Multi-Level Contrastive Learning for Cross-Lingual Alignment

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

[1]National Engineering Research Center for Speech and Language Information Processing, University of Science and Technology of China

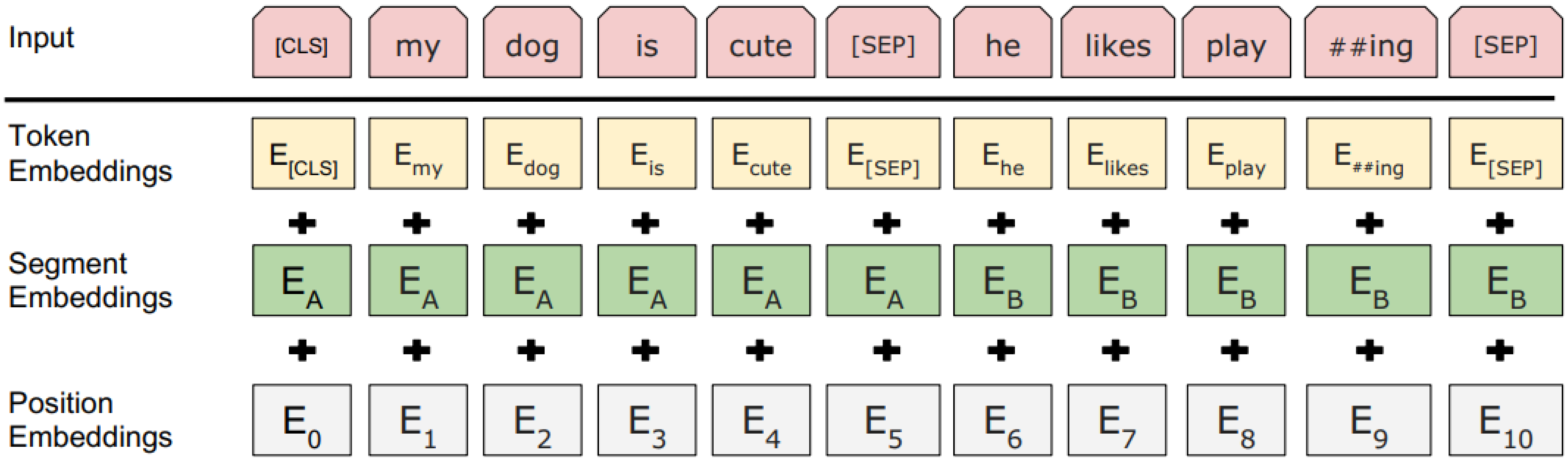
[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research

Outline

- **Introduction**
- Multi-level contrastive learning (ML-CTL)
- Cross-zero noise contrastive estimation loss (CZ-NCE)
- Experiments and analyses
- Conclusion

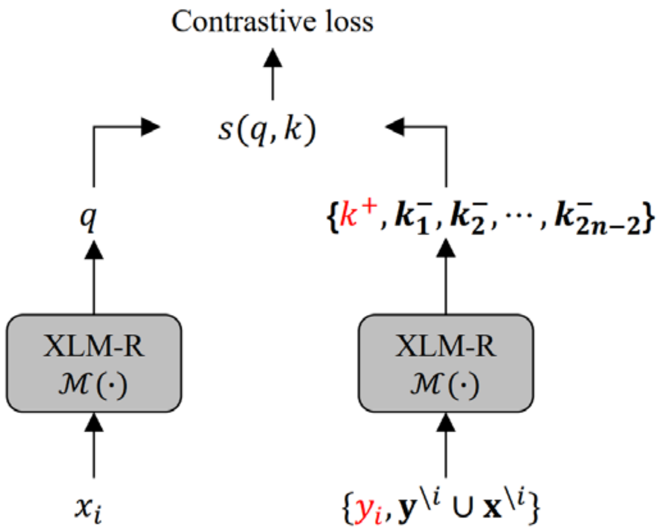
Cross-lingual Language Model

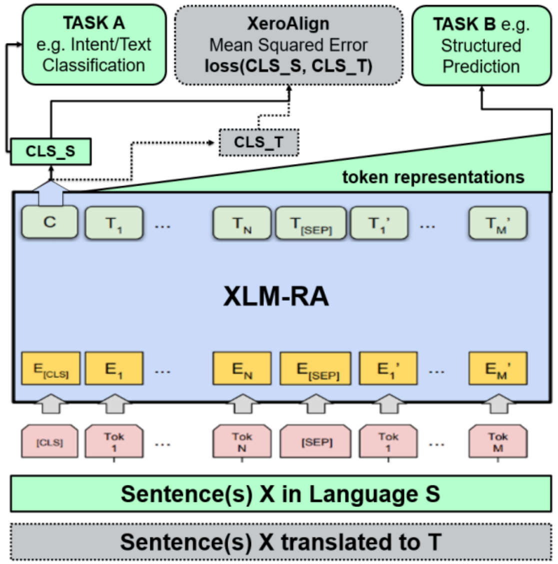
- Transformer based: mBERT, XLM
- MLM: mask language modeling

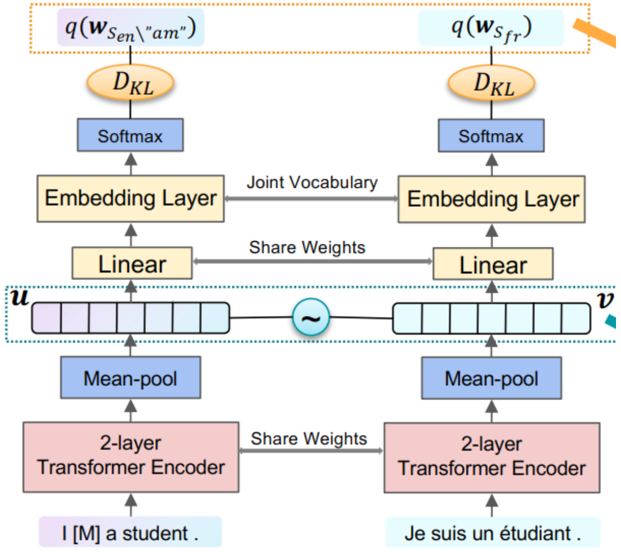


Contrastive Learning

- Issue: no explicit cross-lingual alignment
- Exist solution: contrastive learning (CTL)







Contrastive Learning

- Issue: no explicit cross-lingual alignment
- Exist solution: contrastive learning (CTL)
  - New issue: only sentence-level CTL

sentence-level → sentence-level + word-level

Multi-Level Contrastive Learning

Computational Resources

- Issue: high demand for computational resources

- InfoNCE:

$$L_{infoNCE} = -log(\frac{e^{s^+}}{e^{s^+} + \sum e^{s_i^-}})$$

Especially serious in contrastive learning

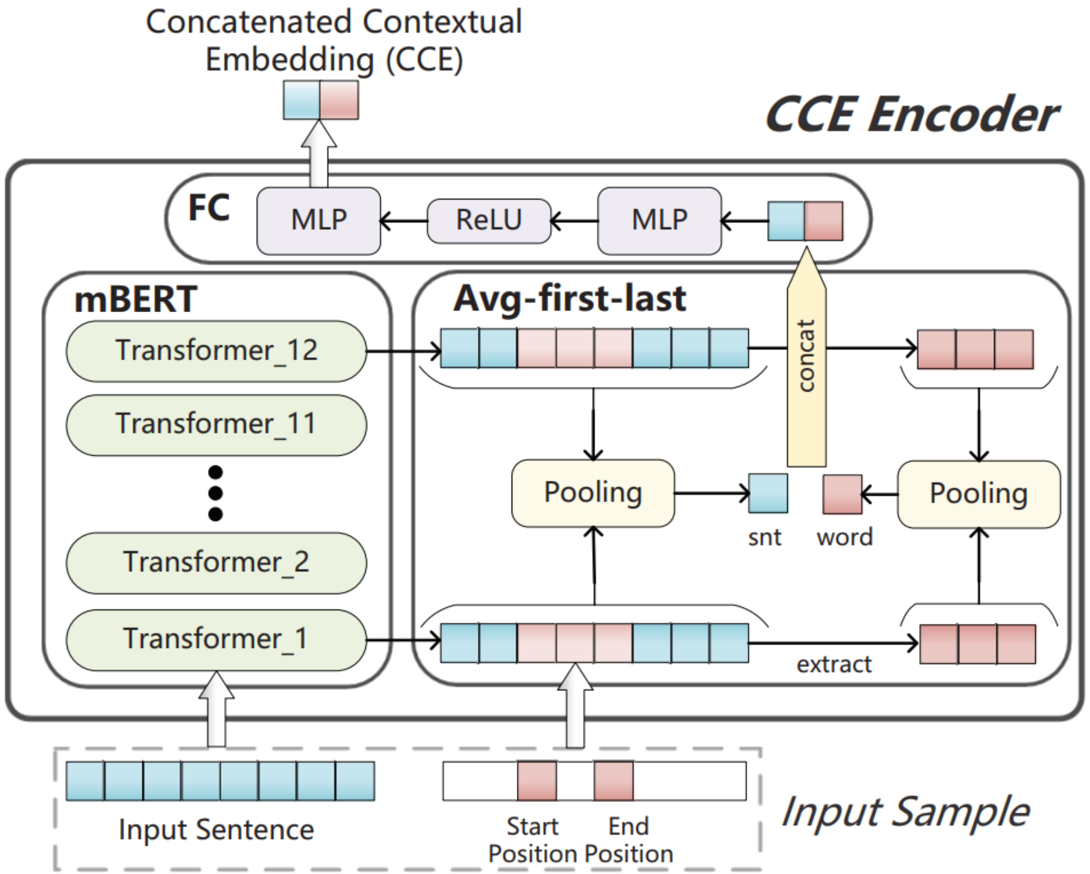
→ Cross-zero NCE loss

Outline

- Introduction
- **Multi-level contrastive learning (ML-CTL)**
- Cross-zero noise contrastive estimation loss (CZ-NCE)
- Experiments and analyses
- Conclusion

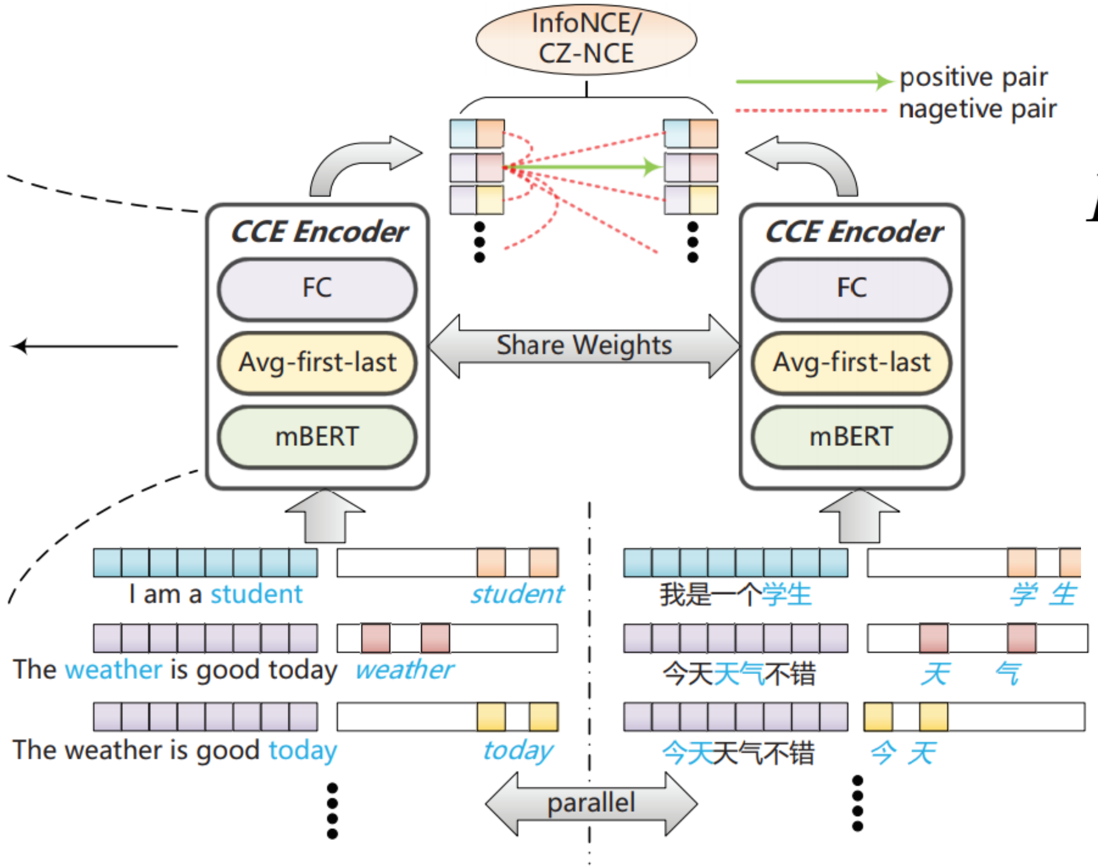
ML-CTL

- Concatenated Contextual Embedding Encoder



ML-CTL

- Multi-level contrastive learning framework



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

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

$$L_{multi_1} = L_{info\_batch} + \alpha * L_{MLM}$$

Outline

- Introduction
- Multi-level contrastive learning (ML-CTL)
- **Cross-zero noise contrastive estimation loss (CZ-NCE)**
- Experiments and analyses
- Conclusion

CZ-NCE

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

CZ-NCE

- Demonstration

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

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

$$L_{multi2} = L_{CZ-NCE\_batch} + \alpha * L_{MLM}$$

Outline

- Introduction
- Multi-level contrastive learning (ML-CTL)
- Cross-zero noise contrastive estimation loss (CZ-NCE)
- **Experiments and analyses**
- Conclusion

Experiments and analyses

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**


(a) mBERT


(b) info-snt


(c) CZ-snt


(d) ML-CTL-CZ

Outline

- Introduction
- Multi-level contrastive learning (ML-CTL)
- Cross-zero noise contrastive estimation loss (CZ-NCE)
- Experiments and analyses
- **Conclusion**

Conclusion

- ML-CTL is proposed to improve the cross-lingual alignment ability of pre-trained language models by applying contrastive learning on concatenated contextual embeddings which contain information of both sentences and words.
- CZ-NCE is proposed to alleviate the impact of the floating-point error with a small training batch size.