Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information

Zehui Lin[†,‡*], **Xiao Pan**[†*], **Mingxuan Wang**[†], **Xipeng Qiu**[‡], **Jiangtao Feng**[†], **Hao Zhou**[†], **Lei Li**[†]
[†]ByteDance AI Lab
{*panxiao.94,wangmingxuan.89,zhouhao.nlp,fengjiangtao,lileilab*}*@bytedance.com*
[‡]School of Computer Science, Fudan University, Shanghai, China
{*linzh18,xpqiu*}*@fudan.edu.cn*

Abstract

We investigate the following question for machine translation (MT): can we develop a single universal MT model to serve as the common seed and obtain derivative and improved models on arbitrary language pairs? We propose mRASP, an approach to pre-train a universal multilingual neural machine translation model. Our key idea in mRASP is its novel technique of random aligned substitution, which brings words and phrases with similar meanings across multiple languages closer in the representation space. We pre-train a mRASP model on 32 language pairs jointly with only public datasets. The model is then fine-tuned on downstream language pairs to obtain specialized MT models. We carry out extensive experiments on 42 translation directions across a diverse settings, including low, medium, rich resource, and as well as transferring to exotic language pairs. Experimental results demonstrate that mRASP achieves significant performance improvement compared to directly training on those target pairs. It is the first time to verify that multiple low-resource language pairs can be utilized to improve rich resource MT. Surprisingly, mRASP is even able to improve the translation quality on exotic languages that never occur in the pre-training corpus. Code, data, and pre-trained models are available at `https://github.com/linzehui/mRASP`.

1 Introduction

Pre-trained language models such as BERT have been highly effective for NLP tasks (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019; Conneau and Lample, 2019; Liu et al., 2019; Yang et al., 2019). Pre-training and fine-tuning has been a successful paradigm. It is intriguing to discover a "BERT" equivalent – a pre-trained model – for machine translation. In this paper, we study the following question: can we develop a single universal MT model and derive specialized models by fine-tuning on an arbitrary pair of languages?

While pre-training techniques are working very well for NLP task, there are still several limitations for machine translation tasks. First, pre-trained language models such as BERT are not easy to directly fine-tune unless using some sophisticated techniques (Yang et al., 2020). Second, there is a discrepancy between existing pre-training objective and down-stream ones in MT. Existing pre-training approaches such as MASS (Song et al., 2019) and mBART (Liu et al., 2020) rely on auto-encoding objectives to pre-train the models, which are different from translation. Therefore, their fine-tuned MT models still do not achieve adequate improvement. Third, existing MT pre-training approaches focus on using multilingual models to improve MT for low resource or medium resource languages. There has not been one pre-trained MT model that can improve for any pairs of languages, even for rich resource settings such as English-French.

In this paper, we propose multilingual Random Aligned Substitution Pre-training (mRASP), a method to pre-train a MT model for many languages, which can be used as a common initial model to fine-tune on arbitrary language pairs. mRASP will then improve the translation performance, comparing to the MT models directly trained on downstream parallel data. In our method, we ensure that the pre-training on many languages and the down-stream fine-tuning share the same model architecture and training objective. Therefore, this approach lead to large translation performance gain. Consider that many languages differ lexically but are closely related at the semantic level, we start by training a large-scale multilingual NMT model across different translation directions, then fine-tuning the model in a specific direction.

[*] Equal contribution. The work was done when the first author was an intern at ByteDance.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2649–2663,
November 16–20, 2020. ©2020 Association for Computational Linguistics

Further, to close the representation gap across different languages and make full use of multilingual knowledge, we explicitly introduce additional loss based on random and aligned substitution of the words in the source and target sentences. Substituted sentences are trained jointly with the same translation loss as the original multilingual parallel corpus. In this way, the model is able to bridge closer the representation space across different languages.

We carry out extensive experiments in different scenarios, including translation tasks with different dataset scales, as well as zero-shot translation tasks. For extremely low resource (<100k), mRASP obtains gains up to +29 BLEU points compared to directly trained models on the downstream language pairs. mRASP obtains consistent performance gains as the size of datasets increases. Remarkably, even for rich resource (>10M, e.g. English-French), mRASP still achieves big improvements. Surprisingly, even when mRASP is fine-tuned on two exotic languages that never occur in the pre-training corpus, the resulting MT model is still much better than the directly trained ones (+3.3 to +14.1 BLEU). We finally conduct extensive analytic experiments to examine the contributing factors inside the mRASP method for the performance gains.

We highlight our contributions as follows:

- We propose mRASP, an effective pre-training method that can be utilized to fine-tune on any language pairs in NMT. It is very efficient in the use of parallel data in multiple languages. While other pre-trained language models are obtained through hundreds of billions of monolingual or cross-lingual sentences, mRASP only introduces several hundred million bilingual pairs. We suggest that the consistent objectives of pre-training and fine-tuning lead to better model performance.

- We explicitly introduce a random aligned substitution technique into the pre-training strategy, and find that such a technique can bridge the semantic space between different languages and thus improve the final translation performance.

- We conduct extensive experiments 42 translation directions across different scenarios, demonstrating that mRASP can significantly boost the performance on various translation tasks. mRASP achieves 14.1 BLEU with only 12k pairs of Dutch and Portuguese sentences even though neither appears in the pre-training data. mRASP also achieves 44.3 BLEU on WMT14 English-French translation. Note that our pre-trained model only use parallel corpus in 32 languages, unlike other methods that also use much more monolingual raw corpus.

2 Methodology

In this section, we introduce our proposed mRASP and the training details.

2.1 mRASP

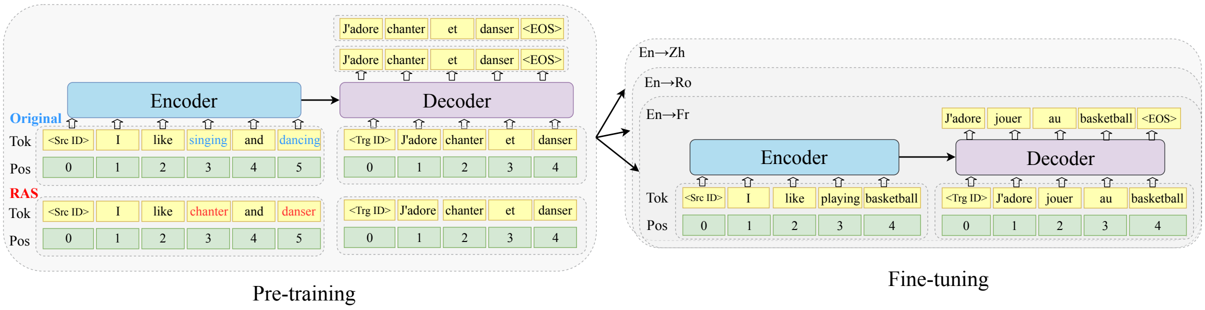
Architecture We adopt a standard Transformer-large architecture (Vaswani et al., 2017) with 6-layer encoder and 6-layer decoder. The model dimension is 1,024 on 16 heads. We replace ReLU with GeLU (Hendrycks and Gimpel, 2016) as activation function on feed forward network. We also use learned positional embeddings.

Methodology A multilingual neural machine translation model learns a many-to-many mapping function f to translate from one language to another. More formally, define $L = \{L_1, \dots, L_M\}$ where L is a collection of languages involving in the pre-training phase. $\mathcal{D}_{i,j}$ denotes a parallel dataset of (L_i, L_j), and $\mathcal{E}$ denotes the set of parallel datasets $\{\mathcal{D}\}_{i=1}^{i=N}$, where N the numbers of the bilingual pair. The training loss is then defined as:

$$\mathcal{L}^{pre} = \sum_{i,j \in \mathcal{E}} \mathbb{E}_{(\mathbf{x}^i, \mathbf{x}^j) \sim \mathcal{D}_{i,j}} [-\log P_\theta(\mathbf{x}^i | C(\mathbf{x}^j))]. \tag{1}$$

where $\mathbf{x}^i$ represents a sentence in language L_i, and θ is the parameter of mRASP, and $C(\mathbf{x}^i)$ is our proposed alignment function, which randomly replaces the words in $\mathbf{x}^i$ with a different language. In the pre-training phase, the model jointly learns all the translation pairs.

Language Indicator Inspired by (Johnson et al., 2017; Ha et al., 2016), to distinguish from different translation pairs, we simply add two artificial language tokens to indicate languages at the source and target side. For instance, the following En→Fr sentence "`How are you? -> Comment vas tu?` " is transformed to "`<en> How are you? -> <fr> Comment vas tu?`"



Pre-training

Fine-tuning

Figure 1: The proposed mRASP method. "Tok" denotes token embedding while "Pos" denotes position embedding. During the pre-training phase, parallel sentence pairs in many languages are trained using translation loss, together with their substituted ones. We randomly substitute words with the same meanings in the source and target sides. During the fine-tuning phase, we further train the model on the downstream language pairs to obtain specialized MT models.

Multilingual Pre-training via RAS Recent work proves that cross-lingual language model pre-training could be a more effective way to representation learning (Conneau and Lample, 2019; Huang et al., 2019). However, the cross-lingual information is mostly obtained from shared subword vocabulary during pre-training, which is limited in several aspects:

- The vocabulary sharing space is sparse in most cases. Especially for dissimilar language pairs, such as English and Hindi, they share a fully different morphology.
- The same subword across different languages may not share the same semantic meanings.
- The parameter sharing approach lacks explicit supervision to guild the word with the same meaning from different languages shares the same semantic space.

Inspired by constructive learning, we propose to bridge the semantic gap among different languages through **Random Aligned Substitution (RAS)**. Given a parallel sentence $(\mathbf{x}^i, \mathbf{x}^j)$, we randomly replace a source word in $\mathbf{x}^i_t$ to a different random language L_k, where t is the word index. We adopt an unsupervised word alignment method MUSE(Lample et al., 2018b), which can translate $\mathbf{x}^i_t$ to $d_{i,k}(\mathbf{x}^i_t)$ in language L_k, where $d_{i,k}(\cdot)$ is the dictionary translating function. With the dictionary replacement, the original bilingual pair will construct a code-switched sentence pair $(C(\mathbf{x}^i), \mathbf{x}^j)$. As the benefits of random sampling, the translation set $\{d_{i,k}(\mathbf{x}^i_t)\}_{k=1}^{k=M}$ potentially appears in the same context. Since the word representation depends on the context, the word with similar meaning across different languages can share a similar representation. Figure 1 shows our alignment methodology.

2.2 Pre-training Data

We collect 32 English-centric language pairs, resulting in 64 directed translation pairs in total. English is served as an anchor language bridging all other languages. The parallel corpus are from various sources: *ted*[1], *wmt*[2], *europarl*[3], *paracrawl*[4], *opensubtitles*[5], *qed*[6]. We refer to our pre-training data as **PC32**(Parallel Corpus 32). **PC32** contains a total size of 197M pairs of sentences. Detailed descriptions and summary for the datasets can be found in Appendix.

For RAS, we utilize ground-truth En-X bilingual dictionaries[7], where X denotes languages involved in PC32. Since not all languages in PC32 have ground-truth dictionaries, we only use available dictionaries.

2.3 Pre-training Details

We use learned joint vocabulary. We learn shared BPE (Sennrich et al., 2016b) merge operations (with 32k merge ops) across all the training data and added monolingual data as a supplement (limit to 1M sentences). We do over-sampling in learning BPE to balance the vocabulary size of languages, whose resources are drastically different in size. We over-sampled the corpus of each language based on the volume of the largest language corpus. We

[1] Compiled by Qi et al. (2018). For simplicity, we deleted zh-tw and zh (which is actually Cantonese), and merged fr-ca with fr, pt-br with pt.

[2] http://www.statmt.org

[3] http://opus.nlpl.eu/Europarl-v8.php

[4] https://paracrawl.eu/

[5] http://opus.nlpl.eu/OpenSubtitles-v2018.php

[6] http://opus.nlpl.eu/QED-v2.0a.php

[7] https://github.com/facebookresearch/MUSE

Extremely Low Resource (<100k)									
Lang-Pairs	En-Be		En-My		En-Af		En-Eo		Avg
Size	20K		29k		41K		67K		
Direction	→	←	→	←	→	←	→	←	
Direct	8.5	9.6	10.2	5.4	8.3	7.2	4.9	6.7	7.6
mRASP	**25.8**	**32.3**	**28.6**	**25.3**	**31.1**	**27.0**	**30.4**	**35.8**	29.5
Δ	+17.3	+22.7	+18.4	+19.9	+22.8	+19.8	+25.5	+29.1	**+21.9**

Low Resource (100k∼1m)									
Lang-Pairs	En-He		En-Tr		En-Ro		En-Cs		Avg
Size	335K		388K		600K		978K		
Direction	→	←	→	←	→	←	→	←	
Direct	19.0	27.6	10.7	19.4	30.5	29.2	19.0	22.7	22.3
mRASP	**32.4**	**44.6**	**21.0**	**33.3**	**39.0**	**37.4**	**23.2**	**29.8**	32.6
Δ	+13.4	+17.0	+10.3	+13.9	+8.5	+8.2	+4.2	+7.1	**+10.3**

Medium Resource (1m∼10m)									
Lang-Pairs	En-Ar		En-Et		En-Bg		En-De		Avg
Size	1.2M		2.3M		3.1M		4.5M		
Direction	→	←	→	←	→	←	→	←	
Direct	14.1	27.6	20.2	24.5	37.4	41.1	29.3	30.8	28.1
mRASP	**19.5**	**38.2**	**25.3**	**31.3**	**39.3**	**44.2**	**30.3**	**34.4**	32.8
Δ	+5.4	+10.6	+5.1	+6.8	+1.9	+3.1	+1.0	+3.6	**+4.7**

Table 1: Fine-tuning performance on *extremely low* / *low* / *medium* resource machine translation settings. The numbers in parentheses indicate the size of parallel corpus for fine-tuning. Pre-training with mRASP and then fine-tuning on downstream MT tasks consistently improves over MT models directly trained on bilingual parallel corpus.

keep tokens occurring more than 20, which results in a subword vocabulary of 64,808 tokens.

In pre-training phase, we train our model with the full pairs of the parallel corpus. Following the training setting in Transformer, we use Adam optimizer with $\epsilon = 1e-8, \beta_2 = 0.98$. A warm-up and linear decay scheduling with a warm-up step of 4000 is used. We pre-train the model for a total of 150000 steps.

For RAS, we use the top 1000 words in dictionaries and only substitute words in source sentences. Each word is replaced with a probability of 30% according to the En-X bilingual dictionaries. To address polysemy, we randomly select one substitution from all candidates.

3 Experiments

This section shows that mRASP obtains consistent performance gains in different scenarios. We also compare our method with existing pre-training methods and outperforms the baselines on En→Ro dataset. The performance further boosts by combining back-translation(Sennrich et al., 2016a) technique. Otherwise stated, for all experiments, we use the pre-trained model as initialization and fine-tune with the downstream target parallel corpus.

3.1 Experiment Settings

Datasets We collect 14 pairs of parallel corpus to simulate different scenarios. Most of the En-X parallel datasets are from the pre-training phase to avoid introducing new information. Most pairs for fine-tuning are from previous years of WMT and IWSLT. Specifically, we use WMT14 for En-De and En-Fr, WMT16 for En-Ro. For pairs like Nl(Dutch)-Pt(Portuguese) that are not available in WMT or IWSLT, we use news-commentary instead. For a detailed description, please refer to the Appendix.

[8] CTNMT only reports the Transformer-base setting.

Lang-Pairs Size	En→De 4.5M	Zh→En 20M	En→Fr 40M
Direct	29.3	24.1	43.2
CTNMT[8] (2020)	30.1	-	42.3
mBART (2020)	-	-	41.0
XLM (2019)	28.8	-	-
MASS (2019)	28.9	-	-
mBERT (2019)	28.6	-	-
mRASP	**30.3**	**24.7**	**44.3**

Table 2: Fine-tuning performance for popular *medium* and *rich* resource MT tasks. For fair comparison, we report detokenized BLEU on WMT newstest18 for Zh→En and tokenized BLEU on WMT newstest14 for En→Fr and En→De. Notice unlike previous methods (except CTNMT) which do not improve in the rich resource settings, mRASP is again able to consistently improve the downstream MT performance. It is the first time to verify that low-resource language pairs can be utilized to improve rich resource MT.

Based on the volume of parallel bi-texts, we divide the datasets into four categories: extremely low resource (<100K), low resource(>100k and <1M), medium resource (>1M and <10M), and rich resource (>10M).

For back translation, we include 2014-2018 newscrawl for the target side, En. The total size of the monolingual data is 3M.

Baseline To better quantify the effectiveness of the proposed pre-training models, we also build two baselines.

mRASP w/o RAS. To measure the effect of alignment information, we also pre-train a model on the same PC32. We do not include alignment information on this pre-training model.

Direct. We also train randomly initialized models directly on downstream bilingual parallel corpus as a comparison with pre-training models.

Fine-tuning We fine-tune our obtained mRASP model on the target language pairs. We apply a dropout rate of 0.3 for all pairs except for rich resource such as En-Zh and En-Fr with 0.1. We carefully tune the model, setting different learning rates and learning scheduler warm-up steps for different data scale. For inference, we use beam-search with beam size 5 for all directions. For most cases, We measure case-sensitive tokenized BLEU. We also report de-tokenized BLEU with SacreBLEU (Post, 2018) for a fair comparison with previous works.

3.2 Main Results

We first conduct experiments on the (extremely) low-resource and medium-resource datasets, where multilingual translation usually obtains significant improvements. As illustrated in Table 1, we obtain significant gains in all datasets. For extremely low resources setting such as En-Be (Belarusian) where the amount of datasets cannot train an NMT model properly, utilizing the pre-training model boosts performance.

We also obtain consistent improvements in low and medium resource datasets. Not surprisingly, We observe that with the scale of the dataset increasing, the gap between the randomly initialized baseline and pre-training model is becoming closer. It is worth noting that, for En→De benchmark, we obtain 1.0 BLEU points gains[9].

To verify mRASP can further boost performance on rich resource datasets, we also conduct experiments on En→Zh and En→Fr. We compare our results with two strong baselines reported by Ott et al. (2018); Li et al. (2019). As shown in Table 2, surprisingly, when large parallel datasets are provided, it still benefits from pre-training models. In En→Fr, we obtain 1.1 BLEU points gains.

Comparing to other Pre-training Approaches We compare our mRASP to recently proposed multilingual pre-training models. Following Liu et al. (2020), we conduct experiments on En-Ro, the only pairs with established results. To make a fair comparison, we report de-tokenized BLEU.

As illustrated in Table 4 , Our model reaches comparable performance on both En→Ro and Ro→En. We also combine Back Translation (Sennrich et al., 2016a) with our PNMT, observing performance boost up to 2 BLEU points, suggesting mRASP is complementary to BT. It should be noted that the competitors introduce much more pre-training data.

mBART contucted experiments on extensive language pairs. To illustrate the superiority of mRASP, we also compare our results with mBART. We use the same test sets as mBART. As illustrated in Table 5, mRASP outperforms mBART for most of language pairs by a large margin. Note that while mBART underperforms baseline for benchmarks En-De and En-Fr, mRASP obtains 4.3 and 2.9 BLEU gains compared to baseline.

[9] We report results of En→De on newstest14. The baseline result is reported in Ott et al. (2018)

	Exotic Pair				**Exotic Full**			
Lang-Pairs	Fr-Zh		De-Fr		Nl-Pt		Da-El	
Size	20K		9M		12K		1.2M	
Direction	→	←	→	←	→	←	→	←
Direct	0.7	3.0	23.5	21.2	0.0	0.0	14.1	16.9
mRASP	**25.8**	**26.7**	**29.9**	**23.4**	**14.1**	**13.2**	**17.6**	**19.9**
	Exotic Source/Target							
Lang-Pairs	En-Mr		En-Gl		En-Eu		En-Sl	
Size	11K		200K		726K		2M	
Direction	→	←	→	←	→	←	→	←
Direct	6.4	6.8	8.9	12.8	7.1	10.9	24.2	28.2
mRASP	**22.7**	**22.9**	**32.1**	**38.1**	**19.1**	**28.4**	**27.6**	**29.5**

Table 3: Fine-tuning MT performance on exotic language corpus. For two the translation direction A→B, *exotic pair*: A and B occur in the pre-training corpus but no pairs of sentences of (A,B) occur; *exotic full*: no sentences in either A nor B occur in the pre-training; *exotic source*: sentences from the target side B occur in the pre-training but not the source side A; *exotic target*: sentences from the source side A occur in the pre-training but not the target side B. Notice that pre-training with mRASP and fine-tuning on those exotic languages consistently obtains significant improvements MT performance in each category.

Model	En→Ro	Ro→En	Ro→En +BT
Direct	34.3	34.0	36.8
XLM (2019)	-	35.6	38.5
MASS (2019)	-	-	**39.1**
BART (2020)	-	-	38.0
XLM-R (2020)	35.6	35.8	-
mBART (2020)	**37.7**	**37.8**	38.8
mRASP	37.6	36.9	38.9

Table 4: Comparison with previous Pre-training models on WMT16 En-Ro. Following (Liu et al., 2020), We report detokenized BLEU. We reaches comparable results on both En→Ro and Ro→En. By combining back translation, the performance further boost for 2 BLEU points on Ro→En.

3.3 Generalization to Exotic Corpus

To illustrate the generalization of mRASP, we also conduct experiments on exotic corpus, which is not included in our pre-training phase. Here we divide exotic corpus into four categories with respect to the source and target side.

- **Exotic Pair** Both source and target languages are individually pre-trained while they have not been seen as bilingual pairs.
- **Exotic Source** Only target language is pre-trained, but source language is not.
- **Exotic Target** Only source language is pre-trained, but the target language is not.
- **Exotic Full** Neither source nor target language is pre-trained.

For each category, we select language pairs of different scales. The results are shown in Table 3. As is shown, mRASP obtains significant gains for each category for different scales of datasets, indicating that even trained with exotic languages, with pre-training initialization, the model still works reasonably well.

Note that in the most challenging case, Exotic Full, where the model does not have any knowledge of both sides, with only 11K parallel pairs for Nl(Dutch)-Pt(Portuguese), the pre-training model still reaches reasonable performance, while the baseline fails to train appropriately. It suggests the pre-train model does learn language-universal knowledge and can transfer to exotic languages easily.

4 Analysis

In this section, we conduct a set of analytical experiments to better understand what contributes to performance gains. Three aspects are studied. First, we study whether the main contribution comes from pre-training or fine-tuning by comparing the performance of fine-tuning and no-fine-tuning. The results suggest that the performance mainly comes from pre-training, while fine-tuning further boosts

Lang-Pairs	En-Gu		En-Kk		En-Tr	
Source	WMT19		WMT19		WMT17	
Direction	→	←	→	←	→	←
Direct	0.0	0.0	0.2	0.8	9.5	12.2
mBART	0.1	0.3	2.5	7.4	17.8	22.5
mRASP	**3.2**	**0.6**	**8.2**	**12.3**	**20.0**	**23.4**
Lang-Pairs	En-Et		En-Fi		En-Lv	
Source	WMT18		WMT17		WMT17	
Direction	→	←	→	←	→	←
Direct	17.9	22.6	20.2	21.8	12.9	15.6
mBART	**21.4**	**27.8**	22.4	28.5	15.9	19.3
mRASP	20.9	26.8	**24.0**	28.0	**21.6**	**24.4**
Lang-Pairs	En-Cs		En-De		En-Fr	
Source	WMT19		WMT19		WMT14	
Direction	→		→		→	
Direct	16.5		30.9		41.4	
mBART	18.0		30.5		41.0	
mRASP	**19.9**		**35.2**		**44.3**	

Table 5: Comprehensive comparison with mBART. mRASP outperforms mBART on MT for all but two language pairs.

the performance. Second, we thoroughly analyze the difference between incorporating RAS at the pre-training phase and pre-training without RAS. The finding shows that incorporating alignment information helps bridge different languages and obtains additional gains. Lastly, we study the effect of data volume in the fine-tuning phase.

The effects with fine-tuning .

In the pre-training phase, the model jointly learns from different language pairs. To verify whether the gains come from pre-training or fine-tuning, we directly measure the performance without any fine-tuning, which is, in essence, zero-shot translation task.

We select datasets covering different scales. Specifically, En-Af (41k) from extremely low resource, En-Ro (600k) from low resource, En-De (4.5M) from medium resource, and En-Fr (40M) from rich resource are selected.

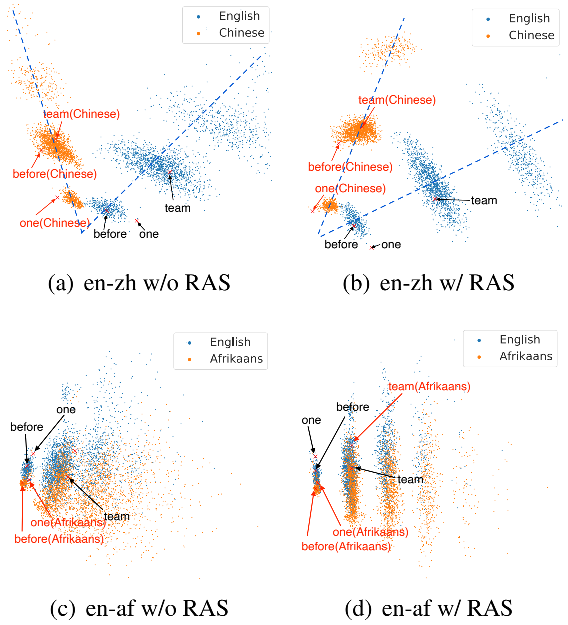
As shown in Table 6 , we find that model without fine-tuning works surprisingly well on all datasets, especially in low resource where we observe model without fine-tuning outperforms randomly initialized baseline model. It suggests that the model already learns well on the pre-training phase, and fine-tuning further obtains additional gains. We suspect that the model mainly tunes the embedding of specific language at the fine-tuning phase while keeping the other model parameters mostly unchanged. Further analytical experiments can be conducted to verify our hypothesis.

Note that we also report pre-trained model without RAS (NA-mRASP). For comparison, we do not apply fine-tuning on NA-mRASP. mRASP consistently obtains better performance that NA-mRASP, implying that injecting information at the pre-training phase do improve the performance.

The effectiveness of RAS technique .

In the pre-training phase, we explicitly incorporate RAS. To verify the effectiveness of RAS, we first compare the performance of mRASP and mRASP without RAS.

As illustrated in Table 7, We find that utilizing RAS in the pre-training phase consistently helps improve the performance in datasets with different scales, obtaining gains up to 2.5+ BLEU points.



(a) en-zh w/o RAS

(b) en-zh w/ RAS

(c) en-af w/o RAS

(d) en-af w/ RAS

Figure 2: Visualization of Word Embedding from mRASP w/o RAS vs mRASP w/ RAS. For both similar language pairs and dissimilar language pairs that have no lexical overlap, the word embedding distribution becomes closer after RAS.

To verify whether the semantic space of different languages draws closer after adding alignment information quantitatively, we calculate the average cosine similarity of words with the same meaning in different languages. We choose the top frequent 1000 words according to MUSE dictionary. Since words are split into subwords through BPE, we

Lang-Pairs	En-Af		En-Ro		En-De		En-Fr	
Size	41K		600k		4.5M		40M	
Direction	→	←	→	←	→	←	→	←
Direct	8.3	7.2	30.5	29.2	29.3	30.8	43.2	39.8
mRASP *w/o RAS & ft*	16.1	23.2	24.4	33.9	22.5	30.9	38.6	37.3
mRASP *w/o ft*	18.5	23.9	25.2	34.7	24.2	31.2	39.6	37.6
mRASP	31.1	27.0	39.0	37.4	30.3	34.4	44.3	45.4

Table 6: MT performance of mRASP with and without the RAS technique and fine-tuning strategy. mRASP includes both the RAS technique and fine-tuning strategy. "*w/o ft*" denotes "without fine-tuning". We also report mRASP without fine-tuning and NAS to compare with mRASP without fine-tuning. Both RAS and fine-tuning proves effective and essential for mRASP.

Lang-Pairs	En-Af		En-Ro		En-De	
Direction	→	←	→	←	→	←
. w/o RAS	30.6	25.4	36.3	36.4	27.7	33.2
mRASP	31.1	27.0	39.0	37.4	30.3	34.4

Table 7: The MT performance of three language pairs with and without alignment information (mRASP w/o RAS) at pre-training phase. We see consistent performance gains for mRASP with RAS.

simply add all subwords constituting the word. As illustrated in Figure 3, we find that for all pairs in the Figure, the average cosine similarity increases by a large margin after adding RAS, suggesting the efficacy of alignment information in bridging different languages. It is worth mentioning that the increase does not only happen on similar pairs like En-De, but also on dissimilar pairs like En-Zh.

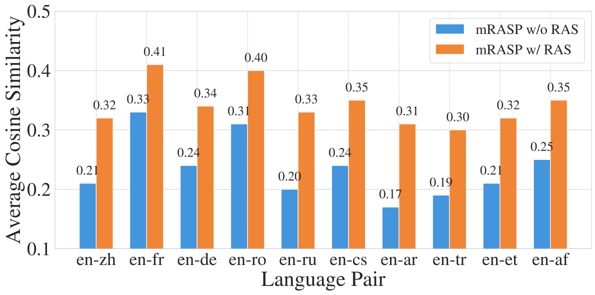


Figure 3: Average cosine similarity No-Alignment (mRASP w/o RAS) vs Alignment (mRASP w/ RAS). The similarity increases after applying the RAS technique, which explains the effectiveness of RAS.

To further illustrate the effect of RAS on semantic space more clearly, we use PCA (Principal Component Analysis) to visualize the word embedding space. We plot En-Zh as the representative for dissimilar pairs and En-Af for similar pairs. More figures can be found in the Appendix.

As illustrated in Figure 2, we find that for both similar pair and dissimilar pair, the overall word embedding distribution becomes closer after RAS. For En-Zh, as the dashed lines illustrate, the angle of the two word embedding spaces becomes smaller after RAS. And for En-Af, we observe that the overlap between two space becomes larger. We also randomly plot the position of three pairs of words, with each pair has the same meaning in different languages.

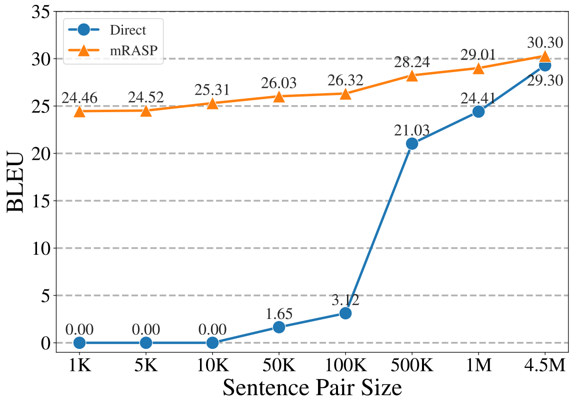


Figure 4: Performance curves for En→De along with the size of parallel pairs. With mRASP pre-trained model, the fine-tuned down-stream MT model is able to obtain descent translation performance even when there is very small corpus to train.

Fine-tuning Volume To study the effect of data volume in the fine-tuning phase, we randomly sample 1K, 5K, 10K, 50K, 100K, 500K, 1M datasets from the full En-De corpus (4.5M). We fine-tune the model with the sampled datasets, respectively. Figure 4 illustrates the trend of BLEU with the increase of data volume. With only 1K parallel pairs, the pre-trained model works surprisingly well, reaching 24.46. As a comparison, the model with random initialization fails on this extremely low resource. With only 1M pairs, mRASP reaches comparable results with baseline trained on 4.5M pairs.

With the size of dataset increases, the performance of the pre-training model consistently increases. While the baseline does not see any improvement until the volume of the dataset reaches 50K. The results confirm the remarkable boosting of mRASP on low resource dataset.

5 Related Works

Multilingual NMT aims at taking advantage of multilingual data to improve NMT for all languages involved, which has been extensively studied in a number of papers such as Dong et al. (2015); Johnson et al. (2017); Lu et al. (2018); Rahimi et al. (2019); Tan et al. (2019). The most related work to mRASP is Rahimi et al. (2019), which performs extensive experiments in training massively multilingual NMT models. They show that multilingual many-to-many models are effective in low resource settings. Inspired by their work, we believe that the translation quality of low-resource language pairs may improve when trained together with rich-resource ones. However, we are different in at least two aspects: *a)* Our goal is to find the best practice of a single language pair with multilingual pre-training. Multilingual NMT usually achieves inferior accuracy compared with its counterpart, which trains an individual model for each language pair when there are dozens of language pairs. *b)* Different from multilingual NMT, mRASP can obtain improvements with rich-resource language pairs, such as English-Frence.

Unsupervised Pretraining has significantly improved the state of the art in natural language understanding from word embedding (Mikolov et al., 2013b; Pennington et al., 2014), pretrained contextualized representations (Peters et al., 2018; Radford et al., 2019; Devlin et al., 2019) and sequence to sequence pretraining (Song et al., 2019). It is widely accepted that one of the most important factors for the success of unsupervised pre-training is the scale of the data. The most successful efforts, such as RoBERTa, GPT, and BERT, highlight the importance of scaling the amount of data. Following their spirit, we show that with massively multilingual pre-training, more than 110 million sentence pairs, mRASP can significantly boost the performance of the downstream NMT tasks.

On parallel, there is a bulk of work on unsupervised cross-lingual representation. Most traditional studies show that cross-lingual representations can be used to improve the quality of monolingual representations. Mikolov et al. (2013a) first introduces dictionaries to align word representations from different languages. A series of follow-up studies focus on aligning the word representation across languages (Xing et al., 2015; Ammar et al., 2016; Smith et al., 2017; Lample et al., 2018b). Inspired by the success of BERT, Conneau and Lample (2019) introduced XLM - masked language models trained on multiple languages, as a way to leverage parallel data and obtain impressive empirical results on the cross-lingual natural language inference (XNLI) benchmark and unsupervised NMT(Sennrich et al., 2016a; Lample et al., 2018a; Garcia et al., 2020). Huang et al. (2019) extended XLM with multi-task learning and proposed a universal language encoder.

Different from these works, *a)* mRASP is actually a multilingual sequence to sequence model which is more desirable for NMT pre-training; *b)* mRASP introduces alignment regularization to bridge the sentence representation across languages.

6 Conclusion

In this paper, we propose a multilingual neural machine translation pre-training model (mRASP). To bridge the semantic space between different languages, we incorporate word alignment into the pre-training model. Extensive experiments are conducted on different scenarios, including low/medium/rich resource and exotic corpus, demonstrating the efficacy of mRASP. We also conduct a set of analytical experiments to quantify the model, showing that the alignment information does bridge the gap between languages as well as boost the performance. We leave different alignment approaches to be explored in the future. In future work, we will pre-train on larger corpus to further boost the performance.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. We would also like to thank Liwei Wu, Huadong Chen, Qianqian Dong, Zewei Sun, and Weiying Ma for their useful suggestion and help with experiments.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

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

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

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1723–1732. The Association for Computer Linguistics.

Xavier Garcia, Pierre Foret, Thibault Sellam, and Ankur P. Parikh. 2020. A multilingual view of unsupervised machine translation. *CoRR*, abs/2002.02955.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798.

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

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2485–2494. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, Kai Feng, Hexuan Chen, Tengbo Liu, Yanyang Li, Qiang Wang, Tong Xiao, and Jingbo Zhu. 2019. The niutrans machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 257–266. Association for Computational Linguistics.

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. *CoRR*, abs/2001.08210.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine*

Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018, pages 84–92. Association for Computational Linguistics.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

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

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 1–9. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 529–535. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 151–164. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

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

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1006–1011. The Association for Computational Linguistics.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. Towards making the most of BERT in neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9378–9385. AAAI Press.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 5754–5764.

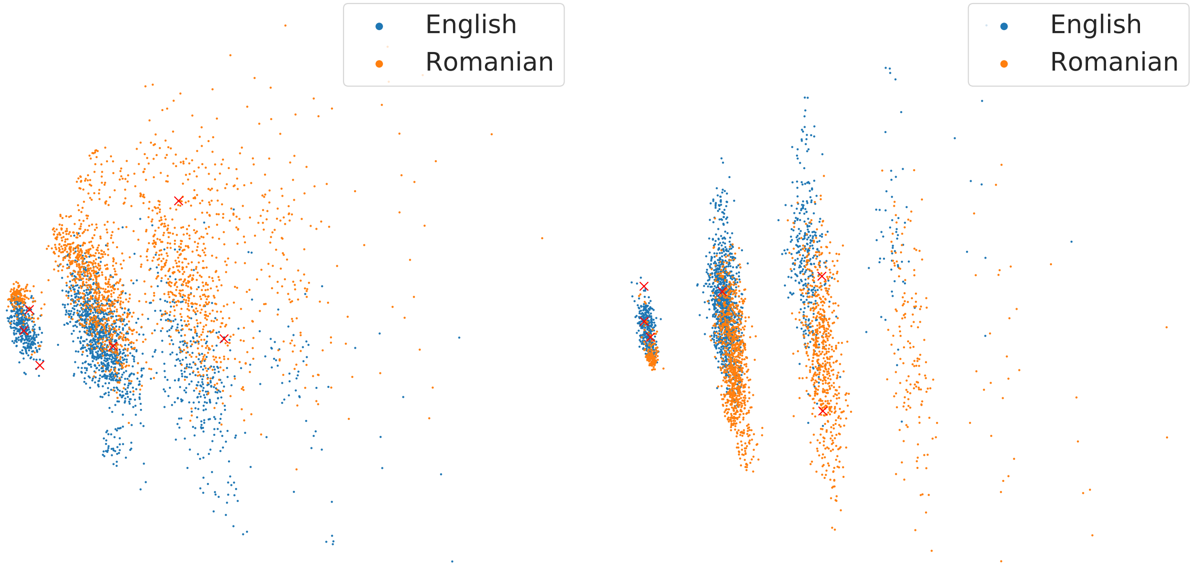
A Appendices

A.1 Visualization of Word Embedding

In addition to visualization of En-Zh and En-Af presented in main body of paper, we also plot visualization of En-Ro, En-Ar, En-Tr and En-De. As shown in Figure 5,6,7,8, the overall word embedding distribution becomes closer after RAS.

A.2 Data Description

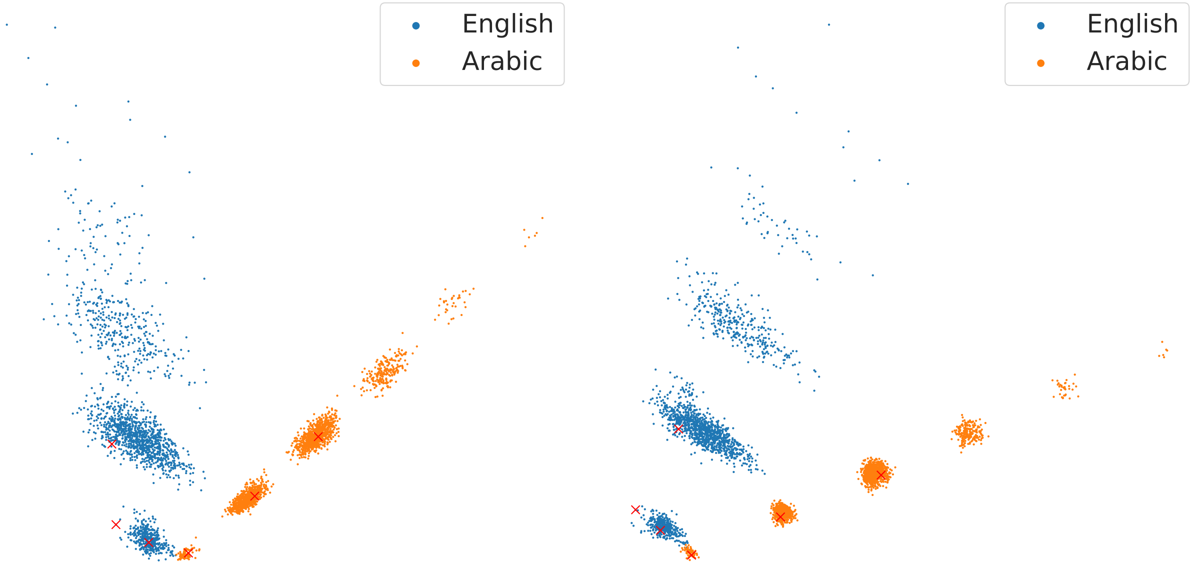
As listed in Table 8, we collect 32 English-centric language pairs, resulting in a total pairs of 110M. The parallel corpus are from various source, ted, wmt, europarl, paracrawl, opensubtitles and qed.



(a) en-ro w/o RAS

(b) en-ro w/ RAS

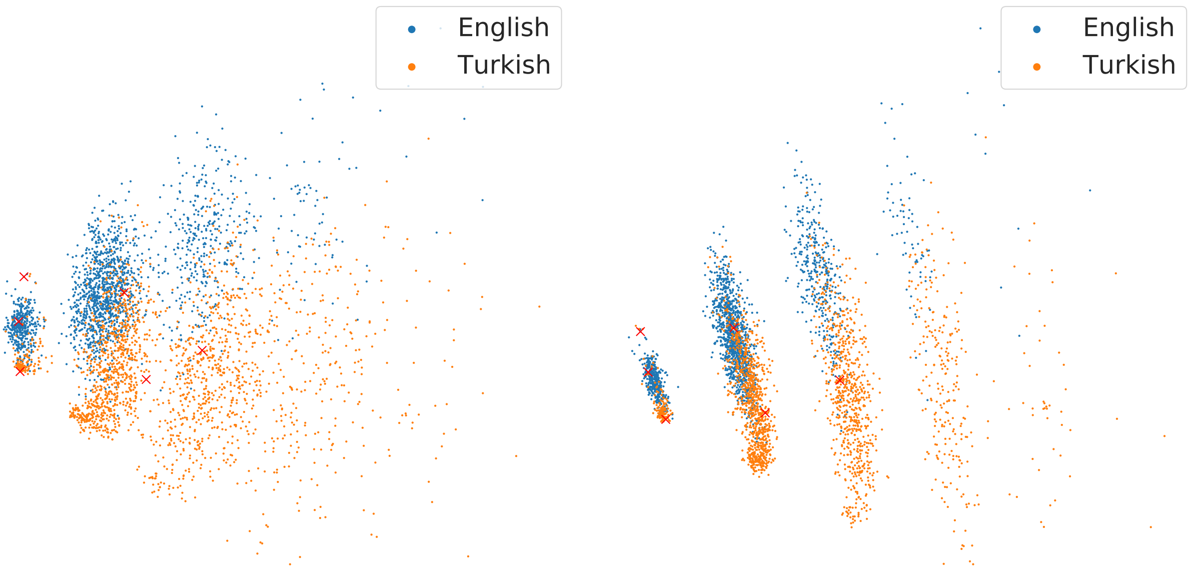
Figure 5: Visualization of Word Embedding from mRASP w/o RAS vs mRASP w/ RAS for English-Romanian



(a) en-ar w/o RAS

(b) en-ar w/ RAS

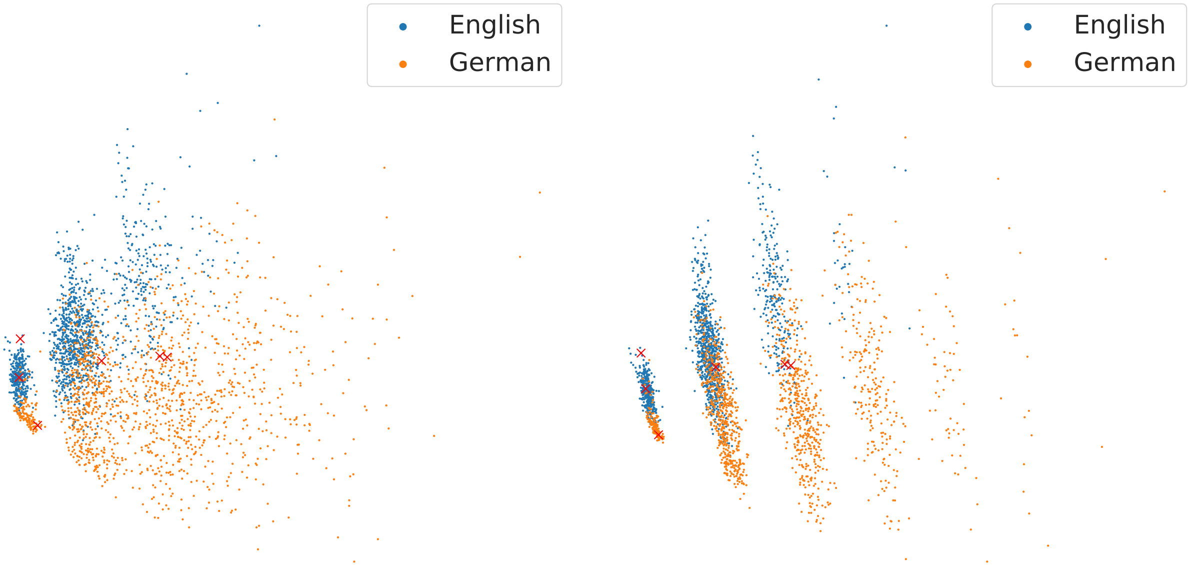
Figure 6: Visualization of Word Embedding from mRASP w/o RAS vs mRASP w/ RAS for English-Arabic



(a) en-tr w/o RAS

(b) en-tr w/ RAS

Figure 7: Visualization of Word Embedding from mRASP w/o RAS vs mRASP w/ RAS for English-Turkish



(a) en-de w/o RAS

(b) en-de w/ RAS

Figure 8: Visualization of Word Embedding from mRASP w/o RAS vs mRASP w/ RAS for English-German

Lang	Ted	Euro	Qed	Ops	WMT	Para	Others	Sum
Af	-	-	-	42429	-	-	-	42429
Ar	214111	-	-	1000788	-	-	-	1214899
Be	4509	-	21080	-	-	-	-	25589
Bg	174444	406934	-	-	-	2586277	-	3167655
Cs	103093	-	-	-	838037	-	-	941130
De	167888	-	-	-	4590101	-	-	4757989
El	134327	1235976	-	-	-	-	-	1370303
Eo	6535	-	-	61043	-	-	-	67578
Es	196026	1965734	-	-	-	-	-	2161760
Et	10738	-	-	-	2176827	132522	-	2320087
Fi	24222	1924942	-	-	2078670	-	-	4027834
Fr	192304	-	-	-	39816621	-	19870	40028795
Gu	-	-	-	-	11671	-	-	11671
He	211819	-	-	123692	-	-	-	335511
Hi	18798	-	-	-	-	-	1555738	1574536
It	204503	1909115	-	-	-	-	-	2113618
Ja	204090	-	-	1872100	-	-	-	2076190
Ka	13193	-	-	187411	-	-	-	200604
Kk	3317	-	-	-	124770	-	-	128087
Ko	205640	-	-	1270001	-	-	-	1475641
Lt	41919	-	-	-	2342917	-	-	2384836
Lv	-	-	-	-	4511715	1019003	-	5530718
Mn	7607	-	23126	-	-	-	-	30733
Ms	5220	-	-	1631386	-	-	-	1636606
Mt	-	-	-	-	-	177244	-	177244
My	21497	-	7518	-	-	-	-	29015
Ro	180484	-	-	-	610444	-	-	790928
Ru	208458	-	-	-	1640777	-	-	1849235
Sr	136898	-	-	-	-	-	-	136898
Tr	182470	-	-	-	205756	-	-	388226
Vi	171995	-	-	3055592	-	-	-	3227587
Zh	199855	-	-	-	25995505	-	-	26195360
Total	3245960	7442701	51724	9244442	84943811	3915046	1575608	110419292

Table 8: Statistics of the dataset PC32 for pre-training. Each entry shows the number of parallel sentence pairs between English and other language X.