Multi-Granularity Contrasting for Cross-Lingual Pre-Training

Shicheng Li[1], Pengcheng Yang[2], Fuli Luo[2], Jun Xie[2]
[1]MOE Key Lab of Computational Linguistics, School of EECS, Peking University
[2]Alibaba DAMO Academy, Hangzhou, China
lisc99@pku.edu.cn
{mingyang.ypc, lfl259702, qingjing.xj}@alibaba-inc.com

Abstract

Cross-lingual pre-training aims at providing effective prior representations for the inputs from multiple languages. With the modeling of bidirectional contexts, recently prevalent language modeling approaches such as XLM achieve better performance than traditional methods based on embedding alignment, which strives to assign similar vector representations to semantic-equivalent units. However, such approaches like XLM capture cross-lingual information based solely on shared BPE vocabulary, resulting in the absence of fine-grained supervision induced by embedding alignment. Inheriting the advantages of the above two paradigms, this work presents a multi-granularity contrasting framework, namely MGC, to learn language-universal representations. While predicting the masked words based on bidirectional contexts, the proposal also encodes semantic equivalents from different languages into similar representations to introduce more fine-grained and explicit cross-lingual information. Two effective contrasting strategies are further proposed, which can be built upon semantic units of multiple granularities covering words, span, and sentences. Extensive experiments demonstrate that our approach can achieve significant performance gains in various downstream tasks, including machine translation and cross-lingual language understanding.

1 Introduction

Cross-lingual pre-training (Lample and Conneau, 2019) has achieved striking success in the field of natural language processing. By providing effective prior representations for the inputs from different languages, it has boosted performance on various downstream tasks such as machine translation and cross-lingual language understanding.

Early efforts regarding cross-lingual pre-training mainly focus on embedding alignment (Mikolov et al., 2013b; Lample et al., 2018), which is targeted at the assignment of similar vector representations to semantic-equivalent units (e.g., the parallel bilingual word or sentence pairs). For instance, Mikolov et al. (2013b) attempt to project pre-trained monolingual word embeddings from two languages into a common semantic space with a simple linear transformation, so that parallel bilingual words share the same representation. This allows the introduction of explicit fine-grained supervision to guarantee the representational similarity of semantic equivalents, but neglects the modeling of bidirectional contexts. Going a step further, recently prevalent approaches of language modeling such as XLM (Lample and Conneau, 2019) remedy this by predicting the masked tokens based on bidirectional contexts (Devlin et al., 2019), and also benefit from larger model capacity (Vaswani et al., 2017). However, the cross-lingual information captured by these language modeling approaches comes solely from the shared BPE vocabulary (Sennrich et al., 2016), resulting in the absence of more fine-grained explicit supervision induced by embedding alignment.

In light of the pros and cons of the above two paradigms, we propose a multi-granularity contrasting (MGC) framework for cross-lingual pre-training. In addition to modeling context bidirectionality with the widely used masked language modeling (MLM) (Devlin et al., 2019), our approach draws upon contrastive learning (Gutmann and Hyvärinen, 2010) to introduce more fine-grained cross-lingual alignment information. The core idea is to enhance the consistency between representations of semantic equivalents (e.g., the aligned word pairs such as "*cat*" in English and "*chat*" in French). To this end, we propose two effective contrasting strategies: *hard* contrasting which constructs pseudo-parallel bilingual word pairs via external word aligner (Dyer et al., 2013),

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1708–1717

and *soft* contrasting which employs multi-head attention (Vaswani et al., 2017) to provide posterior approximation for the representations of the desired semantic equivalents. Considering the inherent multi-granularity of natural language expressions, we build the proposed contrasting framework upon semantic units of various granularities (including *word*, *span*, and *sentence*) to further enrich cross-lingual information and enhance the model's capability of encoding multi-granularity representations.

We conduct experiments on a variety of downstream scenarios, including multiple machine translation and cross-lingual language understanding tasks. Comprehensive experimental results demonstrate that our proposed approach can achieve significant performance gains over baselines. To be more specific, our MGC raises the average accuracy of our implemented XLM-R (Conneau et al., 2019) from 74.4 to 76.0 on XNLI under the setting of cross-lingual transfer and also surpasses various baselines on representative translation tasks such as WMT14 EN-DE and EN-FR.

2 Methodology

In order to introduce more fine-grained and explicit cross-lingual supervision, we propose a multi-granularity contrasting (MGC) framework to learn language-universal representations. We first elaborate on the proposed approach based on *word*-level contrasting, and then extend it to *span*-level and *sentence*-level to further enrich cross-lingual information.

2.1 Overview

We denote a pair of parallel bilingual instance as $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x} = (x_1, \cdots, x_m)$ and $\mathbf{y} = (y_1, \cdots, y_n)$ refer to the source and target sentence, respectively. Then, the transformer (Vaswani et al., 2017) encodes $\mathbf{x}$ to obtain its hidden representations $\mathbf{h_x} = (h_{x_1}, \cdots, h_{x_m})$. The hidden representations $\mathbf{h_y} = (h_{y_1}, \cdots, h_{y_n})$ of $\mathbf{y}$ can be obtained in the same way. In order to introduce more fine-grained and explicit cross-lingual supervision similar to embedding alignment, we expect the semantic-equivalent units (e.g., "*cat*" in English and "*chat*" in French) from different languages to exhibit similar vector representations. Meanwhile, the representations of units with different semantics (e.g., "*cat*" in English and "*car*" in English or "*voiture*" in French) should be distinguished from each other to capture their discriminative specific information.

Motivated by this, we employ contrastive learning (Gutmann and Hyvärinen, 2010) to model such training objectives. Without loss of generality, we elaborate on our proposed approach with the units in the source language as anchors. Formally, we use u to represent the representation of one unit (e.g., "*cat*" in English) in $\mathbf{x}$. The representation of its corresponding semantic equivalent (e.g., "*chat*" in French) in $\mathbf{y}$ is denoted as v^+. The set of negative representations exhibiting different semantic with u is denoted as $\boldsymbol{v}^- = \{v_1^-, \cdots, v_k^-\}$, where k is the number of negative representations. Then, the contrastive loss for the representation tuple $(u, v^+, \boldsymbol{v}^-)$ can be defined as:

$$\mathcal{L}(u, v^+, \boldsymbol{v}^-) = -\log \left(\frac{\exp(u \cdot v^+/\tau)}{Z(u)} \right) \quad (1)$$

where $Z(u) = \exp(u \cdot v^+/\tau) + \sum_{v^-} \exp(u \cdot v^-/\tau)$ is the normalization factor and τ is the temperature controlling the concentration level of the sample distribution. The above equation corresponds to the negative log-likelihood loss of a softmax-based classifier measuring semantic similarity by the dot product. The classifier treats each unit as a distinct class, and aims at classifying u to the class of its semantic equivalent v^+ and vice versa. By maximizing the consistency between the representations of semantic equivalents with such a training objective, the pre-trained models are encouraged to introduce more fine-grained explicit alignment supervision, thereby enhancing their capability of learning language-invariant representations. Meanwhile, the representations of units exhibiting different semantics are penalized to be kept distinguished from each other, so that the model is equipped with the ability to capture specific features of the source inputs.

2.2 Word-Level Contrasting

The word-level contrasting strives to integrate the word-alignment information contained in parallel bilingual instance $(\mathbf{x}, \mathbf{y})$. However, an intractable challenge is that ideal semantic-equivalent word pairs tend to be unavailable in practice. To remedy this, here we propose two effective solutions: *hard* contrasting and *soft* contrasting, detailed as follows.

Hard contrasting The hard contrasting aims at constructing pseudo-parallel bilingual word pairs

via an external word aligner. Specifically, for each word x in the source sentence $\mathbf{x}$, its aligned word in $\mathbf{y}$ is defined as:

$$y^* \approx \hat{y} = \underset{y \in \mathbf{y}}{\operatorname{argmax}} \operatorname{aligner}(y|x) \tag{2}$$

where $\operatorname{aligner}(\cdot|\cdot)$ denotes the alignment probability that can be computed by the word aligners such as fast_align (Dyer et al., 2013). Considering that there exists no semantic equivalent for some words (e.g., "*the*" in English), we construct the semantic-equivalent word sets $\mathcal{N}_{\text{word}}(\mathbf{x}, \mathbf{y})$ as mutually aligned word pairs in $(\mathbf{x}, \mathbf{y})$. For each aligned word pair $(x, y) \in \mathcal{N}_{\text{word}}(\mathbf{x}, \mathbf{y})$, the representations $(u, v^+, \boldsymbol{v}^-)$ in Eq. (1) can be computed as:

$$\begin{cases} u = \ell_2(h_x) \\ v^+ = \ell_2(h_y) \\ \boldsymbol{v}^- = \left\{ \ell_2(h_z) \middle| z \in \mathbf{y} \backslash y \right\} \end{cases} \tag{3}$$

where ℓ_2 represents ℓ_2-normalization and $\mathbf{y}\backslash y$ denotes words in $\mathbf{y}$ other than the word y. Finally, the word-level hard contrasting loss for the source sentence $\mathbf{x}$ is formalized as:

$$\mathcal{L}_{\text{word}}(\mathbf{x}) = \sum_{(x,y) \in \mathcal{N}_{\text{word}}(\mathbf{x},\mathbf{y})} \mathcal{L}(u, v^+, \boldsymbol{v}^-) \tag{4}$$

The loss $\mathcal{L}_{\text{word}}(\mathbf{y})$ for the target sentence $\mathbf{y}$ can be computed in a similar way by swapping $(\mathbf{x}, \mathbf{y})$ to $(\mathbf{y}, \mathbf{x})$. Due to space limitations, here we omit the related details.

Soft contrasting Due to the strict requirements on the quality of constructed pseudo-parallel bilingual word pairs, hard contrasting is prone to suffer from potential error propagation induced by external word aligners. In addition, some source words may correspond to multiple target words, which conflicts with the strict one-to-one alignment of hard contrasting. To tackle the above issues, we propose soft contrasting, aiming at learning word alignment implicitly and jointly to approximate semantic equivalents via the attention mechanism (Vaswani et al., 2017). Specifically, for each word x in the source sentence $\mathbf{x}$, the aggregated representation $\operatorname{MHA}(h_x, \boldsymbol{h}_\mathbf{y})$ can be obtained by performing multi-head attention[1] with h_x serving as the query and $\boldsymbol{h}_\mathbf{y}$ serving as the keys/values. Since multi-head attention naturally assign larger weights to the words in $\mathbf{y}$ that are aligned to x, $\operatorname{MHA}(h_x, \boldsymbol{h}_\mathbf{y})$ can be regarded as an approximation of the representation of semantic equivalent of x. Therefore, the representations $(u, v^+, \boldsymbol{v}^-)$ in Eq. (1) can be defined as:

$$\begin{cases} u = \ell_2(h_x) \\ v^+ = \ell_2\big(\operatorname{MHA}(h_x, \boldsymbol{h}_\mathbf{y})\big) \\ \boldsymbol{v}^- = \left\{ \ell_2\big(\operatorname{MHA}(h_z, \boldsymbol{h}_\mathbf{y})\big) \middle| z \in \mathbf{x} \backslash x \right\} \end{cases} \tag{5}$$

where $\mathbf{x}\backslash x$ refers to the remaining words in $\mathbf{x}$ except x. Soft contrasting not only alleviates the dependence on external word aligners, but also frees the model from the limitations of one-to-one alignment. Additionally, by maximizing the semantic consistency between h_x and $\operatorname{MHA}(h_x, \boldsymbol{h}_\mathbf{y})$, the model is encouraged to learn word alignment in an implicit manner, introducing richer cross-lingual information.

2.3 Span-Level Contrasting

Previous work (Joshi et al., 2019) has demonstrated the superiority of span-level representations over word-level (Devlin et al., 2019) representations due to its strength in language understanding and reasoning. Therefore, we also perform contrasting based on semantic-equivalent spans. Since span gets rid of the limitation that the semantic equivalents of the two languages must share the same number of words, here we focus on the application of hard contrasting. To be specific, given the bilingual instance $(\mathbf{x}, \mathbf{y})$, we induce the phrase table via statistical machine translation tools to obtain span-level semantic equivalents $\mathcal{N}_{\text{span}}(\mathbf{x}, \mathbf{y})$. For each aligned span pair $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \in \mathcal{N}_{\text{span}}(\mathbf{x}, \mathbf{y})$ where $\bar{\boldsymbol{x}} \subset \mathbf{x}$ is a span of $\mathbf{x}$ and $\bar{\boldsymbol{y}} \subset \mathbf{y}$ is a span of $\mathbf{y}$, the representations $(u, v^+, \boldsymbol{v}^-)$ in Eq. (1) can be formulated as:

$$\begin{cases} u = \ell_2\big(\operatorname{MP}(\boldsymbol{h}_{\bar{\boldsymbol{x}}})\big) \\ v^+ = \ell_2\big(\operatorname{MP}(\boldsymbol{h}_{\bar{\boldsymbol{y}}})\big) \\ \boldsymbol{v}^- = \left\{ \ell_2\big(\operatorname{MP}(\boldsymbol{h}_{\bar{\boldsymbol{z}}})\big) \middle| \bar{\boldsymbol{z}} \in \mathbf{y} \backslash \bar{\boldsymbol{y}} \right\} \end{cases} \tag{6}$$

where $\operatorname{MP}(\cdot)$ represents the mean-pooling layer[2] employed to aggregate all hidden representations of multiple words in a span. $\boldsymbol{h}_{\bar{\boldsymbol{x}}} = (h_{\bar{x}_1}, \cdots, h_{\bar{x}_l})$ are hidden representations of span $\bar{\boldsymbol{x}} = (\bar{x}_1, \cdots, \bar{x}_l)$ and similarly for $\boldsymbol{h}_{\bar{\boldsymbol{y}}}$. The span-level contrastive loss of a given bilingual sentence pair $(\mathbf{x}, \mathbf{y})$ is

[1] The semantic similarity between different words from two languages can also be calculated by other approaches such as bilinear attention.

[2] Other similar aggregation layers such as max-pooling or attentive-pooling can also be implemented.

defined as the sum of the losses corresponding to all spans $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$ in $\mathcal{N}_{\text{span}}(\mathbf{x}, \mathbf{y})$, whose calculation is similar to Eq. (4).

2.4 Sentence-Level Contrasting

In order to improve the quality of learned sentence embeddings, we also perform sentence-level contrastive learning to obtain the global supervision signals aggregating all token representations of the entire source input. Our proposed approach strives to pull the sentence representation of $\mathbf{x}$ towards that of its corresponding translation $\mathbf{y}$, and push it away from sentence representations of other instances. However, the direct application of artificially constructed parallel bilingual sentence pairs tends to result in a significant boundary between positive and negative samples, which may lead to vanishing contrasting signals. To remedy this problem, we make use of back-translation to infuse noise in original positive samples to obtain more competitive cross-lingual information. To be more specific, we define the sentence-level semantic equivalents $\mathcal{N}_{\text{sent}}(\mathbf{x}, \mathbf{y})$ as:

$$\mathcal{N}_{\text{sent}}(\mathbf{x}, \mathbf{y}) = \Big\{ (\boldsymbol{x}, \boldsymbol{y}) \big| \boldsymbol{x} \in \{\mathbf{x}, \hat{\mathbf{x}}\}, \boldsymbol{y} \in \{\mathbf{y}, \hat{\mathbf{y}}\} \Big\} \tag{7}$$

where $\hat{\mathbf{x}}$ is the noisy version of $\mathbf{x}$ obtained by means of back-translation[3] and so is $\hat{\mathbf{y}}$. By sampling from the original $\mathbf{x}$ and the back-translated $\hat{\mathbf{x}}$, both sentences from the two languages for contrasting contain a certain amount of noise. This blurs the boundary between the positive and negative representations to some extent, thereby effectively alleviating the vanish of contrasting signals.

As with span-level contrasting, we adopt the mean-pooling layer to aggregate all token representations of a given sentence into its corresponding sentence representation. For each sentence pair $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{N}_{\text{sent}}(\mathbf{x}, \mathbf{y})$, we define the representations $(u, v^+, \boldsymbol{v}^-)$ in Eq. (1) for sentence-level contrasting as:

$$\begin{cases} u = \ell_2\big(\text{MP}(\boldsymbol{h}_{\boldsymbol{x}})\big) \\ v^+ = \ell_2\big(\text{MP}(\boldsymbol{h}_{\boldsymbol{y}})\big) \\ \boldsymbol{v}^- = \Big\{ \ell_2\big(\text{MP}(\boldsymbol{h}_{\boldsymbol{z}})\big) \big| \boldsymbol{z} \neq \mathbf{y} \Big\} \end{cases} \tag{8}$$

where $\boldsymbol{z} \neq \mathbf{y}$ means that the negative representations used for contrasting are derived from other instances in the same mini-batch.

[3] In the implementation, we obtain multiple $\hat{\mathbf{x}}$ by pre-training the *target→source* translation model and performing beam search or top-k sampling.

Following XLM (Lample and Conneau, 2019) and XLM-R (Conneau et al., 2019), to learn from bidirectional contexts, we also adopt masked language modeling (MLM) as one of pre-training tasks. The MLM task aims at predicting the masked words based on the corrupted input. We concatenate a parallel bilingual sentence pair into a single sentence and randomly select 15% tokens as candidates for performing corruption. Of these selected tokens, 80% are replaced with special token `[MASK]`, 10% are kept unchanged, and the remaining are replaced by randomly selected vocabulary tokens. The final training objective is defined as the sum of the above-mentioned multiple contrastive losses as well as the cross-entropy of MLM.

3 Experiments

We conduct experiments on a variety of downstream tasks, which can be divided into two categories: machine translation and cross-lingual language understanding tasks.

3.1 Pre-Training

Datasets We pre-train our model on large-scale datasets involving the 15 languages of XNLI (Conneau et al., 2018)[4]: English, French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili, and Urdu. Following Conneau et al. (2019), we reconstruct Common-Crawl Corpus to obtain monolingual training datasets while the bilingual data is obtained from the OPUS website [5]. We also conduct up/down-sample for all pre-training data with a smoothing parameter. The sentence-piece model (SPM) (Kudo and Richardson, 2018) provided by Conneau et al. (2019)[6] is employed to tokenize all training data.

Model architecture We implement the proposed approach based on the Transformer (Vaswani et al., 2017) encoder with 12 identical layers, each of which consists of a multi-head attention module and a feed-forward network. The model dimension and the number of heads are set to 768 and 12, respectively, with the inner size of the feed-forward network being 3072. We choose GeLU (Hendrycks and Gimpel, 2016) as our activation function. We

[4] https://github.com/facebookresearch/XNLI

[5] http://opus.nlpl.eu/

[6] https://github.com/google/sentencepiece

use the sentence-piece vocabulary provided by Conneau et al. (2019), whose size is 250K.

Training parameters We apply Adam (Kingma and Ba, 2015) optimizer with a learning rate of 5×10^{-4} and adopt invert linear decay schedule to pre-train our models. We employ a dropout with probability to 0.1 for both the hidden states and the attention distribution. The temperature τ in Eq. (1) is set to 0.1 and the coefficients of all contrastive losses are set to 1. We take advantage of the gradient accumulation technique to simulate the batch size of 512. Our model is initialized with the pre-trained checkpoint released by Conneau et al. (2019)[7] and then pre-trained on 8×32GB NVIDIA V100 GPUs with mixed-precision floating-point arithmetic. Overall, it took about 3 weeks to converge.

3.2 Machine Translation

Datasets We conduct experiments on three widely-used machine translation datasets of various training data sizes: IWSLT14 De-En (160K), WMT14 En-De (4.5M), and WMT14 En-Fr (36M). The sentence-piece tokenization with the same vocabulary as pre-training are used to tokenize all translation samples. The BLEU score computed by the *multibleu.perl* script[8] is used as the evaluation metric for all translation tasks.

Model architecture We use the pre-trained model to initialize a 12-layer encoder. The decoder is implemented as a standard 6-layer Transformer (Vaswani et al., 2017) decoder, each layer consisting of a multi-head self-attention module, a multi-head cross-attention module and a feed-forward network. The entire decoder is initialized randomly, and it is jointly trained with the pre-trained encoder. Other model hyper-parameters including the hidden size, the number of heads, the inner size of the feed-forward network, and the choice of activation function are identical to the encoder.

Training parameters We also adopt mixed-precision floating-point arithmetic to train our models on the machine translation task. The experiments are conducted on 8×32GB NVIDIA V100 GPUs. We use an Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ to optimize our model during training. The learning rate is warmed up to 1×10^{-4} linearly in the first 4000 updates and then decays at a rate proportional to the inverse square root of the update number. We average the last 10 checkpoints as the final model and perform beam search with a beam size of 5 during inference. The probability of dropout is set to 0.1 to avoid over-fitting. The length penalty is set to 1.0.

3.3 Cross-Lingual Language Understanding

Dataset In order to verify the effectiveness of our approach on cross-lingual language understanding, we conduct evaluation on XNLI (Conneau et al., 2018) dataset. It is an extension of the English natural language inference dataset MultiNLI (Williams et al., 2018) where the development and test sets come in 15 different languages. The training set contains ~392K English samples and the test set for each language contains $5,010$ samples.

Model architecture We use the same model architecture as under the pre-training setting. Following Conneau et al. (2019), we update all parameters of our model after adding a linear classifier on top of the hidden state of the first token when fine-tuning on XNLI.

Training parameters During fine-tuning, the base transformer model is optimized along with the extra linear classifier using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 0.000025. The dropout rate is set to 0.1. We fine-tune our model on 4 NVIDIA GeForce RTX 2080Ti GPUs with a batch size of 8 sequences per GPU.

4 Results and Analysis

This section presents the detailed experiment results of different systems. We perform the evaluation on a comprehensive suite of benchmark tasks, covering cross-lingual classification and machine translation.

4.1 Cross-Lingual Classification

Following Lample and Conneau (2019), we perform evaluation on the cross-lingual natural language inference (XNLI) benchmark, where the model needs to determine the relation (*entailment*, *contradiction* and *neutral*) between the given *premise* and *hypothesis* sentences. We compare different systems under two settings. (1) Cross-Lingual Transfer: we fine-tune the model on the

[7] https://github.com/pytorch/fairseq/tree/master/examples/xlmr

[8] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

Models	#M	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																	
MBERT*	N	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
XLM (w/o TLM)*	1	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM (w/o TLM)*	N	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
XLM*	N	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	**68.4**	67.3	75.1
UNICODER†	1	85.4	79.2	79.8	78.2	77.3	78.5	76.7	73.8	73.9	75.9	71.8	74.7	70.1	67.4	66.3	75.3
XLM-R*	1	85.8	79.7	80.7	78.7	77.5	**79.6**	**78.1**	74.2	73.8	76.5	**74.6**	**76.7**	**72.4**	66.5	**68.3**	**76.2**
INFOXLM†	1	**86.4**	**80.6**	80.8	**78.9**	**77.8**	78.9	77.6	**75.6**	**74.0**	**77.0**	73.7	**76.7**	72.0	66.4	67.1	**76.2**
XLM-R (reimpl)	1	84.3	78.3	79.1	76.9	75.3	77.8	75.6	74.1	71.9	75.9	72.3	73.5	70.2	64.3	67.1	74.4
MGC-HARD	1	85.9	79.9	**80.9**	78.3	77.5	79.1	76.8	74.0	73.1	76.6	73.0	75.4	71.7	66.3	68.1	75.8
MGC-SOFT	1	86.3	79.6	80.8	78.5	**77.8**	79.3	77.3	73.9	73.4	76.9	73.3	76.1	71.9	66.5	67.9	76.0
Fine-tune multilingual model on all training sets (Translate-Train-All)																	
UNICODER†	1	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
XLM (w/o TLM)*	1	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM*	1	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
XLM-R*	1	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	**73.0**	79.1
INFOXLM†	1	86.1	82.0	82.8	**81.8**	80.9	82.0	80.2	79.0	78.8	80.5	78.3	80.5	**77.4**	73.0	71.6	79.7
XLM-R (reimpl)	1	84.4	81.0	81.5	81.0	80.2	80.9	79.2	77.0	77.9	79.8	77.0	78.5	74.6	71.9	70.4	78.4
MGC-HARD	1	86.0	82.6	82.7	**81.8**	**81.5**	82.6	81.3	78.6	**79.5**	80.9	**80.1**	81.5	76.7	74.0	72.1	80.1
MGC-SOFT	1	**86.5**	**82.7**	**83.0**	81.5	81.3	**82.7**	**81.8**	**79.1**	79.4	**81.3**	79.5	**81.8**	76.6	**74.4**	71.3	**80.2**

Table 1: The performance of different systems on XNLI task. "#M=N" indicates that each language is assigned a separate model based on the performance of the respective dev set, while "#M=1" means only one model is used for all languages. Results with "*" and "†" are taken from Conneau et al. (2019) and Chi et al. (2020), respectively. "(reimpl)" means our own implementation using the same training strategy. The best performance is bolded.

English training set and then directly evaluate on the test sets of the 15 languages. (2) Translate-Train-All: we fine-tune the model on the combined data consisting of English training data and pseudo data that are translated from English to other languages. As indicated by the results in Table 1, our method manages to maintain a consistent improvement over our implemented XLM-R under both settings, raising the average accuracy from 74.4% to 76.0% for cross-lingual transfer. The similar conclusions can be drawn from the translate-train-all setting, where our approach boosts XLM-R by an increment of 1.8 average accuracy and also outperforms all other baselines. By means of multi-granularity contrasting, our approach succeeds in introducing more fine-grained and explicit alignment supervision, which enhances the capability of the model to learn language-universal representations.

In addition, we can also note that there is no definite conclusion about the superiority of the two contrasting strategies for word-level contrastive loss. Hard contrasting attempts to integrate explicit cross-lingual supervision from external word aligners, while soft contrasting aims at enabling the pre-trained model to learn word alignment implicitly via semantic attention. Both strategies contribute to the introduction of more fine-grained and explicit cross-lingual information, thereby lifting the performance of the pre-trained model in downstream scenarios.

4.2 Machine Translation

Table 2 presents the comparison between our approach and several representative systems on machine translation. The results once again confirm that large-scale pre-training can effectively accomplish model transferring and advance the performance of machine translation, as all pre-trained models outperform the unpretrained transformer. In addition, we observe the significant performance gain for our approach compared to the baselines. For instance, it achieves a 1.6% improvement of BLEU score over the base architecture XLM (Lample and Conneau, 2019) on the IWSLT14 DE-EN task. It also surpasses other competitive baselines such as mBERT (Devlin et al., 2019) and MASS (Song et al., 2019) on all three translation benchmarks by a wide margin. The results effectively demonstrates the ability of our approach to learn better representations for semantic equivalents across languages, as well as the versatility of our approach, which can be applied to both language understanding and generation tasks.

Models	IWSLT14 DE-EN	WMT14 EN-DE	WMT14 EN-FR
TRANSFORMER	34.5	28.4	41.9
MBERT	34.8	28.6	–
MASS	35.1	28.9	–
XLM	35.2	28.9	–
ALM	35.5	29.2	–
XLM-R	35.1	30.1	42.4
MGC-HARD	36.4	30.2	**43.1**
MGC-SOFT	**36.8**	**30.6**	42.9

Table 2: The experiment results of different systems on machine translation. The best performance is bolded.

4.3 Ablation Study

We conduct an ablation study on several major components of our approach to explore their influence on cross-lingual pre-training, including the multi-granularity contrastive losses and sentence-level semantic equivalent augmentation with back-translation.

Effect of multi-granularity contrastive losses To understand how much different levels of contrasting account for the overall performance improvement, we train the same model but with different contrastive losses. First, as shown in Table 3, all three levels of contrasting contribute to the superiority of our model. This demonstrates that the incorporation of contrastive learning can truly introduce training signals that are beneficial for cross-lingual pre-training on multiple granularities. Among them, sentence-level contrasting has the largest impact, the removal of which results in a drop of 1.3 BLEU score on WMT14 EN-DE. The reason behind this phenomenon may be that this loss makes up for the relative lack of explicit sentence-level training signals in XLM pre-training.

Effect of sentence-level semantic equivalent augmentation To investigate whether augmenting the semantic equivalents by means of back-translation improves sentence-level contrasting, we compare our model (BTSET SENT-CONTRAST) against a variant where only the original source and target sentence are used to compute the sentence-level contrastive loss (BIPAIR SENT-CONTRAST). The results are presented in Table 4. As can be seen, back-translation leads to an improvement of 0.7 BLEU on WMT14 DE-EN, illustrating its efficacy in alleviating vanishing contrasting signals and boosting cross-lingual pre-training.

Models	WMT14 EN-DE
W/O WORD-CONTRAST	29.8
W/O SPAN-CONTRAST	29.5
W/O SENT-CONTRAST	29.3
FULL MGC-SOFT	30.6

Table 3: The results of ablation study on WMT14 DE-EN translation task.

5 Related Work

The existing efforts performing cross-lingual pre-training mainly consist of two typical paradigms: traditional *embedding alignment* and the recent prevalent *language modeling*.

5.1 Embedding Alignment

Early endeavors regarding cross-lingual pre-training mainly focus on embedding alignment, which aims at learning similar vector representations for semantic-equivalent units. Representative approaches can be categorized into four research lines: regression model, hinge loss, canonical analysis, and linear projection. Based on the observation that the monolingual word embeddings share similar geometric properties across languages, simple but effective linear projection approaches have become mainstream, which aim at aligning two disjoint monolingual vector spaces through a linear transformation. For instance, Mikolov et al. (2013a) propose to learn the desired linear projection by minimizing the mean squared error between the projected source embeddings and the target embeddings. Xing et al. (2015) refine this method by imposing orthogonality constraint and maximizing the cosine similarity. Artetxe et al. (2017) explore the bilingual induction in extremely low-resource scenarios via an effective self-learning framework. Furthermore, unsupervised embedding alignment (Lample et al., 2018; Yang et al., 2019) completely eliminates the dependence on parallel data, which aims to learn cross-lingual word embeddings in the absence of any aligned word pairs. The related approaches can be summarized as: GAN-based distribution matching (Zhang et al., 2017; Lample et al., 2018), non-adversarial distribution matching, heuristic alignment, generalized Pruck analysis and so on. However, the traditional embedding alignment can only learn non-contextualized word representations, which suffers from intractable polysemy problem. Compared with the following language modeling that captures bidirectional contexts and employs large-capacity

Model	WMT14 EN-DE
BIPAIR SENT-CONTRAST	29.9
BTSET SENT-CONTRAST	30.6

Table 4: The comparison of two sentence-level contrasting strategies. "BIPAIR SENT-CONTRAST" and "BTSET SENT-CONTRAST" means sentence-level contrasting with the original bilingual pair and extended back-translation set as positive examples, respectively.

transformer, it tends to result in suboptimal performance in downstream tasks.

5.2 Language Modeling

This research line focuses on masked language modeling (MLM), which aims to predict the masked words based on the corrupted input. In terms of model architecture, one paradigm attempts to capture language-universal representations via *a single encoder*. For instance, Multilingual BERT (Devlin et al., 2019) applies byte-pair encoding (BPE) to merge tokens from 104 different languages into a shared vocabulary and performs MLM on the monolingual sentences. XLM (Lample and Conneau, 2019) extends it to translation language modeling (TLM), which strives to predict the masked words by attending to both source and target sentences. With the mutual attention of bilingual contexts, the model is expected to align representations from two languages in an implicit manner. Unicoder (Huang et al., 2019) introduces several more pre-training tasks such as cross-lingual word recovery, illustrating that these tasks can boost model performance by learning interlingual mapping from more perspectives. ALM (Yang et al.) constructs large-scale instances for masked language modeling by alternatively selecting words from source and target languages. Ren et al. (2019) task the model with predicting the translation of masked n-grams, with the phrase table inferred from monolingual corpora in advance as ground truth. Conneau et al. (2019) pre-train their model using more than two terabytes of filtered CommonCrawl data, demonstrating that large-scale dataset can lead to significant performance gains.

The other research line draws on the idea of the *encoder-decoder* framework and aims to mimic autoregressive generation by generating the target texts based on the given source input. For instance, MASS (Song et al., 2019) jointly trains the encoder and decoder by reconstructing the desired sentence fragment based on the remaining part of the sentence, which enhances the capabilities of the model in feature extraction and language modeling. XNLG (Chi et al., 2019) strives to learn language-universal representations by extending monolingual masked language modeling and denoising autoencoding to cross-lingual settings. mBART (Liu et al., 2020) pre-trains the encoder-decoder by reconstructing the original text based on the corrupted input with an arbitrary noising function, which can be used directly to initialize text generation models or as a denoising strategy for language understanding. However, both lines mentioned above for language modeling focus on projecting the input from different languages into the same semantic space through shared vocabulary and representation models. Compared with traditional embedding alignment, it lacks the introduction of cross-lingual information with more explicit and fine-grained (e.g., word-level) alignment.

Our proposed approach effectively inherits the advantages of both *embedding alignment* and *language modeling*, while avoiding their limitations. It not only captures bidirectional contexts with large-capacity transformer model and MLM task, but also introduces more fine-grained cross-lingual supervision by applying contrastive learning on semantic units of multiple granularities, thereby obtaining significant performance gains.

6 Conclusion

This paper presents a multi-granularity contrastive cross-language pre-training framework, which combines traditional embedding alignment and the recent prevalent language modeling to learn language-universal prior representations . Different from previous work focusing on masked language modeling to capture bidirectional contexts, the proposed approach introduces more fine-grained and explicit cross-lingual supervision by maximizing the representational consistency of semantic equivalents from different languages. Two effective contrasting strategies are proposed, which can be built upon semantic units with different granularity covering word, span, and sentence. Comprehensive empirical evidence illustrates that our approach can achieve consistent improvement on a variety of downstream tasks including machine translation and cross-lingual language understanding.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2019. Cross-lingual natural language generation via pre-training. *arXiv preprint arXiv:1909.10481*.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. In *EMNLP*, pages 2475–2485. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *NAACL HLT 2013*, pages 644–648.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, pages 297–304.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909.00964*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2019. SpanBERT: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP (Demonstration)*, pages 66–71. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Explicit cross-lingual pre-training for unsupervised machine translation. *arXiv preprint arXiv:1909.00180*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.

Jian Yang, Shuming Ma, Dongdong Zhang, ShuangZhi Wu, Zhoujun Li, and Ming Zhou. Alternating language modeling for cross-lingual pre-training.

Pengcheng Yang, Fuli Luo, Peng Chen, Tianyu Liu, and Xu Sun. 2019. MAAM: A morphology-aware alignment model for unsupervised bilingual lexicon induction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3190–3196. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970.