Multi-Level Contrastive Learning for Cross-Lingual Alignment

Beiduo Chen [1] Wu Guo [1] Bin Gu [1] Quan Liu [2] Yongchao Wang [2]

[1]NERC-SLIP, University of Science and Technology of China [2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research

Introduction

- **Object** cross-lingual pre-trained language models
- **Task** zero-shot cross-lingual transfer
- **Issue**
  - few explicit training goals for cross-lingual alignment
  - few fine-grained contrastive learning methods
  - require high computational resources

This paper proposes a multi-level contrastive learning (ML-CTL) method to integrate both **sentence-level** and **word-level** cross-lingual alignment into one training framework. Besides, under the limited computational resources settings, a cross-zero NCE (CZ-NCE) loss is designed by modifying **the lower bound** of the infoNCE loss to alleviate the impact of the floating-point error.

Concatenated Contextual Embedding Encoder

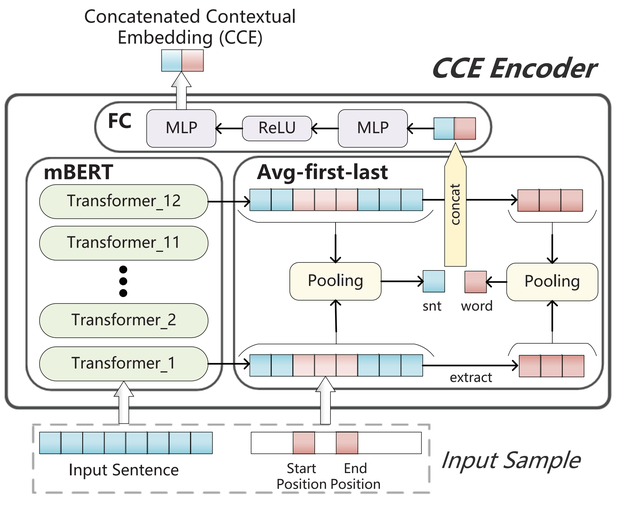

Figure 1. The internal structure of the CCE encoder. We input a sentence with a word's precise position information into the encoder to obtain a CCE.

Word positioned samples (WPSs) are constructed in advance to mark the position of each word (except stop words) in sentences. Then the constructed samples are fed into the CCE encoder consisting of three main modules:

1. **mBERT** a widely used multilingual pre-trained language model
2. **Avg-first-last** an effective way to extract sentence-level and word-level representations
3. **FC** the concatenation process with two MLP layers afterward

Multi-Level Contrastive Learning

For each batch of parallel WPSs $(\mathbf{X}, \mathbf{Y})$ in two languages, we input them separately to the encoder to obtain CCEs $\mathbf{X}_c = \{\mathbf{x}_{c1}, \mathbf{x}_{c2}, ..., \mathbf{x}_{cn}\}, \mathbf{Y}_c = \{\mathbf{y}_{c1}, \mathbf{y}_{c2}, ..., \mathbf{y}_{cn}\}$ where n is the batch size. Each $\mathbf{y}_{ci}$ is treated as a positive sample $\mathbf{k}^+$ for $\mathbf{x}_{ci}$ while a batch of all others $\{{\mathbf{X}_c}^{/\mathbf{x}_{ci}} \cup {\mathbf{Y}_c}^{/\mathbf{y}_{ci}}\}$ are considered as negative samples $\{\mathbf{k}^-\}$ (${\mathbf{X}_c}^{/\mathbf{x}_{ci}}$ denotes the remaining instances of $\mathbf{X}_c$ without $\mathbf{x}_{ci}$). Utilizing the infoNCE loss, the optimization target for each $\mathbf{x}_{ci}$ is achieved:

$$L_{info}(\mathbf{x}_{ci}) = -log(\frac{e^{s(\mathbf{x}_{ci},\mathbf{y}_{ci})}}{e^{s(\mathbf{x}_{ci},\mathbf{y}_{ci})} + \sum_{\mathbf{k}_j^-}^{\{\mathbf{k}^-\}} e^{s(\mathbf{x}_{ci},\mathbf{k}_j^-)}}) \quad (1)$$

$$L_{info\_batch} = \frac{\sum_i^n (L_{info}(\mathbf{x}_{ci}) + L_{info}(\mathbf{y}_{ci}))}{2n} \quad (2)$$

$$L_{multi_1} = L_{info\_batch} + \alpha * L_{MLM} \quad (3)$$

where α is the proportion of MLM. L_{info_batch} will be replaced by L_{CZ-NCE_batch} for situations with limited computational resources.

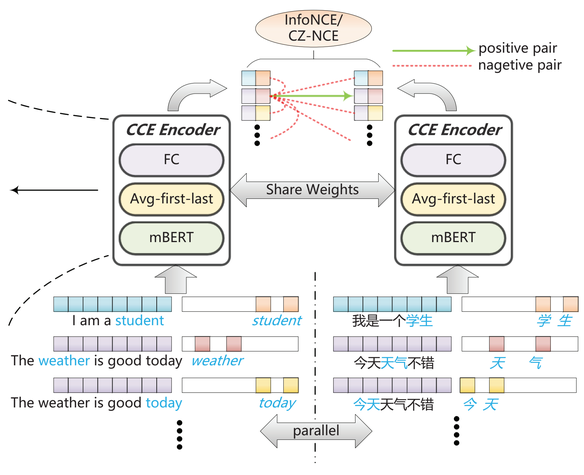

Figure 2. The concept map of ML-CTL which illustrates how to pre-train the model on both sentence-level and word-level with contrastive learning.

Cross-Zero NCE

$$L_{infoNCE} = -log(\frac{e^{s^+}}{e^{s^+} + \sum e^{s_i^-}}) \Rightarrow L_{CZ-NCE} = -log(\frac{e^{s^+}}{\sum e^{s_i^-}}) \quad (4)$$

$$L_{CZ-NCE} = -log(\frac{e^{s^+}}{\sum e^{s_i^-}}) = log(\sum e^{s_i^- - s^+}) = log(\varphi) \quad (5)$$

$$\nabla_\theta L_{CZ-NCE} = \nabla_\theta log(\varphi) = \frac{\nabla_\theta \varphi}{\varphi} = \nabla_\theta(\frac{\varphi}{sg(\varphi)}) \quad (6)$$

$$L_{multi2} = L_{CZ-NCE\_batch} + \alpha * L_{MLM} \quad (7)$$

Results on Xtreme Benchmark

Task	XNLI	PAWS-X	POS	NER	BUCC	TATOEBA
Model\Metric	Acc. (%)	Acc. (%)	F1 (%)	F1 (%)	F1 (%)	Acc. (%)
Main results compared to strong baselines						
mBERT (base)	65.4	81.9	70.3	62.2	56.7	38.7
XLM	**69.1**	80.9	70.1	61.2	56.8	32.6
MMTE	67.4	81.3	72.3	58.3	59.8	37.9
ML-CTL-CZ	67.8	**85.3**	**72.3**	**62.9**	**78.4**	**43.4**
Results of ablation study						
mBERT (base)	65.4	81.9	70.3	62.2	56.7	38.7
info-snt	66.255	84.092	71.544	62.157	76.426	41.148
CZ-snt	66.862	84.485	71.733	62.337	77.403	41.751
ML-CTL-CZ	**67.750**	**85.321**	**72.289**	**62.865**	**78.440**	**43.389**

Table 1. All results of evaluation experiments on pre-trained models.

T-SNE Visualization for Cross-Lingual Ability

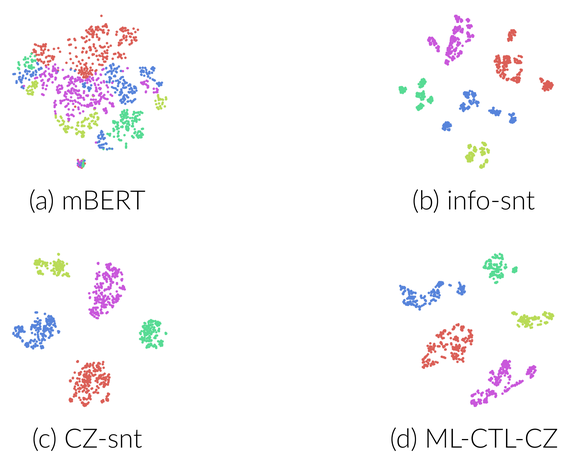

Figure 3. Graphs of t-SNE visualization. Each point represents a token embedding and the tokens in sentences across languages with the same meaning have the same color. ML-CTL-CZ has the optimal cross-lingual ability as the distribution of its tokens has better intra-class compactness and inter-class separability.

Conclusion

1. ML-CTL is proposed to improve the cross-lingual ability of LMs by utilizing both sentence-level and word-level information.
2. CZ-NCE is proposed to improve the performance of contrastive learning under limited computational resources settings.

Mail to: beiduo@mail.ustc.edu.cn

2022 IEEE International Conference on Acoustics, Speech and Signal Processing

Code at: https://github.com/Mckysse/ML-CTL