Zero-Shot Cross-Lingual Abstractive Sentence Summarization through Teaching Generation and Attention

Xiangyu Duan[1,2]*, Mingming Yin[1]*, Min Zhang[1,2], Boxing Chen[3], Weihua Luo[3]

[1] *Institute of Artificial Intelligence, Soochow University, Suzhou, China*
[2] *School of Computer Science and Technology, Soochow University, Suzhou, China*
[3] *Alibaba DAMO Academy, Hangzhou, China*

xiangyuduan@suda.edu.cn; mmyin@stu.suda.edu.cn; minzhang@suda.edu.cn
{boxing.cbx,weihua.luowh}@alibaba-inc.com

Abstract

Abstractive Sentence Summarization (ASSUM) targets at grasping the core idea of the source sentence and presenting it as the summary. It is extensively studied using statistical models or neural models based on the large-scale monolingual source-summary parallel corpus. But there is no cross-lingual parallel corpus, whose source sentence language is different to the summary language, to directly train a cross-lingual ASSUM system. We propose to solve this zero-shot problem by using resource-rich monolingual ASSUM system to teach zero-shot cross-lingual ASSUM system on both summary word generation and attention. This teaching process is along with a back-translation process which simulates source-summary pairs. Experiments on cross-lingual ASSUM task show that our proposed method is significantly better than pipeline baselines and previous works, and greatly enhances the cross-lingual performances closer to the monolingual performances. We release the code and data at `https://github.com/KelleyYin/Cross-lingual-Summarization`.

1 Introduction

Abstractive Sentence Summarization (ASSUM) is a task of condensing the source sentences into the summaries based on the core meaning of the source sentences. ASSUM provides quick access to the important content of the source sentences through the informative re-written summaries. Major ASSUM explorations are monolingual based. There is an urgent demand of cross-lingual ASSUM which produces summaries for people who do not speak the language the same to the source language.

Unlike the monolingual ASSUM receiving extensive studies that are based on the large-scale monolingual ASSUM corpus, the cross-lingual ASSUM is seldom explored due to the lack of training corpus. This zero-shot challenge drives the cross-lingual ASSUM to resort to two existing independent techniques, i.e., the monolingual ASSUM and the bilingual translation. The both techniques should be leveraged together to overcome the difficulty of data scarcity in the cross-lingual ASSUM.

Regarding the techniques of the monolingual ASSUM, neural methods become dominant in this area since the creation of the large-scale ASSUM corpus (Rush et al., 2015; Nallapati et al., 2016; Hu et al., 2015). The corpus consists of huge number of source-summary pairs, and neural methods model these pairs as as a sequence-to-sequence task by encoding the source sentence into vectorized information and decoding it into the abstractive summary.

Regarding the techniques of the bilingual translation, recent years witnessed the method transition from statistical machine translation (SMT) (Koehn et al., 2003) to neural machine translation (NMT). NMT employs the sequence-to-sequence architecture with various implementations such as RNN-based (Sutskever et al., 2014; Bahdanau et al., 2015), CNN-based (Gehring et al., 2017), and Transformer (Vaswani et al., 2017).

Early works on the cross-lingual ASSUM leverage the above two techniques through using bilingual features to cooperate with the monolingual ASSUM based on the data condition that large-scale monolingual ASSUM corpus is not available while large-scale translation corpora are easy to obtain. They utilize bilingual features such as phrase pairs or predicate-argument parallel structures, which are obtained from SMT systems, to achieve extractive or abstractive cross-lingual summarization (Wan, 2011; Yao et al., 2015; Zhang et al., 2016).

* Equal contribution.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3162–3172
Florence, Italy, July 28 - August 2, 2019.

Recently, Ayana et al. (2018) propose the first large-scale corpus-based cross-lingual ASSUM system in which the ASSUM corpus is monolingual. They generate summaries using the monolingual ASSUM system, and train the cross-lingual ASSUM based on these pseudo summaries.

On the contrary, we propose in this paper to use genuine summaries paired with the generated pseudo sources to train the cross-lingual ASSUM system. We use the teacher-student framework in which the monolingual ASSUM system is taken as the teacher and the cross-lingual ASSUM system is the student. The teacher let the student to simulate both the summary word distribution and attention weights according to those of the teacher networks. In comparison to the pseudo summaries used in the work of Ayana et al. (2018), we generate pseudo sources instead and use true summaries to constitute source-summary pairs. This is motivated by the successful application of back-translation which generates pseudo-source paired with true-target for NMT (Sennrich et al., 2016a; Lample et al., 2018).

The main contributions of this paper include:

- We propose teaching both summary word generation distribution and attention weights in the cross-lingual ASSUM networks by using the monolingual ASSUM networks. The distribution teacher is directly from the monolingual ASSUM, while the attention weights teacher is obtained by an attention relay mechanism.

- We use a back-translation procedure that generates pseudo source sentences paired with the true summaries to build a training corpus for the cross-lingual ASSUM. This alleviates the data scarcity that no cross-lingual ASSUM corpus is available.

- Extensive experimental results on two benchmark datasets show that our proposed method is able to perform better than several baselines and related works, and significantly reduce the performance gap between the cross-lingual ASSUM and the monolingual ASSUM.

2 Related Work

2.1 Monolingual ASSUM

There are various methods exploring the effective way to model the monolingual ASSUM process including statistical models (Banko et al., 2000; Cohn and Lapata, 2008) or neural models (Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016). Neural models become dominant in this task since the creation of the large-scale ASSUM corpus (Rush et al., 2015; Nallapati et al., 2016; Hu et al., 2015). On the basis of the sequence-to-sequence neural architecture, there are many further explorations such as using rich linguistic features and large vocabulary set (Nallapati et al., 2016), global training procedures on the sentence level (Ayana et al., 2016; Li et al., 2018; Edunov et al., 2018; Wang et al., 2018), topic enhancement in the summaries (Wang et al., 2018), additional selective gate networks in the encoder (Zhou et al., 2017), and facts fusion measures (Cao et al., 2018).

2.2 Zero Resource Neural Machine Translation

Current state-of-the-art NMT models are effective in modeling the translation process, but they are highly dependent on the large-scale parallel corpus. When applied on zero resource language pairs such as the two languages that do not have direct parallel corpus, the NMT systems perform well below the satisfactory level. To address such problem, three NMT paradigms are explored. The first is the triangular NMT systems that add one additional resource rich language to the zero resource language pair to build a triangular translation scenario (Chen et al., 2017; Zheng et al., 2017; Cheng et al., 2017), the second is the multilingual translation system that concatenates parallel corpora of different language pairs and builds one NMT model for all (Johnson et al., 2017), the third is the unsupervised NMT systems that do not use any parallel data resources (Artetxe et al., 2018; Lample et al., 2018).

Our work is closely related to the first paradigm in which source language, pivot language, and target language form a triangular translation scenario. In our setting, the target language {sentence, summary} pair and the source language sentence form the triangle in which the target language sentence functions as the pivot. We adopt the teacher-student framework that is also applied in Chen et al. (2017), but we have significant difference to them in that we generate pseudo source while they generate pseudo target, which results in different teacher-student networks.

2.3 Cross-lingual Summarization

Early explorations on cross-lingual summarization mainly depend on the traditional monolingual summarization methods, and integrate bilingual parallel informations into the monolingual methods through sentence selection based on translation quality estimation (Wan et al., 2010), sentence ranking based on cross-lingual sentence similarity (Wan, 2011), or abstractive summarization based on phrase pair (Yao et al., 2015) and predicate-argument structure fusing (Zhang et al., 2016).

The first cross-lingual ASSUM system based on the large-scale monolingual ASSUM corpus is proposed by Ayana et al. (2018), which is most related to our work. It is motivated by the triangular NMT systems with pseudo target in the teacher-student networks. In contrast, we use pseudo source and apply different teacher-student networks.

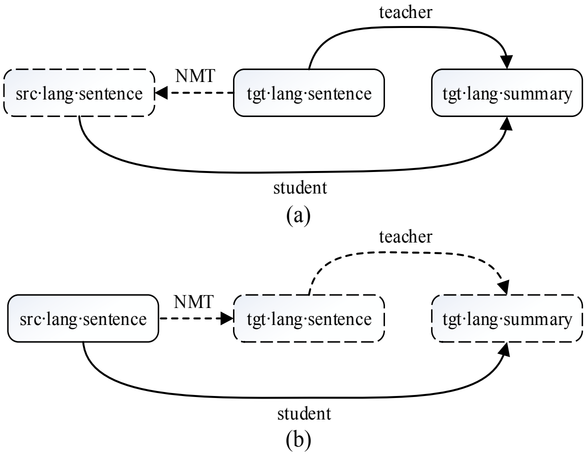

Figure 1: Illustration of the comparison between (a) our overall framework and (b) the framework of Ayana et al. (2018). Solid line boxes denote genuine data, while dashed line boxes denote automatically generated pseudo data. Solid line arrows denote the summarization direction, while dashed line arrows denote pseudo data generation direction. Note that the genuine data is used in the teacher of our framework, while pseudo data is used in the teacher of the framework of Ayana et al. (2018).

3 Our Approach

3.1 Overall Framework

To overcome the data scarcity in the cross-lingual ASSUM, it is easy to come up with the pipeline method at the first thought. The source language sentence can be translated to the target language sentence, followed by target language summarization step to get the final target language summary. Alternatively, the source language sentence can be summarized into source language summary at first, then is translated into the target language summary. Both pipeline methods face the error propagation problem that errors in the early steps will harm the final summary quality.

We propose a jointly optimizing framework that avoids the independent two steps in the pipeline methods. Figure 1 (a) illustrates our overall framework. We introduce a bridge between the source language sentence and the target language summary. The target language sentence functions as the bridge convenient for the information flow from the source sentence to the target summary.

The overall framework mainly consists of two modules: the teacher networks and the student networks. The teacher is the monolingual ASSUM neural networks trained on the large-scale monolingual ASSUM corpus. Note that in our framework, the teacher is strong since the utilized monolingual ASSUM corpus is genuine and no pseudo data is used in the teacher. The student is the cross-lingual ASSUM networks trained to mimic the behavior of the teacher.

To manifest the difference between our framework and the most related framework of Ayana et al. (2018), we depict both in Figure 1. In the framework of Ayana et al. (2018), the source language sentence is automatically translated into the target language sentence, which is automatically summarized into the target language summary. The data on both sides of the teacher networks are pseudo. This is significantly different to our framework in which the teacher networks have the strong data basis that all data on both sides of the teacher networks are genuine. When comparing the student networks, we can find that we adopt pseudo source sentence, while Ayana et al. (2018) adopt pseudo target summary. Furthermore, we also teach the student with the teacher's attention weights via a new attention relay mechanism.

3.2 Back-Translation

Our framework contains a back-translation procedure which is inspired by that used in semi-supervised or unsupervised machine translation (MT) (Sennrich et al., 2016a; Lample et al., 2018). In MT, the back-translation process translates unpaired target text into source text. The resulted pseudo source-target pair serves as additional training data for source-to-target translation. Our proposed back-translation procedure involves triple kinds of data. It translates the target language sentence back into the source language sentence by a third-party NMT system. The generated pseudo source is paired with the true target summary to build a training resource for the student networks. The back-translation procedure is denoted as the dashed arrow NMT in Figure 1(a).

3.3 The Teacher-Student Training Procedure

We use the monolingual ASSUM system as the teacher networks, and use the cross-lingual ASSUM system as the student networks. Both the teacher and the student apply Transformer architecture which is effective for modeling sequence-to-sequence tasks such as machine translation (Vaswani et al., 2017). Two functions of the teacher are set as the learning objective for the student. One is the probability distribution of the summary word generation, the other is the attention weights in the attention mechanism.

Given the source language text *X*, the target language text *Y*, and the target language summary *YS*, the training procedure for the teacher-student framework is presented in the following:

Teaching The Summary Word Generation

Let $P(YS_i|YS_1^{i-1}, Y)$ denote the teacher distribution of the summary word given summary word generation history and *Y*, $P(YS_i|YS_1^{i-1}, \widehat{X})$ denote the student distribution of the summary word given summary word generation history and $\widehat{X}$. $\widehat{X}$ denotes the pseudo source which is generated by the back-summarization procedure. We use cross entropy loss to encourage the similarity between the two distributions:

$$L_{gen} = -P(YS_i|YS_1^{i-1}, Y)\log P(YS_i|YS_1^{i-1}, \widehat{X}) \tag{1}$$

Through Equation (1), the cross-lingual ASSUM learns from the monolingual ASSUM about how to generate summary word under appropriate distribution.

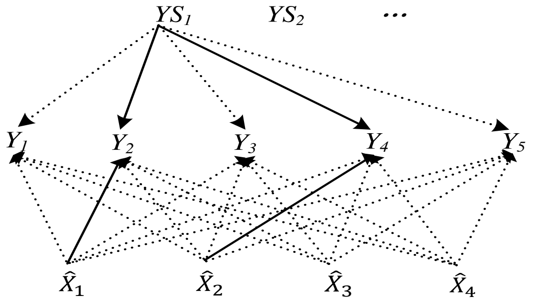

Figure 2: Illustration of the attention relay. The arrows are the attentions with the direction of decoder side word attending to encoder side words. Solid arrows are top-k or the biggest attention weights, and the dashed arrows are the left attention weights. k is 2 in the figure.

Teaching The Attention Weights via Attention Relay

Besides the summary word generation distribution, the attention of the monolingual ASSUM is also a valuable learning resource. But such attention only connects the encoder and the decoder of the monolingual ASSUM system, it has to be relayed to reach the other language to teach the cross-lingual ASSUM system.

The attention relay mechanism is illustrated in Figure 2. The monolingual attention weights of *YS* attending to *Y* is relayed to form the teacher attention weights of *YS* attending to $\widehat{X}$. In particular, Y_2 and Y_4 receive top-2 attention weights from YS_1 in Figure 2, and Y_2 receives biggest attention from $\widehat{X}_1$, Y_4 receives biggest attention from $\widehat{X}_2$. Then the attention weights of YS_1 attending to $\widehat{X}_1$ and $\widehat{X}_2$ are set 1/2. Other attention weights distributed over the rest words of $\widehat{X}$ are set zero. In general case, if top-k attention weights are relayed from *YS* to $\widehat{X}$, then the teacher attention weights over the k words of $\widehat{X}$ are set $1/k$ each, and other attention weights are set zero [1].

We use the Euclidean distance between the teacher attention weights and the student attention

[1] We also use the attention matrix of *YS* attending to *Y* multiplies the attention matrix of $\widehat{X}$ attending to *Y* to form the teacher attention, but we found that the teacher attention weights are evenly distributed, resulting in worse student performance.

weights as the loss to encourage their consistency:

$$L_{att} = \sqrt{\sum_{j} (A_j - \bar{A}_j)^2} \tag{2}$$

where A_j denotes a teacher attention weight, $\bar{A}_j$ denotes a student attention weight.

Note that in our work, the attention only refers to the encoder-decoder attention, not the self-attention in Transformer. Since our teacher networks and the student networks adopt Transformer architecture which contains multi-head attention, we use the average attention that averages attention weights of all heads in the same layer.

3.4 Training and Testing

The training objective is to minimize the joint loss:

$$\mathcal{L} = \lambda L_{gen} + (1 - \lambda) L_{att} \tag{3}$$

where λ is the weight balancing L_{gen} and L_{att}.

During testing, only the student networks are used to decode X into YS. In detail, only $P(YS_i|YS_1^{i-1}, X)$ participates in the beam search, the summary word generation teacher and whole L_{att}-related teacher-student networks are not involved in the testing process.

4 Experiments

We conduct experiments on Chinese-to-English ASSUM, which takes Chinese sentence as input, and outputs English abstractive summary. We build evaluation sets for this task by manually translating English sentences of the existing English evaluation sets into Chinese inputs. To the best of our knowledge, these are the first evaluation sets on which the cross-lingual ASSUM system and the monolingual ASSUM system can be directly compared.

4.1 Datasets

In our experiments, the English ASSUM system and the English-Chinese NMT system are involved. The data for training both systems are presented below.

The data for training the English ASSUM system is from the annotated Gigaword corpus, and we preprocess it identically to Rush *et al.* (2015), which results in around 3.8M training pairs, 190K validation pairs and 1951 test pairs. In this data, the sentence-summary pairs are built by pairing the first sentence of each article with the article's headline. Additionally, DUC-2004 is adopted as another English data set only for testing. It contains 500 documents, and each document has four human-generated reference summaries.

To build the evaluation sets, English sentences of the validation and test sets of Gigaword and DUC2004 are manually translated into Chinese by graduate students of the linguistics department and our institute, who are bilingual with Chinese as the mother tongue. Specifically, in the Gigaword validation set, we randomly select 2000 sentence-summary pairs and manually translate their English sentences into Chinese. The English summaries are not translated. The Chinese sentences are segmented by the word segmentation tool Jieba[2].

Additionally, we also implement some baselines for comparison, some of which utilize a large corpus of Chinese short text summarization (LCSTS) (Hu et al., 2015), which is collected from the Chinese microblogging website Sina Weibo with 2.4M sentence-summary pairs for training and 725 pairs for testing.

4.2 Experimental Configuration

Baseline Systems

- The pipeline of translating source sentence into target sentence at first, then summarizing the target sentence into the summary. We denote this method Pipeline-TS.

- The pipeline of summarizing the source sentence into the source summary, then translating the source summary into the target summary. We denote this method Pipeline-ST.

- The framework of Ayana et al. (2018), which uses pseudo summary for training. We denote it Pseudo-Summary[3].

- The pivot system that enforcing the source-to-pivot system and the pivot-to-target system sharing the same pivot language embedding (Cheng et al., 2017). We denote it Pivot-based.

[2] https://pypi.org/project/jieba/

[3] We also implement the framework that uses the NMT model to teach the cross-lingual ASSUM (Ayana et al., 2018). Since it highly depends on the LCSTS data, whose style is different to our evaluation sets, it performs significantly worse.

	NIST02	NIST03	NIST04	NIST05	NIST08	Avg
Our Transformer Cn2En	45.58	45.19	46.80	46.56	37.27	44.28
Robust Translation Cn2En (Cheng et al., 2018)	46.10	44.07	45.61	44.06	34.94	42.96
Our Transformer En2Cn	39.38	34.48	38.10	36.20	30.80	35.79

Table 1: BLEU of the NMT systems on NIST evaluation sets. Cn2En denotes Chinese-to-English translation, and En2Cn denotes the reverse direction.

System	Gigaword			DUC2004		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
ABS+ (Rush et al., 2015)	29.8	11.9	27.0	28.2	8.5	23.8
Actor-Critic (Li et al., 2018)	36.1	17.4	33.5	29.4	9.8	25.9
StructuredLoss (Edunov et al., 2018)	36.7	17.9	34.3	-	-	-
FactAware (Cao et al., 2018)	37.3	17.7	34.2	-	-	-
Transformer	37.1	18.2	34.4	30.6	10.5	26.6
Transformer$_{\text{bpe}}$	38.1	19.1	35.2	31.2	10.7	27.1

Table 2: Comparison on the monolingual ASSUM performances. "-" denotes that no score is reported in that work.

- Translating the English sentences into the Chinese sentences, and pair these pseudo Chinese sentences with English summaries to build a training corpus for Chinese-to-English ASSUM. We denote it Pseudo-Chinese. We implement it by using Transformer machine translation model to translate the English sentences, and use Transformer architecture to train a Chinese-to-English ASSUM system. Note that this is just the student network without being taught by a teacher network.

Parameter Setup and Evaluation Metric

Transformer is employed as our basis architecture[4] (Vaswani et al., 2017). Six layers are stacked in both the encoder and decoder, and the dimensions of the embedding vectors and all hidden vectors are set 512. We set eight heads in the multi-head attention. The source embedding, the target embedding and the linear sublayer are shared in the teacher networks, while are not shared in the student networks. Byte-pair encoding is employed with a vocabulary of about 32k tokens on English side and Chinese side respectively (Sennrich et al., 2016b).

During evaluation, we employ ROUGE (Lin, 2004) as our evaluation metric. On Gigaword, the full-length F-1 based ROUGE scores are reported. On DUC2004, the recall based ROUGE scores are reported to be consistent with previous works.

NMT Performance

The NMT system involved in all our experiments is Transformer, with the same parameter setup to those of ASSUM systems. It is trained on 1.25M sentence pairs extracted from LDC corpora[5], and is evaluated on NIST sets using multi-bleu.perl. Chinese-to-English results of case-insensitive BLEU and English-to-Chinese results of character-based BLEU are reported in Table 1. Since there are four English references for one Chinese sentence in NIST evaluation sets, we report averaged BLEU of four English input sentences in English-to-Chinese translation.

Compared to Cheng et al. (2018) on Chinese-to-English translation, which targets at robust machine translation and uses the same data to ours, our Transformer significantly outperforms their work, indicating that we build a solid system for machine translation.

4.3 Experimental Results

Monolingual ASSUM Performance

We build a strong monolingual ASSUM system as shown in Table 2. The comparison is made between our basis architecture Transformer and previous works including state-of-the-art monolingual ASSUM systems. The work of ABS+ (Rush et al., 2015) is the pioneer work of using neural models for monolingual ASSUM. The works of Actor-Critic (Li et al., 2018) and Structured-Loss (Edunov et al., 2018) are training methods avoiding exposure bias problems in sequence-to-sequence learning. The work of FactAware (Cao et al., 2018) encode factual informations such as

[4] https://github.com/pytorch/fairseq

[5] The corpora include LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06.

System	Gigaword			DUC2004		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Transformer$_{\text{bpe}}$	38.1	19.1	35.2	31.2	10.7	27.1
Pipeline-TS	25.8	9.7	23.6	23.7	6.8	20.9
Pipeline-ST	22.0	7.0	20.9	20.9	5.3	18.3
Pseudo-Summary (Ayana et al., 2018)	21.5	6.6	19.6	19.3	4.3	17.0
Pivot-based (Cheng et al., 2017)	26.7	10.2	24.3	24.0	7.0	21.3
Pseudo-Chinese	27.9	10.9	25.6	24.4	6.6	21.4
Teaching Generation	29.6	12.1	27.3	25.6	7.9	22.7
Teaching Attention	28.1	11.4	26.0	24.3	7.4	21.7
Teaching Generation+Attention	**30.1**	**12.2**	**27.7**	**26.0**	**8.0**	**23.1**

Table 3: Comparison on the cross-lingual ASSUM performances.

those extracted from openIE and dependency relations into the neural network to get factual summaries.

Transformer with BPE pre-processing (denoted by Transformer$_{\text{bpe}}$) performs consistently better than the related monolingual ASSUM systems. We build the cross-lingual ASSUM system basing on Transformer$_{\text{bpe}}$.

Cross-lingual ASSUM performance

Table 3 mainly presents the results of the cross-lingual ASSUM systems. The first row lists the performance of Transformer$_{\text{bpe}}$, which is the monolingual ASSUM system. It sets the ceiling of the cross-lingual ASSUM performance since the cross-lingual process introduces information loss when using another language.

Comparisons between the Baselines The middle part of Table 3 is about baseline systems. It shows that Pipeline-TS is significantly better than Pipeline-ST. The optimal order of the two steps in the pipeline methods should be translating source sentence at first, then summarizing the translation. The Pseudo-Summary method (Ayana et al., 2018) performs even below the Pipeline-ST method. It indicates that using the pseudo target side is not effective for learning better cross-lingual summarization model. Meanwhile, as Figure 1(b) illustrates, both source side and target side of the teacher network in the framework of Ayana et al. (2018) are pseudo, resulting in less solid data basis for training the student. The pseudo source side is generated by translating LCSTS Chinese sentences.

The two baseline systems that surpass the pipeline systems are Pivot-based system and Pseudo-Chinese system. We re-implement the Pivot-based system but using Transformer instead of RNN, which is used in Cheng et al. (2017). Pseudo-Chinese system is the best baseline system indicating that pseudo source based parallel data is effective for training cross-lingual ASSUM system.

Our Systems VS. the Baselines The bottom part of Table 3 lists the performances of our methods. It manifests that both teaching summary word generation and teaching attention weights are able to improve the performance over the baselines. When the summary word generation and attention weights are taught simultaneously (denoted by Teaching Generation+Attention), the performance is further improved, surpassing the best baseline by more than two points on Gigaword evaluation set and more than one point on DUC2004.

Our Systems VS. the Ceiling Teaching Generation+Attention greatly reduces the gap between the cross-lingual ASSUM performance and the performance ceiling, i.e., the monolingual ASSUM performance shown in the first row. The gap is narrowed from 10.2 ROUGE-1 points to 8 ROUGE-1 points. In fact, our best method performs even better than ABS+, which is the early system for monolingual ASSUM (Rush et al., 2015).

4.4 Experiment Analyses

Hyper-Parameters

λ	ROUGE-1	ROUGE-2	ROUGE-L
0.1	44.8	22.0	41.7
0.3	45.1	22.3	42.0
0.5	45.0	22.2	41.9
0.7	44.9	22.2	41.8
0.9	44.8	21.8	41.7
top-k	ROUGE-1	ROUGE-2	ROUGE-L
2	44.8	22.4	41.9
3	44.9	22.0	42.0
4	45.1	22.3	42.0
5	45.1	22.2	41.8

Table 4: Performances of varying hyper-parameters on the validation set.

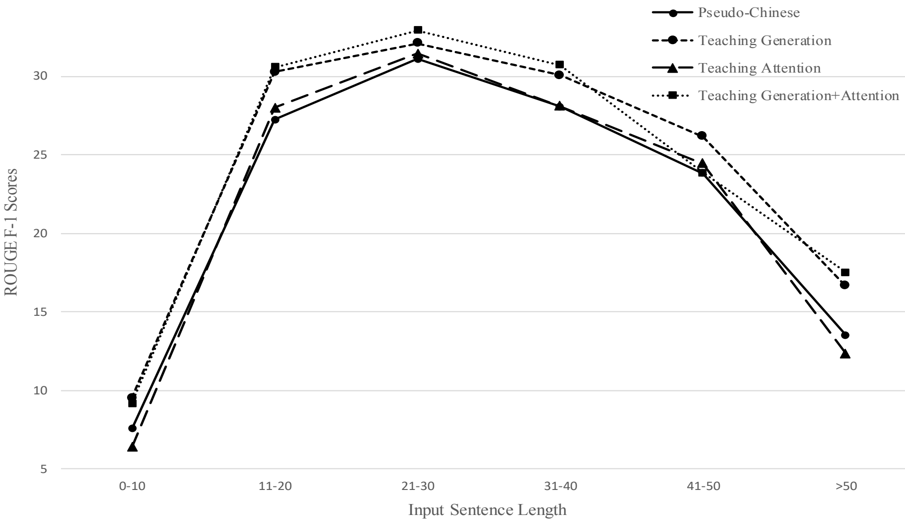

Figure 3: ROUGE-1 scores on different length source sentences in the Gigaword test set.

There are two main hyper-parameters. One is λ in Equation (3) that balances the weights between teaching generation and teaching attention during training. The other is top-k which controls how many portion of the monolingual ASSUM attention can be relayed to the source side as illustrated in Figure 2. Table 4 presents the performance variance when the two hyper-parameters vary.

It shows that the performance is best when λ is 0.3, indicating that training process is balanced towards teaching attention via attention relay. Based on the best λ of 0.3, we explore top-k ranging from 2 to 5. We can find that top-4 monolingual ASSUM attention weights achieve the best performance on the validation set. We select the best hyper-parameters according to Table 4 for testing.

Layers for Attention Relay

Transformer architecture used in our experiment is with six layers on both encoder and decoder. Attention relay can take place on each layer. Since each layer has eight heads for attention computation, we average the weights of all eight heads in the same layer. We study the attention relay effects on all six layers. The results in Table 5 show that relaying attention on the last layer achieves the best performance.

Performances on Different Lengths

We study the performance of each system on sets with different source sentence lengths. The source sentences are divided into six groups according to their lengths. Figure 3 presents the ROUGE-1 scores on the test set. The strongest baseline Pseudo-Chinese is used in this study. It shows that our methods perform better than Pseudo-Chinese on most groups, while teaching attention is slightly worse on the group with the longest length. The sentences with length range 10-50 take up 94.2% of the whole test set. Our methods are consistently better than Pseudo-Chinese on theses sentences.

Layer	ROUGE-1	ROUGE-2	ROUGE-L
1	44.7	21.8	41.6
2	44.7	22.3	41.7
3	45.0	22.0	41.8
4	44.9	22.1	41.3
5	44.9	22.1	41.9
6	45.1	22.3	42.0

Table 5: Validation set performances of using different layers for attention relay.

Qualitative Analysis

Table 6 presents some examples of the cross-lingual ASSUM. The differences between our methods and the strongest baseline Pseudo-Chinese are highlighted. It shows that more accurate summary words are produced in our systems. In contrast, Pseudo-Chinese may produce incorrect words that are even contrary to the meaning of the original sentence.

5 Conclusion

In this paper, we propose a teacher-student framework together with the back-translation procedure to deal with the zero-shot challenge of cross-

Cn-sentence	据 周六 报道 ， 印度 最高 核 专家 对 广岛 日 印度人 的 反核 抗议 不屑一顾 ， 称 激进分子 应该 在 华盛顿 和 莫斯科 喊 口号 。
En-sentence	a india 's top nuclear expert shrugged off antinuclear protests by indians on hiroshima day , saying the activists should instead shout slogans in washington and moscow , a newspaper reported saturday .
Ref-summary	top nuclear scientist shrugs off indian antinuclear protests
Psueo-Chinese	india 's top nuclear expert calls for anti-nuke demo in hiroshima
Teaching-Generation	india 's top nuclear expert warns against nuclear protests
Teaching-Attention	india 's top nuclear scientist defies hiroshima protest
Teaching-Gener+Attn	india 's top nuclear expert defies anti-nuclear protests
Cn-sentence	黎巴嫩 总理 拉菲克 - 哈里里 星期二 指责 英国 支持 以色列 袭击 黎巴嫩真主党 游击队 ， 同时 他 宣布 计划 访问 伦敦 。
En-sentence	lebanese prime minister rafic hariri accused britain on tuesday of supporting the israeli assault on hezbollah guerrillas in lebanon as he announced plans to visit london .
Ref-summary	hariri to visit britain which he accuses of backing israel
Psueo-Chinese	lebanese pm accuses britain of supporting hezbollah
Teaching-Generation	lebanese pm accuses britain of backing hezbollah attacks
Teaching-Attention	lebanese pm accuses britain of supporting hezbollah
Teaching-Gener+Attn	lebanese pm accuses britain of backing israel
Cn-sentence	苏丹 武装部队 发言人 今天 说 ， 政府军 击退 了 叛军 沿 苏丹 东部 边境 发动 的 攻击 。
En-sentence	government troops has repelled an attack by rebel forces along sudan 's eastern borders , the spokesman of the sudanese armed forces said today .
Ref-summary	government forces repel rebel attack in eastern
Psueo-Chinese	sudanese government forces attack rebels in eastern sudan
Teaching-Generation	government troops repulse rebel attack in eastern sudan
Teaching-Attention	sudanese army says it foiled rebel attack on eastern border
Teaching-Gener+Attn	government troops repel rebel attack in eastern sudan

Table 6: Examples of the cross-lingual ASSUM.

lingual ASSUM, which has no parallel data for training. We use monolingual ASSUM which has large-scale training resources as the teacher, and set the cross-lingual ASSUM as the student. Two properties of the teacher are proposed to teach the student. One is the summary word generation probabilities, the other is the attention weights. We also propose attention relay mechanism to form the attention weights of the teacher. Experiments show that our method performs significantly better than several baselines, and is able to significantly reduce the performance gap between the cross-lingual ASSUM and the monolingual ASSUM over the benchmark datasets.

Acknowledgments

The authors would like to thank the anonymous reviewers for the helpful comments. This work was supported by National Key R&D Program of China (Grant No. 2016YFE0132100), National Natural Science Foundation of China (Grant No. 61525205, 61673289), and was also partially supported by the joint research project of Alibaba and Soochow University.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270.

Ayana, Shi-Qi Shen, Yun Chen, Yang Cheng, Zhiyuan Liu, and Maosong Sun. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2319–2327.

Ayana, Shiqi Shen, Yu Zhao, Zhiyuan Liu, and Maosong Sun. 2016. Neural headline generation with sentence-wise optimization. *arXiv preprint arXiv:1604.01904*.

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

Michele Banko, Vibhu O. Mittal, and Michael J. Witbrock. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 310–317.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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 (Volume 1: Long Papers)*, pages 1925–1935.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766.

Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. 2017. Joint training for pivot-based neural machine translation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3974–3980.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 137–144.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 355–364.

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 - Volume 70*, pages 1243–1252. JMLR.org.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, pages 339–351.

Philipp Koehn, Franz J Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

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

Piji Li, Lidong Bing, and Wai Lam. 2018. Actor-critic based training framework for abstractive summarization. *arXiv:1803.11070*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira Dos Santos, Caglar Gulcehre, and Xiang Bing. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

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

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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

Xiaojun Wan. 2011. Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1546–1555.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926.

Li Wang, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. 2018. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4453–4460.

Jinge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 118–127.

Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2016. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM Trans. Audio Speech, Lang. Process*, vol. 24, no. 10.

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*, pages 4251–4257.

Qingyu Zhou, Yang Nan, Furu Wei, and Zhou Ming. 2017. Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104.