A Semantic Alignment System for Multilingual Query-Product Retrieval

Qi Zhang, Zijian Yang, Yilun Huang, Zijian Cai, Ze Chen

Task Introduction

Task 1 of the KDD cup competition aims at ranking the query-product pairs by relevance. The dataset consists of English, Spanish and Japanese and is classified into ESCI categories

Framework

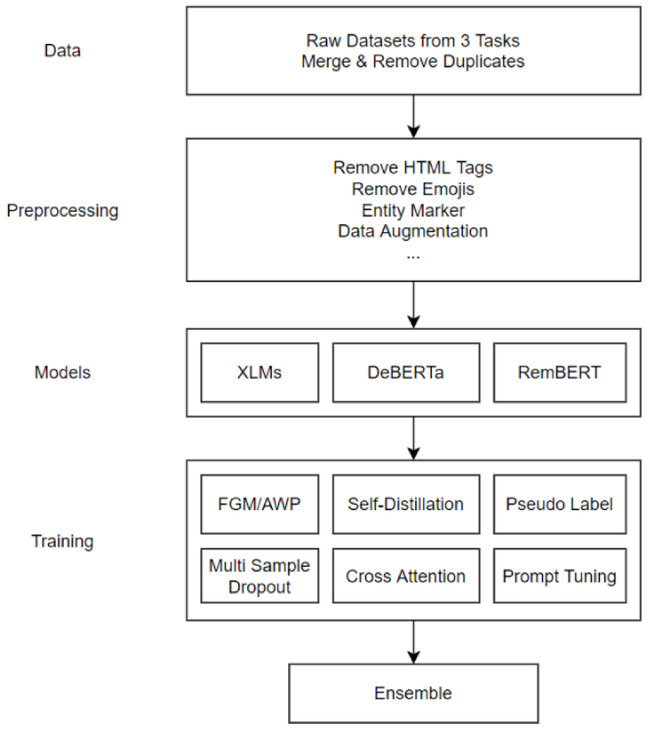


Figure 1. Overall framework

Basic Models

We use the cross-encoder architecture based models. Both multilingual and monolingual pre-trained LMs are adopted.

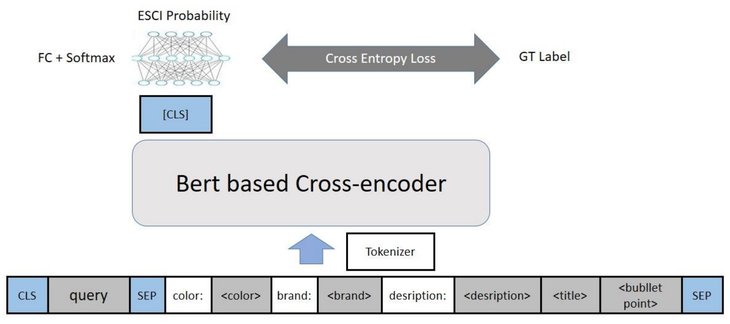
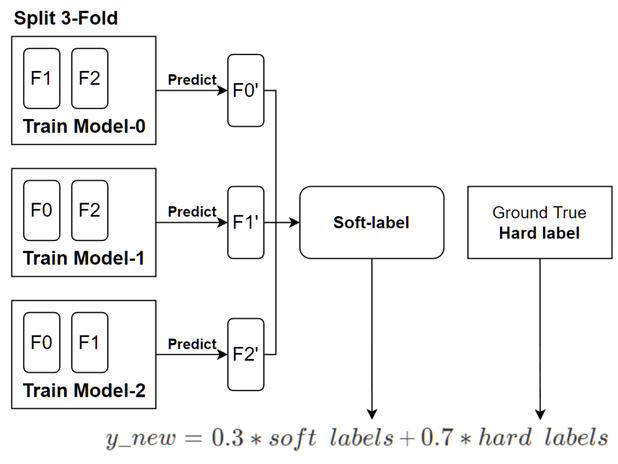


Figure 2. Basic model architecture

Self Distillation

To be specific, we use 3-fold bagging training and make prediction on the out-of-fold datasets to generate the soft labels. And then we merge the soft labels with the ground true hard labels with weights 0.3 and 0.7 to get the new training labels.



$$y\_new = 0.3 * soft\ \ labels + 0.7 * hard\ \ labels$$

Figure 3. self distillation

Pseudo Labeling

To avoid making the training data more noisy, only samples from the public test set with predicted probabilities above 0.7 are used as pseudo labels. And soft labels work better than hard labels during most of our experiments, we guess that hard labels may increase the risk of over-fitting.

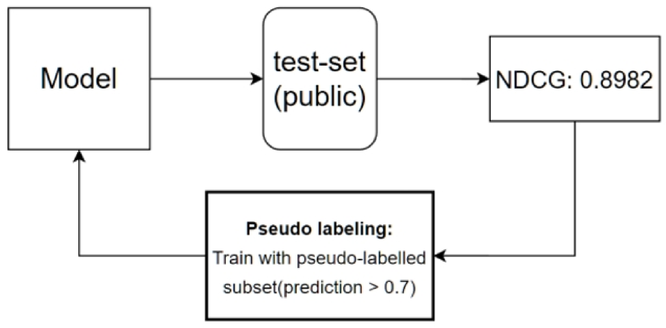


Figure 4. Train model with pseudo-labelled subset

Results

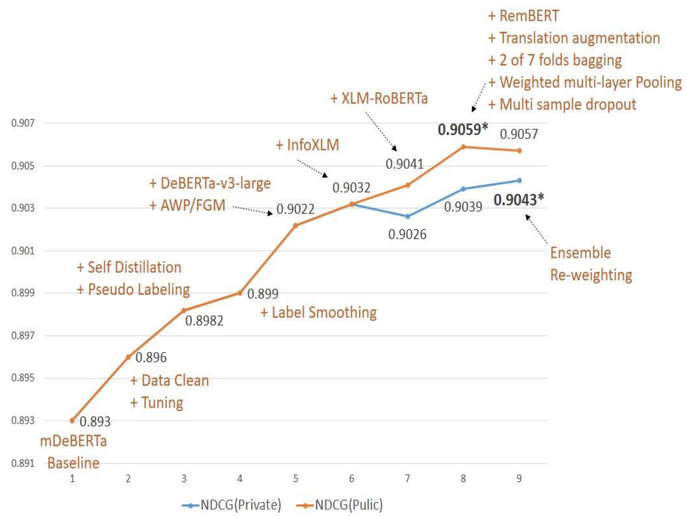


Figure 5. Results of different approaches

Summary and Future Work

- Summary
  - We use multilingual and English pre-trained LMs as backbone, with the combination of data processing and sorts of training optimization.
  - For single model, we achieve NDCG score of 0.9022 on the public leaderboard and 0.9015 on the private leaderboard.
  - At last, we do model ensemble to get the final boost from 0.9015 to 0.9043 on the private leaderboard, which ensures us to win the first place.
- Future Work
  - End to end multilingual model solution.

Acknowledgments

Amazon and AIcrowd organizing team paid a lot of efforts during the whole process of the competition, we really appreciate it for hosting this fantastic competition. And we would like to thank everyone associated with organizing and sponsoring the KDD Cup 2022.