This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2842432, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Zero-shot Cross-Lingual Neural Headline Generation

Ayana[1,3], Shi-qi Shen[1], Yun Chen[4], Cheng Yang[1], Zhi-Yuan Liu[1,2*], Mao-song Sun[1,2]
[1]State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]School of Computer Science and Technology, Heilongjiang University, Harbin, China
[3]Department of Computer Information Management, Inner Mongolia University of Finance and Economics, Hohhot, China
[4]Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China
ayn13@mails.tsinghua.edu.cn, {vicapple22, yun.chencreek, albert.yang33}@gmail.com, liuzy@tsinghua.edu.cn; sms@mail.tsinghua.edu.cn

***Abstract*—Neural headline generation (NHG) has been proven to be effective in generating a fully abstractive headline recently. Existing NHG systems are only capable of producing headline of the same language as the original document. Cross lingual headline generation is an important task since it provides an efficient way to understand the key point of a document in a different language. Due to the lack of those parallel corpora of direct source language articles and target language headlines, we propose to deal with the cross-lingual neural headline generation (CNHG) under the zero-shot scenario. A trivial solution is to translate and summarize the source document in a pipeline way. However, a pipeline solution will lead to error propagation in the translation and summarization phases. This challenge motivates us to build a direct source-to-target CNHG model based on existing parallel corpora of translation and monolingual headline generation. Specifically, we let a parameterized CNHG model (student model) mimic the output of a pre-trained translation or headline generation model (teacher model). To the best of our knowledge, this is the first effort to address CNHG problem. Besides, we construct English-Chinese headline generation evaluation datasets by manual translation. Experimental results on English-to-Chinese cross-lingual headline generation demonstrate that our proposed method significantly outperforms the baseline models.**

***Index Terms*—Neural network, headline generation, cross-lingual headline generation.**

I. Introduction

Headline provides an efficient and effective way for people to obtain the subject information before reading through the whole document. The overwhelming globalization is forcing people with tremendous amount of information in various languages. Since mother tongue remains a better and quicker way to acquire information, providing people with vital information in the mother language is preferred. For instance, Figure 1 provides a news fragment with two headlines, one in English and one in Chinese. To quickly understand the main idea of the text, people usually take a glimpse at the headline first. And those who familiar with English would pay attention to the English headline and vice-versa.

Asian-Pacific summit faces major economic and political challenges
亚 太 首脑 会议 面临 重大 经济 和 政治 挑战

The last time the Asia-Pacific region held its annual summit to promote free trade, Japan's prime minister assured everyone that his economy wouldn't be the next victim of Asia's financial crisis …

Fig. 1: A news article paired with headlines in two languages.

Cross-lingual headline generation aims to produce a headline in a target language (e.g., Chinese) given a document in a different source language (e.g., English). Cross-lingual headline generation is important for efficient information acquisition but has not been well studied. The most related task is cross-lingual summarization [1], [2], [3], [4], in which existing studies focus only on single sentence extraction or compression. Nevertheless, none of them is suitable for headline generation because a good headline needs to present the most valuable information within a short length limit and attract attention using sharp words at the same time. Simply extracting a key sentence from the original document and then translating it into target language is difficult to meet the length limit request. Moreover, the compression procedure could address the problem by getting rid of unnecessary words, but the remaining words may not be sharp enough to present the core idea of the original article.

With the successful application of neural networks in various natural language processing tasks [5], [6], [7], [8], NHG also benefits from neural networks [9], [10], [11], [12], [13], [14] and achieves promising performance. NHG leverages a single, large neural network to generate a headline based on an input document directly. However, previous studies only focus on the same-language headline generation, i.e., the input document and the output headline are in the same language. It is difficult to generate headlines in a different

This work is funded by the Natural Science Foundation of China (NSFC) and the German Research Foundation (DFG) in Project Crossmodal Learning, NSFC 61621136008 / DFC TRR-169, Microsoft Research Asia FY17-RES-THEME-017, and China Association for Science and Technology (2016QNRC001).

*Corresponding author

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2842432, IEEE/ACM Transactions on Audio, Speech, and Language Processing

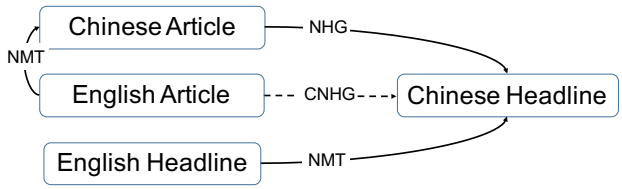

Fig. 2: Model overview of the proposed CNHG methods. A dashed line represents the intended student model without direct parallel corpora, a solid line denotes a pre-trained teacher model with existing parallel corpora.

language because the training data for cross-lingual NHG model is unavailable. One possible solution to this problem is to leverage the pipeline method: We could either translate the original document to target language then generate the headline, or generate the headline then accomplish the translation. Either way involves two models. The training data discrepancy between the two models would have significant influence over the results. Besides, the errors made in the first step would inevitably affect the second step.

To address the model discrepancy and error propagation problems in pipeline methods, we propose a direct source-to-target cross-lingual neural headline generation (CNHG) model and deploy on the task of English-Chinese headline generation based on existing parallel corpora: English headline generation, Chinese headline generation, and English-Chinese translation corpora. As shown in Figure 2, our basic idea is to let the pre-trained neural models with parallel corpora (teacher model) guide the parameter learning of CNHG model without parallel corpora (student model). We investigate three methods for zero-resource CNHG under the teacher-student framework. The first method takes an English-Chinese neural machine translation (NMT) model which is pre-trained on English-Chinese translation corpora as the teacher model. We force the CNHG model to mimic the output of NMT on English headlines. In the second method, we first build a pseudo Chinese article corpora by translating English articles using the NMT model. Then we let the CNHG model imitate the output of the Chinese NHG teacher model, which is pre-trained on Chinese headline generation corpora, on the pseudo Chinese articles. The third method combines them to guide the learning procedure. Our CNHG model directly learns from a pseudo Chinese headline ground truth generated by pre-trained models. We expect that our unified CNHG model can do better on unseen data compared with pre-trained pipeline methods by getting rid of error propagation and model discrepancy problems.

To test the performance of CNHG methods, we build evaluation datasets by manually translating broadly used DUC2003 task-1 and DUC2004 task-1 datasets. These datasets are hoped to benefit future research studies. Experimental results demonstrate that the proposed approaches yield substantial gains over the baseline methods.

II. Background

A. Encoder-Decoder based NHG

Given an input document $\mathbf{x} = (\mathrm{x}_1, \cdots, \mathrm{x}_i, \cdots, \mathrm{x}_M)$, the NHG model aims to take $\mathbf{x}$ as input, and generates a short headline $\mathbf{y} = (\mathrm{y}_1, \cdots, \mathrm{y}_j, \cdots, \mathrm{y}_N)$ with length $N < M$. x_i and y_j represents i-th input word and j-th output word, respectively. The log conditional probability can be formalized as:

$$\log \Pr(\mathbf{y}|\mathbf{x};\theta) = \sum_{j=1}^{N} \log \Pr(\mathrm{y}_j|\mathbf{x}, \mathbf{y}_{<j};\theta), \tag{1}$$

where $\mathbf{y}_{<j} = (\mathrm{y}_1, \cdots, \mathrm{y}_{j-1})$ is partial headline and θ is a set of parameters. The j-th word y_j in a headline is generated in an encoder-decoder framework:

$$\Pr(\mathrm{y}_j|\mathbf{x}, \mathbf{y}_{<j};\theta) \propto \exp\{g(\mathrm{y}_{j-1}, \mathrm{c}_j, \mathrm{s}_j;\theta)\}, \tag{2}$$

where y_{j-1} is the headline word that generated in the last timestep, s_j represents the j-th hidden state computed by the decoder, c_j indicates the j-th context vector for generating y_j, and $g(\cdot)$ is a non-linear function. $\Pr(\cdot)$ represents the function to calculate the generation probability in the rest of the paper. The model parameters are trained to maximize the log likelihood over a large-scale parallel corpora $\mathcal{D} = \{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}_{t=1}^{T}$:

$$\mathcal{L}(\theta) = \sum_{t=1}^{T} \log \Pr(\mathbf{y}^{(t)}|\mathbf{x}^{(t)};\theta), \tag{3}$$

where $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t)}$ denote t-th source document and target headline, T represents corpus size.

For the English NHG task, the widely adopted training data [9] comes from Gigaword [15][1]. For Chinese, the training data is from [16]. Nevertheless, direct large-scale cross-lingual article-to-headline training data is nonexistent, which is hindering the exploration of end-to-end CNHG model.

B. Teacher-Student Framework

A teacher-student framework is usually involved with model distillation in which the student model is trained to simulate the output of a teacher model, or ensemble of teachers. The standard training approach is to minimize the distance (typically L_2, cross entropy or KL-divergence) between the student and teacher model:

$$\mathcal{J}(\theta_T;\theta_S) = G(\Pr(\mathbf{y}|\mathbf{x};\theta_T), \Pr(\mathbf{y}|\mathbf{x};\theta_S)), \tag{4}$$

where $G(\cdot)$ is a function that measures the distance between two distribution probabilities, θ_S and θ_T represent student and teacher network parameters. In our work, we take the intended CNHG model as the student model, pre-trained NMT and NHG models as the teacher models, KL-divergence as the distance measure to accomplish CNHG task.

[1]The access link of English Gigaword is https://catalog.ldc.upenn.edu/LDC2012T21

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2842432, IEEE/ACM Transactions on Audio, Speech, and Language Processing

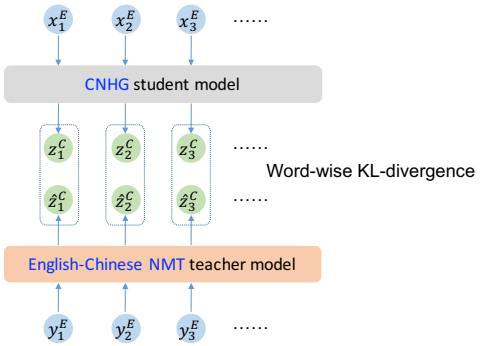

Fig. 3: The zero-resource CNHG student model with the English-Chinese NMT model as the teacher model. Notations are described in Table I.

III. Model

We propose to model the intended English-Chinese CNHG model based on a teacher-student framework with no direct training data. Given an English-Chinese translation parallel corpus $\mathcal{D}_{\mathbf{x^{ES}},\mathbf{y^{CS}}}$, we obtain an NMT teacher model $\Pr(\mathbf{y^{CS}}|\mathbf{x^{ES}};\hat{\theta}_{\text{NMT}})$, where $\hat{\theta}_{\text{NMT}}$ is a set of learned parameters. We also acquire an NHG teacher model $\Pr(\mathbf{y^C}|\mathbf{x^C};\hat{\theta}_{\text{NHG}})$, based on a given Chinese article-headline parallel corpus $\mathcal{D}_{\mathbf{x^C},\mathbf{y^C}}$. $\hat{\theta}_{\text{NHG}}$ refers to the Chinese NHG model parameters. Then, the teacher models guide the learning of student model, i.e., the $\Pr(\mathbf{z^C}|\mathbf{x^E};\theta_{\text{CNHG}})$ with the English article-headline parallel corpus $\mathcal{D}_{\mathbf{x^E},\mathbf{y^E}}$ under three assumptions: (a) the translation of a target headline would have the same generation probability with the intended cross language headline, (b) the headline of a translated source article would have the same generation probability with the intended cross language headline, (c) combining two teacher models as in (a) and (b) would improve the performance of the target model. Table I describes the notations used in our methods.

A. NMT Teacher Model

The underlying assumption is that if an article $\mathbf{x^E}$ and a headline $\mathbf{y^E}$ constitute a parallel article-headline pair, then the intended cross-lingual headline $\hat{\mathbf{z}}^{\mathbf{C}}$ of $\mathbf{x^E}$ should have the same generation probability with the translation of the headline $\mathbf{y^E}$. Figure 3 shows the illustration. Given the English article-headline parallel corpus $\mathcal{D}_{\mathbf{x^E},\mathbf{y^E}}$, the training objective based on the assumption is defined as:

$(\mathbf{x^E},\mathbf{y^E})$	English article and headline of parallel corpus for English NHG model
$(\mathbf{x^{ES}},\mathbf{y^{CS}})$	English sentence and Chinese sentence of parallel corpus for English-Chinese NMT model
$(\mathbf{x^C},\mathbf{y^C})$	Chinese article and headline of parallel corpus for Chinese NHG model
$\hat{\mathbf{x}}^{\mathbf{C}}$	Chinese translation of English article
$\hat{\mathbf{y}}^{\mathbf{C}}$	Chinese translation of English headline
$\hat{\mathbf{z}}^{\mathbf{C}}$	Intended Chinese headline

TABLE I: Notation table.

$$\mathcal{J}_{\text{NMT}}(\theta_{\text{CNHG}}) = \sum_{\langle \mathbf{x^E},\mathbf{y^E}\rangle} \mathbb{E}_{\hat{\mathbf{z}}^{\mathbf{C}}|\mathbf{y^E};\hat{\theta}_{\text{NMT}}}\left[K(\mathbf{x^E},\mathbf{y^E},\hat{\mathbf{z}}^{\mathbf{C}},\hat{\theta}_{\text{NMT}},\theta_{\text{CNHG}})\right], \tag{5}$$

where

$$K(\mathbf{x^E},\mathbf{y^E},\hat{\mathbf{z}}^{\mathbf{C}},\hat{\theta}_{\text{NMT}},\theta_{\text{CNHG}}) = \sum_{j=1}^{|\hat{\mathbf{z}}^{\mathbf{C}}|} \text{KL}\left((\Pr(\hat{z}_j^{\text{C}}|\mathbf{y^E},\hat{\mathbf{z}}^{\mathbf{C}}_{<j};\hat{\theta}_{\text{NMT}})||\Pr(\hat{z}_j^{\text{C}}|\mathbf{x^E},\hat{\mathbf{z}}^{\mathbf{C}}_{<j};\theta_{\text{CNHG}})\right). \tag{6}$$

where $\text{KL}(\cdot)$ is the KL-divergence function which is defined at word-level:

$$\text{KL}\left((\Pr(\hat{z}_j^{\text{C}}|\mathbf{y^E},\hat{\mathbf{z}}^{\mathbf{C}}_{<j};\hat{\theta}_{\text{NMT}})||\Pr(\hat{z}_j^{\text{C}}|\mathbf{x^E},\hat{\mathbf{z}}^{\mathbf{C}}_{<j};\theta_{\text{CNHG}})\right) = \sum_{\hat{z}_j^{\text{C}}\in\mathcal{V}_{z^C}} \Pr(\hat{z}_j^{\text{C}}|\mathbf{y^E},\hat{\mathbf{z}}^{\mathbf{C}}_{<j};\hat{\theta}_{\text{NMT}}) \log \frac{\Pr(\hat{z}_j^{\text{C}}|\mathbf{y^E},\hat{\mathbf{z}}^{\mathbf{C}}_{<j};\hat{\theta}_{\text{NMT}})}{\Pr(\hat{z}_j^{\text{C}}|\mathbf{x^E},\hat{\mathbf{z}}^{\mathbf{C}}_{<j};\theta_{\text{CNHG}})}, \tag{7}$$

in which $\mathcal{V}_{z^C}$ denotes the Chinese vocabulary. Since the NMT teacher model is fixed, the training objective is equivalent to:

$$\mathcal{J}_{\text{NMT}}(\theta_{\text{CNHG}}) = -\sum_{\langle \mathbf{x^E},\mathbf{y^E}\rangle} \mathbb{E}_{\hat{\mathbf{z}}^{\mathbf{C}}|\mathbf{y^E};\hat{\theta}_{\text{NMT}}}\left[R(\mathbf{x^E},\mathbf{y^E},\hat{\mathbf{z}}^{\mathbf{C}},\hat{\theta}_{\text{NMT}},\theta_{\text{CNHG}})\right], \tag{8}$$

where

$$R(\mathbf{x^E},\mathbf{y^E},\hat{\mathbf{z}}^{\mathbf{C}},\hat{\theta}_{\text{NMT}},\theta_{\text{CNHG}}) = \sum_{j=1}^{|\hat{\mathbf{z}}^{\mathbf{C}}|} \sum_{\hat{z}_j^{\text{C}}\in\mathcal{V}_{z^C}} \Pr(\hat{z}_j^{\text{C}}|\mathbf{y^E},\hat{\mathbf{z}}^{\mathbf{C}}_{<j};\hat{\theta}_{\text{NMT}}) \times \log\Pr(\hat{z}_j^{\text{C}}|\mathbf{x^E},\hat{\mathbf{z}}^{\mathbf{C}}_{<j};\theta_{\text{CNHG}}). \tag{9}$$

Then, our goal is to find a set of parameters that minimizes the training objective:

$$\hat{\theta}_{\text{CNHG}} = \arg\min_{\theta_{\text{CNHG}}}\left\{\mathcal{J}_{\text{NMT}}(\theta_{\text{CNHG}})\right\}. \tag{10}$$

Given the learned model parameters $\hat{\theta}_{\text{NMT}}$, the standard decision rule for finding the translation with the highest probability for a headline $\mathbf{y^E}$ is given by:

$$\hat{\mathbf{z}}^{\mathbf{C}} = \arg\max_{\hat{z}^C}\left\{\Pr(\hat{\mathbf{z}}^{\mathbf{C}}|\mathbf{y^E};\hat{\theta}_{\text{NMT}})\right\}. \tag{11}$$

B. NHG Teacher Model

Instead of taking the NMT model as the teacher model, we further investigate to utilize an NHG model as the teacher model, as shown in Figure 4. We assume that if $\hat{\mathbf{x}}^{\mathbf{C}}$ is the translation of an article $\mathbf{x^E}$, then the intended cross-lingual headline $\hat{\mathbf{z}}^{\mathbf{C}}$ would have the same generation probability with the headline of $\hat{\mathbf{x}}^{\mathbf{C}}$. The training objective definition based on this assumption is similar to Eq.(5), and the KL-divergence is defined at word-level as well. As the NHG teacher model

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2842432, IEEE/ACM Transactions on Audio, Speech, and Language Processing

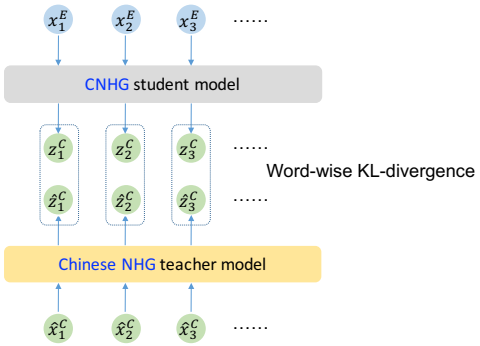

Fig. 4: The zero-resource CNHG student model with the Chinese NHG model as the teacher model. Notations are described in Table I.

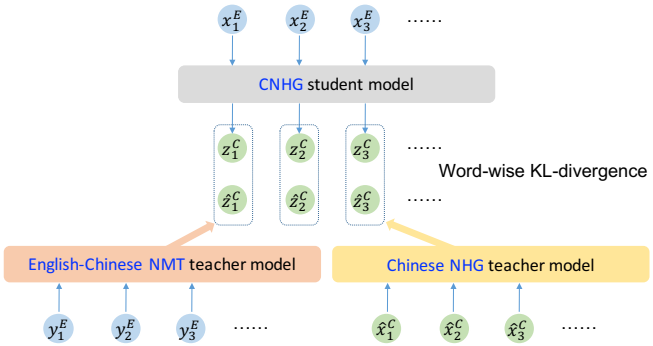

Fig. 5: The zero-resource CNHG student model with the English-Chinese NMT model and the Chinese NHG model as the teacher models. Notations are described in Table I.

does not change during training, the final training objective is equivalent to:

$$\mathcal{J}_{\mathrm{NHG}}(\theta_{\mathrm{CNHG}}) = -\sum_{\langle \mathbf{x}^{\mathbf{E}}, \mathbf{y}^{\mathbf{E}} \rangle} \mathbb{E}_{\hat{\mathbf{z}}^{\mathbf{C}}|\hat{\mathbf{x}}^{\mathbf{C}};\hat{\theta}_{\mathrm{NHG}}} \left[Q(\mathbf{x}^{\mathbf{E}}, \hat{\mathbf{x}}^{\mathbf{C}}, \hat{\mathbf{z}}^{\mathbf{C}}, \hat{\theta}_{\mathrm{NHG}}, \theta_{\mathrm{CNHG}}) \right], \tag{12}$$

where

$$\begin{aligned} &Q(\mathbf{x}^{\mathbf{E}}, \hat{\mathbf{x}}^{\mathbf{C}}, \hat{\mathbf{z}}^{\mathbf{C}}, \hat{\theta}_{\mathrm{NHG}}, \theta_{\mathrm{CNHG}}) \\ &= \sum_{j=1}^{|\hat{\mathbf{z}}^{\mathbf{C}}|} \sum_{\hat{z}_j^{\mathrm{C}} \in \mathcal{V}_{z^C}} \Pr(\hat{z}_j^{\mathrm{C}}|\hat{\mathbf{x}}^{\mathbf{C}}, \hat{\mathbf{z}}^{\mathbf{C}}_{<j}; \hat{\theta}_{\mathrm{NHG}}) \\ &\qquad \times \log \Pr(\hat{z}_j^{\mathrm{C}}|\hat{\mathbf{x}}^{\mathbf{C}}, \hat{\mathbf{z}}^{\mathbf{C}}_{<j}; \theta_{\mathrm{CNHG}}). \end{aligned} \tag{13}$$

The goal becomes finding a set of parameters that minimizes the training objective:

$$\hat{\theta}_{\mathrm{CNHG}} = \arg\min_{\theta_{\mathrm{CNHG}}} \{\mathcal{J}_{\mathrm{NHG}}(\theta_{\mathrm{CNHG}})\}. \tag{14}$$

Finding the translation $\hat{\mathbf{x}}^{\mathbf{C}}$ of input article $\mathbf{x}^{\mathbf{E}}$ is given by:

$$\hat{\mathbf{x}}^{\mathbf{C}} = \arg\max_{\hat{\mathbf{x}}^{\mathbf{C}}} \left\{ \Pr(\hat{\mathbf{x}}^{\mathbf{C}}|\mathbf{x}^{\mathbf{E}}; \hat{\theta}_{\mathrm{NMT}}) \right\}. \tag{15}$$

and finding the headline for a translation $\hat{\mathbf{x}}^{\mathbf{C}}$ under the Chinese NHG model $\hat{\theta}_{\mathrm{NHG}}$ is given by:

$$\hat{\mathbf{z}}^{\mathbf{C}} = \arg\max_{\hat{\mathbf{z}}^{\mathbf{C}}} \left\{ \Pr(\hat{\mathbf{z}}^{\mathbf{C}}|\hat{\mathbf{x}}^{\mathbf{C}}; \hat{\theta}_{\mathrm{NHG}}) \right\}. \tag{16}$$

C. NMT+NHG Teacher Model

As the NMT and NHG models both can be utilized as the teacher model to guide the student model, we further investigate the combined "teaching" ability of the two models, as shown in Figure 5. The training objective consists of two parts: KL-divergence between NMT teacher model and student model, KL-divergence between NHG teacher model and student model. In this way, our approach is capable of considering the two pre-trained teacher models.

$$\begin{aligned} &\mathcal{J}_{\mathrm{NMT+NHG}}(\theta_{\mathrm{CNHG}}) \\ &= -\sum_{\langle \mathbf{x}^{\mathbf{E}}, \mathbf{y}^{\mathbf{E}} \rangle} \Big\{ \alpha \mathbb{E}_{\hat{\mathbf{z}}^{\mathbf{C}}|\mathbf{y}^{\mathbf{E}};\hat{\theta}_{\mathrm{NMT}}} \left[R(\mathbf{x}^{\mathbf{E}}, \mathbf{y}^{\mathbf{E}}, \hat{\mathbf{z}}^{\mathbf{C}}, \hat{\theta}_{\mathrm{NMT}}, \theta_{\mathrm{CNHG}}) \right] \\ &+ (1-\alpha) \mathbb{E}_{\hat{\mathbf{z}}^{\mathbf{C}}|\hat{\mathbf{x}}^{\mathbf{E}};\hat{\theta}_{\mathrm{NHG}}} \left[Q(\mathbf{x}^{\mathbf{E}}, \hat{\mathbf{x}}^{\mathbf{C}}, \hat{\mathbf{z}}^{\mathbf{C}}, \hat{\theta}_{\mathrm{NHG}}, \theta_{\mathrm{CNHG}}) \right] \Big\} \end{aligned} \tag{17}$$

where the definition of functions $R(\cdot)$ and $Q(\cdot)$ remain the same as in Eq.(9) and Eq.(13), respectively. α is a hyper-parameter that balances between the NMT and NHG models.

D. Teacher Distribution Approximation

When we utilize the teacher model to generate a teacher distribution, in Eq.(11) for instance, there is a major difficulty, i.e., the intractable search space. Enumerating all possible sequences is impossible. As a result, we utilize three approximation methods to address this problem.

Single Word Embedding Sampling This is a standard solution to approximate the full search space [17]. Let $\hat{\mathbf{z}}^{\mathbf{C}} = (\cdots, \hat{\mathrm{z}}^{\mathrm{C}}_{j-1}, \hat{\mathrm{z}}^{\mathrm{C}}_{j}, \hat{\mathrm{z}}^{\mathrm{C}}_{j+1}, \cdots)$ be a sampled headline. When calculating the output distribution $\Pr(\hat{\mathrm{z}}^{\mathrm{C}}_{j}|\mathbf{y}^{\mathbf{E}}, \hat{\mathbf{z}}^{\mathbf{C}}_{<j}; \hat{\theta}_{\mathrm{NMT}})$ according to Eq.(6), the single word embedding sampling method utilizes the word embedding of $\hat{\mathrm{z}}^{\mathrm{C}}_{j-1}$. The word $\hat{\mathrm{z}}^{\mathrm{C}}_{j-1}$ is selected based on a multinomial distribution defined by $\Pr(\hat{z}^{\mathrm{C}}_{j-1})$. We denote it as **Single sampling**. This method not only could introduce more diverse data, but also is less time consuming.

Expected Word Embeddings Sampling The single word embedding sampling method only takes one past-sampled word as input during generating. As a result, it inevitably brings error propagation problem: sampling mistakes will affect the next step of sampling. Inspired by [18], we utilize the expected word embeddings to address the problem. In this method, the emitted output word embedding that used to calculate the j-th output distribution $\Pr(\hat{\mathrm{z}}^{\mathrm{C}}_{j}|\mathbf{y}^{\mathbf{E}}, \hat{\mathbf{z}}^{\mathbf{C}}_{<j}; \hat{\theta}_{\mathrm{NMT}})$ is the weighted average of word embeddings over the entire target vocabulary, according to the probability distribution $\Pr(\hat{\mathrm{z}}^{\mathrm{C}}_{j-1})$, instead of the single word embedding itself. The expected word embeddings allow us to consider the full vocabulary and approximate the full search space better. We denote it as **Expected sampling**.

Greedy Decoding Another simpler way to approximate the teacher distribution is to utilize the mode of the teacher model, as suggested in [19]. This method simply runs beam search with beam size = 1 on $\mathbf{y}^{\mathbf{E}} \in \mathcal{D}_{\mathbf{x}^{\mathbf{E}},\mathbf{y}^{\mathbf{E}}}$ to obtain the teacher distribution. Specifically, at each time step of decoding, this method always chose the word with the highest generation probability to be the input of the next time step. This method is a greedy decoding process, and we denote it as **Greedy decoding**.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2842432, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Dataset	Statistics		
	art.num	art.avg.tok	head.avg.tok
DUC2003	624	35.37	11.17
DUC2004	500	35.56	11.37

TABLE II: Data statistics of the DUC datasets. The art.num, art.avg.tok and head.avg.tok refer to article numbers, average token numbers in each article and average token numbers in each translated headline respectively.

IV. Experiments

A. Setup

Teacher Model Setup We evaluate our approaches on English-Chinese headline generation task. Note that in our approaches, there are two pre-trained teacher models involved, i.e., the English-Chinese NMT model and the Chinese NHG model.

For the English-Chinese NMT model, the training set[2] consists of 1.25M sentence pairs with 27.9M Chinese words and 34.5M English words. We use the NIST 2002 dataset as the development set to select model parameters. The evaluation metric is BLEU [20], calculated by the *multi-bleu.perl* script.

For the Chinese NHG model, the training set and development set all come from LCSTS [16]. This dataset provides Chinese article-headline pairs, which are collected from Sina Weibo[3]. We utilize the word-based Chinese NHG model, and the word segmentation is accomplished by THULAC[4]. The Part-I of LCSTS serves as the training set, which consists of 2.40M article-headline pairs with 131.02M article words and 21.86M headline words. We randomly sample 800 headline-article pairs from the Part-II LCSTS as the development set to pick model parameters. The evaluation metric is ROUGE [21] which is calculated by *ROUGE-1.5.5.pl* script.

Student Model Setup For the intended English-Chinese CNHG model, we leverage the English Gigaword [15] to build the training data, as introduced in [9][5]. This training set includes 3.8M article-headline pairs with 11.91M article words and 3.13M headline words. This dataset is also used as the training data for an English NHG model which we adopt in the baseline method.

For all involved models, including the teacher and student models, we adopt the same model architecture. Specifically, a bi-directional GRU recurrent neural network [22] is used as the encoder and an attention based GRU recurrent neural network [6] is utilized as the decoder. We use the AdaDelta algorithm [23] for optimizing model parameters. The English-Chinese NMT model, the Chinese NHG model, and the English-Chinese CNHG model share the same Chinese vocabulary, and the size is limited to 50K. English vocabulary size is limited to 30k. The word embedding size, the encoder, and the decoder hidden state dimension are set as 512, 1024 and 1024 respectively. When inferencing, the beam-size is set to 5.

Data Construction For validating and testing the performance of the intended CNHG model, we manually translate DUC2003 task-1 and DUC2004 task-1 data. One professional translator is requested to generate translations for reference headlines. We give the following rules to guide the translation: (1) Let the translations faithful to the original headlines. (2) Make the translations as concise as possible. (3) Keep the proper nouns unchanged to avoid ambiguity. Although in the original DUC2003 and DUC2004 data, each article is paired with 4 reference headlines, we only translate 1 reference headline. We take DUC2003 dataset as the development dataset to select model parameters and take DUC2004 as the test set. Table II shows detailed statistics.

Evaluation Metric The evaluation metric we utilize for headline generation tasks is the ROUGE [21], which reports recall, precision and F1 scores. Previous studies related to English headline generation either report recall score with length limit or full-length F1 scores from ROUGE-1, ROUGE-2 and ROUGE-L [9], [12], [14]. For Chinese headline generation task, the authors usually report full-length F1 scores [16], [10]. The recall scores of ROUGE are sensitive to length as it favors longer headlines. The F1 scores, on the other hand, could provide fairer results by penalizing longer headlines that are noisy [14]. In our work, we use full-length F1 scores from ROUGE-1, ROUGE-2 and ROUGE-L to evaluate our systems for fair comparison.

B. Baselines

We compare our approaches with the following baseline methods:

1) **Baseline-TS** (**T**ranslate then **S**ummarize): Specifically, this method first utilizes the English-Chinese NMT model $\Pr(\hat{\mathbf{x}}^{\mathbf{C}}|\mathbf{x}^{\mathbf{E}};\hat{\theta}_{\mathrm{NMT}})$ to translate English articles into Chinese, then obtain the corresponding headline with the Chinese NHG model $\Pr(\mathbf{z}^{\mathbf{C}}|\hat{\mathbf{x}}^{\mathbf{C}};\hat{\theta}_{\mathrm{NHG}})$.
2) **Baseline-ST** (**S**ummarize then **T**ranslate): Instead of translating first, this method utilizes an English NHG model $\Pr(\hat{\mathbf{y}}^{\mathbf{E}}|\mathbf{x}^{\mathbf{E}};\hat{\theta}_{\mathrm{NHG\_EN}})$ to obtain the English headline, then generates the corresponding Chinese headline with the English-Chinese NMT model $\Pr(\hat{\mathbf{z}}^{\mathbf{C}}|\hat{\mathbf{y}}^{\mathbf{E}};\hat{\theta}_{\mathrm{NMT}})$.
3) **Baseline-PSEUDO**: This builds a pseudo parallel corpora adapted for CNHG. We translate the English headlines $\mathbf{y}^{\mathbf{E}}$ from the English article-headline corpus $\mathcal{D}_{\mathbf{x}^{\mathbf{E}},\mathbf{y}^{\mathbf{E}}}$ into Chinese headlines using the English-Chinese NMT model $\Pr(\hat{\mathbf{y}}^{\mathbf{C}}|\mathbf{y}^{\mathbf{E}};\hat{\theta}_{\mathrm{NMT}})$ with greedy decoding. The original English articles $\mathbf{y}^{\mathbf{E}}$ from $\mathcal{D}_{\mathbf{x}^{\mathbf{E}},\mathbf{y}^{\mathbf{E}}}$ and translated Chinese headlines $\hat{\mathbf{y}}^{\mathbf{C}}$ constitute a pseudo English-Chinese headline generation corpus $\hat{\mathcal{D}}_{\mathbf{x}^{\mathbf{E}},\hat{\mathbf{y}}^{\mathbf{C}}}$.

C. Effect of Approximation Method

We propose three methods to approximate the teacher distribution. To investigate the performance of different approximation methods, we conduct experiments on top of the

[2]The training set includes LDC2002E18, LDC2003E07, LDC2003E14, part of LDC2004T07, LDC2004T08 and LDC2005T06.

[3]http://www.weibo.com

[4]http://thulac.thunlp.org/

[5]The corresponding pre-processing script is available at https://github.com/facebookarchive/NAMAS

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2842432, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Approximation methods	$R1$	$R2$	RL
Single sampling	13.92	2.92	12.99
Expected sampling	13.43	2.74	12.53
Greedy decoding	**14.14**	**3.04**	**13.32**

TABLE III: Effect of using different approximation methods on DUC2003 development dataset. $R1$, $R2$ and RL refer to F1 score of ROUGE-1, ROUGE-2 and ROUGE-L respectively.

NMT teacher model, and Table III shows the experimental results on DUC2003 development dataset.

We observe that the Single sampling method performs better than the Expected sampling method. One possible explanation is that when we utilize the expected word embeddings to approximate the teacher distribution, the teacher model itself does not change during training. The decoder side word embeddings remain fixed rather than tuned along with the training.

The greedy decoding method outperforms the other two methods over all ROUGE scores. In this method, when the student model is imitating the teacher model, only the candidate with the highest probability in the full space is used to predict the next output word at each time step. This finding suggests that the locally optimal choice at each stage is more important than data diversity in CNHG method. According to the experimental results, we use greedy decoding as the default approximation method in our experiments.

D. Effect of Hyperparameter α

To let the student CNHG model simultaneously learn from the English NMT model and the Chinese NHG model, we set a hyper-parameter α to balance between them. To explore the influence of hyper-parameter α on the performance of CNHG model, we conduct experiments with the α set as 0.1, 0.3, 0.5, 0.7, and 0.9 respectively. The higher the value, the more the student model learn from the English NMT model. The experimental results are shown in Table V. We observe that when α is set to 0.7, the model achieves the highest ROUGE-1, ROUGE-2 and ROUGE-L scores. The scores are also higher than the NMT teacher models on ROUGE-1 and ROUGE-L scores, and the NHG teacher model over all ROUGE scores. One possible reason is that using one teacher model may lead to over-fitting during training. Combining two models with proper ratio would reduce the uncertainty of the model and make the overall performance more stable and consistent. As a result, we use $\alpha = 0.7$ in the following experiments.

E. Main Results

Table IV shows the headline generation performance on DUC datasets. We have the following observations.

Firstly, the Baseline-ST method performs better than the Baseline-TS method. This can be partly attributed to the model discrepancy problem: the English-Chinese NMT and the Chinese NHG models are quite different in terms of vocabulary and parameter space because the English-Chinese translation and Chinese article-headline parallel corpora are loosely-related or even unrelated.

Secondly, the Baseline-TS and Baseline-ST methods achieve lower performance than those of other approaches (except for the NHG method). It is mainly because the Baseline-ST and Baseline-TS systems are pipeline methods and they inevitably suffer from cascaded translation errors: the mistakes made in the first step will be propagated to the second step.

Thirdly, the Baseline-PSEUDO and NHG teacher method obtain general performance. Although Baseline-PSEUDO and NHG teacher method are not pipeline based methods, they are trained with a pseudo corpus. They only use the source translation with the highest probability to build the pseudo parallel corpus, which may also cause severe error propagation problem in training.

Finally, the NMT teacher and NMT+NHG teacher methods significantly outperform the other methods on both DUC2003 and DUC2004 datasets. This finding suggests that performance of CNHG approach greatly benefits from the direct training and training corpora with less noise.

F. Case Study

Table VI shows headline examples of baseline systems and our proposed methods.

In the Baseline-ST method, the headline generation step outputs redundant words “News analysis”, “(by xiong UNK) UNK UNK contributed reporting”. As a result, the corresponding translation includes redundant words “新闻 分析 (*News analysis*)”, “(UNK [10]” as well. In Baseline-TS method, the error of missing key information “Asia-Pacific region” error is propagated to the next headline generation step. This observation indicates that the errors made in the first step would inevitably propagate to the second step in pipeline methods.In pipeline methods, the “UNK” generated in the first step would inevitably affect the second step. Table VII demonstrates the “UNK” statistics in the pipeline methods.

The Baseline-PSEUDO and NHG teacher methods repeat words “亚洲 金融 危机 (*Asian financial crisis*)” and “经济 (*economy*)” respectively. This suggests that the pseudo corpora may harm the fluency in CNHG methods.

From the results, we observe that the NMT and NHG+NMT methods do not generate repeated or unnecessary words. Hence they are able to generate more fluent and headlines comparing to other methods. However, they still suffer from the missing key information problem.

V. Related work

A. Neural Headline Generation

End-to-end neural headline generation (NHG), has attracted increasing attention in recent several years. Researchers have been attempting to improve the performance of NHG from different aspects. For instance, source article representation methods [12], [14], [24] , encoder choices [9], [12], [25], decoder adoptions [9], [12], limited vocabulary problem solutions [10], [11], output length controlling problem [25] and training strategies [26].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2842432, IEEE/ACM Transactions on Audio, Speech, and Language Processing

	DUC2003			DUC2004		
	$R1$	$R2$	RL	$R1$	$R2$	RL
Baseline-TS	10.49	1.59	9.89	11.41	1.51	10.50
Baseline-ST	11.93	1.84	11.04	12.84	1.66	11.60
Baseline-PSEUDO	12.28	2.23	11.60	12.61	1.64	11.95
NHG teacher	11.15	2.11	10.54	11.15	1.88	10.51
NMT teacher	14.14	**3.04**	13.32	13.86	2.64	13.10
NMT+NHG teacher	**14.23**	3.00	**13.34**	**14.64**	**3.09**	**13.86**

TABLE IV: Experimental results on DUC datasets. DUC2003 is the development set, and the DUC2004 is the test set.

α	$R1$	$R2$	RL
0.1	10.79	2.02	10.24
0.3	12.37	2.33	11.63
0.5	13.04	2.93	12.25
0.7	**14.23**	**3.00**	**13.34**
0.9	13.95	2.72	12.99

TABLE V: Effect of hyper-parameter α on DUC2003 development dataset.

B. Cross Language Summarization

The cross-lingual headline generation is closely related to cross-lingual summarization. [1] proposes to score sentences from the original document, then translate selected sentences into target language to form a summary. [2] designs two graph based extractive summarization models which consider bilingual information. [3] presents a machine translation inspired scoring paradigm to construct the summary. [4] introduces bilingual concepts and facts utilizing translation and parsing information and generates cross lingual multi-document summarization.

C. Zero-shot Learning

Although there are plenty of data available, fine-grained annotated data are still missing in many cases. Humans have the ability to solve a task even when observing no examples of the task, i.e., zero-shot learning. There are various zero-shot learning related studies in natural language processing, for instance document retrieval [27], spoken language understanding [28] and neural machine translation [29], [30], [31], [32]. The closest related work to ours is that of the zero-shot NMT[29], [32].

VI. Conclusions

In this paper, we propose a direct end-to-end CNHG model which can address the training data discrepancy problem and error propagation problem in pipeline methods under a zero-shot scenario. Let the CNHG model be the student model, and we assume it would have close generation probability with the pre-trained NMT and NHG teacher models. Based on this assumption, we introduce three methods to guide the learning process of the CNHG student model. To evaluate the performance of the proposed approaches, we build corresponding development dataset and test dataset for the English-Chinese cross-lingual headline generation task by manually translating the standard DUC2003 task-1 and DUC2004 task-2 datasets. Experiments on the datasets show that our proposed NMT and NMT+NHG models can significantly outperform the baseline systems.

There are still many open problems to be explored as future work: (1) One problem in the neural sequence generation is the limited vocabulary size, which leads to the appearance of "UNK". There are many successful work [14], [10] that utilize the pointer network [33] in NHG models to address this problem. We will explore the proper way to integrate the technique in the cross-lingual scenario. (2) Besides article-headline pairs, there are also rich plain monolingual text data not considered in CNHG training. We will investigate the probability of integrating these plain texts to enhance CNHG for semi-supervised learning.

References

[1] X. Wan, H. Li, and J. Xiao, "Cross-language document summarization based on machine translation quality prediction," in *Proceedings of ACL*, 2010.

[2] X. Wan, "Using bilingual information for cross-language document summarization," in *Proceedings of ACL*, 2011.

[3] J.-g. Yao, X. Wan, and J. Xiao, "Phrase-based compressive cross-language summarization," in *Proceedings of EMNLP*, 2015.

[4] J. Zhang, Y. Zhou, and C. Zong, "Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[5] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *Proceedings of NIPS*, 2015.

[6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of ICLR*, 2015.

[7] H. Chen, M. Sun, C. Tu, Y. Lin, and Z. Liu, "Neural sentiment classification with user and product attention." in *Proceedings of EMNLP*, 2016.

[8] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances." in *Proceedings of ACL*, 2016.

[9] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proceedings of EMNLP*, 2015.

[10] J. Gu, Z. Lu, H. Li, and V. O. Li, "Incorporating copying mechanism in sequence-to-sequence learning," in *Proceedings of ACL*, 2016.

[11] C. Gulcehre, S. Ahn, R. Nallapati, B. Zhou, and Y. Bengio, "Pointing the unknown words," in *Proceedings of ACL*, 2016.

[12] S. Chopra, M. Auli, A. M. , and S. Harvard, "Abstractive sentence summarization with attentive recurrent neural networks," in *Proceedings of NAACL*, 2016.

[13] L. Yu, J. Buys, and P. Blunsom, "Online segment to segment neural transduction," in *Proceedings of EMNLP*, 2016.

[14] R. Nallapati, B. Zhou, and C. dos Santos, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *Proceedings of CoNLL*, 2016.

[15] C. Napoles, M. Gormley, and B. Van Durme, "Annotated gigaword," in *Proceedings of AKBC-WEKEX*, 2012.

[16] B. Hu, Q. Chen, and F. Zhu, "Lcsts: A large scale chinese short text summarization dataset," in *Proceedings of EMNLP*, 2015.

[17] S. Shen, Y. Cheng, Z. He, W. He, H. Wu, M. Sun, and Y. Liu, "Minimum risk training for neural machine translation," in *Proceedings of ACL*, 2016.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2842432, IEEE/ACM Transactions on Audio, Speech, and Language Processing

English article		The last time the Asia-Pacific region held its annual summit to promote free trade, Japan's prime minister assured everyone that his economy wouldn't be the next victim of Asia's financial crisis.
English reference headline		Asian-Pacific summit faces major economic and political challenges
Chinese reference headline		亚 太 首脑 会议 面临 重大 经济 和 政治 挑战
Baseline-ST	Step-1	News analysis : Asia-Pacific summit to promote free trade (by xiong UNK) UNK UNK contributed reporting .
	Step-2	新闻 分析 : UNK 首脑 峰会 促进 自由 贸易 活动 (UNK [10] *(News analysis: UNK summit promotes free trade activities (UNK [10])*
Baseline-TS	Step-1	去年 UNK 召开 年度 会议 促进 自由 贸易 , 日本 首相 保证 , 他 的 经济 不 会 成为 亚洲 金融 危机 的 牺牲品 。 *(Last year UNK hold annual meeting to promote free trade, Japanese prime minister assured that his economy would not be a victim of the Asian financial crisis.)*
	Step-2	日本 首相 : 不 会 成为 亚洲 金融 危机 的 牺牲品 *(Japan prime minister: will not be victim of Asian financial crisis)*
Baseline-PSEUDO		亚洲 金融 危机 使 亚洲 金融 危机 受到 影响 *(Asian financial crisis has affected the Asian financial crisis)*
NMT teacher		亚 太 经合 组织 首脑 会议 将 促进 自由 贸易 *(Asia-Pacific summit will promote free trade)*
NHG teahcer		日本 经济 增长 的 中国 经济 *(Japan's economy growth in China's economy)*
NMT+NHG teacher		亚 太 经合 组织 首脑 会议 开幕 *(Asia-Pacific summit opened)*

TABLE VI: Example headlines from each system and the corresponding English translation are given in parentheses, by italics. Considering the readability, we conduct a post-processing step to change the tokenized English contexts into normal form. The Baseline-ST and Baseline-TS systems are two steps pipeline systems; hence we list the corresponding results from each step as well. In Baseline-ST, Step-1 is the English headline of the original article generated by the English NHG model, and the Step-2 is the Chinese headline obtained by the English-Chinese NMT model. In Baseline-TS, Step-1 is the Chinese translation of original article generated by the English-Chinese NMT model, and the Step-2 is the Chinese headline obtained by the Chinese NHG model.

		Baseline-TS	Baseline-ST
avg.all	Step-1	16.71	4.34
	Step-2	33.27	26.80
max.per	Step-1	50.50	52.17
	Step-2	100.00	100.00

TABLE VII: "UNK" statistics in the pipeline methods. The avg.all and max.per refer to average "UNK" percentage in all results and max "UNK" percentage in one result respectively.

[18] T. Kočiský, G. Melis, E. Grefenstette, C. Dyer, W. Ling, P. Blunsom, and K. M. Hermann, "Semantic parsing with semi-supervised sequential autoencoders," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[19] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proceedings of EMNLP*, 2016.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of ACL*, 2002.

[21] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out: Proceedings of the ACL-04 workshop*, 2004.

[22] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of EMNLP*, 2014.

[23] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *Computer Science*, 2012.

[24] S. Takase, J. Suzuki, N. Okazaki, T. Hirao, and M. Nagata, "Neural headline generation on abstract meaning representation," in *Proceedings of ACL*, 2016.

[25] Y. Kikuchi, G. Neubig, R. Sasano, H. Takamura, and M. Okumura, "Controlling output length in neural encoder-decoders," in *Proceedings of EMNLP*, 2016.

[26] Ayana, S.-Q. Shen, Y.-K. Lin, C. hao Tu, Y. Zhao, Z.-Y. Liu, and M.-S. Sun, "Recent advances on neural headline generation," *Journal of Computer Science and Technology*, vol. 32, no. 4, pp. 768–784, 2017.

[27] R. Funaki and H. Nakayama, "Image-mediated learning for zero-shot cross-lingual document retrieval," in *Proceedings of EMNLP*, 2015.

[28] M. Yazdani and J. Henderson, "A model of zero-shot learning of spoken language understanding," in *EMNLP*, 2015.

[29] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado *et al.*, "Google's multilingual neural machine translation system: enabling zero-shot translation," *arXiv preprint arXiv:1611.04558*, 2016.

[30] H. Nakayama and N. Nishida, "Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot," *Machine Translation*, 2016.

[31] O. Firat, B. Sankaran, Y. Al-Onaizan, F. T. Y. Vural, and K. Cho, "Zero-resource translation with multi-lingual neural machine translation," *arXiv preprint arXiv:1606.04164*, 2016.

[32] Y. Chen, Y. Liu, Y. Cheng, and O. V. Li, "A teacher-student framework for zero-resource neural machine translation," in *Proceedings of ACL*, 2017.

[33] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," in *Proc. Advances in Neural Information Processing Systems*, 2015.

Ayana is a PhD student of the Department of Computer Science and Technology, Tsinghua University. She got her M.E. degree in 2009 from the College of Computer Science and Technology, Inner Mongolia University. Her research interest is document summarization.

Shi-qi Shen is a senior researcher of Wechat, Tencent. He got his PhD degree in computer science from the Department of Computer Science and Technology, Tsinghua University in 2017. His research interests are in the area of machine translation and deep learning for natural language processing. .

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2842432, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Yun Chen is a Ph.D. student in Department of Electrical and Electronic Engineering at The University of Hong Kong since 2014, under the supervision of Prof. Victor O.K. Li. She received her bachelor degree from Tsinghua University in 2013. She has research interests in machine learning approaches that are both linguistically motivated, and tailored to natural language processing, especially neural machine translation.

Cheng Yang is a 4-th year PhD student of the Department of Computer Science and Technology, Tsinghua University. He got his B.E. degree from Tsinghua University in 2014. His research interests include natural language processing and network representation learning.

Zhi-yuan Liu is an asssociate professorat the Department of Computer Science and Technology in Tsinghua University. He received his PhD degree from the Department of Computer Science and Technology, Tsinghua University in 2011. His research areas include natural language processing, knowledge graph and social computation.

Mao-song Sun is a professor at the Department of Computer Science and Technology in Tsinghua University. He received his PhD degree in Computational Linguistics from City University of Hong Kong in 2004. His research interests include natural language processing, Web intelligence and machine learning.