2022 하반기 Sol Project 3기 최종 발표

STS 척도의 표준화를 통한 모델 개선

인턴 서광욱, 황은해
멘토 최진혁 연구원님

Contents

STS란?

- 개념

  유사문장평가(Semantic Textual Similarity)의 약자로서,
  텍스트의 의미적 유사도를 측정하는 Task를 뜻함.

  유사어 사용 여부가 아닌
  두 문장 내용을 이해하고, **두 내용이 유사한지 판단**하는 것이 중요.

- 예시

문장1	너 설마 약속 취소할 생각을 하는 건 아니겠지?
문장2	약속을 취소하는 건 진짜 그릇된 행동이야.

두 문장의 다른 어휘가 많아,
기계는 완전 다른 내용이라고 판단하기 쉬움.

하지만 **두 문장 모두 핵심 내용은**
'약속을 취소하면 안된다.'를 말하고 있음.

STS란?

- STS 점수 척도

  두 문장간의 **의미 동등성을 점수로 표현**

  0점(연관없음) ~ 5점(연관있음)
  3점 이상이면 연관 있음으로 판단

점수	내용
5	두 문장은 중요한 내용과 중요하지 않은 내용이 동등하다.
4	두 문장은 거의 동일하다. 일부 중요하지 않은 내용은 다르다.
3	두 문장은 대략 같다. 중요한 콘텐츠는 서로 비슷하지만 중요하지 않은 콘텐츠의 차이도 무시할 수 없다.
2	두 문장은 동등하지 않다. 중요한 콘텐츠는 서로 비슷하지 않고 일부 중요하지 않은 콘텐츠만 공유한다.
1	두 문장은 동등하지 않다. 중요한 내용과 중요하지 않은 내용이 서로 비슷하지 않다. 두 문장은 주제만 공유한다.
0	두 문장은 동등하지 않다. 중요하지 않은 내용과 심지어 주제도 공유하지 않고 있다.

- STS dataset 구조 간단하게 살펴보기

guid	source	sentence1	sentence2	label
klue-sts-v1_dev_00000	airbnb-rtt	무엇보다도 호스트분들이 너무 친절하셨습니다.	무엇보다도, 호스트들은 매우 친절했습니다.	4.9
klue-sts-v1_dev_00001	airbnb-sampled	주요 관광지 모두 걸어서 이동가능합니다.	위치는 피렌체 중심가까지 걸어서 이동 가능합니다.	1.4
klue-sts-v1_dev_00002	policy-sampled	학생들의 균형 있는 영어능력을 향상시킬 수 있는 학교 수업을 유도하기 위해 2018학년도 수능부터 도입된 영어 영역 절대평가는 올해도 유지한다.	영어 영역의 경우 학생들이 한글 해석본을 암기하는 문제를 해소하기 위해 2016학년도부터 적용했던 EBS 연계 방식을 올해도 유지한다.	1.3
klue-sts-v1_dev_00003	airbnb-rtt	다만, 도로와 인접해서 거리의 소음이 들려요.	하지만, 길과 가깝기 때문에 거리의 소음을 들을 수 있습니다.	3.7
klue-sts-v1_dev_00004	paraKQC-para	형이 다시 캐나다 들어가야 하니 가족모임 일정은 바꾸지 마세요.	가족 모임 일정은 바꾸지 말도록 하십시오.	2.5

문장 도메인 | 비교할 문장 쌍 | 라벨점수

점수 척도 및 dataset 출처 : KLUE: Korean Language Understanding Evaluation (2021)

현존하는 한국어 STS benchmark dataset은 2가지
⇒ KLUE, korSTS

KLUE

[데이터 개수]
13,224개

[생성 방법]
Airbnb 리뷰, policy(공식 뉴스),
paraKQC(smart home queries)
+ RTT로 문장쌍 추가 생성

[특징]

- ✓ 원문이 한국어로 된 데이터
- ✓ 한국인 작업자들이 직접 점수를 라벨링함

✓ 연구에 가장 적합한 데이터로 KLUE 채택

korSTS

[데이터 개수]
약 8,700개

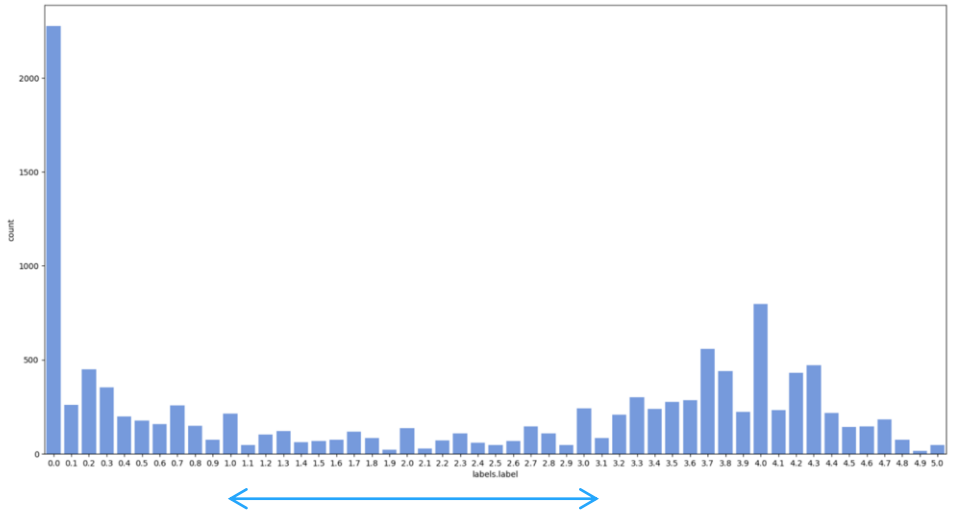
[생성 방법]
영어로 된 STS benchmark를 재번역

[문제점]

- ✓ 영어 문장을 그대로 번역해서 생성함
- ✓ 번역으로 인해 발생하는 근본적 문제가 그대로 존재
- ✓ 원 데이터의 라벨 점수를 그대로 사용

01 균일하지 않은 점수 분포

KLUE train set 분포



1점 ~ 3점 사이의 학습 데이터
매우 부족

해당 데이터를 통해 학습을 진행하면,
데이터가 부족한 구간은 **잘못된 학습을 할 가능성 ↑**

02 잘못된 번역으로 생성된 데이터들

① 문맥상 이해하기 힘든 단어의 등장

문장1	집도 호스트님도 뭐하나 빠짐없이 베스트였습니다.
문장2	집과 숙주가 최고였어요.

② 비 완전한 문장

문장1	지난 2005년 노무현 전 대통령이 친환경 수소경제 구현을 위한 마스터플랜을 수립한 이후 14년여 만이다.
문장2	2005년에 겨우 14년 이후에 대한 종합 계획을 노무현 전 대통령은 녹색 수소 경제.

.
.

이외 잘못된 번역으로 생성된 다수의 문장들 존재

RTT로 생성된 데이터 중
잘못된 번역으로 인해 **라벨 작업 난이도 ↑**
데이터의 신뢰도 ↓

03 주관적 판단이 개입되는 점수 척도

- 모호한 STS 점수 척도

점수	내용
4	두 문장은 거의 동일하다. 일부 중요하지 않은 내용은 다르다.
3	두 문장은 대략 같다. 중요한 콘텐츠는 서로 비슷하지만 중요하지 않은 콘텐츠의 차이도 무시할 수 없다.
2	두 문장은 동등하지 않다. 중요한 콘텐츠는 서로 비슷하지 않고 일부 중요하지 않은 콘텐츠만 공유한다.

대략 같다와
거의 동일이 어떤 차이죠?

중요한 내용과 중요하지
않은 내용을 어떻게 구별하죠?

- 비슷한 유형이지만 다른 점수 분포

S1	S2	label
수건에서 **쾌쾌한** 냄새가 난 것과 집 문을 열기 어려웠던점, 그리고 침대가 불편했습니다.	수건의 **신선한** 냄새, 집의 문을 여는 어려움, 그리고 침대는 불편했습니다.	2.8
위치, 시설, **베란다 뷰** 정말정말 모든게 완벽합니다.	숙소의 위치,시설, **청결도** 모든게 완벽했다.	3.4

0.6 점의 차이 발생

중요한 내용과 유사한 정도를
개인의 주관으로 판단할 가능성 ↑

04 오류율이 가장 높은 2.0 ~ 3.5 구간

- 선행 연구 검토

유사 문장 말뭉치 분석을 통한 유사도 인식에 관한 연구 (2021, 어문연구학회)

… 2.4점에서 3점 사이에 불일치 문장이 거의 분포한다.
이는 유사 문장의 경계, 즉 3점 주변 문장 쌍들의 쌍방
함의 관계를 판단하기가 어렵다는 것이다. …. (생략)

- KLUE train set으로 학습한 구간 별 binary 판단 표

bin	TP	TN	FP	FN	accuracy
1.3 – 1.9	0	37	6	0	0.860465
2.0 – 2.5	0	31	24	0	0.563636
2.6 – 3.0	9	8	34	2	0.320755
3.1 – 3.5	48	0	0	7	0.872727
3.6 – 4.0	53	0	0	2	0.963636
4.1 – 4.5	55	0	0	0	1.0
4.6 – 5.0	44	0	0	0	1.0

유독 낮은 2.0 – 3.5 구간의 accuracy

연구의 **핵심 과정은 2가지 단계**로 나뉨

STS 데이터 척도 재정립 및 재라벨링

척도 재정립

데이터 구간 산정

데이터 전처리

재라벨링

데이터 보완

변경된 데이터의 유효성 검증 및 모델 성능 개선

모델 선정

모델 실험 설계

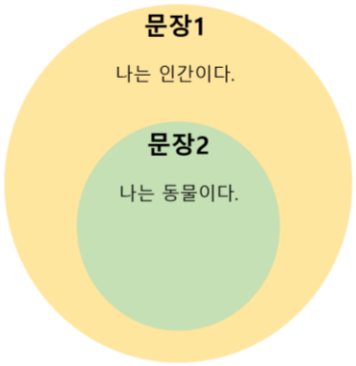
모델 성능 개선

03 연구 1 : STS 데이터 척도 재정립 및 재라벨링

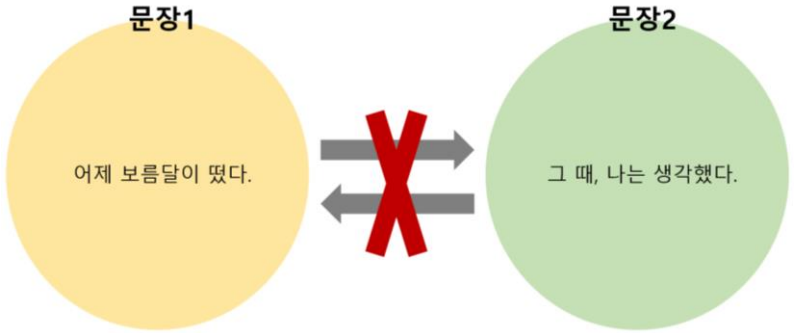
(1) STS 라벨 척도 재정립

- 상호함의

- 일방함의



- 비함의



- 새로 정립한 STS 기준과 기존 KLUE 작업 가이드라인

점수	새로 정립한 기준	기존 KLUE 작업 가이드라인
4		핵심 내용의 의미는 포함되어 있다. 보조 내용이 약간 포함되어 있지 않다.
3.5	핵심맥락 상호함의(2) + 보조맥락 상호함의(2)	
3	핵심맥락 상호함의(2) + 보조맥락 일방함의(1) 핵심맥락 일방함의(1) + 보조맥락 상호함의(2)	핵심 내용의 의미가 약간 달라진다. 또는 핵심 내용은 그대로지만, 보조 내용의 큰 의미 차이가 있다.
2.5	핵심맥락 상호함의(2) + 보조맥락 비함의(0) 핵심맥락 일방함의(1) + 보조맥락 일방함의(1) 핵심맥락 일방함의(1) + 보조맥락 비함의(0) 핵심맥락 비함의(0) + 병렬 정보 많은경우	
2	핵심맥락 비함의(0)	핵심 내용은 달라진다. 하지만, 핵심 내용 주제 일부/핵심 내용 수식어에서 공통 주제를 많이 공유한다.

(2) 문제 있는 데이터 구간 선정

- 2.0 ~ 3.5 구간의 낮은 정확도

bin	TP	TN	FP	FN	accuracy
1.3 – 1.9	0	37	6	0	0.860465
2.0 – 2.5	0	31	24	0	0.563636
2.6 – 3.0	9	8	34	2	0.320755
3.1 – 3.5	48	0	0	7	0.872727
3.6 – 4.0	53	0	0	2	0.963636
4.1 – 4.5	55	0	0	0	1.0
4.6 – 5.0	44	0	0	0	1.0

모델이 binary 판단을 틀리는 데이터
&
작업자들이 발견한 오라벨 데이터

모두 2.0 ~ 3.5 구간에 집중됨

해당 구간의 모호한 라벨 기준

평균을 통한 라벨 점수 도출

- 초기 라벨링 계획 무산

초기계획

크라우드소싱을 통하여
재정립한 척도를 기반으로 라벨링 작업 진행

-> 정규화된 점수를 통해 재라벨링할 계획

시간적 한계 발생

기존의 계획 무산

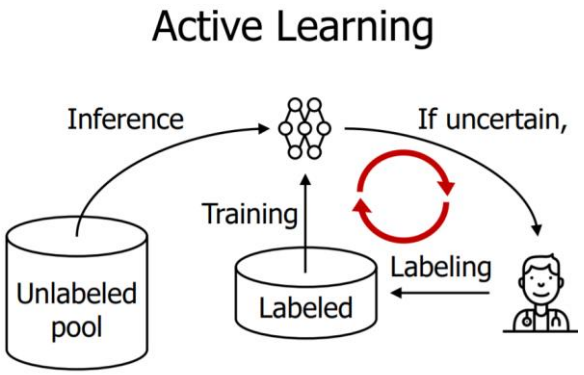
결론

모든 데이터를 재라벨링 하기는 힘듦.

2.0 – 3.5 구간 중 모델에게
잘못된 정답을 전달하는
데이터들을 선별할 필요

(2) 문제 있는 데이터 구간 선정 - 412개의 재라벨 대상 train데이터 선정방법

- Active learning에서 활용되는 방식을 응용

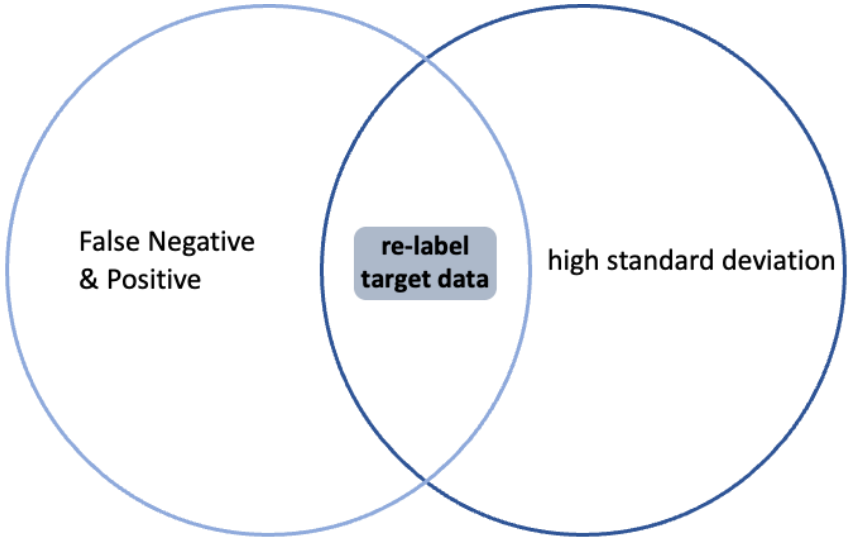


불확실성이 높은 데이터들을 추출하고
해당 데이터에 대해서만 annotation을 진행하는 방식

모델이 스스로 학습을 하는 과정에서 예측 불확실성이
높은 데이터를 추출하여 재라벨을 요구

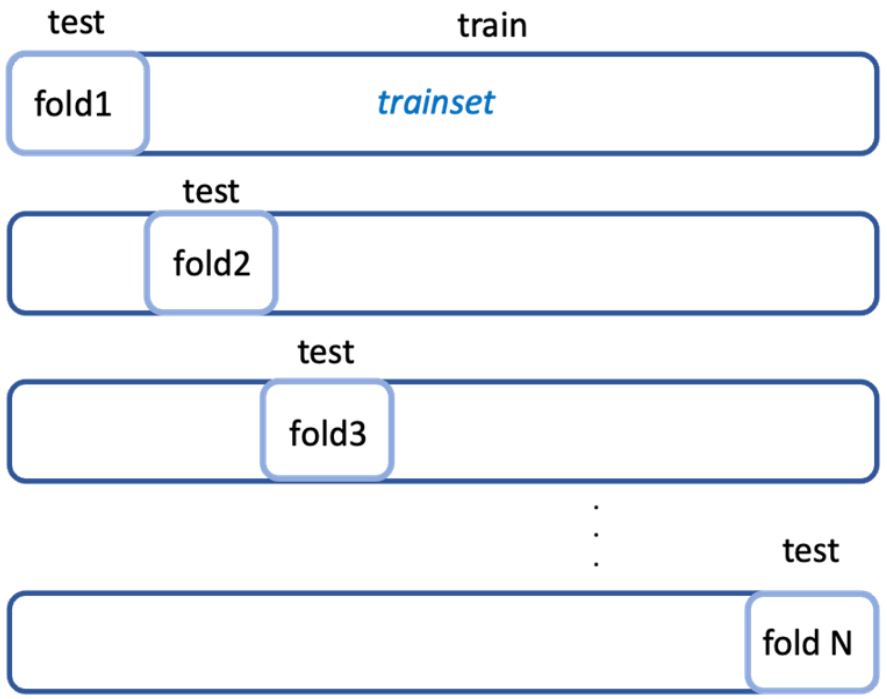
- 재라벨 대상 데이터

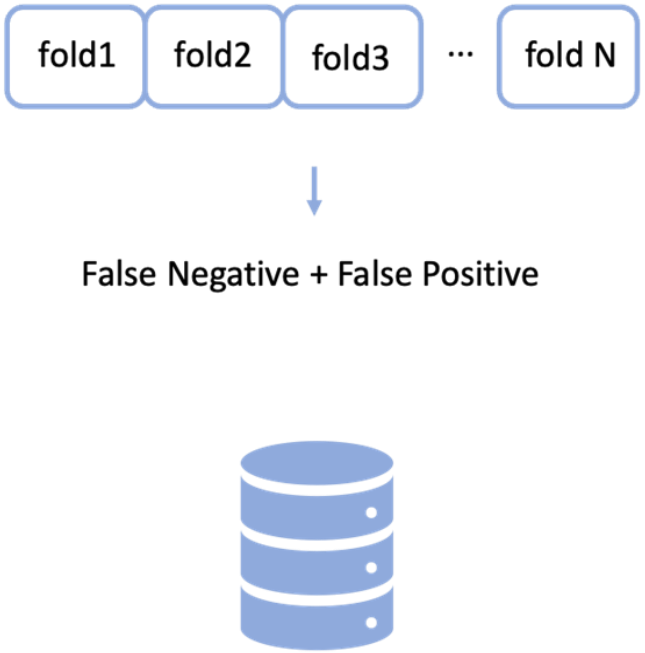
: 학습된 모델이 binary 판단을 틀리는 데이터
& 예측불확실성이 높은 데이터



(2) 문제 있는 데이터 구간 선정 - 412개의 재라벨 대상 train데이터 선정방법

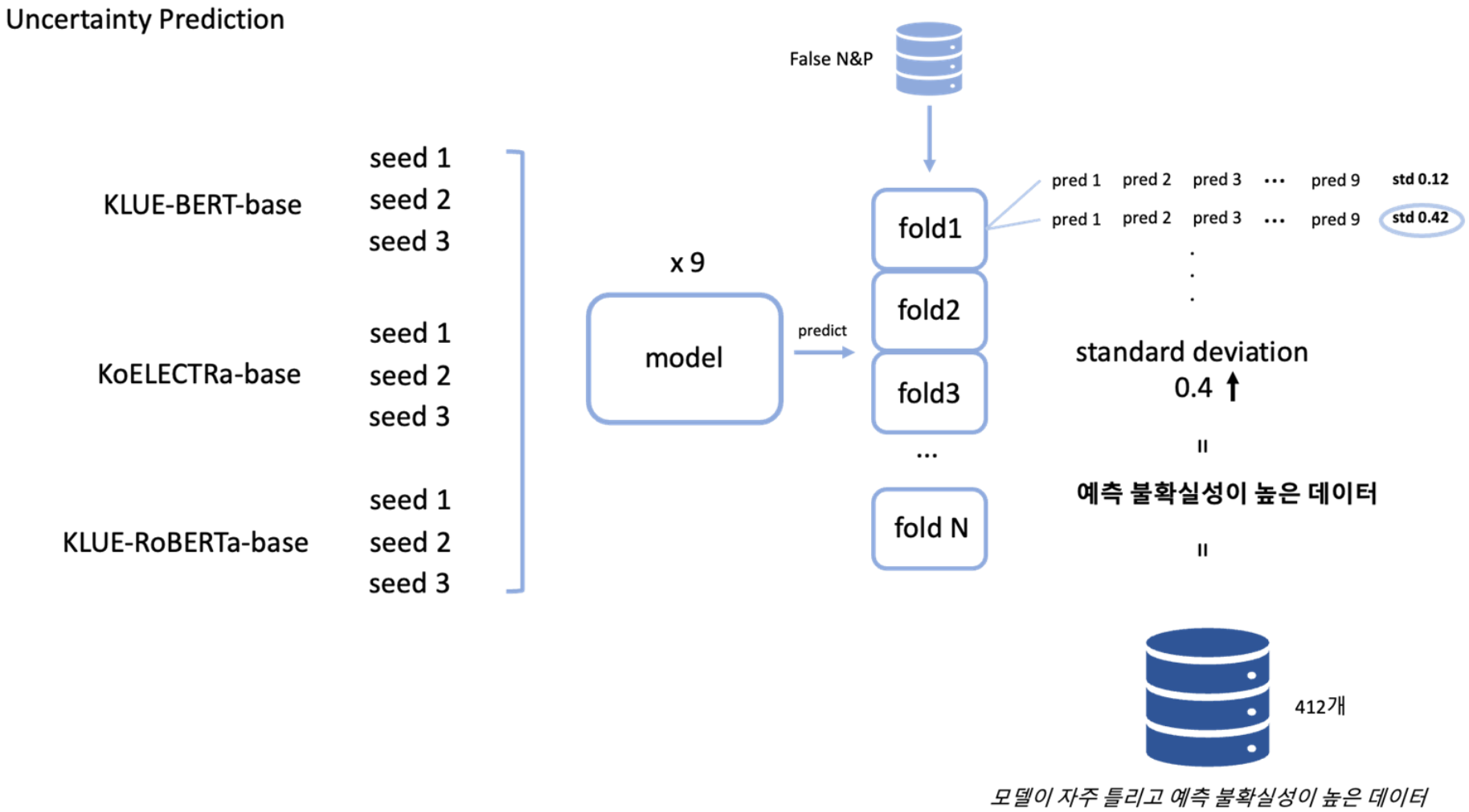
Cross – Validation





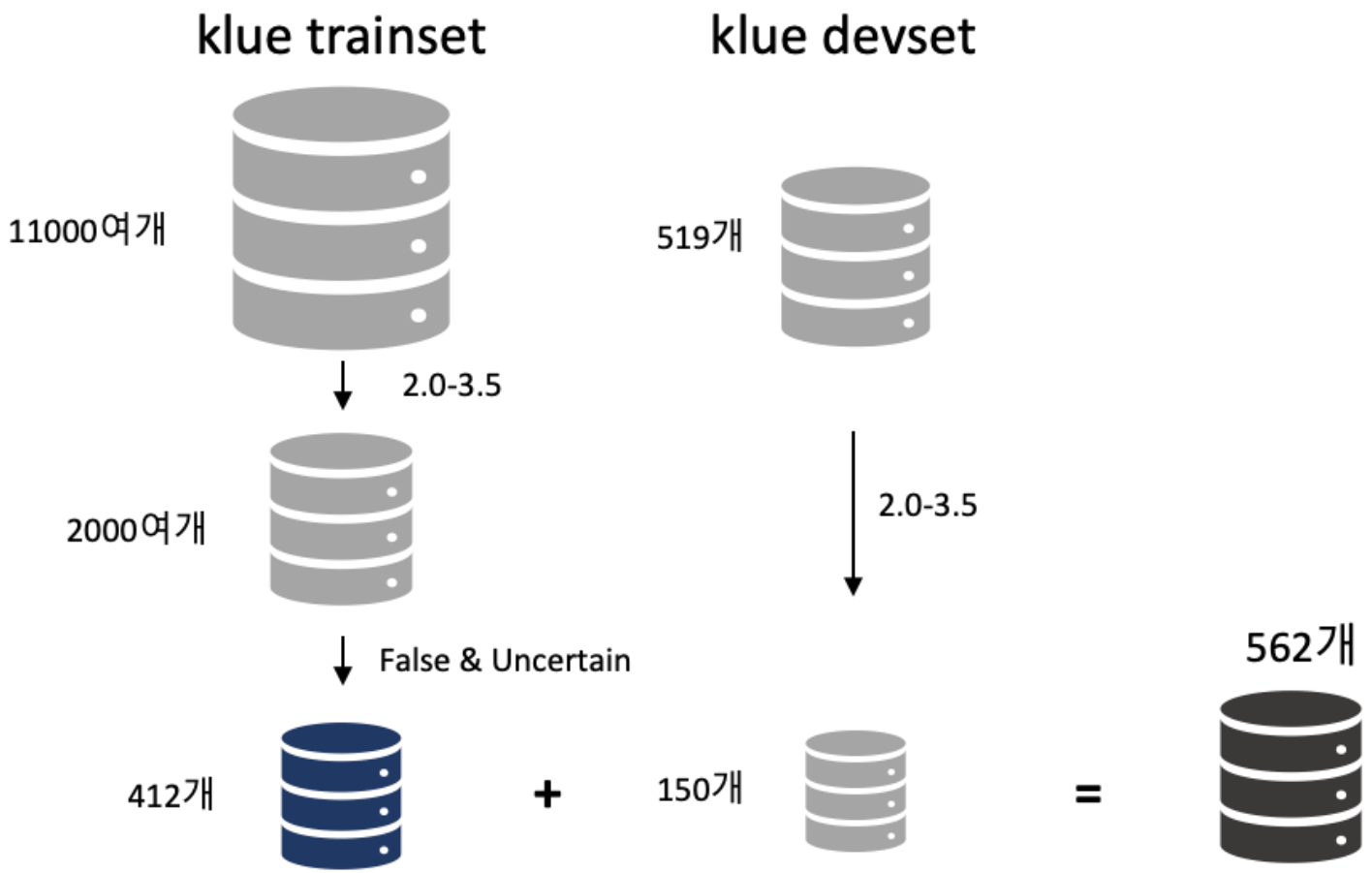
* each fold = 2.0-3.5 label

(2) 문제 있는 데이터 구간 선정 - 412개의 재라벨 대상 train데이터 선정방법



모델이 자주 틀리고 예측 불확실성이 높은 데이터

(2) 문제 있는 데이터 구간 선정

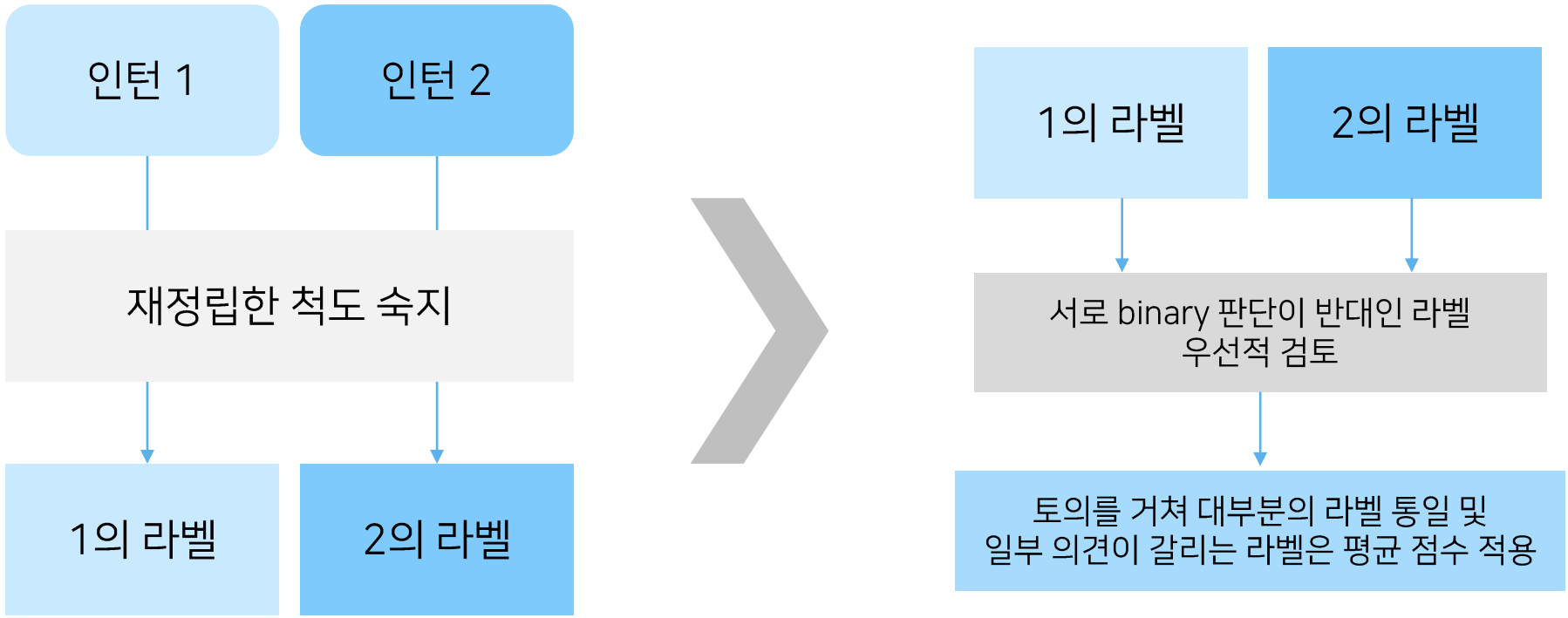


(3) 데이터 전처리

source	번역 오류가 존재하는 데이터들		재 번역한 데이터들
airbnb-rtt	호스트와 연락이 잘 닿는 점이 좋습니다.	>	호스트와 연락이 잘 되어서 좋습니다.
airbnb-rtt	특히 바로 옆방 소리는 대화를 알아들을 수 있을 정도로 잘 들렸습니다.	>	특히 바로 옆 방의 소리가 너무 좋아서 대화 내용을 이해할 수 있었습니다.
policy-rtt	소상공인지원센터에 '정책자금 확인서' 온라인 발급 시스템 도입과 센터 내 번호표 발급기 비치 등 고객 편의 제도를 도입한다.	>	소상공인지원센터에 '정책자금 확인'을 위한 온라인 발급 시스템을 도입하고 센터에 번호표 발급기를 설치하는 등 고객 편의 시스템을 도입합니다.
airbnb-rtt	오븐이나 인덕션 사용은 유투브에서 찾으시면 나와요.	>	오븐이나 인덕션은 유튜브에서 찾을 수 있습니다.
airbnb-rtt	숙박시설에서 바라본 경치가 아름답고 희미합니다. 저는 심지어 파밀리아 대성당도 볼 수 있어요.	>	기숙사에서 보는 경치는 멋지고 여러분은 파밀리아 대성당도 볼 수 있습니다.
airbnb-rtt	샤워하는 물이 좀 불편했어요.	>	샤워실 물이 안 나와서 조금 불편했어요.
policy-rtt	무엇이 다자녀 가정을 더 좋게 만들까요?	>	무엇이 다자녀 가정을 더 낫게 만들까요?
policy-rtt	온라인 커뮤니케이션 태도는 게임을 타이핑함으로써 자연스럽게 배웁니다.	>	타이핑 게임을 통해 자연스럽게 배울 수 있는 온라인 커뮤니케이션 태도입니다.

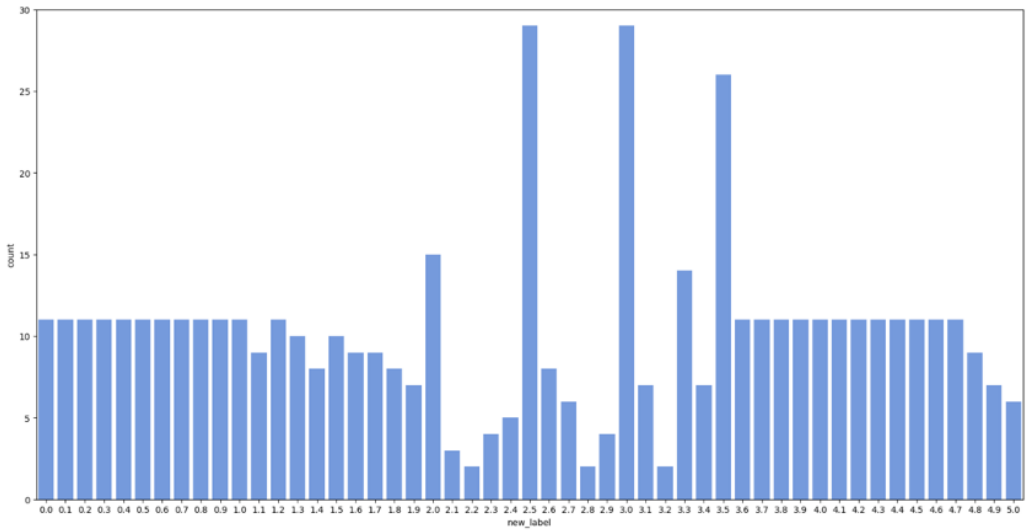
총 train set 36개 + devset 6개의 오번역 데이터들을 올바른 문맥을 가진 데이터로 재번역해줌

(4) 재라벨링

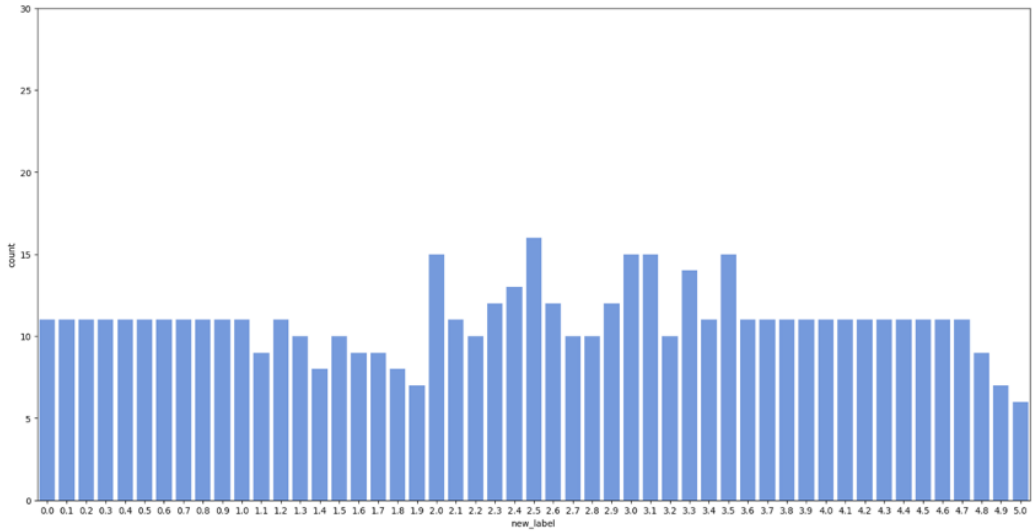

인턴 1
인턴 2
재정립한 척도 숙지
1의 라벨
2의 라벨
1의 라벨
2의 라벨
서로 binary 판단이 반대인 라벨
우선적 검토
토의를 거쳐 대부분의 라벨 통일 및
일부 의견이 갈리는 라벨은 평균 점수 적용

(5) 재라벨링 이후 데이터 보완 - devset 분포 맞춰주기

[재라벨 후 분포가 불균형한 devset]



[분포를 맞춰준 devset]



3.0미만의 낮은 유사도라벨은 **워드오버랩이 높은 데이터**들을 이용하고,
3.0이상의 높은 유사도라벨은 **워드오버랩이 낮은 데이터**들을 이용해

우선적으로 새로운 devset에 포함

04 연구 2 : 변경된 데이터의 유효성검증, 모델 성능개선

(1) 모델 선정

- cross-encoder vs bi-encoder

Model trained with	**Pearson R**	**F1 score**	**MSE**
Bi-encoder / KLUE-RoBERTa-base			
original klue-trainset	87.60	80.03	-
new trainset	88.08	80.65	-
Cross-encoder / KLUE-RoBERTa-base			
original klue-trainset	90.55	81.38	0.41
new trainset	**91.14**	**82.98**	**0.37**

* trained with 10 random_seed and 3epochs

1. Fine-tuning의 성능

   cross-encoder > bi-encoder

2. cross-encoder의 단점

   [연산방식]
   입력문장 pair 전체를 나머지 문장 pair들과 비교하며 연산

   [연산시간]
   각 문장 임베딩을 학습하는 bi-encoder보다 오래 걸림

3. 결론

   [bi-encoder]
   문장 유사도를 이용한 검색 시스템, 랭크 방식 등의 활용과정에 좀 더 가벼운 bi-encoder가 적합

   [cross-encoder]
   각 새로운 데이터 셋을 제작하고 **데이터 셋의 변화에 따른 성능을 평가하는 과제**에서는 cross-encoder로 **정밀하게 성능을 측정**하는 것이 적함

⇒ Cross – encoder 방식 채택

04 연구 2 : 변경된 데이터의 유효성검증, 모델 성능개선

(1) 모델 선정

- BERT / ELECTRA / RoBERTa

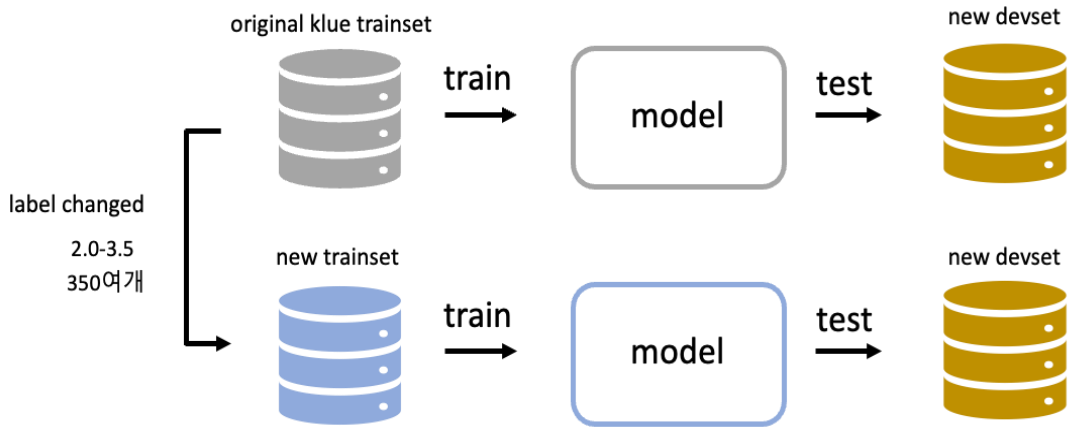
Model trained with	**Pearson R**	**F1 score**	**MSE**
KLUE-BERT-base			
original klue-trainset	88.32	81.09	0.44
new trainset	89.62	80.30	0.41
KoELECTRa-base			
original klue-trainset	88.72	81.12	0.40
new trainset	90.54	**83.12**	**0.37**
KLUE-RoBERTa-base			
original klue-trainset	90.55	81.38	0.41
new trainset	**91.14**	82.98	**0.37**

* trained with 10 random_seed and 3epochs

⇒ 성능비교를 통해 가장 높은 성능을 보이는 roberta로 전부 실험 진행

(2) 변경된 데이터의 유효성 검증

- 재라벨에 따른 모델 성능 변화 측정



- 우리의 가설

1. 새롭게 라벨링한 데이터들을 일관된 기준으로 제대로 라벨링을 했다면 새로운 set으로 학습한 모델이 더 좋은 성능을 보일 것
2. 적은 수의 데이터 **변화지만 모델이 가지는 불확실성이 높은 데이터들을 선별**했기 때문에 어느 정도의 효과가 있을 것을 기대
3. 특히 **타겟으로 잡은 2.0-3.5 구간** (고쳐준 라벨이 포함된 구간)의 **성능이 더 큰 폭으로 향상될 것**

- training arguments

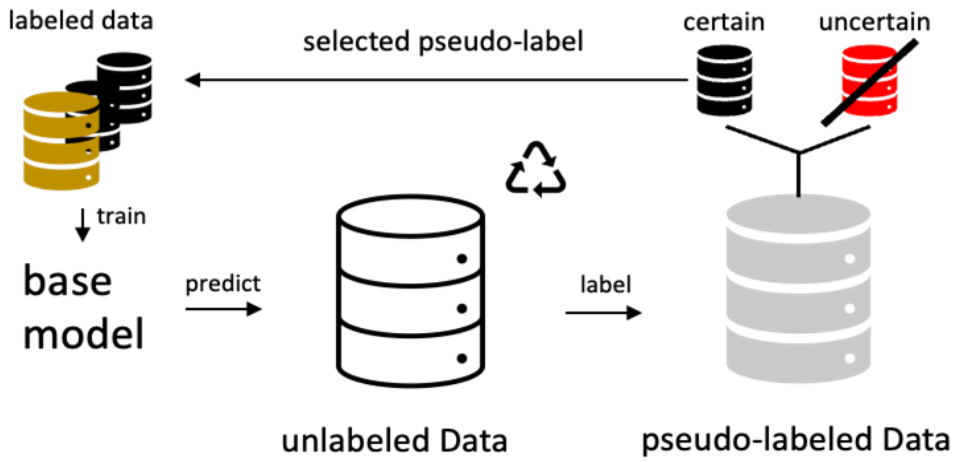
num_epochs	*3*
batch_size	*32*
learning_rate	*5e-5*
evaluation step	*len(train) * 0.1*
warmup_steps	*len(train) * num_epochs / train_batch_size * 0.1*
results	*average of 10 random seed*

- GPU Quadro P4000 x 2

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 418.88       Driver Version: 418.88       C
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage |
|===============================+======================+======================+
|   0  Quadro P4000        On   | 00000000:3B:00.0 Off |
| 46%   31C    P8     5W / 105W |   6254MiB /  8119MiB |
+-------------------------------+----------------------+----------------------+
|   1  Quadro P4000        On   | 00000000:D8:00.0 Off |
| 46%   34C    P8     5W / 105W |    950MiB /  8119MiB |
+-------------------------------+----------------------+----------------------+
```

(3) 성능 개선 방안 : Pseudo labeling data를 통한 semi-supervised learning

- wrapper method - self_training



👀 Pseudo Label을 통한 **Self-Training방식**이란?

: 학습에 사용할 **라벨데이터의 불균형을 해소하는 것**에 효과적인 준지도학습기법.

: 부족한 라벨 데이터에 unlabeled data를 활용해 증강하는 방식.

✓ 증강과정에서 **pseudo-label 중 확실성이 높은 라벨**을 고르는 과정이 필요.

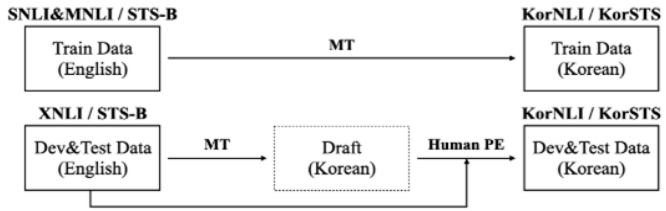
✓ self-training은 **이상치에 민감** => classification에서는 지속적으로 모델을 업데이트하여 **확률값이 높은 것들을 선택**하여 가져감.

(4) 모델 성능 개선방안 모색 - unlabeled data

[korSTS - test]

KorNLI and KorSTS:
New Benchmark Datasets for Korean Natural Language Understanding

Jiyeon Ham*, Yo Joong Choe*, Kyubyong Park*, Ilji Choi, Hyungjoon Soh
Kakao Brain
{jiyeon.ham,yj.choe,kyubyong.park,ilji.choi,hj.soh}@kakaobrain.com



- ✓ 영어 Benchmark set인 STS-B를 한국어로 번역하여 생성
- ✓ train/dev/test 합쳐서 8000여개 정도
- ✓ 인간의 번역 검수가 이루어진 test셋의 데이터만 사용

[EXOBRAIN - paraKAIST]

ETRI 한국전자통신연구원
Electronics and Telecommunications Research Institute

엑소브레인 패러프레이즈 말뭉치 (KAIST)	한국어 패러프레이즈 말뭉치: Korean Paraphrase Corpus(KAIST) · 한국어 패러프레이즈 인식 및 평가를 위한 주석 가이드라인 및 말뭉치 · 말뭉치 구성: 패러프레이즈 관계 2,000문장 쌍과 출처, 유사도(0-5)/난이도(상/중/하) 표준 태깅, 의미(실질) 형태소 정보 태깅

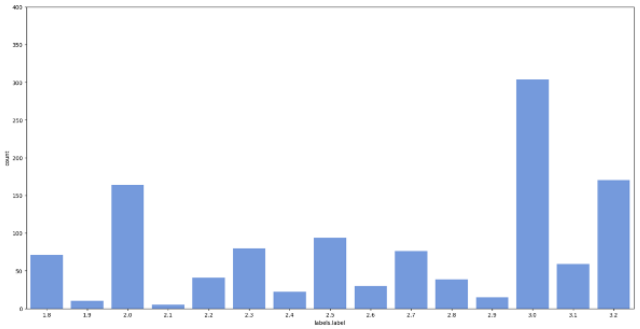
- ✓ 엑소브레인에서 제작한 한국어 페러프레이즈 문장쌍 데이터
- ✓ 문장쌍 2000여개로 이루어져있고 유사도와 난이도 태깅이 되어있음

(4) 모델 성능 개선방안 모색 – 데이터 증강 구간 설정

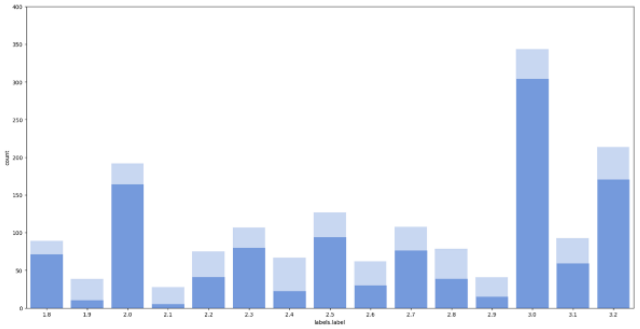
- Klue train set으로 학습한 구간 별 binary 판단 정확도 표

bin	TP	TN	FP	FN	accuracy
1.3 - 1.7	0	44	2	0	0.956521
1.8 - 2.2	0	39	12	0	0.764706
2.3 - 2.7	0	33	30	0	0.523809
2.8 - 3.2	25	3	19	15	0.451613
3.3 - 3.7	58	0	0	4	0.935484
3.8 - 4.2	54	0	0	1	0.981818
4.3 - 4.7	55	0	0	0	1.0
4.8 - 5.0	22	0	0	0	1.0

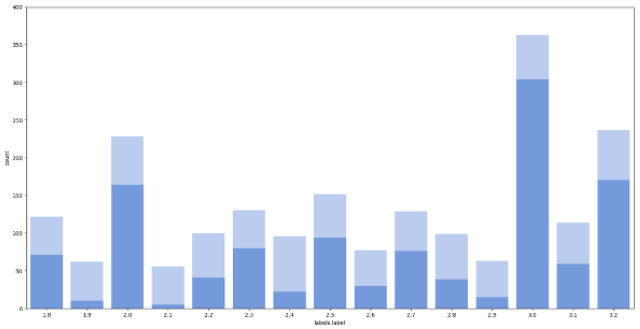
- ✓ augment를 적용하지 않은 모델로 테스트셋을 평가 했을 때 가장 **binary 정답률이 낮은 구간이 1.8-3.2 구간**인 것을 확인함
- ✓ trainset의 **1.8-3.2 구간**에 해당하는 데이터들이 **압도적으로 적기 때문에 발생하는 문제**라 판단
- ✓ 따라서 **1.8-3.2구간의 학습 데이터를 pseudo label을 통해 늘려주는 것을 목표로 augmentation**을 적용



new trainset
1.8-3.2



new trainset 1.8-3.2
+ korsts_test 485개

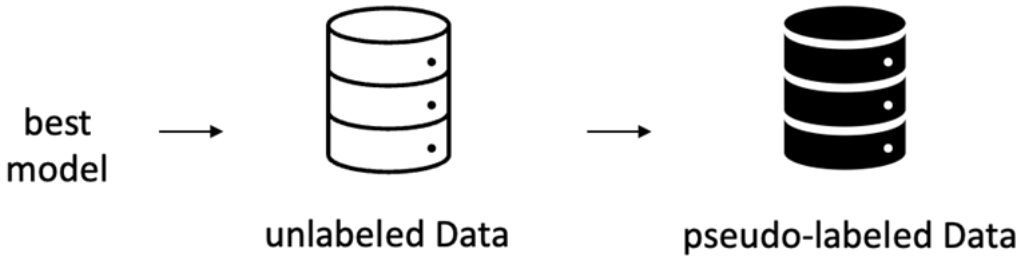


new trainset 1.8-3.2
+ korsts_test
+ paraKAIST 851개

(4) 모델 성능 개선방안 모색 – Label Selection

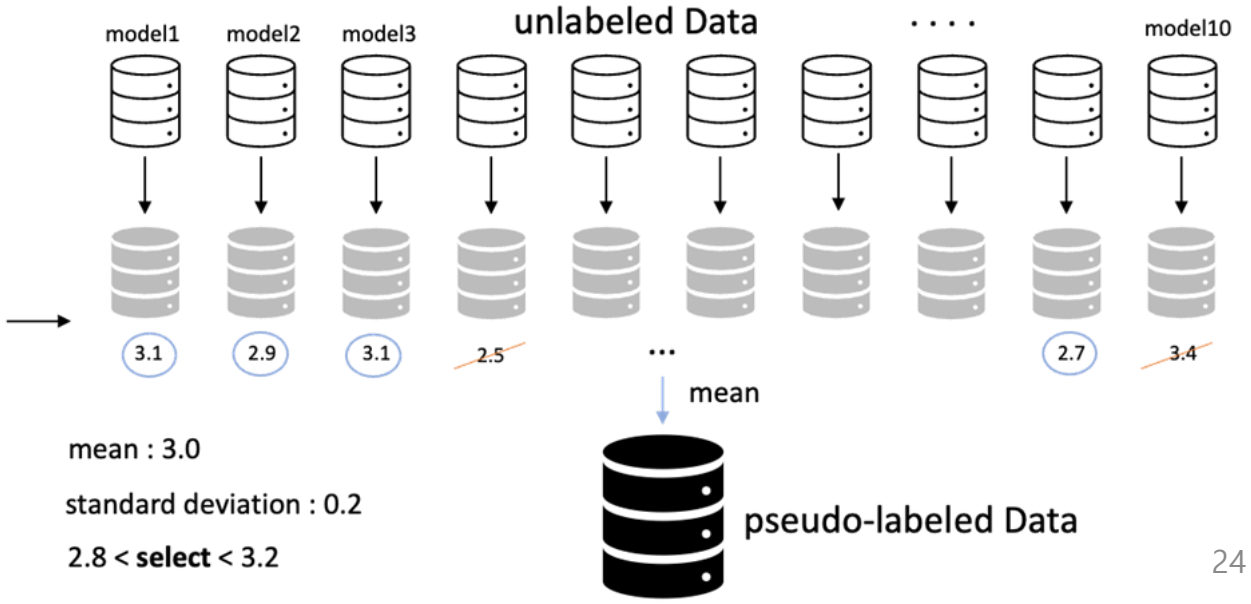
1. use BestModel for pseudo-labeling

seed	pearsonR	1,8-3.2 pears	1.8-3.2 False	f1 score
seed1	0.9144	0.4104	76	0.8375
seed2	0.9179	0.4478	82	0.8298
seed3	0.9104	0.4663	87	0.8154
seed4	0.9125	0.4202	84	0.8226
seed5	0.9169	0.4299	85	0.8327
seed6	0.9226	0.4535	70	0.8235
seed7	0.9019	0.4457	95	0.8021
seed8	0.9154	0.4435	84	0.8317
seed9	0.9148	0.4301	79	0.8431
seed10	0.9146	0.4283	88	0.8274



2. label ensembling

seed	pearsonR	1,8-3.2 pears	1.8-3.2 False	f1 score
seed1	0.9144	0.4104	76	0.8375
seed2	0.9179	0.4478	82	0.8298
seed3	0.9104	0.4663	87	0.8154
seed4	0.9125	0.4202	84	0.8226
seed5	0.9169	0.4299	85	0.8327
seed6	0.9226	0.4535	70	0.8235
seed7	0.9019	0.4457	95	0.8021
seed8	0.9154	0.4435	84	0.8317
seed9	0.9148	0.4301	79	0.8431
seed10	0.9146	0.4283	88	0.8274



Train set 11600여개 중 **350여개 라벨 교정**

original dataset vs **new dataset** evaluation

Model trained with	**Pearson R**	**F1 score**	**MSE**	**2.0-3.5 Pearson**	**2.0-3.5 ACC**
Cross-encoder / *KLUE-RoBERTa-base*					
original klue-trainset	90.55	81.38	0.41	39.42	59.36
new trainset	**91.14**	**82.98**	**0.37**	**43.73**	**62.68**

재라벨링 타겟이었던
2.0 – 3.5 구간의 피어슨 계수와 acc 상승

&

전체구간의 성능 또한 함께 상승

05 연구 결과

Augmentation result

Model trained with	**Pearson R**	**F1 score**	**1.8-3.2 Pearson R**	**Data Augmentation**	**Number of Augmented data**	**Aug range**
without data augmentation						
original klue-trainset	90.55	81.38	39.45	x	x	x
new trainset	91.14	82.65	41.75	x	x	x
with data augmentation						
use best-model prediction	91.61	83.57	43.59	+ korsts-test	485	1.8-3.2
label ensemble	91.98	83.24	46.44	+ korsts-test	485	1.8-3.2
use best-model prediction	91.57	**84.04**	46.60	+ korsts-test & paraKAIST	851	1.8-3.2
label ensemble	**92.05**	83.30	**49.53**	+ korsts-test & paraKAIST	847	1.8-3.2

✓ 데이터 증강을 통한 **준지도학습 기법으로도 성능향상**을 보임.

✓ **한가지 데이터에 대해서만 증강 / 수를 늘려 두가지 모두에 대해 증강한 경우**
-> 두가지 **모두** 증강 전보다 **성능향상**
-> 단순 최고 성능모델의 라벨을 이용하는 것이 아닌 앙상블을 통해 불확실성이 높은 라벨을 제거하고 데이터를 증강한 경우가 모두 더 좋은 성능을 보임

05 연구 결과 - 산출물

STS 재라벨링 가이드라인

점수	새로 정립한 기준
4	
3.5	핵심맥락 상호함의(2) + 보조맥락 상호함의(2)
3	핵심맥락 상호함의(2) + 보조맥락 일방함의(1) 핵심맥락 일방함의(1) + 보조맥락 상호함의(2)
2.5	핵심맥락 상호함의(2) + 보조맥락 비함의(0) 핵심맥락 일방함의(1) + 보조맥락 일방함의(1) 핵심맥락 일방함의(1) + 보조맥락 비함의(0) 핵심맥락 비함의(0) + 병렬 정보 많은경우
2	핵심맥락 비함의(0)

맥락 간 함의 관계를 기반으로
유사성을 판단하는 라벨 가이드

재라벨링 데이터 파일

KLUE train set

KLUE dev set

문장쌍에 대해 새로 매긴 라벨 점수
핵심맥락 함의 점수
보조맥락 함의 점수 등 포함

데이터 유효성 검증 및 모델 개선에 사용된 코드

```
## CROSS-ENCODER ##


# linear learning-rate warmup steps
warmup_steps = math.ceil(len(train_dataloader) * num_epochs / train_batch_size * 0.1)
logging.info("Warmup-steps: {}".format(warmup_steps)) # 학습 로그 표시



# Training
cross_encoder.fit(
    train_dataloader = train_dataloader,
    evaluator=val_evaluator,
    epochs=num_epochs,
    evaluation_steps=int(len(train_dataloader)*0.1),
    optimizer_params = {'lr':5e-5},
    warmup_steps=warmup_steps,
    output_path=model_save_path,
    show_progress_bar = True
)
```

```
2022-12-15 00:32:35 - Warmup-steps: 4
Epoch: 100% 3/3 [10:45<00:00, 215.03s/it]
Iteration: 100% 347/347 [03:32<00:00, 1.59it/s]
Iteration: 100% 347/347 [03:33<00:00, 1.43it/s]
Iteration: 100% 347/347 [03:29<00:00, 1.48it/s]
```

연구 의의

- ✓ **라벨링 난이도가 높은 점수구간의 명확한 라벨링 기준 정립**

  난이도가 높은 3점 부근의 점수대에 기계와 인간이 '유사성'을 판단할 수 있는 보다 명확한 기준과 근거를 마련함

- ✓ **양질의 STS 라벨 데이터 제작**

  라벨 기준 제작자들이 직접 라벨링을 실시해 라벨의 신뢰도 보장

- ✓ **적은 수의 라벨교정과 데이터 증강만으로 유의미한 성능변화를 도출**

- ✓ **기존 벤치마크의 문제점을 교정한 새로운 벤치마크로의 발전가능성**

- ✓ **모델 성능향상과 데이터 교정 및 증강을 위한 다양한 방법의 시도와 유의미한 실험결과 도출**

한계점

- ✓ **다소 적은 양의 재라벨 데이터**

  2.0 – 3.5 구간의 모든 데이터에 재라벨링을 실시한다면 더 큰 성능 향상을 기대

- ✓ **라벨링 가이드의 난이도 상승**

  난이도가 높은 3점 부근의 특성 상 가이드가 명확해짐에 따라 기존에 비해 라벨링 작업 난이도 상승

- ✓ **해결하지 못한 워드 오버랩 기반 예측경향**

  다량의 관련데이터 필요

- ✓ **여전히 불균형한 훈련 데이터**