# 의미론적 feature 공간상에서의 negative sampling을 통한 검색 성능 개선

HCLT 2022

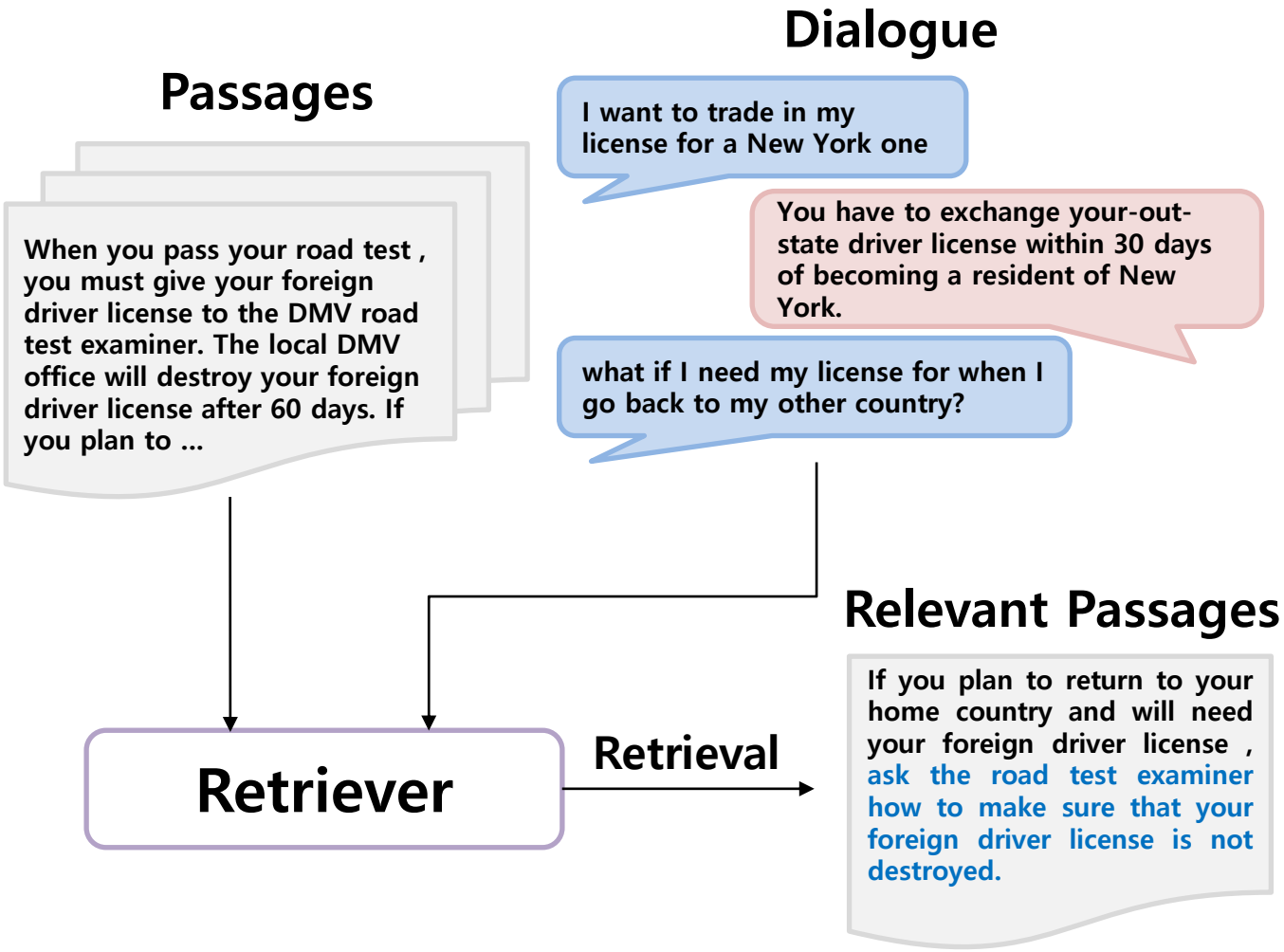
이정두[1], 홍범석[2], 최원석[2], 한영섭[2], 전병기[2], 나승훈[1]

[1] 전북대학교 [2] LG U+

# 목 차

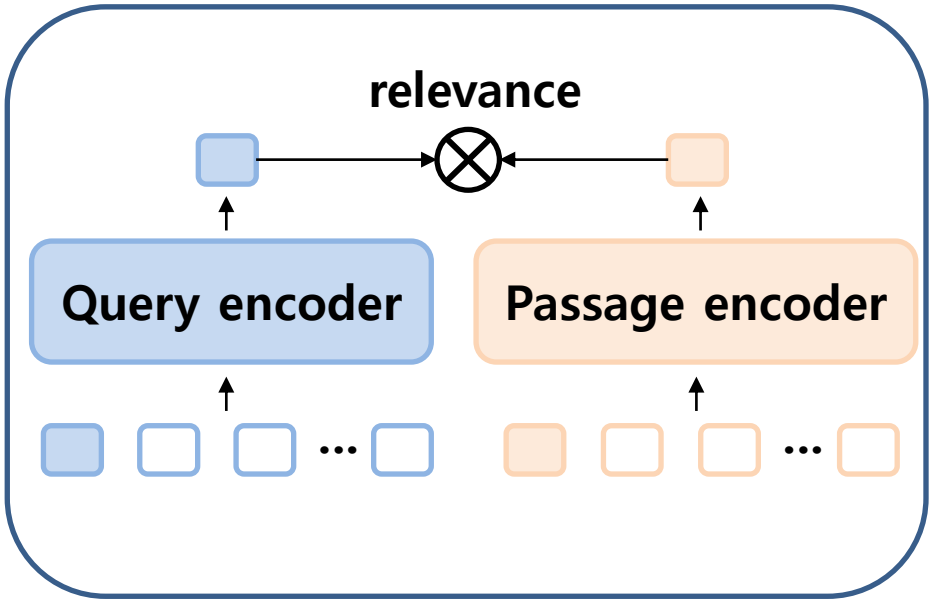
# Introduction

- 개요도



# Introduction

- **Retriever**
  - query 와 passage 를 독립적으로 encoding




$$sim_\theta(q,d) = E_\theta(q)^T E_\theta(d)$$

$$\mathcal{L} = -\log \frac{e^{sim(q^i,d^i)}}{e^{\mathrm{sim}(q^i,d^i)} + \sum_{j=1,j\neq i}^{n} e^{\mathrm{sim}(q^i,d^j)}}$$

# 관련 연구

- **기존 Negative sampling 전략**

  - **in-batch**
    배치 내에 각 example 이 {q, p+} 로 positive pair 로 이루어 졌을 때, 자기 자신을 제외한 passage 를 negative 로 두는 방법

  - **TF-IDF (or BM25)**
    통계 기반 검색을 통해 얻은 topk 개의 passage 를 negative sample 로 사용하는 방법

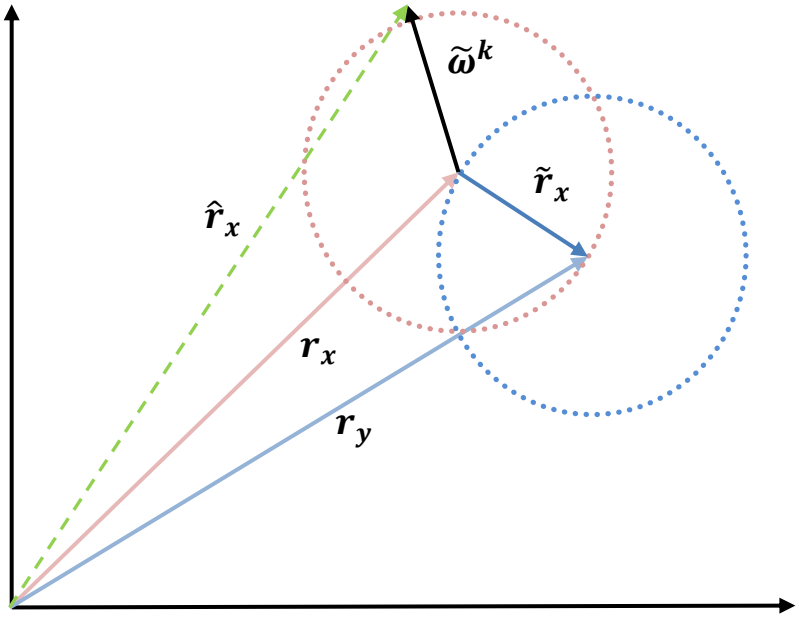
  - **ANCE [Xiong et al. '20][1]**
    훈련을 진행하며 전체 passage 를 다시 encoding 하여 계속해서 모델이 어려워하는 hard negative 를 얻는 방법

[1]Approximate nearest neighbor negative contrastive learning for dense text retrieval[Xiong et al '20]

# 제안 방법

- **Feature space 에서의 negative sampling**

  - **수정된 Mixed Gaussian Recurrent Chain(MGRC) sampling**
    positive feature 의 semantic space 의 경계선상에서 hard negative sample 을 얻는 방법




$(x, y)$: positive pair

$\boldsymbol{r_x}$: query 의 중심 vector

$\boldsymbol{r_y}$: positive passage 의 중심 vector

$\tilde{\boldsymbol{r}}_x$: query 에 대한 semantic space 반경

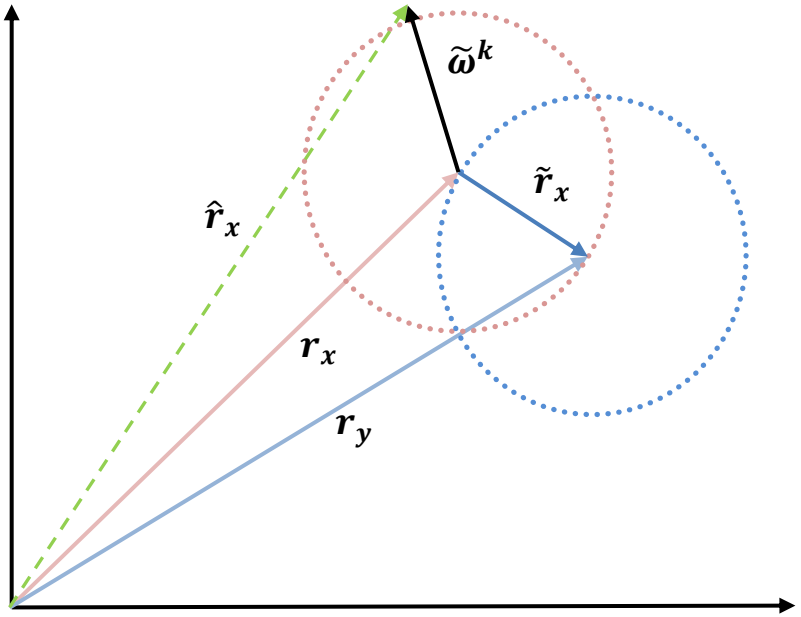
$\hat{\boldsymbol{r}}_x$: negative sample vector

Learning to Generalize to More: Continuous Semantic Augmentation for Neural Machine Translation[Wei et al '20]

# 제안 방법

- **Feature space 에서의 negative sampling**

  - ➢ **수정된 MGRC sampling**




$$\hat{r}_x^k = r_x + \tilde{\omega}^k \odot \tilde{r}$$

$$\tilde{r} = r_y - r_x$$

$$\tilde{\omega}^k = \mathrm{sgn}(\omega^k)(|\omega^k| + 1)$$

$$\begin{cases} \omega^k \sim \mathcal{N}(0, \mathrm{diag}(\mathcal{W}_r^2)) & \text{if } k = 1 \\ \omega^k \sim p(\omega|\omega^1, \cdots, \omega^{k-1}) & \text{if } k > 1 \end{cases}$$

$$p = \eta\mathcal{N}(0, \mathrm{diag}(\mathcal{W}_r^2)) + (1.0 - \eta)\mathcal{N}\left(\frac{1}{k-1}\sum_{i=1}^{k-1}\omega^i, 1\right)$$

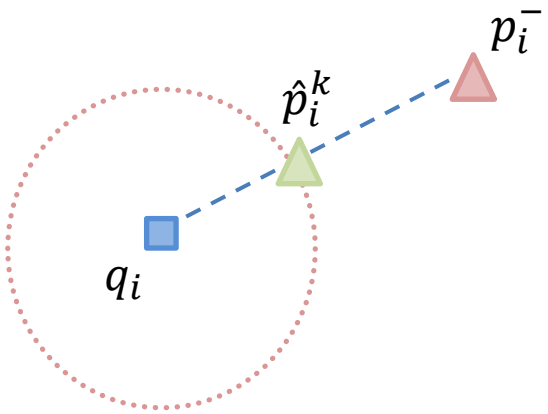
$$\mathcal{W}_r = (1 - \lambda)\frac{|\tilde{r}| - \min(|\tilde{r}|)}{\max(|\tilde{r}|) - \min(|\tilde{r}|)}$$

$$\lambda = \frac{1}{100}\sum_{j \in [-100, -1]} e^{-\mathcal{L}^j}$$

# 제안 방법

- **Feature space 에서의 negative sampling**

  - **Feature space mixing[Kalantidis et al '20][2]**
    encoding 된 positive feature 와 negative feature 를 linear-interpolation 시켜 hard negative sample 을 얻는 방법




$$\hat{p}_i^k = \frac{\tilde{p}_i^k}{\|\tilde{p}_i^k\|_2}, where\ \tilde{p}_i^k = \alpha_k p_i^+ + (1-\alpha_k)p_i^-$$

$$\alpha_k \in (0, 0.5)$$

$p_i^+$: $q_i$에 대한 positive passage feature
$p_i^-$: $q_i$에 대한 negative passage feature

[2]Hard Negative Mixing for Contrastive Learning[NeurIPS 2020]

# 데이터

- **Multidoc2dial**
  - 문서 기반 대화 데이터

- **데이터 통계**

| Split | Instance Num | Passage Num |
|---|---|---|
| Train | 21451 | 3820 |
| Validation | 4201 | |
| Test | 4094 | |

# 실 험

- 실험 세팅

| | Baseline-ours | Ours | Others |
|---|---|---|---|
| **Batch_size** | 128 | 128 | 128 |
| **Learning rate** | 2e-5 | 2e-5 | - |
| **Encoder** | Roberta-base | Roberta-base | Bert-base |
| **Shared-encoder** | O | O | X |
| **In-batch negative** | O | O | O |
| **TF-IDF(or BM25)** | X | X | O |
| **Feature space mixing** | X | O | X |
| **MGRC sampling** | X | O | X |

# 실 험

- **실험 결과**

| Performance on validation set | | | | | |
|---|---|---|---|---|---|
| **Model** | **Architecture** | **R@1** | **R@5** | **R@10** | **R@100** |
| Baseline-official | DPR | 49.0 | 72.3 | 80.0 | - |
| [CPII-NLP] – 1st | DPR | 44.5 | 71.4 | - | - |
| [zsw_dyy_lgz] – 2nd | DPR | 42.8 | 68.0 | 77.1 | - |
| [CMU-QA] – 4th | DistilSPLADE | - | - | 78.6 | 94.9 |
| Baseline-Ours | DPR | 51.1 | 77.9 | 85.2 | 96.9 |
| Ours | DPR | **51.8** | **78.8** | **85.6** | **97.4** |

- Passage-retriever 를 사용한 팀들과 비교
- 1, 2, 4위 팀은 Test Phase of SEEN leaderboard 상에서의 순위

1st - Grounded Dialogue Generation with Cross-encoding Re-ranker, Grounding Span Prediction, and Passage Dropout[Li et al '22]
2nd - G4: Grounding-guided Goal-oriented Dialogues Generation with Multiple Documents[Zhang et al '22]
4th - R3 : Refined Retriever-Reader Pipeline for Multidoc2dial[Bansal et al '22]

# 실 험

- **실험 결과**

| Performance on validation set | | | | |
|---|---|---|---|---|
| | **R@1** | **R@5** | **R@10** | **R@100** |
| Ours | 51.8 | 78.8 | 85.6 | 97.4 |
| w/o Feature space mixing | 52.2 | 78.6 | 85.4 | 97.0 |
| w/o MGRC sampling | 51.1 | 77.9 | 85.4 | 97.1 |

| Performance on test set | | | | | |
|---|---|---|---|---|---|
| | **R@1** | **R@5** | **R@10** | **R@100** | **MRR** |
| Baseline-Ours | 50.5 | **80.0** | **87.4** | 97.7 | 63.4 |
| Ours | **51.4** | 79.2 | 86.9 | **97.9** | **63.8** |

# 감사합니다.