Transformer 기반의 Global pLACE(GLACE) Matching 개발기

고성필 박성진
NAVER GLACE

NAVER DEVIEW 2023

CONTENTS

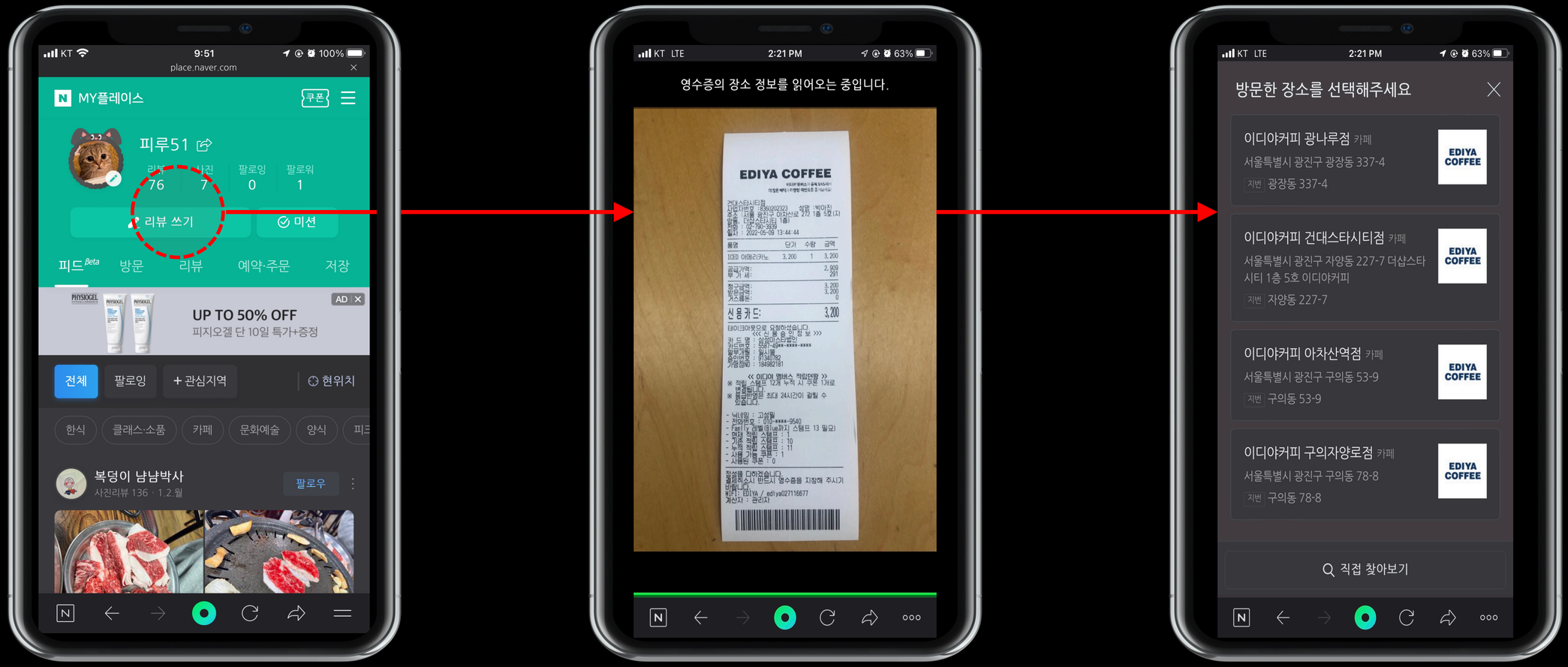
1.OCR POI Matching

CONTENTS

1.1 검색엔진으로 문제 해결이 어려운 배경

NAVER
DEVIEW
2023

GLACE CIC에서 운영하는 서비스인 마이플레이스의 영수증 인증 기능



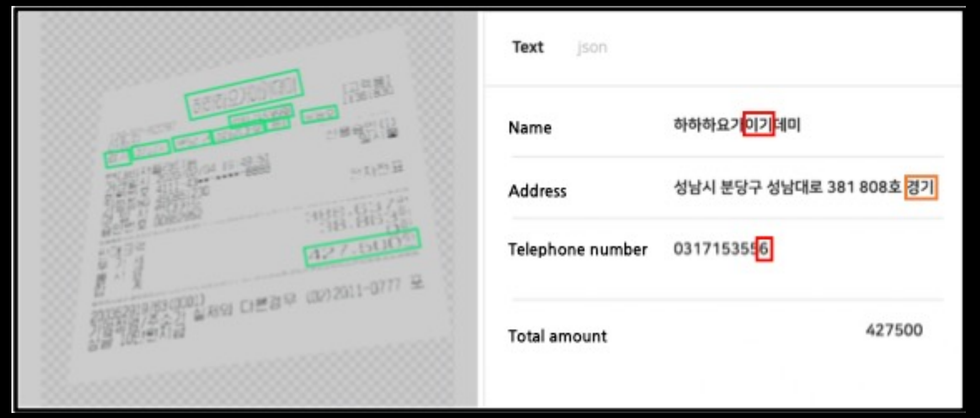
1.1 검색엔진으로 문제 해결이 어려운 배경

NAVER
DEVIEW
2023

Noisy sample Issue

- 왜곡된 영수증 이미지로 OCR을 통해 텍스트가 추출된 경우 오탈자 포함

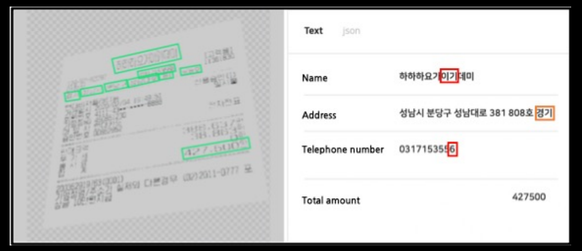
- Database에 등록된 가게 정보와 영수증 정보가 일치하지 않는 경우 존재



1.1 검색엔진으로 문제 해결이 어려운 배경

NAVER DEVIEW 2023

(a) Image noise



Receipt Image

(1) OCR

```
{
    "text": [
    ...
    {
        "text": " 하하하요가이기데미",
        "2d_location": [0.5, 0.15]
    }, ...,{
        "text": "0317153556",
        "2d_location": [0.9, 0.9]
        }, ...
    ]
}
```

(b) OCR noise

(2) Parsing

(c) Parsing noise

Keys	Extracted Values	Ground Truth
name	하하하요가이기데미	하하하요가아카데미
tel	0317153556	0317153558
address	분당구 성남대로 381 808호 경기	경기 성남시 분당구 성남대로 381 808호

(3) Matching

...

```
{
  "name" : "하하하요가학원",
  "tel" : "0317153558"
  "address" : NULL,
  "type" : "academy",
  "is_franchise" : false,
   ...
}
```

(d) DB noise

```
{
  "name" : "하하하 Cafe 태평로점",
  "address" : "서울 태평로1가 세종대로 21",
  "tel" : "02-346-362",
  "type" : "cafe",
  "is_franchise" : true,
   ...
}
```

...

Store Database

1.2 Baseline - CLOVA POI Matching

NAVER DEVIEW 2023

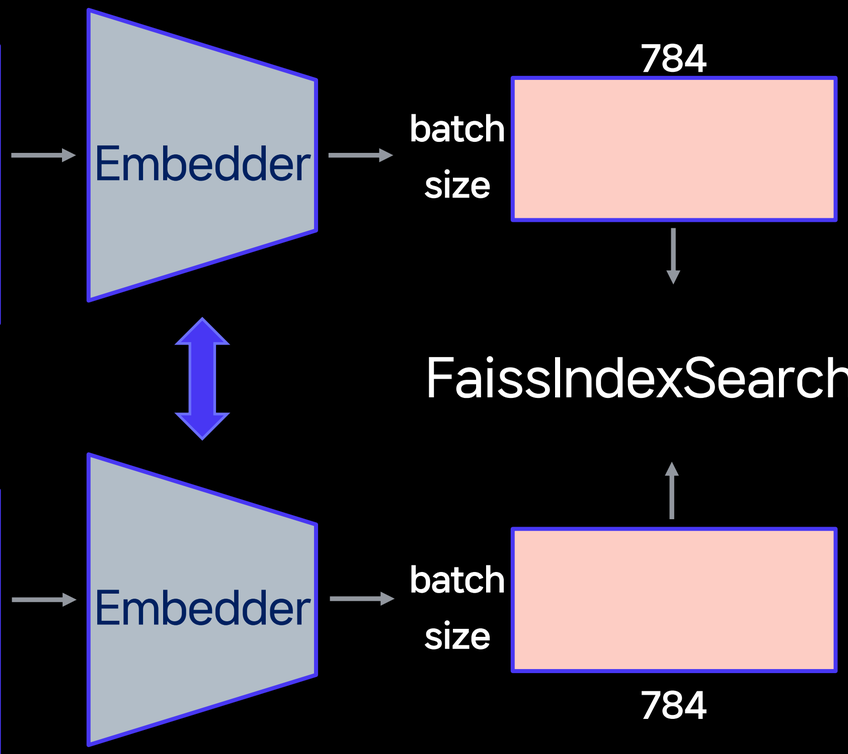
POI: Place of Interest

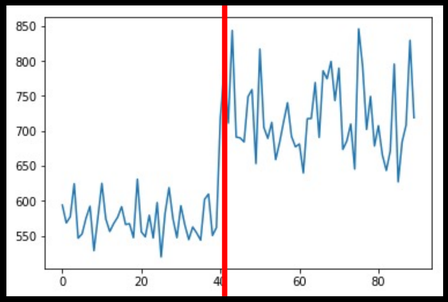
parsed ocr


```
{
  "name": "하하하요가이기데미",
  "tel": "0317153556",
  "address": "분당구 성남대로 381 808호 경기"
}
```


poi info


```
{
  "name": "하하하요가아카데미",
  "tel": "0317153558",
  "address": "경기 성남시 분당구 성남대로 381 808호"
}
```



Semi-Structured Query Grounding for Document-Oriented Databases with Deep Retrieval and Its Application to Receipt and POI Matching [AAAI/w 2022]

1.2 Baseline – 학습데이터 구축

NAVER
DEVIEW
2023

모델 고도화를 위한 학습데이터 구축

- 플레이스는 하루에도 많은 수의 업체가 개/폐업이 진행됨으로써 영수증의 특징도 매일 달라질 것이라고 예측
- 영수증 인증 서비스를 운영하면서 쌓인 parsed ocr과 매칭된 poi info를 통해 학습 진행
- {parsed-ocr – poi info} 를 pair로 약 4천만 개 trainset 구축

parsed ocr

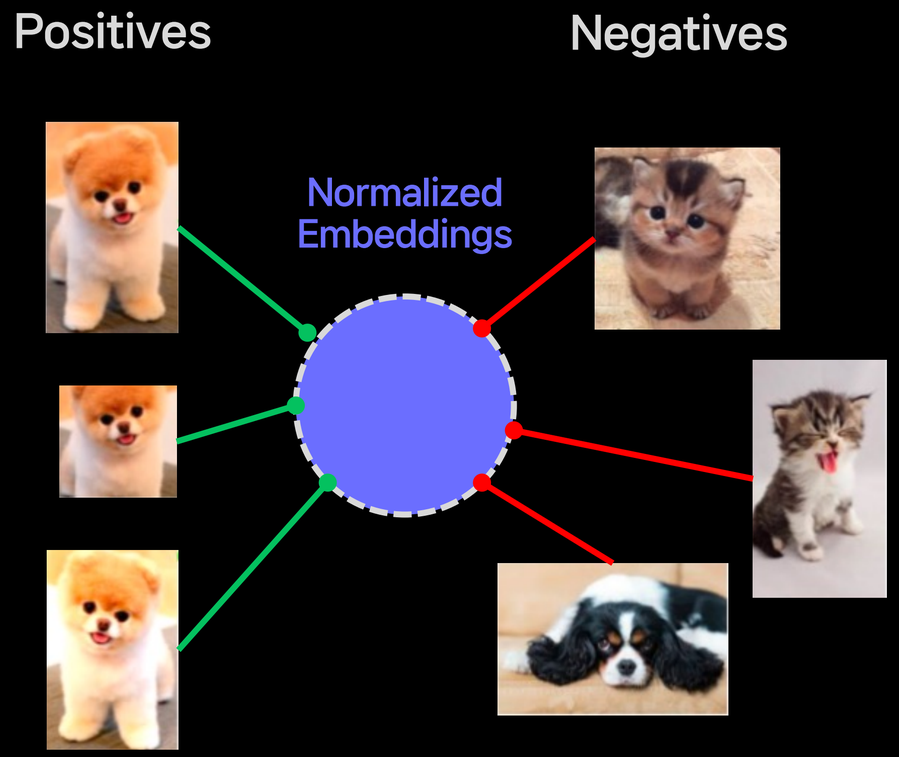

```
{
  "name": "하하하요가이기데미",
  "tel": "0317153556",
  "address": "분당구 성남대로 381
  808호 경기"
}
```


poi info

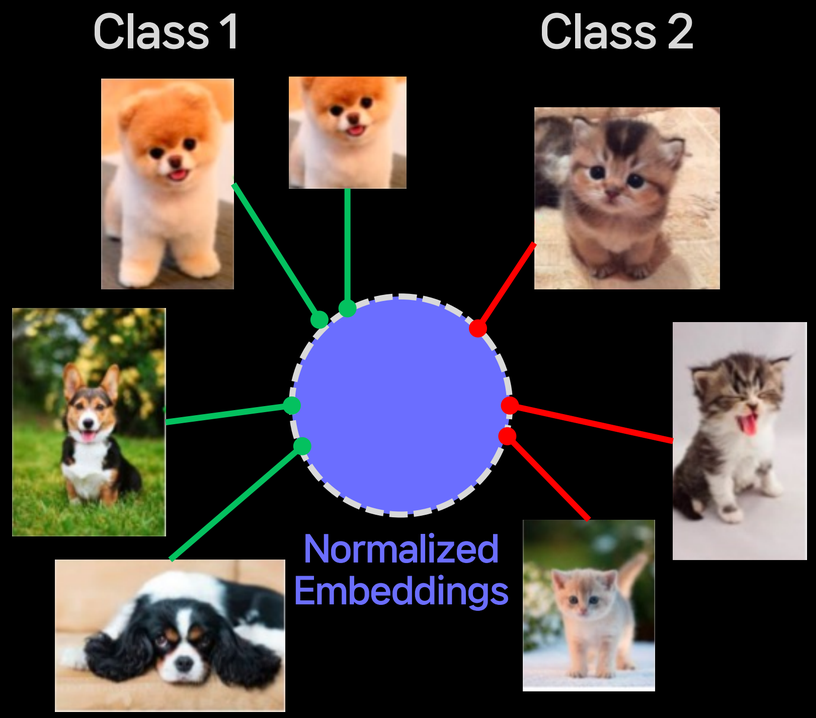

```
{
  "name": "하하하요가아카데미",
  "tel": "0317153558",
  "address": "경기 성남시 분당구
  성남대로 381 808호"
}
```

1.3 성능 고도화 - Supervised Contrastive Learning

NAVER
DEVIEW
2023



Self Supervised
Contrastive

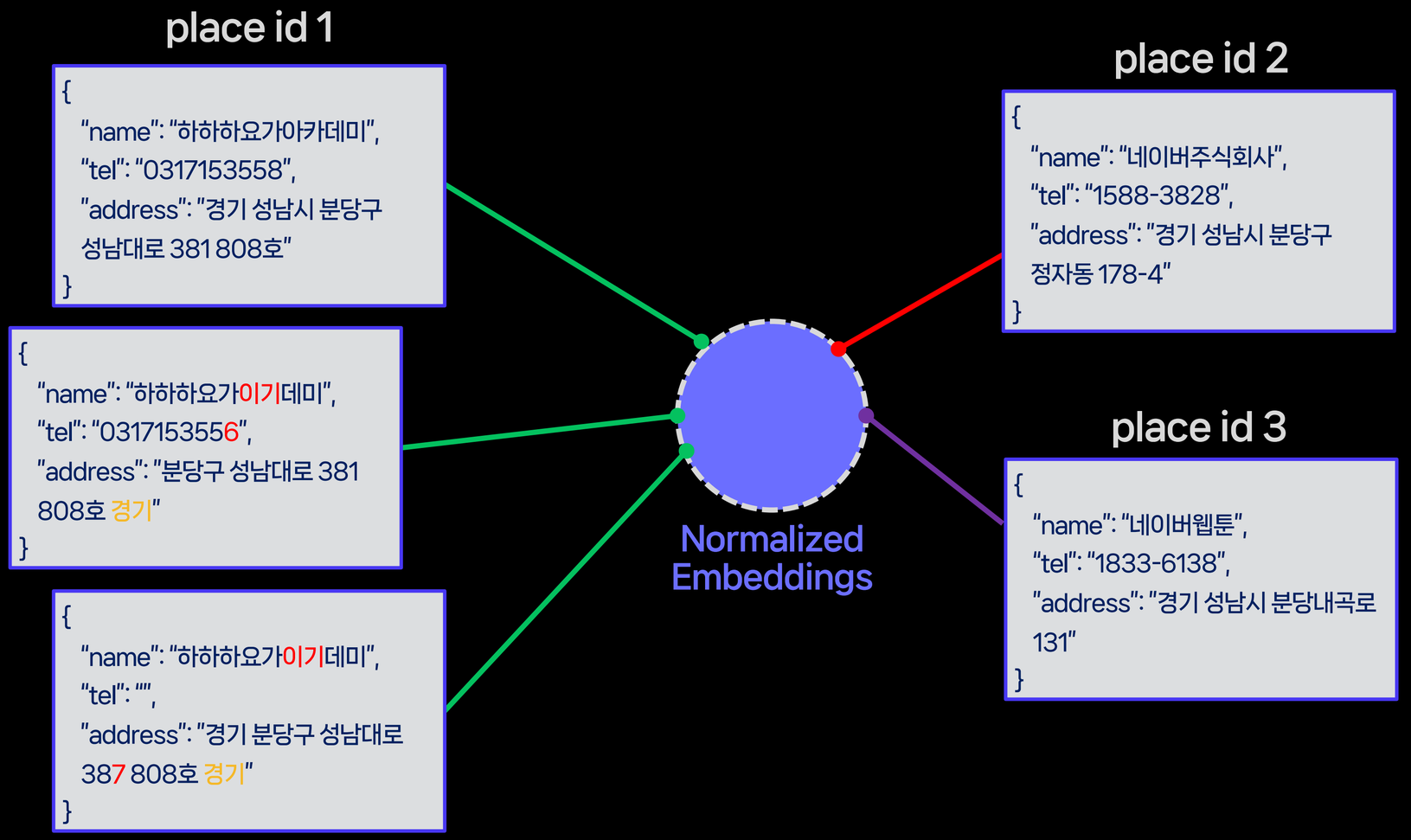


Supervised Contrastive

Supervised Contrastive Learning (NeurIPS 2020)

1.3 성능 고도화 - Supervised Contrastive Learning

NAVER
DEVIEW
2023



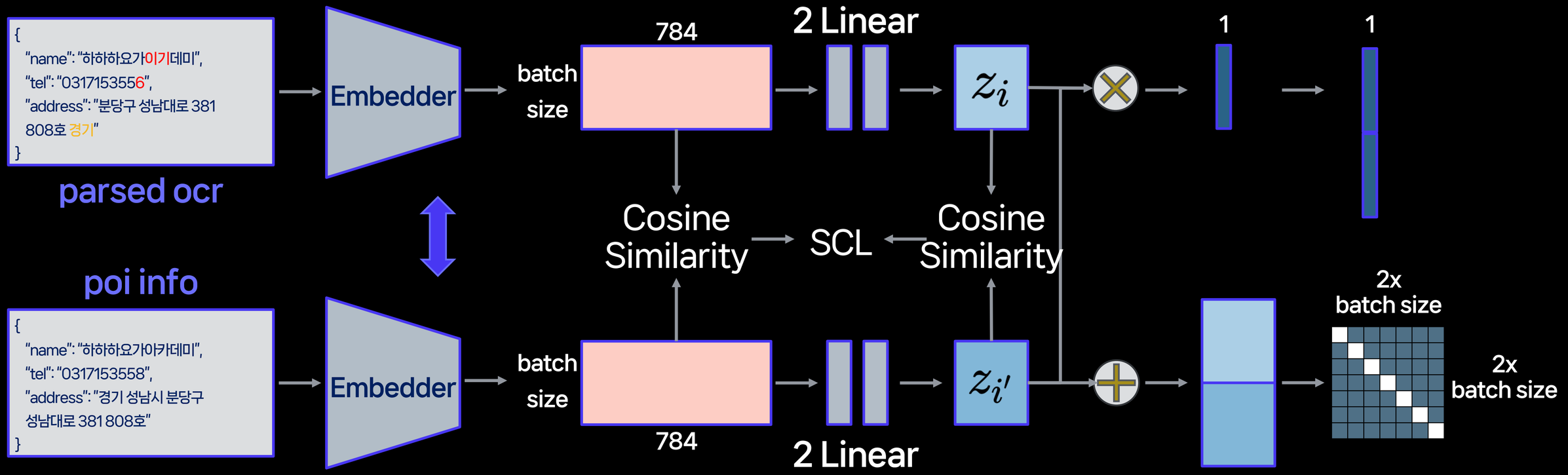
Supervised Contrastive

Supervised Contrastive Learning (NeurIPS 2020)

1.3 성능 고도화 – Pairwise supervised contrastive learning

NAVER DEVIEW 2023

- To better capture the **high-level semantic feature** of text data with **parsed OCR** & **poi info**
- We adopt **2 linear layer** after Embedder model for extracting projected features
- Calculate the cosine similarity between projected features, then apply **SCL**



Pairwise Supervised Contrastive Learning of Sentence Representations [EMNLP 2021]

1.3 성능 고도화 – Pairwise supervised contrastive learning

NAVER DEVIEW 2023

- Apply **instance discrimination loss (ID)** for implicitly separating positive pair apart from other samples.
- In general, contrastive loss learns semantic similar samples (z_i , $z_{i'}$) closely, and dissimilar samples (z_i , z_j) separately

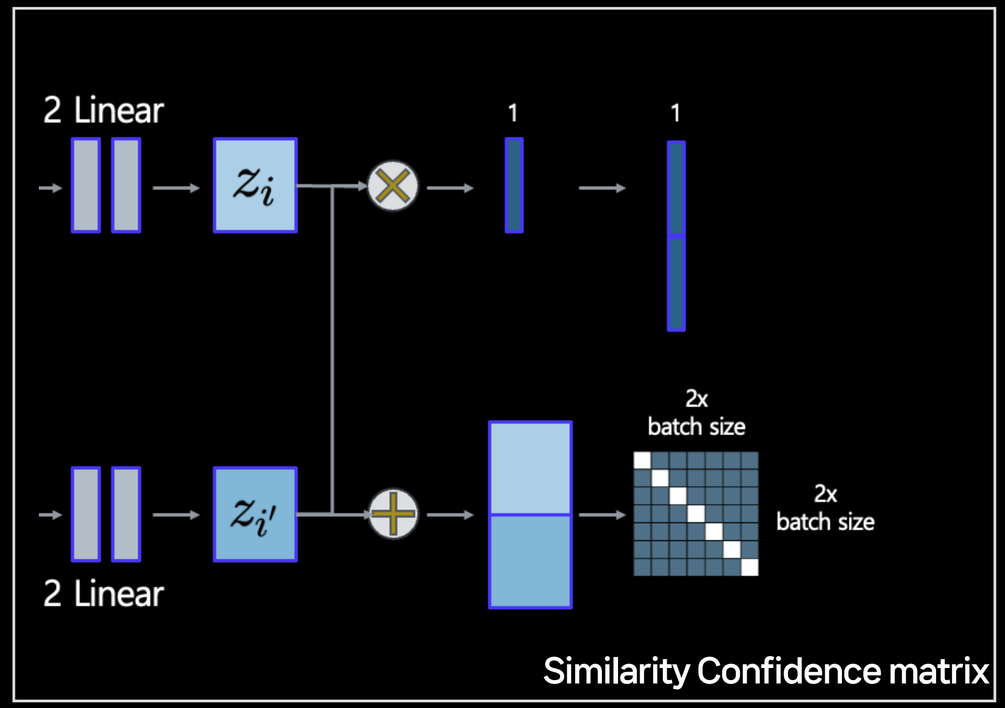
Contrastive Loss

$$\ell_{\text{const}}^{i} = -\log \frac{\exp(s(z_i, z_{i'})/\tau)}{\sum_{j \in \mathcal{I} \setminus i} \exp(s(z_i, z_j)/\tau)}$$

- However, we apply **ID** for learning not only samples (z_i , $z_{i'}$), but also (z_i , z_j) separately.

Instance discrimination Loss

$$\mathcal{L}_{\text{ID}} = \frac{1}{P_M} \sum_{i=1}^{M} \mathbf{1}_{y_i=1} \cdot \left(\ell_{\text{ID}}^{i} + \ell_{\text{ID}}^{i'}\right) \qquad \ell_{\text{ID}}^{i} = -\log \frac{\exp(s(z_i, z_{i'})/\tau)}{\sum_{j \in \mathcal{T} \setminus i} \exp(s(z_i, z_j)/\tau)}$$



Pairwise Supervised Contrastive Learning of Sentence Representations [EMNLP 2021]

1.3 성능 고도화 – 2 stage learning

NAVER DEVIEW 2023

- We originally use **PLM*** named **LaRva** to train massive parameter models with Korean corpus by CLOVA
- However, our {**parsed OCR – poi info**} pair dataset doesn't fit PLM 's Korean corpus
- **Domain discrepancy** exists between our data and Korean corpus

poi info


```
{
  "name": "하하하요가아카데미",
  "tel": "0317153558",
  "address": "경기 성남시 분당구
  성남대로 381 808호"
}
```


Domain discrepancy

Korean corpus

- Motivated by COCO-LM, we apply Corrective Language Modeling (CLM) to train PLM in order to better capture **token-level semantics from poi-dataset**

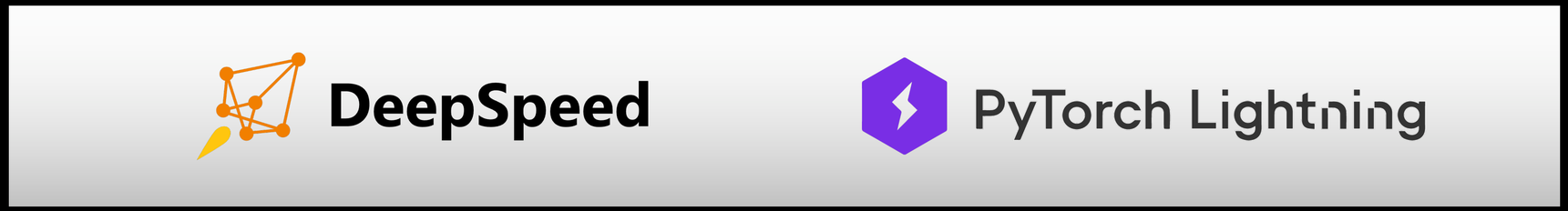
PLM: Pretrained Language Model

COCO-LM: Correcting and contrasting text sequences for language model pretraining [NeurIPS 2021]

1.3 성능 고도화 – DeepSpeed를 활용한 Big-batch train

NAVER DEVIEW 2023

- In contrastive learning, **larger batch size, more negative samples in batch** enhances feature extraction performance.
- We use **DeepSpeed** in **Pytorch-Lightning** to gain memory benefits and increase batch size for contrastive learning
- **Pytorch-Lightning** can scale up models without the boilerplate.

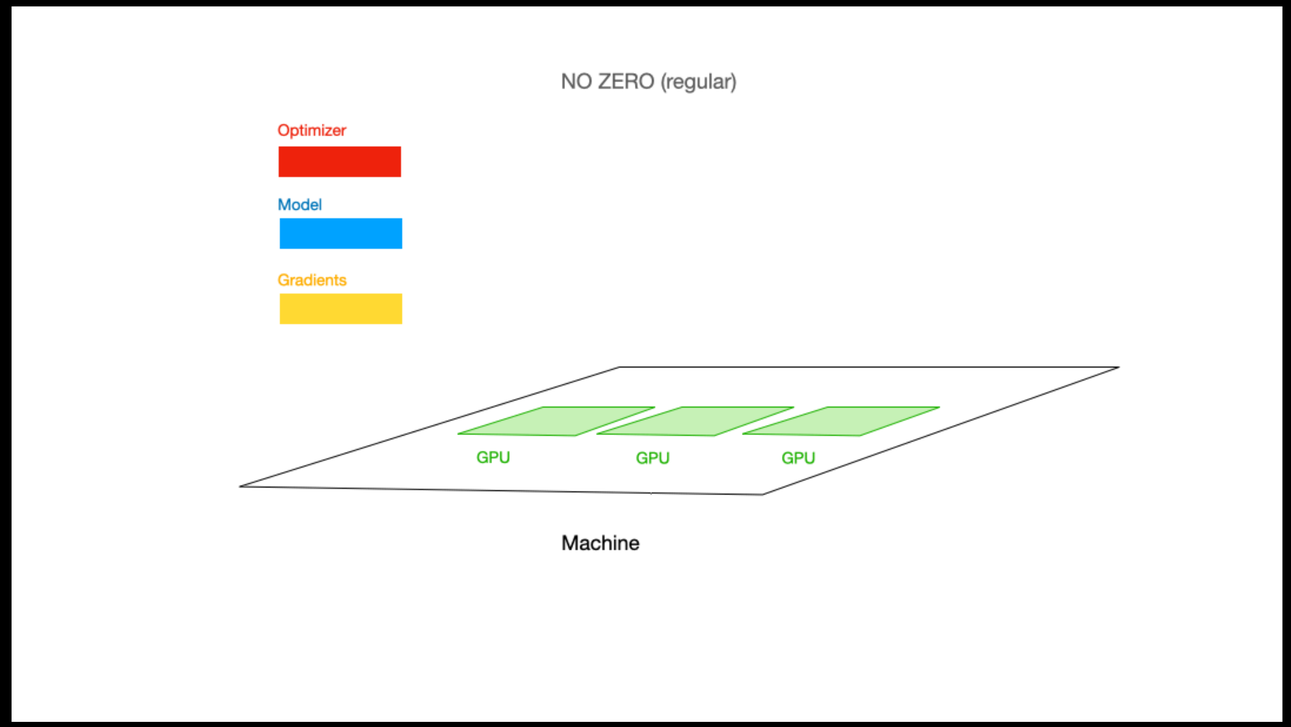


ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. [SC20]

1.3 성능 고도화 – DeepSpeed를 활용한 Big-batch train

NAVER DEVIEW 2023

- We use **DeepSpeed ZeRO Stage 2**, which partitions optimizer states and gradients across device
- We can increase 320 local batch size to 512 local batch size in A100 GPUs.



ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. [SC20]

1.3 성능 고도화 – 평가 지표

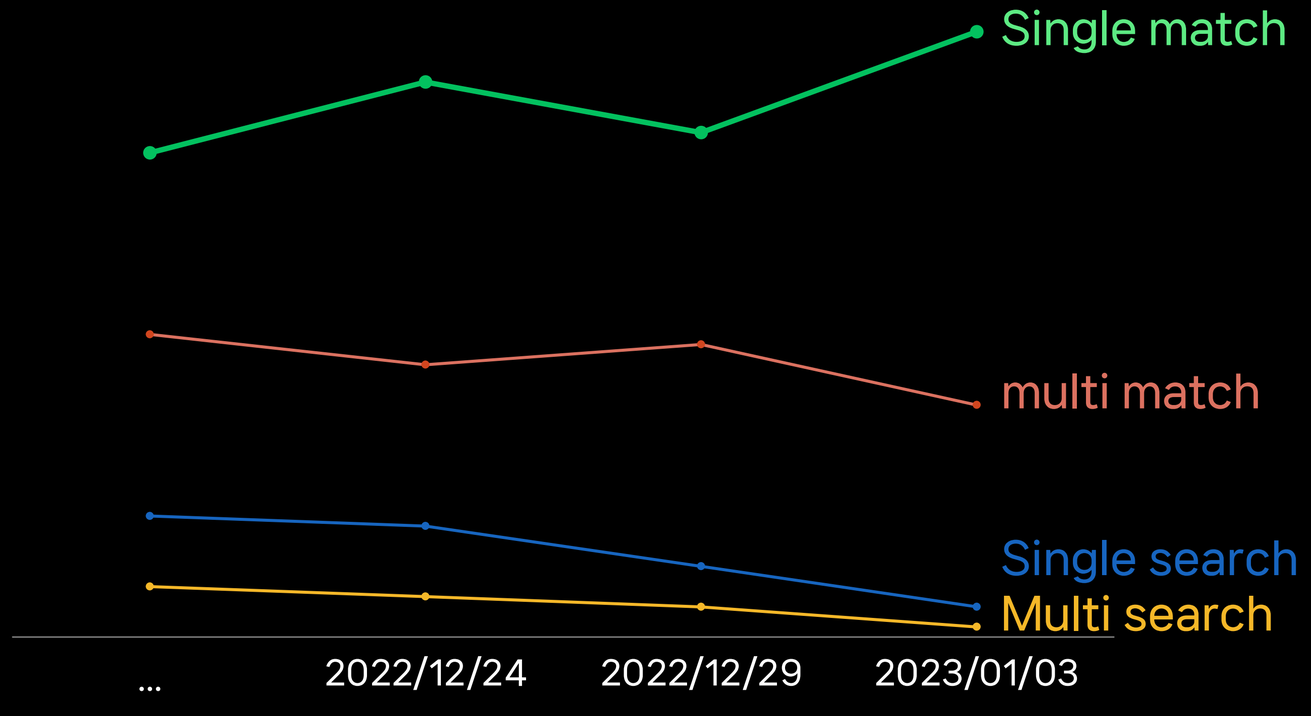
NAVER
DEVIEW
2023

Single match, Multi match

- **Single match(자동매칭)**: 업체를 1개 내려주고 해당 업체가 맞음
- **Multi match(복수매칭)**: 업체를 복수 개로 내려주고 해당 업체가 있음

Single search, Multi search

- **Single search(자동매칭-직접검색)** : 업체를 1개 내려주고 틀림
- **Multi search(복수매칭-직접검색)** : 업체를 복수 개로 내려주고 해당 업체가 없음



1.3 성능 고도화 – 평가 지표

NAVER
DEVIEW
2023

Evaluation metrics

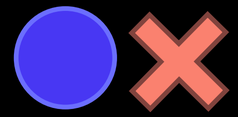
- **Accuracy**: TopK에서 정답을 맞춘 비율
- **mAP@k**: TopK 내의 예측 순서 가중치에 따른 precision metric

mAP@k

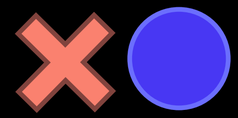
- 추천시스템과 동일하게 정답 place가 상위에 rank될수록 좋은 **matching model**
- 후보 내의 예측 순서까지 반영한 **mAP@k**를 개발하여 성능평가에 추가

정답: 1003112

예측: 1003112
: 1

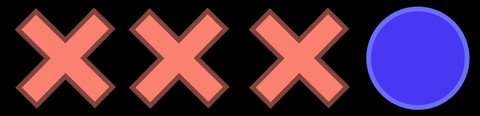


예측: 1003112, 1003115
: (1x1 + 1/2x0) / 2 = 1/2



예측: 1003115, 1003112
: (1x0 + 1/2x1) / 2 = 1/4

예측: 1003112, 1003115, 2523123,
: (1x1 + 1/2x0 + 1/3x0) / 3 = 1/3



예측: 1003115, 1003227, 2523123, 10031121
: (1x0 + 1/2x0 +1/3x0 + 1/4x1) / 4 = 1/16

1.3 성능 고도화 – 실험 결과

NAVER DEVIEW 2023

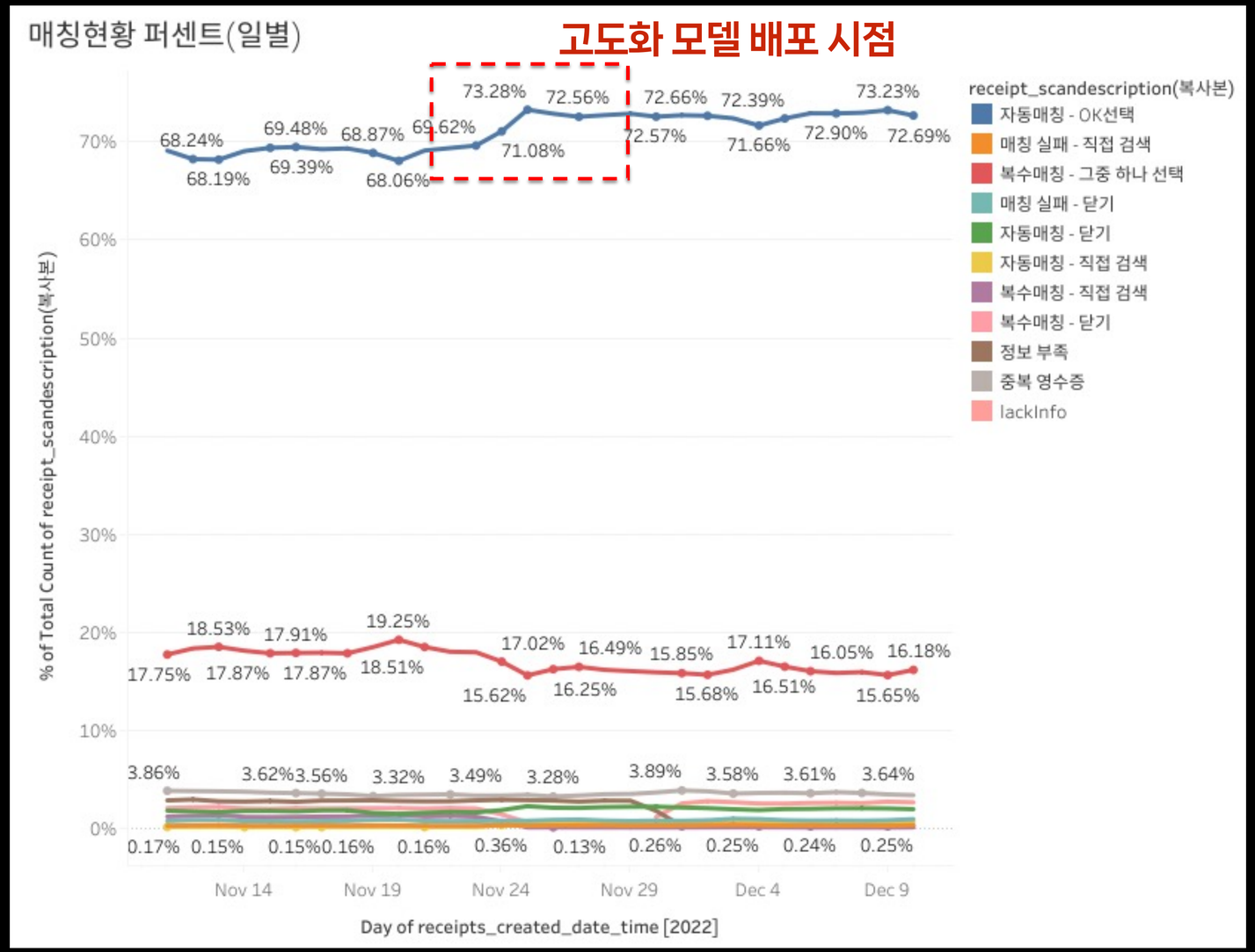
Ablation studies

- We validate **unseen 2,221 test data** without post-processing
- Our Pairwise Supervised Contrastive (**PairSupCon**) learning model successfully alleviates the issues of lower single-match
- Using DeepSpeed outperforms in single-matched and mAP@4 with compared to baseline model
- **Trade-off** between accuracy and mAP@K

#	single match	multi match	single search	multi search	Acc (%)	mAP@4
Supervised Contrastive Learning model (baseline)	858	1133	2	203	89.64 %	0.5416
baseline + PairSupCon	1,296	712	15	183	90.41 %	0.6815
baseline + PairSupCon + 2-stage	1,702	304	40	173	90.32 %	0.8080
baseline + PairSupCon + DeepSpeed (512bs)	1,774	227	46	175	90.09 %	0.8306
baseline + PairSupCon + 2-stage + DeepSpeed (512bs)	1,820	150	92	159	88.70 %	0.8420

1.3 성능 고도화 – 서비스 적용

NAVER
DEVIEW
2023

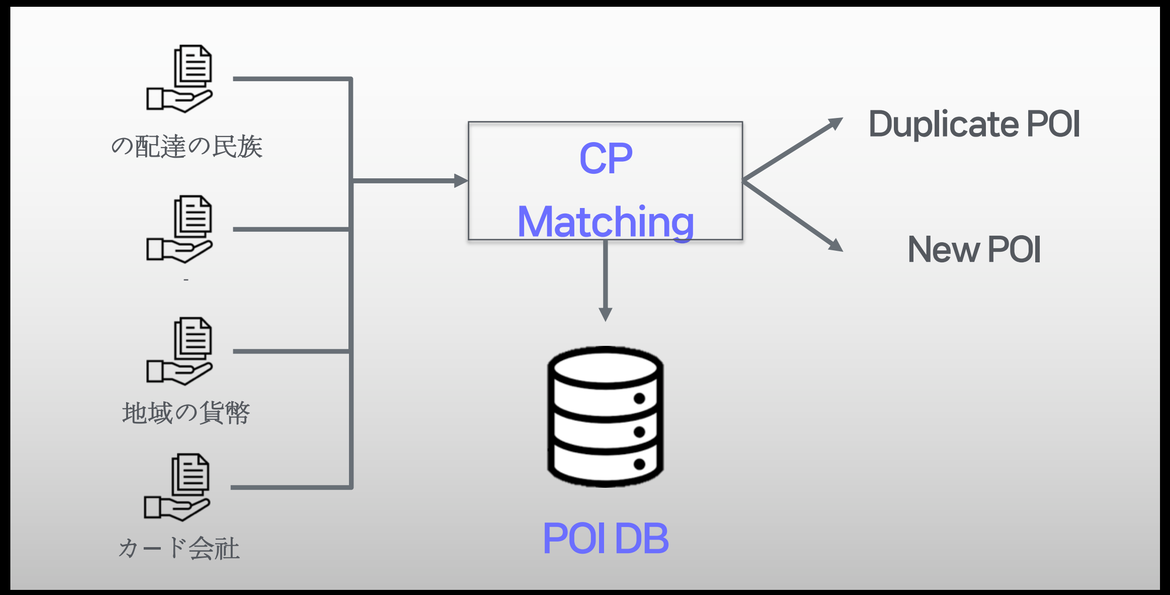


1.4 POI Matching 확장

NAVER
DEVIEW
2023

Expansion project of OCR POI Matching model

- **CP Matching** is database search system when providing place information (i.e., name, address, business number, etc) to check duplicated data in DB
- We utilize our advanced grounding model to train CP Matching framework
- We also use our **PairSupCon** model, **DeepSpeed**, and **2-stage PLM** when training POI Matching



CP input


```
{
  "name": "이디야커피백석대북카페점,
  "tel": "041-555-2964",
  "address": "충청남도 천안시 동남구 백석대학로 1-19",
  "latitude": "36.83864212",
  "longitude": "127.18274242",
  "biznum": "5862600601"
}
```


POI info


```
{
  "poi_id": 38343914
  "name": "이디야커피 백석대북카페점,
  "tel": "041-555-2964",
  "address": "충청남도 천안시 동남구 문암로 77",
  "latitude": "36.8423514",
  "longitude": "127.182746642",
  "biznum": "5862600601"
}
```


* CP: Contents Provider

1.4 POI Matching 확장

NAVER DEVIEW 2023

Expansion project of OCR POI Matching model

- We validate **6,500 testset** and **3,954 unseen testset** with label and compare our model with existing rule-base methods in terms of guarantee ratio and accuracy
- Our method achieves a better performance.

6,500 testset	Guarantee	Doubt	None	Grt ratio	Dbt ratio	None ratio
rulebase logic	3,769	1,443	1,242	57.98 %	22.20 %	19.12 %
Ours (PairSupcon + DeepSpeed + 2-stage PLM)	5,359	501	640	82.45 %	7.71 %	9.85 %

3,954 testset	Guarantee match	Guarantee mismatch	Doubt match	Doubt mismatch	Guarantee Acc (%)	Doubt Acc (%)
rulebase logic	2,700	29	421	89	98.94 %	82.55 %
Ours (PairSupcon + DeepSpeed + 2-stage PLM)	2,972	60	232	8	98.02 %	96.67 %

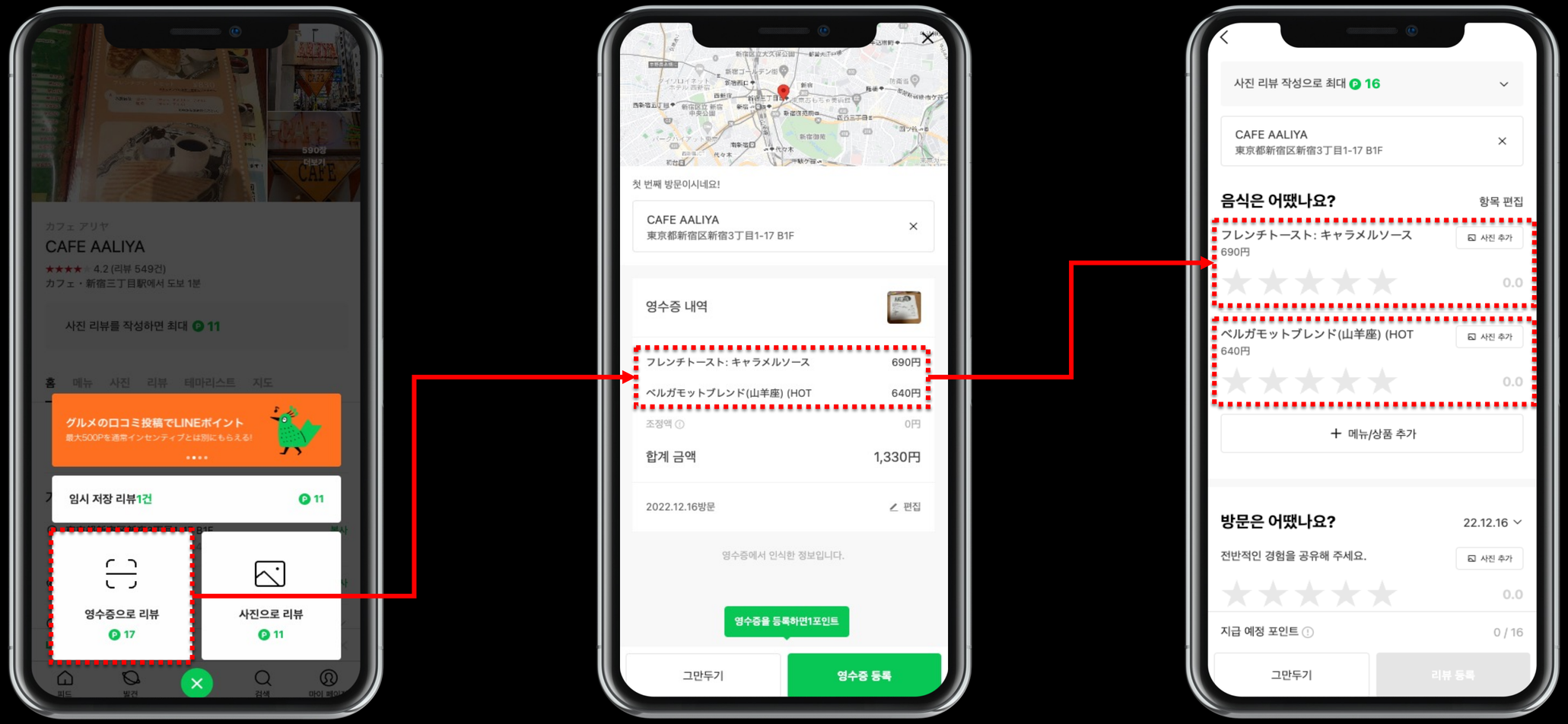
2. Menu Grounding

CONTENTS

2.1 Background

NAVER
DEVIEW
2023

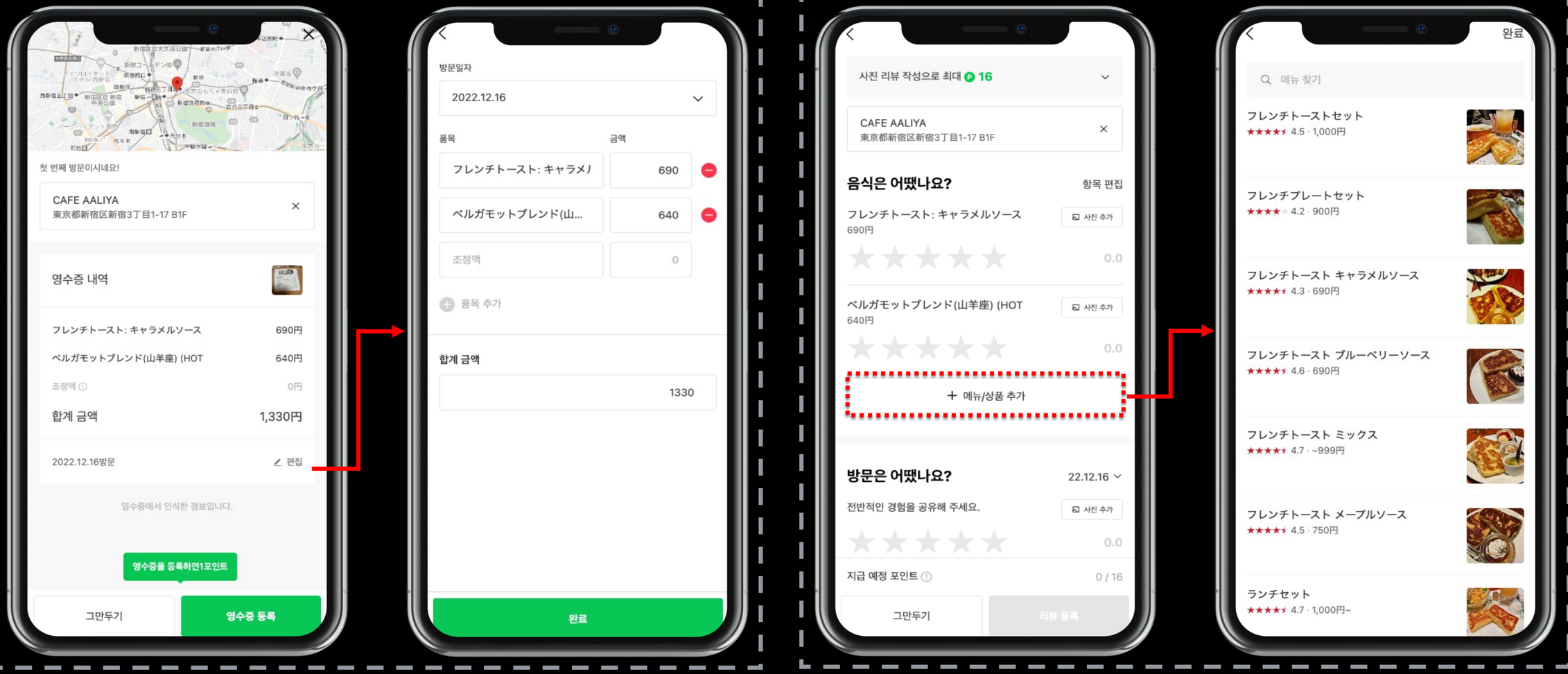
Line Place 서비스에서 영수증과 리뷰를 통해 메뉴 데이터 인입



2.1 Background

NAVER
DEVIEW
2023

Line Place 서비스에서 영수증과 리뷰를 통해 메뉴 데이터 인입



2.1 Background

NAVER
DEVIEW
2023

ISSUE: 메뉴 검수 리소스의 부담

- 인입 하려는 사용자 데이터(raw_menu_text)를 DB에 적재하는 과정
- 서비스 초반에는 exact 매칭으로만 운영
- exact 매칭 외 나머지 메뉴는 rule base / human resource로 운영
- 유저, 리뷰 수 증가에 따라 검수 리소스가 많아짐

-> ML을 활용하여 자동으로 menu를 matching하자

2.2 Dataset

NAVER
DEVIEW
2023

Place 기준

- standard_menu_name: 검수 완료 메뉴명, raw_menu_name: 사용자 인입 메뉴명

```
place_a= {
  "41710738": [
    {
      "raw_text": "ポイントカードレブレッソブレッド",
      "standard_menu_name": "レブレッソブレッド",
    },
    {
      "raw_text": "ストロベリー&カスタードトースト",
      "standard_menu_name": "ストロベリー&カスタードトースト",
    },
    {
      "raw_text": "ホットカフェラテ",
      "standard_menu_name": "カフェラテ",
    },
    {
      "raw_text": "Iカフェラテ",
      "standard_menu_name": "カフェラテ",
    },
    {
      "raw_text": "dBLTサンド",
      "standard_menu_name": "BLTサンド",
    },
    {
      "raw_text": "BLTサンド",
      "standard_menu_name": "BLTサンド",
    }
  ]
}
```

standard
menu_name

```
place_a = {
    "41710738": [
        {                          # 레브레소브레드
          "standard_menu_name": "レブレッソブレッド",
          "raw_menu_names": ["ポイントカードレブレッソブレッド"]
        },
        {                         # ストロベリー&カスタード トースト
          "standard_menu_name": "ストロベリー&カスタードトースト",
          "raw_menu_names": ["ストロベリー&カスタードトースト"],
        },
        {
          "standard_menu_name": "カフェラテ",
          "raw_menu_names": ["ホットカフェラテ", "Iカフェラテ"],
        },                        # 핫카페라떼,       I카페라떼
        {
          "standard_menu_name": "BLTサンド",
          "raw_menu_names": ["dBLTサンド","BLTサンド"]
        },
    ]
}
```

2.2 Dataset

NAVER
DEVIEW
2023

Place 기준

- standard_menu_name: 검수 완료 메뉴명, raw_menu_name: 사용자 인입 메뉴명
- exact 매칭 외 다양한 표현의 raw_menu_name들

```
all_place_has_caffelatte = [
    {
        "place_id": "41832184",                  "place_id": "42078011",
        "raw_menu_name": "アイスカフェラテR",     "raw_menu_name": "カフェラテS",
    },              # 아이스카페라떼R                          # 카페라떼S
    {
        "place_id": "50101769",                  "place_id": "41804615",
        "raw_menu_name": "アイスカフェラテ(540)", "raw_menu_name": "Sカフェ ラテ",
    },              # 아이스카페라떼(540)                      # S카페라떼
    {
        "place_id": "42088026",                  "place_id": "36781318",
        "raw_menu_name": "カフェラテ(Hot)",       "raw_menu_name": "アイスカフェラテM",
    },              # 카페라떼(Hot)                           # 아이스카페라떼M
]

```

```
all_place_has_takoyaki = [
    {
        "place_id": "48421323",      "place_id": "57941063",
        "raw_menu_name": "たこ焼",    "raw_menu_name": "たこ焼4個",
    },                  # 타코야끼                  # 타코야끼4개
    {
        "place_id": "42102659",      "place_id": "33852647",
        "raw_menu_name": "たこ焼",    "raw_menu_name": "TOたこ焼8個",
    },                  # 타코야끼                  # TO타코야끼8개
    {
        "place_id": "41865966",      "place_id": "338209307812526 47",
        "raw_menu_name": "たこ焼き8コ",   "raw_menu_name": "Tたこ焼8テイク",
    },                  # 타코야끼8개                  # T타코야끼8테이크
]

```

2.2 Dataset

NAVER
DEVIEW
2023

Train / Valid set (0.95:0.05)

- data pair = [standard_menu_name, raw_menu_name]
- 동일한 standard_menu_id인 경우 positive sample로 정의 (SCL)
- 총 43만 place의 394만 메뉴명에 대한 428만 건의 data pair

Supervised Contrastive Learning (NeurIPS 2020)

2.3 Menu Matching (training)

NAVER
DEVIEW
2023

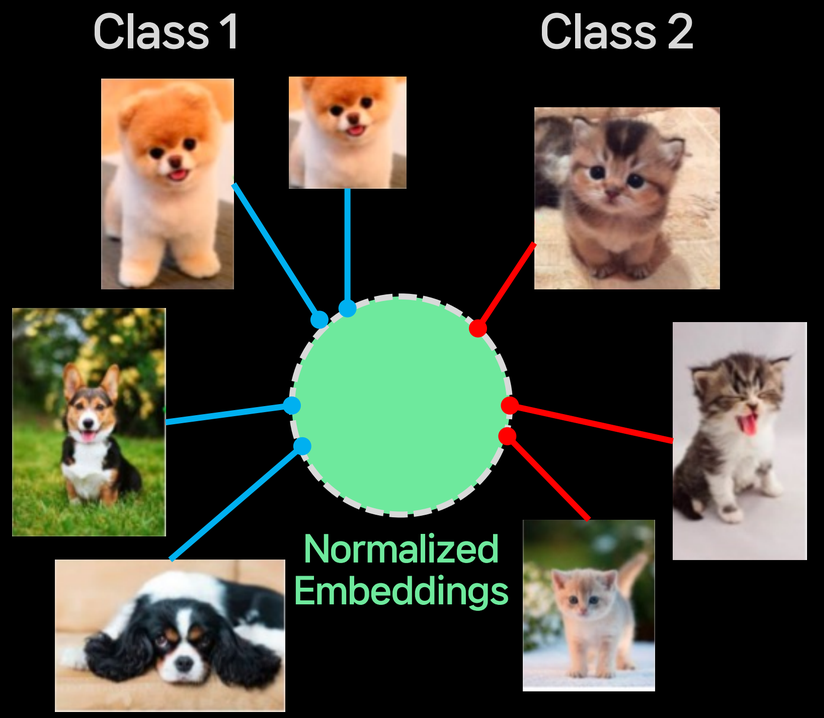
standard_menu_name

```
{
  1: みそ汁,
  2: オランジェ,
  3: 生キャラメルソフト,
  ...
}
```

raw_menu_name

```
{
  1: 無)みそ汁,
  2: オランジェ,
  3: キャラメルソフト,
  ....
}
```

Embedder

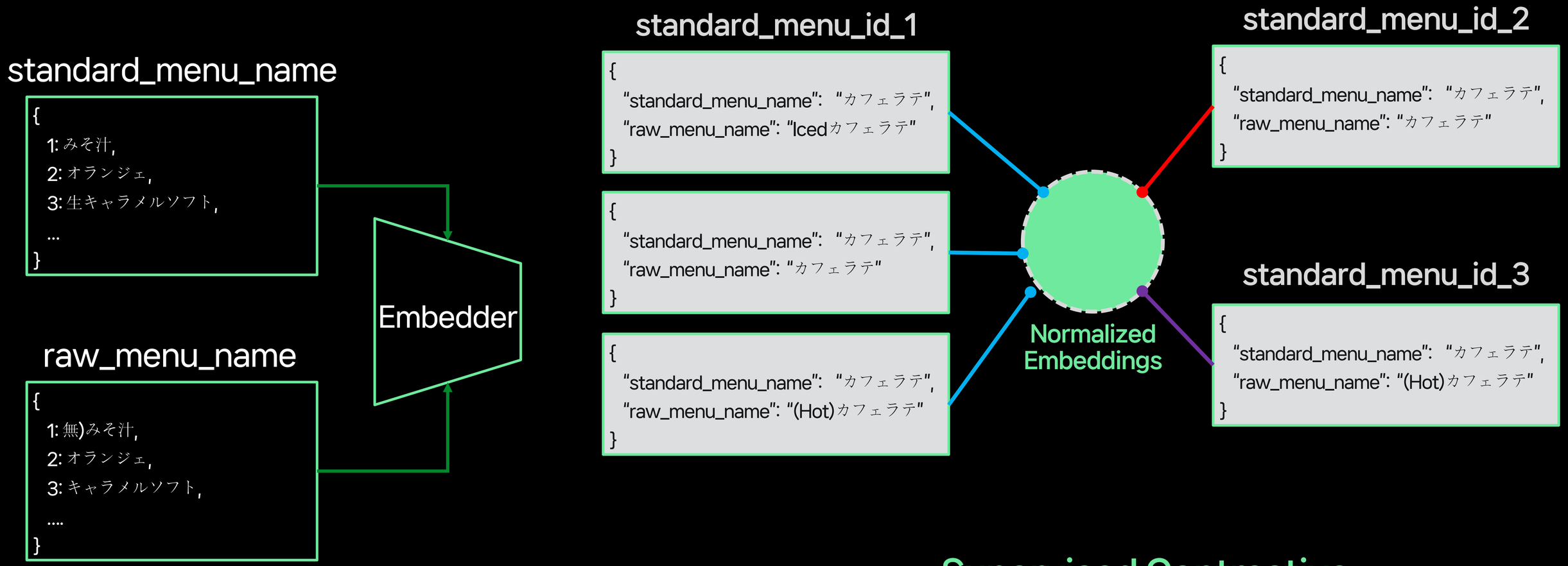


Supervised Contrastive

Supervised Contrastive Learning (NeurIPS 2020)

2.3 Menu Matching (training)

NAVER
DEVIEW
2023

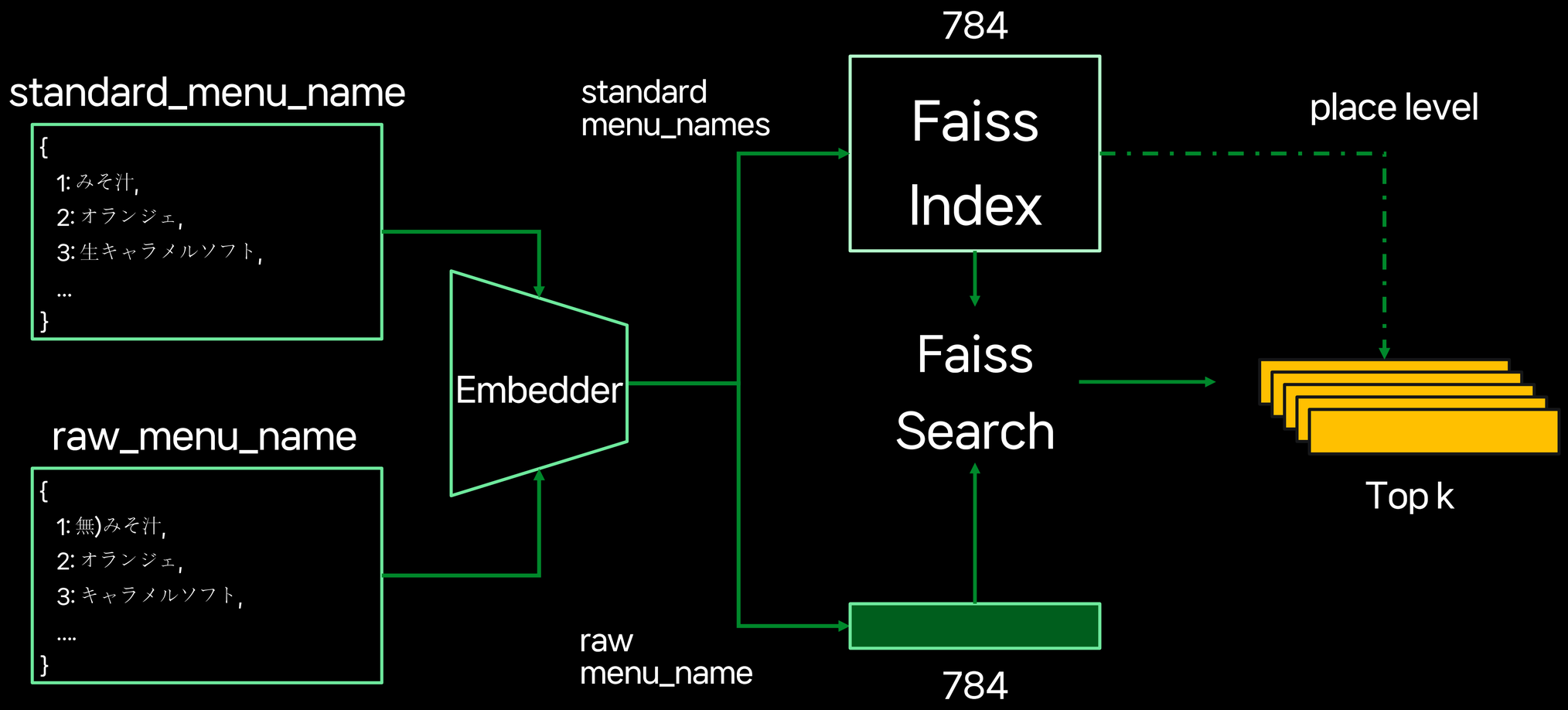


Supervised Contrastive

Supervised Contrastive Learning (NeurIPS 2020)

2.3 Menu Matching (inference)

NAVER
DEVIEW
2023



Supervised Contrastive Learning (NeurIPS 2020)

2.4 Results

NAVER
DEVIEW
2023

Methods	top1	top2	top3	top4
Baseline - SCL	0.9995	0.9999	1	1
MLM + CLM + SCL	0.9991	0.9999	0.9999	0.9999
MLM + CLM + SCL + second stage w pooler	0.9995	0.9999	0.9999	1
MLM + CLM + SCL + second stage w/o pooler	0.9999	0.9999	0.9999	1

Candidated models...

- COCO-LM [NeurIPS 2021]
- Pairwise supervised contrastive learning [EMNLP 2021]

COCO-LM: Correcting and contrasting text sequences for language model pretraining [NeurIPS 2021]

Pairwise Supervised Contrastive Learning of Sentence Representations [EMNLP 2021]

2.4 Results

NAVER
DEVIEW
2023

Methods	top1	top2	top3	top4
Baseline - SCL	0.9995	0.9999	1	1

Error case

- Place마다 standard menu name 상이
- 온도, 사이즈, 매장/포장 등등 다양한 option 표현
- Reranking logic의 필요성

```
1
2  # 1 사이즈, 온도
3  raw_menu_name:S.アイスコーヒー -> S.아이스커피
4  prediction:Sアイスコーヒー -> S아이스커피
5  standard_menu_name:コーヒー -> 커피
6  # 2 사이즈, 구성
7  raw_menu_name: チキンカレーSサイズ -> 치킨카레S사이즈
8  prediction: チキンカレーセット -> 치킨카레세트
9  standard_menu_name: チキンカレー -> 치킨 카레
10 # 3 수량
11 raw_menu_name: たこ焼 8ヶ2点 -> 타코야키 8개 2점
12 prediction: たこ焼 8ヶ -> 타코야키 8개
13 standard_menu_name:たこ焼 -> 타코야키
14 # 4 사이즈
15 raw_menu_name: 鶏親子丼(並) -> 오야꼬동(보통)
16 prediction: 鶏親子丼(大) -> 오야꼬동(대)
17 standard_menu_name:鶏親子丼 -> 오야꼬동
```

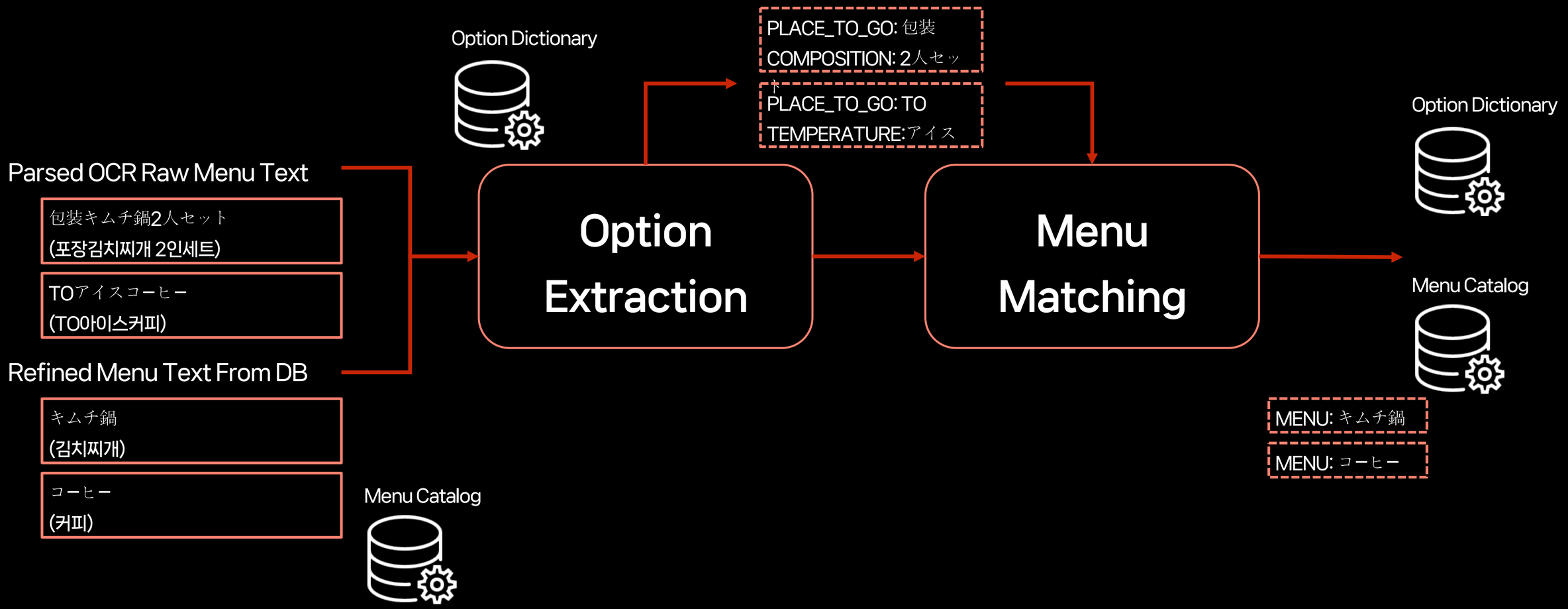
3. Option Extraction

CONTENTS

3.1 Background

NAVER
DEVIEW
2023

GOAL: 사용자 인입 메뉴명의 옵션을 추출하자



3.1 Background

NAVER
DEVIEW
2023

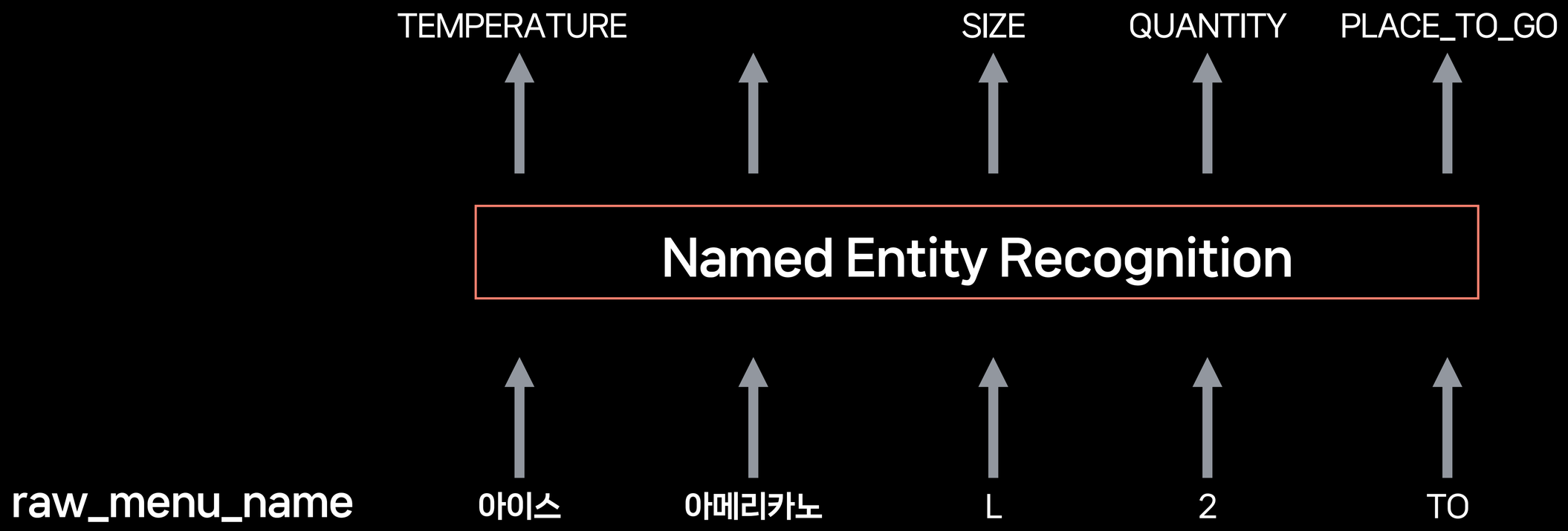
다양한 Option의 종류

- COMPOSITION
- PROMOTION
- SIZE
- WEIGHT
- QUANTITY
- PLACE_TO_GO
- TEMPERATURE

```
data = {
  "standard_menu_name": "カフェラテ", -> 카페라떼
  "raw_menu_names": [
    "カフェラテICE", -> 카페라떼 ICE
    "カフェラテ(アイス)", -> 카페라떼(아이스)
    "カフェラテHOT", -> 카페라떼HOT
    "Sカフェラテ(H)", -> S카페라떼(H)
    "カフェラテ HOT", -> 카페라떼 HOT
    "10202カフェラテ", -> 10202 카페라떼
    "冬得半額Mアイスカフェラテ", -> 동득반값M아이스카페라떼
    "EI アイス カフェフ", -> EI 아이스 카페프
    "カフェラテ レギュラー", -> 카페라떼 레귤러
    "Mサイズ カフェラテ", -> M사이즈 카페라떼
    "カフェラテ-HOT", -> 카페라떼-HOT
    "[H]カフェラテ", -> [H]카페라떼
    "カフェラテ(H/I)", -> 카페라떼(H/I)
    "L·カフェラテ", -> L.카페라떼
    "Lアイスカフェラテ", -> L아이스카페라떼
    "カフェラテアイス", -> 카페라떼 아이스,
    ...
  ]
}
```

3.1 Sequence Labeling (NER)

NAVER
DEVIEW
2023

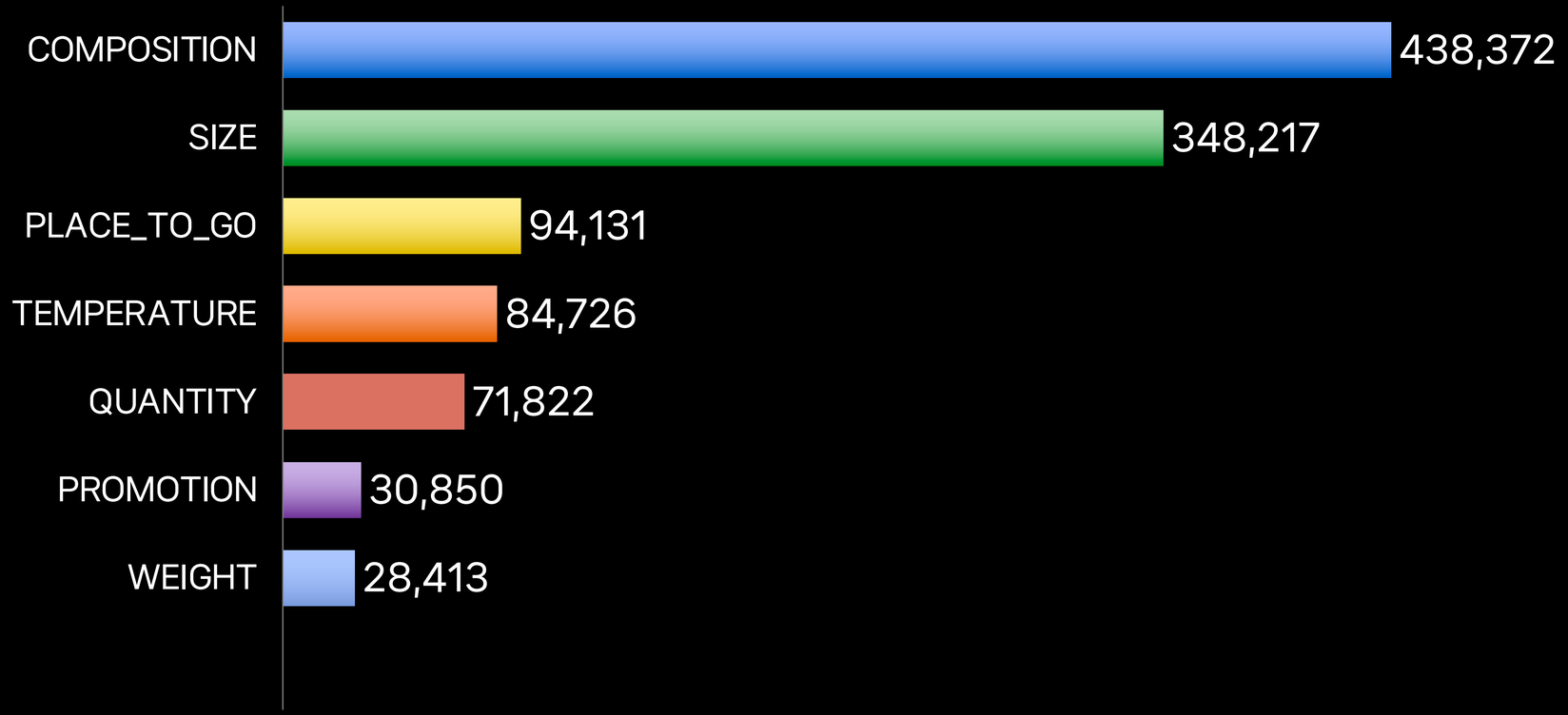


3.2 Dataset

NAVER
DEVIEW
2023

Rule Base Option labels (1,179 patterns)

- 총 43만 place의 394만 메뉴명에 대한 428만 건의 raw_menu_name중에서
- Option이 존재하는 33만 건 + random sample 5만 건

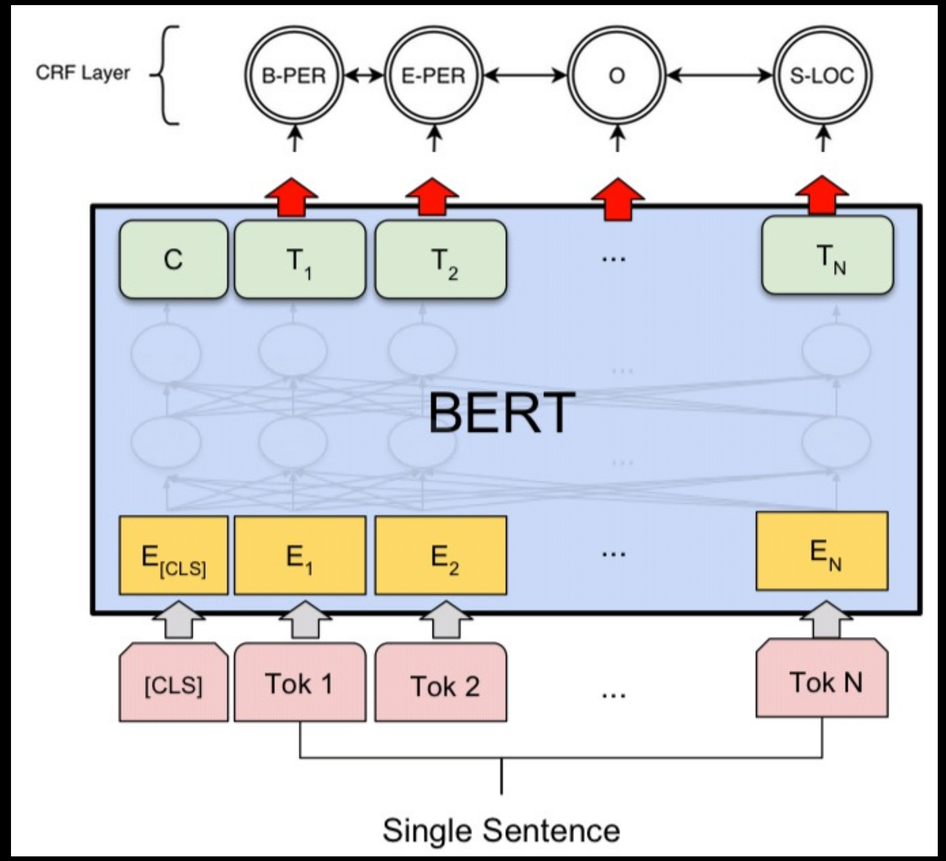


3.3 Option Extraction model

NAVER
DEVIEW
2023

Baseline: mBERT(base) + CRF

- LaRva Japanese models
- Roberta models
- Base, Large, Char,...



3.4 Result(test set)

NAVER
DEVIEW
2023

Baseline: mBERT(base) + CRF

```
# 아이스 커피
{"raw_menu_name": "S.アイスコーヒー"},
{"text": "S", "start": 0, "end": 1, "entity": "SIZE"},
{"text": "アイス", "start": 2, "end": 5, "entity": "TEMPERATURE"},
# 치킨카레S사이즈
{"raw_menu_name": "チキンカレーSサイズ"},
{"text": "Sサイズ", "start": 6, "end": 10, "entity": "SIZE"},
# 타코야끼8개2점
{"raw_menu_name": "たこ焼き8個2点"},
{"text": "8個", "start": 4, "end": 6, "entity": "QUANTITY"},
{"text": "2点", "start": 6, "end": 8, "entity": "QUANTITY"},
# 오야꼬동(보통)
{"raw_menu_name": "おやこ洞（普通）"}
{"text": "普通", "start": 5, "end": 7, "entity": "SIZE"},
```

	precision	recall	f1-score	support
COMPOSITION	0.69	0.77	0.73	963
PLACE_TO_EAT	0.83	0.91	0.87	280
PROMOTION	0.89	0.86	0.87	372
QUANTITY	0.94	0.94	0.94	282
SIZE	0.86	0.89	0.87	572
TEMPERATURE	0.83	0.94	0.88	332
WEIGHT	0.86	0.95	0.91	265
micro avg	0.81	0.87	0.84	3066
macro avg	0.84	0.90	0.87	3066
weighted avg	0.81	0.87	0.84	3066

3.5 Reranking logic

NAVER
DEVIEW
2023

Test set

- Menu Matching의 valid set에서 top1 prediction 실패 케이스 (110개)
- 110개의 raw_menu_name을 지닌 data pair를 valid set에서 resample -> 400 건

Methods	top1	top2	top3	top4
Baseline - SCL	0.733	0.993	1	1
Remove option (raw_menu_name)	0.765	0.985	0.995	1
Remove option (standard_menu_name)	0.718	0.975	0.993	1
Remove option (both)	0.73	0.975	0.99	1

3.5 Reranking logic

NAVER DEVIEW 2023

```
before = {
    "raw_menu_name": "チキンカレーSサイズ",
    "standard_menu_name": "チキンカレー",
    "gt": 2,
    "cand": [
        {"cand_1_menu_name": "チキンカレーセット", "cand_1_cosine": 0.867},
        {"cand_2_menu_name": "チキンカレー", "cand_2_cosine": 0.832},
        {"cand_3_menu_name": "ポーク&キーマ", "cand_3_cosine": -0.109},
    ],
}
```

```
after = {
    "raw_menu_text_name": "チキンカレー",
    "standard_menu_name": "チキンカレー",
    "gt": 1,
    "cand": [
        {"cand_1_menu_name": "チキンカレー", "cand_1_cosine": 1},
        {"cand_2_menu_name": "チキンカレーセット", "cand_2_cosine": 0.704},
        {"cand_3_menu_name": "ポーク&キーマ", "cand_3_cosine": -0.077},
    ],
}
```

```
before = {            # 도넛 뷔페(프리드링크 포함) 성인
    "raw_menu_name": "ドーナツビュッフェ(フリードリンク付)大人",
    "standard_menu_name": "桜ドーナツ",  # 벚꽃 도넛
    "gt": 2,
    "cand": [
        {"cand_1_menu_name": "むぎゅっとドーナツオリジナル", "cand_1_cosine": 0.312},
        {"cand_2_menu_name": "桜ドーナツ", "cand_2_cosine": 0.23},
        {"cand_3_menu_name": "ホットセイボリーパイBBQフランクフルト", "cand_3_cosine": 0.187},
        {"cand_4_menu_name": "ブレンドコーヒー", "cand_4_cosine": 0.12},
    ],
}
```

```
after = {
    "raw_menu_name": "ドーナツ()",  # 도넛
    "standard_menu_name": "桜ドーナツ",
    "gt": 1,
    "cand": [
        {"cand_1_menu_name": "桜ドーナツ", "cand_1_cosine": 0.453},
        {"cand_2_menu_name": "むぎゅっとドーナツオリジナル", "cand_2_cosine": 0.435},
        {"cand_3_menu_name": "ブレンドコーヒー", "cand_3_cosine": 0.016},
        {"cand_4_menu_name": "ホットセイボリーパイBBQフランクフルト", "cand_4_cosine": -0.05},
    ],
}
```

3.5 Reranking logic

NAVER
DEVIEW
2023

Tuning 1

- Option으로만 이루어진 경우, 모든 token이 삭제됨
- 길이에 대한 lower bound 추가

```
# lower bound
if len(new_text)/len(raw_text) >= alpha:
    return new_text
else:
    return raw_text
```

Methods	top1	top2	top3	top4
Baseline - SCL	0.733	0.993	1	1
Remove option (raw_menu_name)	0.765	0.985	0.995	1
Remove option (raw_menu_name, alpha=0.2)	0.765	0.993	1	1
-	-	-	-	-

3.5 Reranking logic

NAVER
DEVIEW
2023

Tuning 2

- Option의 부재 (45%)
- 히라가나, 가타카나, 한자어, 영어 표현 이음동의어 문제

```
{ 히라가나 라면:らぁめん, 가타가나 라면: ラーメン
  "raw_menu_name" : "納豆らぁめん",
  "standard_menu_name": "なっとうラーメン" (cos_sim: 0.565),
  "model_top_1": "らぁめん" (cos_sim: 0.695),
}

{ 히라가나+한자어 라면:らぁ麺, 가타가나 라면: ラーメン
  "raw_menu_name" : "特製醤油らぁめん",
  "standard_menu_name": "特製醤油ラーメン" (0.775),
  "model_top_1": "特製醤油らぁ麺" (0.904),
}

{ 가타가나 초콜릿케이크: チョコレートケーキ"
  "raw_menu_name" : "CHOCOLATECAKE",
  "standard_menu_name": "チョコレートケーキ" (0.422),
  "model_top_1": "CHICKENTURKISHSET" (0.482),
}
```

3.5 Reranking logic

NAVER
DEVIEW
2023

Tuning 2

- Option의 부재 (45%)
- 히라가나, 가타카나, 한자어, 영어 표현 -> 히라가나로 표현 통일

Methods	top1	top2	top3	top4
Baseline - SCL	0.733	0.993	1	1
Remove option (raw_menu_name)	0.765	0.985	0.995	1
Remove option (raw_menu_name, alpha=0.2)	0.765	0.993	1	1
Remove option (raw_menu_name, alpha=0.2, hiragana)	0.810	0.988	0.988	1

Q & A

Thank You