A Lifelong Multilingual Multi-granularity Semantic Alignment Approach via Maximum Co-occurrence Probability

Xin Liu[1], **Hongwei Sun**[2], **Shaojie Dai**[3,1], **Bo Lv**[3,1],
Youcheng Pan[1], **Hui Wang**[1,*] **and Yue Yu**[1,*]
[1]Peng Cheng Laboratory, [2] Mudanjiang Normal University,
[3]University of Chinese Academy of Sciences
{hit.liuxin, panyoucheng4}@gmail.com, emailshw0319@yeah.net,
{daishaojie22,lvbo19}@mails.ucas.ac.cn, {wangh06,yuy}@pcl.ac.cn

Abstract

Cross-lingual pre-training methods mask and predict tokens in a multilingual text to generalize diverse multilingual information. However, due to the lack of sufficient aligned multilingual resources in the pre-training process, these methods may not fully explore the multilingual correlation of masked tokens, resulting in the limitation of multilingual information interaction. In this paper, we propose a lifelong multilingual multi-granularity semantic alignment approach, which continuously extracts massive aligned linguistic units from noisy data via a maximum co-occurrence probability algorithm. Then, the approach releases a version of the multilingual multi-granularity semantic alignment resource, supporting seven languages, namely English, Czech, German, Russian, Romanian, Hindi and Turkish. Finally, we propose how to use this resource to improve the translation performance on WMT14∼18 benchmarks in twelve directions. Experimental results show an average of 0.3∼1.1 BLEU improvements in all translation benchmarks. The analysis and discussion also demonstrate the superiority and potential of the proposed approach.

Keywords: lifelong, multilingual, multi-granularity, alignment, maximum co-occurrence probability

1. Introduction

The alignment between languages is the key message for machine translation, and it encourages the models to learn the correlation of different languages and to achieve multilingual interaction (Mao et al., 2022; Tang et al., 2022; Adjali et al., 2022). Typically, the models could acquire the multilingual alignment information through the multilingual texts during the cross-lingual pre-training (Wei et al., 2021; Chi et al., 2021; Batheja and Bhattacharyya, 2022) or through the parallel corpora during the fine-tuning (Fernandez and Adlaon, 2022).

However, although most models could learn accurate multilingual alignment information through the parallel corpora, they are usually limited by the insufficient scale of the parallel corpora and thus cannot learn sufficiently (Wang and Li, 2021; Chimoto and Bassett, 2022). In contrast, the cross-lingual pre-training methods based on various training strategies and large-scale multilingual texts alleviate this problem to some extent and inject the alignment information into the models (Yang et al., 2020a; Luo et al., 2021). So these models could present the multilingual correlation of different languages more or less. Benefiting from this alignment information in the pre-training process, these methods have shown promising performances in multilingual machine translation (Lin et al., 2020; Pan et al., 2021). However, constrained by not using explicit multilingual alignment resources during

English linguistic unit	Alignment	German linguistic unit	
over-dimensioning	⟷	*überdimensionie*	**word**
my team-mate	⟷	*meinem teamkollegen*	**phrase**
in current discussions	⟷	*in den aktuellen diskussionen*	**phrase**
even in the best	⟷	*selbst in den besten*	**segment**
eliab his oldest brother heard	⟷	*eliabsein ältester bruderhörte ihn*	**segment**
full of mercy and good fruits	⟷	*voll barmherzigkeit und guter früchteunparteiisch*	**short sentence**
none of the windows system tools like registry editor	⟷	*keiner der windows-system-tools wie den registrierungs-editor*	**short sentence**

Figure 1: Illustration of English-German multi-granularity alignment linguistic units in the resource built by this approach.

pre-training, these pre-training methods may not be able to explore the multilingual correlation of different tokens in multilingual texts as comprehensively and accurately as in parallel corpora (Yang et al., 2021). Thus, this undoubtedly presents an opportunity and raises an urgent need for a high-quality multilingual alignment resource for the further advances of the cross-lingual pre-training methods.

To address the need, this paper proposes a lifelong multilingual multi-granularity semantic alignment approach via maximum co-occurrence probability in noisy parallel data and uses it to build a semantic alignment resource. A linguistic unit is a sequence of consecutive tokens in a sentence, So it may be a word, phrase, segment, or short sentence. The approach collects a group of noisy pairs that contain the same linguistic unit in one language

* Corresponding author

LREC-COLING 2024, pages 684–694
20-25 May, 2024.

and computes the co-occurrence probability of one candidate linguistic unit in the other languages. The co-occurrence probability is the probability that one linguistic unit appears in all sentences, so one candidate linguistic unit will have a higher probability if it occurs in most sentences. Figure 1 presents the English-German multi-granularity alignment linguistic units in the resource built by this approach. The approach can satisfy the above need from three aspects, namely the scale and quality, the linguistic diversity, and the lifelong property. Taking the large-scale noisy data as the data source and constraining the aligned unit with maximum co-occurrence probability ensure the scale and quality. The multi-granularity and multilingualism reflect the linguistic diversity. Through a lifelong stream of noisy data, the approach can continuously expand the resource with new languages or aligned units. Additionally, the resource can also be used in multilingual machine translation scenarios to boost translation performance through the combination of pre-training and fine-tuning strategies. In summary, we highlight our contributions as follows:

- This paper proposes a lifelong multilingual multi-granularity semantic alignment approach that only relies on the co-occurrence constraints in the multilingual noisy data, and can identify massive semantically aligned linguistic units at various granularity through the maximum occurrence probability continuously and unsupervised.

- The proposed approach releases a version of the **l**ifelong **m**ultilingual **m**ulti-**g**ranularity **s**emantic **a**lignment **r**esource (called $\text{LM}_\text{g}^2\text{SAR}$). In this version, $\text{LM}_\text{g}^2\text{SAR}$ supports the multilingual alignment between seven languages, namely English (en), Czech (cs), German (de), Russian (ru), Romanian (ro), Hindi (hi) and Turkish (tr). Meanwhile, it also supports the continuous expansion in scale, language coverage, and granularity. The resource will be publicly available[1].

- This paper conducts exhaustive experiments on the aligner comparisons and the bi-direction translation tasks between English and the above six languages. Compared to the other aligners, the approach shows higher alignment accuracy. The models using $\text{LM}_\text{g}^2\text{SAR}$ have shown significant improvements in almost all translation directions. In addition, we perform objective analysis and discussion as strong evidence of the value and significance of this work.

[1]https://github.com/Gdls/MCoPSA

2. Related Works

The mainstream pre-training methods rely on different mechanisms, techniques, or tools (Wu et al., 2022; Dou and Neubig, 2021) to learn multilingual alignment information. For example, MARGE (Lewis et al., 2020) learned with an unsupervised multilingual multi-document paraphrasing objective. Luo et al. (2021) plugged a cross-attention module into the Transformer encoder to build language interdependence. mRASP (Lin et al., 2020) introduced a random aligned substitution technique into the pre-training to bridge the semantic space. Yang et al. (2020b) performed lexicon induction with unsupervised word embedding mapping technique to learn the cross-lingual alignment information from monolingual corpus (Bajaj et al., 2022). Tang et al. (2022) specifically highlighted the importance of word embedding alignment by guiding similar words in different languages. Yang et al. (2020a) performed the word alignment with the GIZA++ (Casacuberta and Vidal, 2007) toolkit to code-switch the sentences of different languages to capture the cross-lingual context of words and phrases. Yang et al. (2021) proposed to use FastAlign (Dyer et al., 2013) as the prior knowledge to guide cross-lingual word prediction. These mechanisms, techniques, or tools have boosted the capabilities of these methods on generalizing alignment information, but due to the absence of accurate and sufficient alignment resources, there is still a lot of room for improving their capabilities.

Normally, the common multilingual alignment resources are the parallel corpora, which come from the public releases (Ziemski et al., 2016), web mining (Tiedemann and Nygaard, 2004), or competitions. These resources are usually aligned at the sentence level and can be used to train the translation models directly. The other resources mainly focus on the word or phrase level (Imani et al., 2022), e.g., the multilingual paraphrase database (Ganitkevitch and Callison-Burch, 2014), the multilingual lexical database (Giguet and Luquet, 2006), the multilingual multi-word expression corpora (Han et al., 2020), automatic similarity-based dataset (Yousef et al., 2022) and unpublished synonym dictionary (Pan et al., 2021). Although these resources could provide multilingual alignment information, the scale or linguistic diversity may not meet the need for the pre-training methods.

In view of the above, this paper proposed the lifelong multilingual multi-granularity semantic alignment approach to build a semantic alignment resource. Compared to the previous methods and resources, the approach considers the scale, diversity, and other linguistic properties. Meanwhile, the

Algorithm 1 The maximum co-occurrence probability based semantic alignment algorithm.

procedure *MCoPSA*(u_l, $\mathfrak{D}_{\mathfrak{N}}$)

  Assert $u_l \in X$ and $u_l^X = u_l$;

  Initialize $G(u_l^X) = (p_0, ..., p_n)$ from $\mathfrak{D}_{\mathfrak{N}}$;

  Initialize lists L = [], uL = [], dict D = {}

  Select $t_0^{\mathcal{Y}}, t_1^{\mathcal{Y}}$ from $G(u_l^X)$;

  $G(u_l^X)$ = $G(u_l^X)$ - p_0 - p_1;

  L.append($t_0^{\mathcal{Y}}$, $t_1^{\mathcal{Y}}$);

  uL.extend(*CSFunc*($t_0^{\mathcal{Y}}$, $t_1^{\mathcal{Y}}$));

  Update *cnt*(uL);

  for ${u_l}_i{}^Y \in uL$ **do**

    $D[{u_l}_i^Y]$ = *cnt*$[{u_l}_i^Y]$/*len*(L);

  end for

  while $G(u_l^X)$ is not Ø **do**

    Select $t_i^{\mathcal{Y}}$ from $G(u_l^X)$;

    $G(u_l^X)$ = $G(u_l^X)$ - p_i;

    for $t_j^{\mathcal{Y}}$ in L **do**

      uL.extend(*CSFunc*($t_i^{\mathcal{Y}}$, $t_j^{\mathcal{Y}}$));

      Update *cnt*(uL);

    end for

    L.gappend($t_j^{\mathcal{Y}}$);

    for ${u_l}_k{}^Y \in uL$ **do**

      $D[{u_l}_k^Y]$ = *cnt*$[{u_l}_k^Y]$/*len*(L);

    end for

  end while

  ${u_l}^Y$ = *maxProb*(D,ϱ);

  Return (u_l^X, ${u_l}^Y$);

end procedure

resource provides more specific and sufficient semantic alignment information than those alignment techniques in the pre-training methods.

3. Methods

In this section, we first introduce the **m**aximum **co**-**o**ccurrence **p**robability based **s**emantic **a**lignment algorithm (MCoPSA), which is the core of the proposed approach. Next, we present the statistics on the first version of the semantic alignment resources (LM^2_gSAR) built by this approach.

3.1. The maximum co-occurrence probability based semantic alignment algorithm

The core idea of the MCoPSA algorithm is as follows: there is a group of translated pairs from noisy data, and each pair consists of sentences in two languages. A linguistic unit of one language exists in all sentences in the group, and the algorithm calculates the co-occurrence probability of each candidate linguistic unit in all sentences in the other language. The co-occurrence probability means the probability of one candidate linguistic unit appearing in all the sentences of the group. Then, the algorithm selects the candidate with the maximum co-occurrence probability as the aligned linguistic unit.

The calculation procedure of MCoPSA is listed in Algorithm 1. MCoPSA takes one linguistic unit in one language and noisy data as input and selects a group of pairs from noisy data that contains the linguistic unit. Initially, MCoPSA selects two sentences in the other language from the group into a list and computes the occurrence probability of one linguistic unit in two sentences. Next, MCoPSA continues to select one sentence and updates the occurrence probability of one linguistic unit with these sentences in the list. Finally, MCoPSA outputs a linguistic unit with the maximum co-occurrence probability.

In Algorithm 1, u_l denotes the linguistic unit. $\mathfrak{D}_{\mathfrak{N}}$ denotes the noisy parallel data, which are translated sentence pairs but the translation is usually inaccurate because two sentences may be partially aligned. $\mathcal{X}$ and $\mathcal{Y}$ denotes two languages, and $p = (s^{\mathcal{X}}, t^{\mathcal{Y}})$ denotes a pair of sentence x from $\mathcal{X}$ and sentence t from $\mathcal{Y}$. A linguistic unit of language X is denoted as u_l^X. $G(u_l) = (p_0, ..., p_i, ...p_n)$ is a group of pairs with $u_l^X \in s_i^{\mathcal{X}}$ and $u_l^Y \in t_i^{\mathcal{Y}}$. The function *CSFunc*($\cdot$) takes two sentences as input and outputs all the linguistic units that appear in two sentences simultaneously. The function *cnt*($\cdot$) takes a list of linguistic units as input and outputs a dictionary to store the occurrence and frequency of each unit in the current step. The function *maxProb*($\cdot$) takes a dictionary D and a penalty factor ϱ as input, where D stores the co-occurrence probability of each unit, and outputs a linguistic unit ${u_l}^Y$ with the maximum co-occurrence probability. The penalty factor ϱ considers the length(l), frequency(f), and similarity(s) of the candidate units. For one linguistic unit ${u_l}_k^Y$, the corresponding ϱ_k value is calculated based on Equation 1. Here, the normalization function(N($\cdot$)) is based on all the candidate units ${u_l}^Y$. With penalty factor ϱ, we update the co-occurrence probability with $D[{u_l}_k^Y]$ = $D[{u_l}_k^Y] \times \varrho_k$ to re-calculate the co-occurrence probability, which may reduce the effect of units such as stop-words.

$$\varrho_k = \frac{s({u_l}^X, {u_l}_k^Y) \times N(f_k^Y, f^Y)}{|l^X - l^Y|} \tag{1}$$

Figure 2 presents an example of English and Romanian pairs to illustrate the processing of Algorithm 1. The linguistic unit u_l is from English, and we have u_l^{EN}= *calculation error*. There are four EN-RO pairs in its group $G(u_l) = (p_0, p_1, p_2, p_3)$. When p_0 and p_1 are selected, the algorithm will output the candidate linguistic units, namely "*erori de calclu*", "*sunt*", "*de*", and their co-occurrence probabilities in the current step. Next, when p_2 comes, the algorithm updates the candidate list and their

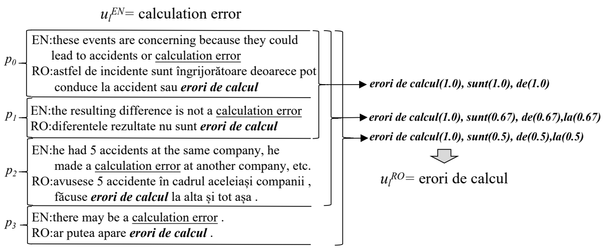

Figure 2: An example to illustrate the processing of Algorithm 1. The underlined part in the English sentences is the input linguistic unit, and the bold and italic part in the Romanian sentences is the output(aligned linguistic unit). The float in brackets is the co-occurrence probability of each candidate unit in the current step.

co-occurrence probabilities. In this step, a new unit "*la*" is appended. The algorithm repeats the calculation with the coming of new pairs. Finally, when receiving the last one p_3, the algorithm outputs the final candidate linguistic units with the ranking of co-occurrence probability and takes the one with the maximum co-occurrence probability as the aligned linguistic unit, namely u_l^{RO}= *erori de calclu*. So, in this example, "*calculation error*" and "*erori de calclu*" is the semantic alignment between English and Romanian.

3.2. The statistics on LM^2_gSAR

Based on the MCoPSA algorithm, this paper released the first version of the **l**ifelong **m**ultilingual **m**ulti-**g**ranularity **s**emantic **a**lignment **r**esource (LM^2_gSAR). This section will detail the statistics on LM^2_gSAR from three aspects: the languages, the scale, and the linguistic diversity.

In this release, LM^2_gSAR supports seven languages, namely English, Czech, German, Russian, Romanian, Hindi and Turkish. Meanwhile, it is built on the noisy bilingual data from published CCMatrix v1 (Schwenk et al., 2021) between English and the other six languages. When building it with the en-XX bilingual data, the MCoPSA algorithm takes the English linguistic units as input and outputs its alignment in language XX, where XX is one of the other six languages.

From the scale, Table 1 lists the statistics on the scale of the bilingual data used in the MCoPSA algorithm for building LM^2_gSAR, including the volume of the bilingual data and its percentage in CCMatrix v1. Table 2 shows the statistics on the scale of the aligned linguistic units in LM^2_gSAR between seven languages.

Since the original scale of the bilingual data in CCMatrix v1 varies greatly, we randomly sampled a certain percentage for each en-XX bilingual data.

Languages	Volume	PinND (%)
en-de	21.5M	8.7
en-ru	20.6M	14.7
en-cs	15.6M	27.7
en-ro	15.1M	27.2
en-tr	14.2M	29.2
en-hi	5.5M	36.4

Table 1: The scale of the bilingual data from published CCMatrix v1 (Schwenk et al., 2021) used in the MCoPSA algorithm for building LM^2_gSAR. The magnitude "M" in the second column is million. The third column "PinND" is the final percentage of CCMatrix data used in this work.

	en	cs	de	ru	ro	tr
cs	3.30M	-				
de	2.53M	0.33M	-			
ru	3.04M	0.19M	0.13M	-		
ro	3.11M	0.53M	0.30M	0.17M	-	
tr	1.95M	0.36M	0.21M	0.11M	0.34M	-
hi	1.01M	0.19M	0.13M	0.33M	0.20M	0.18M

Table 2: The statistics on the scale of the aligned linguistic units in LM^2_gSAR between seven languages.

It is 10% for en-de, 20% for en-ru, 30% for en-cs, en-ro, and en-tr, and 40% for en-hi. Table 1 presents the final scale and percentage after the post-processing, e.g., deduplication, discarding, merging, etc. In Table 2, the first column indicates that the scale of the aligned linguistic units between English and each language is in millions, which is an encouraging scale. In the construction, the MCoPSA algorithm takes the English linguistic unit as input and outputs its aligned unit. So we take the English linguistic unit as a bridge and can easily get the aligned linguistic unit between any two languages except English. The remaining columns in Table 2 present the scale of the aligned linguistic units between the other six languages. Obviously, the scale between any two of these languages is more than one hundred thousand, and it is also a promising scale.

From the linguistic diversity, Table 3 shows the statistics on the granularity distribution of the aligned linguistic units in LM^2_gSAR on the en-XX alignment. In each row, Table 3 presents the percentage of this granularity in all alignments. For accuracy, we only report the statistics on the en-XX alignment since the granularity in the English linguistic unit is known during the construction. In Table 3, most of the aligned units belong to the phrase or segment level, which is because the phrases and segments are widely used through the word combination in language expression. But the words are usually finite, so the percentage is stable. As for the short sentence-level granularity, it

Aligment	Multi-granurality(%)				
	w-1	p-2	p-3	s-4	ss-5+
en-cs	3	21	44	31	1
en-de	4	23	42	30	1
en-ru	4	20	44	31	1
en-ro	3	19	42	34	2
en-tr	3	30	44	21	2
en-hi	4	28	43	24	1
Avg	3.5	23.5	43.2	28.5	1.3

Table 3: The statistics on the granularity distribution(%) of the aligned linguistic units in LM_g^2SAR based on the en-XX alignment. 'w-1': word-level by unigram, 'p-2' and 'p-3': phrase-level by bi-gram or tri-gram, 's-4': segment-level by four-gram, and 'ss-5+': short sentence-level.

is really rare. Table 2 also indicates the multilingualism in LM_g^2SAR. Though LM_g^2SAR only relies on the en-XX bilingual data, it can still find the alignment between any two languages.

4. Experiments and Results

4.1. Experimental datasets and metrics

In this work, we validate the proposed approach through two experiments, namely the aligner comparison experiment and the machine translation experiment. In the aligner comparison experiment, we select the well-known statistical aligners (GIZA++, Fast-Align) and neural aligner (Awesome-align) to compare with our proposed methods on the same test set. In this process, we first randomly collected a group of en-XX corpora that are not used in Section 3.2 and selected the top 500 language units in each en-XX corpus based on the term frequency and inverse document frequency. Second, we recruited some language experts with English and XX backgrounds, and for each pair, they manually annotated the golden alignment of the English language unit in XX sentences, which serves as the evaluation test set. Next, we applied the proposed MCoPSA algorithm, GIZA++, Fast-Align, and Awesome-align to compute the alignments of the top 500 language units. Finally, we performed the evaluation using the metrics of the alignment error rates (AER).

In the machine translation experiment, we apply the resource that the approach built to the machine translation tasks. We select the WMT datasets including twelve translation directions as the evaluation benchmarks, namely en-de (4.5M), en-ru (1.1M), and en-hi (32K) in WMT14, en-ro (0.6M) in WMT16, en-tr (0.2M) in WMT17, and en-cs (11M) in WMT18. Based on the scale of the training data in each dataset, we follow the division in Tang et al. (2021) and Lin et al. (2020) to divide the datasets into four categories: extremely low resource (<100K), low resource(>100k and <1M), medium resource (>1M and <10M), and high resource (>10M). These datasets are publicly available, and anyone can easily access the same training, validation, and test sets for reproduction or comparison. For all evaluation benchmarks, we take the BLEU score as the metrics and it is computed with the official sacreBLEU (Post, 2018) with default tokenization.

4.2. Baseline and comparison methods

In the aligner comparison experiment, the statistical aligners for comparison are GIZA++ (Casacuberta and Vidal, 2007) and Fast-Align (Dyer et al., 2013), and the neural aligner is Awesome-align (Batheja and Bhattacharyya, 2022). In these experiments, we followed the default setting of each method.

- GIZA++ is an extension of the program GIZA (part of the SMT toolkit EGYPT). We used the version[2] released by Och and Ney (2003).
- Fast-align is a simple, fast, unsupervised word aligner. We used the version released by Dyer et al. (2013) from the Github page[3].
- Awesome-align is a tool that can extract word alignments from multilingual BERT. We used the version released by Dou and Neubig (2021) from the Github page[4].

In machine translation experiments, three famous open-source multilingual models are selected as the baseline, namely mBART (Liu et al., 2020), M2M100 (Fan et al., 2021), and mT5 (Xue et al., 2021). In these experiments, all the codes and checkpoints for these models are from the public Hugging Face hub. One reason is that these models are all multilingual models and cover enough languages to evaluate our LM_g^2SAR as it grows continuously. Another reason is that these models come from different types of pre-training tasks, which can demonstrate the quality of LM_g^2SAR from different aspects.

- mBART is one of the first methods for pre-training a complete sequence-to-sequence model by denoising full texts in multiple languages. The initial checkpoint of mBART model we used in this work is mbart-large-cc25[5].
- M2M-100 is a Many-to-Many multilingual translation model that can translate directly between

[2] http://www2.statmt.org/moses/giza/GIZA++.html
[3] https://github.com/clab/fast_align
[4] https://github.com/neulab/awesome-align
[5] https://huggingface.co/facebook/mbart-large-cc25

any pair of 100 languages. The initial checkpoint of M2M-100 model we used in this work is m2m100_418M[6].

- mT5 is pre-trained on a new Common Crawl-based dataset covering 101 languages and has shown SOTA performance on many multilingual benchmarks. The initial checkpoint of mT5 model we used in this work is mt5-base[7].

4.3. Experimental setup

Some experimental settings or hyperparameters for the machine translation task in this work are listed below: all experiments with pre-training or fine-tuning are based on three baseline models. In these experiments, the tokenizer in each model is the default one, namely MBartTokenizer, M2M100Tokenizer, and T5Tokenizer. The training batch size is 4∼16 in all experiments. The max sequence length is 1024. The beam size for decoding is 5. Checkpoints are saved every 1000 steps in high and medium resource benchmarks, and every 100-500 steps for low and extremely-low resource benchmarks. We use the AdamW (Loshchilov and Hutter, 2017) optimizer with an initial learning rate of 5e-5. Early stopping is used when the training loss converges during the pre-training and fine-tuning process, and we select the hyperparameters based on the validation set.

4.4. Training strategy in machine translation experiment

We adopt a new strategy to (pre-)train a baseline model with LM^2_gSAR, called LM^2_gSAR-based pre-training and fine-tuning.

In this training strategy, we first apply the alignment substitution technique (AST) with LM^2_gSAR to prepare the pre-training corpus. In this part, the monolingual sentences used to construct the alignment substitution with LM^2_gSAR come from the corresponding bilingual training data. Therefore, during the LM^2_gSAR pre-training, no additional monolingual or bilingual data is introduced; the same data source is utilized in the fine-tuning phase. A baseline model is pre-trained with the corpus. Next, the pre-trained model is fine-tuned with the training data. Finally, the trained model is evaluated on the test set. The AST technique is similar to the previous works (Lin et al., 2020; Yang et al., 2020a) that given a monolingual sentence S, AST substitutes the linguistic units in S with the corresponding alignments in LM^2_gSAR to produce a mixed language sentence S_x. During the pre-training, the input of the model is S_x, and the output is its original sentence S. For example, in pair p_3 of Figure 2, for en→ro direction, the input S_x is "there may be a ***erori de calclu***" and the output is "there may be a calculation error". A similar operation goes in the other direction. Since a sentence may contain multiple linguistic units that can be substituted, we take a random combination from the smallest granularity to the biggest each time. Table 4 lists the statistics on the pre-training corpus.

[6]https://huggingface.co/facebook/m2m100_418M
[7]https://huggingface.co/google/mt5-base

Benchmark	Scale(->/<-)	PinT(%)	AvgLSu	AvgLSe	AvgP(%)
en-cs	1.2M/1.2M	10.4	2.3	9.7	23.8
en-de	3.0M/3.0M	66.5	5.6	23.1	24.1
en-ru	0.2M/0.2M	11.8	1.9	15.9	11.9
en-ro	0.6M/0.6M	92.9	5.3	23.1	22.3
en-tr	0.1M/0.1M	70.1	2.9	21.4	13.5
en-hi	5.0K/7.6K	19.2	1.1	2.1	51.2

Table 4: The statistics on the pre-training corpus for each benchmark. "Scale(->/<-)" is the pre-training corpus scale for both translation directions. "PinT" indicates the percentage of the pre-training corpus scale in each benchmark to its training data scale. "AvgLSu" is the average length of the substitution, and "AvgLSe" is the average length of the original sentence. "AvgP" is AvgLSu/AvgLSe.

	GIZA++	Fa-Align	Aw-Align	MCoPSA
en-cs	48.4	42.1	30.8	**19.8**
en-de	61.2	58.7	29.0	**17.2**
en-hi	67.2	66.0	30.3	**16.2**
en-ro	53.4	50.2	24.4	**16.4**
en-ru	50.1	46.1	21.8	**15.2**
en-tr	63.2	72.6	30.3	**12.4**
Avgs	57.3	55.9	24.4	**16.2**

Table 5: The AER scores on each en-XX test set of the MCoPSA, GIZA++, FastAlign (Fa-Align), and Awesome-Align (Aw-Align).

	word	phrase	segment	Avgs
GIZA++	52.2	48.4	58.1	52.9
Fa-Align	54.5	47.4	56.1	52.7
Aw-Align	**14.1**	24.1	20.2	19.5
MCoPSA	25.4	**13.1**	**13.4**	**17.3**

Table 6: The AER score of each method at word, phrase, and segment granularity of the linguistic unit on the en-ru test set. The "Avgs" column is the average of the three granularities and is therefore slightly different from that in Table 5.

4.5. Experimental results and analysis

Table 5 lists the alignment error rates(AER) of the MCoPSA and three aligners, and the last line gives the average score. Obviously, MCoPSA presents the best performances on each set. In particular, though MCoPSA is an unsupervised algorithm as GIZA++ and Fast-Align, it can still surpass the supervised aligner Awesome-Align. The main reason may be that the MCoPSA algorithm can fully ex-

Benchmark	mBART		M2M100		mT5		AvgΔ
	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$	
en-cs^{H}	16.6	28.9	17.2	30.1	18.1	27.9	
	17.5$^{\uparrow}$	**30.5(+1.6)**$^{\uparrow}$	17.7$^{\uparrow}$	29.4	**18.5(+0.4)**$^{\uparrow}$	29.7$^{\uparrow}$	+0.8
en-de^{M}	26.5	32.5	26.9	31.6	24.4	28.9	
	27.3(+0.8)$^{\uparrow}$	**33.0(+0.5)**$^{\uparrow}$	26.9	32.2$^{\uparrow}$	24.6$^{\uparrow}$	29.6$^{\uparrow}$	+0.5
en-ru^{M}	33.0	32.2	33.0	32.4	26.8	26.8	
	34.1(+1.1)$^{\uparrow}$	**32.6(+0.4)**$^{\uparrow}$	33.1$^{\uparrow}$	32.4	26.9$^{\uparrow}$	27.0$^{\uparrow}$	+0.4
en-ro^{L}	25.6	36.2	26.4	36.7	22.8	32.9	
	26.8(+1.2)$^{\uparrow}$	**37.0(+0.8)**$^{\uparrow}$	26.3	36.6	22.9$^{\uparrow}$	33.2$^{\uparrow}$	+0.4
en-tr^{L}	19.1	22.7	19.8	22.7	13.1	17.2	
	20.0(+0.9)$^{\uparrow}$	23.1$^{\uparrow}$	19.8	**23.3(+0.6)**$^{\uparrow}$	13.3$^{\uparrow}$	16.8	+0.3
en-hi^{El}	0.9	1.1	10.3	13.6	0.2	0.1	
	2.3$^{\uparrow}$	2.1$^{\uparrow}$	**13.1(+2.8)**$^{\uparrow}$	**14.5(+0.9)**$^{\uparrow}$	0.3$^{\uparrow}$	0.3$^{\uparrow}$	+1.1

Table 7: The BLEU scores of the baseline models under different training strategies on the test sets of each WMT benchmark.The blocks in "Benchmark" corresponds to the "high(H)/medium(M)/low(L)/extremely low(El)" resource according to their official training data volume. The bold is the best BLEU score in this direction. Here, "__": with fine-tuning only, " ":with LM^2_gSAR pre-training and fine-tuning. "$\rightarrow$/$\leftarrow$" is the translation direction. "AvgΔ": average of the difference of all models between the upper line and below line, and "+" and "$\uparrow$" mean improvement.

plore the correlation of language units between the parallel data based on the co-occurrence constraint. Besides, the performance difference between MCoPSA and the others also indicates that the alignments by MCoPSA are of better quality and more promising for the pre-training stage in real-world scenarios. To further investigate their ability, we report the AER score of each method at word, phrase, and segment granularity of the linguistic unit on the en-ru test set in Table 6, which helps to indicate the ability of each method on different granularity. The GIZA++ and Fast-Align seem to have the similar performance on each granularity, while the Awesome-Align performs best on word-level. However, the proposed McoPSA show a much better performance on phrase and segment-level, and this may be the main reason why the proposed McoPSA achieves the best results in Table 5 and 6.

Table 7 lists the BLEU scores of the baseline models under different training strategies on the test sets of each WMT benchmark. For each test set, we provide two lines of results from three baseline models. The upper one is the results of fine-tuning the baseline models with the official training data, and the below one is the results of the LM^2_gSAR-based pre-training and fine-tuning. In table 7, we highlight the lines of fine-tuning with underline and the lines of LM^2_gSAR-based pre-training and fine-tuning with wave line, respectively. For clarity, we denote a model with only fine-tuning as model^{f} and that with LM^2_gSAR-based pre-training and fine-tuning as model$^{p\cdot f}$.

From the results, we have the following observation: 1) Each model with the LM^2_gSAR pre-training and fine-tuning shows better performances than the only fine-tuning one on all benchmarks, with an average of 0.3$\sim$1.1 BLEU improvement in the six benchmarks (See "AvgΔ" column). This is strong evidence that LM^2_gSAR contributes to the translation task. 2) Almost all models$^{p\cdot f}$ show some improvement over models^{f} that were only fine-tuned (see results with "$\uparrow$"). In particular, the improvement of the best results in bold on each benchmark is significant. Even though in some directions, such as M2M100 in en<-cs, en->de, en-ro, and mT5 in en-<tr, models$^{p\cdot f}$ is slightly worse than models^{f}, the results are still very competitive. 3) The table also indicates the LM^2_gSAR pre-training is somewhat helpful for the high/medium/low/extremely low resource translations, and the results show a consistent improvement trend. 4) In the table, we have bolded the best BLEU scores for both directions, and the best results are almost from mBART$^{p\cdot f}$. The M2M100$^{p\cdot f}$ and mT5$^{p\cdot f}$ also perform better in most cases. It is worth noticing that on en-hi benchmarks, the M2M100$^{p\cdot f}$ far exceeds the others in both directions. This may benefit from the mechanism in M2M100[8] that its initial parameters are pre-trained with pseudo-parallel data, and the LM^2_gSAR pre-training in this work can further strengthen its ability.

At present, the LM^2_gSAR pre-training in this experiment is just an initial attempt, the improvement is still significant. Once we expand the scale of the pre-training, it will be really encouraging. Considering the improvement from multiple dimensions, the experimental results show LM^2_gSAR has a signifi-

[8]One may notice that the BLEU scores in this experiment are different from those in M2M100 paper (Fan et al., 2021). One reason is that the evaluation benchmarks between this work and the original M2M100 paper are different, and there is almost no overlap. The other reason is that they reported the scores of the M2M100-1.2B model while we used the M2M100-418M.

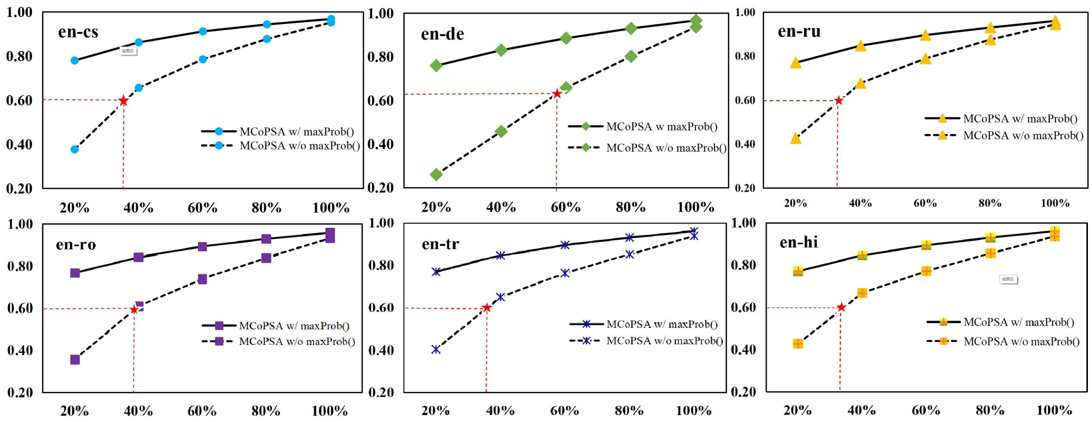

Figure 3: The average LaBSE score curve of each fold on the en-XX languages. The X-axis is the distribution interval of each fold, and the Y-axis is the LaBSE score. MCoPSA w/(w/o) *maxProb*(): MCoPSA algorithm with(without) the *maxProb*(·) function.

cant contribution to translation tasks.

4.6. Significant test

In this work, we also performed the significance test on machine translation task based on the models that report the best BLEU scores in both directions (see the bold results in Table 7). The well-known Wilcoxon signed-rank test was used to measure whether the improvement between the corresponding data distributions in two samples is significant.

In the Wilcoxon signed-rank test, we first randomly sampled 50% data in each test set 20 times and used the model$^{p\cdot f}$ and model^{f} to predict the translations on the sample data. Second, we scored the translation results with the sacreBLEU script to obtain the BLEU score on each direction of each benchmark. After sampling 20 times, we had a sequence of BLEU scores with the length of 20 for model$^{p\cdot f}$ and model^{f}, respectively. Finally, the corresponding BLEU score sequences for model$^{p\cdot f}$ and model^{f} were input into the "wilcox.test()" function in R Tutorial, and the function will output the P-value of two sequences to indicate the significance. If P-value$<$0.05, the improvement between model$^{p\cdot f}$ and model^{f} is significant, otherwise not. Finally, in Table 7, on each translation direction of each benchmark, the improvement between model$^{p\cdot f}$ and model^{f}on BLEU score is significant (P-value$<$0.05).

5. Discussion

In section 3.2, this paper has proven the advantages of scale and linguistic diversity of the proposed approach via statistics on LM^2_gSAR. This section discusses the quality and lifelong property.

5.1. Quality control in MCoPSA agorithm

In the MCoPSA algorithm, the function *maxProb*(·) provides a series of post-processing operations to output the co-occurrence probability. We find that it is a key point to control the quality of the aligned linguistic units in MCoPSA. To prove the quality, we perform an analysis experiment on en-XX languages. In this experiment, we use the public LaBSE (Feng et al., 2022) to evaluate the aligned linguistic units that MCoPSA extracted with or without the *maxProb*(·) function. LaBSE can map languages to a shared vector space and compute their similarities. Then given the en-XX aligned linguistic units, LaBSE will output a similarity score.

First, we collected the aligned linguistic units based on "MCoPSA w/o *maxProb*(·)" in en-XX languages. Second, we used LaBSE to compute the similarity scores and ranked them in ascending order. Then, we divided the ranked units into five-folds, and each fold contains 20% of the whole units. Finally, we averaged the LaBSE scores in each fold. The average LaBSE score curves (Dotted curve) for each fold on en-XX languages are presented in Figure 3. The dotted curves indicate that the scores range from 0.2 to 1.0, and a certain percentage (≈40%) of data falls into 0.2∼0.6. Based on our manual statistics, the aligned linguistic units with a LaBSE score of less than 0.6 are quite noisy. Next, we repeated the operations with "MCoPSA w/ *maxProb*(·)" to recollect and recompute the LaBSE score of the aligned linguistic units. The solid curves in Figure 3 indicate the distributions of the scores, which range from 0.7 to 1.0, which brings a great improvement for each language. The score ranges between two curves prove that the *maxProb*(·) function plays a key role in quality control. Meanwhile, the LaBSE scores are over 0.7 and beyond our statistics of 0.6, indicating that the aligned linguistic units via the MCoPSA algorithm have good quality.

5.2. Lifelong property of the approach

The term "lifelong" is an important property of the proposed approach and a key differentiator from other known methods. It refers to the sustainability and extensibility of the proposed approach, which is mainly reflected in the language extension and continuous scale expansion of the resource.

First, in this paper, the proposed approach released the first version of LM_g^2SAR, which supports seven languages. But from Table 1 in Section 3.2, we know that the approach only relies on the en-XX bilingual data. The multilingualism in Table 2 also shows that the approach presents positive effects between languages. So the approach can easily extend new languages into LM_g^2SAR through their bilingual data, and make connections between the new language and other languages to improve linguistic diversity. Second, Table 1 also shows that the noisy bilingual data used in this paper is only a part of the original library, and the linguistic units are far from reaching the upper bound. Thus, with the expansion of the parallel data, the approach can expand the scale of the resource, and supplement more linguistic units to perfect its resource.

6. Conclusion

In this paper, to alleviate the problem of lacking sufficient alignment resources in the pre-training methods, we proposed a lifelong multilingual multi-granularity semantic alignment approach via maximum co-occurrence probability in the noisy parallel data and released a version of its corresponding resource. We also conducted experiments to prove the ability of the MCoPSA algorithm compared to the traditional aligners and elaborate on how to use the resource to prove its effectiveness in machine translation tasks. The experimental results, analysis, and discussion also prove the superiority of the proposed approach and resource.

In the future, we will continue to optimize the approach from the quality and linguistic diversity. Meanwhile, we will release more versions of the resource with the optimized approach to support more languages and provide a bigger scale. Besides, we will explore the strategies for utilizing the resource to contribute to the pre-training methods. At the same time, the approaches and resources will gradually be opened to the public.

7. Acknowledgements

We thank all the reviewers for their efforts to make the paper comprehensive and solid. This work is supported by the National Key Research and Development Program of China (Grant No. 2021ZD0112905), the Major Key Project of PCL (Grant No. PCL2023A09), the National Natural Science Foundation of China (Grant No. 62206140), and the China Postdoctoral Science Foundation (Grant No. 2022M711726).

8. Bibliographical References

Omar Adjali, Emmanuel Morin, and Pierre Zweigenbaum. 2022. Building comparable corpora for assessing multi-word term alignment. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 3103–3112, Marseille, France. European Language Resources Association.

Goonmeet Bajaj, Vinh Nguyen, Thilini Wijesiriwardene, Hong Yung Yip, Vishesh Javangula, Amit Sheth, Srinivasan Parthasarathy, and Olivier Bodenreider. 2022. Evaluating biomedical word embeddings for vocabulary alignment at scale in the UMLS Metathesaurus using Siamese networks. In Proceedings of the Third Workshop on Insights from Negative Results in NLP, pages 82–87, Dublin, Ireland. Association for Computational Linguistics.

Akshay Batheja and Pushpak Bhattacharyya. 2022. Improving machine translation with phrase pair injection and corpus filtering. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5395–5400, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Francisco Casacuberta and Enrique Vidal. 2007. Giza++: Training of statistical translation models.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, He-Yan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3418–3430.

Everlyn Chimoto and Bruce Bassett. 2022. Very low resource sentence alignment: Luhya and Swahili. In Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022), pages 1–8, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2112–2128.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 644–648.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. J. Mach. Learn. Res., 22(107):1–48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Jenn Leana Fernandez and Kristine Mae M. Adlaon. 2022. Exploring word alignment towards an efficient sentence aligner for Filipino and Cebuano languages. In Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022), pages 99–106, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In Proceedings of the 6th Language Resources and Evaluation Conference.

Emmanuel Giguet and Pierre-Sylvain Luquet. 2006. Multilingual lexical database generation from parallel texts in 20 european languages with endogenous resources. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 271–278.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. Multimwe: Building a multi-lingual multi-word expression (mwe) parallel corpora. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 2970–2979.

Ayyoob Imani, Lütfi Kerem Senel, Masoud Jalili Sabet, François Yvon, and Hinrich Schuetze. 2022. Graph neural networks for multiparallel word alignment. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1384–1396, Dublin, Ireland. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020. Pre-training via paraphrasing. Advances in Neural Information Processing Systems, 33:18470–18481.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2649–2663.

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. Transactions of the Association for Computational Linguistics, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. CoRR, abs/1711.05101.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. Veco: Variable and flexible cross-lingual pre-training for language understanding and generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3980–3994.

Zhuoyuan Mao, Chenhui Chu, Raj Dabre, Haiyue Song, Zhen Wan, and Sadao Kurohashi. 2022. When do contrastive word alignments improve many-to-many neural machine translation? In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1766–1775, Seattle, United States. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1):19–51.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation.

In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 244–258.

Matt Post. 2018. A call for clarity in reporting bleu scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Édouard Grave, Armand Joulin, and Angela Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6490–6500.

Henry Tang, Ameet Deshpande, and Karthik Narasimhan. 2022. Align-mlm: Word embedding alignment is crucial for multilingual pre-training. arXiv preprint arXiv:2211.08547.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3450–3466.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal. European Language Resources Association (ELRA).

Mingxuan Wang and Lei Li. 2021. Pre-training methods for neural machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 21–25.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In International Conference on Learning Representations.

Di Wu, Liang Ding, Shuo Yang, and Mingyang Li. 2022. MirrorAlign: A super lightweight unsupervised word alignment model via cross-lingual contrastive learning. In Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 83–91, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498.

Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. 2020a. Alternating language modeling for cross-lingual pre-training. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 9386–9393.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020b. Csp: Code-switching pre-training for neural machine translation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2624–2636.

Ziqing Yang, Wentao Ma, Yiming Cui, Jiani Ye, Wanxiang Che, and Shijin Wang. 2021. Bilingual alignment pre-training for zero-shot cross-lingual transfer. In Proceedings of the 3rd Workshop on Machine Reading for Question Answering, pages 100–105.

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022. Automatic translation alignment for Ancient Greek and Latin. In Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, pages 101–107, Marseille, France. European Language Resources Association.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).