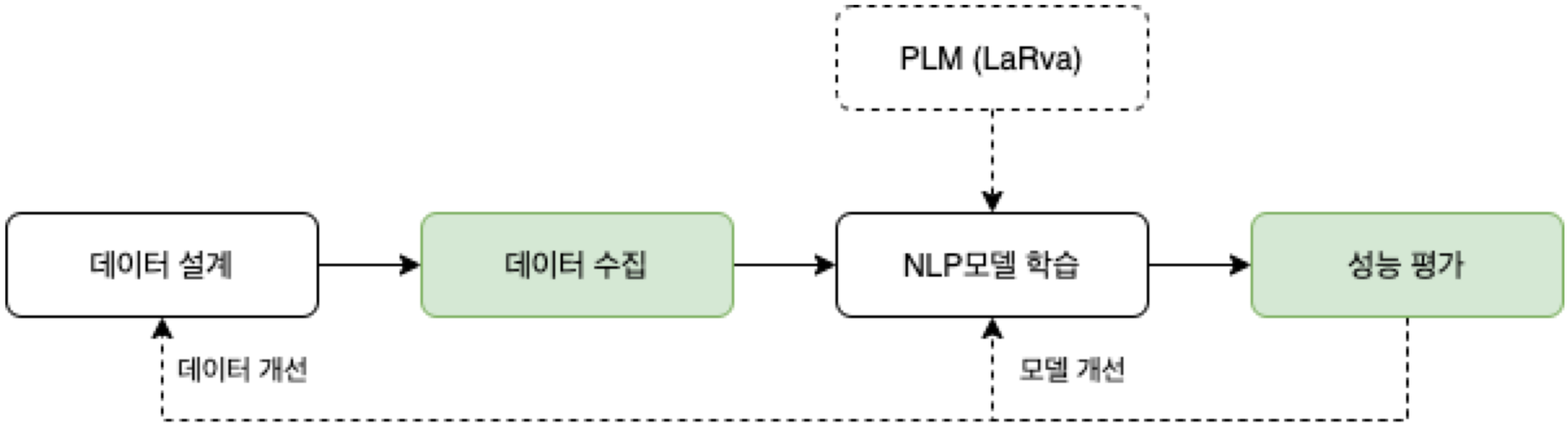
유강민 NAVER AI LAB
박동주 NAVER CLOVA

CONTENTS

1. 데이터 수집, 힘들지 않으세요?

1.1 Industry에서 데이터 수집하는 방법

사전학습된 언어모델을 이용한 파인튜닝 (fine-tuning) 패러다임



1.1 Industry에서 데이터 수집하는 방법

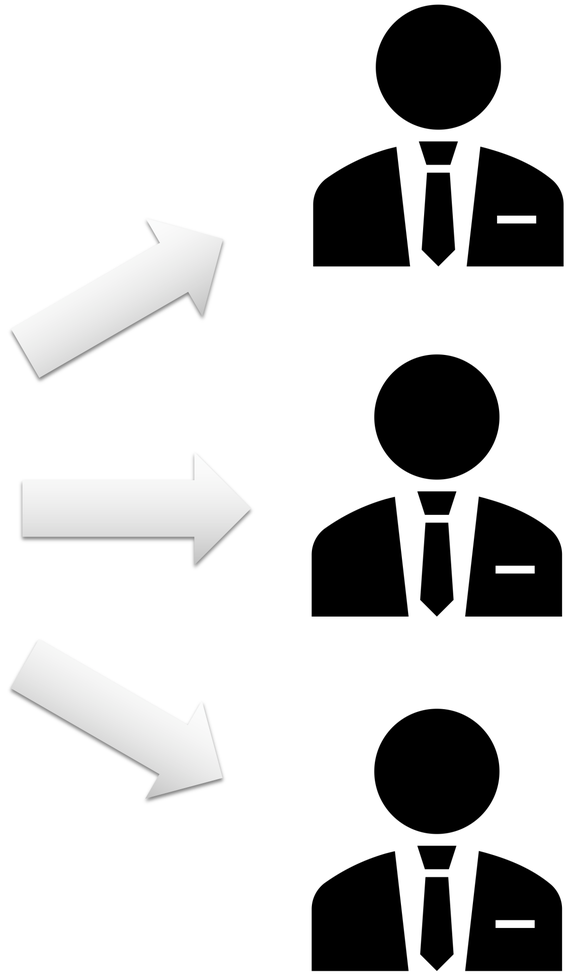
DEVIEW 2021

데이터 설계

- 문제 정의
- 모델 입출력 설계
- 클래스 정의

데이터 수집

- 수집 가이드라인 설계
- 수집 프로세스 설계
- 용역 통한 수집 실시
- 수집 결과 취합



1.1 Industry에서 데이터 수집하는 방법

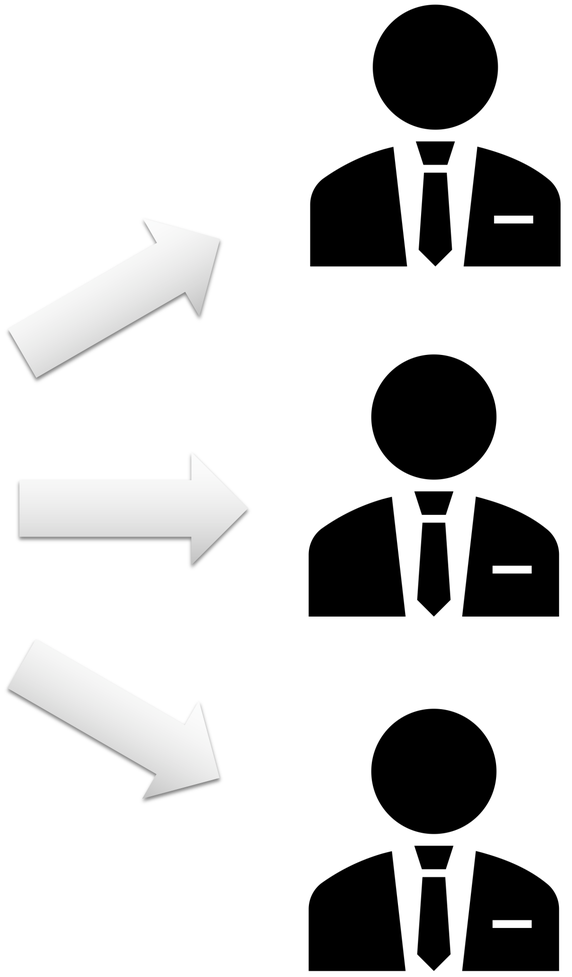
DEVIEW 2021

데이터 설계

- 문제 정의
- 모델 입출력 설계
- 클래스 정의

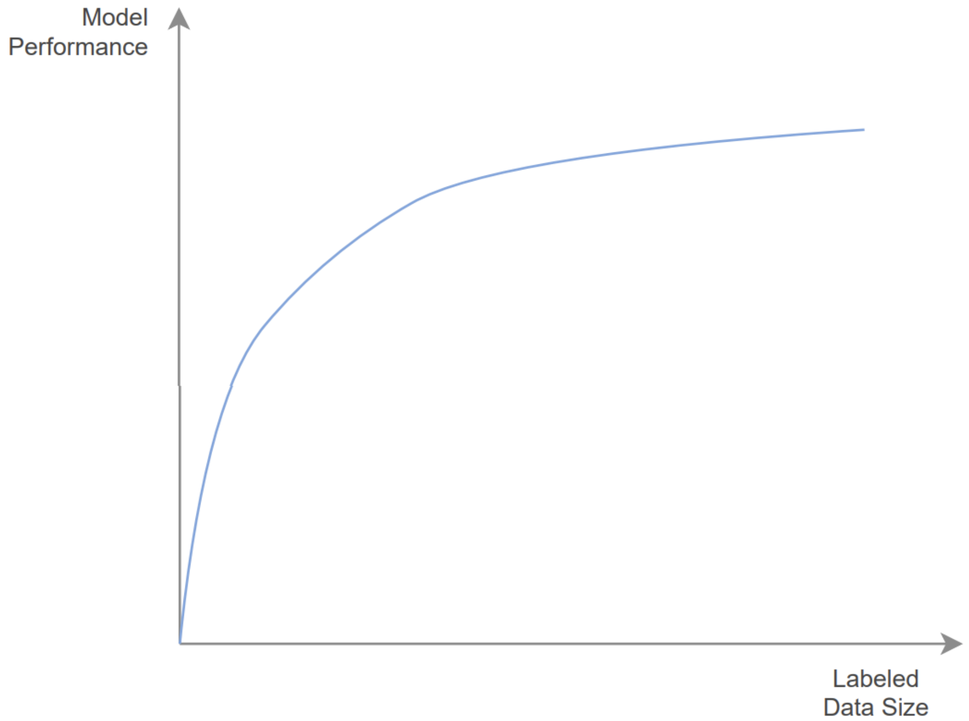
데이터 수집

- 수집 가이드라인 설계
- 수집 프로세스 설계
- 용역 통한 수집 실시
- 수집 결과 취합



1.2 데이터는 많을 수록 좋지만…

- 일반적으로 수집된 데이터 양과 모델 성능은 단조증가(monotonically increasing) 관계를 형성 -> Data-Performance Curve
- 데이터가 많아질 수록 수확체감(diminishing returns) 발생



1.2 데이터는 많을 수록 좋지만…

DEVIEW 2021

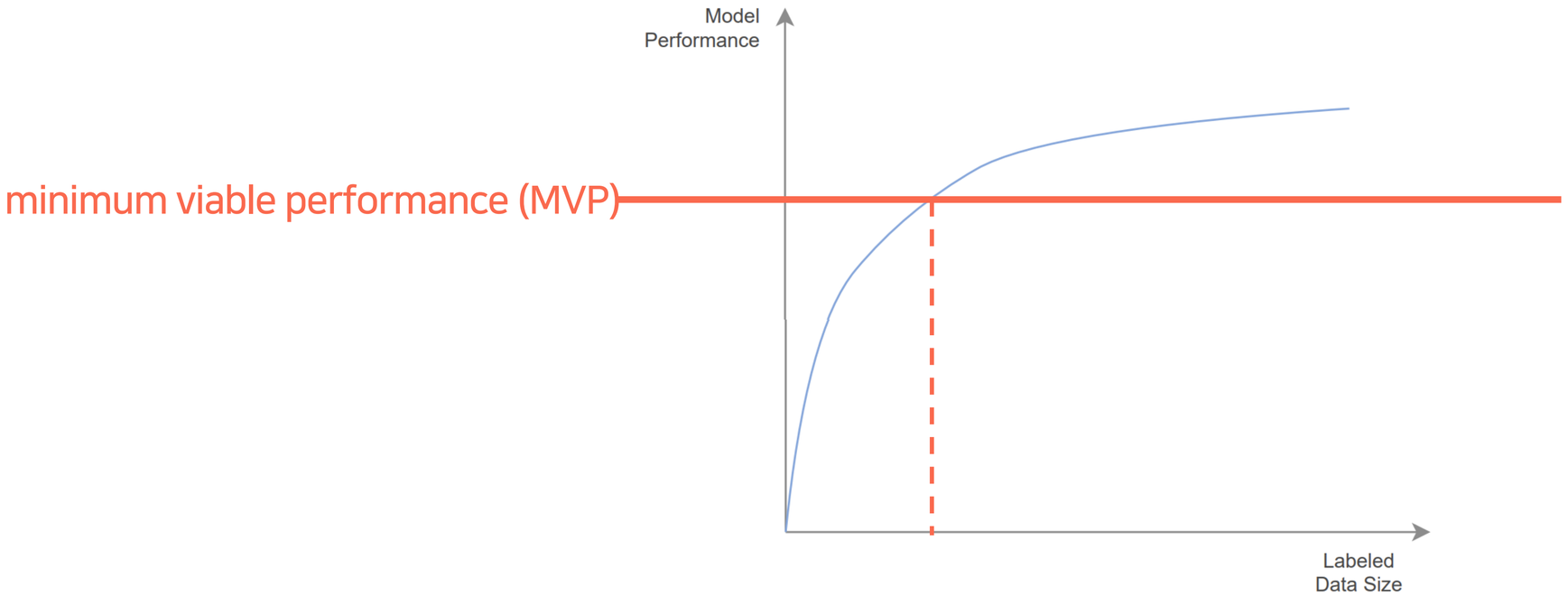
- 모델의 살이 되는 데이터, 첫 단추도 중요하지만 상용화를 통해 실전 데이터를 모아 지속적으로 모델 개선을 하는 것이 더욱 중요함

"Deploying to production means you're halfway there."

- Andrew Ng

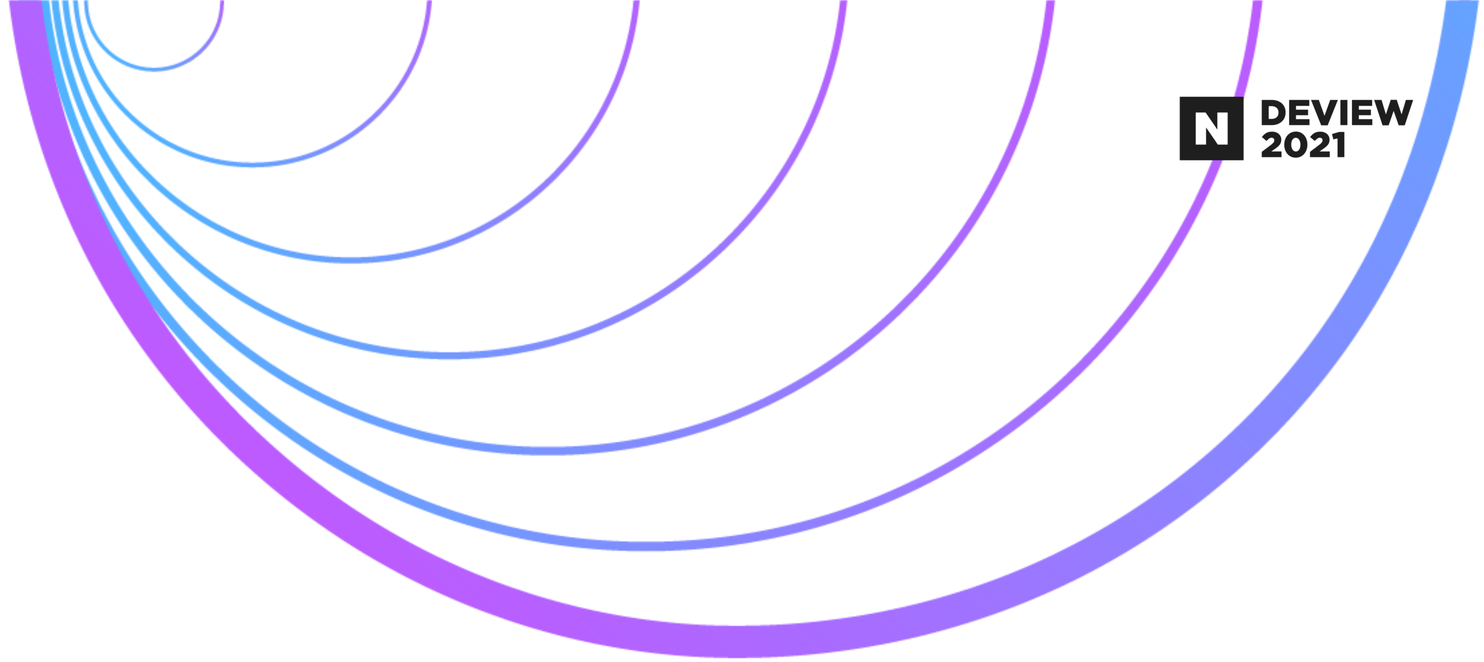
1.2 데이터는 많을 수록 좋지만…

- 최소 기능 성능(minimum viable performance)에 필요한 레이블 데이터만 우선적으로 수집
- 서비스 통해 레이블 없는 데이터 지속적으로 축적



1.3 데이터 수집의 딜레마

- MVP 달성을 위한 최소 데이터를 어떻게 가늠할 수 있는가?
- Quality over Quantity: 질 좋은 데이터를 확보하기 위해 데이터 수집 프로세스에 필요한 의사결정을 잘하는 방법은?
- MVP 이상으로 확보된 데이터(초기 수집 + 서비스 데이터)를 어떻게 활용할 것인가?



2. Semi-supervised Learning

2.1 AI는 혼자서도 배울 수 있어요

DEVIEW 2021

AI의 학습 방법

Supervised Learning

Semi Supervised Learning

Unsupervised Learning

2.1 AI는 혼자서도 배울 수 있어요

DEVIEW 2021

AI의 학습 방법

Data	Label
	곰
	호랑이
	닭
	오리

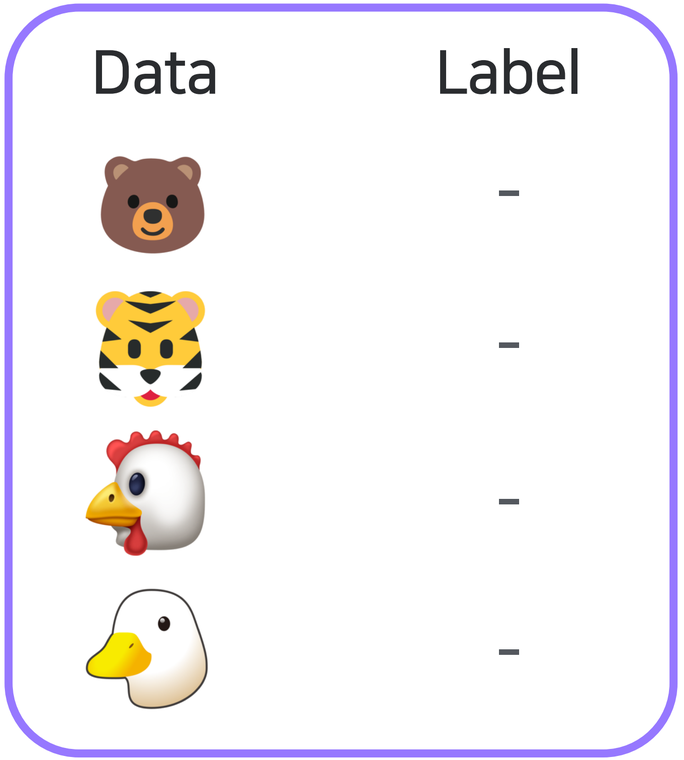
Supervised Learning

2.1 AI는 혼자서도 배울 수 있어요

DEVIEW 2021

AI의 학습 방법

Supervised Learning



Unsupervised Learning

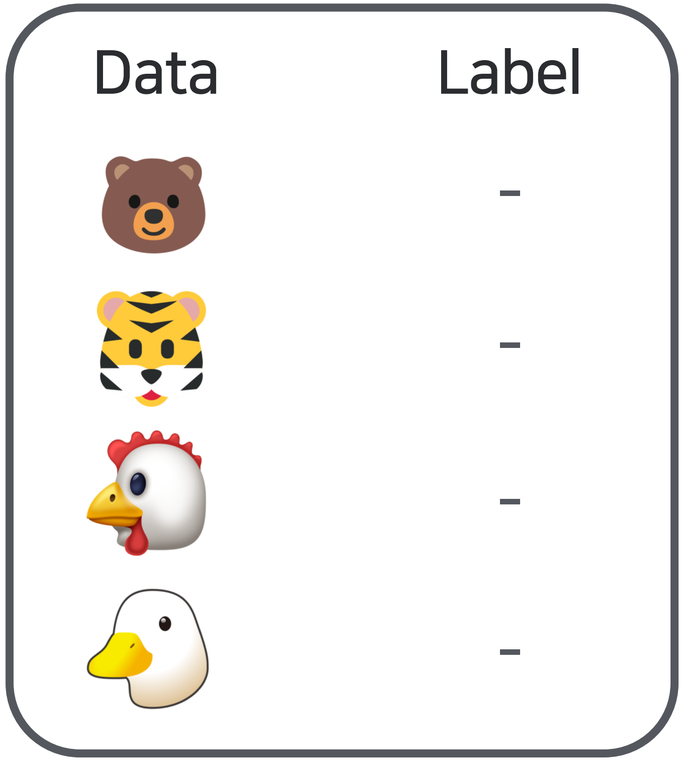
2.1 AI는 혼자서도 배울 수 있어요

DEVIEW 2021

AI의 학습 방법

Supervised Learning

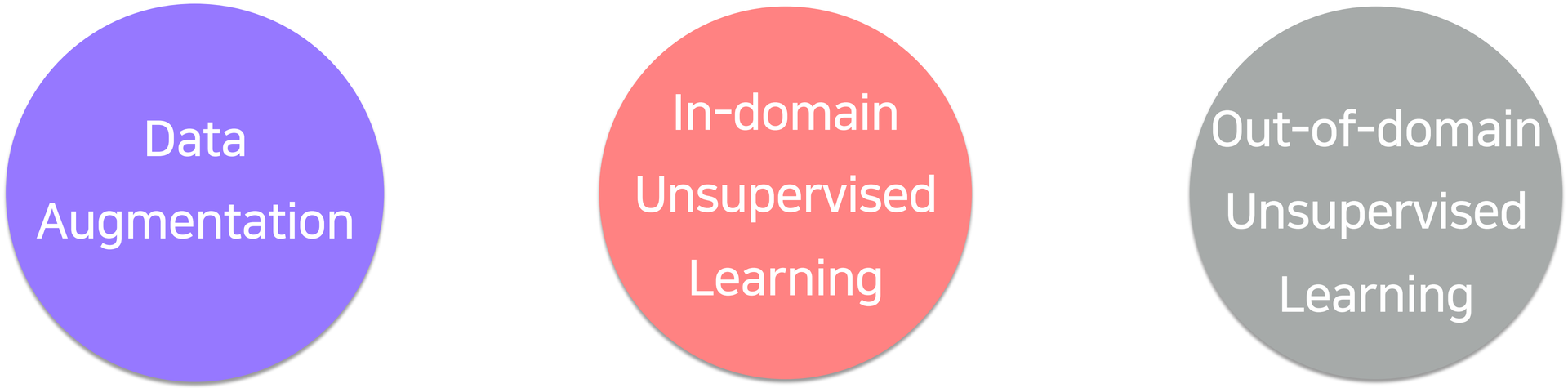
Semi-Supervised Learning



Unsupervised Learning

Semi-Supervised Learning (SSL)
어떻게 구성될까요?

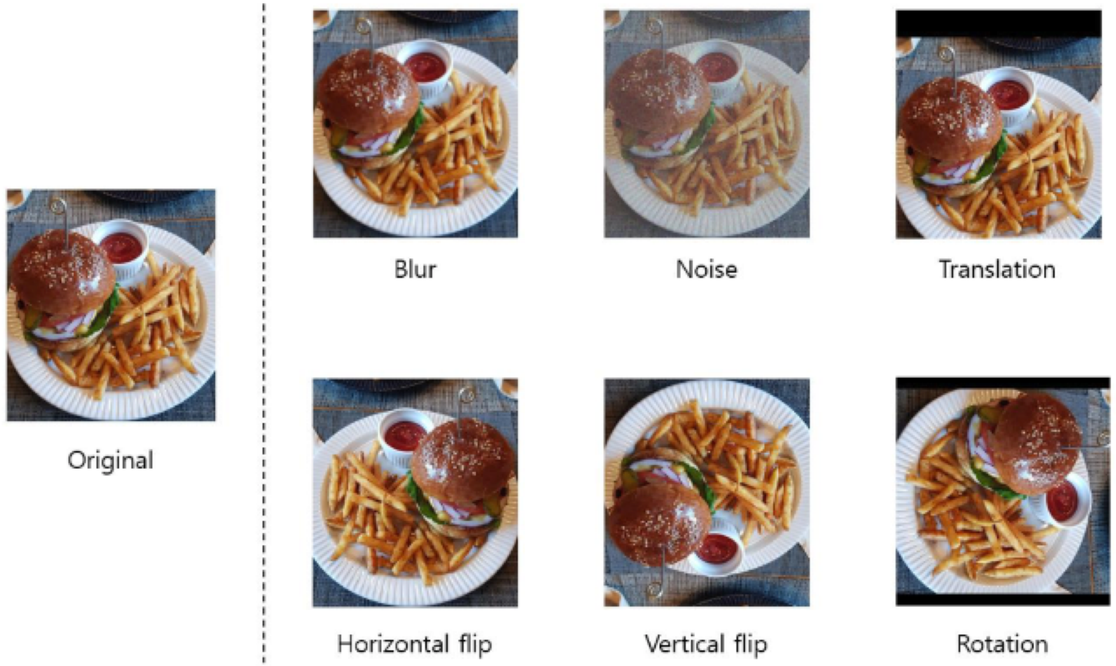
2.2 SSL의 구성 요소


Data Augmentation
In-domain Unsupervised Learning
Out-of-domain Unsupervised Learning

2.2.1 Data Augmentation

가지고 있는 데이터가 부족하다면?

- Data Augmentation



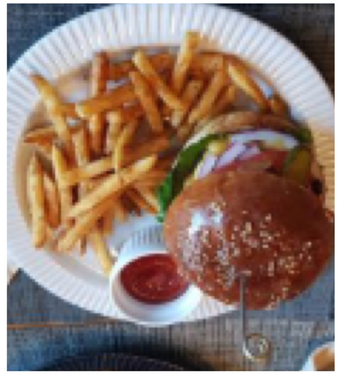
이미지 Augmentation 방법

2.2.1 Data Augmentation

좋은 Augmentation 이란?

- 기존 데이터의 중요한 특징을 잘 가지고 있으며
- 클래스 또는 클래스 정보를 잘 유지하고
- 새로운 학습 정보를 줄 수 있는 Augmentation

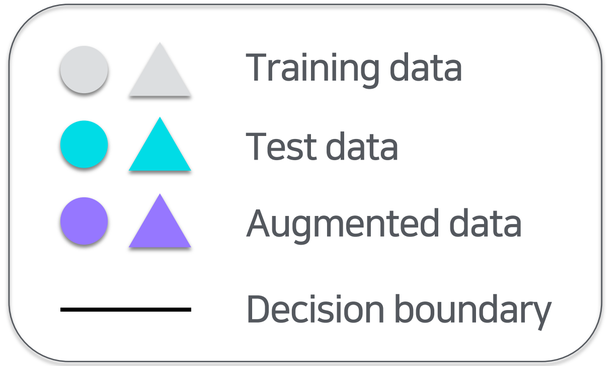
Augmentation

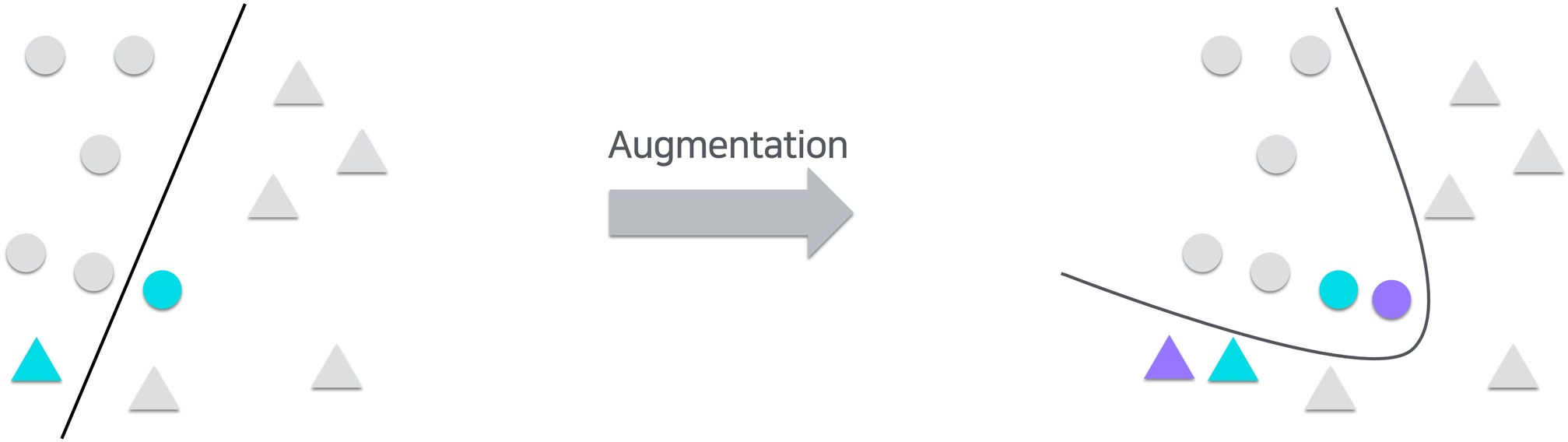


2.2.1 Data Augmentation

DEVIEW 2021

Decision Boundary 관점에서의 augmentation





2.2.1 Data Augmentation

대표적인 Text data augmentation 방법

Operation	Sentence
None	나는 맛있는 사과를 좋아한다.
Synonym Replacement	나는 맛있는 사과를 선호한다.
Random Insertion	나는 맛있는 사과를 매우 좋아한다.
Random Swap	사과를 맛있는 나를 좋아한다.
Random Deletion	나는 사과를 좋아한다.

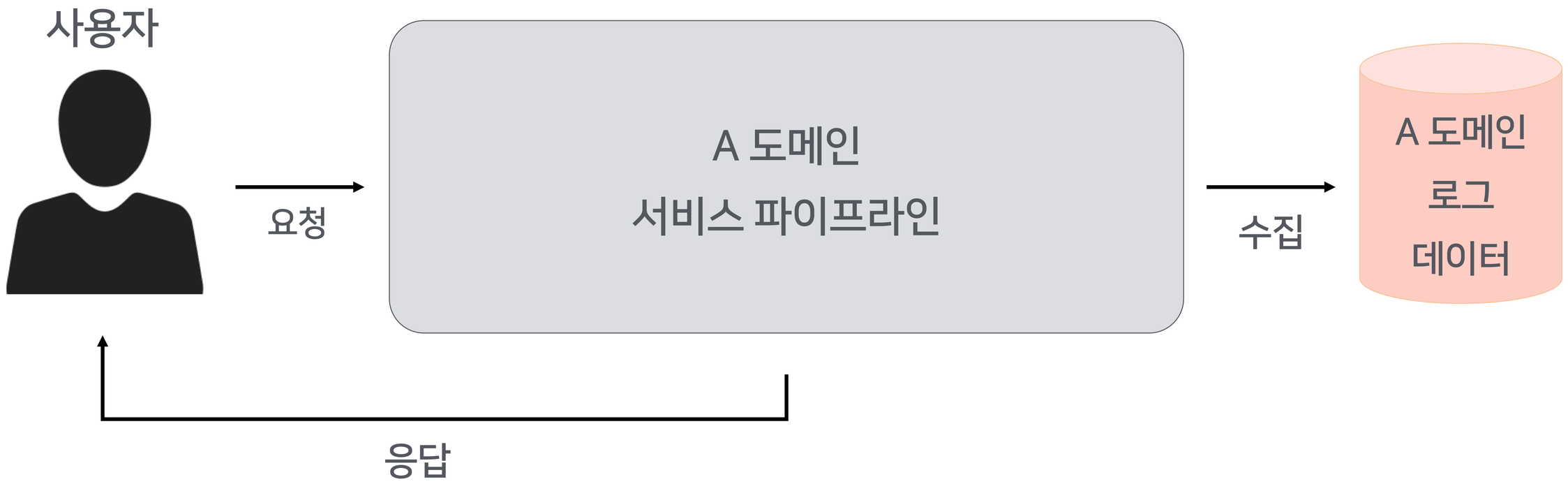
Easy Data Augmentation (EDA)

Back-translation

Data augmentation 사용!
그래도 데이터 부족...
함께 사용할 만한 다른 방법은 없을까?

2.2.2 In-domain Unsupervised Learning

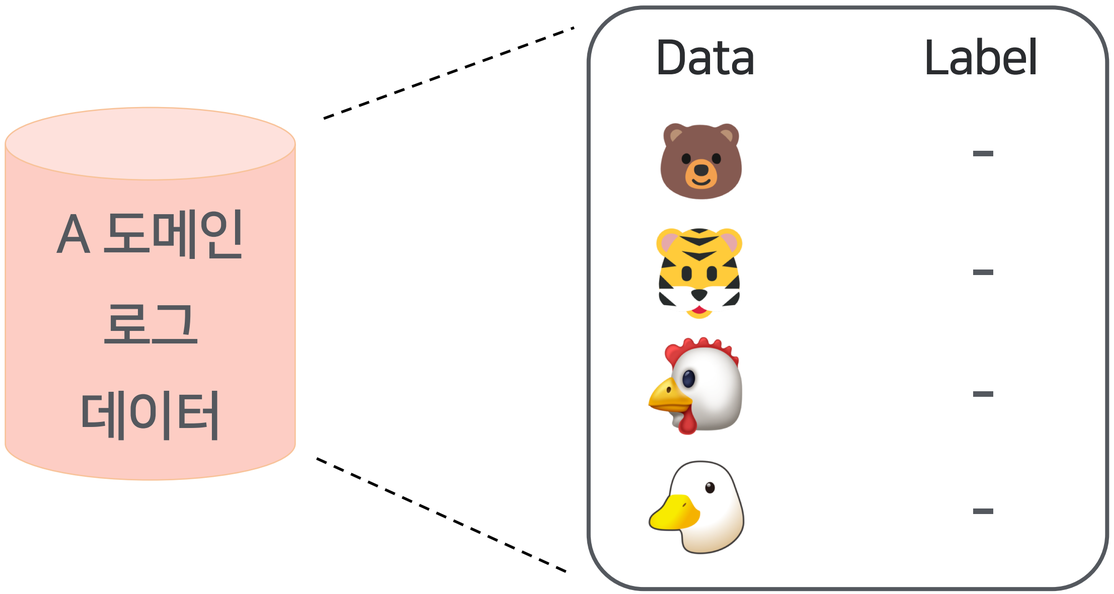
서비스할 수록 모이는 정제되지 않은 데이터, 버리지 말고 모델 모이로 !

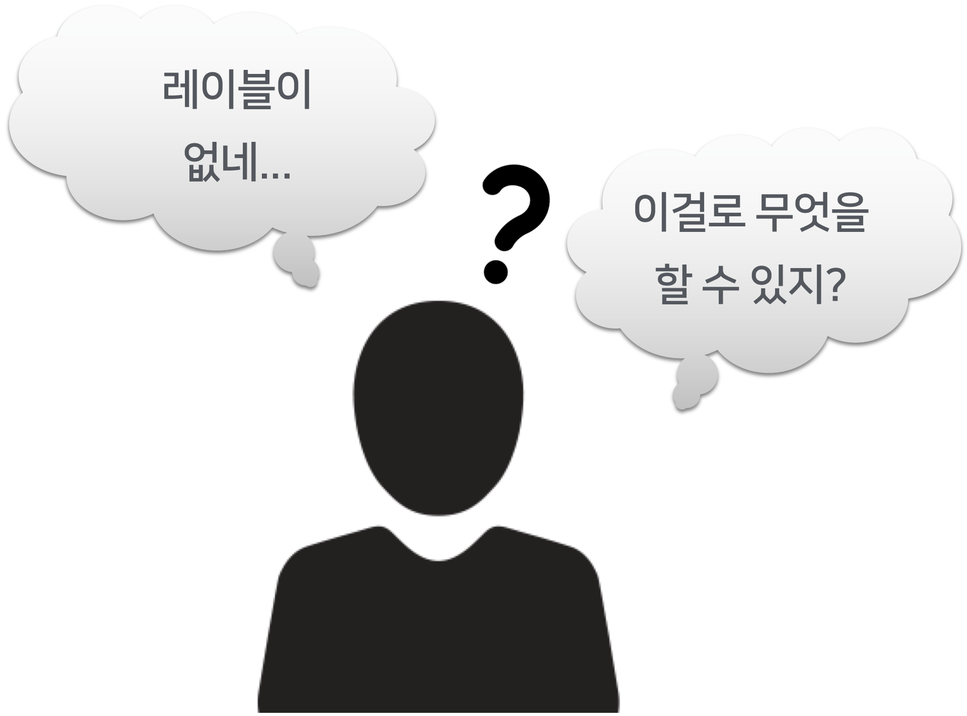


2.2.2 In-domain Unsupervised Learning

DEVIEW 2021

확보한 데이터로 무엇을 할 수 있을까?



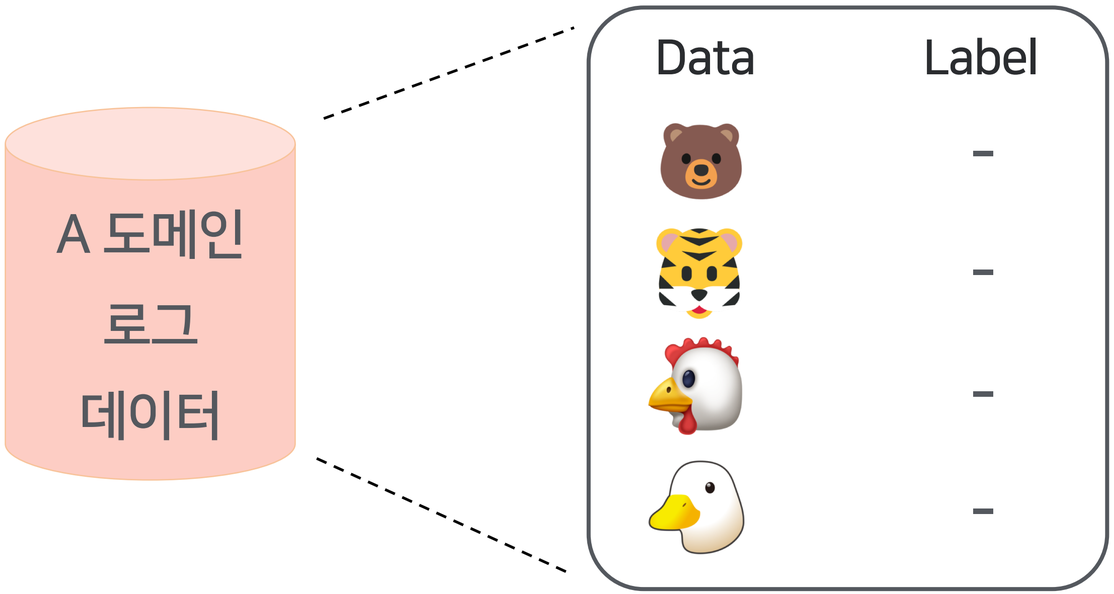


수 많은 데이터 확보

2.2.2 In-domain Unsupervised Learning

DEVIEW 2021

확보한 데이터로 무엇을 할 수 있을까?



수 많은 데이터 확보

2.2.2 In-domain Unsupervised Learning

DEVIEW 2021

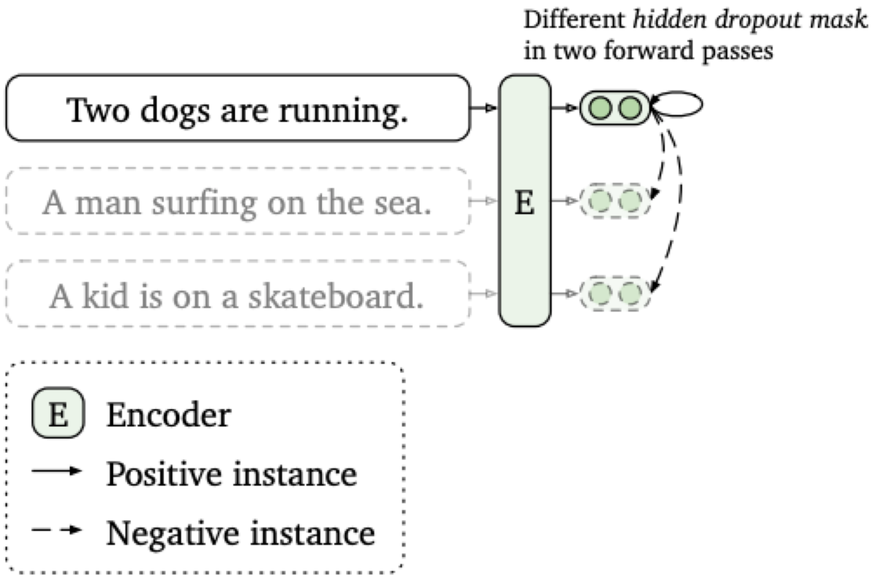
Consistency Regularization

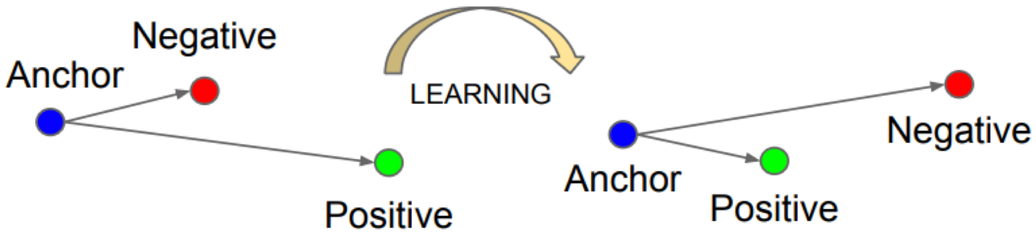
- 원본 데이터와 노이즈를 추가한 데이터간의 모델의 출력을 유사하게 만드는 학습 방법
- 방법에 따라 노이즈를 추가한 데이터만을 사용하기도 함
- 데이터의 representation 자체를 학습할 뿐만 아니라 다양한 노이즈에 대해 강건해짐

2.2.2 In-domain Unsupervised Learning

Contrastive Learning

- 주어진 데이터의 positive sample과 negative sample 을 통한 Metric learning 기반 학습 방법
- Positive sample에 대해 가까워지도록, negative sample에 대해서는 멀어지도록 학습





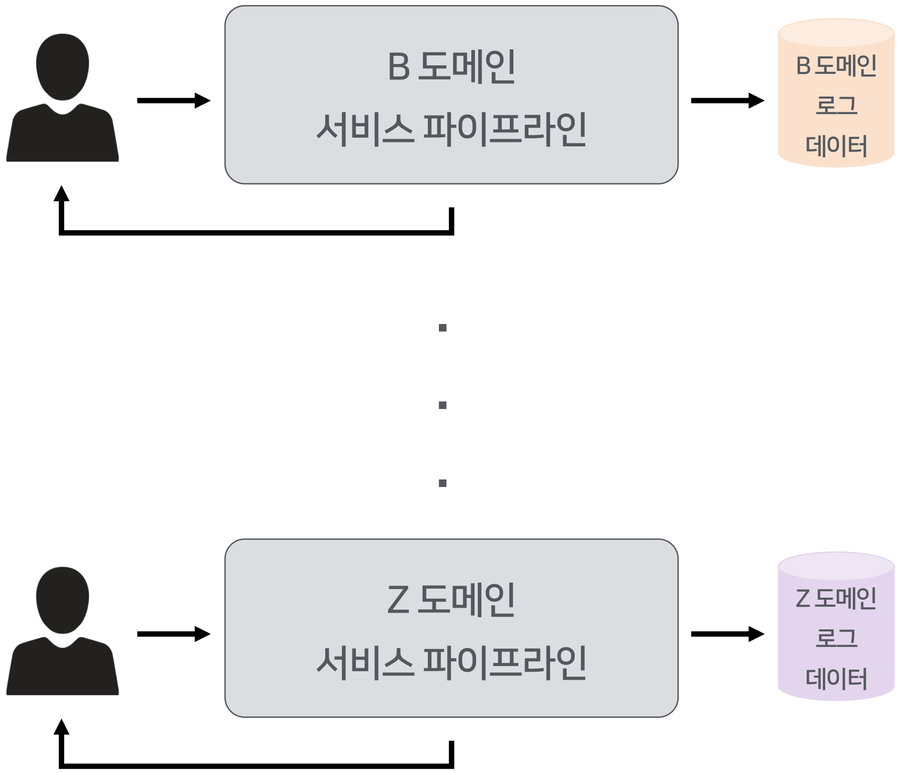
FaceNet: A Unified Embedding for Face Recognition and Clustering (Florian Schroff et al, 2015)

SimCSE: Simple Contrastive Learning of Sentence Embedding (Tianyu Gao et al, 2021)

서비스 로그 또는 동일한 도메인
데이터가 존재하지 않는다면?

2.2.3 Out-of-domain Unsupervised Learning

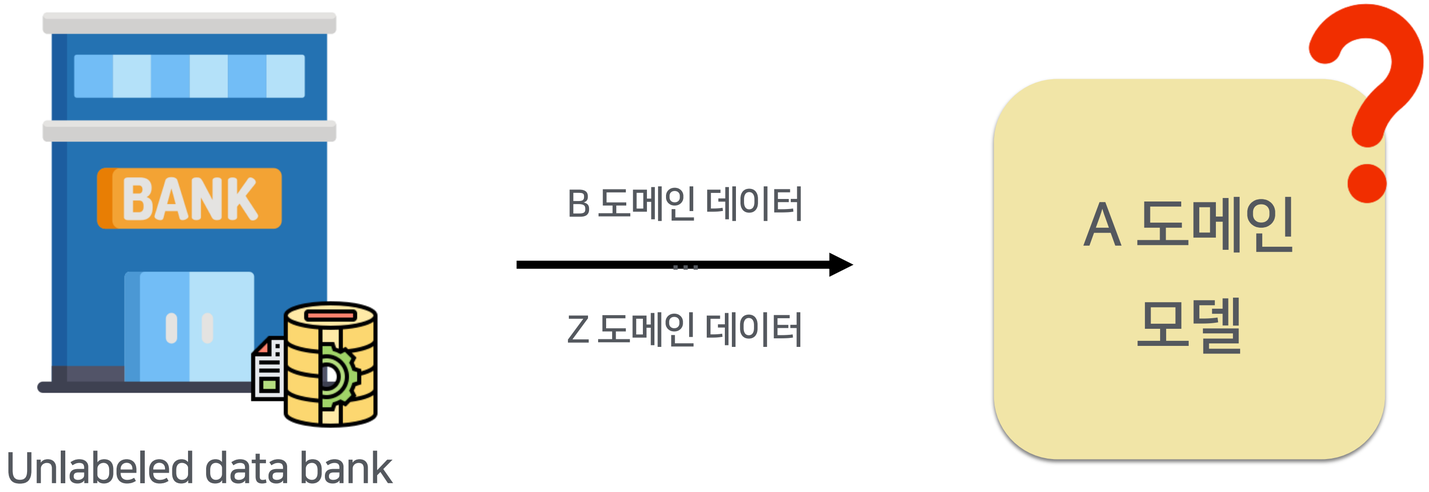
다른 도메인의 데이터를 현재 도메인의 데이터처럼 !



Unlabeled data bank

2.2.3 Out-of-domain Unsupervised Learning

많은 데이터를 확보했지만 이렇게 많은 데이터를 다 사용할 수 있을까?

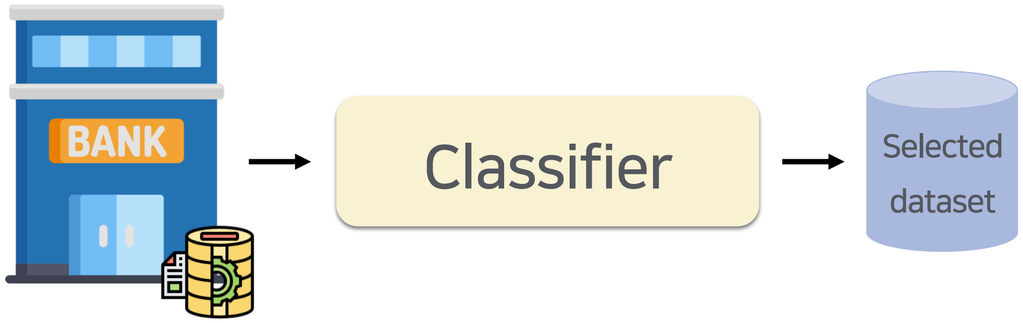


필요한 데이터를 선택하는 Data selection 방법이 필요

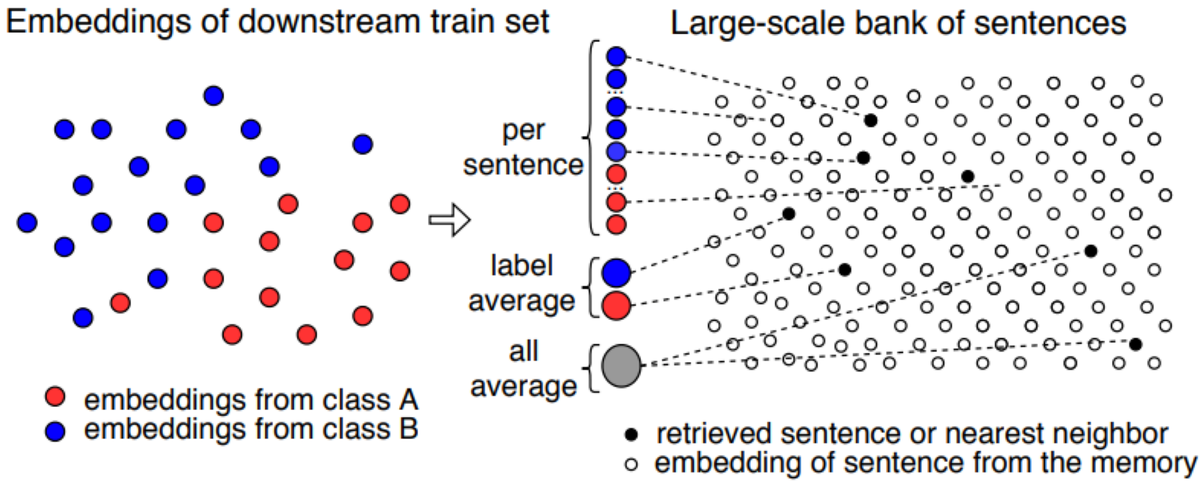
2.2.3 Out-of-domain Unsupervised Learning

DEVIEW 2021

대표적인 Data selection 방법



Classifier-based selection



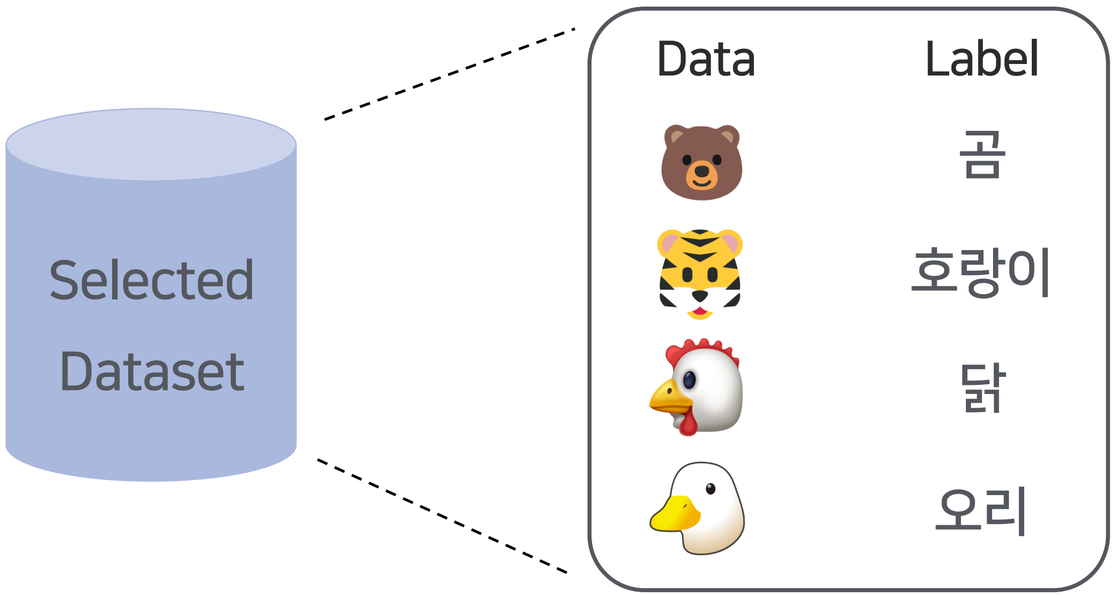
Retrieval-based selection

Self-training Improves Pre-training for Natural Language Understanding (Jingfei Du et al, 2020)

2.2.3 Out-of-domain Unsupervised Learning

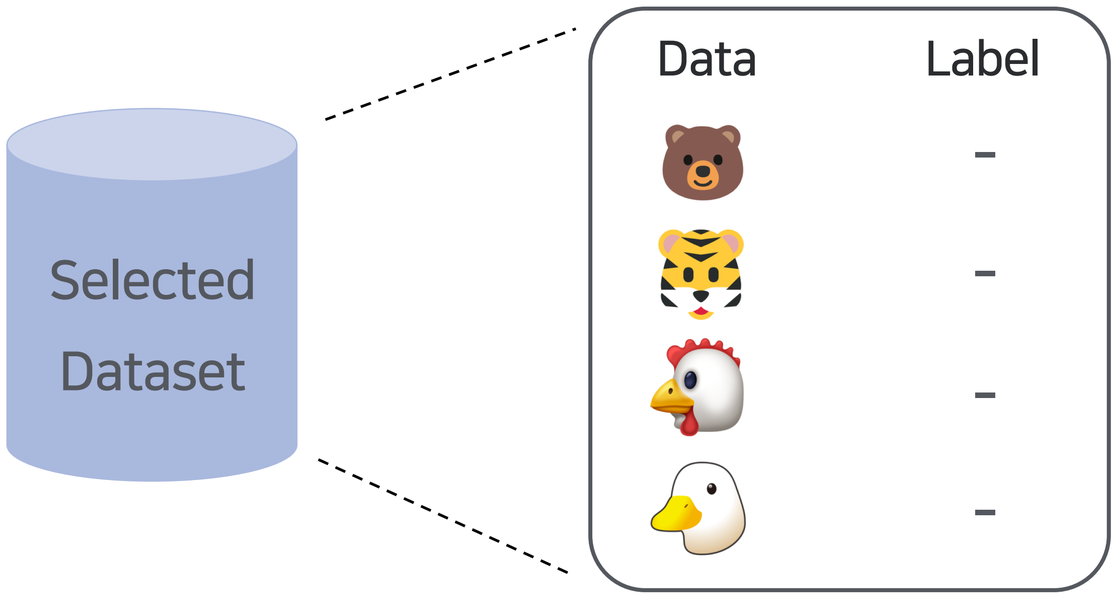
DEVIEW 2021

그렇다면 이렇게 선택된 데이터는 어떻게 사용할 수 있을까?



Pseudo- / soft- 레이블이 부착되었다면

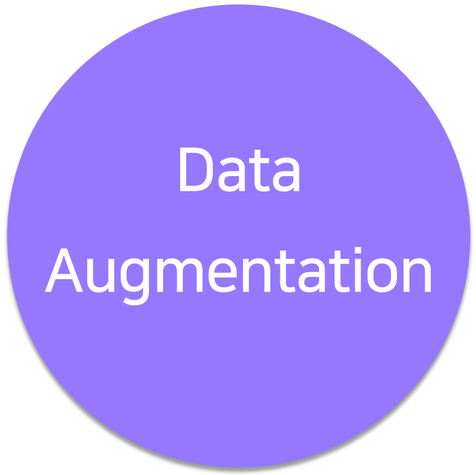
(Semi-) Supervised learning



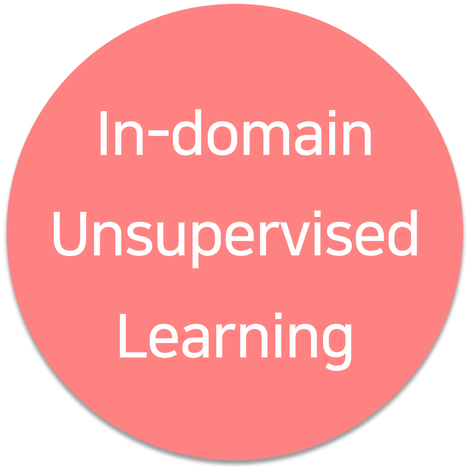
레이블이 부착되지 않았다면

In-domain unsupervised learning

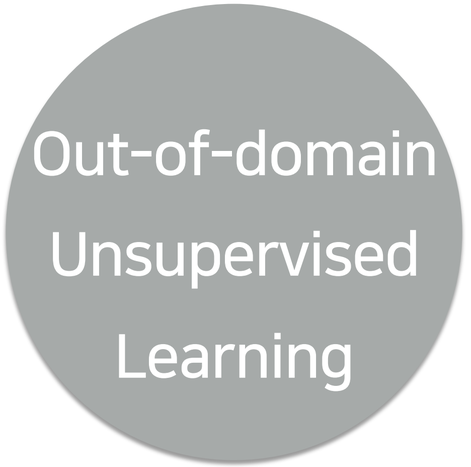
2.2 SSL의 구성 요소



가지고 있는 데이터가 조금 부족하다면?



서비스 데이터가 있다면?



동일한 도메인 데이터가 존재하지 않는다면?

DEVIEW 2021

Industrial NLP에서
Semi-Supervised Learning은 무적 !?

2.3 Industrial NLP에서 SSL의 한계

1. Augmented / Perturbed data가 원본 데이터의 클래스를 보존하지 못하거나 원본 데이터와 관계가 없을 수 있음

이 영화 정말 재미있네! (긍정)

Augmentation

이 영화 정말 재미없네! (긍정)

정말 배고프네 (긍정)

2.3 Industrial NLP에서 SSL의 한계

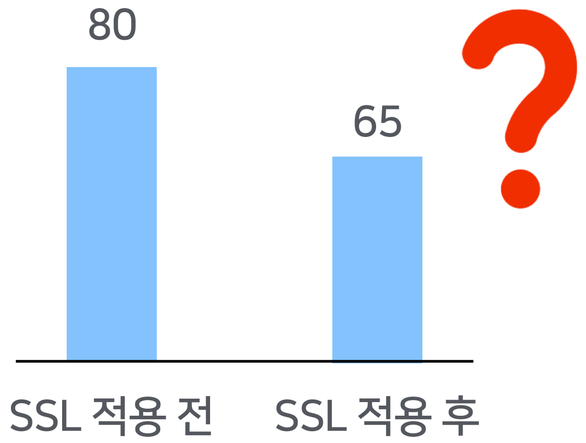
DEVIEW 2021

2. 항상 성능향상을 보장할 수 없고 오히려 떨어질 수도 있는 리스크가 존재하여 도입 여부에 대한 신중한 판단 필요

지금 SSL을 사용해도 되나?

성능이 오를 수 있을까?

일단 사용해보자 !



2.3 Industrial NLP에서 SSL의 한계

3. 새로운 도메인의 경우 로그 데이터에서도 적절한 데이터를 구하기 힘들 수 있음

Unlabeled data bank

Data selection method

?

"공룡과의 대화" 도메인 모델

강력한 언어모델
HyperCLOVA 를 이용하자!

3. HyperMatch: HyperCLOVA-powered SSL

3.1 HyperMatch 구조

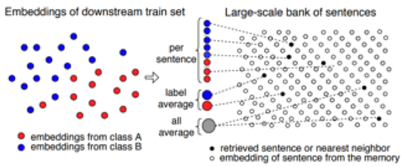
Unlabeled data bank

EDA | Back-translation | HyperMix | ...

Augmentation Techniques

HyperCLOVA

Classifier



Data selection algorithm

HyperMatch SSL Core

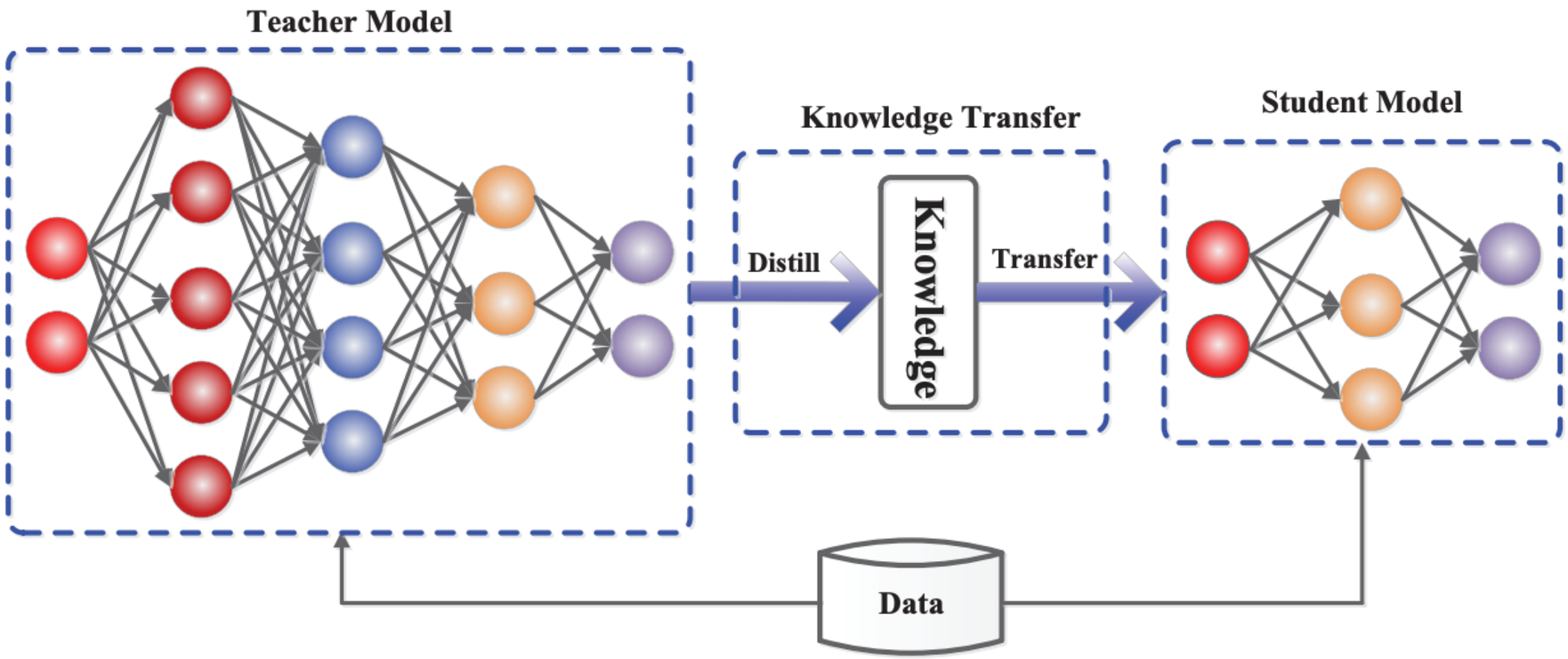
- Consistency regularization
- Contrastive learning
- Augmentation anchoring
- ...

3.1.1 HyperCLOVA의 활용

DEVIEW 2021

HyperCLOVA가 가지고 있는 언어 지식을 어떻게 이용할 수 있을까?

- HyperCLOVA의 지식을 추출하자 - Knowledge Distillation



Knowledge Distillation: A Survey (Jianping Gou et al, 2021)

3.1.1 HyperCLOVA의 활용

DEVIEW 2021

HyperCLOVA가 가지고 있는 언어 지식을 어떻게 이용할 수 있을까?

- 프롬프트를 이용하여 HyperCLOVA의 지식 추출

입력 (Prompt)

출력

```
1  Translate English to French:          ← task description
2  sea otter => loutre de mer            ← examples
3  peppermint => menthe poivrée          ←
4  plush girafe => girafe peluche        ←
5  cheese =>                             ← prompt
```

Language Models are Few-Shot Learners (Tom B. Brown et al, 2020)

3.1.1 HyperCLOVA의 활용

DEVIEW 2021

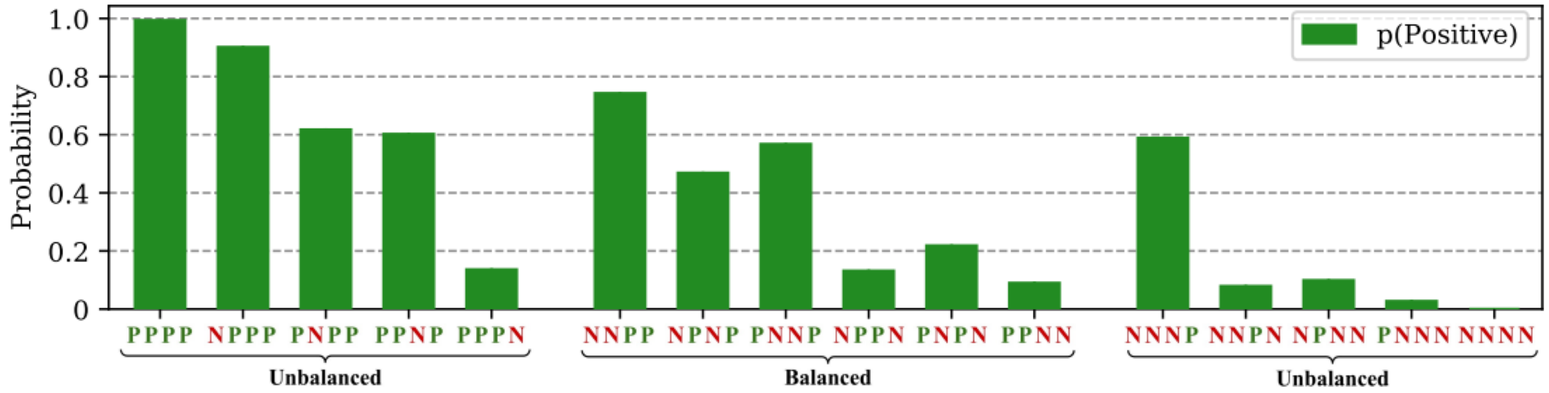
HyperCLOVA 와 같은 Large-scale language model 의 중요한 특성

- 입력으로 주어지는 프롬프트에 아주 민감

Input : 재밌네	Sentiment : 긍정
Input : 꿀잼이야	Sentiment : 긍정
Input : 노잼	Sentiment :

HyperCLOVA

긍정 ?



프롬프트 구성에 사용되는 레이블(Sentiment) 과
순서에 따른 “긍정” 레이블이 나올 확률

Calibrate Before Use: Improving Few-Shot Performance of Language Models (Tony Z. Zhao et al, 2020)

3.2 HyperCLOVA를 이용한 데이터 증강 및 생성

3.2.1 HyperCLOVA를 이용한 데이터 증강

DEVIEW 2021

1. 데이터와 레이블이 있는 경우 HyperCLOVA를 통한 증강

다음은 영화리뷰와 감정 예시입니다. 감정은 '긍정', '중립', '부정' 중 하나 입니다.

영화리뷰 : 이 영화는 너무 재미있고 감동적이에요! (긍정)
영화리뷰 : 음... 시간 보내기에는 좋지만 꼭 보진 않아도 됩니다 (중립)
영화리뷰 : 이걸 왜 돈주고 봐야하지? (부정)
영화리뷰 :

HyperCLOVA

재미 없어서 시간도 아깝고 돈도 아깝다 (부정)

3.2.1 HyperCLOVA를 이용한 데이터 증강

DEVIEW 2021

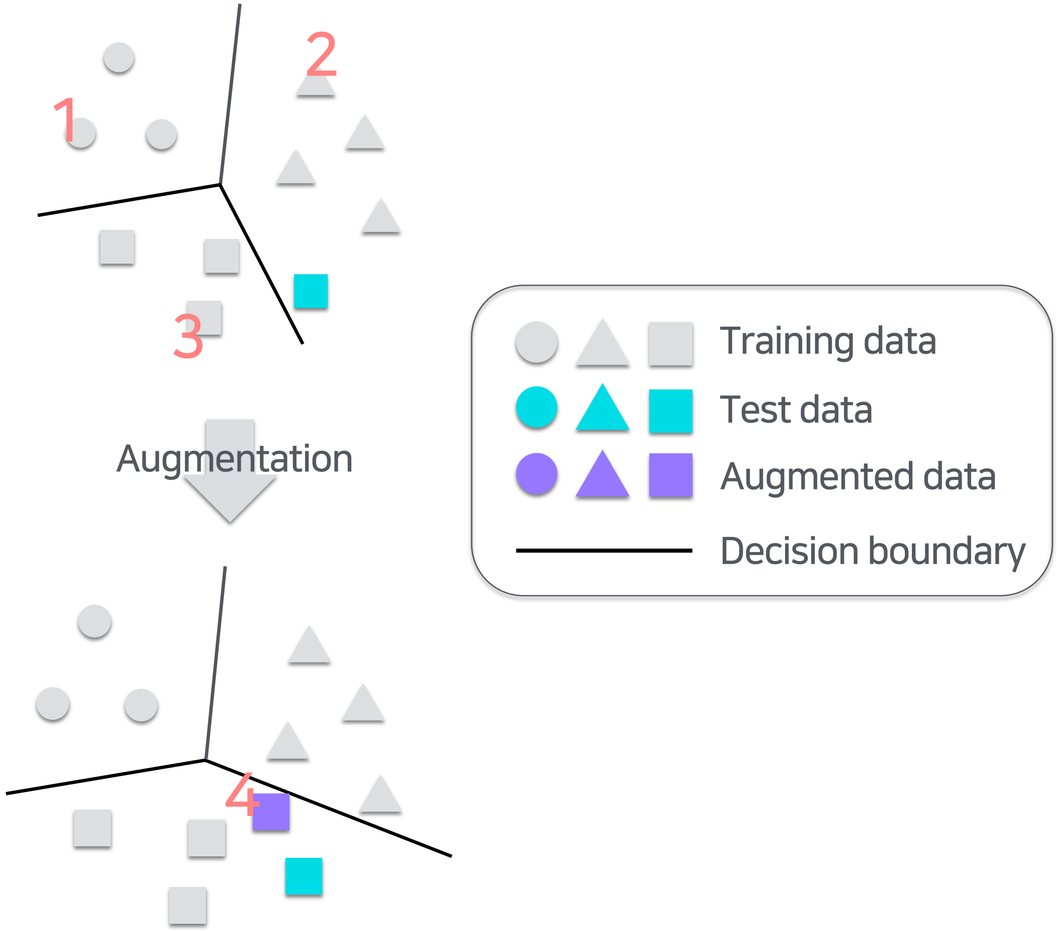
1. 데이터와 레이블이 있는 경우 HyperCLOVA를 통한 증강

다음은 영화리뷰와 감정 예시입니다. 감정은 '긍정', '중립', '부정' 중 하나 입니다.

영화리뷰 : 이 영화는 너무 재미있고 감동적이에요! (긍정) ··· 1
영화리뷰 : 음... 시간 보내기에는 좋지만 꼭 보진 않아도 됩니다 (중립) ··· 2
영화리뷰 : 이걸 왜 돈주고 봐야하지? (부정) ··· 3
영화리뷰 :

HyperCLOVA

재미 없어서 시간도 아깝고 돈도 아깝다 (부정) ··· 4



3.2.1 HyperCLOVA를 이용한 데이터 증강

DEVIEW 2021

HyperCLOVA가 생각하는 레이블 분포는 무엇일까? - HyperMix

다음은 영화리뷰와 감정 예시입니다. 감정은 '긍정', '중립', '부정' 중 하나 입니다.

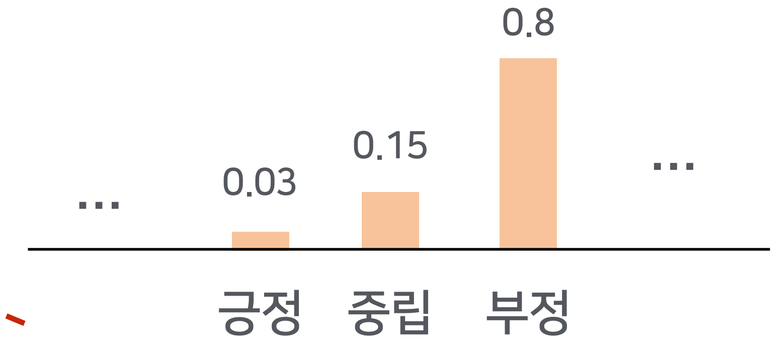
영화리뷰 : 이 영화는 너무 재미있고 감동적이에요! (긍정)
영화리뷰 : 음... 시간 보내기에는 좋지만 꼭 보진 않아도 됩니다 (중립)
영화리뷰 : 이걸 왜 돈주고 봐야하지? (부정)
영화리뷰 :

HyperCLOVA

재미 없어서 시간도 아깝고 돈도 아깝다 (긍정: 3%, 중립: 15%, 부정: 80%)

재미 없어서 시간도 아깝고 돈도 아깝다

HyperCLOVA



3.2.1 HyperCLOVA를 이용한 데이터 증강

DEVIEW 2021

2. 데이터는 있으나 레이블이 없을 경우에는?

1. 레이블 없이 생성하여 SSL을 위한 Unlabeled data로 활용하자

다음의 영화리뷰 예시들을 바탕으로 새로운 영화 리뷰를 작성하세요.

영화리뷰 : 이 영화는 너무 재미있고 감동적이에요!
영화리뷰 : 음... 시간 보내기에는 좋지만 꼭 보진 않아도 됩니다
영화리뷰 : 이걸 왜 돈주고 봐야하지?
영화리뷰 :

HyperCLOVA

와~ 정말 최고에요! 안보면 후회합니다

3.2.1 HyperCLOVA를 이용한 데이터 증강

DEVIEW 2021

2. 데이터는 있으나 레이블이 없을 경우에는?

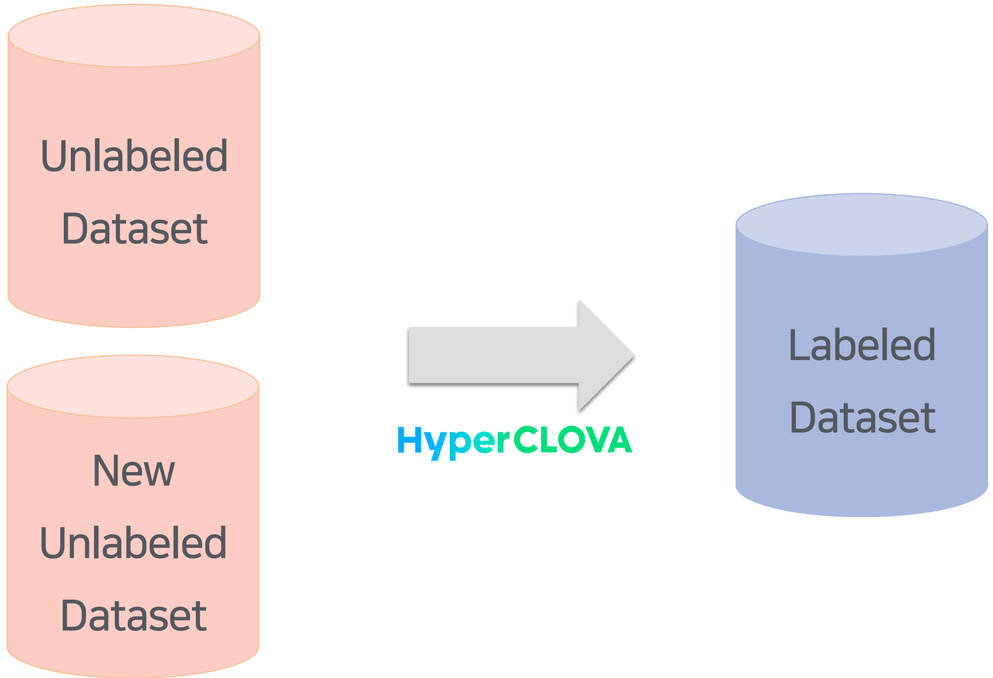
2. 생성 후 HyperCLOVA를 통해 레이블을 만들어주자

다음의 영화리뷰를 '긍정', '중립', '부정' 중 하나의 '감정'으로 분류하세요.

영화리뷰 : 와~ 정말 최고예요! 안보면 후회합니다
감정 :

HyperCLOVA

긍정



3.2.2 HyperCLOVA를 이용한 데이터 생성

DEVIEW 2021

3. 데이터가 있을때는 OK, 하지만 데이터가 하나도 없을 때는?

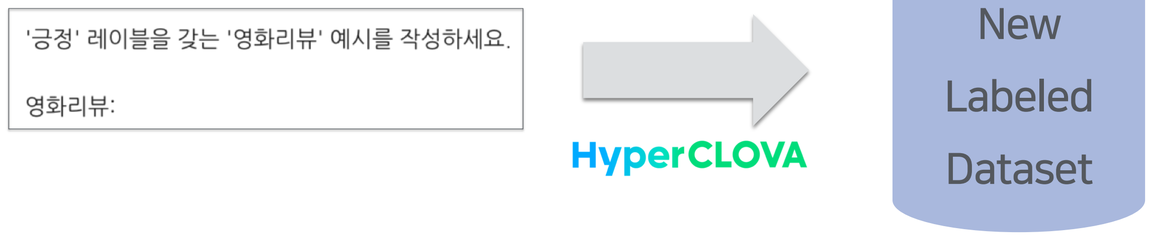
- HyperCLOVA로 알맞는 데이터를 생성하자

'긍정' 레이블을 갖는 '영화리뷰' 예시를 작성하세요.

영화리뷰:

HyperCLOVA

"이 영화는 내가 본 최고의 공포영화다."

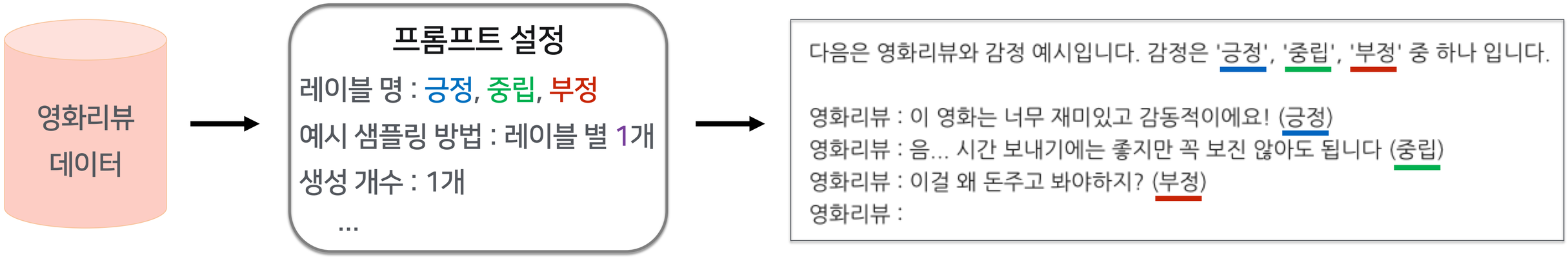


3.2.3 HyperCLOVA의 프롬프트 구성

DEVIEW 2021

HyperCLOVA로 많은걸 할 수 있네 !? 그렇다면 프롬프트는 어떻게?

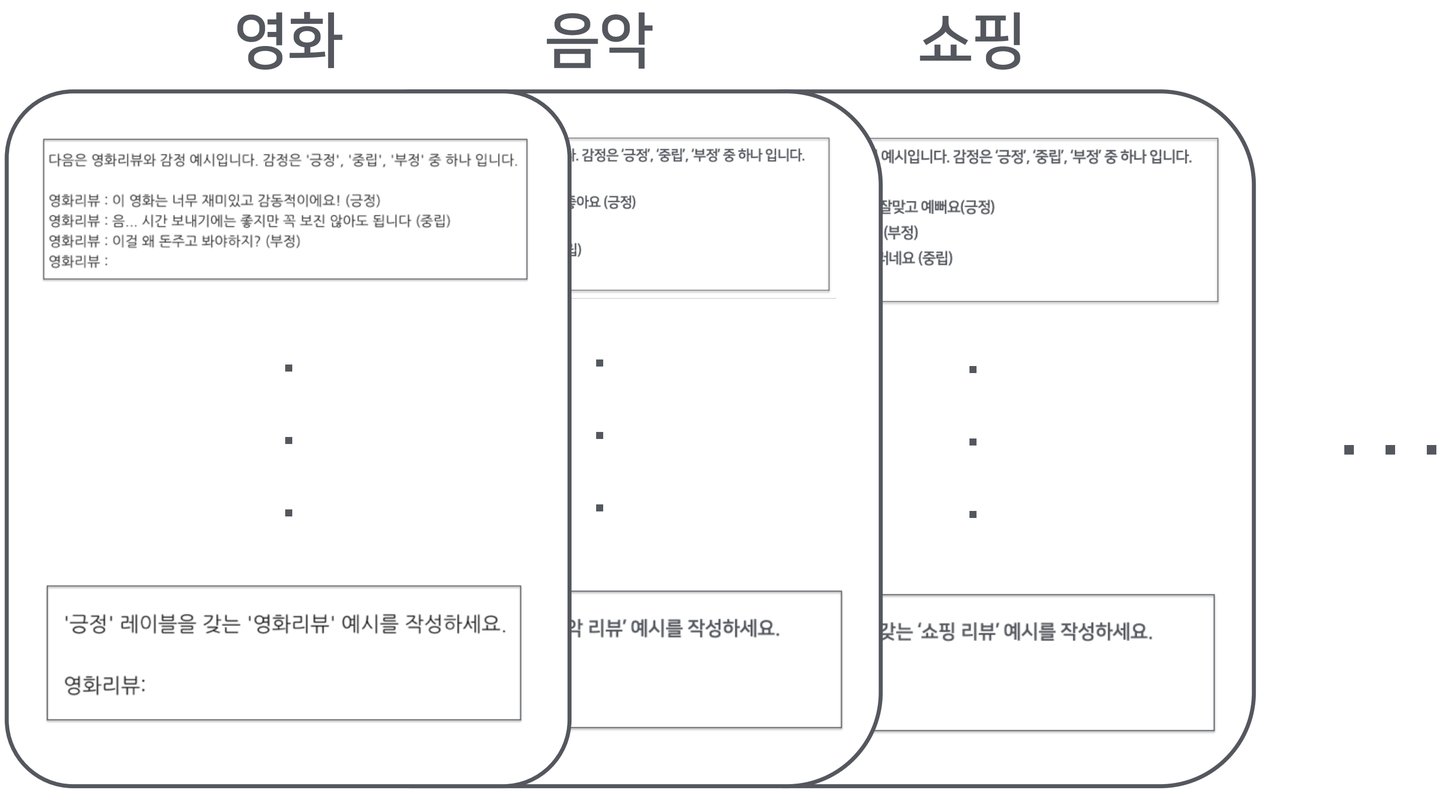
- 원하는 설정을 정의하고 HyperMatch 내부에서 동작하도록 하자



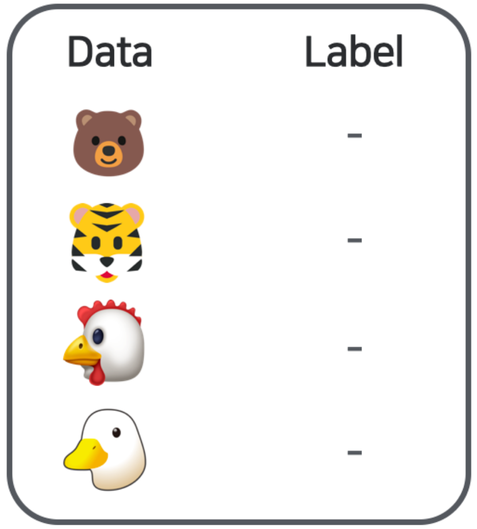
3.2.3 HyperCLOVA의 프롬프트 구성

HyperCLOVA로 많은걸 할 수 있네 !? 그렇다면 프롬프트는 어떻게?

- 좋은 프롬프트는 재활용하자



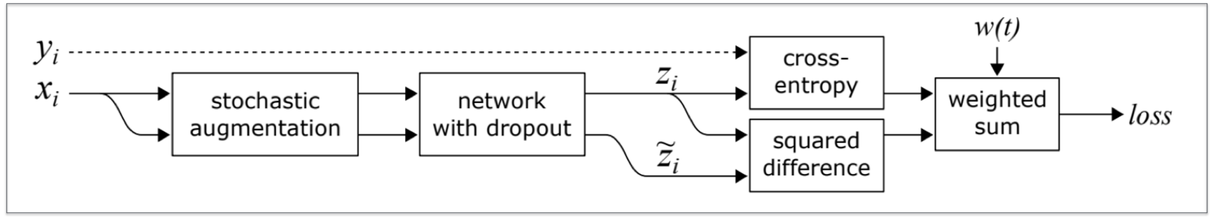
3.3 Unsupervised 에서 Semi-supervised 로



Unsupervised Learning

Consistency regularization

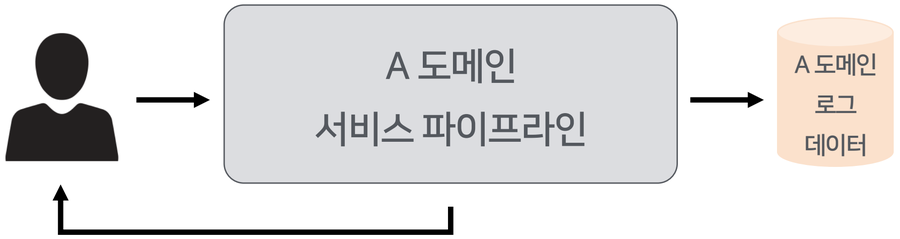
Semi-Supervised Learning



Π-model

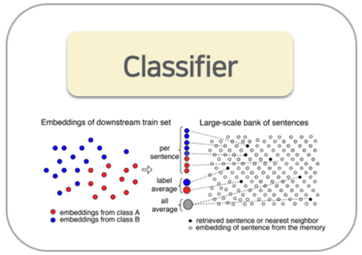
Temporal ensembling for semi-supervised learning (Laine, Samuli, and Timo Aila, 2016)

3.3 Unsupervised 에서 Semi-supervised 로



In-domain Unsupervised Learning

Unlabeled data bank



Data selection algorithm

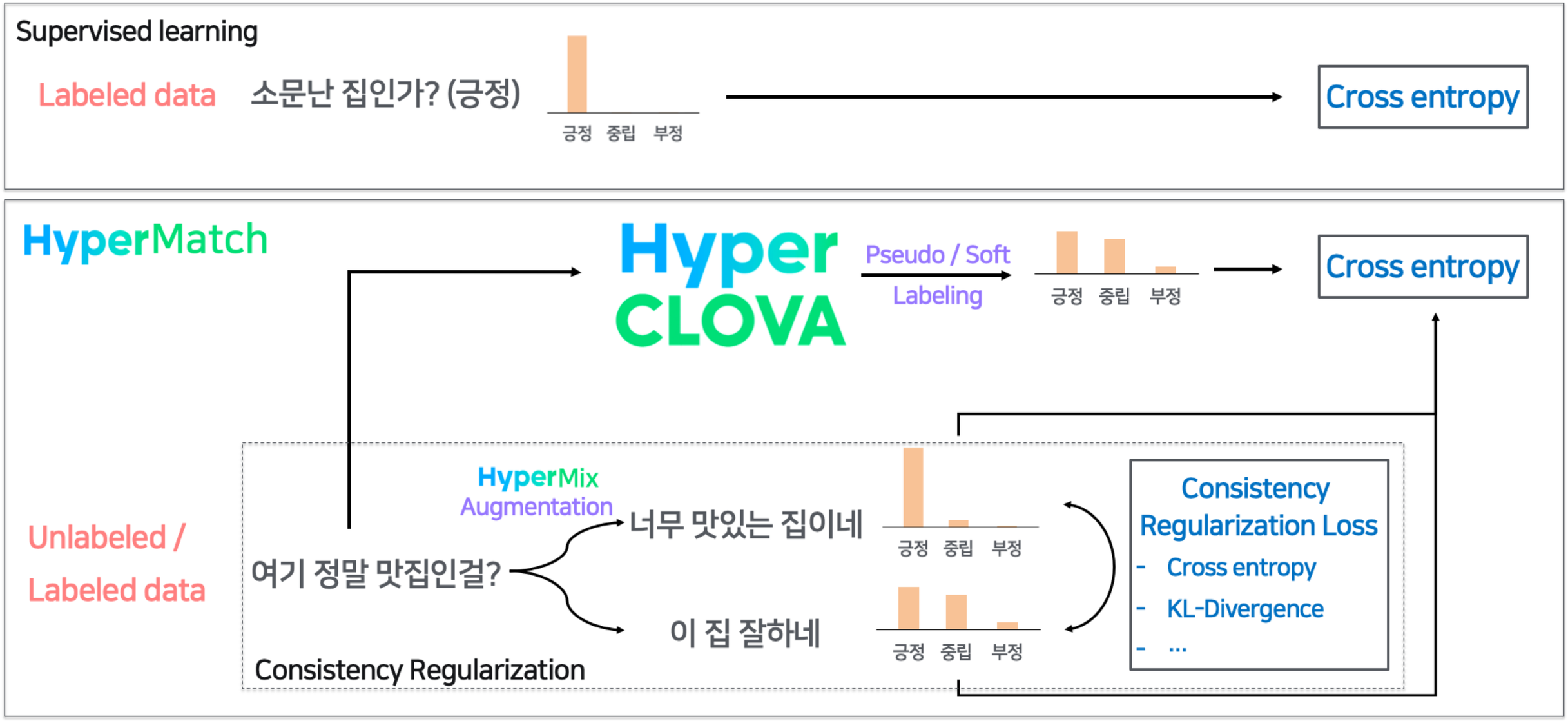
Out-of-domain Unsupervised Learning

HyperCLOVA

In-domain Semi-Supervised Learning (ISSL)

Out-of-domain Semi-Supervised Learning (OSSL)

3.4 HyperCLOVA를 이용한 HyperMatch



HyperCLOVA

+

Semi-Supervised Learning

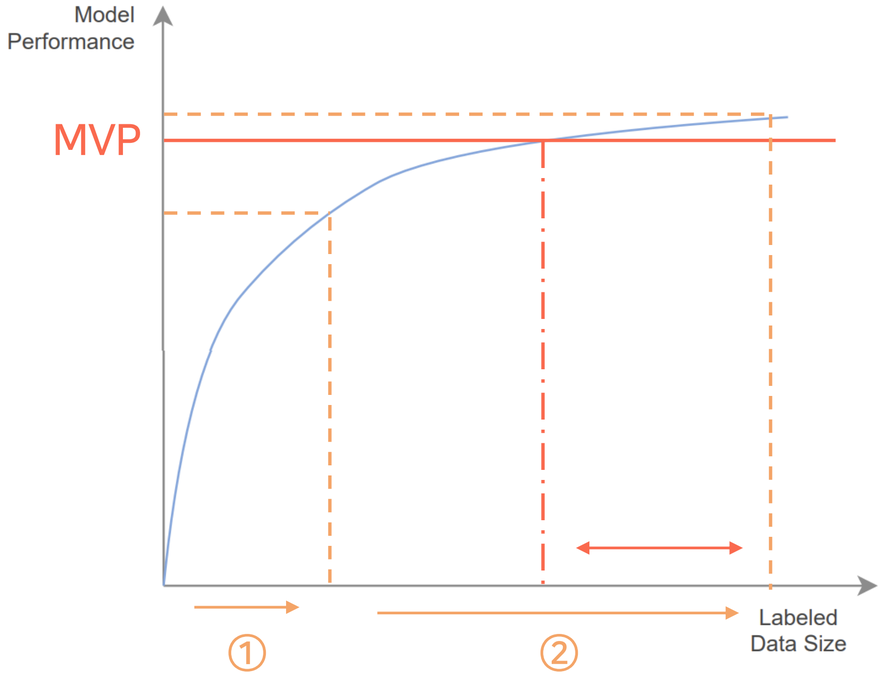
=

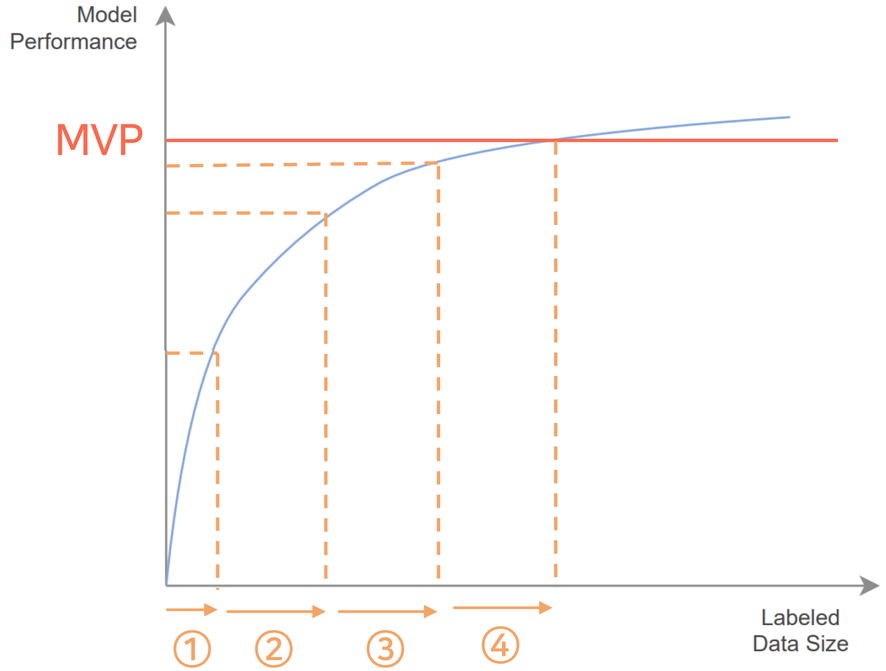
HyperMatch

4. HyperMatch 실전기 및 기대효과

4.1 똑똑하게 수집하는 법

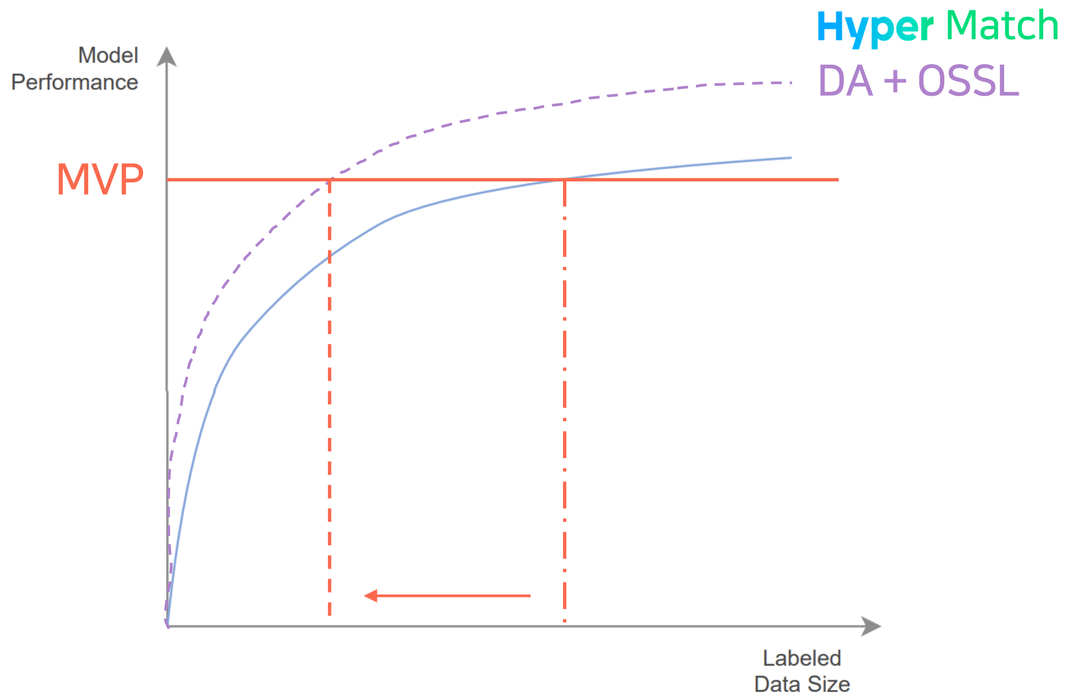
1. 짧은 주기로 데이터 설계 + 데이터 수집 프로세스 사이클 잦은 반복으로 데이터 질 개선 도모 및 MVP 달성 우선시





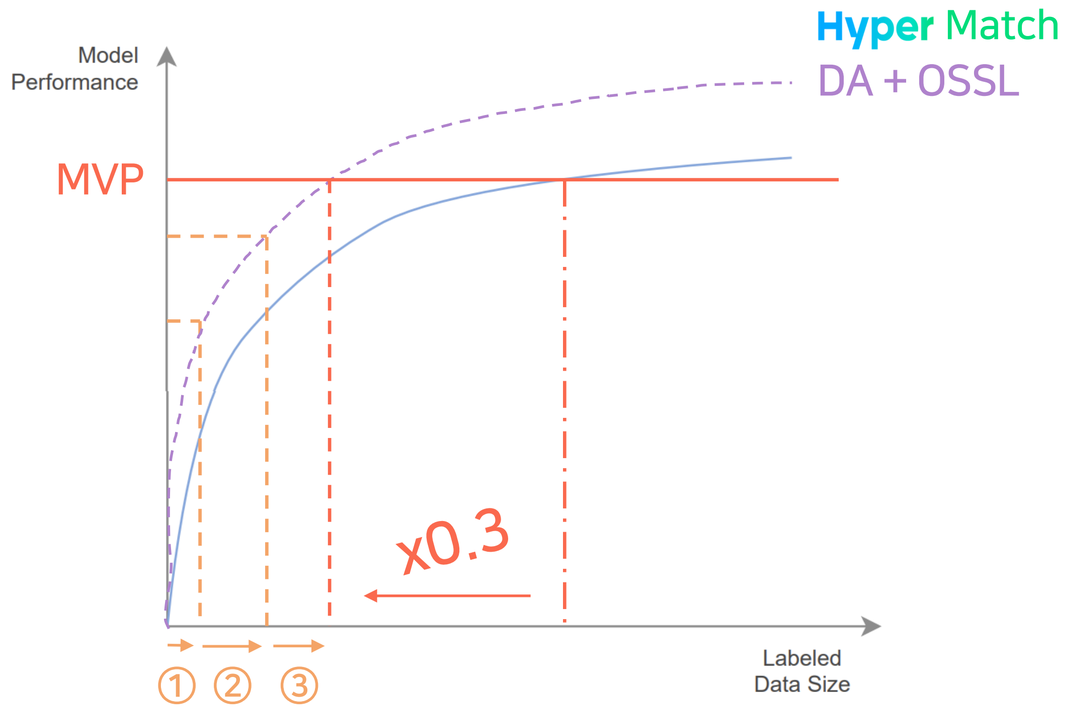
4.1 똑똑하게 수집하는 법

2. HyperMatch 데이터증강(DA) 및 Out-of-domain SSL (OSSL)을 적극활용하여 수집된 레이블 데이터로부터 달성 가능한 모델 성능 극대화



4.1 똑똑하게 수집하는 법

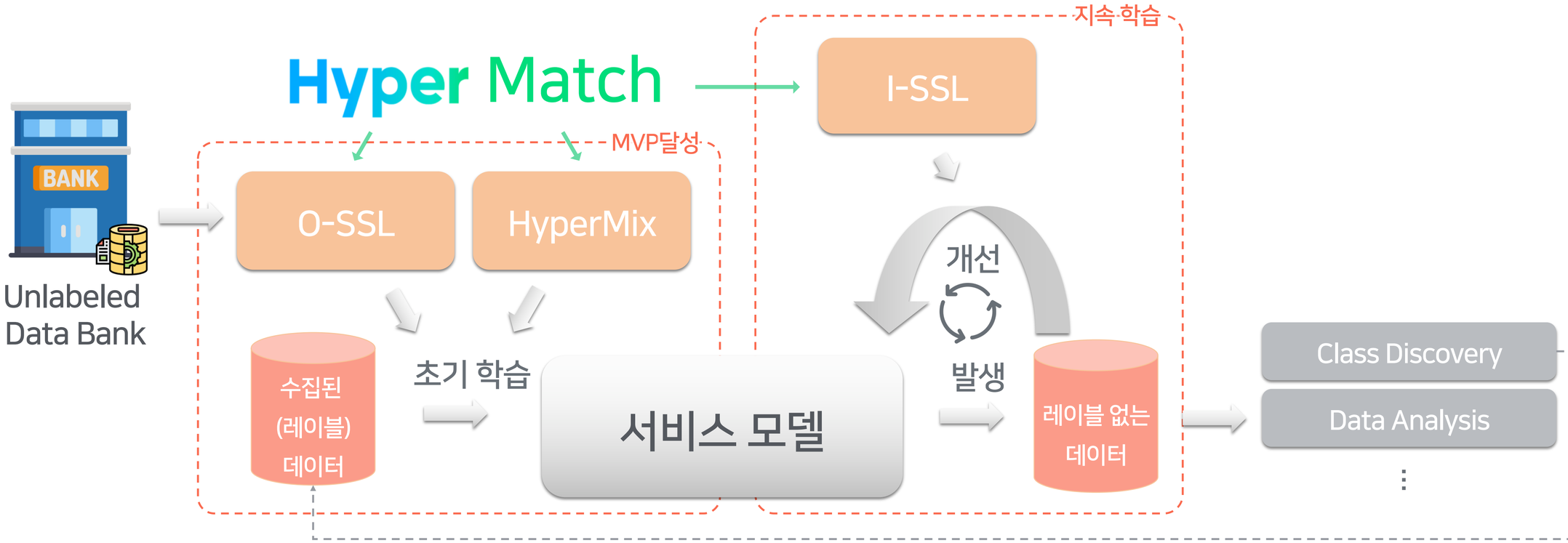
3. DA+OSSL을 수집 사이클과 병행하여 최소 MVP 달성을 위한 레이블 데이터 량 모색-
   - 대화감정분석 모델링 문제에서 MVP 데이터에 필요한 레이블 데이터 70% 감소



4.2 서비스할 수록 똑똑해지는 AI

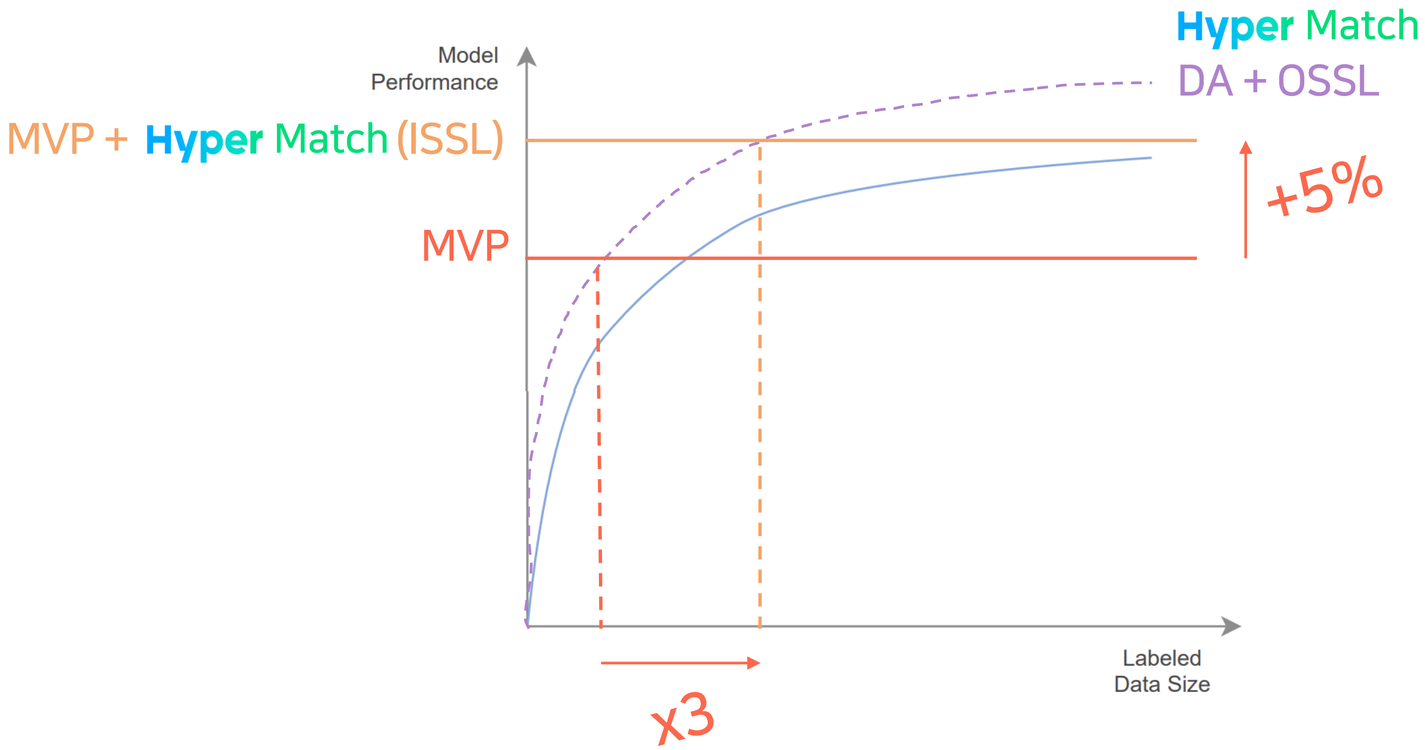
- 모델을 서비스하면서 지속적으로 쌓이는 레이블 없는 데이터를 활용하여 HyperMatch ISSL 적용 및 모델 재학습 개선
- 레이블 없는 데이터에서 Class Discovery (e.g. UNICON*)을 통해 클래스 확장 가능

* “UNICON: LABEL 없이 고객문의 유형을 분석 및 설계해보자”



4.2 서비스할 수록 똑똑해지는 AI

- HyperMatch의 In-domain SSL을 통해 마치 추가적으로 레이블 데이터를 확보한 듯한 성능 개선 효과 달성 (MVP 초과분)
- DA+OSSL의 데이터 감소 효과 (x0.3)와 ISSL의 데이터 증폭 효과 (x3)를 합쳐 x9 효과 달성 가능



4.3 결론 - Takeaways

- Industrial NLP에서 MVP 달성의 중요성, 그리고 MVP달성을 위한 노력을 최소화하는 법
- 데이터증강 + 반지도학습을 통해 labeled + unlabeled 데이터 가치 극대화!
- HyperCLOVA를 이용한 반지도학습 솔루션 연구개발
- 실전 서비스 적용에서 최대 x9 데이터 증폭 효과 검증

Hyper Match

Thank You!