A Semantic Alignment System for Multilingual Query-Product Retrieval

Team **www**

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

OPDAI, Interactive Entertainment Group of Netease Inc., Guangzhou, China

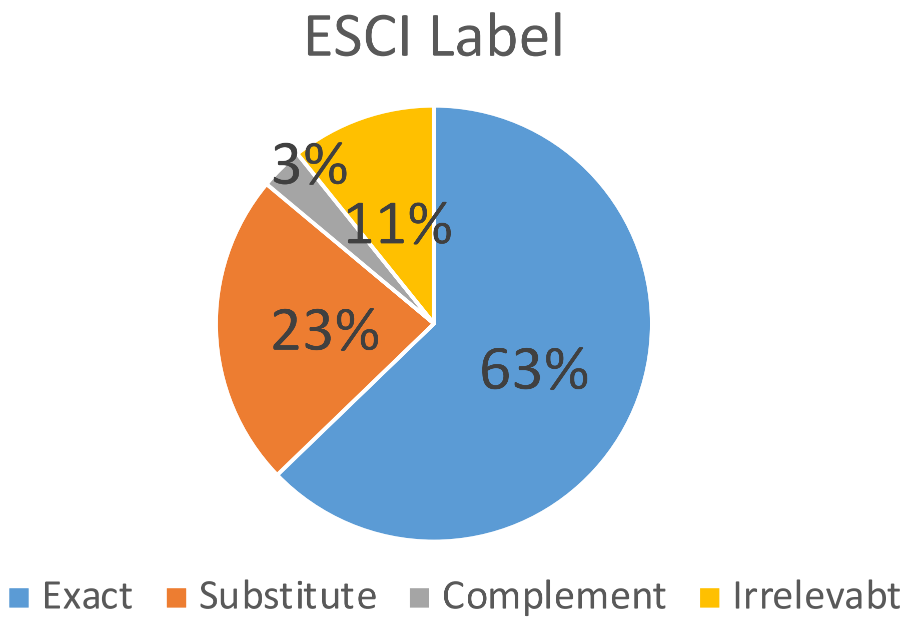
Outline

- Task Introduction
- Data Analysis
- Overall Framework
  - Data Processing
  - Model Architecture
- Basic models
- Training Optimization
- Results
- Summary and Future Work

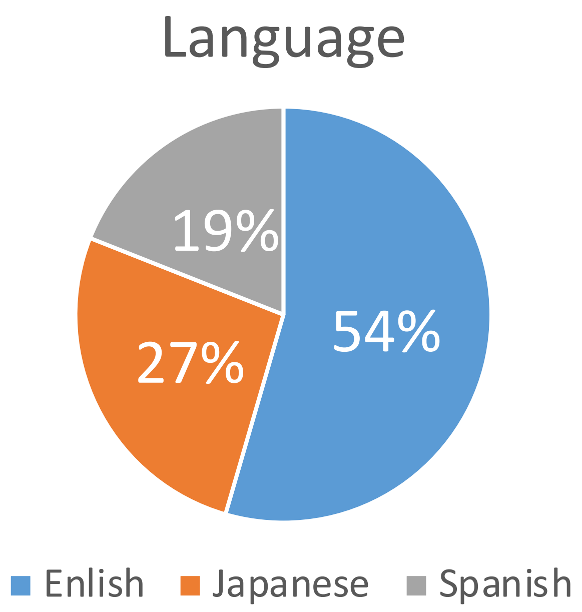
Task Introduction

- Task 1 aims at ranking the query-product pairs by **relevance**.
- Query-product dataset consists of **English, Spanish and Japanese**.
- Query-product dataset is classified into **Exact, Substitute, Complement,** or **Irrelevant** (ESCI) categories.
- Evaluation is based on **NDCG**.
- How to better understand the query-product **semantic relevance is the major challenge**.

Data Analysis



label distribution

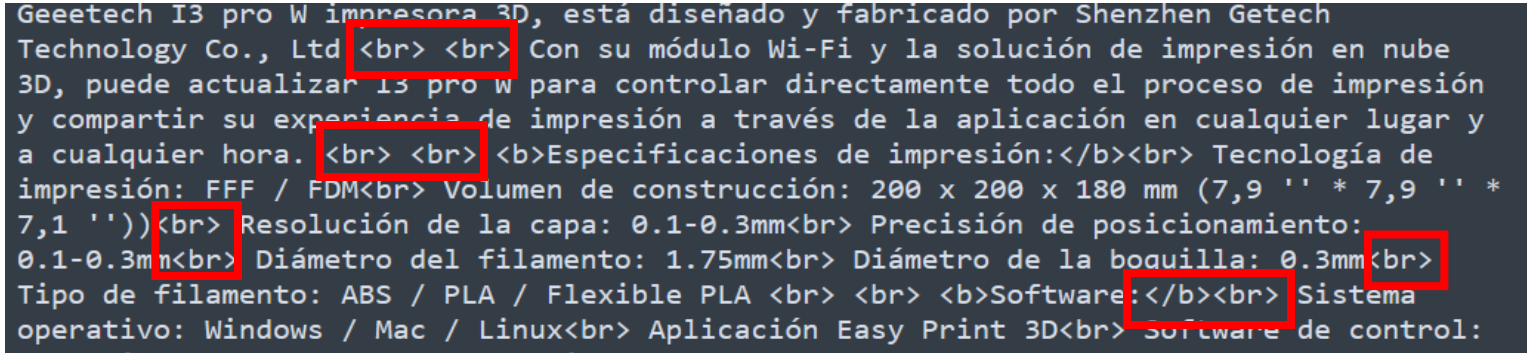


language distribution

- The label distribution and the language distribution are both imbalanced.
- And according to our statistics, 54% of product brands focus on providing only one product, while only 7.1% of brands provide more than 10 products.
- More than 80% of the color names are only customized for a single product.

Data Preprocess

- Remove those **HTML marks** and **emoji**.



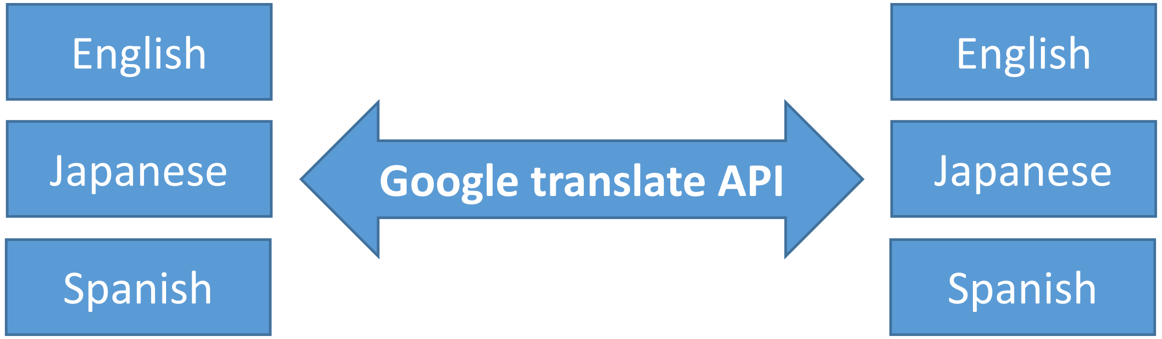
Geeetech I3 pro W impresora 3D, está diseñado y fabricado por Shenzhen Getech
Technology Co., Ltd

 Con su módulo Wi-Fi y la solución de impresión en nube
3D, puede actualizar I3 pro W para controlar directamente todo el proceso de impresión
y compartir su experiencia de impresión a través de la aplicación en cualquier lugar y
a cualquier hora.

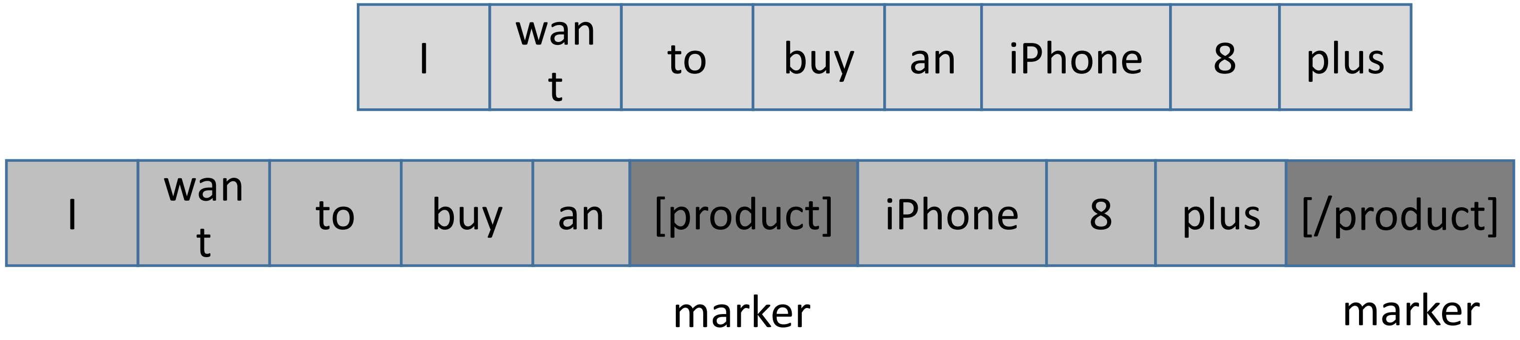
 Especificaciones de impresión:
 Tecnología de
impresión: FFF / FDM
 Volumen de construcción: 200 x 200 x 180 mm (7,9 '' * 7,9 '' *
7,1 ''))
 Resolución de la capa: 0.1-0.3mm
 Precisión de posicionamiento:
0.1-0.3mm
 Diámetro del filamento: 1.75mm
 Diámetro de la boquilla: 0.3mm
Tipo de filamento: ABS / PLA / Flexible PLA

 Software:
 Sistema
operativo: Windows / Mac / Linux
 Aplicación Easy Print 3D
 Software de control:

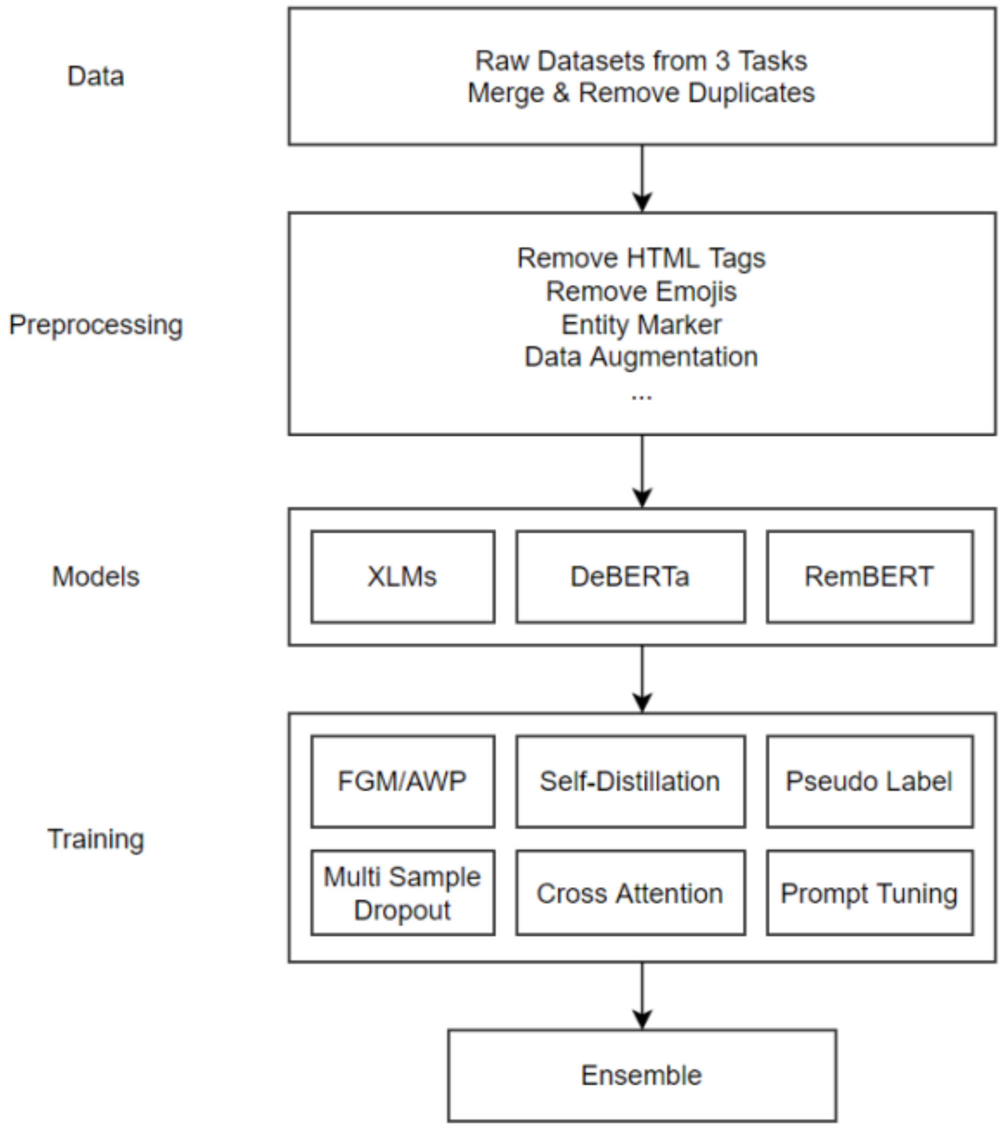
- Translate all of the data into English, Spanish and Japanese separately to do **data augmentation**.



- Incorporate the NER information using **Entity Marker**.

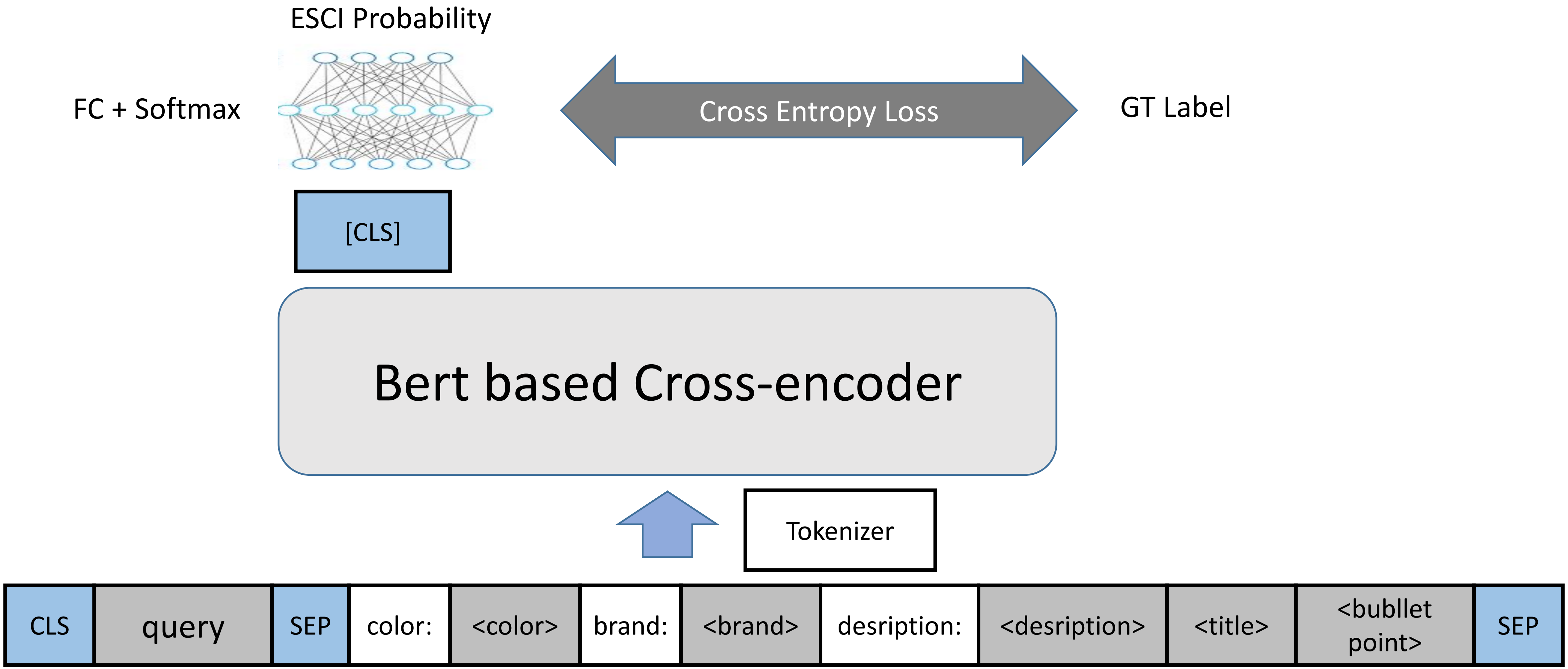


Overall Framework

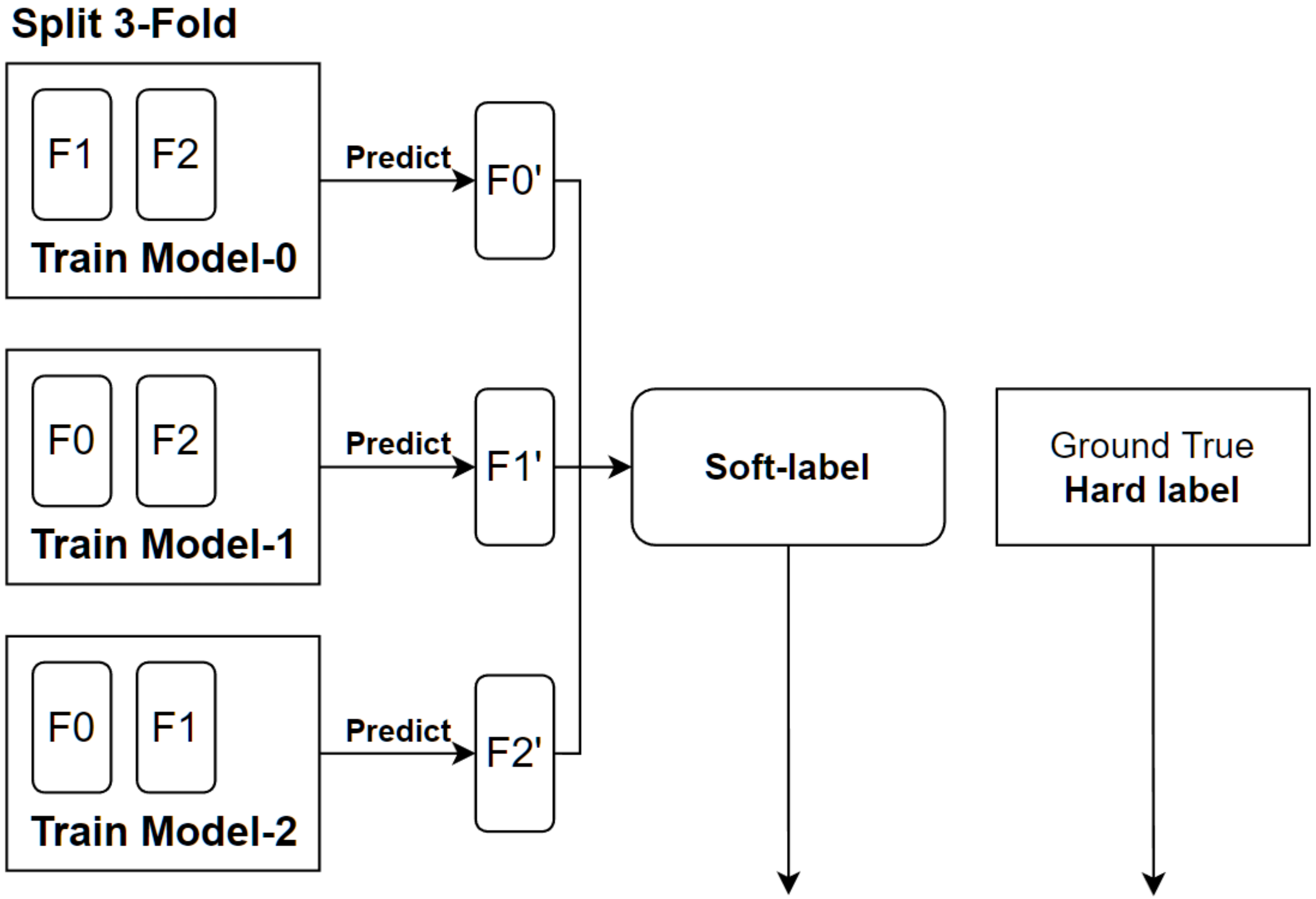


- Step 1: Data merge from 3 tasks.
- Step 2: Data preprocess: cleaning, augmentation, entity marker.
- Step 3: Model fine tune from both multilingual LMs and monolingual LMs.
- Step 4: Different training strategies applied.
- Step 5: Ensemble results from different models and strategies.

Basic Models



Training Optimization



- **Self Distillation**

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

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.

Training Optimization

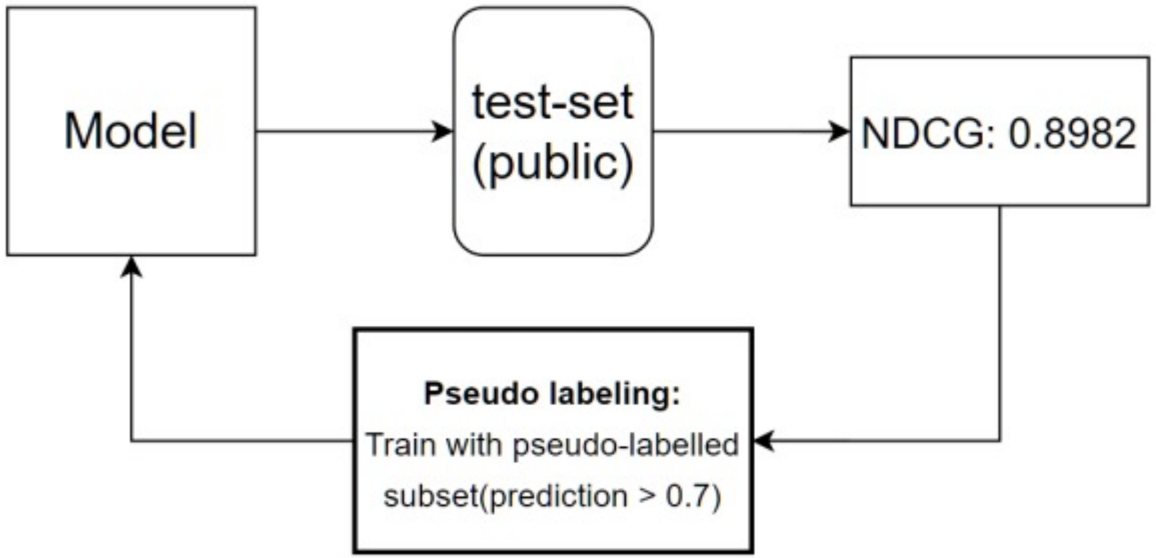


Figure 3: Train model with pseudo-labelled subset

- **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.

Training Optimization

- **Adversarial Training**

To gain robustness of models, we use Adversarial Weight Perturbation (AWP) in training steps that adversarially perturbs both model weights and the embeddings when the loss is below some threshold (like 0.6).

Besides, we also tried Fast Gradient Method (FGM) which performs slightly worse than AWP does in public leaderboard.

Methodology	NDCG (Public)
AWP	0.9022
FGM	0.9019

Training Optimization

- **Multi-sample dropout & Grouped layer-wise learning rate decay**

There are several effective regularization learning strategies to avoid overfitting of deep neural network, which can not only accelerate training and improve generalization ability, but also achieve lower error rates and losses.

Multi-sample dropout

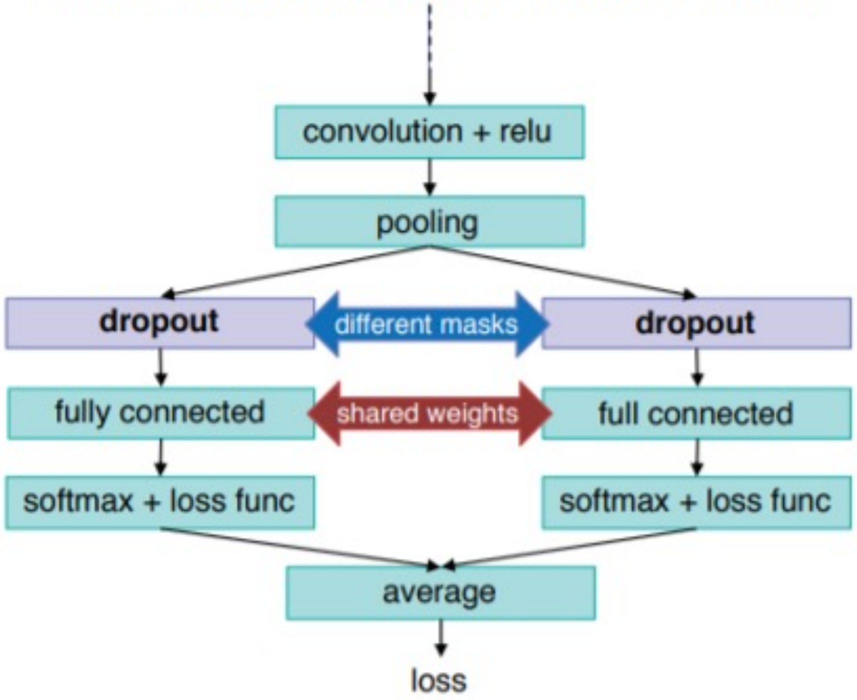


Table: Grouped layer-wise learning rate decay

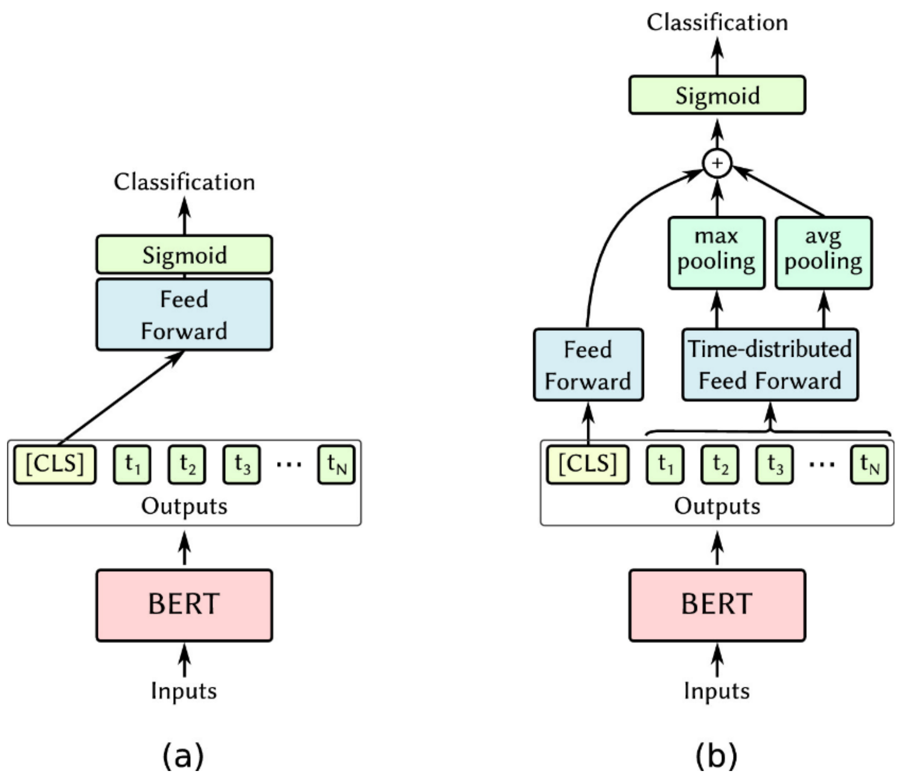
Model Layers	Learning Rate
0-5	5e-6
6-11	1e-5
12-17	1e-5
18-23	2e-5

Training Optimization

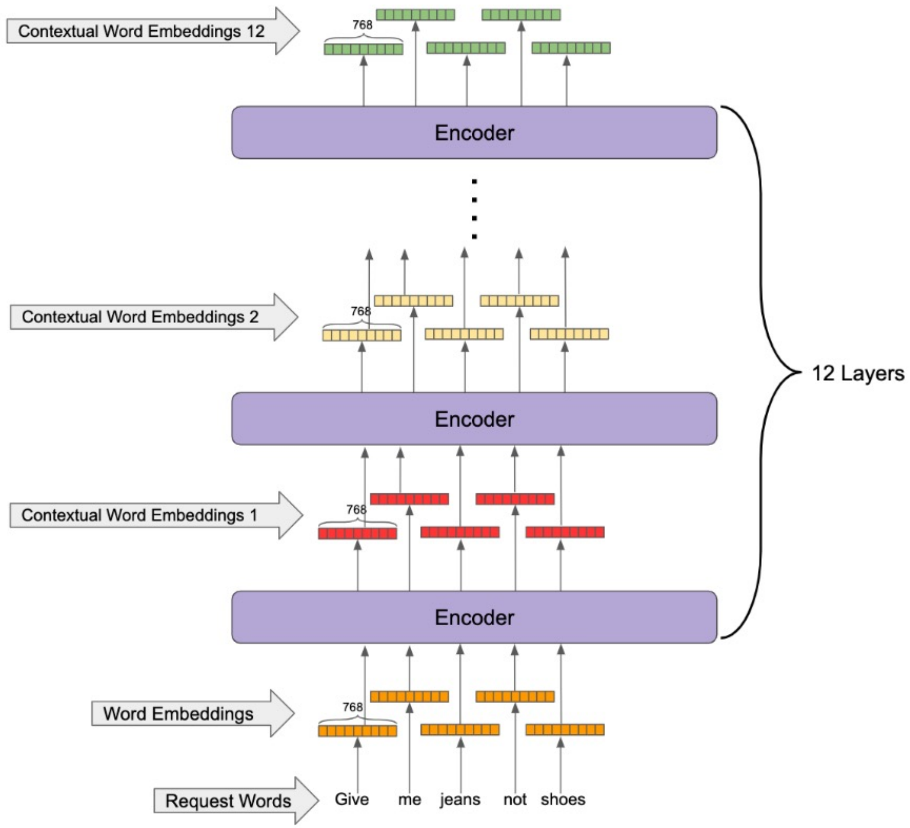
- **Weighted multi-layer Pooling**

Utilizing intermediate representations from various layers always provide better performance as it can help in incorporating more information.

Last Hidden State Output



Hidden State Output

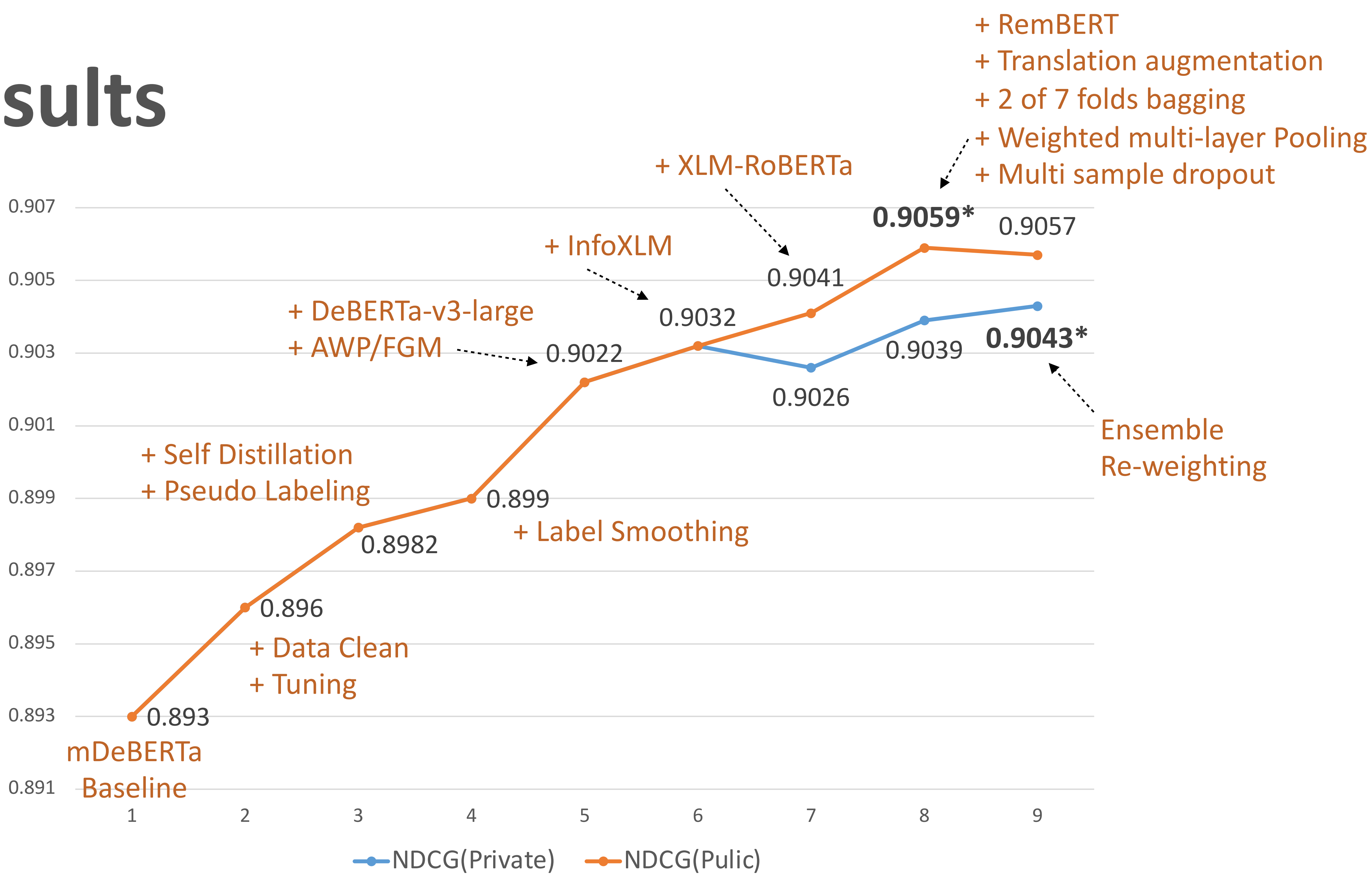


Ensemble

- Ensemble weights are mainly determined by the public scores and also the local cross-validation scores.
- Lower the weights of the models with high correlation coefficients.
- Our score is improved from **0.9022** to **0.9057** on the public leaderboard, and from **0.9015** to **0.9043** on the private leaderboard after ensemble.

NDCG (Public)	NDCG (Private)
0.9057	**0.9043**

Results



+ RemBERT
+ Translation augmentation
+ 2 of 7 folds bagging
+ Weighted multi-layer Pooling
+ Multi sample dropout
+ XLM-RoBERTa
0.9059*
0.9057
+ InfoXLM
0.9041
0.9032
+ DeBERTa-v3-large
+ AWP/FGM
0.9022
0.9039
0.9043*
0.9026
Ensemble
Re-weighting
+ Self Distillation
+ Pseudo Labeling
0.899
0.8982
+ Label Smoothing
0.896
+ Data Clean
+ Tuning
0.893
mDeBERTa
Baseline
0.907
0.905
0.903
0.901
0.899
0.897
0.895
0.893
0.891
1
2
3
4
5
6
7
8
9
NDCG(Private)
NDCG(Pulic)

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.

Thank You!