Jointly Learning to Align and Summarize for Neural Cross-Lingual Summarization

Yue Cao, Hui Liu, Xiaojun Wan
Wangxuan Institute of Computer Technology, Peking University
Center for Data Science, Peking University
The MOE Key Laboratory of Computational Linguistics, Peking University
{yuecao,xinkeliuhui,wanxiaojun}@pku.edu.cn

Abstract

Cross-lingual summarization is the task of generating a summary in one language given a text in a different language. Previous works on cross-lingual summarization mainly focus on using pipeline methods or training an end-to-end model using the translated parallel data. However, it is a big challenge for the model to directly learn cross-lingual summarization as it requires learning to understand different languages and learning how to summarize at the same time. In this paper, we propose to ease the cross-lingual summarization training by jointly learning to align and summarize. We design relevant loss functions to train this framework and propose several methods to enhance the isomorphism and cross-lingual transfer between languages. Experimental results show that our model can outperform competitive models in most cases. In addition, we show that our model even has the ability to generate cross-lingual summaries without access to any cross-lingual corpus.

1 Introduction

Neural abstractive summarization has witnessed rapid growth in recent years. Variants of sequence-to-sequence models have shown to obtain promising results on English (See et al., 2017) or Chinese summarization datasets. However, **Cross-lingual summarization**, which aims at generating a summary in one language from input text in a different language, has been rarely studied because of the lack of parallel corpora.

Early researches on cross-lingual abstractive summarization are mainly based on the summarization-translation or translation-summarization pipeline paradigm and adopt different strategies to incorporate bilingual features (Leuski et al., 2003; Orasan and Chiorean, 2008; Wan et al., 2010; Wan, 2011) into the pipeline model.

Recently, Shen et al. (2018) first propose a neural cross-lingual summarization system based on a large-scale corpus. They first translate the texts automatically from the source language into the target language and then use the teacher-student framework to train a cross-lingual summarization model. Duan et al. (2019) further improve this teacher-student framework by using genuine summaries paired with the translated pseudo source sentences to train the cross-lingual summarization model. Zhu et al. (2019) propose a multi-task learning framework to train a neural cross-lingual summarization model.

Cross-lingual summarization is a challenging task as it requires learning to understand different languages and learning how to summarize at the same time. It would be difficult for the model to directly learn cross-lingual summarization. In this paper, we explore this question: can we ease the training and enhance the cross-lingual summarization by establishing alignment of context representations between two languages?

Learning cross-lingual representations has been proven a beneficial method for cross-lingual transfer for some downstream tasks (Klementiev et al., 2012; Artetxe et al., 2018; Ahmad et al., 2019; Chen et al., 2019). The underlying idea is to learn a shared embedding space for two languages to improve the model's ability for cross-lingual transfer. Recently, it has been shown that this method can also be applied to context representations (Aldarmaki and Diab, 2019; Schuster et al., 2019). In this paper, we show that the learning of cross-lingual representations is also beneficial for neural cross-lingual summarization models.

We propose a multi-task framework that jointly learns to summarize and align context-level representations. Concretely, we first integrate monolingual summarization models and cross-lingual summarization models into one unified model and then

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6220–6231
July 5 - 10, 2020.

build two linear mappings to project the context representation from one language to the other. We then design several relevant loss functions to learn the mappers and facilitate the cross-lingual summarization. In addition, we propose some methods to enhance the isomorphism and cross-lingual transfer between different languages. We also show that the learning of aligned representation enables our model to generate cross-lingual summaries even in a fully unsupervised way where no parallel cross-lingual data is required.

We conduct experiments on several public cross-lingual summarization datasets. Experiment results show that our proposed model outperforms competitive models in most cases, and our model also works on the unsupervised setting. To the best of our knowledge, we are the first to propose an unsupervised framework for learning neural cross-lingual summarization.

In summary, our primary contributions are as follow:

- We propose a framework that jointly learns to align and summarize for neural cross-lingual summarization and design relevant loss functions to train our system.
- We propose a procedure to train our cross-lingual summarization model in an unsupervised way.
- The experimental results show that our model outperforms competitive models in most cases, and our model has the ability to generate cross-lingual summarization even without any cross-lingual corpus.

2 Overview

We show the overall framework of our proposed model in Figure 1. Our model consists of two encoders, two decoders, two linear mappers, and two discriminators.

Suppose we have an English source text $\boldsymbol{x} = \{x_1, \ldots, x_m\}$ and a Chinese source text $\boldsymbol{y} = \{y_1, \ldots, y_n\}$, which consist of m and n words, respectively. The English encoder $\phi_{E_\mathcal{X}}$ (res. Chinese encoder $\phi_{E_\mathcal{Y}}$) transforms $\boldsymbol{x}$ (res. $\boldsymbol{y}$) into its context representation $\boldsymbol{z_x}$ (res. $\boldsymbol{z_y}$), and the decoder $\phi_{D_\mathcal{X}}$ (res. $\phi_{D_\mathcal{Y}}$) reads the memory $\boldsymbol{z_x}$ (res. $\boldsymbol{z_y}$) and generates the corresponding English summary $\tilde{\boldsymbol{x}}$ (res. Chinese summary $\tilde{\boldsymbol{y}}$).

The mappers $M_\mathcal{X} : \mathcal{Z}_{\boldsymbol{x}} \rightarrow \mathcal{Z}_{\boldsymbol{y}}$ and $M_\mathcal{Y} : \mathcal{Z}_{\boldsymbol{y}} \rightarrow \mathcal{Z}_{\boldsymbol{x}}$ are used for transformations between $\boldsymbol{z_x}$ and $\boldsymbol{z_y}$, and the discriminators $D_\mathcal{X}$ and $D_\mathcal{Y}$ are used for discriminating between the encoded representations and the mapped representations.

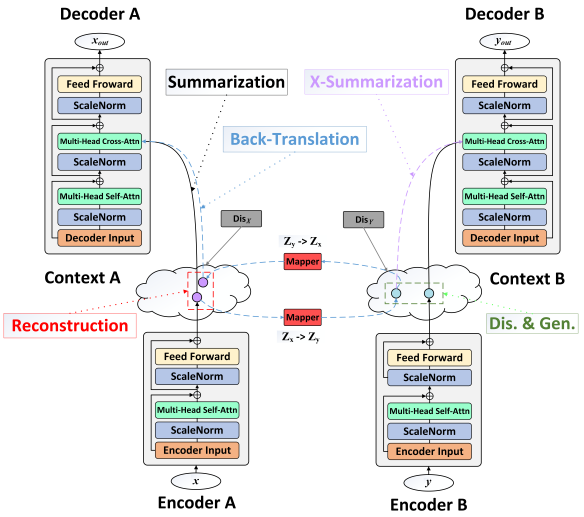


Figure 1: The overall framework of our proposed model.

Taking English-to-Chinese summarization for example, our model generates cross-lingual summaries as follows: First we use the English encoder to get the English context representations, then we use the mapper to map English representations into Chinese space. Lastly the Chinese decoder is used to generate Chinese summaries.

In Section 3, we describe the techniques we adopt to enhance the cross-lingual transferability of the model. In Section 4 and Section 5, we describe the unsupervised training objective and supervised training objective for cross-lingual summarization, respectively.

3 Model Adjustment for Cross-Lingual Transfer

3.1 Normalizing the Representations

In our model, we adopt Transformer (Vaswani et al., 2017) as our encoder and decoder, which is the same with previous works (Duan et al., 2019; Zhu et al., 2019). The encoder and decoder are connected via cross-attention. The cross-attention is implemented as the following dot-product attention module:

$$\text{Attention}\,(S, T) = \text{softmax}\left(\frac{TS^\top}{\sqrt{d_k}}\right) S \quad (1)$$

where S is the packed encoder-side contextual representation, T is the packed decoder-side contextual representation and d_k is the model size.

In the dot-product module, it would be beneficial if the contextual representations of the encoder and decoder have the same distributions. However, in the cross-lingual setting, the encoder and decoder deal with different languages and thus the distributions of the learned contextual representations may be inconsistent. This motivates us to explicitly learn alignment relationships between languages.

To make the contextual representations of two languages easier to be aligned, we introduce the normalization technique into the transformer model. Normalizing the word representations has been proved an effective technique on word alignment (Xing et al., 2015). After normalization, two sets of embeddings are both located on a unit hypersphere, which makes them easier to be aligned.

We achieve this by introducing the pre-normalization technique and replacing the LayerNorm with ScaleNorm (Nguyen and Salazar, 2019):

$$o_{\ell+1} = \text{LayerNorm}\left(o_\ell + F_\ell\left(o_\ell\right)\right)$$
$$\Downarrow$$
$$o_{\ell+1} = o_\ell + F_\ell(\text{ScaleNorm}\left(o_\ell\right))$$

where F_ℓ is the ℓ-th layer and o_ℓ is its input. The formula for calculating ScaleNorm is:

$$\text{ScaleNorm}(x; g) = g \cdot x / \|x\| \tag{2}$$

where g is a hyper-parameter.

An additional benefit of ScaleNorm is that after being normalized, the dot-product of two vectors $u^\top v$ is equivalent to their cosine distance $\frac{u^\top v}{\|u\|\|v\|}$, which may benefit the attention module in Transformer. We will conduct experiments to verify this.

3.2 Enhancing the Isomorphism

A key assumption of aligning the representations of two languages is the **isomorphism** of learned monolingual representations. Some researchers show that the isomorphism assumption weakens when two languages are etymologically distant (Søgaard et al., 2018; Patra et al., 2019). However, Ormazabal et al. (2019) show that this limitation is due to the independent training of two separate monolingual embeddings, and they suggest to jointly learn cross-lingual representations on monolingual corpora. Inspired by Ormazabal et al. (2019), we take the following approaches to address the isomorphism problem.

First, we combine the English and Chinese summarization corpora and build a unified vocabulary. Second, we share encoders and decoders in our model. Sharing encoders and decoders can also enforce the model to learn shared contextual representations across languages. For the shared decoder, to indicate the target language, we set the first token of the decoder to specify the language the module is operating with. Third, we train several monolingual summarization steps before cross-lingual training, as shown in the first line in Alg. 1. The pre-trained monolingual summarization steps also allow the model to learn easier monolingual summarization first, then further learn cross-lingual summarization, which may reduce the training difficulty.

4 Unsupervised Training Objective

We describe the objective of unsupervised cross-lingual summarization in this section. The whole training procedure can be found in Alg. 1.

Summarization Loss Given an English text-summary pair $\boldsymbol{x}$ and $\boldsymbol{x'}$, we use the encoder $\phi_{E_\mathcal{X}}$ and the decoder $\phi_{D_\mathcal{X}}$ to generate the hypothetical English summary $\tilde{\boldsymbol{x}}$ that maximizes the output summary probability given the source text: $\tilde{\boldsymbol{x}} = \arg\max_{\bar{\boldsymbol{x}}} P(\bar{\boldsymbol{x}} \mid \boldsymbol{x})$. We adopt maximum log-likelihood training with cross-entropy loss between hypothetical summary $\tilde{\boldsymbol{x}}$ and gold summary $\boldsymbol{x'}$:

$$\boldsymbol{z_x} = \phi_{E_\mathcal{X}}(\boldsymbol{x}), \quad \tilde{\boldsymbol{x}} = \phi_{D_\mathcal{X}}(\boldsymbol{z_x})$$
$$\mathcal{L}_{\text{summ}_\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x'}) = -\sum_{t=1}^{T} \log P\left(x'_t \mid \tilde{x}_{<t}, \boldsymbol{z_x}\right) \tag{3}$$

where T is the length of $\boldsymbol{x'}$. The Chinese summarization loss $\mathcal{L}_{\text{summ}_\mathcal{Y}}$ is similarly defined for the Chinese encoder $\phi_{E_\mathcal{Y}}$ and decoder $\phi_{D_\mathcal{Y}}$.

Generative and Discriminative Loss Given an English source text $\boldsymbol{x}$ and a Chinese source text $\boldsymbol{y}$, we use the encoder $\phi_{E_\mathcal{X}}$ and $\phi_{E_\mathcal{Y}}$ to obtain the contextual representations $\boldsymbol{z_x} = \{z_{x_1}, \ldots, z_{x_m}\}$ and $\boldsymbol{z_y} = \{z_{y_1}, \ldots, z_{y_n}\}$, respectively. For Zh-to-En summarization, we use the mapper $M_\mathcal{Y}$ to map $\boldsymbol{z_y}$ into the English context space: $\boldsymbol{z_{y \to x}} = M_\mathcal{Y}(\boldsymbol{z_y})$. We hope the mapped distribution $\boldsymbol{z_{y \to x}}$ and the real English distribution $\boldsymbol{z_x}$ could be as similar as possible such that the English decoder can deal with cross-lingual summarization just like dealing with monolingual summarization.

To learn this mapping, we introduce two discriminators and adopt the adversarial training (Goodfellow et al., 2014) technique. We optimize the

mappers at the sentence-level[1] rather than word-level, which is inspired by Aldarmaki and Diab (2019) where they found learning the aggregate mapping can yield a more optimal solution compared to word-level mapping.

Concretely, we first average the contextual representations:

$$\tilde{z}_{y\to x} = \frac{1}{n}\sum_{i=1}^{n}\left(z_{y\to x}\right)_i\,, \quad \tilde{z}_x = \frac{1}{m}\sum_{i=1}^{m} z_{x_i} \quad (4)$$

Then we train the discriminator $D_{\mathcal{X}}$ to discriminate between $\tilde{z}_{y\to x}$ and $\tilde{z}_x$ using the following discriminative loss:

$$\begin{aligned}\mathcal{L}_{dis_{\mathcal{X}}}\left(\tilde{z}_{y\to x}, \tilde{z}_x\right) = &-\log P_{D_{\mathcal{X}}}\left(\text{src}=0|\tilde{z}_{y\to x}\right)\\ &-\log P_{D_{\mathcal{X}}}\left(\text{src}=1|\tilde{z}_x\right)\end{aligned} \quad (5)$$

where $P_{D_{\mathcal{X}}}\left(\text{src}\,|\tilde{z}\right)$ is the predicted probability of $D_{\mathcal{X}}$ to distinguish whether $\tilde{z}$ is coming from the real English representation $(\text{src}=1)$ or from the mapper $M_{\mathcal{Y}}$ $(\text{src}=0)$.

In our framework, the encoder $\phi_{E_{\mathcal{X}}}$ and mapper $M_{\mathcal{Y}}$ together make up the generator. The generator tries to generate representations which would confuse the discriminator, so its objective is to maximize the discriminative loss in Eq. 5. Alternatively, we train the generator to minimize the following generative loss:

$$\begin{aligned}\mathcal{L}_{gen_{\mathcal{Y}}}\left(\tilde{z}_{y\to x}, \tilde{z}_x\right) = &-\log P_{D_{\mathcal{X}}}\left(\text{src}=1|\tilde{z}_{y\to x}\right)\\ &-\log P_{D_{\mathcal{X}}}\left(\text{src}=0|\tilde{z}_x\right)\end{aligned} \quad (6)$$

The discriminative loss $\mathcal{L}_{dis_{\mathcal{Y}}}\left(\tilde{z}_{x\to y}, \tilde{z}_y\right)$ for $D_{\mathcal{Y}}$, generative loss $\mathcal{L}_{gen_{\mathcal{X}}}\left(\tilde{z}_{x\to y}, \tilde{z}_y\right)$ for $\phi_{E_{\mathcal{Y}}}$ and $M_{\mathcal{X}}$ are similarly defined.

Notice that since we use vector averaging and adopt the linear transformation, it does not matter whether we apply the linear mapping before or after averaging the contextual representations, and the learned sentence-level mappers can be directly applied to word-level mappings.

Cycle Reconstruction Loss Theoretically, if we do not add additional constraints, there exist infinite mappings that can align the distribution of $\tilde{z}_x$ and $\tilde{z}_y$, and thus the learned mappers may be invalid. In order to learn better mappings, we introduce the cycle reconstruction loss and back-translation loss to enhance them.

[1]The "sentence" in this paper can refer to the sequence containing multiple sentences.

Given $\boldsymbol{z_x}$, we first use $M_{\mathcal{X}}$ to map it to the Chinese space, and then use $M_{\mathcal{Y}}$ to map it back:

$$\boldsymbol{z_{x\to y}} = M_{\mathcal{X}}(\boldsymbol{z_x}), \quad \hat{\boldsymbol{z}}_{\boldsymbol{x}} = M_{\mathcal{Y}}(\boldsymbol{z_{x\to y}}) \quad (7)$$

We force $\boldsymbol{z_x}$ and $\hat{\boldsymbol{z}}_{\boldsymbol{x}}$ to be consistent, constrained by the following cycle reconstruction loss:

$$\mathcal{L}_{\text{cyc}_{\mathcal{X}}}\left(\boldsymbol{z_x}, \hat{\boldsymbol{z}}_{\boldsymbol{x}}\right) = \|\boldsymbol{z_x} - \hat{\boldsymbol{z}}_{\boldsymbol{x}}\| \quad (8)$$

The cycle reconstruction loss $\mathcal{L}_{\text{cyc}_{\mathcal{Y}}}$ for $\boldsymbol{z_y}$ and $\hat{\boldsymbol{z}}_{\boldsymbol{y}}$ is similarly defined.

Back-Translation Loss The cycle-reconstructed representation $\hat{\boldsymbol{z}}_{\boldsymbol{x}}$ in Eq. 8 can be regarded as augmented data to train the decoder, which is similar to the back-translation in the Neural Machine Translation area.

Concretely, we use the decoder $\phi_{D_{\mathcal{X}}}$ to read $\hat{\boldsymbol{z}}_{\boldsymbol{x}}$ and generate the hypothetical summary $\hat{\boldsymbol{x}}$. The back-translation loss is defined as the cross-entropy loss between $\hat{\boldsymbol{x}}$ and gold summary $\boldsymbol{x'}$:

$$\begin{aligned}\hat{\boldsymbol{x}} &= \phi_{D_{\mathcal{X}}}(\hat{\boldsymbol{z}}_{\boldsymbol{x}})\\ \mathcal{L}_{\text{back}_{\mathcal{X}}}(\hat{\boldsymbol{z}}_{\boldsymbol{x}}) &= -\sum_{t=1}^{T}\log P\left(x'_t \mid \hat{x}_{<t}, \hat{\boldsymbol{z}}_{\boldsymbol{x}}\right)\end{aligned} \quad (9)$$

The back-translation loss enhances not only the generation ability of the decoder but also the effectiveness of the mapper. The back-translation loss $\mathcal{L}_{\text{back}_{\mathcal{Y}}}$ for $\hat{\boldsymbol{z}}_{\boldsymbol{y}}$ is similarly defined.

Total Loss The total loss for optimizing the encoder, decoder, and mapper of the English side is weighted sum of the above losses:

$$\mathcal{L}_{\mathcal{X}} = \mathcal{L}_{\text{summ}_{\mathcal{X}}} + \lambda_1\mathcal{L}_{gen_{\mathcal{X}}} + \lambda_2\mathcal{L}_{\text{cyc}_{\mathcal{X}}} + \lambda_3\mathcal{L}_{\text{back}_{\mathcal{X}}} \quad (10)$$

where λ_1, λ_2, and λ_3 is the weighted hyper-parameters.

The total loss of the Chinese side is similarly defined, and the complete loss of our model is the sum of English loss and Chinese loss:

$$\mathcal{L} = \mathcal{L}_{\mathcal{X}} + \mathcal{L}_{\mathcal{Y}} \quad (11)$$

The total loss for optimizing the discriminators is:

$$\mathcal{L}_{dis} = \mathcal{L}_{dis_{\mathcal{X}}} + \mathcal{L}_{dis_{\mathcal{Y}}} \quad (12)$$

5 Supervised Training Objective

The supervised training objective contains the same summarization loss in unsupervised training objective (Eq. 3). In addition, it has X-summarization loss and reconstruction loss.

Algorithm 1 Cross-lingual summarization

Input: English summarization data $\mathcal{X}$ and Chinese summarization data $\mathcal{Y}$.

Pre-train English and Chinese monolingual summarization several epochs on $\mathcal{X}$ and $\mathcal{Y}$.
for $i = 0$ **to** max_iters **do**
  Sample a batch from $\mathcal{X}$ and a batch from $\mathcal{Y}$
  if unsupervised **then**
    for $k = 0$ **to** dis_iters **do**
      Update $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$ on $\mathcal{L}_{dis}$ in Eq. 5.
    (a) Update $\phi_{E_{\mathcal{X}}}, \phi_{E_{\mathcal{Y}}}, \phi_{D_{\mathcal{X}}}$, and $\phi_{D_{\mathcal{Y}}}$ on $\mathcal{L}_{\text{summ}}$ in Eq. 3.
    (b) Update $\phi_{E_{\mathcal{X}}}, \phi_{E_{\mathcal{Y}}}, M_{\mathcal{X}}$, and $M_{\mathcal{Y}}$ on $\mathcal{L}_{\text{gen}}$ in Eq. 6.
    (c) Update $\phi_{E_{\mathcal{X}}}, \phi_{E_{\mathcal{Y}}}, M_{\mathcal{X}}$, and $M_{\mathcal{Y}}$ on $\mathcal{L}_{\text{cyc}}$ in Eq. 8.
    (d) Update $M_{\mathcal{X}}, M_{\mathcal{Y}}, \phi_{D_{\mathcal{X}}}$, and $\phi_{D_{\mathcal{Y}}}$ on $\mathcal{L}_{\text{back}}$ in Eq. 9.
  else if supervised **then**
    (a) Upate $\phi_{E_{\mathcal{X}}}, \phi_{E_{\mathcal{Y}}}, \phi_{D_{\mathcal{X}}}$, and $\phi_{D_{\mathcal{Y}}}$ on $\mathcal{L}_{\text{summ}}$ in Eq. 3.
    (b) Update $\phi_{E_{\mathcal{X}}}, \phi_{E_{\mathcal{Y}}}, \phi_{D_{\mathcal{X}}}$, and $\phi_{D_{\mathcal{Y}}}$ on $\mathcal{L}_{\text{xsumm}}$ in Eq. 13.
    (c) Update $\phi_{E_{\mathcal{X}}}, \phi_{E_{\mathcal{Y}}}, M_{\mathcal{X}}$, and $M_{\mathcal{Y}}$ on $\mathcal{L}_{\text{rec}}$ in Eq. 14.

X-Summarization Loss Given a parallel English source text $\boldsymbol{x}$ and Chinese summary $\boldsymbol{y'}$. We use $\phi_{E_{\mathcal{X}}}$, $M_{\mathcal{X}}$, and $\phi_{D_{\mathcal{Y}}}$ to generate the hypothetical Chinese summary $\tilde{\boldsymbol{y}}$, then train them with cross-entropy loss:

$$\boldsymbol{z_x} = \phi_{E_{\mathcal{X}}}(\boldsymbol{x}),\ \boldsymbol{z_{x \to y}} = M_{\mathcal{X}}(\boldsymbol{z_x}),\ \tilde{\boldsymbol{y}} = \phi_{D_{\mathcal{Y}}}(\boldsymbol{z_{x \to y}})$$
$$\mathcal{L}_{\text{xsumm}_{\mathcal{X}}}(\boldsymbol{x}, \boldsymbol{y'}) = -\sum_{t=1}^{T} \log P\left(y'_t \mid \tilde{y}_{<t}, \boldsymbol{x}\right) \tag{13}$$

The X-summarization loss for a Chinese text y and English summary x' is similarly defined.

Reconstruction Loss Since the cross-lingual summarization corpora are constructed by translating the texts to the other language, the English texts and the Chinese texts are parallel to each other. We can build a reconstruction loss to align the sentence representation for the parallel English and Chinese texts.

Specifically, supposing $\boldsymbol{x}$ and $\boldsymbol{y}$ are parallel source English and Chinese texts, we first use $\phi_{E_{\mathcal{X}}}$ and $\phi_{E_{\mathcal{Y}}}$ to obtain contextual representations $\boldsymbol{z_x}$ and $\boldsymbol{z_y}$, respectively. Then we average the contextual representations to get their sentence representations and use the mappers to map them into the other language. Since the English and Chinese texts are translations to each other, the semantics of their sentence representations should be the same. Thus we design the following reconstruction loss:

$$\tilde{z}_x = \frac{1}{m}\sum_{i=1}^{m} z_{x_i}, \quad \tilde{z}_{y \to x} = \frac{1}{n}\sum_{i=1}^{n} \left(z_{y \to x}\right)_i$$
$$\mathcal{L}_{\text{rec}_{\mathcal{X}}}\left(\tilde{z}_x, \tilde{z}_{y \to x}\right) = \left\|\tilde{z}_x - \tilde{z}_{y \to x}\right\| \tag{14}$$

and $\mathcal{L}_{\text{rec}_{\mathcal{Y}}}$ is similarly defined.

Notice that the generative and discriminative loss, cycle-construction loss, and back-translation loss are unnecessary here because we can directly use aligned source text with objective 14 to align the context representations.

Total Loss The total loss for training the English side is:

$$\mathcal{L}_{\mathcal{X}} = \mathcal{L}_{\text{xsumm}_{\mathcal{X}}} + \lambda_1 \mathcal{L}_{\text{summ}_{\mathcal{X}}} + \lambda_2 \mathcal{L}_{\text{rec}_{\mathcal{X}}} \tag{15}$$

where λ_1 and λ_2 is the weighted hyper-parameters. The total loss of the Chinese side is similarly defined.

6 Experiments

6.1 Experiment Settings

We conduct experiments on English-to-Chinese (En-to-Zh) and Chinese-to-English (Zh-to-En) summarizations. Following Duan et al. (2019), we translate the source texts to the other language to form the (pseudo) parallel corpus. Since they do not release their training data, we translate the source text ourselves through the Google translation service. Notice that Zhu et al. (2019) translate the summaries rather than source texts.

Since Duan et al. (2019) use Gigaword and DUC2004 datasets for experiments while Zhu et al. (2019) use LCSTS and CNN/DM for experiments, we conduct experiments on all the 4 datasets. When comparing with Duan et al. (2019) and Zhu et al. (2019), we use the same number of translated parallel data for training. Due to limited computing resources, we only do unsupervised experiments on gigaword and LCSTS datasets.

Notice that the test sets provided by Zhu et al. (2019) are unprocessed, therefore we have to process the test samples they provided ourselves.

6.2 Dataset

Gigaword English Gigaword corpus (Napoles et al., 2012) contains 3.80M training pairs, 2K validation pairs, and 1,951 test pairs. We use the human-translated Chinese source sentences provided by (Duan et al., 2019) to do Zh-to-En tests.

DUC2004 DUC2004 corpus only contains test sets. We use the model trained on gigaword corpus to generate summaries on DUC2004 test sets. We use the 500 human-translated test samples provided by (Duan et al., 2019) to do Zh-to-En tests.

LCSTS LCSTS (Hu et al., 2015) is a Chinese summarization corpus, which contains 2.40M training pairs, 10,666 validation pairs, and 725 test pairs. We use 3K cross-lingual test samples provided by Zhu et al. (2019) to do Zh-to-En tests.

CNN/DM CNN/DM (Hermann et al., 2015) contains 287.2K training pairs, 13.3K validation pairs, and 11.5K test pairs. We use the 3K cross-lingual test samples provided by Zhu et al. (2019) to do En-to-Zh cross-lingual tests.

6.3 Evaluation Metrics

We use ROUGE-1 (unigram), ROUGE-2 (bigram), and ROUGE-L (LCS) F1 scores as the evaluation metrics, which are most commonly used evaluation metrics in the summarization task.

6.4 Competitive Models

For unsupervised cross-lingual summarization, we set the following baselines:

- **Unified** It jointly trains English and Chinese monolingual summarizations in a unified model and uses the first token of the decoder to control whether it generates Chinese or English summaries.
- **Unified+CLWE** It builds a unified model and adopts pre-trained unsupervised cross-lingual word embeddings. The cross-lingual word embeddings are obtained via projecting embeddings from source language to target language. We use Vecmap[2] to learn the cross-lingual word embeddings.

For supervised cross-lingual summarization, we compare our model with (Shen et al., 2018), (Duan et al., 2019), and Zhu et al. (2019). We also consider the following baselines for comparison:

- **Pipe-TS** The Pipe-TS baseline first uses a Transformer-based translation model to translate the source text to the other language, then uses a monolingual summarization model to generate summaries. To make this baseline stronger, we replace the translation model with the Google translation system and name it as **Pipe-TS***.
- **Pipe-ST** The Pipe-ST baseline first uses a monolingual summarization model to generate the summaries, then uses a translation model to translate the summaries to the other language. We replace the translation model with the Google translation system as **Pipe-ST***.
- **Pseudo** The Pseudo baseline directly trains a cross-lingual summarization model by using the pseudo parallel cross-lingual summarization data.
- **XLM Pretraining** This method is proposed by Lample and Conneau (2019), where they pretrain the encoder and decoder on large-scale multilingual text using causal language modeling (CLM), masked language modeling (MLM), and translation language modeling (TLM) tasks. [3]

6.5 Implementation Details

For transformer architectures, we use the same configuration as Vaswani et al. (2017), where the number of layers, model hidden size, feed-forward hidden size, and the number of heads are 6, 512, 1024, and 8, respectively. We set $g = \sqrt{d_{\text{model}}} = \sqrt{512}$ in ScaleNorm. The mapper is a linear layer with a hidden size of 512, and the discriminator is a two-layer linear layer with a hidden size of 2048.

We use the NLTK[4] tool to process English texts and use jieba[5] tool to process Chinese texts. The vocabulary size of English words and Chinese words are 50,000 and 80,000 respectively. We set $\lambda_1 = 1, \lambda_2 = 5, \lambda_3 = 2$ in unsupervised training and $\lambda_1 = 0.5, \lambda_2 = 5$ in supervised training according to the performance of the validation set. We set $dis\_iters = 5$ in Alg. 1.

[2] https://github.com/artetxem/vecmap

[3] This baseline was suggested by the reviewers, and the results are only for reference since it additionally uses a lot of pre-training text.

[4] https://github.com/nltk/nltk

[5] https://github.com/fxsjy/jieba

Method	Zh-to-En									En-to-Zh		
	Gigaword			DUC2004			LCSTS			CNN/DM		
	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
Pipe-TS	22.27	6.58	20.53	21.29	5.96	17.99	27.26	10.41	21.72	-	-	-
Pipe-ST	28.27	11.90	26.50	25.73	8.19	21.60	36.48	18.87	31.44	25.95	11.01	23.29
Pipe-TS*	22.52	6.67	20.76	21.83	6.11	18.42	29.29	11.09	23.18	-	-	-
Pipe-ST*	29.56	12.50	26.42	26.66	8.51	22.37	38.26	19.56	32.93	27.82	11.78	24.97
Pseudo*	30.93	13.25	27.29	27.03	8.49	23.08	38.61	19.76	34.63	35.81	14.96	32.07
(Shen et al., 2018)	21.5	6.6	19.6	19.3	4.3	17.0	-	-	-	-	-	-
(Duan et al., 2019)	30.1	12.2	27.7	26.0	8.0	23.1	-	-	-	-	-	-
(Zhu et al., 2019)	-	-	-	-	-	-	40.34	22.65	36.39	38.25	20.20	34.76
(Zhu et al., 2019) w/ LDC	-	-	-	-	-	-	40.25	22.58	36.21	**40.23**	**22.32**	**36.59**
XLM Pretraining	**32.28**	**14.03**	**28.19**	**28.27**	**9.40**	**23.78**	**42.75**	**22.80**	**38.73**	39.11	17.57	34.14
Ours	32.04	13.60	27.91	27.25	8.71	23.36	40.97	23.20	36.96	38.12	16.76	33.86

Table 1: Rouge F1 scores (%) on cross-lingual summarization tests. "XLM Pretraining" and "Zhu et al. (2019) w/ LDC" use additional training data. Our model significantly ($p < 0.01$) outperforms all pipeline methods and pseudo-based methods.

We use Adam optimizer (Kingma and Ba, 2014) with $\beta = (0.9, 0.98)$ for optimization. We set the learning rate to $3e-4$ and adopt the warm-up learning rate (Goyal et al., 2017) for the first 2,000 steps, the initial warm-up learning is set to $1e-7$. We adopt the dropout technique and set the dropout rate to 0.2.

7 Results and Analysis

7.1 Unsupervised Cross-Lingual Summarization

The experiment results of unsupervised cross-lingual summarization are shown in Table 2, and it can be seen that our model significantly outperforms all baselines by a large margin. By training a unified model of all languages, the model's cross-lingual transferability is still poor, especially for the gigaword dataset. Incorporating cross-lingual word embeddings into the unified model can improve the performance, but the improvement is limited. We think this is due to that the cross-lingual word embeddings learned by Vecmap cannot leverage the contextual information. Due to space limitations, we present case studies in the Appendix.

After checking the generated summaries of the two baseline models, we find that they can generate readable texts, but the generated texts are far away from the theme of the source text. This indicates that the encoder and decoder of these baselines have a large gap, such that the decoder cannot understand the output of the encoder. We also find that summaries generated by our model are obviously more relevant, demonstrating that aligned representations between languages are helpful.

But we can also see that there is still a gap between our unsupervised results (Table 2) and supervised results (Table 1), indicating that our model has room for improvement.

Method	LCSTS			Gigaword		
	R1	R2	RL	R1	R2	RL
Unified	13.52	1.35	10.02	5.25	0.87	2.09
Unified+CLWE	14.02	1.49	12.10	6.51	1.07	2.92
Ours	**20.11**	**5.46**	**16.07**	**13.75**	**4.29**	**11.82**

Table 2: Rouge F1 scores (%) on unsupervised cross-lingual summarization tests. Our model outperforms all baselines significantly ($p < 0.01$).

7.2 Supervised Cross-Lingual Summarization

The experiment results of supervised cross-lingual summarization are shown in Table 1. Due to the lack of corpus for training Chinese long document summarization model, we do not experiment with the Pipe-TS model on the CNN/DM dataset.

By comparing our results with pipeline-based or pseudo baselines, we can find that our model outperforms all these baselines in all cases. Our model achieves an improvement of 0∼3 Rouge scores over the Pseudo model trained directly with translated parallel cross-lingual corpus, and 1.5∼4 Rouge-1 scores over those pipeline models. We also observe that models using the Google translation system all perform better than models using the Transformer-based translation system. This may because the Transformer-based translation system will bring some "UNK" tokens, and the transformer-based translation system trained by ourselves does not perform as well as the Google translation system. In addition, Pipe-ST models perform better than Pipe-TS models, which is con-

Method	Info. ↑	Con. ↑	Flu. ↑
Reference	**3.60**	3.50	3.80
PipeST*	3.56	3.51	**4.00**
PipeTS*	3.37	3.80	3.81
Pseudo	3.27	3.81	3.89
Ours (supervised)	3.56	**3.93**	3.94
Ours (unsupervised)	2.18	3.34	2.87

Table 3: Results of the human evaluation on the gigaword dataset.

Method	Info. ↑	Con. ↑	Flu. ↑
Reference	**3.58**	3.57	**4.21**
PipeST*	3.38	3.45	4.13
PipeTS*	3.38	3.93	3.78
Pseudo	3.46	3.90	4.05
Ours (supervised)	3.55	**4.03**	4.13

Table 4: Results of the human evaluation on the CNN/DM dataset.

sistent with the conclusions of previous work. This is because (1) the translation process may discard some informative clauses, (2) the domain of the translation corpus is different from the domain of summarization corpus, which will bring the domain discrepancy problem to the translation process, and (3) the translated texts are often "translationese" (Graham et al., 2019). The Pseudo model performs better than Pipe-TS models but performs similarly as Pipe-ST models.

By comparing our results with others, we can find that our model outperforms Shen et al. (2018) and Duan et al. (2019) on both gigaword and DUC2004 test sets, and it outperforms Zhu et al. (2019) on the LCSTS dataset. But our Rouge scores are lower than Zhu et al. (2019) on the CNN/DM dataset, especially the Rouge-2 score. However, our model performs worse than pre-trained models.

7.3 Human Evaluation

The human evaluation was also performed. Since we cannot get the summaries generated by other models, we only compare with our baselines in the human evaluation. We randomly sample 50 examples from the gigaword (Zh-to-En) test set and 20 examples from the CNN/DM (En-to-Zh) test set. We ask five volunteers to evaluate the quality of the generated summaries from the following three aspects: (1) **Informative**: how much does the generated summaries cover the key content of the source text? (2) **Conciseness**: how concise are the generated summaries? (3) **Fluency**: how fluent are the generated summaries? The scores are between 1-5, with 5 being the best. We average the scores and show the results in Table 3 and Table 4.

Our model exceeds all baselines in informative and conciseness scores, but get a slightly lower fluency score than Pipe-ST*. We think this is because the Google translation system has the ability to identify grammatical errors and generate fluent sentences.

Method	Gigaword			CNN/DM		
	R1	R2	RL	R1	R2	RL
Ours (supervised)	**32.04**	**13.60**	**27.91**	38.12	**16.76**	33.86
w/o summ. loss	30.36*	12.84*	26.41*	36.37*	15.97*	32.11*
w/o mappers	31.95	13.46	27.88	**38.28**	16.73	**33.93**
w/o ScaleNorm	31.27*	13.29	27.22*	37.01*	16.30*	32.87*
w/o pre. steps	31.33*	13.30	27.35*	37.23*	16.39	33.01*
Unshare enc/dec	30.10*	12.71*	26.28*	35.93*	15.86*	31.82*

Table 5: Results of ablation tests in supervised setting. Statistically significant improvement ($p < 0.01$) over the complete model are marked with *.

7.4 Ablation Tests

To study the importance of different components of our model, we also test some variants of our model. For supervised training, we set variants of (1) without (monolingual) summarization loss, (2) without mappers[6], (3) replace ScaleNorm with LayerNorm, (4) without pre-trained monolingual steps, and (5) unshare the encoder and decoder. For unsupervised training, we additionally set variants without cyc-reconstruction loss or back-translation loss. The results of ablation tests of supervised and unsupervised cross-lingual summarization are shown in Table 5 and Table 6, respectively.

It seems that the role of mappers does not seem obvious in the case of supervised training. We speculate that this may be due to the joint training of monolingual and cross-lingual summarizations, and directly constraining the context representations before mapping can also yield shared (aligned) representations. But mappers are crucial for unsupervised cross-lingual summarization. For supervised cross-lingual summarization, except for mappers, all components contribute to the improvement of the performance. The performance decreases after removing any of the components. For unsupervised cross-lingual summarization, all components contribute to the improvement of the performance and the mappers and shared encoder/decoder are key components.

[6]In this case, we directly constrain the parallel $\boldsymbol{z_x}$ and $\boldsymbol{z_y}$ to be the same.

Method	LCSTS			Gigaword		
	R1	R2	RL	R1	R2	RL
Ours (unsupervised)	**20.10**	**5.46**	**16.07**	**13.75**	**4.29**	**11.82**
w/o mappers	14.79*	2.29*	12.36*	6.26*	1.02*	3.11*
w/o cyc. loss	17.51*	4.70*	13.95*	7.21*	1.31*	4.04*
w/o back. loss	19.37	5.23	15.44	13.20	4.11	11.27
w/o ScaleNorm	19.24*	5.21	15.37*	13.15*	4.08	11.21
w/o pre. steps	19.70	5.24	15.72	13.13	4.10	10.91
Unshare enc/dec	12.28*	0.97*	10.37*	4.88*	0.82*	1.91*

Table 6: Results of the ablation tests of unsupervised cross-lingual summarization. Statistically significant improvement ($p < 0.01$) over the complete model are marked with *.

8 Related Work

8.1 Cross-Lingual Summarization

Early researches on cross-lingual abstractive summarization are mainly based on the monolingual summarization methods and adopt different strategies to incorporate bilingual information into the pipeline model (Leuski et al., 2003; Orasan and Chiorean, 2008; Wan et al., 2010; Wan, 2011; Yao et al., 2015).

Recently, some neural cross-lingual summarization systems have been proposed for cross-lingual summarization (Shen et al., 2018; Duan et al., 2019; Zhu et al., 2019). The first neural-based cross-lingual summarization system was proposed by Shen et al. (2018), where they first translate the source texts from the source language to the target language to form the pseudo training samples. A teacher-student framework is adopted to achieve end-to-end cross-lingual summarization. Duan et al. (2019) adopt a similar framework to train the cross-lingual summarization model, but they translate the summaries rather than source texts to strengthen the teacher network. Zhu et al. (2019) propose a multi-task learning framework by jointly training cross-lingual summarization and monolingual summarization (or machine translation). They also released an English-Chinese cross-lingual summarization corpus with the aid of online translation services.

8.2 Learning Cross-Lingual Representations

Learning cross-lingual representations is a beneficial method for cross-lingual transfer.

Conneau et al. (2017) use adversarial networks to learn mappings between languages without supervision. They show that their method works very well for word translation, even for some distant language pairs like English-Chinese. Lample et al. (2018) learn word mappings between languages to build an initial unsupervised machine translation model, and then perform iterative back-translation to fine-tune the model. Aldarmaki and Diab (2019) propose to directly map the averaged embeddings of aligned sentences in a parallel corpus, and achieve better performances than word-level mapping in some cases.

9 Conclusions

In this paper, we propose a framework that jointly learns to align and summarize for neural cross-lingual summarization. We design training objectives for supervised and unsupervised cross-lingual summarizations, respectively. We also propose methods to enhance the isomorphism and cross-lingual transfer between languages. Experimental results show that our model outperforms supervised baselines in most cases and outperforms unsupervised baselines in all cases.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61772036), Tencent AI Lab Rhino-Bird Focused Research Program (No.JR201953), and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We thank the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2440–2452.

Hanan Aldarmaki and Mona Diab. 2019. Context-aware cross-lingual mapping. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3906–3911.

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

Xilun Chen, Ahmed Hassan, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *arXiv preprint arXiv:1906.09833*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy. 2003. Cross-lingual c* st* rd: English access to hindi information. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3):245–269.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100. Association for Computational Linguistics.

Toan Q Nguyen and Julian Salazar. 2019. Transformers without tears: Improving the normalization of self-attention. *arXiv preprint arXiv:1910.05895*.

Constantin Orasan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual romanian-english multi-document summariser. In *LREC 2008*.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. *arXiv preprint arXiv:1906.05407*.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. *arXiv preprint arXiv:1908.06625*.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, and Mao-song Sun. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(12):2319–2327.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, 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*, 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.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.

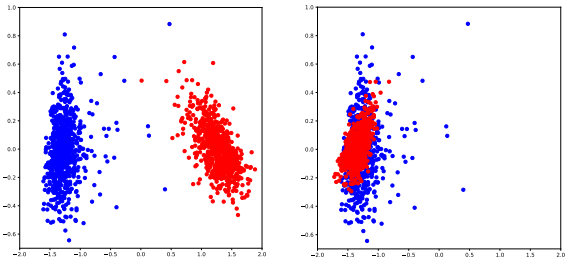
Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 118–127.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3045–3055.

A Visualization

We use the PCA (Wold et al., 1987) algorithm to visualize the pre- and post-aligned context representations of our model in Figure 2. The left picture shows the original distribution of two languages, and the right picture shows the distribution after we map Chinese representations to English.

Figure 2 reveals that the representations of the two languages are originally separated but become aligned after our proposed procedure, which demonstrates that our proposed alignment procedure is effective.



(a) Pre-alignment (b) Post-alignment

Figure 2: Visualization of the pre- and post-aligned context representations. The blue dots are English context representations and the red dots are Chinese context representations.

B Case Studies

We show four cases of Chinese-to-English summarization in Table 7. Since most of the summaries generated by other unsupervised baselines are meaningless (e.g., far away from the theme of the source text, all tokens are "UNK" and so on), we don't show their results here.

Text: 野生动物专家称，除非政府发起全面打击猖獗偷猎的战争，否则印度大象将会灭绝. (*wildlife experts say indian elephants will go extinct unless government launches full-scale war against sting poaching*)
Reference: india elephant may be facing extinction : experts by <unk> **Pipe-ST:** wildfile expert says indian elephant will die out **Pipe-TS:** india to kill elephants in war on poaching **Pseudo:** indian elephants face extinction unless government launches war against poaching **Ours (supervised):** india elephants face extinction over poaching **Ours (unsupervised):** india elephants rise to extinct
Text: 一份媒体报道，一名日本男子周日在台湾上吊自杀，原因是亚洲冠军没能在世界杯上获得一场胜利. (*report claimed that a japanese man hanged himself in taiwan on sunday because the asian champion failed to win a victory at the word cup*)
Reference: fan hangs himself for nation 's dismal world cup performance **Pipe-ST:** japanese man hangs himself in taiwan as asian champion fails to win **Pipe-TS:** world cup winner commits suicide **Pseudo:** man commits suicide because of world cup failure **Ours (supervised):** man hangs himself after world cup failure **Ours (unsupervised):** failed to secure a single champions
Text: 澳大利亚教练罗比-迪恩斯对上周末在这里对阵意大利的袋鼠测试前被新西兰击败的球队做了八次改变. (*australian coach robbie deans made eight changes to a team defeated by new zealand before the kangaroo test against italy here last weekend*)
Reference: <unk> : deans rings changes for aussies azzurri test **Pipe-ST:** australian coach changes team eight times before kangaroo test **Pipe-TS:** australia make eight changes for italy test **Pseudo:** deans makes eight changes for new zealand **Ours (supervised):** australia make eight changes ahead of italy test **Ours (unsupervised):** weekend ahead of wallabies test against Italy here
Text: 凯尔特人中场保罗哈特利在经历了一个星期痛苦的欧洲之旅后，于周五为苏格兰足球发起了一场激情的辩护. (*celtic midfielder paul hartley launched a passionate defence for scottish football on friday after a week of painful european travel*)
Reference: football : scottish football is not a joke says celtic star **Pipe-ST:** paul hartley launches passionate defense **Pipe-TS:** celtic 's hartley launches passionate defense **Pseudo:** celtic 's hartley launches passionate defense for scotland **Ours (supervised):** celtic 's hartley defends scottish football **Ours (unsupervised):** celtic midfielder paul week of european misery

Table 7: Case studies of Chinese-to-English summarization.