Second place solution of Amazon KDD Cup 2022: ESCI Challenge for Improving Product Search

Team qinpersevere:
Xiaolei Qin
Nan Liang
Hongbo Zhang
Wuhe Zou
Weidong Zhang

Presenter:
Xiaolei Qin

网易互娱 AILab

Introduction

- **Dataset**: Shopping Queries Data Set

  1. a large dataset of difficult search queries and products with ESCI relevance judgements(Exact, Substitute, Complement, Irrelevant)

  2. multilingual, including English, Japanese and Spanish

  3. task2 and task3 share the same larger dataset and task1 is a reduced version of what is deemed to be “easy” queries

- **Three Tasks:**

  Given a user specified query

  1. Task1: Query-Product Ranking(NDCG)
     rank the products by semantic matching

  2. Task2: Multiclass Product Classification(micro-F1)
     classify each product as being an Exact, Substitute, Complement or Irrelevant match

  3. Task3: Product Substitute Identification(micro-F1)
     identify the substitute products

网易互娱 AILab

Pretrained language model

- US

cross-encoder/ms-marco-MiniLM-L-12-v2
bert-large-uncased
roberta-large
microsoft/deberta-v3-base
microsoft/deberta-v3-large

- JP

rinna/japanese-roberta-base
cl-tohoku/bert-base-japanese
cl-tohoku/bert-large-japanese

- multilingual

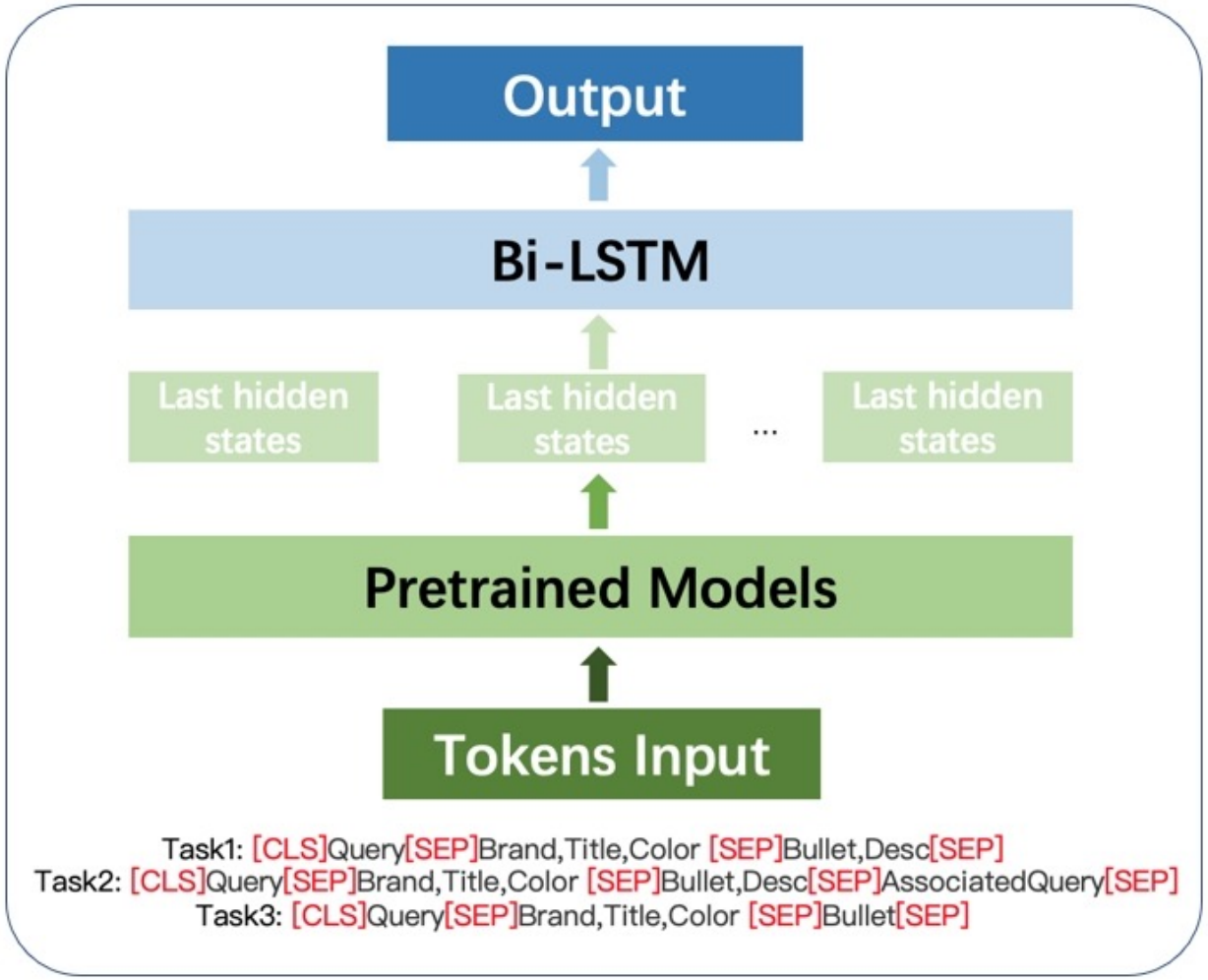
bert-base-multilingual-cased
MoritzLaurer/mDeBERTa-v3-base-mnli-xnli
microsoft/mdeberta-v3-base
xlm-roberta-large

- ES

dccuchile/bert-base-spanish-wwm-uncased
bertin-project/bertin-roberta-base-Spanish
CenlA/albert-large-spanish

网易互娱 AILab

Model Architecture



Model	Public	Private
microsoft/deberta-v3-base microsoft/mdeberta-v3-base(5 folds)	0.888	
microsoft/deberta-v3-large microsoft/mdeberta-v3-base(1 of 5folds)	0.8946	0.8919
microsoft/deberta-v3-large microsoft/mdeberta-v3-base(5 folds)	0.9003	0.8985
microsoft/deberta-v3-large microsoft/mdeberta-v3-base(4 of 5folds) + roberta-large xlm-roberta-large(2 of 5folds)	0.9019	0.9002

Exact -> 1.0

Substitute -> 0.1

Complement -> 0.01

Irrelevant -> 0.0

网易互娱 AILab

Don’t Stop Pretraining

```
python run_mlm.py \
    --model_name_or_path roberta-base \
    --train_file path_to_train_file \
    --validation_file path_to_validation_file \
    --per_device_train_batch_size 8 \
    --per_device_eval_batch_size 8 \
    --do_train \
    --do_eval \
    --output_dir /tmp/test-mlm
```

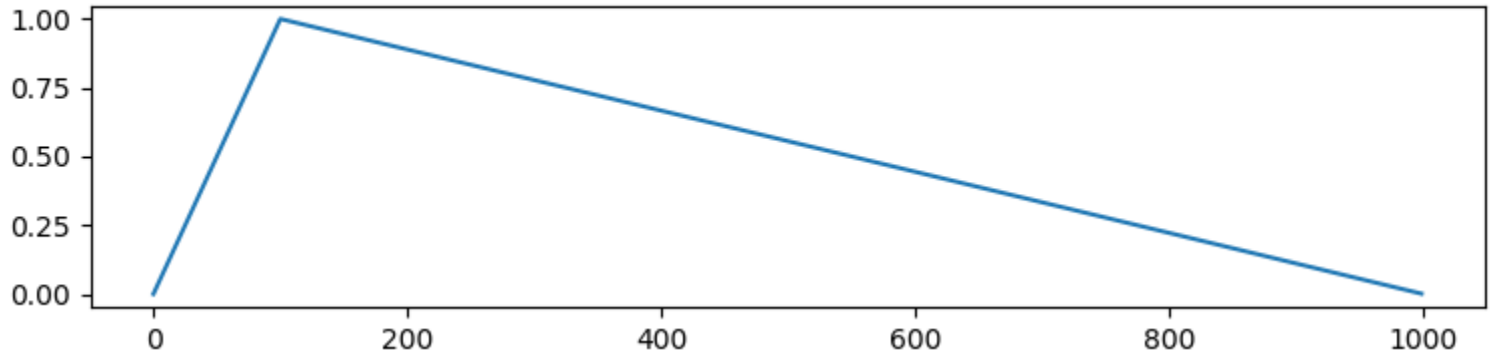
- 3 epochs
- (Query</s>)title,bullet,desc
- us pretrain
- jp + es pretrain
- us + jp + es pretrain

Model	Public	Private
microsoft/deberta-v3-large microsoft/mdeberta-v3-base(4 of 5 folds) + roberta-large xlm-roberta-large(2 of 5 folds)	0.9019	0.9002
1 + Roberta pretrain	0.9017	0.9026
us: deberta-large(4 of 5folds) + Roberta-large(4 of 5folds) **jp:** mdeberta(2 of 5folds) **es:** mdeberta(2 of 5folds) **jpes:** mdeberta(5 folds) **usjpes:** xlm-Roberta-large(5 folds)	**0.9047**	**0.9036**

网易互娱 AILab

Training Strategies

- **Freeze embeddings**
- **Different learning rates**
  deberta-large: 7e-6 others: 2e-5 output layers: 1e-3
- **Linear schedule with warmup**



- **Inference acceleration**: Sort by length

网易互娱 AILab

Training Strategies

- Multi-task learning

Task2

$$L = 0.5 * L_{esci} + 0.125 * (L_e + L_s + L_c + L_i)$$

Task3

$$L = 0.5 * L_s + 0.5 * L_esci$$

Discussion

- **Data augmentation**
  translation(jp->us, es->us)
  Text generation(use bart to generate query from product title and bullet point)

- **Keywords/key sentences of long texts**

- **Cross features**

Query	Product	Label
q1	P1	exact
q1	p2	substitute
q1	p3	irrelevant
q2	p4	complement
q2	p5	exact
q2	p1	exact

q2

Query	Product	Brand	Label
q1	P1	b1	exact
q1	p2	b2	substitute
q1	p3	b3	irrelevant
q1	p4	b1	complement
q1	p5	b1	exact
q1	p6	b2	exact

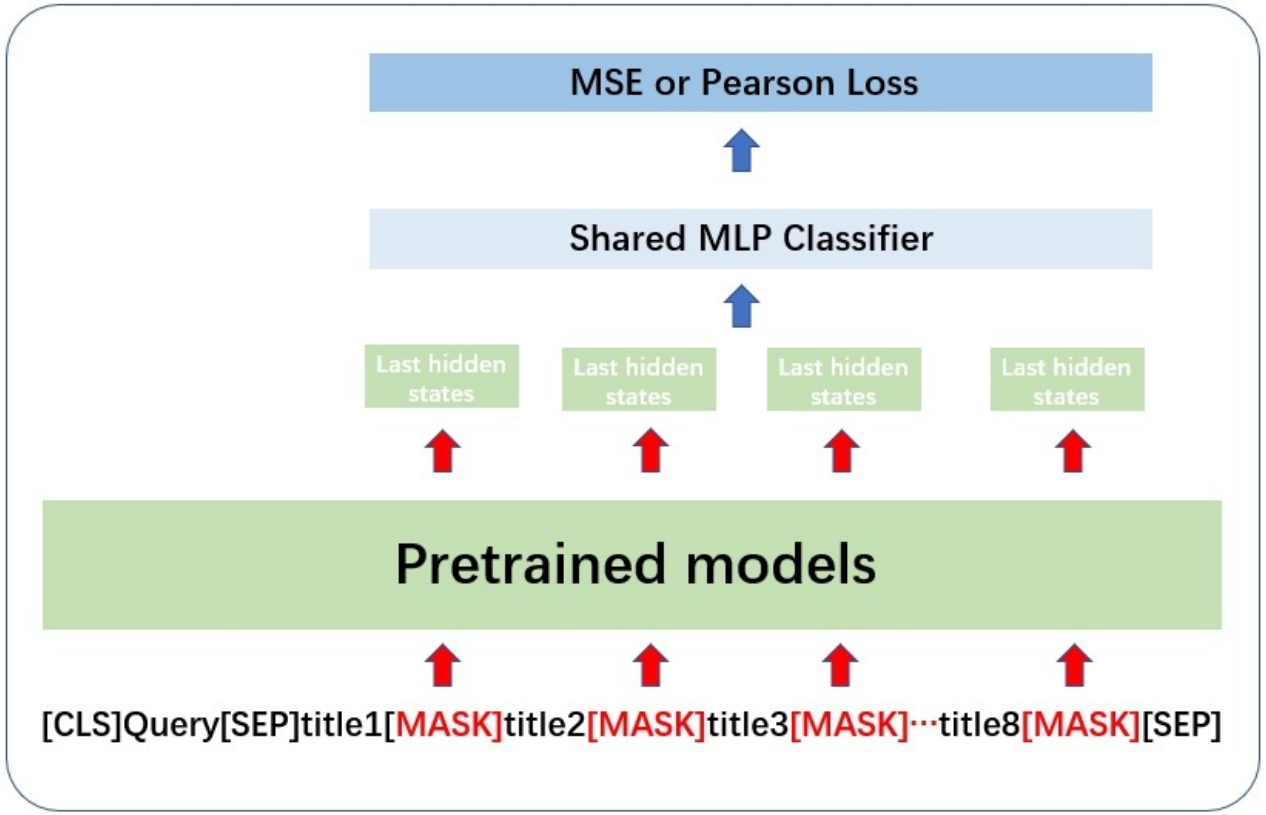
b1

网易互娱 AILab

Discussion

- **Pairwise loss**
  1. Slow to converge
  2. Sensitive to noise label

- **Postprocess**



网易互娱 AILab

Thank You

网易互娱
AILab