TRIP: Accelerating Document-level Multilingual Pre-training via Triangular Document-level Pretraining on Parallel Data Triplets

Hongyuan Lu, Haoyang Huang, Shuming Ma, Dongdong Zhang, Wai Lam, Zhaochuan Gao, Anthony Aue, Arul Menezes, Furu Wei

Abstract

Despite the success of multilingual sequence-to-sequence pre-training, most existing approaches rely on document-level monolingual corpora in many different languages, sentence-level bilingual corpora, and sometimes synthetic document-level bilingual corpora. This hampers the performance with cross-lingual document-level translation tasks such as document-level translation. Hence, we propose to mine and leverage document-level trilingual parallel corpora to improve sequence-to-sequence multilingual pre-training. We present Triangular Document-level Pre-training (TRIP) as the first in the field to accelerate the conventional monolingual and bilingual objectives into a trilingual objective with a novel method called Grafting. Experiments show that TRIP achieves several strong state-of-the-art (SOTA) scores on three multilingual document-level machine translation benchmarks and one cross-lingual abstractive summarization benchmark, including consistent improvements by up to 3.11 d-BLEU points and 8.9 ROUGE-L points.

1 Introduction

Conventional multilingual pre-training achieved promising results on machine translation (Liu et al., 2020) and cross-lingual classification (Xue et al., 2021). These pre-training paradigms usually rely on monolingual corpora in many different languages, with denoising objectives such as sentence permutation and span masking (Liu et al., 2020). Following the calls that the unsupervised scenario is not strictly realistic for cross-lingual learning (Artetxe et al., 2020), multilingual pre-training advanced into a supervised setting through sentence-level bilingual translation pairs (Chi et al., 2021; Reid and Artetxe, 2022) to provide a stronger signal for pre-training. Among these pioneering works, document-level multilingual pre-training with parallel data is currently an understudied topic. This direction is particularly significant for tasks that necessitate contextual comprehension, such as document-level machine translation and cross-lingual summarization. As a workaround, DOCmT5 (Lee et al., 2022) resorts to using synthetic bilingual translation pairs to scale up document-level multilingual pre-training.

In addition to the lack of study for document-level multilingual pre-training with parallel data, prior works also overlooked the value of trilingual parallel data for multilingual pre-training. Compared to bilingual parallel data, trilingual parallel data is expected to better capture different linguistic clues and coherence among different languages such as past tense and gendered expressions, which can enhance the model pre-training on aspects of document-level cross-lingual understanding and resolve cross-lingual ambiguities.

To this end, we present TRIP, a document-level multilingual pre-training method using trilingual parallel corpora. Because there is no publicly available document-level trilingual corpus, we propose a novel method to construct trilingual document pairs from document-level bilingual corpora. Subsequently, we augment the conventional multilingual pre-training by (i) Grafting two documents presented in two different languages into one mixed document, and (ii) predicting the remaining one language as the reference translation.

7845

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 7845–7858
December 6-10, 2023 ©2023 Association for Computational Linguistics
We conduct experiments on document-level machine translation on TED Talks (Cettolo et al., 2015), News benchmark (News-commentary) and Europarl (Koehn, 2005), and cross-lingual abstractive summarization on Wikilingua (Ladhak et al., 2020; Gehrmann et al., 2021). We found that TRIP clearly improves previous multilingual pre-training paradigms that use monolingual and bilingual objectives (Lee et al., 2022), and achieves strong SOTA results on both tasks.

In summary, we make three key contributions:

- **TRIP** proposes a novel trilingual pre-training objective through **Grafting** for multilingual pre-training, along with a novel method to construct trilingual data from parallel corpora.
- **TRIP** yields SOTA scores on both multilingual document-level MT and cross-lingual abstractive summarization.
- We conduct in-depth analyses on document-level cross-lingual understanding and compare **TRIP** to commercial systems.

## 2 Triangular Document-level Pre-training

We start by introducing the conventional methodologies previously used by the monolingual and bilingual objectives for multilingual pre-training:

- **Denoising Pre-training**: Sentence permutation (Liu et al., 2020) and span corruption (Xue et al., 2021) are effective denoising pre-training objectives for document-level multilingual pre-training.

- **Translation Pre-training**: Making the use of sentence-level translation pairs is a bilingual pre-training strategy for multilingual models (Kale et al., 2021; Tang et al., 2021).

### Constructing a Trilingual Objective

In comparison, **TRIP** is the first in the field to introduce a trilingual objective for multilingual pre-training. The core to making better use of trilingual data is to **Grafting** the documents by splitting the documents written in two different languages but with the same meaning half by half and concatenating each half to form a new document that retains the same meaning written in two different languages. **TRIP** then applies sentence permutation and span corruption on the Grafted documents.

Conventional monolingual and bilingual pre-training objectives overlooked the value to take such an advantage (Liu et al., 2020; Reid and Artetxe, 2022) of linguistic clues from different languages. In contrast, **TRIP** fuses authentic trilingual data, in which linguistic clues such as past tense and gendered nouns are usually preserved.

We present in Figure 1 to illustrate how **TRIP** operates to make use of linguistic clues through trilingual data. Given three documents with the same meaning written in Chinese, Japanese, and English, two of the documents are split and concatenated. The concatenation is randomly permuted at the sentence level, and the remaining unchanged document is used as the translation reference. Here, Chinese is tenseless, and **TRIP** effectively fuses useful linguistic clues for past tense written in Japanese and English into the Chinese text to resolve cross-lingual ambiguities.

### Table 1

<table>
<thead>
<tr>
<th>Models</th>
<th>Denoising Pre-training</th>
<th>Translation Pre-training</th>
<th>Trilingual Document Pairs</th>
<th>Objective</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBART</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>mT5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>mT6</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>PARADISE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>DOCmT5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>TRIP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 1: **Comparisons of various multilingual pre-training methods.** We denote the intermediate value as ○. For example, mT5 uses span corruption solely without sentence permutation, so we put a value of ○ for the column of Denoising Pre-training for mT5. The columns of **Denoising Pre-training** and **Translation Pre-training** refer to the pre-training objectives we introduce at the start of Section 2.

3**Grafting** refers to joining two plants together by cutting and using scion (the upper part of the grafting) as the top and the understock (the lower part of the grafting) as the root.

7846
**Creating Trilingual Document Pairs**

As there is no public corpus with trilingual document pairs, TRIP creates MTDD (Microsoft Trilingual Document Dataset), a high-quality trilingual parallel corpus with document translation pairs across 67 languages, 4,422 bilingual directions, and 99,628 trilingual combinations. The corpus is sourced from high-quality news documents scraped from an in-house website\(^5\) timestamped from April 2021 to July 2022. The whole procedure is composed of two steps: (i) creating bilingual document pairs and (ii) creating trilingual document pairs based on the bilingual document pairs.

To obtain bilingual document pairs, we follow ParaCrawl (Bañón et al., 2020) to translate all the documents we have into English using a lightweight word-based machine translation model.\(^4\)

As a pre-training method, TRIP is robust and does not require the sentences in trilingual document pairs to be perfectly aligned in their orderings. Filtering the non-perfect pairs can throw away the data and deteriorate the performance gains.

The resulting translation is used for pairing only and the documents are paired and thresholded with similarity scores such as tf-idf computed on their English translation (Bañón et al., 2020). To improve efficiency, we attempt to pair documents only if they are timestamped within a small window such as one week. The motivation is that the semantic news with the same meaning in different languages are often reported within a small timestamp window in high probabilities. The resulting document pairs are further thresholded and filtered with LASER (Artetxe and Schwenk, 2018),\(^6\) which is a multilingual sentence representation.

Given the bilingual data constructed as above,
we follow previous works (Bañón et al., 2020; El-Kishky et al., 2020). These previous works leverage URL addresses for constructing bilingual data. In contrast, we use URL addresses to construct trilingual data pairs by matching and linking. Figure 2 depicts a detailed illustration.

For space reasons, we present statistics to illustrate the scale of MTDD in Table 10 in Appendix D. We also note that existing MTmC4 (Lee et al., 2022) used by DOCmT5 can be less favourable for our experiments as (i) MTmC4 is composed of synthetic data that could be of lower quality, (ii) MTmC4 is not publicly available at the time of writing, and (iii) MTmC4 can lead to potential data leakage for the test sets on TED Talks.

3 Experiments

3.1 TRIP Pre-training

Model Configuration We use a Transformer architecture that is composed of 24 Transformer encoder layers and 12 interleaved decoder layers. In addition, it has an embedding size of 1024, and a dropout rate of 0.1. The feed-forward network is configured to have a size of 4096 with 16 attention heads. For parameter initialization, we follow Ma et al. (2021) and Yang et al. (2021) to train a sentence-level MT system. The motivation is that previous studies have shown that the hybrid training of sentence-level and document-level MT can improve the performance of document-level translation (Sun et al., 2022). We call it the Baseline Model in the remaining of this paper.

Data and Pre-processing As described in Section 2, we create a trilingual document-level corpus, MTDD, for TRIP pre-training with the use of trilingual document pairs. We create a list of keywords to automatically clean and remove noisy text such as claims and advertisements. We follow Ma et al. (2021) to use SentencePiece (Kudo and Richardson, 2018) for tokenization, and we use the same SentencePiece model as Yang et al. (2021). Following the previous works, we prefix the inputs with a language tag that indicate the target language of the generation for both pre-training and fine-tuning.

Training Details We use the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ for our multilingual pre-training. The learning rate is set as $1e^{-5}$ with a warmup step of 4000. We use the label smoothing cross-entropy for our translation loss and we set label smoothing with a ratio of 0.1 for model training. All of our pre-trainings are conducted on 16 NVIDIA V100 GPUs. We set the batch size as 512 tokens per GPU. To simulate a larger batch size, we update the model every 128 steps. For the Grafting operation $\mathcal{T}$ defined for TRIP, we split the documents 50% by 50%.

3.2 Multilingual Document-level MT

3.2.1 TED Talks

Experimental Settings Following DOCmT5, we use the IWSLT15 Campaign for the evaluation of TED Talks. Prior systems have reported scores on only 1 or 2 translation directions (Lee et al., 2022; Sun et al., 2022), and DOCmT5 supports only the translation direction into English (X $\rightarrow$ En). We report more language directions while DOCmT5 only evaluates on (Zh $\rightarrow$ En). Following DOCmT5, we split all documents into a maximum of 512 tokens for all train/dev/test sets during training and inference. We use the official parallel training data from IWSLT15 without any additional monolingual data, with the official 2010 dev set and 2010-2013 test set for evaluation (Lee et al., 2022). We compute d-BLEU (Papineni et al., 2002; Liu et al., 2020; Bao et al., 2021), a BLEU score for documents.
We use SacreBLEU for evaluation.\footnote{https://github.com/mjpost/sacrebleu}

**Baseline Systems** We report strong baselines evaluated at both sentence and document levels, including SOTA models DOCmT5+\footnote{(Lee et al., 2022), MARGE\footnote{(Miculicich et al., 2018), MARGE\footnote{(Lewis et al., 2020a), and the Baseline Model that}} (Fan et al., 2022), M2M-100 (Fan et al., 2022), mBART (Liu et al., 2020), HAN+\footnote{SOTA models PARADISE (Reid and Artetxe, 2022), a pre-trained model that uses dictionary denoising on monolingual data, as its weights are not publicly available so far. During our trials, we found that monolingual dictionary denoising can degrade document-level systems. We think that it could better serve sentence-level tasks such as sentence-level MT and cross-lingual classification as conducted in its original paper. See Appendix C for the number of model parameters.}

### Table 3: Results for document-level MT on the News benchmark in the direction of (X → En).

<table>
<thead>
<tr>
<th>Model</th>
<th>Fr→En</th>
<th>De→En</th>
<th>Zh→En</th>
<th>Cs→En</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentence-level MT Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2M-100</td>
<td>31.58</td>
<td>25.65</td>
<td>18.47</td>
<td>28.17</td>
<td>25.97</td>
</tr>
<tr>
<td>mBART</td>
<td>29.93</td>
<td>29.31</td>
<td>18.33</td>
<td>30.15</td>
<td>26.93</td>
</tr>
<tr>
<td>Baseline Model</td>
<td>35.59</td>
<td>34.71</td>
<td>27.23</td>
<td>37.39</td>
<td>33.73</td>
</tr>
<tr>
<td><strong>Document-level MT Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2M-100</td>
<td>32.67</td>
<td>25.78</td>
<td>17.85</td>
<td>29.06</td>
<td>26.34</td>
</tr>
<tr>
<td>mBART</td>
<td>30.14</td>
<td>26.35</td>
<td>15.01</td>
<td>29.79</td>
<td>25.32</td>
</tr>
<tr>
<td>Baseline Model</td>
<td>36.38</td>
<td>34.24</td>
<td>25.58</td>
<td>36.97</td>
<td>33.29</td>
</tr>
<tr>
<td>Baseline Model+</td>
<td>38.47</td>
<td>35.20</td>
<td>26.74</td>
<td>37.26</td>
<td>34.42</td>
</tr>
<tr>
<td>TRIP (Ours)</td>
<td>39.49</td>
<td>35.48</td>
<td>27.58</td>
<td>38.06</td>
<td>35.15</td>
</tr>
</tbody>
</table>

TRIP effectively improves language pairs that are unseen during pre-training. We also found that (i) the Baseline Model+ clearly surpasses the Baseline Model and (ii) TRIP clearly surpasses the Baseline Model+. This observation indicates two points: (i) the bilingual data in MTDD used to construct the trilingual data are of high quality and (ii) the trilingual objective with the Grafting mechanism is superior to the conventional bilingual objectives for multilingual pre-training.

### 3.2.2 News

**Experimental Settings** For evaluation on the News benchmark, we follow Sun et al. (2022) to use News Commentary v11 as the training set. For Cs and De, we use newstest2015 as the dev set, and newstest2016/newstest2019 as the test set respectively. For Fr, we use newstest2013 as the dev set and newstest2015 as the test set. For Zh, we use newstest2019 as the dev set and newstest2020 as the test set. We use the same dataset preprocessing and evaluation metric as for the TED Talks.

**Baseline Systems** As the weights for DOCmT5 are not available at the time of writing, we compare our system to various strong baselines such as M2M-100, mBART, the Baseline Model, and the Baseline Model+. The scores are obtained by fine-tuning the official checkpoints.

**Results** Table 3 shows obvious and consistent improvements by up to 3.11 d-BLEU points (from 36.38 to 39.49) with TRIP for (Fr → En) compared to the Baseline Model.

### 3.2.3 Europarl

**Experimental Settings** For the Europarl dataset (Koehn, 2005), we follow Sun et al. (2022) to use Europarl-v7, and we experiment with the setting of (X → En) where we test nine languages: Da, De, El, Es, Fr, It, Nl, Pt, and Sv. Like previous works (Bao et al., 2021; Sun et al., 2022), the dataset is randomly partitioned into train/dev/test divisions. Additionally, we split by English document IDs to avoid information leakage.

**Baseline Systems** As the weights for DOCmT5 are not available at the time of writing, we compare our system to various strong baselines such as M2M-100, mBART, the Baseline Model, and
Table 4: Results for document-level machine translation on Europarl in the direction of (X → En).

<table>
<thead>
<tr>
<th>Model</th>
<th>Da→En</th>
<th>De→En</th>
<th>El→En</th>
<th>Es→En</th>
<th>Fr→En</th>
<th>It→En</th>
<th>Ni→En</th>
<th>Pt→En</th>
<th>Sv→En</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentence-level MT Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2M-100</td>
<td>50.40</td>
<td>47.38</td>
<td>52.28</td>
<td>52.03</td>
<td>48.26</td>
<td>49.70</td>
<td>46.78</td>
<td>49.84</td>
<td>52.34</td>
</tr>
<tr>
<td>mBART</td>
<td></td>
<td>48.28</td>
<td></td>
<td></td>
<td>49.16</td>
<td>50.83</td>
<td>47.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline Model</td>
<td>48.94</td>
<td>47.25</td>
<td>53.46</td>
<td>50.57</td>
<td>47.68</td>
<td>49.49</td>
<td>45.95</td>
<td>50.65</td>
<td>52.77</td>
</tr>
<tr>
<td><strong>Document-level MT Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2M-100</td>
<td>50.33</td>
<td>47.00</td>
<td>52.24</td>
<td>52.14</td>
<td>48.13</td>
<td>49.71</td>
<td>46.65</td>
<td>40.68</td>
<td>52.28</td>
</tr>
<tr>
<td>mBART</td>
<td></td>
<td>47.70</td>
<td></td>
<td></td>
<td>48.98</td>
<td>50.62</td>
<td>46.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline Model</td>
<td>49.85</td>
<td>47.64</td>
<td>53.34</td>
<td>51.32</td>
<td>48.46</td>
<td>50.26</td>
<td>47.12</td>
<td>50.13</td>
<td>52.42</td>
</tr>
<tr>
<td>Baseline Model(+)</td>
<td>49.90</td>
<td>47.75</td>
<td>53.75</td>
<td>51.78</td>
<td>48.70</td>
<td>50.37</td>
<td>47.18</td>
<td>50.49</td>
<td>52.49</td>
</tr>
<tr>
<td>TRIP (Ours)</td>
<td>51.13</td>
<td>48.30</td>
<td>54.38</td>
<td>52.29</td>
<td>49.36</td>
<td>51.23</td>
<td>48.07</td>
<td>51.03</td>
<td>53.43</td>
</tr>
</tbody>
</table>

Case Study Table 5 presents three case studies that demonstrate and compare the outputs between TRIP and the baseline systems. We highlight the correct translation in aqua and the wrong translation in hot pink. In addition to the comparison to the Baseline Models, we also present the outputs from popular commercial translation systems Google Translate, Microsoft Translator, and DeepL Translate. Each case demonstrates that TRIP is the best in terms of three characteristics respectively: (i) tense consistency (Jiang et al., 2022; Sun et al., 2022) across the sentences, (ii) noun-related issues (Jiang et al., 2022) such as singular and plural consistency as well as attaching definite article ‘the’ to a previously mentioned object ‘light’, and (iii) conjunction presence that indicates the relationship between sentences and makes the translation natural and fluent (Xiong et al., 2019; Sun et al., 2022). While some translations in the third case are acceptable, missing coordinating conjunction does not precisely capture the relationship between sentences and can make the translation less fluent.

3.2.4 Coherence and Consistency Evaluation

BlondDe Evaluation Figure 3 depicts the evaluations on TED Talks with BlondDe scores (Jiang et al., 2022), an evaluation metric designed for document-level MT which considers coherence and consistency issues that require the model to resolve cross-lingual ambiguities. Consistent improvements can be observed in all the directions on TED Talks with TRIP, meaning that TRIP generates more coherent and consistent translations than the baseline does. As discussed in Section 2, we postulate that these improvements attribute to the Grafting mechanism that resolves cross-lingual ambiguities by exploiting useful linguistic clues in trilingual data. This improves translation in coherence and consistency as reflected in the BlondDe scores. We demonstrate case studies for more analysis of coherence and consistency issues.

3.2.5 Large Language Models

Table 7 compares TRIP to popular ChatGPT (GPT-3.5-TURBO)\(^8\) on TED Talks. We use a prompt: “Translate the following text into English:”. The results indicate ChatGPT still lags behind supervised system TRIP on document-level MT. This conclusion aligns with the previous study on sentence-level MT (Zhu et al., 2023), and we postulate that the reason is ChatGPT fails in handling contexts perfectly for document-level MT.

\(8\)https://chat.openai.com/chat
Case 1: Tense Consistency (Jiang et al., 2022; Sun et al., 2022)

Source: But it's a fairly abstract discussion, and at some point when there was a pause, Octavio said, "Paul, maybe we could watch the TEDTalk.

Reference: So the TEDTalk hadn't been in very simple terms,...

Google Translate: But it's a roughly abstract discussion when at some point Octavio said, "Paul, maybe we can watch the TEDTalk." The TEDTalk and it is in a simple way,...

Microsoft Translator: But it's a roughly abstract discussion when, at certain intervals, Octavio said, "Paul, maybe we can watch TEDTalk. TEDTalk explains it in a simple way,...

DeepL Translate: But it's a broadly abstract discussion, and when there were certain breaks, Octavio said, "Paul, maybe we can watch TEDTalk. TEDTalk explains it in a very simple way,...

Baseline Model (Sentence-level): But it's kind of an abstract discussion, and at some point, Octavio said, "Paul, maybe we can watch the TEDTalk." And the TEDTalk simply explained that,...

Baseline Model (Document-level): But it's sort of an abstract discussion. And at some point, Octavio said, "Paul, maybe we can watch TEDTalk." The TEDTalk explains it in a very simple way,...

TRIP: But it's a sort of abstract discussion, and at some point in the intermission, Octavio said, "Paul, maybe we can watch the TEDTalk." And the TEDTalk made it clear,...

Case 2: Noun-related Issues

Source: ...but in the middle, there is one thing, it is a tomato. It is not red. It is not yellow, and it is not even a tomato,...

Reference: ...the light is bouncing around inside the tomato. It continues to glow. It doesn't become dark. Why is that? Because the tomato is actually ripe,...

Google Translate: as the light passes over the tomatoes. It keeps shining. It didn't get darker. Why? Because the tomatoes are ripe and the light is reflected inside the tomatoes,...

Microsoft Translator: as the light travels over the tomatoes. It kept shining. It didn't get darker. Why? Because the tomatoes are ripe and light is reflected inside the tomatoes,...

DeepL Translate: as the light travels over the tomatoes, it kept shining. It doesn't dim. Why? Because the tomatoes are ripe and the light is reflected inside the tomatoes,...

Baseline Model (Sentence-level): as the light goes over the tomato. It's always glowing. It's not darkening. Why? Because the tomato is ripe and light is reflected inside the tomato,...

Baseline Model (Document-level): as the light passes over the tomato. It keeps flashing. It doesn't get darker. Why? Because the tomatoes are ripe and the light is reflected inside the tomato,...

TRIP: as the light passes over the tomato. It's flashing all the time. It's not getting darker. Why? Because the tomato is ripe and the light is reflected inside the tomato,...

Case 3: Conjunction Presence

Source: ...I'm going to explain something, because I have already mentioned things that are important for this discussion. Let me try to illustrate what democracy might or might not look like,...

Reference: I want to suggest to you that I've been talking a lot about the future. And I want to think about the possibilities of what democracy might look like, or might have looked like, or what it already looks like if we could get more mothers involved,...

Google Translate: I want to remind that I’ve talked a lot about my predecessors. I want to think about what democracy would look like, or it already looks like what possibilities of what democracy might look like, or what it already looks like if we could get more mothers involved,...

Microsoft Translator: I want to remind you that I’ve talked a lot about my predecessors. Would also like to consider what democracy would look like, or it already be What kind of possibilities if we can involve more mothers,...

DeepL Translate: I want to remind you that I’ve talked a lot about things that have come before. I want to think about the possibilities of what democracy would look like, or what it already looks like if we could get more mothers involved in,...

Baseline Model (Sentence-level): I want to remind you that I’ve talked a lot about things before. I want to think about the possibilities of what democracy might look like, or what it would be like if we could get more mothers involved in,...

Baseline Model (Document-level): I’d like to remind you that I’ve talked about a lot of things before. I also like to think about the possibilities of what democracy might look like, or what it might be like if we could get more mothers involved in,...

TRIP: I want to remind you that I’ve talked a lot about the past. And I want to think about the possibilities of what democracy might look like, or already looks like, if we can get more mothers involved,...

Table 5: Cases from TED Talks demonstrate that TRIP captures better tense consistency, noun-related issues, and conjunction presence. We highlight the correct translation in aqua (the darker one when printed in B&W), and the mistakes in hot pink (the lighter one when printed in B&W). Google Translate: https://translate.google.com/, Microsoft Translator: https://www.bing.com/translator, DeepL Translate: https://www.deepl.com/translator. Time-stamped on 15th June 2023, can be subject to change.

3.3 Cross-lingual Abstractive Summarization

Experimental Settings: We follow the same setting used by DOCmT5 (Lee et al., 2022) to evaluate cross-lingual abstractive summarization on the benchmark of Wikilingua (Ladhak et al., 2020).

The only difference is that they put a special prefix "Summarize X to Y" where X and Y are the source and target language tags for summarization like mT5. We put a target language tag as the prefix. We use the F1 measure for ROUGE-1/ROUGE-2/ROUGE-L scores (Lin, 2004) for evaluation.

Baseline Systems: We report the scores for DOCmT5 taken from Lee et al. (2022), and we use prior SOTA scores from the official GEM benchmark (Gehrmann et al., 2021) for mT5, ByT5 (Xue et al., 2022). We also employ mBART and the Baseline Models as the baselines. See Appendix C for the number of model parameters.

Results: Table 6 demonstrates that TRIP clearly exceeds previous SOTA systems in several directions, including up to 8.9 ROUGE-L points in (Hi → En) compared to DOCmT5. Hence, we conclude that TRIP is an effective multilingual pre-training framework for cross-lingual abstractive summarization. We postulate that the improvement is attributed to the trilingual pre-training objective overlooked by previous works such as DOCmT5.

Also, we found that using bilingual data for Baseline Model† seems less beneficial on Wikilingua for cross-lingual abstractive summarization. TRIP
Table 6: Results for cross-lingual abstractive summarization on Wikilingua in (X → En). We report the scores of F-measure for ROUGE-1/ROUGE-2/ROUGE-L. *: the score is not reported. †: the scores are taken from Lee et al. (2022) and the official GEM benchmark (Gehrmann et al., 2021): https://gem-benchmark.com/results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Tr→En</th>
<th>Vi→En</th>
<th>Ru→En</th>
<th>En→En</th>
<th>Hi→En</th>
<th>Fr→En</th>
<th>Id→En</th>
</tr>
</thead>
<tbody>
<tr>
<td>mT5-XL†</td>
<td>40.0/18.3/33.3</td>
<td>37.6/14.9/31.2</td>
<td>37.2/14.6/30.9</td>
<td>41.2/17.2/34.6</td>
<td>-/-/-</td>
<td>-/-/-</td>
<td>-/-/-</td>
</tr>
<tr>
<td>ByT5†</td>
<td>35.9/15.8/29.8</td>
<td>32.7/12.2/27.2</td>
<td>31.4/11.0/26.2</td>
<td>35.1/13.5/29.1</td>
<td>-/-/-</td>
<td>-/-/-</td>
<td>-/-/-</td>
</tr>
<tr>
<td>mBART†</td>
<td>34.4/13.0/28.1</td>
<td>32.0/11.1/26.4</td>
<td>33.1/11.0/27.8</td>
<td>38.3/15.4/32.4</td>
<td>-/-/-</td>
<td>-/-/-</td>
<td>-/-/-</td>
</tr>
<tr>
<td>DOC-mT5†</td>
<td>37.7/16.7/31.4</td>
<td>32.4/11.9/27.0</td>
<td>33.6/12.8/28.5</td>
<td>36.8/15.0/31.5</td>
<td>-/-/-</td>
<td>-/-/-</td>
<td>-/-/-</td>
</tr>
<tr>
<td>Baseline Model</td>
<td>42.4/19.7/36.4</td>
<td>38.5/15.8/32.9</td>
<td>34.9/13.4/29.7</td>
<td>36.9/14.3/31.5</td>
<td>-/-/-</td>
<td>-/-/-</td>
<td>-/-/-</td>
</tr>
<tr>
<td>Baseline Model+</td>
<td>42.6/19.6/36.6</td>
<td>38.8/16.1/33.1</td>
<td>34.9/13.3/29.6</td>
<td>37.1/14.3/31.5</td>
<td>-/-/-</td>
<td>-/-/-</td>
<td>-/-/-</td>
</tr>
</tbody>
</table>

TRIP (Ours) | 45.3/22.5/39.0 | 46.0/17.3/34.4 | 36.6/14.6/30.8 | 38.7/15.9/32.7 | 42.8/19.9/36.8 | 38.5/16.0/32.9 | 39.4/16.4/33.3 |

Table 7: Comparison of TRIP to ChatGPT on the task of document-level machine translation on TED Talks in the direction of (X → En). The results are snapshotted in May 2023 and can be subject to change.

<table>
<thead>
<tr>
<th>Model</th>
<th>Fr→En</th>
<th>De→En</th>
<th>Zh→En</th>
<th>Vi→En</th>
<th>Cs→En</th>
<th>Th→En</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChatGPT</td>
<td>40.47</td>
<td>36.76</td>
<td>23.31</td>
<td>28.26</td>
<td>30.29</td>
<td>20.94</td>
<td>30.01</td>
</tr>
<tr>
<td>TRIP</td>
<td>54.13</td>
<td>49.94</td>
<td>28.45</td>
<td>41.19</td>
<td>42.73</td>
<td>34.92</td>
<td>42.00</td>
</tr>
</tbody>
</table>

Clearly surpasses the Baseline Model+. This observation indicates that the trilingual objective with the Grafting mechanism is superior to the conventional bilingual objectives for multilingual pre-training.

Case Study Table 8 in Appendix shows three case studies that TRIP outputs better abstractive cross-lingual summarization. For space reasons, we leave more details in Appendix B.

4 Related Work

4.1 Multilingual Pre-training

Multilingual pre-training has achieved great success. Previous works can be categorized into two streams: monolingual pre-training (Conneau et al., 2020; Liu et al., 2020, 2021) and bilingual pre-training (Huang et al., 2019; Chi et al., 2021; Ouyang et al., 2021; Tang et al., 2021; Chi et al., 2021; Reid and Artetxe, 2022; Lee et al., 2022). Monolingual pre-training uses monolingual corpora in many different languages and perturbs the inputs with sentence permutation (Liu et al., 2020) and span corruption (Xue et al., 2021) and requires the model to reconstruct the original input. Reid and Artetxe (2022) also proposes dictionary denoising on monolingual data. For bilingual pre-training, Tang et al. (2021) uses clean sentence-level bilingual translation pairs on pre-trained models to improve MT. Chi et al. (2021) extends mT5 with objectives such as translation span corruption. DOCmT5 (Lee et al., 2022) creates synthetic bilingual translation pairs and uses sentence permutation for a document-level multilingual pre-training.

4.2 Document-level Cross-lingual Tasks

Document-level MT and cross-lingual abstractive summarization are the two document-level cross-lingual tasks that we investigate in this paper.

Document-level MT (Miculicich et al., 2018; Maruf et al., 2019, 2021; Lu et al., 2022) is a challenging translation task, possibly due to the long input problem (Pouget-Abadie et al., 2014; Koehn and Knowles, 2017) when directly modelling the long document and the necessity in understanding contexts (Voita et al., 2018, 2019). Therefore, many works focus on using sentence-level models with a smaller contextual window to simulate document-level MT (Zheng et al., 2020; Chen et al., 2020). This paper follows the challenging setting (Bao et al., 2021; Lee et al., 2022) that directly optimizes a document-level model with a longer context window that provides a richer source of context, which is also a double-edged sword that could be harder due to the long input problem.

Abstractive summarization is a generation task that requires an understanding of texts (Chopra et al., 2016; Fan et al., 2018). We focus on a cross-lingual setting where source and target are written in different languages (Ladhak et al., 2020).

5 Conclusions

We present a novel sequence-to-sequence multilingual document-level pre-training methodology called TRIP, which is the first in our field to propose a trilingual objective for multilingual pre-training through Grafting. We also propose a novel method to construct high-quality trilingual document pairs. Experimental results indicate that TRIP achieves competitive SOTA scores on both multilingual document-level machine translation and cross-lingual abstractive summarization. Future works could improve TRIP to include polygonal parallel...
translation pairs in multilingual pre-training. We plan to release the model checkpoints and a manually annotated benchmark created using our created document-level corpus MTDD to facilitate future research on multilingual document-level MT.

Limitations

**TRIP** TRIP leverages high-quality document-level trilingual translation pairs for pre-training on multilingual models. It is usually harder to collect high-quality trilingual data than to collect monolingual data written in different languages used by conventional methods. While we can possibly relax the quality bar for the data, additional experiments should be done to verify this view.

**MTDD** We create MTDD, a corpus that is composed of trilingual document pairs. It could be further extended to include polygonal parallel document pairs to provide a stronger signal for multilingual pre-training. We leave this to future works.

Large Language Models Large language models (LLMs) such as ChatGPT have shown good translation abilities (Lu et al., 2023), while they still lag behind supervised systems (Jiao et al., 2023; Zhu et al., 2023). We conduct a limited comparison to them, as they are much larger in their number of parameters than the systems described in this work.

Ethics Statement

We honour and support the EMNLP Code of Ethics. The datasets used in this work are well-known and widely used, and the dataset pre-processing does not use any external textual resource. We also curate a corpus for pre-training language models. Although we have made our best efforts in reducing potentially offensive and toxic data, the models are subject to generating offensive context. But the issues mentioned above are widely known to exist for these models commonly. Any content generated does not reflect the view of the authors.

References


Case 1

**Source**
Ayakkabılarnı (ve bağcıklarınla tabanlıklarını) kuruması için orta derecede ıskal alan bir yere koy. Sıcak bir yere (örneğin, radyatörun yanına) ya da doğrudan güneş şarına koyma çünkü çünkü bu, ayakkabılarnın kurutucuya koymak tavsiye edilmez çünkü kurutucu, ayakkabı tabanlarını yumutabilir.

**Source (Google-translated)**
Put your shoes (and your laces and insoles) in a moderately light place to dry. Do not place it in a hot place (for example, near a radiator) or in direct sunlight as this may damage the shoes. Putting your shoes in the dryer is not recommended because the dryer can warp the soles of your shoes.

**Reference**
Air-dry your shoes.

**Baseline Model (Document-level)**
Let your shoes (and the laces) air dry.

**TRIP**
Let your shoes (and the laces) air dry.

---

**Case 2**

**Source**

**Source (Google-translated)**
To do this, tap the white speech bubble on a green background. It should be on one of their Home Screens. Tap It’s in the upper-left corner of the Messages screen. If you have an open chat, tap the < button in the upper left corner of the screen to return to the Messages menu. Tap It’s in the lower right corner of your screen. The selected messages are deleted.

**Reference**
Open your iPhone’s messages. Tap Edit. Select each conversation you wish to delete. Tap Delete.

**Baseline Model (Document-level)**
Open your iPhone’s Settings. Tap Messages. Tap Delete Messages.

**TRIP**
Open Messages. Tap the Messages tab. Tap Delete. Tap Delete to confirm.

---

**Case 3**

**Source**
Bazıları için geçmiş yaşamlar gidilecek bir yer değil, seni sen yapan şeyin bir komûndur. İnsanlığın tarihi boyunca birçok kültür reenkarnasyonu inancının merkezine koymuşuktur. İslam ve Hristiyanlık reenkarnasyona inanmamıza neden olmasa da, Hinduistler, bazı Meseviler ve bazı Budistler buna inanır. En iyisi kendini bir dîne tümüyle adamlakansa (çünkî dîner çok kısıtlayıcı olabilir) kendi yolunu keşfetmen. Kendi manevi doûarlarnı kendin bul.

**Source (Google-translated)**
For some, past lives are not a place to go but part of what makes you who you are. Throughout the history of humanity, many cultures have put reincarnation at the center of their beliefs. Although Islam and Christianity do not believe in reincarnation, Hindus, some Jews, and some Buddhists do. It’s best to explore your own path rather than devote yourself entirely to a religion (because religions can be too restrictive). Find your own spiritual truth.

**Reference**
Become spiritual.

**Baseline Model (Document-level)**
Understand that some people believe in reincarnation. Find your own way.

**TRIP**
Explore your own spiritual journey.

Table 8: Three case studies from Wikilingua (Tr → En) demonstrate that TRIP outputs better summarization.
Figure 4: Results on TED Talks in (X → X) with our TRIP checkpoint pre-trained in (X → En) directions only. The scores are written in TRIP as the former and the Baseline Model as the latter. Rows represent the source languages and columns represent the target languages. We highlight in aqua when TRIP wins (darker one when printed in B&W) and in hot pink (lighter one when printed in B&W) when the Baseline Model wins.

### A Unseen (X → X) Language Pairs on MT

Figure 4 reports the performance on TED Talks in the direction of (X → X) with our TRIP check-point pre-trained in (X → En) directions with our corpus. The row represents the translation source language and the column represents the translation target language. TRIP clearly improves most of these translation directions which are unseen during pre-training. This indicates that fact the TRIP can generalize the cross-lingual understanding ability to unseen language pairs. This aligns with the fact reported in Lee et al. (2022).

### B Case Study on Summarization

Table 8 shows that TRIP outputs better summarization in (i) precisely capturing the context in Case 1, (ii) outputting consistent nouns, i.e., ‘messages’ instead of ‘settings’ in Case 2 and (iii) producing concise and accurate summarization in Case 3. This highlights that TRIP captures better cross-lingual understanding than the baseline system, which effectively mitigates cross-lingual ambiguities.

### C Number of Model Parameters

Table 9 presents the number of model parameters for the pre-trained models used in our experiments. ∗: these models all use the model architecture of mT5-Large, and we report the number of model parameters taken from the original paper of mT5 reported by Xue et al. (2021).

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2M-100</td>
<td>418M</td>
</tr>
<tr>
<td>mBART</td>
<td>611M</td>
</tr>
<tr>
<td>MARGE</td>
<td>963M</td>
</tr>
<tr>
<td>mT5</td>
<td>1.23B*</td>
</tr>
<tr>
<td>DOCmT5</td>
<td>1.23B*</td>
</tr>
<tr>
<td>ByT5-Small</td>
<td>300M</td>
</tr>
<tr>
<td>ByT5-Base</td>
<td>582M</td>
</tr>
<tr>
<td>ByT5-Large</td>
<td>1.23B*</td>
</tr>
<tr>
<td>mT5-XL</td>
<td>3.74B</td>
</tr>
<tr>
<td>Baseline Model</td>
<td>862M</td>
</tr>
<tr>
<td>Baseline Model+</td>
<td>862M</td>
</tr>
<tr>
<td>TRIP (Ours)</td>
<td>862M</td>
</tr>
</tbody>
</table>

Table 9: Comparison in the number of parameters for the pre-trained models used in our experiments. ∗: these models all use the model architecture of mT5-Large, and we report the number of model parameters taken from the original paper of mT5 reported by Xue et al. (2021).

Table 10: A language list in ISO code for the top 12 language directions for the bilingual high-quality pre-training data to illustrate the scale of size.

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Size/GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Es</td>
<td>En</td>
<td>3.22</td>
</tr>
<tr>
<td>Es</td>
<td>Ca</td>
<td>2.07</td>
</tr>
<tr>
<td>Fr</td>
<td>Es</td>
<td>1.47</td>
</tr>
<tr>
<td>En</td>
<td>De</td>
<td>1.25</td>
</tr>
<tr>
<td>Ca</td>
<td>Es</td>
<td>1.12</td>
</tr>
<tr>
<td>Ru</td>
<td>Uk</td>
<td>0.87</td>
</tr>
<tr>
<td>Source</td>
<td>Target</td>
<td>Size/GB</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>Pt</td>
<td>Es</td>
<td>2.71</td>
</tr>
<tr>
<td>Uk</td>
<td>Ru</td>
<td>1.60</td>
</tr>
<tr>
<td>Es</td>
<td>Pt</td>
<td>1.47</td>
</tr>
<tr>
<td>Pt</td>
<td>En</td>
<td>1.14</td>
</tr>
<tr>
<td>Fr</td>
<td>En</td>
<td>1.03</td>
</tr>
<tr>
<td>Pt</td>
<td>Fr</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 10: A language list in ISO code for the top 12 language directions for the bilingual high-quality pre-training data to illustrate the scale of size.

Table 9 presents the number of model parameters for the pre-trained models used in our experiments.

For the scores of ByT5 presented in Table 6, we report the maximum scores for each direction among ByT5-Base, ByT5-Small, and ByT5-Large. This is due to space reasons. See https://gem-benchmark.com/results for the tailored scores.

### D MTDD Corpus Scale

Table 10 presents the top-12 English-centric bilingual data statistics to illustrate the scale of MTDD. The total size of the data is about 40/80 GB respectively for the bilingual and the trilingual data applied with Grafting.