KDD CUP 2022
MULTICLASS PRODUCT CLASSIFICATION
TEAM MetaSoul SOLUTION

Zhichao Feng, Jiawei Lu, Junwei Cheng, Ke Hou, Kaiyuan Li, Pengfei Wang, Yadong Zhu
Beijing University of Posts and Telecommunications, Beijing Shuyuanling Technology Co., Ltd.

CONTENTS

Abstract

In this challenge, our team designed solutions the Task2 Multi-Product Classification, which gives queries and corresponding lists of products and requires contestants to classify each product as an exact, substitute, complementary, or irrelevant match for the query.

We utilized single tower BERT as our backbone and add tricks to it. The approaches we tried fall into three main categories: model methods, training methods and evaluation methods. We finally ensembled three optimal models in the code submission round and achieved micro-F1 score of 0.8207, which won the 5th place.

Methods

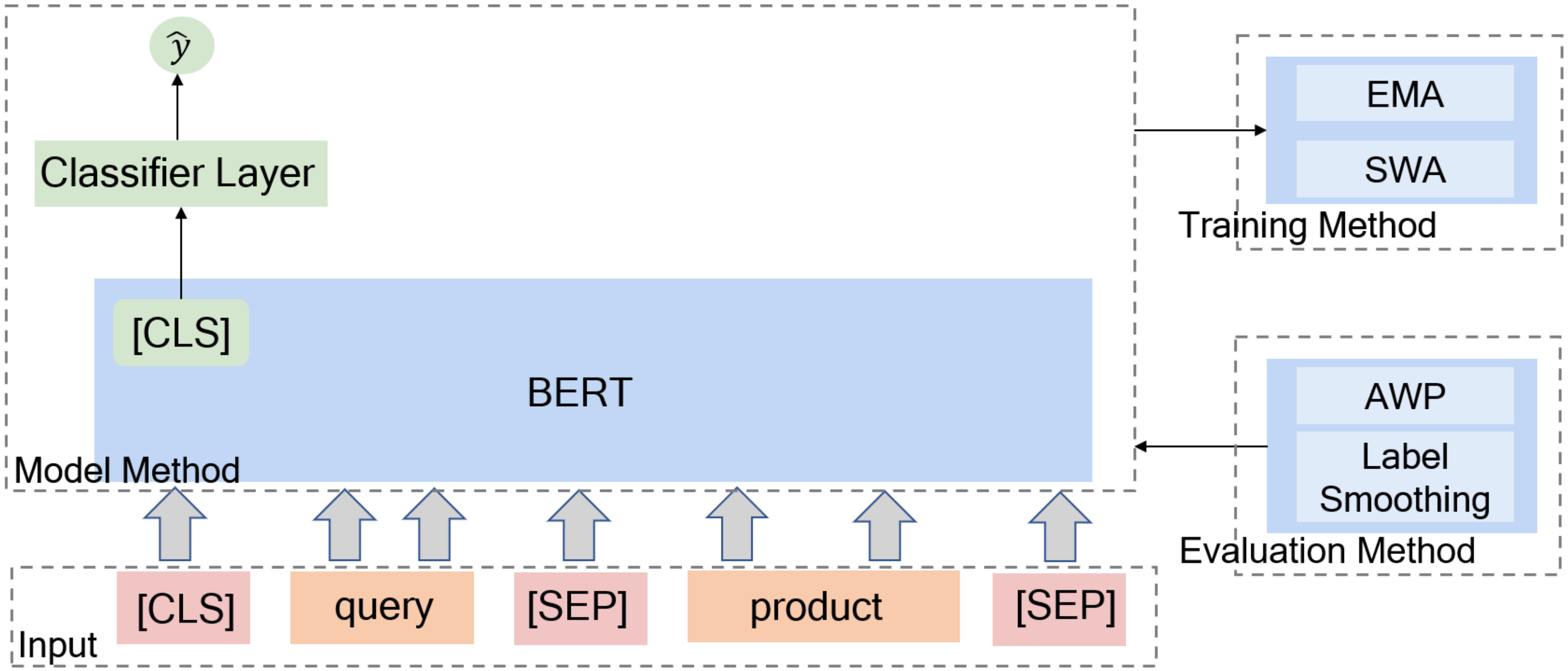


Figure 1: The structure of our model. Our backbone is a single-tower model. On the right of the figure are some of the tricks we tried. The output of classifier layer is probability of four classes.

Methods

- **Model Method**
  - DeBERTa
  - XLM-RoBERTa
  - Multi-Sample Dropout
- **Training Method**
  - Adversarial Weight Perturbation(AWP)
  - Label Smoothing
- **Evaluation Method**
  - Exponential Moving Averages(EMA)
  - Stochastic Weight Averaging(SWA)
- **Post Processing**
  - Model Parameters Ensemble

$$w_i \mapsto w_i + \delta w_i = w_i + \gamma \frac{\nabla_i}{\|\nabla_i\|} \|w_i\|, \quad |\delta w_i| \le \epsilon |w_i| \tag{1}$$

where w_i denotes to the i-th parameter of model, ∇_i denotes to the gradient of it.

Equation 1. The formula of AWP

$$v_t = \beta \cdot v_{t-1} + (1 - \beta) \cdot \theta_t \tag{2}$$

Equation 2. The formula of EMA

$$w_{SWA} \leftarrow \frac{w_{SWA} \cdot n_{models} + w}{n_{models} + 1} \tag{3}$$

Equation 3. The formula of SWA

Experiments

Table 1: Summary of the sampled dataset, including the number of unique queries, the number of judgements, and the average number of judgements per query.

Language	#Queries	#Judgements	Avg. Depth
English	7,479	140,220	18.7
Spanish	1,143	26,422	23.1
Japanese	1,378	33,446	24.3
US+ES+JP	10,000	200,088	20.0

Table 2: Summary of the full dataset, including the number of unique queries, the number of judgements, and the average number of judgements per query.

Language	#Queries	#Judgements	Avg. Depth
English	68,139	1,272,626	18.7
Spanish	10,624	249,721	23.5
Japanese	12,687	312,397	24.6
US+ES+JP	91,450	1,834,744	20.1

Table 3: Performance of mDeBERTa with different learning rates on sampled dataset.

learning rate	micro-F1(%)
1e-6	**74.11**
5e-6	74.05
2e-5	73.32

Table 4: Performance of models with different tricks on sampled dataset.

model	micro-F1(%)
mDeBERTa	74.05
mDeBERTa + EMA	74.15
mDeBERTa + SWA	74.18
mDeBERTa + AWP	74.99
mDeBERTa + Multi-Sample Dropout	74.26
mDeBERTa + Label Smoothing	74.16

Model Ensemble

Table 5: Offline micro-F1(%) of models on full dataset. SM represents single model and MPE represents model parameters ensemble.

model	SM	MPE
mDeBERTa + AWP	76.37	76.41
DeBERTa-Large(only English)	77.3	77.4
XLM-RoBERTa + multidropout	75.55	75.76

Table 6: Online performance of models on full dataset. Ensemble consist of the three models in Table 5.

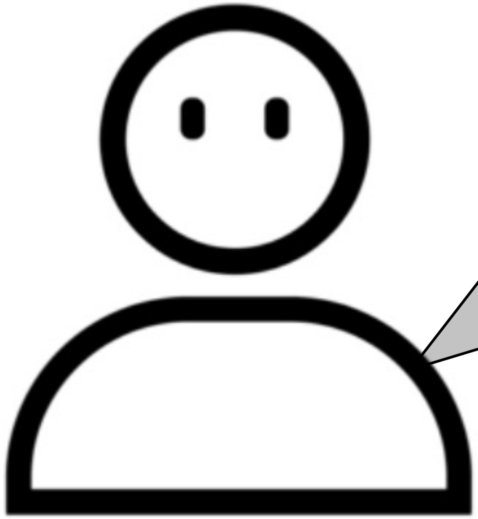
model	Public micro-F1(%)	Private micro-F1(%)
mDeBERTa + AWP	81.48	81.72
Ensemble	81.82	82.07

English Logits =

0.6*mDeBERTa + 0.4*DeBERTa-Large

Spanish and Japanese Logits =

0.7*mDeBERTa + 0.3*XLM-RoBERTa



Thank You！

Email: fengzc@bupt.edu.cn

Code release: https://github.com/guijiql/kddcup2022