EAG: Extract and Generate Multi-way Aligned Corpus for Complete Multi-lingual Neural Machine Translation

Yulin Xu[1]*, Zhen Yang[1]*, Fandong Meng[1], and Jie Zhou[1]
[1]Pattern Recognition Center, WeChat AI, Tencent Inc, China
{xuyulincs}@gmail.com
{zieenyang,fandongmeng,withtomzhou}@tencent.com

Abstract

Complete Multi-lingual Neural Machine Translation (C-MNMT) achieves superior performance against the conventional MNMT by constructing multi-way aligned corpus, i.e., aligning bilingual training examples from different language pairs when either their source or target sides are identical. However, since exactly identical sentences from different language pairs are scarce, the power of the multi-way aligned corpus is limited by its scale. To handle this problem, this paper proposes "Extract and Generate" (EAG), a two-step approach to construct large-scale and high-quality multi-way aligned corpus from bilingual data. Specifically, we first extract candidate aligned examples by pairing the bilingual examples from different language pairs with highly similar source or target sentences; and then generate the final aligned examples from the candidates with a well-trained generation model. With this two-step pipeline, EAG can construct a large-scale and multi-way aligned corpus whose diversity is almost identical to the original bilingual corpus. Experiments on two publicly available datasets i.e., WMT-5 and OPUS-100, show that the proposed method achieves significant improvements over strong baselines, with +1.1 and +1.4 BLEU points improvements on the two datasets respectively.

1 Introduction

Multilingual Neural Machine Translation (MMMT) (Dong et al., 2015; Firat et al., 2017; Johnson et al., 2017; Aharoni et al., 2019) has achieved promising results on serving translations between multiple language pairs with one model. With sharing parameters of the model, MNMT can facilitate information sharing between similar languages and make it possible to translate between low-resource and zero-shot language pairs. Since the majority of available MT training data are English-centric, i.e., English either as the source or target language, most non-English language pairs do not see a single training example when training MNMT models (Freitag and Firat, 2020). Therefore, the performance of MNMT models on non-English translation directions still left much to be desired: 1) Lack of training data leads to lower performance for non-English language pairs(Zhang et al., 2021); 2) MNMT models cannot beat the pivot-based baseline systems which translate non-English language pairs by bridging through English (Cheng et al., 2016; Habash and Hu, 2009).

Recently, Freitag and Firat (2020) re-kindle the flame by proposing C-MNMT, which trains the model on the constructed multi-way aligned corpus. Specifically, they extract the multi-way aligned examples by aligning training examples from different language pairs when either their source or target sides are identical (i.e., pivoting through English, for German→English and English→French to extract German-French-English examples). Since they directly extract the multi-way aligned examples from the bilingual corpus, we refer to their approach as the *extraction-based* approach. Despite improving the performance, the scale of multi-way aligned corpus extracted by Freitag and Firat (2020) is always limited compared to English-centric bilingual corpus, e.g., only 0.3M German-Russian-English multi-way aligned corpus extracted from 4.5M German-English and 33.5M English-Russian bilingual corpus. A simple idea for remedying this problem is to add the roughly-aligned corpus by extracting the training examples when either their source or target sides are highly similar. However, our preliminary experiments show that the performance of the model decreases dramatically when we train the model with appending the roughly-aligned corpus.[1] One possible solution, referred

*Equal contribution. Work was done when Yulin Xu was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.

[1]Detailed descriptions about the preliminary experiment

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8141 - 8153

to as the *generation-based* approach, is to generate the multi-way aligned examples by distilling the knowledge of the existing NMT model, e.g., extracting German-English-French synthetic three-way aligned data by feeding the English-side sentences of German-English bilingual corpus into the English-French translation model. Although the *generation-based* approach can theoretically generate non-English corpus with the same size as original bilingual corpus, its generated corpus has very low diversity as the search space of the beam search used by NMT is too narrow to extract diverse translations (Wu et al., 2020; Sun et al., 2020; Shen et al., 2019), which severely limits the power of the *generation-based* approach.

In order to combine advantages of the two branches of approaches mentioned above, we propose a novel two-step approach, named EAG (Extract and Generate), to construct large-scale and high-quality multi-way aligned corpus for C-MNMT. Specifically, we first extract candidate aligned training examples from different language pairs when either their source or target sides are highly similar; and then we generate the final aligned examples from the pre-extracted candidates with a well-trained generation model. The motivation behind EAG is two-fold: 1) Although identical source or target sentences between bilingual examples from different language pairs are scarce, highly similar sentences in source or target side are more wide-spread; 2) Based on the pre-extracted candidate aligned examples which have highly similar source or target sentences, EAG can generate the final aligned examples by only refining the sentences partly with a few modifications. Therefore, the non-English corpus constructed by EAG has almost identical diversity to the original bilingual corpus. Experiments on the publicly available data sets, i.e., WMT-5 and OPUS-100, show that the proposed method achieves substantial improvements over strong baselines.

2 Background

Bilingual NMT Neural machine translation (Sutskever et al., 2014; Cho et al., 2014; Vaswani et al., 2017) achieves great success in recent years due to its end-to-end learning approach and large-scale bilingual corpus. Given a set of sentence pairs $D = \{(x, y) \in (X \times Y)\}$, the NMT model is trained to learn the parameter θ by maximizing the log-likelihood $\sum_{(x,y)\in D} logP(y|x;\theta)$.

MNMT Considering training a separate model for each language pair is resource consuming, MNMT (Dong et al., 2015; Johnson et al., 2017; Gu et al., 2020) is introduced to translate between multiple language pairs using a single model (Johnson et al., 2017; Ha et al.; Lakew et al., 2018). We mainly focus on the mainstream MNMT model proposed by Johnson et al. (2017), which only introduces an artificial token to the input sequence to indicate which target language to translate.

C-MNMT C-MNMT is proposed to build a complete translation graph for MNMT, which contains training examples for each language pair (Freitag and Firat, 2020). A challenging task remaining is how to get direct training data for non-English language pairs. In Freitag and Firat (2020), non-English training examples are constructed by pairing the non-English sides of two training examples with identical English sides. However, this method can't get large-scale training examples since the quantity of exactly identical English sentences from different language pairs is small. Another feasible solution is to generate training examples with pivot-based translation where the source sentence cascades through the pre-trained source $\rightarrow$ English and English $\rightarrow$ target systems to generate the target sentence (Cheng et al., 2016). Despite a large quantity of corpus it can generate, its generated corpus has very low diversity (Wu et al., 2020; Sun et al., 2020; Shen et al., 2019).

3 Methods

The proposed EAG has a two-step pipeline. The first step is to extract the candidate aligned examples from the English-centric bilingual corpus. The second step is to generate the final aligned examples from the candidates extracted in the first step.

3.1 Extract candidate aligned examples

Different from Freitag and Firat (2020) who extract non-English training examples by aligning the English-centric bilingual training examples with identical English sentences, we extract the candidate aligned examples by pairing two English-centric training examples with highly similar English sentences. Various metrics have been proposed to measure the superficial similarity of two sentences, such as TF-IDF (Aizawa, 2003; Huang et al., 2011), edit distance (Xiao et al., 2008; Deng

can be found in Section 5.1.

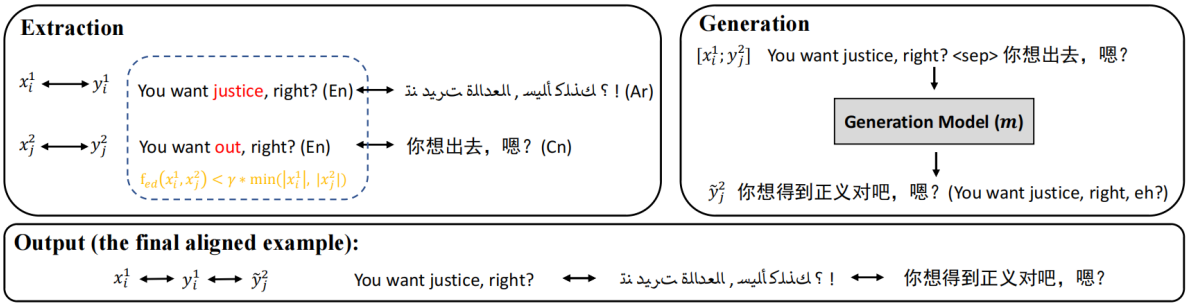

Figure 1: Examples constructed by EAG. "$x_i^1 \leftrightarrow y_i^1$" and "$x_j^2 \leftrightarrow y_j^2$" represent the bilingual examples in English $\rightarrow$ Arabic and English $\rightarrow$ Chinese respectively. $\tilde{y}_j^2$ is the generated Chinese sentence, which is aligned to x_i^1 and y_i^1. For a clear presentation, the Google translation (in English) for the generated $\tilde{y}_j^2$ is also provided.

et al., 2013), etc. In this paper, we take edit distance as the measurement to decide the superficial similarity of two English sentences. Three main considerations are behind. Firstly, since edit distance measures the similarity of two sentences with the minimum number of operations to transform one into the other, it tends to extract sentences with similar word compositions and sentence structures. Secondly, since edit distance only utilizes three operations, i.e., removal, insertion, or substitution, it is easier to mimic these operations in the process of generating the final aligned examples (we leave the explanation in the next subsection). Finally, unlike TF-IDF which only considers word bags in two sentences, edit distance also considers the word order in each sentence.

Formally, given two English-centric bilingual corpora from two different language pairs $\{X^1, Y^1\}$ and $\{X^2, Y^2\}$, where X^1 and X^2 are English sides, Y^1 and Y^2 belong to language L_a and L_b respectively. For sentence pair $(x_i^1, y_i^1) \in \{X^1, Y^1\}$ and $(x_j^2, y_j^2) \in \{X^2, Y^2\}$, we take $(x_i^1, y_i^1, x_j^2, y_j^2)$ as a candidate aligned example if the two English sentences x_i^1 and x_j^2 meets:

$$f_{ed}(x_i^1, x_j^2) \leq \gamma * min(|x_i^1|, |x_j^2|), \gamma \in (0, 1) \quad (1)$$

where f_{ed} refers to the function of edit distance calculation, $|x|$ represents the length of the sentence x, γ is the similarity threshold which can be set by users beforehand to control the similarity of sentences in the candidate aligned examples. With setting $\gamma = 0$, we can directly extract the same multi-way aligned examples with Freitag and Firat (2020). With larger γ, more candidate aligned examples can be extracted for looser restriction. Accordingly, there are more noises in the extracted candidate aligned examples.

3.2 Generate final aligned examples

In the extracted candidate aligned example $(x_i^1, y_i^1, x_j^2, y_j^2)$, (x_j^2, y_j^2) is not well aligned to (x_i^1, y_i^1) if $f_{ed}(x_i^1, x_j^2)$ does not equal to zero. To construct the final three-way aligned example, we search for one sentence pair $(\tilde{x}_j^2, \tilde{y}_j^2)$ in the language pair $\{X^2, Y^2\}$, where $\tilde{x}_j^2$ has the same meaning to x_i^1 (thus $(x_i^1, y_i^1, \tilde{y}_j^2)$ is a three-way aligned example). Unfortunately, it is very difficult for us to directly find such a sentence pair in the large search space. However, considering $\tilde{x}_j^2$ and x_i^1 are both in English, we can take an extreme case where $\tilde{x}_j^2$ is identical to x_i^1 in the superficial form. Now, the remained question is that we need to search for the sentence $\tilde{y}_j^2$ in language L_b, which has the same meaning to x_i^1. By comparing $(x_i^1, \tilde{y}_j^2)$ with (x_j^2, y_j^2), as x_i^1 can be transformed from x_j^2 with the operations performed by edit distance, it is naturally to suppose that we can find such a $\tilde{y}_j^2$ which can be transformed from y_j^2 with these operations similarly. Therefore, we can limit the search space for $\tilde{y}_j^2$ with two restrictions: Firstly, sentence $\tilde{y}_j^2$ has the same meaning with x_i^1; Secondly, $\tilde{y}_j^2$ is transformed from y_j^2 with the operations performed by edit distance. Considering the restrictions mentioned above, we apply an NMT model m to search and generate $\tilde{y}_j^2$. There are two main questions left to be resolved: how to train such a model m and how to generate $\tilde{y}_j^2$ with a well-trained m.

Training Motivated by the recent success of self-supervised training (Devlin et al., 2018; Conneau and Lample, 2019; Song et al., 2019; Yang et al., 2020) in natural language processing, we automatically construct the training corpus for m from the candidate aligned examples. Given the candidate aligned example $(x_i^1, y_i^1, x_j^2, y_j^2)$, the training ex-

ample for m is built as:

$$([x_j^2; \hat{y}_j^2], y_j^2) \quad (2)$$

where y_j^2 is the target sentence, the concatenation of x_j^2 and $\hat{y}_j^2$ is the source-side input. $\hat{y}_j^2$ is the noisy form of y_j^2 which we build by mimicking the operations of edit distance, i.e, performing insertion, removal, or substitution on some pieces of y_j^2 randomly. Specifically, with probability β, each position of sentence y_j^2 can be noised by either removed directly, inserted or substituted with any other words in the dictionary W_b, which is constructed from the corpus Y^2. With the self-constructed training examples, the model m is trained to generate the target sentence, which is recovered from the right-side of the concatenated input with the operations performed by edit distance, and has the same meaning to the left-side of the input.

Generating With a well-trained m, we generate the final aligned examples by running the inference step of m. Formally, for the final aligned example $(x_i^1, y_i^1, \tilde{y}_j^2)$, the sentence $\tilde{y}_j^2$ is calculated by:

$$\tilde{y}_j^2 = m([x_i^1; y_j^2]) \quad (3)$$

where $[\cdot;\cdot]$ represents the operation of concatenation, and $m(x)$ refers to running the inference step of m with x fed as input. With this generation process, $\tilde{y}_j^2$ is not only has the same meaning to x_i^1 (thus also aligned to y_i^1), but also keeps the word composition and sentence structure similar to y_j^2. Therefore, EAG can construct the final aligned corpus for each non-English language pair, and keep the diversity of the constructed corpus almost identical to the original English-centric corpus. For a clear presentation, Algorithm 1 in Appendix A.2 summarizes the process of generating the final aligned examples. We also provide a toy example in Figure 1 to illustrate how the proposed EAG works.

4 Experiments and Results

For fair comparison, we evaluate our methods on the publicly available dataset WMT-5, which is used by Freitag and Firat (2020). Additionally, we test the scalability of our method by further conducting experiments on Opus-100, which contains English-centric bilingual data from 100 language pairs (Zhang et al., 2020). In the extraction process, we run our extraction code on the CPU with 24 cores and 200G memory.[2] In the generation process, we take transformer-big (Vaswani et al., 2017) as the configuration for m, and m is trained with the self-constructed examples mentioned in Section 3.2 on eight V100 GPU cards.[3]

We choose Transformer as the basic structure for our model and conduct experiments on two standard configurations, i.e, transformer-base and transformer-big. All models are implemented based on the open-source toolkit fairseq (Ott et al., 2019) and trained on the machine with eight V100 GPU cards.[4] All bilingual models are trained for 300,000 steps and multi-lingual models are trained for 500,000 steps. We add a language token at the beginning of the input sentence to specify the required target language for all of the multi-lingual models. For the hyper-parameters β and γ, we set them as 0.5 and 0.3 by default and also investigate how their values produce effects on the translation performance.

4.1 Experiments on WMT-5

4.1.1 Datasets and pre-processing

Following Freitag and Firat (2020), we take WMT13EnEs, WMT14EnDe, WMT15EnFr, WMT18EnCs and WMT18EnRu as the training data, the multi-way test set released by WMT2013 evaluation campaign (Bojar et al., 2014) as the test set. The size of each bilingual training corpus (the non-English corpus constructed by Freitag and Firat (2020) included) is presented in Table 1. For the bilingual translation task, the source and target languages are jointly tokenized into 32,000 sub-word units with BPE (Sennrich et al., 2016). The multi-lingual models use a vocabulary of 64,000 sub-word units tokenized from the combination of all the training corpus. Similar to Freitag and Firat (2020), we use a temperature-based data sampling strategy to over-sample low-resource language pairs in standard MNMT models and low-resource target-languages in C-MNMT models (temperature $T = 5$ for both cases). We use BLEU scores (Papineni et al., 2002) to measure the model performance and all BLEU scores are calculated with sacreBLEU (Post, 2018).[5]

[2] Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz

[3] Detained training process for m can be found in the Appendix A.1.

[4] We upload the code as supplementary material for review, and it will be released publicly upon publication.

[5] sacreBLEU signatures: BLEU+case.mixed+lang.SRC-TGT+numrefs.1+smooth.exp+tok.intl+version.1.5.1

	cs	de	en	es	fr	ru
cs		0.7	47	0.8	1	0.9
de	0.7		4.5	2.3	2.5	0.3
en	47	4.5		13.1	38.1	33.5
es	0.8	2.3	13.1		10	4.4
fr	1	2.5	38.1	10		4.8
ru	0.9	0.3	33.5	4.4	4.8	

Table 1: WMT: Available training data (in million) after constructing non-English examples by (Freitag and Firat, 2020).

	cs	de	en	es	fr	ru
cs		2.2	47	2.5	4.1	3.2
de	2.2		4.5	7.1	6.1	1.4
en	47	4.5		13.1	38.1	33.5
es	2.5	7.1	13.1		22	10.1
fr	4.1	6.1	38.1	22		11.0
ru	3.2	1.4	33.5	10.1	11.0	

Table 2: WMT: Available training data (in million) after constructing non-English examples by EAG.

4.1.2 Corpus constructed by EAG

Table 2 shows the training data after constructing non-English examples from English-centric corpus by the proposed EAG. By comparing Table 2 with Table 1, we can find that EAG can construct much more multi-way aligned non-English training examples than Freitag and Firat (2020), e.g., EAG constructs 1.4M bilingual training corpus for the language pair German → Russian which is almost up to 4 times more than the corpus extracted by Freitag and Firat (2020). In all, EAG constructs no less than 1M bilingual training examples for each non-English language pair.

4.1.3 Baselines

In order to properly and thoughtfully evaluate the proposed method, we take the following five kinds of baseline systems for comparison:

Bilingual systems (Vaswani et al., 2017) Apart from training bilingual baseline models on the original English-centric WMT data, we also train bilingual models for non-English language pairs on the direct bilingual examples extracted by Freitag and Firat (2020).

Standard MNMT systems (Johnson et al., 2017) We train a standard multi-lingual NMT model on the original English-centric WMT data.

Bridging (pivoting) systems (Cheng et al., 2016) In the bridging or pivoting system, the source sentence cascades through the pre-trained source → English and English → target systems to generate the target sentence.

Extraction-based C-MNMT systems (Freitag and Firat, 2020) Freitag and Firat (2020) construct the multi-way aligned examples by directly extracting and pairing bilingual examples from different language pairs with identical English sentences.

Generation-based C-MNMT systems The generation-based C-MNMT baselines construct non-English bilingual examples by distilling the knowledge of the system which cascades the source → English and English → target models. Different from the bridging baselines which just feed the test examples into the cascaded system and then measure the performance on the test examples, the generation-based C-MNMT baselines feed the non-English sides of the bilingual training examples into the cascaded systems and then get the non-English bilingual training examples by pairing the inputs and outputs. The combination of the generated non-English corpus and original English-centric corpus is used to train the C-MNMT model.

4.1.4 Results

We first report the results of our implementations and then present the comparisons with previous works. In our implementations, we take the transformer-base as the basic model structure since it takes less time and computing resources for training. To make a fair comparison with previous works, we conduct experiments on transformer-big which is used by baseline models.

Results of our implementation Table 3 shows the results of our implemented systems. Apart from the average performance of the translation directions from each language to others, we also report the average performance on the English-centric and non-English language pairs.[6] As shown in Table 3, we can find that the proposed EAG achieves better performance than all of the baseline systems. Compared to the extraction-based C-MNMT, the proposed method achieves an improvement up to 1.1 BLEU points on non-English language pairs.

[6]Readers can find the detailed results for each language pair in the Appendix A.3.

System	En-X	De-X	Fr-X	Ru-X	Es-X	Cs-X	English-centric	non-English
bilingual system (Vaswani et al., 2017)	28.8	22.4	24.6	19.6	25.9	21.1	30.3	20.4
MNMT system (Johnson et al., 2017)	28.6	16	15.4	14.2	15.0	20.3	30.0	12.3
pivot system (Cheng et al., 2016)	28.8	25.7	26.1	25.2	27.4	26.4	30.3	24.7
generation-based C-MNMT	28.8	26.2	26.7	26.0	27.5	26.8	30.3	25.3
extraction-based C-MNMT (Freitag and Firat, 2020)	29.3	27.2	27.3	26.7	28.6	28.2	30.6	26.8
EAG	**29.6***	**28.3***	**28.2***	**27.7***	**29.5***	**29.7***	**30.8**	**27.9***

Table 3: The translation performance for different systems on WMT data. 'L-X' means the set of translation directions from language L to other five languages. 'English-centric' and 'non-English' refer to the set for English-centirc and non-English language pairs respectively. For each set, bold indicates the highest value, and $*$ means the gains are statistically significant with $p < 0.05$ compared with extraction-based C-MNMT.

The generation-based C-MNMT performs worse than the extraction-based one even if it generates much larger corpus. Since there is no any training example for non-English language pairs in standard MNMT, the standard MNMT system achieves inferior performance to the pivot and bilingual systems on non-English translation directions. However, with the constructed non-English training examples, EAG achieves 3.2 and 7.5 BLEU points improvements compared with the pivot and bilingual systems respectively.

source \ target	cs	de	en	es	fr	ru
cs		27.6 +1.8	31.9 -0.1	31.6 +1.5	33.8 +2.4	28.4 +1.5
de	25.8 +1.9		31.4 +0.2	31.2 +1.3	33.3 +1.5	25.1 +1.7
en	26.7 -0.2	27.4 +0.3		35.1 +0.1	36.0 +0.5	26.6 +0.2
es	25.9 +1.0	26.4 +0.7	35.1 +0.2		36.8 +0.8	25.6 +0.7
fr	24.9 +1.2	26.5 +1.3	34.6 +0.2	33.8 +0.5		23.7 +0.2
ru	24.9 +0.6	25.1 +2.4	30.1 +0.3	30.4 +1.8	31.0 +0.9	

Table 4: Results for transformer-big trained on corpus constructed by EAG. The small numbers are the difference with respect to Freitag and Firat (2020).

Results compared with previous works Table 4 shows the results of the proposed EAG. We can find that the proposed EAG surpasses Freitag and Firat (2020) almost on all of the translation directions, and achieves an improvement with up to 2.4 BLEU points on the Russian-to-German direction.

4.2 Experiments on Opus-100

Datasets and pre-processing Zhang et al. (2020) first create the corpus of Opus-100 by sampling from the OPUS collection (Tiedemann, 2012). Opus-100 is an English-centric dataset which contains 100 languages on both sides and up to 1M training pairs for each language pair. To evaluate the performance of non-English language pairs, Zhang et al. (2020) sample 2000 sentence pairs of test data for each of the 15 pairings of Arabic, Chinese, Dutch, French, German, and Russian. Following Zhang et al. (2020), we report the sacreBLEU on the average of the 15 non-English language pairs.[7] The statistics about the non-English corpus constructed by Freitag and Firat (2020) and EAG are presented in Table 5. We can find that EAG is able to construct much more bilingual corpus for non-English language pairs (almost nine times more than Freitag and Firat (2020) for each language pair). We use a vocabulary of 64,000 sub-word units for all of the multi-lingual models, which is tokenized from the combination of all the training corpus with SentencePiece.

	Freitag and Firat (2020)	EAG
Ar-X	0.19	1.12
De-X	0.14	1.01
Fr-X	0.15	1.25
Cn-X	0.16	1.03
Ru-X	0.18	0.94
Nl-X	0.18	0.98

Table 5: The amount of non-English examples (in million) constructed from bilingual examples in OPUS-100. "L-X" means the total number of the corpus for the directions from language L to other five.

Results Apart from the baselines mentioned above, we also compare with other two systems proposed by Zhang et al. (2020) and Fan et al. (2020). Zhang et al. (2020) propose the online

[7] Signature: BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.1

System	non-English
MNMT system	4.5
pivot system	13.1
generation-based C-MNMT	13.8
extraction-based C-MNMT	16.5
Zhang et al. (2020)	14.1
Fan et al. (2020)	**18.4**
EAG	17.9*

Table 6: Results on Opus-100. We directly cite their results for Zhang et al. (2020) and Fan et al. (2020). $*$ means the gains of EAG are significant compared with extraction-based C-MNMT ($p < 0.05$)

back-translation for improving the performance of non-English language pairs in Opus-100. Fan et al. (2020) build a C-MNMT model, named $m2m_{100}$, which is trained on 7.5B training examples built in house. Following Zhang et al. (2020), we take the transformer-base as the basic model structure for the experiments and results are reported in Table 6. We can find that EAG achieves comparable performance to Fan et al. (2020) which utilizes much more data than ours. This is not a fair comparison as the data used Fan et al. (2020) is 75 times as much as ours. Additionally, our model surpasses all other baseline systems and achieves +1.4 BLEU points improvement compared with the extraction-based C-MNMT model.

5 Analysis

We analyze the proposed method on Opus-100 and take the transformer-base as the model structure.

5.1 Effects of the hyper-parameters

The similarity threshold γ and the noise ratio β are important hyper-parameters in EAG. In this section, we want to test how these two hyper-parameters affect the final translation performance and how they work with each other. We investigate this problem by studying the translation performance with different γ and β, where we vary γ and β from 0 to 0.7 with the interval 0.2. We report the average BLEU score for the translation directions from Arabic to other five languages on the development sets built in house. Figure 2 shows the experimental results. With $\beta = 0$, it means that the generation process is not applied and we directly train the NMT model with the extracted roughly aligned examples. And this is the setting of our motivated experiments mentioned in Section 1. We can find that, the final performance drops sharply when we directly train the model with the roughly aligned sentence pairs. For each curve in Figure 2, we can find that the model achieves the best performance when the γ is around β, and then the performance decreases with γ growing. A relatively unexpected result is that the model usually achieves the best performance when $\beta = 0.5$ rather than when $\beta = 0.7$ (with a larger β, m is trained to handle more complex noise). We conjecture the main reason is that the noise in the training data when $\beta = 0.7$ is beyond the capacity of m, which makes m converge poorly during training. Overall, with β set as 0.5 and γ set as 0.3, the model achieves the best performance.

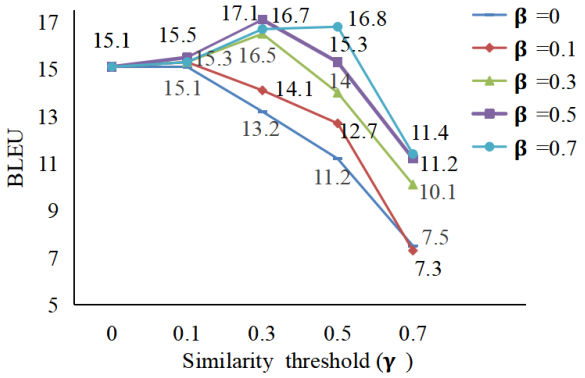

Figure 2: Experimental results on testing the hyper-parameters. With γ set as 0, this is the setting of the baseline system of Freitag and Firat (2020).

#	system	BLEU
0	**EAG**	**17.9**
1	w/o m	15.7
2	w/ m as Transformer-big	18.1
3	w/ m as Transformer-deep	17.8

Table 7: Results on the ability of m.

5.2 The ability of m

We test how the ability of m affects the final performance. Apart from the Transformer-base, i.e., the default setting used in EAG, we also test other two settings, namely Transformer-big (Vaswani et al., 2017) and Transformer-deep (20 encoder layers and 4 decoder layers). With different settings, m is expected to perform different abilities in the generation process. The experimental results are presented in Table 7. We can find that if we remove m, the final performance drops dramatically (comparing #0 with #1). This shows that the generation step

x^1- y^1	Did you have anything to do with Bobby Jordan. ↔ هل فعلت له شيء (نادروج بيباب) ؟
x^2- y^2	Did you have anything to do with it? ↔ 你跟这事有关吗
Generated $\tilde{y}^2$	你跟鲍比·乔丹有关吗？ (Are you related to Bobby Jordan?)
x^1- y^1	You want justice, right? ↔ تريد العدالة ، أليس كذلك ؟ !
x^2- y^2	You want out, right? ↔ 你想出去，嗯？
Generated $\tilde{y}^2$	你想得到正义对吧，嗯？ (You want justice right, eh?)
x^1- y^1	Item 56 of the provisional agenda* ↔ البند 56 من جدول الأعمال المؤقت*
x^2- y^2	Item 100 of the provisional agenda* ↔ 临时议程项目 100
Generated $\tilde{y}^2$	临时议程项目 56 (Provisional agenda item 56)

Figure 3: Examples constructed by EAG. "x^1-y^1" and "x^2-y^2" represent the bilingual examples in English → Arabic and English → Chinese respectively. $\tilde{y}^2$ is the generated Chinese sentence. The words in red color are the different word compositions between x^1 and x^2. The Google translations (in English) for $\tilde{y}^2$ are also provided.

plays a significant role in the proposed EAG. However, by comparing among #0, #2 and #3, we can find that the ability for m shows little effect on the final performance. Taking all of #0, #1, #2 and #3 into consideration, we can reach a conclusion that, the generation step is very important for EAG and a simple generation model, i.e., a baseline NMT model, is enough to achieve strong performance.

System	w/o BT	w/ BT
extraction-based C-MNMT	16.5	17.6
EAG	**17.9**	**18.8**

Table 8: Results on back-translation. We report the average BLEU score on the non-English language pairs.

5.3 Back-translation

We are very curious about how EAG works with back-translation (BT). To investigate this problem, we utilize the extraction-based C-MNMT model to decode the non-English monolingual sentences in the candidate aligned examples extracted by EAG, and then get the synthetic non-English sentence pairs by pairing the decoded results with the original sentences. The reversed sentence pairs are appended into the training corpus for the MNMT models. The experimental results are presented in Table 8. We find that BT improves the performances of both the two systems. Additionally, BT can work as a complementary to the proposed EAG.

5.4 Case study and weaknesses

Figure 3 presents some examples constructed by EAG, each of which includes the extracted candidate aligned example and the generated sentence for Arabic → Chinese. The extracted candidate aligned example contains two bilingual examples, which are extracted from Arabic → English and Chinese → English respectively. In Figure 3, the two bilingual examples in case one are extracted as a candidate aligned example as their English sentences have high similarity. And there is a composition gap between x^1 and x^2 since "Bobby Jordan" is mentioned in x^1, but not in x^2. By comparing the generated Chinese sentence $\tilde{y}^2$ with the original sentence y^2, we can find that $\tilde{y}^2$ is modified from y^2 by inserting the Chinese words " 鲍比乔丹", which has the same meaning with "Bobby Jordan". Therefore, the generated $\tilde{y}^2$ is aligned to x^1 and y^1. In case 2, the Chinese word "出去 (out)" in y^2 has been replaced with Chinese words "得到正义 (justice)" in $\tilde{y}^2$, which makes the $\tilde{y}^2$ aligned to x^1 and y^1. Case 3 in Figure 3 behaves similarly.

While achieving promising performance, the proposed approach still has some weaknesses in the real application: 1) The two-step pipeline performed by EAG is somewhat time-consuming compared to Freitag and Firat (2020); 2) The generated multi-way aligned examples by EAG are sometimes not strictly aligned as the generation process does not always perform perfectly.

6 Conclusion and Future work

In this paper, we propose a two-step approach, i.e., EAG, to construct large-scale and high-quality multi-way aligned corpus from English-centric bilingual data. To verify the effectiveness of the proposed method, we conduct extensive experiments on two publicly available corpora, WMT-5 and Opus-100. Experimental results show that the proposed method achieves substantial improvements over strong baselines consistently. There are three promising directions for the future work. Firstly, we plan to test whether EAG is applicable to the domain adaptation problem in NMT. Sec-

ondly, we are interested in applying EAG to other related tasks which need to align different example pairs. Finally, we want to investigate other model structures for the generation process.

Acknowledgments

The authors would like to thank the anonymous reviewers of this paper, and the anonymous reviewers of the previous version for their valuable comments and suggestions to improve our work.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884.

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2016. Neural machine translation with pivot languages. *CoRR*, abs/1611.04928.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069.

Dong Deng, Guoliang Li, Jianhua Feng, and Wen-Syan Li. 2013. Top-k string similarity search with edit-distance constraints. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 925–936. IEEE.

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

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 (Volume 1: Long Papers)*, pages 1723–1732.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T Yarman Vural, and Yoshua Bengio. 2017. Multi-way, multilingual neural machine translation. *Computer Speech & Language*, 45:236–252.

Markus Freitag and Orhan Firat. 2020. Complete multilingual neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 550–560. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2020. Improved zero-shot neural machine translation via ignoring spurious correlations. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 1258–1268. Association for Computational Linguistics (ACL).

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. Toward multilingual neural machine translation with universal encoder and decoder. *Institute for Anthropomatics and Robotics*, 2(10.12):16.

Nizar Habash and Jun Hu. 2009. Improving arabic-chinese statistical machine translation using english as pivot language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181.

Cheng-Hui Huang, Jian Yin, and Fang Hou. 2011. A text similarity measurement combining word semantic information with tf-idf method. *Jisuanji Xuebao(Chinese Journal of Computers)*, 34(5):856–864.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

SM Lakew, M Cettolo, and M Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *27th International Conference on Computational Linguistics (COLING)*, pages 641–652.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *International Conference on Machine Learning*, pages 5719–5728. PMLR.

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

Zewei Sun, Shujian Huang, Hao-Ran Wei, Xin-yu Dai, and Jiajun Chen. 2020. Generating diverse translation by manipulating multi-head attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8976–8983.

I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. *NIPS*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

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 *NIPS*.

Xuanfu Wu, Yang Feng, and Chenze Shao. 2020. Generating diverse translation from model distribution with dropout. *arXiv preprint arXiv:2010.08178*.

Chuan Xiao, Wei Wang, and Xuemin Lin. 2008. Ed-join: an efficient algorithm for similarity joins with edit distance constraints. *Proceedings of the VLDB Endowment*, 1(1):933–944.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. 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.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1628–1639. Association for Computational Linguistics.

Mingliang Zhang, Fandong Meng, Yunhai Tong, and Jie Zhou. 2021. Competence-based curriculum learning for multilingual machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2481–2493, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

A.1 Structure and training process for m

We take transformer-big as the configuration for m. The word embedding dimension, head number, and dropout rate are set as 1024, 16, and 0.3 respectively. The model is trained on 8 V100 GPU cards, with learning rate, max-token, and update-freq set as 0.01, 8192, and 10 respectively.

We train m on the self-constructed examples with early-stopping. For the original example (x^2, y^2), we feed m with the input format "x^2 <sep> $\tilde{y}^2$", where $\tilde{y}^2$ is the noised form of y^2, "<sep>" is a special token utilized to denote the sentence boundary. Similar to the traditional NMNT model, m is trained to predict the original target sentence y^2.

A.2 Algorithm for our approach

The algorithm for the proposed EAG is detailed as A.2.

A.3 Concrete BLEU score for each language pair

In this section, we present the concrete BLEU score for each language pair on the corpus of WMT-5. Table 9, 10, 11, 12, and 13 show the translation performance for the bilingual system, standard MNMT system, pivot system, extraction-based system and the EAG respectively.

Algorithm 1 Generating final aligned corpus: Given the aligned candidates $\{x^1, y^1, x^2, y^2\}$; an NMT model m, noisy probability β, word list W_b; return the final aligned corpus $\{y^1, \tilde{y}^2\}$

procedure NOISING(y_i^2, β, W_b)
    $\hat{y}_i^2 \leftarrow y_i^2$
    for $t \in 0, \ldots, |\hat{y}_i^2| - 1$ **do**
        generate random float $\alpha \in (0, 1)$
        if $\alpha < \beta$ **then**
            perform insertion, removal or substitution on the position t of $\hat{y}_i^2$ based on W_b
    return $\hat{y}_i$
procedure TRAINING(x^2, y^2)
    initialize m randomly
    while not convergence **do**
        for $i \in 1, \ldots, |x^2|$ **do**
            $\hat{y}_i^2 \leftarrow$ NOISING(y_i^2, β)
            train m with the example $([x_i^2; \hat{y}_i^2], y_i^2)$
    return m
procedure GENERATING(x^1, x^2, y^1, y^2, m)
    for $i \in 1, \ldots, |x^1|$ **do**
        get $\tilde{y}_i^2$ by performing the inference step of the well-trained m with input $[x_i^1; y_i^2]$
        get the final aligned example by pairing y_i^1 with $\tilde{y}_i^2$
    return $\{y^1, \tilde{y}^2\}$

	cs	de	en	es	fr	ru
cs		17.1	30.8	21.2	22.4	14.3
de	16.2		30.4	28.0	29.3	8.1
en	25.2	25.9		33.2	35.6	24.5
es	16.3	23.1	35.4		35.1	20.0
fr	15.2	22.7	33.5	33.0		18.7
ru	13.2	7.3	29.0	23.5	25.0	

Table 9: The BLEU score for the bilingual systems on WMT-5.

	cs	de	en	es	fr	ru
cs		19.5	30.5	22.0	20.8	8.9
de	6.1		31.4	17.5	21.0	4.0
en	24.2	27.1		33.2	34.4	24.1
es	4.4	8.3	34.4		19.4	8.9
fr	3.8	10.9	32.5	23.9		6.2
ru	4.5	10.0	29.0	19.2	8.4	

Table 10: The BLEU score for the standard MNMT on WMT-5.

	cs	de	en	es	fr	ru
cs		22.8	30.8	27.1	29.2	22.3
de	21.4		30.4	26.5	29.4	20.8
en	25.2	25.9		33.2	35.6	24.5
es	23.1	23.0	35.4		32.6	22.9
fr	21.7	22.8	33.5	30.5		22.0
ru	21.6	20.4	29.0	27.1	28.0	

Table 11: The BLEU score for the pivot system on WMT-5.

	cs	de	en	es	fr	ru
cs		24.9	30.9	29.3	31.2	26.0
de	23.5		30.0	30.2	31.5	23.3
en	25.6	26.8		34.1	35.2	25.1
es	24.2	25.4	35.0		35.3	24.6
fr	22.9	24.9	34.0	32.5		23.3
ru	24.2	22.6	29.7	28.6	29.8	

Table 12: The BLEU score for the extraction-based C-MNMT on WMT-5.

	cs	de	en	es	fr	ru
cs		26.6	30.5	30.9	32.8	27.6
de	25.0		30.1	30.1	32.5	24.0
en	25.1	26.9		34.7	35.5	25.6
es	24.9	25.7	35.3		36.5	25.2
fr	23.9	25.4	34.2	33.8		23.7
ru	24.7	23.8	30.0	29.8	30.4	

Table 13: The BLEU score for EAG on WMT-5.