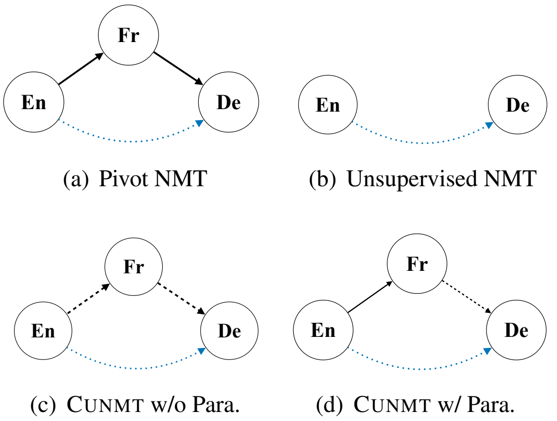
Cross-lingual Supervision Improves Unsupervised Neural Machine Translation

Mingxuan Wang[1] **Hongxiao Bai**[2] **Hai Zhao**[2] **Lei Li**[1]
[1]ByteDance AI Lab, Beijing, China
{wangmingxuan.89,lileilab}@bytedance.com
[2] Department of ComputeScience and Engineering, Shanghai Jiao Tong University
{baippa, zhaohai} @cs.sjtu.edu.cn

Abstract

We propose to improve unsupervised neural machine translation with cross-lingual supervision (CUNMT), which utilizes supervision signals from high resource language pairs to improve the translation of zero-source languages. Specifically, for training En−Ro system without parallel corpus, we can leverage the corpus from En−Fr and En−De to collectively train the translation from one language into many languages under one model. Simple and effective, CUNMT significantly improves the translation quality with a big margin in the benchmark unsupervised translation tasks, and even achieves comparable performance to supervised NMT. In particular, on WMT'14 En-Fr tasks CUNMT achieves 37.6 and 35.18 BLEU score, which is very close to the large scale supervised setting and on WMT'16 En-Ro tasks CUNMT achieves 35.09 BLEU score which is even better than the supervised Transformer baseline.



(a) Pivot NMT (b) Unsupervised NMT

(c) CUNMT w/o Para. (d) CUNMT w/ Para.

Figure 1: Different settings for zero-resource NMT. Full edges indicate the existence of parallel training data. Dashed blue edges indicate the target translation pair. "CUNMT w/o Para." jointly train several unsupervised pairs in one model with unsupervised cross-lingual supervision. "CUNMT w/ Para." train unsupervised directions with supervised cross-lingual supervision, such as jointly train unsupervised En−De with supervised En−Fr.

1 Introduction

Neural machine translation (NMT) has achieved great success and reached satisfactory translation performance for several language pairs (Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017). Such breakthroughs heavily depend on the availability of colossal amounts of bilingual sentence pairs, such as the some 40 million parallel sentence pairs used in the training of WMT14 English French Task. As bilingual sentence pairs are costly to collect, the success of NMT has not been fully duplicated in the vast majority of language pairs, especially for zero-resource languages. Recently, (Artetxe et al., 2018b; Lample et al., 2018a; **?**) tackled this challenge by training unsupervised neural machine translation (UNMT) models using only monolingual data, which achieves considerably high accuracy, but still not on par with that of the state of the art supervised models.

Most previous works focused on modeling the architecture through parameter sharing or proper initialization to improve UNMT. We argue that the drawback of UNMT mainly stems from the lack of supervised signals, and it is beneficial to transfer multilingual information across languages. In this paper, we take a step towards practical unsupervised NMT with cross-lingual supervision (CUNMT) — making the most of the signal from other language. We investigate two variants of multilingual supervision for UNMT. *a)* CUNMT w/o Para.: a general setting where unrelated monolingual data can be introduced. For example, using monolingual Fr data to help the training of En-De (Figure 1(c)). *b)* CUNMT w/ Para.: a relatively strict setting where other bilingual language pairs can be introduced. For example, we can naturally leverage parallel En-Fr data to facilitate the unsupervised En-De transla-

Proceedings of NAACL HLT 2021: IndustryTrack Papers, pages 89–96
June 6–11, 2021. ©2021 Association for Computational Linguistics

tion (Figure 1(d)).

We introduce cross-lingual supervision which aims at modeling explicit translation probabilities across languages. Taking three languages as an example, suppose the target unsupervised direction is $\texttt{En} \rightarrow \texttt{De}$ and the auxiliary language is $\texttt{Fr}$. Our target is to model the translation probability $p(\texttt{De}|\texttt{En})$ with the support of $p(\texttt{Fr}|\texttt{En})$ and $p(\texttt{De}|\texttt{Fr})$. For forward cross-lingual supervision, the system $\text{NMT}_{\texttt{Fr}\rightarrow\texttt{De}}$ serves as a teacher, translating the $\texttt{Fr}$ part of parallel data $(\texttt{En}, \texttt{Fr})$ to $\texttt{De}$. The resulted synthetic data $(\texttt{En}, \texttt{Fr}, \texttt{De})$ can be used to improve our target system $\text{NMT}_{\texttt{En}\rightarrow\texttt{De}}$. For backward cross-lingual supervision, we translate the monolingual $\texttt{De}$ to $\texttt{Fr}$ with $\text{NMT}_{\texttt{De}\rightarrow\texttt{Fr}}$, and then translate $\texttt{Fr}$ to $\texttt{En}$ with $\text{NMT}_{\texttt{Fr}\rightarrow\texttt{En}}$. The resulted synthetic bilingual data $(\texttt{De}, \texttt{En})$ can be used for $\text{NMT}_{\texttt{En}\rightarrow\texttt{De}}$ as well.

Our contributions can be summarized as follow: a) Empirical evaluation of CUNMT on six benchmarks verifies that it surpassed individual MT models by a large margin of more than 3.0 BLEU points on average, and also bested several strong competitors. Particularly, on WMT'16 $\texttt{En}$-$\texttt{Ro}$ tasks, CUNMT surpass the supervised baseline by 0.7 BLEU, showing the great potential for UNMT. b) CUNMT is very effective in the use of additional training data. MBART or MASS introduces billions of sentences, while CUNMT only introduces tens of millions of sentences and achieves super or comparable results. It shows the importance of introducing explicit supervision.

2 The Proposed CUNMT

CUNMT is based on a multilingual machine translation model involving supervised and unsupervised methods with a triangular training structure. The original unsupervised NMT depends only on monolingual corpus, therefore the performances of these translation directions cannot be guaranteed.

Formally, given n different languages L_i, x_i denotes a sentence in language L_i. D_i denotes a monolingual dataset of L_i, and $D_{i,j}$ denotes a parallel dataset of (L_i, L_j). We use $\mathcal{E}$ to indicate the set of all translation directions with parallel data and $\mathcal{W}$ to indicate the set of all unsupervised translation directions respectively. The goal of CUNMT is to minimize the log likelihood of both unsupervised and supervised directions:

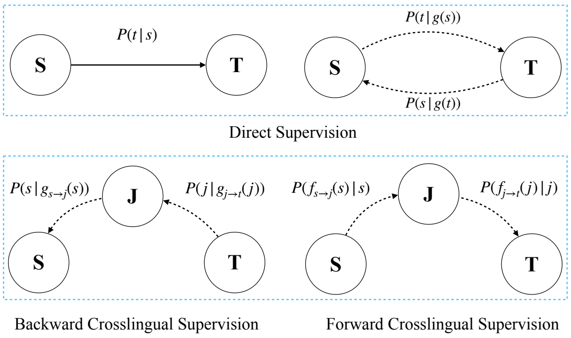

S
T
$P(t\,|\,s)$
S
T
$P(t\,|\,g(s))$
$P(s\,|\,g(t))$
Direct Supervision
$P(s\,|\,g_{s\rightarrow j}(s))$
J
$P(j\,|\,g_{j\rightarrow t}(j))$
S
T
$P(f_{s\rightarrow j}(s)\,|\,s)$
J
$P(f_{j\rightarrow t}(j)\,|\,j)$
S
T
Backward Crosslingual Supervision
Forward Crosslingual Supervision

Figure 2: Forward and backward cross lingual translation for auxiliary data. The dashed blue arrow indicates target unsupervised direction. The solid arrow indicates using the parallel data. The dashed black arrow indicates generating synthetic data.

$$\mathcal{L}^{\text{CUNMT}} = \sum_{i,j \in \mathcal{W}} \mathcal{L}^{U}_{i\rightarrow j} + \sum_{i,j \in \mathcal{E}} \mathcal{L}^{S}_{i\rightarrow j} + \sum_{i,j \in \mathcal{W}+\mathcal{E}} \hat{\mathcal{L}}_{i\rightarrow j} \tag{1}$$

where $\mathcal{L}^{U}_{i\rightarrow j}$ is the unsupervised direct supervision, and $\mathcal{L}^{S}_{i\rightarrow j}$ is the direct supervised supervision, and $\hat{\mathcal{L}}_{i\rightarrow j}$ is the indirect supervision.

2.1 Direct & Cross-lingual Supervision

Direct supervision We will first introduce the notion of direct supervision loss, which only consider the translation probability between two different languages.

For supervised machine translation models, given parallel dataset $D_{s,t}$ with source language L_s and target language L_t, we use $\mathcal{L}^{S}_{s\rightarrow t}$ to denote the supervised training loss from language L_s to language L_t. The training loss for a single sentence can be defined as:

$$\mathcal{L}^{S}_{s\rightarrow t} = \mathbb{E}_{(x_s,x_t)\sim D_{s,t}}[-\log P(x_t|x_s)]. \tag{2}$$

For unsupervised machine translation models, only monolingual dataset D_s and D_t are given. We use $\mathcal{L}^{U}_{s\rightarrow t}$ to denote the unsupervised training loss from language L_s to language L_t. We use $\mathcal{B}_{s\rightarrow t}$ to denote this back translation procedure. After that, we can use these data to train the model with supervised method from L_s to L_t. The losses of the dual structural are:

$$\begin{aligned} \mathcal{L}^{\mathcal{B}}_{t\rightarrow s} &= \mathbb{E}_{x_s\sim D_s}[-\log P(x_s|g_{s\rightarrow t}(x_s))], \\ \mathcal{L}^{\mathcal{B}}_{s\rightarrow t} &= \mathbb{E}_{x_t\sim D_t}[-\log P(x_t|g_{t\rightarrow s}(x_t))], \end{aligned} \tag{3}$$

where $g_{s\rightarrow t}(x_s)$ translate the sentence in language L_s to L_t, that is, the back translation of x_s. Then the total loss of an unsupervised machine translation is:

$$\mathcal{L}^{U} = \mathcal{L}^{\mathcal{B}}_{t\to s} + \mathcal{L}^{\mathcal{B}}_{s\to t}. \tag{4}$$

Cross-lingual supervision When extend to the multilingual scenario, it is natural to introduce indirect supervision across languages. Given n different languages, for each language pair (L_i, L_j), we can easily obtain the translation probability of $P(x_i|x_j)$ through the direct supervised model $\mathcal{L}^S$ or $\mathcal{L}^U$. We use $\hat{\mathcal{L}}_{s\to t}$ to indicate the indirect supervised loss, which can be defined as:

$$\hat{\mathcal{L}}_{s\to t} = \sum_{i=0, i\neq s,t}^{n} \lambda_i \hat{\mathcal{L}}_{s\to i\to t} \tag{5}$$

where λ is the coefficient. T

Due to the lack of triples data (L_i, L_k, L_j), it is difficult to directly estimate the cross translation loss $\hat{\mathcal{L}}_{s\to i\to t}$. We therefor propose the backward and forward indirect supervision to calculate the cross loss:

$$\begin{aligned}\hat{\mathcal{L}}_{s\to j\to t} = \mathbb{E}_{x_t\sim D_t}[-\log P(x_t|g_{t\to j\to s}(x_t))] \\ + \mathbb{E}_{x_s\sim D_s}[-\log P(f_{s\to j\to t}(x_s)|x_s)]\end{aligned} \tag{6}$$

where $g_{t\to j\to s}(x_t)$ is the indirect backward translation which translate x_t to language L_s and $f_{s\to j\to s}(x_t)$ is the indirect forward translation which translate x_s to language L_t.

2.2 Training Procedure of CUNMT

The procedure of CUNMT includes two main steps: multi-lingual pre-training and iterative multi-lingual training.

Multi-lingual Pre-training Due to the ill-posed nature, it is also important to find a good initialization to associate the source side languages and the target side languages. We propose a Multi-lingual Pre-training approach, which jointly train the unsupervised auto-encoder and supervised machine translation. Intuitively, the multi-lingual joint pre-training can take advantage of transfer learning and thus benefit the low resource languages. Apart form the monolingual data, pre-training can also leverage the bilingual parallel data. We suggest the supervised data provides strong signal to optimize the network, which also advantage the unrelated unsupervised NMT pre-training. For example, it is beneficial to use the supervised En-Fr model to initialize the unsupervised De-Fr model.

Indirect Supervised Training The goal is to train a single system that minimize the jointly loss function of $\mathcal{L}^{\text{CUNMT}}$.

Generally, CUNMT can be applied to a restrict unsupervised scenario where only monolingual are provided, and also can be extended to a unrestricted scenario where parallel data are introduced. For the sake of simplicity, we describe our method on three language pairs, which can be easily extended to more language pairs. Suppose that the three languages are denoted as the triad $(\texttt{En}, \texttt{Fr}, \texttt{De})$, and we have monolingual data for all the three languages and also bilingual data for En-Fr. The target is to train an unsupervised En →Fr system. The detailed method is as follows:

1. Sample batch of monolingual $x_{\texttt{En}}, x_{\texttt{Fr}}, x_{\texttt{De}}$ sentences from $D_{\texttt{En}}, D_{\texttt{Fr}}, D_{\texttt{De}}$
2. Sample batch of parallel sentence from $D_{\texttt{En,Fr}}$ to generate supervised data $\mathcal{S}$
3. Back translate $x_{\texttt{En}}, x_{\texttt{Fr}}, x_{\texttt{De}}$ to generate pseudo data $\mathcal{B}$
4. Indirect back translate $x_{\texttt{En}}, x_{\texttt{Fr}}, x_{\texttt{De}}$ to generate pseudo data $\mathcal{B}^i$
5. Indirect forward translate $x_{\texttt{En}}, x_{\texttt{Fr}}, x_{\texttt{De}}$ to generate pseudo data $\mathcal{F}^i$
6. Merge $\mathcal{B}$, $\mathcal{B}^i$, $\mathcal{F}^i$ and $\mathcal{S}$ to jointly train CUNMT.
7. Repeat 1-6 until convergence.

For indirect or direct supervision, we follow the Equation (6), which will adopts one step forward translation if parallel data is provided. Since we train all directions in one model, the pseudo data will include all directions. In this setting, it contains: En ↔ Fr, En ↔ De, Fr ↔ De with both direct and indirect directions.

3 Experiments

3.1 Datasets and Settings

We conduct experiments including $(\texttt{De}, \texttt{En}, \texttt{Fr})$, $(\texttt{Fr}, \texttt{En}, \texttt{De})$, and $(\texttt{Ro}, \texttt{En}, \texttt{Fr})$. For monolingual data of English, French and German, 20 million sentences from available WMT monolingual News Crawl datasets were randomly selected. For Romanian monolingual data, all of the available Romanian sentences from News Crawl dataset were used and and were supplemented with WMT16 monolingual data to yield a total of in 2.9 million sentences. For parallel data, we use the standard WMT 2014 English-French dataset consisting of about 36M sentence pairs, and the

	(Fr,En,De)		(De,En,Fr)		(Ro,En,Fr)	
	En-Fr	Fr-En	En-De	De-En	En-Ro	Ro-En
Supervised Transformer	41.0	-	34.0	38.6	34.3	34.0
Comparison systems of UNMT						
UNMT (Lample et al., 2018c)	25.1	24.2	17.2	21.0	21.2	19.4
EMB (Lample and Conneau, 2019)	29.4	29.4	21.3	27.3	27.5	26.6
MLM (Lample and Conneau, 2019)	33.4	33.3	26.4	34.3	33.3	31.8
MASS (Song et al., 2019)	37.5	34.9	28.3	**35.2**	**35.2**	33.1
MBART (Liu et al., 2020)	-	-	**29.8**	34.0	35.0	30.5
CUNMT						
CUNMT w/o Para.	32.90	31.93	23.03	31.01	33.23	32.34
CUNMT w/ Para.	34.37	32.77	23.99	31.98	33.95	33.15
CUNMT + Forward	35.88	33.64	26.50	33.11	34.12	33.61
CUNMT + Backward + Forward	**37.60**	**35.18**	27.60	34.10	35.09	**33.95**

Table 1: Main results comparisons. MASS uses large scale pre-training and back translation during fine-tuning. MBART employ large scale multi-lingual pretraining with billions sentences. The last four lines are the results of our method.

standard WMT 2014 English-German dataset consisting of about 4.5M sentence pairs. For analyses, we also introduce the standard WMT 2017 English-Chinese dataset consisting of 20M sentence pairs. Consist with previous work, we report results on newstest 2014 for English-French pair, and on newstest 2016 for English-German and English-Romanian.

In the experiments, CUNMT is built upon Transformer models. We use the Transformer with 6 layers, 1024 hidden units, 16 heads. We train our models with the Adam optimizer, a linear warm-up and learning rates varying from 10^{-4} to 5×10^{-4}. The model is trained on 8 NVIDIA V100 GPUs. We implement all our models in PyTorch based on the code of (Lample and Conneau, 2019)[1]. All the results are evaluated on BLEU score with Moses scripts, which is in consist with the previous studies.

3.2 Main Results

The main results of similar pairs are shown in Table 1. We make comparison with three strong unsupervised methods:

- MLM (Lample and Conneau, 2019) uses large scale cross-lingual data to train the mask language model and then fine-tune on unsupervised NMT.
- MASS (Song et al., 2019) is a sequence to sequence model pre-trained with billions of monolingual data.
- MBART (Liu et al., 2020) introduces tens of billions monolingual data to pre-train a deep Transformer model.

CUNMT *is very efficient in the use of multi-lingual data.* While the pretrained language model is obtained through several hundred times larger monolingual or cross-lingual corpus, CUNMT achieves superior or comparable results with much less cost.

The model was improved by using synthetic data of cross translation that is based on the jointly trained model. The results of "CUNMT + Forward" are from the model tuned by only 1 epoch with about 100K sentences. This method is fast and the performances are surprisingly effective. The "CUNMT + Forward + Backward" denotes that, besides forward translation, we also use monolingual data and cross translate it to the source language. This method yielded the best performance by outperforming the "CUNMT w/o Para." by more than 3 BLEU score on average. The improvements show great potential for introducing indirect cross lingual supervision for unsupervised NMT.

When compared with supervised approaches, CUNMT shows very promising performance. For the large scale WMT14 En-Fr tasks, the gap between CUNMT and supervised baseline is closed to 3.4 BLEU score. And for the medium WMT16 En-Ro task, CUNMT performs even better than the supervised approach.

[1] https://github.com/facebookresearch/XLM

4 Analyses

In this part, we conduct several studies on CUNMT to better understand its setting.

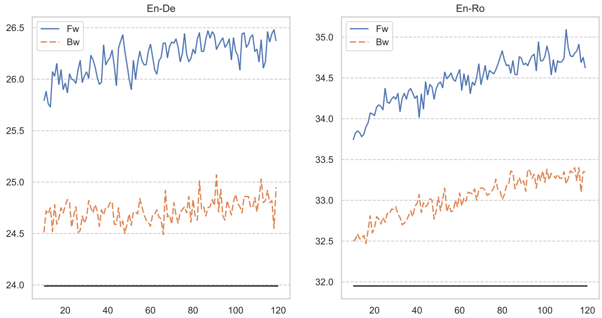


Figure 3: Results comparison for CUNMT fine-tuning with different auxiliary data. “Bw” only adopts cross-lingual backward translation synthetic data, and “Fw” only adopts cross-lingual forward translation synthetic data. The black horizontal is the baseline of UNMT. The horizontal axis is epoch and the vertical axis is the BLEU score. Epoch size is 100K sentences.

Backward or Forward Here we have explored the effect of cross-lingual backward supervision and cross-lingual forward supervision, and plot the performance curves along with the training procedure in Figure 3. The comparison system is CUNMT trained only with monolingual data. To make a fair comparison, we use “CUNMT w/ Para.” as the baseline model and fine-tuning it with only indirect forward supervision or indirect backward supervision. We conduct experiments on WMT16 En-De and En-Ro tasks. Clearly, the forward supervision outperforms the backward one with big margins, which shows the importance of introducing the forward supervision for multilingual UNMT. It is still interesting to find that only introducing the indirect backward translation achieves better results than the unsupervised baseline.

We suppose the reasons for the performance gap is that, *a)* The UNMT baseline has included the traditional direct back translation, therefore the information gain from indirect backward translation is limited compared to the forward translation. *b)* The indirect forward translation provides a more direct way to model the relation across different languages. The results in consist with the previous research that pivot translation can help low resource language translation.

Robustness on Parallel Data Scale As shown in Table 4, CUNMT is robust to the parallel data scale. The results also dovetail with the unsupervised En-Fr experiments in Table 1. As it turns out the smaller parallel data of En-De was able to significantly improve the performance of unsupervised En-Fr translation. We then reduce the scale of the parallel data En-De and surprisingly find that even with only 25% supervised data, CUNMT still works well. The experiments demonstrate that CUNMT is robust and has great potential to be applied to practical systems.

Auxiliary Direction	En-Ro	Ro-En
En-De	34.86	33.18
En-De (50%)	34.72	32.85
En-De (25%)	34.52	32.33

Table 2: Robustness of Parallel Data Scale. Mainly evaluated on unsupervised En-Ro direction with different auxiliary parallel data settings.

Auxiliary Direction	En-Ro	Ro-En
En-Fr	35.09	33.95
En-De	34.86	33.18
En-Zh	33.85	32.86
En-De-Fr	35.26	34.20

Table 3: Effects of the Auxiliary Language. Mainly evaluated on unsupervised En-Ro direction with different parallel data settings.En-Fr,En-De and En-Zh are the auxiliary parallel data for training En-Ro. En-De-Fr is the combination of the En-De and En-Fr parallel data.

Importance of the Auxiliary Language Table 3 shows effects of the auxiliary language. We first switch the parallel data from En-Fr to En-De, the performance is almost consistent. We then switch the parallen data to En − Zh, where Zh is dissimilar with Ro, the performance increases. This is in line with our expectations, that similar languages make it easier for transfer learning. Finally, we extend the parallel data to En-De and En-Fr, and achieves further benefits. Compared with , we suggest the language similarity is more important than the auxiliary data scale.

Benefits as All in One Model In table 4, the performance of supervised directions are shown to illustrate the effects on which jointly training a single system has First, we test the baseline supervised system, that is, only $En \rightarrow Fr$ and $Fr \rightarrow En$ are conducted on the model. Due to difference in model architecture, the performance

System	En-Fr	Fr-En
Supervised Training	39.70	36.62
CUNMT + Forward	39.26	36.82
CUNMT + Backward	39.12	36.20

Table 4: Translation performance on supervised directions of CUNMT.

of CUNMT is slightly lower than that of its state of the art counterparts. Also, some techniques such as model average are not applied, and two directions are trained in one model. In CUNMT, the performance of supervised directions drops a little, but in exchange, the performances of zero-shot directions are greatly improved and the model is convenient to serve for multiple translation directions.

Strategies of Synthetic Data Generation For the synthetic data generation, the reported results are from greedy decoding for time efficiency. We compared the effects of sample strategies on the language setting of (Ro, En, De) where En-De is the supervised direction. The results based on beam search generation for En → Ro is 34.86, and 33.18 for En → Fr in terms of BLEU. Compared with greedy decoding, the performance of beam search is slightly inferior. A possible reason is that the beam search makes the synthetic data further biased on the learned pattern. The results suggest that CUNMT is exceedingly robust to the sampling strategies when performing forward and backward cross translation.

5 Related Work

Multilingual NMT It has been proven low resource machine translation can adopt methods to utilize other rich resource data in order to develop a better system. These methods include multilingual translation system (Firat et al., 2016; Johnson et al., 2017), teacher-student framework (Chen et al., 2017), or others (Zheng et al., 2017). Apart from parallel data as an entry point, many attempts have been made to explore the usefulness of monolingual data, including semi-supervised methods and unsupervised methods which only monolingual data is used. Much work also has been done to attempt to marry monolingual data with supervised data to create a better system, some of which include using small amounts of parallel data and augment the system with monolingual data (Sennrich et al., 2016; He et al., 2016; Wang et al., 2018; Gu et al., 2018; Edunov et al., 2018; Yang et al., 2020). Others also try to utilize parallel data of rich resource language pairs and also monolingual data (Ren et al., 2018; Wang et al., 2019; Al-Shedivat and Parikh, 2019; Lin et al., 2020). (Ren et al., 2018) also proposed a triangular architecture, but their work still relied on parallel data of low resource language pairs. With the joint support of parallel and monolingual data, the performance of a low resource system can be improved.

Unsupervised NMT In 2017, pure unsupervised machine translation method with only monolingual data was proven to be feasible. On the basis of embedding alignment (Artetxe et al., 2017; Lample et al., 2018b), (Lample et al., 2018a) and (Artetxe et al., 2018b) devised similar methods for fully unsupervised machine translation. Considerable work has been done to improve the unsupervised machine translation systems by methods such as statistical machine translation (Lample et al., 2018c; Artetxe et al., 2018a; Ren et al., 2019; Artetxe et al., 2019), pretraining models (Lample and Conneau, 2019; Song et al., 2019), or others (Wu et al., 2019), and all of which greatly improve the performance of unsupervised machine translation.

Our work attempts to utilize both monolingual and parallel data, and combine unsupervised and supervised machine translation through multilingual translation method into a single model CUNMT to ensure better performance for unsupervised language pairs.

6 Conclusion

In this work, we propose a multilingual machine translation framework CUNMT incorporating distant supervision to tackle the challenge of the unsupervised translation task. By mixing different training schemes into one model and utilizing unrelated bilingual corpus, we greatly improve the performance of the unsupervised NMT direction. By joint training, CUNMT can serve all translation directions in one model. Empirically, CUNMT has been proven to deliver substantial improvements over several strong UNMT competitors and even achieve comparable performance to supervised NMT. In the future, we plan to build a universal CUNMT system that is applicable in a wide span of languages.

References

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT), Volume 1 (Long and Short Papers)*, pages 1184–1197, Minneapolis, Minnesota.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3632–3642, Brussels, Belgium.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 194–203, Florence, Italy.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *International Conference on Learning Representations (ICLR)*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 489–500, Brussels, Belgium.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–277, Austin, Texas.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT), Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems (NeurIPS) 29*, pages 820–828.

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

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

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5039–5049, Brussels, Belgium.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. *arXiv preprint arXiv:2010.03142*.

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. *arXiv preprint arXiv:2001.08210*.

Shuo Ren, Wenhu Chen, Shujie Liu, Mu Li, Ming Zhou, and Shuai Ma. 2018. Triangular architecture for rare language translation. In *Proceedings of the*

56th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers), pages 56–65, Melbourne, Australia.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with smt as posterior regularization. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 33:241–248.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

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)*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936, Long Beach, California, USA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS) 30*, pages 5998–6008.

Yijun Wang, Yingce Xia, Li Zhao, Jiang Bian, Tao Qin, Guiquan Liu, and Tie-Yan Liu. 2018. Dual transfer learning for neural machine translation with marginal distribution regularization. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pages 5553–5560, New Orleans, USA.

Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2019. Multi-agent dual learning. In *International Conference on Learning Representations (ICLR)*.

Jiawei Wu, Xin Wang, and William Yang Wang. 2019. Extract and edit: An alternative to back-translation for unsupervised neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT), Volume 1 (Long and Short Papers)*, pages 1173–1183, Minneapolis, Minnesota.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9378–9385.

Hao Zheng, Yong Cheng, and Yang Liu. 2017. Maximum expected likelihood estimation for zero-resource neural machine translation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4251–4257.