MULTI-LEVEL CONTRASTIVE LEARNING FOR CROSS-LINGUAL ALIGNMENT

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

[1]NELSLIP, University of Science and Technology of China, Hefei, China
[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research

ABSTRACT

Cross-language pre-trained models such as multilingual BERT (mBERT) have achieved significant performance in various cross-lingual downstream NLP tasks. This paper proposes a multi-level contrastive learning (ML-CTL) framework to further improve the cross-lingual ability of pre-trained models. The proposed method uses translated parallel data to encourage the model to generate similar semantic embeddings for different languages. However, unlike the sentence-level alignment used in most previous studies, in this paper, we explicitly integrate the word-level information of each pair of parallel sentences into contrastive learning. Moreover, cross-zero noise contrastive estimation (CZ-NCE) loss is proposed to alleviate the impact of the floating-point error in the training process with a small batch size. The proposed method significantly improves the cross-lingual transfer ability of our basic model (mBERT) and outperforms on multiple zero-shot cross-lingual downstream tasks compared to the same-size models in the Xtreme benchmark.

Index Terms— Cross-language pre-trained model, contrastive learning, multi-level, cross-zero NCE, cross-lingual alignment

1. INTRODUCTION

Recently, cross-lingual pre-trained language models with the structure of transformers like multilingual BERT (mBERT) and cross-lingual language model (XLM) have enabled effective cross-lingual transfer and performed surprisingly well on plenty of downstream tasks [1, 2]. These models are firstly pre-trained on large-scale corpus covers over 100 languages mainly using the multilingual masked language modeling (MMLM) [1] algorithm, and then fine-tuned on English-supervised training data for downstream tasks aiming to improve the performance on low-resource languages [3]. However, such a pre-training method only encourages implicit cross-lingual alignment in the vector space [4, 5]. Although XLM [2] introduced the translation language modeling (TLM) method to further enhance cross-lingual alignment, some scholars believe that TLM only learns the structural pattern of parallel sentence pairs without truly understanding the meaning of sentences [6, 7].

Thanks to the National Natural Science Foundation of China (Grant No. U1836219) for funding.

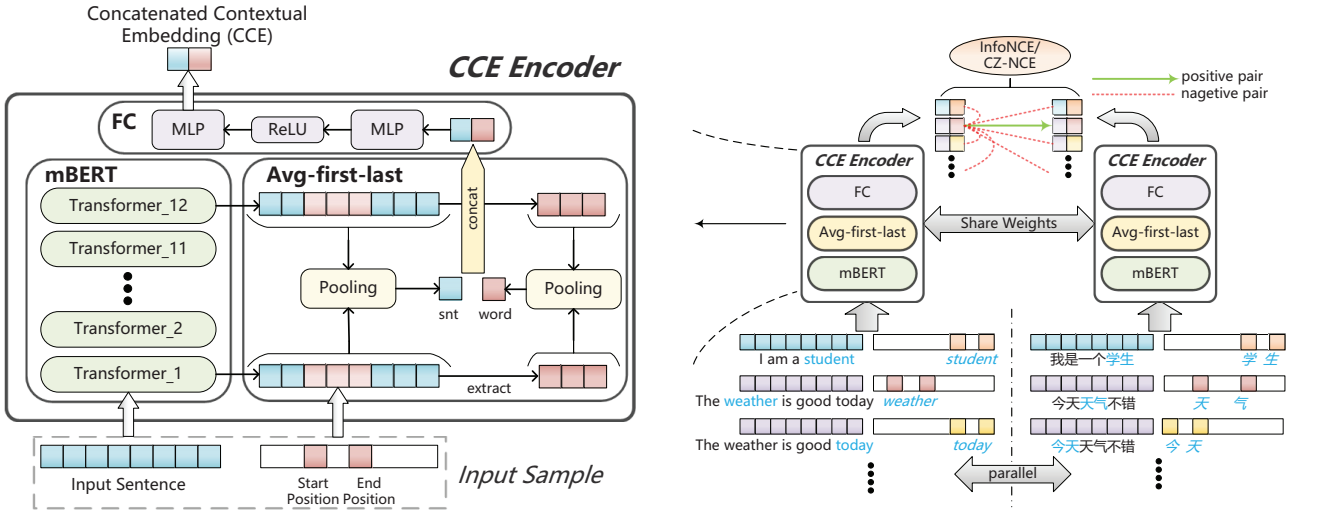
A feasible approach to improve the model's ability on cross-lingual transfer is to incorporate translated parallel sentences into pre-training for explicit alignment. Many studies have used the contrastive learning (CTL) [8, 9] to solve this problem as translated parallel sentences can naturally be used as positive pairs in CTL for encouraging the model to learn explicit cross-lingual alignment [10, 11]. However, most of them directly use the [CLS] token in BERT [1] to represent the meanings of sentences for cross-language alignment only at the sentence-level, without considering alignment at the word-level. Wei *et al.* use the idea of bag-of-words to address this problem [6], however, where the advantage of BERT's bidirectional structure to generate dynamic contextual word embeddings is overshadowed by the fixed word embeddings.

To overcome these problems, we propose a multi-level contrastive learning (ML-CTL) method to integrate both sentence-level and word-level cross-lingual alignment into one training framework. In concrete terms, we first construct samples that consist both of sentences and positions of words, and then fed them into a designed encoder that contains a basic language model to get concatenated contextual embeddings (CCE) representing both sentence information and contextual word information. After pre-trained with CCEs by contrastive learning, the basic model's cross-lingual ability can be significantly improved by learning the structural pattern of parallel sentence pairs as well as aligning the semantic meanings of parallel words in the sentences.

In practice, CTL with commonly used information noise contrastive estimation (infoNCE) [12] loss requires a large batch size to improve performance [13] which is too expensive for cases with limited computational resources. However, the infoNCE loss value of a mini-batch will soon approach zero during training with a small batch size and the floating-point error will seriously deteriorate the final performance [14]. Hence, we propose cross-zero NCE (CZ-NCE) loss by modifying the lower bound of the infoNCE loss to alleviate the impact of the floating-point error.

We carry out evaluation experiments for the proposed pre-trained models on multiple zero-shot cross-lingual tasks in the Xtreme benchmark [15]. The results demonstrate that the proposed methods achieve a significant performance improvement on a strong baseline.

In summary, we present two novel contributions:



(a) CCE encoder (b) ML-CTL

Fig. 1. (a) shows the internal structure of the CCE encoder. We input a sentence with a word's precise position information into the encoder to obtain a concatenated embedding. (b) is the concept map of ML-CTL which illustrates how to pre-train the model on both sentence-level and word-level with contrastive learning.

(1) 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.
(2) CZ-NCE is proposed to alleviate the impact of the floating-point error with a small training batch size.

2. MULTI-LEVEL CONTRASTIVE LEARNING

To integrate the word-level alignment information into the sentence-level cross-lingual pre-training, the word positioned sample (WPS) is first constructed and then the concatenated contextual embedding (CCE) of WPS is extracted for the following multi-level contrastive learning (ML-CTL).

2.1. Word Positioned Sample Construction

In conventional CTL training, one sentence is considered as a training sample [10, 11]. For the proposed ML-CTL, the start and end positions of each word (without stop words) are appended to the sentence to form a new word positioned sample (WPS). Using a bilingual dictionary that has been removed stop words, a parallel WPSs pair is obtained with accurate word alignment, as Algorithm 1 shows.

2.2. Concatenated Contextual Embedding

As shown in Fig. 1a, mBERT as part of the encoder is applied to extract concatenated contextual embedding (CCE). For the sentence-level representation, we choose the average of all tokens from the first and last transformer layers of mBERT (*avg-first-last*), as it has been proved to be a good way to represent the meaning of a sentence [17]. Because high-frequency words may dominate the meaning of the whole sentence [18], it is not a good idea to rely only on sentence embeddings for cross-lingual alignment.

Algorithm 1 Parallel WPSs construction

use a parallel sentences pair for example
Fetch parallel sentences pair $\{\mathbf{A} \leftrightarrow \mathbf{B}\}$ with words string $\mathbf{A} = \{a_1, a_2, ..., a_N\}, \mathbf{B} = \{b_1, b_2, ..., b_M\}$
Bilingual Dictionary $\mathbf{V}_{\mathbf{A}\times\mathbf{B}}\{a_i \leftrightarrow b_i\}$
TK = BertTokenizer[16, 1] # a tokenizer for BERT

# get parallel words pair and their positions
for a_i in $\mathbf{V}_\mathbf{A}$ **do**
  if $\mathbf{A}$.count(' a_i ')==TK($\mathbf{A}$).count(TK('a_i'))==1 **then**
    # 'count' returns the occurrence number
    get $\mathbf{S} = (\mathbf{A}, start_{a_i}, end_{a_i})$
    # record the word's position in the sentence
    get $b_i = \mathbf{V}_{\mathbf{A}\times\mathbf{B}}[a_i]$
    if $\mathbf{B}$.count(' b_i ')==1 **then**
      if TK($\mathbf{B}$).count(TK('b_i'))==1 **then**
        get $\mathbf{S^t} = (\mathbf{B}, start_{b_i}, end_{b_i})$
        return $\mathbf{S}, \mathbf{S^t}$
        # return a pair of positive parallel WPSs

Using WPS, we can extract the word embedding that has contextual semantics due to the BERT's bidirectional structure. The extracted word embedding is more suitable to be the representation for alignment than the conventional fixed word embedding. To be consistent with the sentence embedding, we also adapt *avg-first-last* pattern.

Finally, these two embeddings are spliced and fed into the following FC layer to obtain the final CCE.

2.3. Multi-Level Contrastive Learning

As shown in Fig. 1b, 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 infoNCE loss [12], the optimization target for each $\mathbf{x}_{ci}$ is achieved:

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

where $s(\cdot)$ denotes the cosine similarity calculation with a denominator (temperature parameter t set as 0.07).

Then the total loss for a batch of samples is shown as:

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

InfoNCE loss brings positive samples closer together and negative samples farther away, which fits well with cross-lingual alignment. It is worth noting that the WPSs with parallel sentences but different words are also treated as negative pairs, as depicted in Fig. 1b. We hypothesize that this design is more effective in word-level alignment as the distance between negative word pairs can be further extended. In order not to damage the language information within mBERT itself, we also add MLM [1] as an auxiliary task in the real training. The total loss is calculated as follow:

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

where α is the proportion of MLM.

3. CROSS-ZERO NCE

As an accepted conclusion, contrastive learning requires a large batch size to improve the learning effect [13]. However, in most cases with limited computational resources, we can only pre-train the model with a small batch size. The model can quickly distinguish positive and negative samples correctly during pre-training with a small batch size because there are too few negative ones as interferences and CCE can further increase the discriminability. Assuming s^+ and s_i^- denote distances between positive and negative samples respectively in infoNCE loss function, while e^{s^+} becomes much larger than e^{s^-} during training with a small batch size, the loss will soon approach 0, which is at the same magnitude with the floating-point error. It may seriously affect the training of contrastive learning [14].

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

In Eq. 4, the e^{s^+} in the denominator sets a lower bound 0 for infoNCE loss. Therefore, we consider removing the e^{s^+} in the denominator and modifying infoNCE loss to cross-zero NCE (CZ-NCE) loss for keeping the loss value away from zero during most of the training time.

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

As Eq. 5 shows, the learning goal of CZ-NCE is still the same as infoNCE. In the following, we prove the effectiveness of CZ-NCE on alleviating the disturbance of the float-point error. Assuming $\varphi = \sum e^{s_i^- - s^+}$, we perform the following calculation:

$$L_{CZ-NCE} = -log(\frac{e^{s^+}}{\sum e^{s_i^-}}) = log(\sum e^{s_i^- - s^+}) = log(\varphi) \quad (6)$$

$$\nabla_\theta L_{CZ-NCE} = \nabla_\theta log(\varphi) = \frac{\nabla_\theta \varphi}{\varphi} = \nabla_\theta(\frac{\varphi}{sg(\varphi)}) \quad (7)$$

where ∇_θ denotes the gradient calculation and $sg(\cdot)$ stands for the stop-gradient operator that is defined as identity at forward computation time and has zero partial derivatives, thus effectively constraining its operand to a non-updated constant.

Assuming a new loss $\rho = \frac{\varphi}{sg(\varphi)}$, it is easy to notice that $\rho \equiv 1$ for any batch of inputs (note the gradient of ρ is not flat) which is far from 0 and less affected by the floating-point error. Since the back-propagation of neural network training only matters with the gradient of loss instead of the value, CZ-NCE has just the same effect with ρ for model's training, indicating that CZ-NCE can indeed alleviate the impact of the float-point error.

Same as before, we set the total learning loss as follow:

$$L_{multi2} = L_{CZ-NCE\_batch} + \alpha * L_{MLM} \quad (8)$$

4. EXPERIMENTS AND RESULTS

4.1. Datasets and Experiments Setup

The United Nations (UN) Parallel Corpus v1.0 is used as a training set, consisting of manually translated UN documents from 1990 to 2014 for the six official UN languages: Arabic, Chinese, English, French, Russian, and Spanish [20]. This dataset has been used for lots of pre-training of LM [7, 2, 6]. We create several bilingual dictionaries from Google translate and then randomly construct 250M parallel WPSs in total evenly distributed over six languages. We apply the proposed methods to pre-training the basic model mBERT and generally set the learning rate as 2e-6 while the batch size is 64. We use Adam optimizer [21] and set α as 0.1.

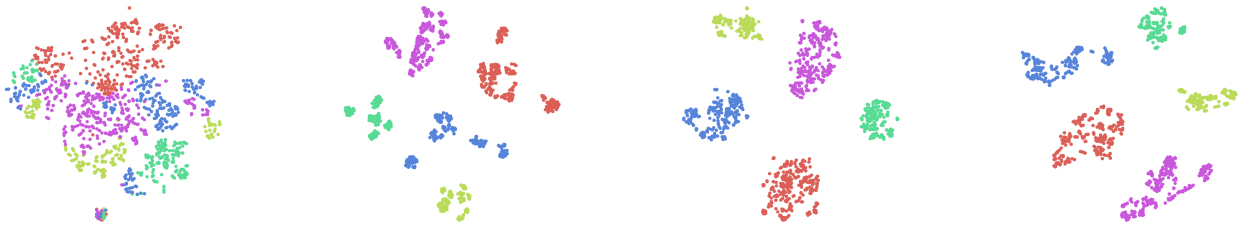
To analyze the cross-lingual performance of the proposed pre-trained models, we choose several zero-shot downstream tasks in Xtreme benchmark [15]: XNLI and PAWS-X for sentence classification with metric as accuracy (Acc.), UD-POS and PANX(NER) for structured prediction with metric as F1 Score(F1), as well as BUCC and Tatoeba for retrieval with metrics as F1 and Acc. All these tasks are widely used in the cross-lingual assessment. We first fine-tune the part of modified mBERT in our pre-trained model on English labeled data and then apply it to predicting on non-English unlabeled data. All the configurations of fine-tuning are set the same as Xtreme benchmark for fair reference.

4.2. Main Results

For the proposed ML-CTL system with CZ-NCE loss (denoted as ML-CTL-CZ), the CCE encoder is first initialized with mBERT's parameters and then pre-trained by our methods. We also employ two strong baselines XLM [2] and MMTE [19] with the same size as mBERT for comparison. As shown in the upper half of Table 1, the proposed ML-CTL-CZ significantly improves the performance of the basic

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

Task	XNLI	PAWS-X	POS	NER	BUCC	TATOEBA
Model\Metrics	Acc. (%)	Acc. (%)	F1 (%)	F1 (%)	F1 (%)	Acc. (%)
Main results compared to strong baselines						
mBERT [1] (base)	65.4	81.9	70.3	62.2	56.7	38.7
XLM [2]	**69.1**	80.9	70.1	61.2	56.8	32.6
MMTE [19]	67.4	81.3	72.3	58.3	59.8	37.9
ML-CTL-CZ (ours)	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

Fig. 2. 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. From the graphs, ML-CTL-CZ has the optimal cross-lingual ability as the distribution of its tokens has better intra-class compactness and inter-class separability.

model (mBERT) and achieves optimal results on multiple downstream tasks compared to the same-size models. The reason for the impressive improvement on BUCC might be that the goal of this task is to look for parallel pairs in a pile of sentences, exactly in line with our pre-training goal.

4.3. Ablation Study

To investigate the effect of the proposed methods, we perform the ablation study as shown in the bottom half of Table 1. All the enhanced systems are initialized with mBERT's parameters. *info-snt* denotes the pre-trained model that applies only to sentence embeddings (no word embeddings concatenated) with infoNCE loss. *CZ-snt* replaces the former system's loss with CZ-NCE loss. For the ML-CTL-CZ system, both the word-level and sentence-level information are used with CZ-NCE loss. From the results, CZ- NCE loss can provide improvements over the infoNCE loss in a small training batch size. CCE with multi-level information can increase the discriminability between positive and negative pairs and further improve the cross-language performance of the model.

4.4. T-SNE Visualization for Cross-lingual Ability

We use t-SNE visualization [22] to intuitively demonstrate the cross-lingual ability of different models in the ablation study. Sentence groups with five different meanings in 6 languages (30 sentences totally) are randomly selected from the UN corpus and fed into different models. Fig. 2 shows the results. Compared with large overlaps among different classes for the basic model mBERT(Fig. 2a), we observe a clear disentanglement in the feature space with distinctive boundaries between each class for CTL systems. As illustrated, the tokens in Fig. 2c&d have better intra-class compactness and inter-class separability than those in Fig. 2b. The tokens in the same class shown in Fig. 2d are further closer which indicates that the ML-CTL-CZ has achieved better word-level alignment. These graphs demonstrate the power of the proposed methods.

5. CONCLUSION

In this paper, we have proposed a framework of multi-level contrastive learning to further help cross-lingual models learn universal representations across languages. As demonstrated by the experiments, word-level information is a strong complement to sentence-level information in cross-lingual alignment. Furthermore, CZ-NCE loss is proposed to reduce the floating-point error near the zero point in the case of training with a small batch size. The proposed model achieves better performance on multiple zero-shot cross-language tasks than the same-size models in Xtreme benchmark. Through the ablation study and t-SNE visualization, we have clearly demonstrated the effectiveness of the proposed methods.

6. REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] Guillaume Lample and Alexis Conneau, "Cross-lingual language model pretraining," *arXiv preprint arXiv:1901.07291*, 2019.

[3] Sebastian Ruder, Ivan Vulić, and Anders Søgaard, "A survey of cross-lingual word embedding models," *Journal of Artificial Intelligence Research*, vol. 65, pp. 569–631, 2019.

[4] Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah, "First align, then predict: Understanding the cross-lingual ability of multilingual bert," *arXiv preprint arXiv:2101.11109*, 2021.

[5] Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim, "What's so special about bert's layers? a closer look at the nlp pipeline in monolingual and multilingual models," *arXiv preprint arXiv:2004.06499*, 2020.

[6] Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo, "On learning universal representations across languages," *arXiv preprint arXiv:2007.15960*, 2020.

[7] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[8] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, 2020.

[9] Andriy Mnih and Yee Whye Teh, "A fast and simple algorithm for training neural probabilistic language models," *arXiv preprint arXiv:1206.6426*, 2012.

[10] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou, "Infoxlm: An information-theoretic framework for cross-lingual language model pre-training," *arXiv preprint arXiv:2007.07834*, 2020.

[11] Liang Wang, Wei Zhao, and Jingming Liu, "Aligning cross-lingual sentence representations with dual momentum contrast," *arXiv preprint arXiv:2109.00253*, 2021.

[12] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[14] Junya Chen, Zhe Gan, Xuan Li, Qing Guo, Liqun Chen, Shuyang Gao, Tagyoung Chung, Yi Xu, Belinda Zeng, Wenlian Lu, et al., "Simpler, faster, stronger: Breaking the log-k curse on contrastive learners with flatnce," *arXiv preprint arXiv:2107.01152*, 2021.

[15] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson, "Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 4411–4421.

[16] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

[17] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li, "On the sentence embeddings from pre-trained language models," *arXiv preprint arXiv:2011.05864*, 2020.

[18] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu, "Consert: A contrastive framework for self-supervised sentence representation transfer," *arXiv preprint arXiv:2105.11741*, 2021.

[19] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al., "Massively multilingual neural machine translation in the wild: Findings and challenges," *arXiv preprint arXiv:1907.05019*, 2019.

[20] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen, "The united nations parallel corpus v1. 0," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 3530–3534.

[21] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.