Towards Neural Machine Translation with Partially Aligned Corpora

Yining Wang[†‡], **Yang Zhao**[†‡], **Jiajun Zhang**[†‡], **Chengqing Zong**[†‡*] and **Zhengshan Xue**[#]
[†]National Laboratory of Pattern Recognition, CASIA, Beijing, China
[‡]University of Chinese Academy of Sciences, Beijing, China
[*]CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China
[#]Toshiba (China) Co.,Ltd.
{yining.wang, yang.zhao, jjzhang, cqzong}@nlpr.ia.ac.cn
xuezhengshan2@toshiba.com.cn

Abstract

While neural machine translation (NMT) has become the new paradigm, the parameter optimization requires large-scale parallel data which is scarce in many domains and language pairs. In this paper, we address a new translation scenario in which there only exists monolingual corpora and phrase pairs. We propose a new method towards translation with partially aligned sentence pairs which are derived from the phrase pairs and monolingual corpora. To make full use of the partially aligned corpora, we adapt the conventional NMT training method in two aspects. On one hand, different generation strategies are designed for aligned and unaligned target words. On the other hand, a different objective function is designed to model the partially aligned parts. The experiments demonstrate that our method can achieve a relatively good result in such a translation scenario, and tiny bitexts can boost translation quality to a large extent.

1 Introduction

Neural machine translation (NMT) proposed by Kalchbrenner et al.(2013), Sutskever et al.(2014) and Cho et al.(2014) has achieved significant progress in recent years. Different from traditional statistical machine translation(SMT) (Koehn et al., 2003; Chiang, 2005; Liu et al., 2006; Zhai et al., 2012) which contains multiple separately tuned components, NMT builds an end-to-end framework to model the whole translation process. For several language pairs, NMT is reaching significantly better translation performance than SMT (Luong et al., 2015b; Wu et al., 2016).

In general, in order to obtain an NMT model of great translation quality, we usually need large-scale parallel data. Unfortunately, the large-scale parallel data is always insufficient in many domains and language pairs. Without sufficient parallel sentence pairs, NMT tends to learn poor estimates on low-count events.

Figure 1: An example of our partially aligned training data, in which the source sentence and target sentence are not parallel but they include two parallel parts (highlight in blue and red respectively).

Actually, there have been some effective methods to deal with the situation of translating language pairs with limited resource under different scenarios (Johnson et al., 2016; Cheng et al., 2017; Sennrich et al., 2016a; Zhang and Zong, 2016). In this paper, we address a new translation scenario in which we do not have any parallel sentences but have massive monolingual corpora and phrase pairs. The previous methods are hard to be used to learn an NMT model under this situation. In this paper, we propose a novel method to learn an NMT model using only monolingual data and phrase pairs.

Our main idea is that although there does not exist the parallel sentences, we can derive the sentence pairs which are non-parallel but contain the parallel parts (in this paper, we call these sentences as **partially aligned sentences**) with the

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 384–393,
Taipei, Taiwan, November 27 – December 1, 2017

monolingual data and phrase pairs. Then we can utilize these partially aligned sentences to train an NMT model. Figure 1 shows an example of our data. Source sentence and target sentence are not fully aligned but contain two translation fragments: ("外交部发言人", "foreign ministry deputy") and ("在例行的记者招待会上说", "speaking at a regular press"). Intuitively, these kinds of sentence pairs are useful in building an NMT model.

To use these partially aligned sentences, the training method should be different from the original methods which are designed for parallel corpora. In this work, we adapt the conventional NMT training method mainly from two perspectives. On one hand, different generation strategies are designed for aligned and unaligned target words. For aligned words, our method guides the translation process based on both the context of source side and previously predicted words. When generating the unaligned target words , our model only depends on the words previously generated without considering the context of source side. On the other hand, we redesign the objective function so as to emphasize the partially aligned parts in addition to maximizing the log-likelihood of the target sentence.

The contributions of our paper are twofold:

1) Our approach addresses a new translation scenario, where there only exists monolingual data and phrase pairs. We propose a method to train an NMT model under this scenario. The method is simple and easy to implement, which can be used in arbitrary attention-based NMT framework.

2) Empirical experiments on the Chinese-English translation tasks under this scenario show that our method can achieve a relatively good result. Moreover, if we only add a tiny parallel corpus, the method can obtain significant improvements in terms of translation quality.

2 Review of Neural Machine Translation

Our approach can be easily applied to any end-to-end attention-based NMT framework. In this work, we follow the neural machine translation architecture by Bahdanau et al. (2015), which we will summarize in this section.

Given the source sentence $X = \{x_1, x_2, ..., x_{Tx}\}$ and the target sentence $Y = \{y_1, y_2, ..., y_{Ty}\}$. The goal of machine translation is to transform source sentence into the target sentence. The end-to-end NMT framework consists of two recurrent neural networks, which are respectively called encoder and decoder. First, the encoder network encodes the X into context vectors C. Then, the decoder network generates the target translation sentences one word each time based on the context vectors C and target words previously generated. More specifically, that is $p(y_i|y_{<i}, C)$.

In encoding stage, it transforms X into a sequence of vectors $h_{enc} = \{h_1^k, h_2^k, h_3^k, ..., h_T^k\}$ using m stacked LSTM (Hochreiter and Schmidhuber, 1997) layers. Finally, the encoder chooses the hidden states of the top encoder layer as $h_{top} = \{h_1^m, h_2^m, h_3^m, ..., h_T^m\}$ that we will use in attention mechanism to calculate context vector later.

In decoding stage, it generates one target word at a time from conditional probability $p(y_i|y_{<i}, C; \theta)$ also via m stacked LSTM layers parameterized by θ. Supposing we have obtained the context vector, the conditional probability $p(y_i|y_{<i}, C; \theta)$ is calculated as follows:

$$\begin{aligned} p(y_i|y_{<i}, C; \theta) &= p(y_i|y_{<i}, c_i) \\ &= softmax(g(W_{y_i}, z_i^m, c_i)) \end{aligned} \tag{1}$$

Where W_{y_i} is embedding of the target word, z_i^m is current hidden states of top layer in decoder network. Note that the first hidden states of decoder z_0^k are set to the last hidden states of encoder as follows:

$$z_0^k = h_T^k \tag{2}$$

c_i can be computed as a weighted sum of the source-side h_s as follows:

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_i^m \tag{3}$$

Where a_{ij} is alignment probability, which can be calculated in multiple ways (Luong et al., 2015a). In our method, we use a simple single-layer feed forward network. This alignment probability measures how relevant *i*-th context vector of source sentence is in deciding the current symbol in translation. The probability will be further normalized:

$$a_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{Tx} exp(e_{ik}))} \tag{4}$$

A detailed introduction of the encoder-decoder framework is described in Bahdanau et al. (2015). In order to train NMT system, we use parallel data to optimize the network parameters by maximizing the conditional log-likelihood:

$$L(\theta, D) = \frac{1}{N}\sum_{n=1}^{N}\sum_{i}^{T_y} log(p(y_i^{(n)}|p(y_{<i}^{(n)}, X^{(n)}, \theta)) \tag{5}$$

3 NMT with Partially Aligned Data

In §2 we gave a brief description of the attention-based NMT models whose network parameters are trained using parallel sentence pairs. However, in the translation scenario where there only exists monolingual corpora and phrase pairs, the conventional NMT framework is hard to be used in training a model. In this section, we first explain how we actually obtain the partially aligned corpora with aligned positions using phrase pairs and monolingual corpora, then introduce our method to train the NMT models using the partially aligned sentences according to the particular properties of the corpora.

3.1 Constructing partially Aligned Corpora

Assuming there exists abundant phrase pairs and monolingual sentences in source and target languages, we define our approach to extract partially aligned sentence pairs for training.

Given a phrase pair (ph_s, ph_t), ph_s may appear in a source-side monolingual sentence X, and ph_t may appear in a target-side monolingual sentence Y. Then, X and Y are non-parallel but contain the parallel part. We call these kinds of data the partially aligned sentences. In this work, we collect the partially aligned sentences by searching the phrase pairs in both of the source and target monolingual data simultaneously. In order to reduce the time of the searching process, the monolingual training corpora are first split into many parts. Then, we retrieve the source phrase in each part to restrict the source range of partially aligned corpus. With the retrieved results, we can search the final results of the partially aligned sentences pairs easily. In this way we can construct our corpora, in which only one or more phrases are aligned in every sentence pairs. We denote a partially aligned sentence ($X = x_1, x_2, ...x_{Tx}$,$Y = y_1, y_2, ..., y_{Ty}$), in which a set of the phrase pairs aligned with each other. We call these pairs partially aligned part:

$$\begin{aligned} P_x^{(k)} &= x_{k1}, ..., x_{kp} \\ P_y^{(k)} &= y_{k1}, ..., x_{kq} \end{aligned} \tag{6}$$

$P_x^{(k)}$ and $P_y^{(k)}$ are the phrases in the source and target sentences respectively, and they are translation equivalents.

3.2 Model Descriptions

In §3.1, we acquired the partially aligned corpora with the phrase pairs and monolingual sentences. Now, we need to use them to train the NMT model. As the traditional NMT model is designed for the parallel sentences, it is not suitable for partially aligned sentences. Thus we redesign the traditional NMT model as follows. Figure 2 shows the basic framework of our training method. Our model has 4 different parts from conventional NMT model, including initial states, generation process, objective functions and vocabulary size.

3.2.1 Initial States

The first difference is the initial hidden states of decoder. In the conventional NMT model, the initial hidden states of decoder z_0^k is set to the last hidden state in encoder h_{Tx}^k, including initial states, generation process, objective functions and vocabulary size. shown in Eq. (2). For parallel sentences, this setting is reasonable, while for partially aligned sentence, this initial method is inappropriate. The reason is that the hidden state of last word in source sentence is irrelevant to the target sentence, considering the fact that under our scenario, the target side sentences may entirely uncorrelated to the source sentences. Thus, in our model, z_0^k is set to a zero vector as follows:

$$Z_0^k = \mathbf{0} \tag{7}$$

In Figure 2, the hidden state of "<start>" symbol is set to zero vector when y_1 does not belong to parallel part of the sentence pair.

3.2.2 Generation Process

The second difference is the generation process. In the conventional NMT system, the model generates each target word based on the context vector c_i and previously predicted words $y_{<i}$ as shown in Eq. (1). This generation strategy is unsuitable for the partially aligned corpora, since there exists many unaligned target words. Intuitively,

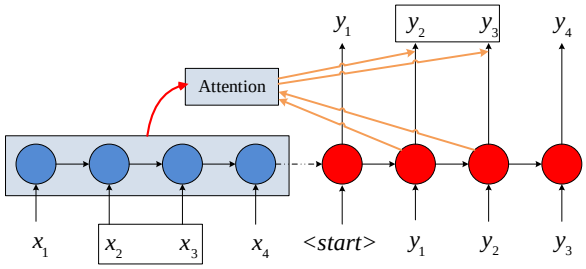

Figure 2: The framework of our training method for partially aligned sentence, in which (x_2,x_3) and (y_2,y_3) are parallel parts.

when the model generates the non-parallel parts, it is unnecessary to take the context vector c_i into consideration. As Figure 2 illustrated, (x_2,x_3) and (y_2,y_3) are parallel parts in this partially aligned sentence pair, and we use context vector which is generated by attention mechanism only when the decoder outputs y_2 and y_3. Therefore, the decoder in our model can be described as follows:

$$c_i = \begin{cases} \sum_j^{Tx} a_{ij} h_i & if\ y_i \in P_y^{(k)} \\ 0 & if\ y_i \notin P_y^{(k)} \end{cases} \quad (8)$$

where a_{ij} is calculated as Eq. (2), $P_y^{(k)}$ is the target partial part, as shown in Eq. (5). In Eq. (8), our model generates the aligned target words based on the context vector c_i and previously predicted words $y_{<i}$. When generating the unaligned target words, the model sets the context vector c_i to zero, indicating that the model generates these words only based on the LSTM-based RNN language model.

3.2.3 Objective Function

Third, we redesign the objective function. Given the parallel data, the objective function is to maximize the log-likelihood of each target word as shown in Eq. (5). For the partially aligned sentence, besides the source and target sentence, we know the phrase alignment information. Hence, apart from maximizing the log-likelihood of each target word, we also hope to make the source and target words in partially aligned part align to each other. As shown in Figure 2, when predicting the words y_2 and y_3, we want to attend more information of corresponding words x_2 and x_3. Thus, we inject an auxiliary object function to achieve it. More specifically, our objective function is designed as follows:

$$L_P(\theta, D) = \frac{1}{N}\sum_{n=1}^{N}\{\sum_i^{T_y} log(p(y_i^{(n)}|p(y_{<i}^{(n)}, X^{(n)}; \theta)) + \lambda \times \triangle(a^{(n)}, \widehat{a}^{(n)}); \theta)\} \quad (9)$$

Where $a^{(n)}$ is defined in Eq. (4), $\triangle$ is a loss function that encourages the agreement between $a^{(n)}$ and $\widehat{a}^{(n)}$. $\widehat{a}^{(n)}$ is the supervised attention determined by alignment relationship between P_X and P_Y, and can be calculated as follows:

$$\widehat{a}_{i,j}^{(n)} = \begin{cases} 1 & if\ x_j \in P_X\ and\ y_i \in P_Y \\ 0 & others \end{cases} \quad (10)$$

$\lambda > 0$ is a hyper-parameter that balances the preference between likelihood and agreement. In this paper, it is set to 0.3.

As shown in Eq. (8), our objective function does not only consider to maximize the log-likelihood of the target sentence, but also encourages the alignment a_{ij} produced by NMT to have a larger agreement with the prior alignment information. This objective function is similar to that used by the supervised attention method (Mi et al., 2016a; Liu et al., 2016). Inspired by Liu et al. (2016), the agreement between $a^{(n)}$ and $\widehat{a}^{(n)}$ can be defined in different ways:

- Multiplication (MUL)

$$\triangle(a^{(n)}, \widehat{a}_{i,j}^{(n)}; \theta) = -\sum_{i=0}^{Ty}\sum_{j=0}^{Tx} a(\theta)_{i,j}^{(n)} \times \widehat{a}_{i,j}^{(n)} \quad (11)$$

  where $\widehat{a}_{i,j}^{(n)}$ is computed by Eq. (10)

- Mean Squared Error (MSE)

$$\triangle(a^{(n)}, \widehat{a}_{i,j}^{(n)}; \theta) = -\sum_{i=0}^{Ty}\sum_{j=0}^{Tx} \frac{1}{2}(a(\theta)_{i,j}^{(n)} - \widehat{a}_{i,j}^{(n)})^2 \quad (12)$$

3.2.4 Limited Vocabulary

The last difference is the vocabulary size during decoding. To make use of phrase pairs as much as possible, we extract a number of special phrase pairs whose source and target are both one word.

In decoding phase, as what Mi et al. (2016b) have done, we can use these special phrase pairs to reduce the vocabulary size when computing the final score distribution. In this way, we can not only acquire more accurate translation of each word, but also accelerate the decoding speed. The vocabulary size can be reduced as follows:

$$V = V_1 \cup V_2 \tag{13}$$

Where V_1 contains the most frequently target words and V_2 is a target words set. This set V_2 is made up of all the target words of the special phrase pairs whose corresponding source words belong to the source sentence.

4 Experiment

In this section, we perform the experiment on Chinese-English translation tasks to test our method.

4.1 Dataset

We evaluate our approach on large-scale monolingual data set from LDC corpus, which includes 13M Chinese sentences and 10M English sentences. Table 1 shows the detailed statistics of our training data. To test our model, we use NIST 2003(MT03) as development set, and NIST 2004-2006(MT04-06) as test set. The evaluation metric is case-insensitive BLEU (Papineni et al., 2002) as calculated by the $multi\text{-}bleu.perl$.

Corpus		Chinese	English
monolingual	#Sent.	13.33M	10.03M
	#Word	327.10M	276.07M
	Vocab	1.83M	1.07M

Table 1: The statistics of monolingual dataset on the LDC corpus.

4.2 Data Preparing and Preprocessing

Considering the fact that the amount of manually annotated phrase pairs is not enough, in order to imitate the environment of experiment, we extract phrase pairs from parallel corpora automatically to make up for the shortage of quantity. To do this, we use Moses (Koehn et al., 2007) in its training step to learn a phrase table from LDC corpus, which includes 0.63M sentence pairs. In order to simulate the experiment as far as possible, we adopt three strategies to filter low quality phrase pairs: 1) the phrases containing the punctuation should be filtered out. (The special phrase parirs introduced in §3.2.4 should be retain) 2) the length of source phrase and target phrase should be greater than 3. 3) only the phrase pairs whose translation probability exceed 0.5 should be retain. In this way, we can get 3M phrase pairs in our experiment. According to our analysis, the average length of phrases are 4.15 and 4.70 on source and target side respectively.

When we search the phrase pairs in monolingual sentences, an obstacle is that one phrase pair will get different source sentences with same target sentence or same source sentence with different target sentences. Therefore, for one phrase pair, we have to restrict the number n of both source sentences and target sentences. To balance the search speed of the phrase pairs in monolingual corpora and the amount of partially aligned sentences, we set the hyper-parameter n to 7. We can search for 5M partially aligned sentences in our experiment. We also calculate the average length ratio of aligned phrases against the whole sentence, which is only 21% and 23% respectively on source and target side.

To ensure the quality of the partially aligned corpora, we also set the number of phrases that aligned to each other in one sentence pair must be greater than a threshold. Here, the threshold is set to 2. That is to say the partially aligned sentence pair should contain at least two aligned phrase pairs.

4.3 Training Details

We build our described method based on the Zoph_RNN toolkit[1] written in C++/CUDA. Both encoder and decoder consist of two stacked LSTM layers. We set minibatch size to 128. The word embedding dimension of both source and target sides is 1000, and the dimensions of hidden layers unit is set to 1000. In our baseline model, we limit the vocabulary of both source and target languages to 30K most frequent words, and other words are replaced by a special symbol "UNK". We run our model on the training corpus 20 iterations in total with stochastic gradient decent algorithm. We set learning rate to 0.1 at the beginning and halve the threshold while the perplexity increases on the development set. Dropout is applied to our model, and the rate is set to 0.2. For

[1] https://github.com/isi-nlp/ZophRNN

#	System	MT03	MT04	MT05	MT06	Ave
1	Phrase NMT Model	3.64	4.25	3.55	3.77	3.80
2	Partially Aligned Model(MUL)	3.80	4.37	3.75	4.24	4.04
3	Partially Aligned Model(MSE)	5.11	5.04	4.26	4.95	4.84
4	Partially Aligned Model(MSE) + LimitedVocab	**6.63**	**6.81**	**5.59**	**5.77**	**6.20**
5	Phrase NMT model + LimitedVocab	3.78	4.33	3.63	3.94	3.92

Table 2: Translation results (BLEU score) for different translation methods.

testing, we employ beam search algorithm, and the beam size is 12.

4.4 Training Methods

We conduct our experiment on the dataset mentioned above, and we list the training methods used as follows:

1) **Phrase NMT Model:** As mentioned above, the only parallel resource is phrase pairs. We use attention-based NMT system to train only on the 3M phrase pairs to get our baseline result.

2) **Partially Aligned Model(MUL):** We train our NMT model using the objective function of multiplication method on the partially aligned sentences.

3) **Partially Aligned Model(MSE):** We train our NMT model using the objective function of Mean Squared Error(MSE) method.

4) **Partially Aligned Model(MSE) + LimitedVocab:** It is similar to Partially Aligned Model(MSE) and the only difference is that we restrict the final score distribution on a limited target vocabulary, which is described in §3.2.4.

5) **Phrase NMT Model + LimitedVocab:** It is the method that LimitedVocab is used in Phrase NMT model.

5 Results and Analysis

5.1 Phrase NMT Model vs. Partially Aligned Method

We present the translation results in BLEU scores of different systems in Table 2. Our first concern is whether the proposed model can actually improve the translation quality. As Table 2 shows, we find that our partially aligned model (both MUL supervised method and MSE supervised method) is superior to the Phrase NMT Model, which indicates that our Partially Aligned method is effective in improving the translation quality.

Source sentence:	布什说：“此项计划将对劳动大众提供减税优惠。”
Phrase NMT model:	The bush says the <UNK> tax concessions be made a possible member
Our model:	Bush said, the scheme will provide tax concessions to the community.

Figure 3: A translation example comparing between phrase NMT model and our partially aligned method.

Figure 3 lists a comparison example of Phrase NMT model and our model. Obviously, our model can achieve the correct translation while the Phrase NMT model generates the unfaithful result. It demonstrates that our model is actually having the ability to learn translation adequacy from aligned parts and fluency from both aligned and unaligned parts.

In §3.2, we described two objective functions in our model. We focus on the difference of these two approaches. As a comparison, MSE method (Row 3) outperforms the MUL method (Row 2) with an average improvement of 0.8 BLEU points, indicating that MSE method is more effective as an object function for our partially aligned model. In the following experiment, we adopt the MSE method as the new objective function used in our model.

5.2 Effect of Limited Vocabulary

In Table 2, an interesting result is that using a reduced vocabulary can significantly improve the performance (+1.36 BLEU points), but it can only achieve 0.12 BLEU points improvement for Phrase NMT model. According to Mi et al. (2016b), this approach is useful in conventional

NMT model. Our result is in agreement with their findings, and the improvement is more prominent in our partially aligned model. Under our scenario, compared to the parallel corpora, fewer parallel parts appear in sentence pairs. The faithfulness of our translation result is relatively poor while the fluency is relatively good. With the limited relevant vocabulary, the faithfulness of the translation results is much improved.

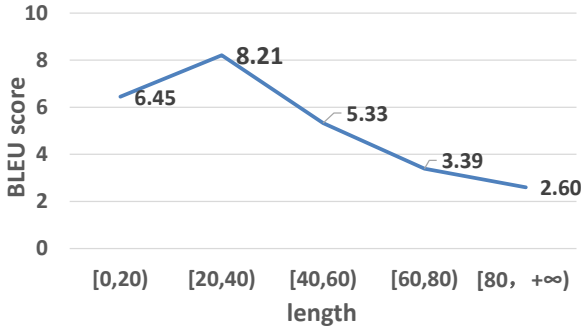

Figure 4: Translation results (BLEU score) of different lengths.

5.3 Result of Different Sentence Lengths

The performance of our partially aligned model with different lengths is another problem we care about. We randomly select 1000 sentences from translation results of test set (MT03-08), which are trained by the Partially Aligned Model(MSE)+ReduceDict method. We classify them into five categories according to the length. Figure 4 shows the results of the experiment.

We find that the sentences with shorter length (lower than 40) yield better results than the long sentences. When the length of sentences exceeds 80, the quality of translation is rather poor. The reason is that we mainly use phrase pairs in our method, and the length of phrase is relative short compared to whole sentence. So our model is more suitable for the translation of short sentences. When we translate long sentences, the parameters in our model are not adjusted for tuning, and our approach can not produce translation of high quality.

5.4 Effect of Adding Small Parallel Corpus

We concern that when we have a tiny parallel corpus, whether the small scale parallel corpus can boost the translation performance of the partially aligned method. Here, we fine-tune the partially aligned translation model on these corpora. The details of these corpora are introduced in Table 3.

Corpus		Chinese	English
parallel	#Sent.	0.1M	
	#Word	3.00M	3.86M
	Vocab	0.07M	0.04M

Table 3: The statistics of small-scale parallel datasets.

The result is presented in Table 4. We can observe that the translation result tends to be poor by only using a small-scale parallel corpora. It indicates that conventional NMT system cannot learn a good model on the small-scale datasets. However, when fine-tuning our partially aligned model with this small parallel corpus, we can get a surprising improvement. The results suggest that when under a scenario in which we only have monolingual corpora and phrase pairs, even a few bitexts can boost translation quality to a large extent.

We investigate the effect of the different corpora size on the final translation results. According to Table 4, when the number of parallel sentences is quite small (lower than 60K), we can acquire a measurable improvement (more than 10 BLEU) compared to the conventional NMT system result. Especially, when the size of sentence pairs is 40K and 60K, our method obtains the enormous improvement over the NMT model by +13.82 BLEU points and +13.21 BLEU points respectively. When using more than 60K sentence pairs, we still get a relatively high promotion of translation quality. However, the promotion is not very remarkable as Row1-3 reveal in Table 4. We can see when the number of parallel corpora is 100K(Row 5), the improvement over NMT Model is +3.95 BLEU points, which indicates that as the size of parallel corpora increases, the improvement of fine-tuning model is decreasing.

6 Related Work

Most of existing work in neural machine translation focus on integrating SMT strategies (He et al., 2016; Zhou et al., 2017; Wang et al., 2017; Shen et al., 2015), handling rare words (Li et al., 2016; Sennrich et al., 2016b; Luong et al., 2015b) and designing the better framework (Tu et al., 2016; Luong et al., 2015a; Meng et al., 2016). As for translation scenarios, training NMT model under

#Sent.	Method	MT03	MT04	MT05	MT06	Ave
20K	**NMT Model**	1.60	1.22	1.05	1.70	1.39
	Partially Aligned Model(MSE) + Para	12.36	15.07	11.64	14.61	13.42
40K	**NMT Model**	1.87	2.00	1.47	2.24	1.90
	Partially Aligned Model(MSE) + Para	14.12	17.84	13.66	17.26	15.72
60K	**NMT Model**	3.72	5.04	3.49	4.47	4.18
	Partially Aligned Model(MSE) + Para	15.54	19.62	15.16	19.25	17.39
80K	**NMT Model**	7.96	11.85	8.16	10.53	9.63
	Partially Aligned Model(MSE) + Para	17.16	21.18	16.65	20.61	18.90
100K	**NMT Model**	14.50	18.21	14.29	17.49	16.12
	Partially Aligned(MSE) Model + Para	18.30	22.50	17.92	21.55	20.07

Table 4: Effect of different data size of parallel corpus. Method NMT Model means the result of conventional NMT system trained on these low-count parallel sentences. Partially Aligned Model(MSE) + Para means the result of our model fine-tuned by these parallel sentences.

different scenarios has drawn intensive attention in recent years. Actually, there have been some effective methods to deal with them. We divide the related work into three categories:

6.1 Pivot-based Scenario

Pivot-based scenario assumes that there only exists source-pivot and pivot-target parallel corpora, which can be used to train source-to-pivot and pivot-to-target translation models. Cheng et al. (2017) propose to translate source language into pivot language, and then the pivot language will be translated into target language. According to the fact that parallel sentences should have close probabilities of generating a sentence in a third language, Chen et al. (2017) construct a Teacher-Student framework, in which existing pivot-target NMT model guides the learning process of the source-target model.

6.2 Multilingual Scenario

In multilingual scenario, there exists multiple language pairs but no source-target sentence pairs. Johnson et al. (2016) use parallel corpora of multiple languages to train a universal NMT model. This universal model learns translation knowledge from multiple different languages, which makes zero-shot translation feasible. Firat et al. (2016) present a multi-way, multilingual model to resolve the zero-resource translation. They use other language to train a multi-way NMT model. The model generates pseudo parallel corpora to fine-tune attention mechanism, so as to achieve better translation.

6.3 Monolingual Data Scenario

In this scenario, an NMT model of good quality has been trained on existing parallel corpora, but a preferable translation result is still in need by incorporating additional data resource. Gülçehre et al. (2015) propose to incorporate target-side corpora as a language model. Sennrich et al. (2016a) attempt to enhance the decoder network model of NMT by incorporating the target-side monolingual data. Luong et at. (2016) explore the sequence autoencoders and skip-thought vectors method to exploit the monolingual data of source language. Zhang and Zong (2016) propose two approaches, self-training algorithm and multi-task learning framework, to incorporate source-side monolingual data. Besides that, Cheng et al. (2016) have explored the usage of both source and target monolingual data using a semi-supervised method to reconstruct both source and target side monolingual language, where two NMT frameworks will be used.

Above methods are designed for different scenarios, and their work can achieve great results on these scenarios. However, when in the scenario we propose in this work, that is we only have monolingual sentences and some phrase pairs, their methods are hard to be utilized to train an NMT model. Under this scenario, monolingual data can be acquired easily, and high quality phrase pairs can be obtained using some effective methods (Zhang et al., 2014). To learn a good NMT

model in our translation scenario, we adapt the conventional training procedure by designing a different generation mechanism and a different objective function.

7 Conclusion

In this paper, we have presented a new translation scenario for NMT in which we have only monolingual data and bilingual phrase pairs. We obtain large-scale partially aligned sentence pairs from the monolingual data and phrase pairs by an information retrieval algorithm. The generation process and objective function are specially designed in NMT training to take full advantage of the partially aligned corpora. The empirical experiments show that the proposed method is capable to achieve a relatively good result. We further find that only a little amount of parallel sentences can significantly boost the translation quality.

We also notice that the proposed approach with only partially aligned data cannot obtain high translation quality. In the future, we plan to design better approaches to model the partially aligned corpus. We also attempt to evaluate our approach on other language pairs, especially low-resource language pairs.

Acknowledgments

The research work has been funded by the Natural Science Foundation of China under Grant No. 61333018 and No. 61402478, and it is also supported by the Strategic Priority Research Program of the CAS under Grant No. XDB02070007

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *In Proceedings of ICLR 2015*.

Yun Chen, Yong Cheng, Yang Liu, and Li Victor, O.K. 2017. A teacher-student framework for zero-resource neural machine translation. *In Proceedings of ACL 2017*.

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2017. Joint training for pivot-based neural machine translation. *In Proceedings of IJCAI 2017*.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. *In Proceedings of ACL 2016*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. *In Proceedings of ACL 2005*.

Kyunghyun Cho, Bart van Merriënboer Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. *In Proceedings of EMNLP 2014*.

Orhan Firat, Baskaran Sankaran, Yaser Alonaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. *In Proceedings of EMNLP 2016*.

Caglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loıc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR, abs/1503.03535*.

Wei He, Zhongjun He, Hua Wu, and Haifeng Wang. 2016. Improved neural machine translation with smt features. *In Proceedings of AAAI 2016*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Vi é gas, Martin Wattenberg, and Greg Corrado. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. *In Proceedings of EMNLP 2013*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. *In Proceedings of ACL 2007*, pages 177–180.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. *In Proceedings of ACL-NAACL 2013*.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. Towards zero unknown word in neural machine translation. *In Proceedings of IJCAI 2016*.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Neural machine translation with supervised attention. *In Proceedings of COLING 2016*.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. *In Proceedings of ACL 2006*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. *In Proceedings of ICLR 2016*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015a. Effective approaches to attention-based neural machine translation. *In Proceedings of EMNLP 2015*.

Minh-Thang Luong, Ilya Sutskever, Quoc V Le, Oriol Vinyals, and Wojciech Zaremba. 2015b. Addressing the rare word problem in neural machine translation. *In Proceedings of ACL 2015*.

Fandong Meng, Zhengdong Lu, Hang Li, and Qun Liu. 2016. Interactive attention for neural machine translation. In *In Proceedings of COLING 2016*.

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016a. Supervised attentions for neural machine translation. *In Proceedings of EMNLP 2016*.

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016b. Vocabulary manipulation for neural machine translation. *In Proceedings of ACL 2016*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *In Proceedings of ACL 2002*, pages 311–318.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. *In Proceedings of ACL 2016*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. *In Proceedings of ACL 2016*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2015. Minimum risk training for neural machine translation. *In Proceedings of ACL 2015*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *In Proceedings of NIPS 2014*.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Coverage-based neural machine translation. *In Proceedings of ACL 2016*.

Xing Wang, Zhengdong Lu, Zhaopeng Tu, Hang Li, Deyi Xiong, and Min Zhang. 2017. Neural machine translation advised by statistical machine translation. *In Proceedings of AAAI 2017*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Feifei Zhai, Jiajun Zhang, Yu Zhou, Chengqing Zong, et al. 2012. Tree-based translation without using parse trees. *In Proceedings of COLING 2012*.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually-constrained phrase embeddings for machine translation. *In Proceedings of ACL 2014*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. *In Proceedings of EMNLP 2016*.

Long Zhou, Wenpeng Hu, Jiajun Zhang, and Chengqing Zong. 2017. Neural system combination for machine translation. *In Proceedings of ACL 2017*.