Attend, Translate and Summarize: An Efficient Method for Neural Cross-Lingual Summarization

Junnan Zhu[1,2], **Yu Zhou**[1,2,3], **Jiajun Zhang**[1,2], **Chengqing Zong**[1,2]
[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Beijing Fanyu Technology Co., Ltd
{junnan.zhu, yzhou, jjzhang, cqzong}@nlpr.ia.ac.cn

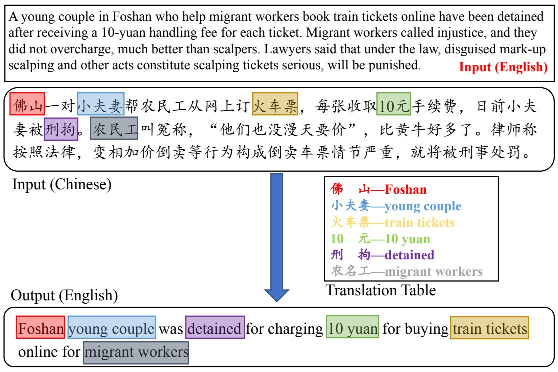
Abstract

Cross-lingual summarization aims at summarizing a document in one language (e.g., Chinese) into another language (e.g., English). In this paper, we propose a novel method inspired by the translation pattern in the process of obtaining a cross-lingual summary. We first attend to some words in the source text, then translate them into the target language, and summarize to get the final summary. Specifically, we first employ the encoder-decoder attention distribution to attend to the source words. Second, we present three strategies to acquire the translation probability, which helps obtain the translation candidates for each source word. Finally, each summary word is generated either from the neural distribution or from the translation candidates of source words. Experimental results on Chinese-to-English and English-to-Chinese summarization tasks have shown that our proposed method can significantly outperform the baselines, achieving comparable performance with the state-of-the-art.

1 Introduction

Cross-lingual summarization is to produce a summary in a target language (e.g., English) from a document in a different source language (e.g., Chinese). Cross-lingual summarization can help people efficiently understand the gist of an article written in an unfamiliar foreign language.

Traditional cross-lingual summarization methods are pipeline-based. These methods either adopt summarize-then-translate (Orasan and Chiorean, 2008; Wan et al., 2010) or employ translate-then-summarize (Leuski et al., 2003; Ouyang et al., 2019). The pipeline-based approach is intuitive and straightforward, but it suffers from error propagation. Due to the difficulty of acquiring cross-lingual summarization dataset, some previous researches focus on zero-shot methods (Ayana et al., 2018; Duan et al., 2019), i.e., using machine translation or monolingual summarization or both to teach the cross-lingual system.


A young couple in Foshan who help migrant workers book train tickets online have been detained after receiving a 10-yuan handling fee for each ticket. Migrant workers called injustice, and they did not overcharge, much better than scalpers. Lawyers said that under the law, disguised mark-up scalping and other acts constitute scalping tickets serious, will be punished.

佛山一对小夫妻帮农民工从网上订火车票，每张收取10元手续费，日前小夫妻被刑拘。农民工叫冤称，“他们也没漫天要价”，比黄牛好多了。律师称按照法律，变相加价倒卖等行为构成倒卖车票情节严重，就将被刑事处罚。

佛　山—Foshan
小夫妻—young couple
火车票—train tickets
10　元—10 yuan
刑　拘—detained
农名工—migrant workers

Foshan young couple was detained for charging 10 yuan for buying train tickets online for migrant workers

Figure 1: An example of the translation pattern in a sample extracted from Zh2EnSum (Zhu et al., 2019) which is a Chinese-to-English cross-lingual summarization dataset. It shows that some words in the summary are translated from the source words (in the same color). The translation table also gives the corresponding relation to these words. Best viewed in color.

Recently, Zhu et al. (2019) propose to use round-trip translation strategy to obtain large-scale cross-lingual summarization datasets. They incorporate machine translation and monolingual summarization into the training of cross-lingual summarization using multi-task learning to improve the summary quality with a quite promising performance. However, we find that there exist the following problems: (1) The multi-task methods adopt extra large-scale parallel data from other related tasks, such as monolingual summarization or machine translation. These methods are heavily dependent on data, making it difficult to migrate to languages with low resources. (2) The multi-task methods either simultaneously train cross-lingual summarization and monolingual summarization or alternately train cross-lingual summarization and machine translation, resulting in a quite time-consuming training process.

To alleviate the above problems, we observe some examples extracted from the cross-lingual summarization dataset. We find that there exists a translation pattern in the cross-lingual summaries, as shown in Figure 1. Inspired by the translation pattern, we can first attend to some specific segments in the input sequence, then translate them into the target language, and integrate this bilingual information into the final summary. Therefore, in this paper, we explore an efficient method consistent with the translation pattern.

To achieve that goal, we propose a novel method (Figure 2) that allows either generating words from the vocabulary or selecting words from the translation candidates of the words in the source article. Specifically, we first employ the encoder-decoder attention distribution to help determine which source word should be translated. Then we present three strategies, i.e., Naive, Equal, and Adapt, to obtain the translation probability from a probabilistic bilingual lexicon. The translation distribution can be acquired based on the encoder-decoder attention distribution and the translation probability. Next, we add an extra translation layer to calculate a translating probability. The final distribution is the weighted sum (weighed by the translating probability) of the translation distribution and the neural distribution.

Our main contributions are as follows:

- We introduce a novel and efficient method which integrates the operation of attending, translating, and summarizing.

- We present three effective strategies to acquire the translation probability. It has shown that all these strategies can significantly improve the performance over the baseline.

- Experimental results demonstrate that our method can achieve remarkable improvements over baselines and achieve comparable performance with the state-of-the-art on both English-to-Chinese and Chinese-to-English cross-lingual summarization tasks.

- Our method has two advantages over the state-of-the-art[1]: (1) We only adopt an additional probabilistic bilingual lexicon instead of a large-scale parallel machine translation dataset, which significantly relaxes the model's dependence on data. (2) Our model has a much smaller model size and a much faster training speed.

2 Background

In this paper, we implement our method based on Transformer (Vaswani et al., 2017) encoder-decoder framework, where the encoder first maps the input sequence $X = (x_1, x_2, \cdots, x_n)$ into a sequence of continuous representations $z = (z_1, z_2, \cdots, z_n)$ and the decoder generates an output sequence $Y = (y_1, y_2, \cdots, y_m)$ from the continuous representations. The encoder and decoder are trained jointly to maximize the conditional probability of target sequence given a source sequence:

$$L_\theta = \sum_{t=1}^{N} \mathrm{logP}(\mathrm{y_t}|\mathrm{y_{<t}}, \mathrm{X}; \theta) \tag{1}$$

Transformer is composed of stacked encoder and decoder layers. The encoder layer is a self-attention block followed by a position-wise feed-forward block. Compared with the encoder layer, the decoder layer has an extra encoder-decoder attention block. For self-attention and encoder-decoder attention, a multi-head attention block is used to obtain information from different representation subspaces at different positions. Each head corresponds to a scaled dot-product attention, which operates on query Q, key K, and value V:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

where d_k is the dimension of the key.

Finally, the output values are concatenated and projected by a feed-forward layer to get final values:

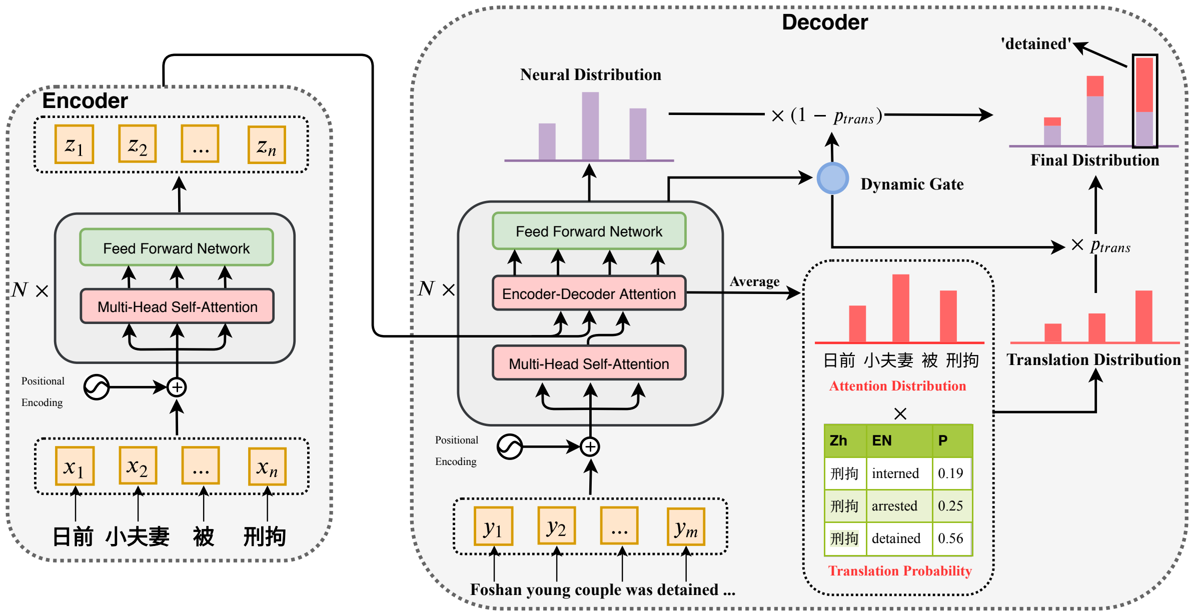
$$\begin{aligned} \mathrm{MultiHead}(Q, K, V) &= \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_\mathrm{h})W^O \\ \text{where head}_i &= \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{3}$$

where W^O, W_i^Q, W_i^K, and W_i^V are learnable matrices, and h is the number of heads.

3 Our Model

Inspired by the phenomenon that some words contained in a cross-lingual summary can be obtained by translating some source words (Figure 1), we introduce a novel cross-lingual summarization

[1] A multi-task method (Zhu et al., 2019) which trains cross-lingual summarization and machine translation using alternating training strategy.



Zh	EN	P
刑拘	interned	0.19
刑拘	arrested	0.25
刑拘	detained	0.56

Figure 2: Overview of our method. We first use encoder-decoder attention distribution to attend to some words and obtain the translation candidates from a probabilistic bilingual lexicon. Then a *translating probability* p_{trans} is calculated, which balances the probability of generating words from the neural distribution with that of selecting words from the translation candidates of the source text. The final distribution is obtained by the weighted sum (weighed by p_{trans}) of the neural distribution P_{N} and the translation distribution P_{T}. Best viewed in color.

method. It first attends to some source words, then obtains the translation candidates of them, and finally generates words from the translation candidates or the neural distribution. Our proposed method is a hybrid between Transformer and an additional translation layer, which is depicted in Figure 2 and described as follows.

Attend. Inspired by the pointer-generator network (See et al., 2017), we employ the encoder-decoder attention distribution α_t^h (the last layer) to help focus on some salient words in the source text. Since α_t^h is a multi-head attention, we take the mean value over the heads as follow:

$$\alpha_t = \frac{1}{h} \sum_h \alpha_t^h \tag{4}$$

Translate. With the attention distribution on the source words, we also need to know what should each source word be translated into. To achieve that, we obtain a probabilistic bilingual lexicon $P^L(w_1 \Rightarrow w_2)$ from existing machine translation corpora and then acquire the translation probability P_{T} based on $P^L(w_1 \Rightarrow w_2)$.

Acquisition of the probabilistic bilingual lexicon. There are many different ways to get the probabilistic bilingual lexicon, such as learning from bilingual corpora (Dyer et al., 2013; Chandar A P et al., 2014; Artetxe et al., 2016) and learning from monolingual corpora (Conneau et al., 2018; Zhang et al., 2017; Artetxe et al., 2018). To facilitate access to the high-quality probabilistic bilingual lexicon, we apply the method described in Dyer et al. (2013). Specifically, we first extract word alignments L using the fast-align tool (Dyer et al., 2013) on the bilingual parallel corpus[2] for machine translation in both source-to-target and target-to-source directions. To improve the quality of the word alignments, we only keep the alignments existing in both directions. Next, the lexicon translation probability $P^L(w_1 \Rightarrow w_2)$ is the average of source-to-target and target-to-source probabilities calculated through maximum likelihood estimation on word alignments L. We filter the lexicon pairs (w_1, w_2), where $P^L(w_1 \Rightarrow w_2) < 0.05$, and renormalize the lexicon translation probabilities to get the final probabilistic bilingual lexicon.

We propose the following three different strategies (Figure 3) to obtain the translation probability:

(1) **Naive.** We directly use the probability in the probabilistic bilingual lexicon as the translation probability. We limit the number of translation candidates of the word w_1 to at most *m*. Specifically,

[2] We employ the 2.08M sentence pairs from the LDC corpora which includes LDC2000T50, LDC2002L27, LDC2002T01, LDC2002E18, LDC2003E07, LDC2003E14, LDC2003T17, LDC2004T07.

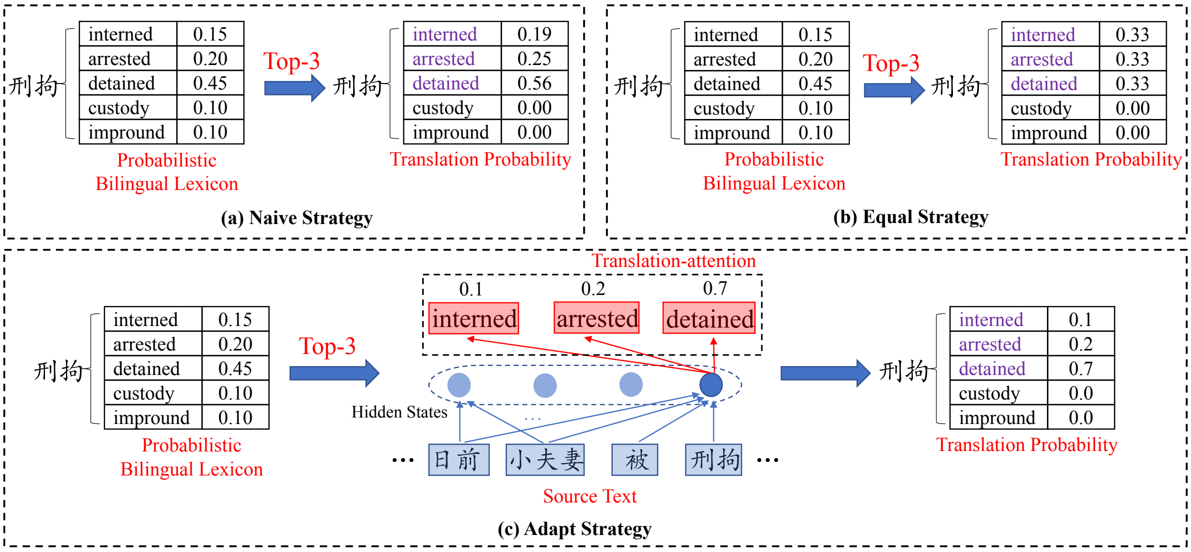


Figure 3: Overview of our three strategies to obtain the translation probability from the probabilistic bilingual lexicon. We take *m*=3 for example.

we sort the translation candidates of word w_1 in descending order according to the lexicon translation probability and then take the top-*m*. Finally, the lexicon translation probability will be normalized to get the translation probability:

$$P_{\mathrm{T}}(w_1 \Rightarrow w_2) = \frac{P^L(w_1 \Rightarrow w_2)}{\sum_{w_j} P^L(w_1 \Rightarrow w_j)} \tag{5}$$

(2) **Equal.** The Naive strategy will bring about a problem that the decoder tends to select the words with the high probability from the translation candidates of source word w_1, and those with low translation probability will hardly be selected. To alleviate this, we set the translation probability of w_1's translation candidates to be equal. Therefore, which translation candidate will eventually be selected depends on the probability of these translation candidates in the neural distribution. This strategy can be considered to achieve the goal of small vocabulary with the help of translation knowledge.

(3) **Adapt.** This strategy aims to select the correct translation candidates by source-side context. Specifically, we first limit the number of translation candidates to at most *m*, which is consistent with the two strategies above. Then we propose a *translation-attention* which is a multi-head attention block, where the hidden state of the source word w_1 is fed as the query and the target-side embedding of the corresponding translation candidates will be treated as the keys and values.

$$P_{\mathrm{T}}(w_1 \Rightarrow w_2) = \mathrm{Attention}(w_1, w_2^{\mathrm{tgt}}, w_2^{\mathrm{tgt}}) \tag{6}$$

where w_2^{tgt} is the target-side embedding of word w_2. We also take the mean value of the multi-head *translation-attention* as the final translation probability. Since the hidden state of the source word w_1 is obtained by the self-attention on the source-side, this context-aware strategy can help the model learn to choose the correct translation adaptively with the help of the source-side context.

Summarize. We use $\mathrm{H}_{\mathrm{dec}}$ to represent the decoder hidden state at timestep *t* and d_{model} to denote the dimension of the hidden states. We employ a translation layer to determine the *translating probability* $p_{\mathrm{trans}} \in [0, 1]$ via a dynamic gate:

$$p_{\mathrm{trans}} = \sigma(\mathbf{W}_2(\mathbf{W}_1 \mathrm{H}_{\mathrm{dec}} + b_1) + b_2) \tag{7}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_{\mathrm{model}} \times d_{\mathrm{model}}}$ and $\mathbf{W}_2 \in \mathbb{R}^{1 \times d_{\mathrm{model}}}$ are learnable matrices, $b_1 \in \mathbb{R}^{d_{\mathrm{model}}}$ and $b_2 \in \mathbb{R}^1$ are bias vectors, σ is the sigmoid function. Then p_{trans} is regarded as a soft switch to determine whether to generate a word w by sampling from the neural distribution or directly select a word from the translation candidates of the source words. Therefore, the final probability distribution can be calculated as follow:

$$\begin{aligned} P(w) = p_{\mathrm{trans}} & \sum_{i: w_i = w_{\mathrm{src}}} \alpha_{t,i} P_{\mathrm{T}}(w_{\mathrm{src}} \Rightarrow w) \\ & + (1 - p_{\mathrm{trans}}) P_{\mathrm{N}}(w) \end{aligned} \tag{8}$$

where $P_{\mathrm{T}}(w_{\mathrm{src}} \Rightarrow w)$ denotes the translation probability of word w_{src} to word w and P_{N} means the neural distribution.

4 Experiments

4.1 Datasets

In this study, we focus on Chinese-to-English and English-to-Chinese cross-lingual summarization. We test our proposed method on En2ZhSum and Zh2EnSum datasets[3] released by Zhu et al. (2019). En2ZhSum is an English-to-Chinese summarization dataset, which contains 370,687 English documents (755 tokens on average) paired with multi-sentence English (55 tokens on average) and Chinese summaries (96 Chinese characters on average). The dataset is split into 364,687 training pairs, 3,000 validation pairs, and 3,000 test pairs. Zh2EnSum is a Chinese-to-English summarization dataset, which contains 1,699,713 Chinese short texts (104 Chinese characters on average) paired with Chinese (18 Chinese characters on average) and English short summaries (14 tokens on average). The dataset is split into 1,693,713 training pairs, 3,000 validation pairs, and 3,000 test pairs. Both the English-to-Chinese and Chinese-to-English test sets are manually corrected.

4.2 Experimental Settings

We follow the setting of the vocabularies described in Zhu et al. (2019). In En2ZhSum, we surround each target sentence with tags "<t>" and "</t>". If there is no special explanation, the limit on the number of translation candidate m in our models is set to 10. All the parameters are initialized via Xavier initialization method (Glorot and Bengio, 2010). We train our models using configuration *transformer_base* (Vaswani et al., 2017), which contains a 6-layer encoder and a 6-layer decoder with 512-dimensional hidden representations. Each mini-batch contains a set of document-summary pairs with roughly 3,072 source and 3,072 target tokens. We apply Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.998$, and $\epsilon = 10^{-9}$. For evaluation, we use beam search with a beam size 4 and length penalty 0.6. All our methods are trained and tested on a single NVIDIA TITAN XP.

4.3 Comparative Methods

We compare our method with the following relevant methods (Zhu et al., 2019):

- **GETran**: It first translates the original article into the target language by Google Translator and then summarizes the translated text via LexRank (Erkan and Radev, 2004).
- **GLTran**: It first summarizes the original article via a Transformer-based monolingual summarization model and then translates the summary into the target language by Google Translator.
- **TNCLS**: It denotes the Transformer-based neural cross-lingual summarization system.

The above methods only employ the cross-lingual summarization dataset, and we also compare our method with the following two methods (Zhu et al., 2019) that use extra datasets in other tasks.

- **CLSMS**: It refers to the multi-task method, which simultaneously trains cross-lingual summarization and monolingual summarization.
- **CLSMT**: It is the multi-task method which adopts the alternating training strategy (Dong et al., 2015) to train cross-lingual summarization and machine translation jointly.

We denote our method as ATS:

- **ATS**: It refers to our method with three different strategies (Naive, Equal, and Adapt).

4.4 Experimental Results

We evaluate all models with the standard ROUGE metric (Lin, 2004), reporting the F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L. All ROUGE scores are reported by the 95% confidence interval measured by the official script[4]. Besides, we evaluate the equality of English summaries in Zh2EnSum with MoverScore (Zhao et al., 2019) which compares system output against references based on their semantics rather than surface forms. Zhao et al. (2019) have shown that MoverScore has a higher correlation with human judgment than ROUGE on evaluating English summaries.

Results on Zh2EnSum and En2ZhSum. Table 1 shows the results of different models on Zh2EnSum test set, while Table 2 gives the results on En2ZhSum test set. We use "subword-subword" and "word-character" segmentation granularities in Zh2EnSum and En2ZhSum, respectively.

We find that ATS can significantly outperform the baseline TNCLS on both Zh2EnSum and

[3] http://www.nlpr.ia.ac.cn/cip/dataset.htm

[4] The parameter for ROUGE script here is "-c 95 -r 1000 -n 2 -a".

	Model	RG-1	RG-2	RG-L	MVS
Baseline	GETran	24.34	9.14	20.13	0.64
	GLTran	35.45	16.86	31.28	16.90
	TNCLS	**38.85**	**21.93**	**35.05**	**19.43**
Baseline +Extra Data	CLSMS	**40.34**	**22.65**	**36.39**	**21.09**
	CLSMT	40.25	22.58	36.21	21.06
ATS	Naive	40.40	23.82†	36.63	21.86*
	Equal	40.10	23.36*	36.22	21.41
	Adapt	**40.68**	**24.12†**	**36.97**	**22.15**

Table 1: ROUGE F1 scores (%) and MoverScore scores (%) on Zh2EnSum test set. RG and MVS refer to ROUGE and MoverScore, respectively. We adopt "subword-subword" segmentation granularity here. The improvement of all ATS models over the baseline TNCLS is statistically significant ($p < 0.01$). * (†) indicates that the improvement over CLSMS is statistically significant where $p < 0.05$ (0.01).

	Model	RG-1	RG-2	RG-L
Baseline	GETran	28.19	11.40	25.77
	GLTran	32.17	13.85	29.43
	TNCLS	**36.82**	**18.72**	**33.20**
Baseline +Extra Data	CLSMS	38.25	20.20	34.76
	CLSMT	**40.23**	**22.32**	**36.59**
ATS	Naive	40.19	21.84	36.46
	Equal	39.98	21.63	36.29
	Adapt	**40.47**	**22.21**	**36.89**

Table 2: ROUGE F1 scores (%) on En2ZhSum test set. RG refers to ROUGE for short. We adopt "word-character" segmentation granularity here. The improvement of all ATS models over both TNCLS and CLSMS is statistically significant ($p < 0.01$).

En2ZhSum. Furthermore, ATS can significantly outperform CLSMS and CLSMT on Zh2EnSum while achieving comparable performance with CLSMS and CLSMT on En2ZhSum. However, both CLSMS and CLSMT employ large-scale parallel datasets of other tasks during the training process, limiting the generality of the models. In contrast, our method only requires an extra probabilistic bilingual lexicon, which significantly reduces the dependence on data. Among the variants of ATS, the ATS with Adapt strategy has the best performance. The reason is quite straightforward since the Adapt strategy helps to choose the right translation with the help of the source-side context. The Equal strategy performs worst, but its advantage over the Naive strategy is that it is not affected by the prior probability in probabilistic bilingual lexicon. In other words, the Equal strategy only makes use of the corresponding relationship between source language words and target language words, making it effective even if there is only a bilingual vocabulary dictionary. In summary, all three of our strategies can bring about significant improvement, which demonstrates that our method is robust to the acquisition method of translation candidates.

Src-Tgt	Model	Size (M)	Train (S)
Zh-En	TNCLS	134.92	21
	CLSMS	211.41	48
	CLSMT	208.84	63
	ATS-NE	136.55	27
	ATS-A	137.60	30
En-Zh	TNCLS	113.74	24
	CLSMS	190.23	65
	CLSMT	148.16	72
	ATS-NE	114.00	24
	ATS-A	115.05	25

Table 3: Model size (number of trainable parameters and M denotes mega) and training time of various models. Train (S) denotes how many seconds required for each model to train the 100-batch cross-lingual summarization task of the same batch size (3072). ATS-NE refers to our method with the Naive or Equal strategy. ATS-A is the one with Adapt strategy.

Model size and training time. The model size and training time of various models are given in Table 3. As it is shown, ATS is comparable with Transformer from both model size and training time. For model size, ATS is significantly less than the multi-task methods CLSMS and CLSMT. Especially on the Zh2En task, the size of multi-task models is nearly twice that of ATS. For training time, ATS is roughly half of the multi-task methods on both Zh2En and En2Zh tasks. Therefore, compared with the multi-task methods, ATS can significantly reduce the model size and improve the training efficiency.

In conclusion, our ATS models have achieved significant improvements over the baseline TNCLS on both Zh2EnSum and En2ZhSum, which can demonstrate the effectiveness of our approach. Furthermore, ATS achieves comparable performance with the state-of-the-art. Compared with the state-of-the-art, ATS can not only relax model's dependence on datasets but also reduce model size and improve training efficiency.

The impact of *m*. To study the impact of m (the limit on the number of translation candidates), we conduct an experiment on how the model performance changes when m varies from 10 to 5 or a

Model	m	Zh2En				En2Zh		
		RG-1	RG-2	RG-L	MVS	RG-1	RG-2	RG-L
ATS-A	1	40.93	24.17	37.11	22.31	39.85	21.45	36.12
	5	**41.05**	**24.31**	**37.28**	**22.77**	40.27	21.96	36.60
	10	40.68	24.12	36.97	22.15	**40.47**	**22.21**	**36.89**

Table 4: Results of ATS on Zh2EnSum and En2ZhSum under different hyperparameters, where m is the limit on the number of translation candidates. RG and MVS refer to ROUGE and MoverScore, respectively. We adopt "subword-subword" and "word-character" segmentation granularities in Zh2En and En2Zh models, respectively.

Model	Unit	RG-1	RG-2	RG-L	MVS
TNCLS	w-w	37.70	21.15	34.05	19.43
	sw-sw	38.85	21.93	35.05	19.07
ATS-A	w-w	39.65	23.79	36.05	22.06
	sw-sw	40.68	24.12	36.97	22.15

Table 5: Results of models on Zh2EnSum with different segmentation granularities. Unit represents the granularity combination of text units. *w* and *sw* denote "word" and "subword" (Sennrich et al., 2016), respectively. The improvement of all ATS models over TNCLS is statistically significant ($p < 0.01$).

Task	Unit	$p^{\text{macro}}_{\text{trans}}$	$p^{\text{micro}}_{\text{trans}}$	r^{macro}	r^{micro}
Zh2En	sw-sw	21.41	20.71	21.86	21.00
Zh2En	w-w	21.17	20.46	21.90	21.05
En2Zh	w-c	14.91	14.84	14.27	14.05

Table 6: Statistics on p_{trans} in ATS-A models. $p^{\text{macro}}_{\text{trans}}$ (%) and $p^{\text{micro}}_{\text{trans}}$ (%) respectively represent the macro-average and micro-average translating probability during decoding. r^{macro} (%) and r^{micro} (%) respectively represent the ratio of words where $p_{\text{trans}} > 0.5$ during decoding.

more aggressive value 1. The results are presented in Table 4. In Zh2En experiment, the ATS-A (m=5) performs best while ATS-A (m=1) performs comparably with ATS-A (m=10). In En2Zh experiment, the ATS-A (m=5) performs comparably with ATS-A (m=10) while the performance drops a bit when m=1. The above results illustrate that (1) A slightly larger m enables the model to learn when to search for translation candidates from the source words and which ones to choose, leading to improve the quality of the final summaries. (2) When m=1, the translation probability will contain some noise, but our method is still significantly better than the baseline, which further demonstrates the effectiveness and robustness of our method.

The impact of segmentation granularity. To study the effect of different segmentation granularities on the performance, we compare the performance of the model trained with "word-word" and "subword-subword" segmentation granularities on Zh2EnSum dataset. The results are given in Table 5. From ROUGE, our method brings about a similar degree of improvement over the baseline when using these two segmentation granularities. From MoverScore, it can be found that our method brings slightly greater improvement over the baseline when using the "subword-subword" segmentation granularity than using the "word-word" segmentation granularity. MoverScore metric compares system output against references based on their semantics, thus we believe ATS-A (sw-sw) can improve the semantic accuracy of the generated summary to a greater extent than ATS-A (w-w). Although the obtained probabilistic bilingual lexicon is of lower quality when using a smaller segmentation granularity, the source side covers more units, thus more translation candidates are exposed, making up for the noise in the probabilistic bilingual dictionary. In summary, our method can improve the performance under the above two different segmentation granularities, which illustrates that our method is robust to the segmentation granularity.

Translating Probability. Table 6 gives the statistics of translating probability in different ATS-A models. As it is shown, there is little difference in average translating probability under different segmentation granularities. However, the translation probabilities in tasks with different language directions are quite different. It is worth noting that the ration of words with translating probability greater than 0.5 does not mean that so many words are generated from translation operations, since the final distribution of summary words is jointly determined by translating probability, translation probability, encoder-decoder attention distribution, and neural distribution.

Human Evaluation. We conduct the human evaluation on 25 random samples extracted from each of Zh2EnSum and En2ZhSum, respectively.

Model	Zh2En			En2Zh		
	IF	CC	FL	IF	CC	FL
TNCLS	3.34	4.00	3.78	3.08	3.28	3.12
CLSMS	3.56	4.12	3.92	3.28	3.40	3.36
CLSMT	3.44	4.08	4.04	**3.38**	**3.56**	3.48
ATS-A	**3.64**	**4.16**	**4.18**	3.36	3.54	**3.52**

Table 7: Human evaluation results. IF, CC, and FL represent informativeness, conciseness, and fluency, respectively.

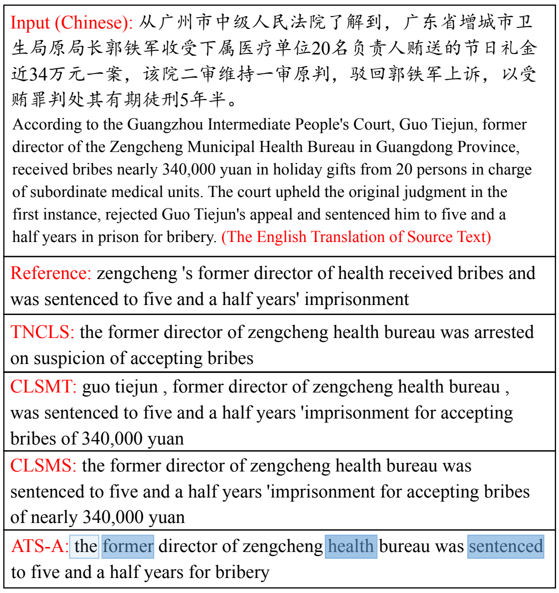
We compare the summaries generated by ATS (Adapt strategy) with other methods (including TNCLS, CLSMS, and CLSMT). Three graduate students are recruited to rate the generated summaries according to the references. Each summary is assessed from the three independent aspects: (1) How informative is the summary? (2) How concise is the summary? (3) How fluent and grammatical is the summary? Each aspect is scored from 1 (worst) to 5 (Best). The average results are given in Table 7.

We can find that the informativeness score, conciseness score, and fluency score of ATS-A are significantly better than those of the baseline TNCLS, which further demonstrates the effectiveness of our method. In Zh2En task, ATS-A outperforms CLSMT from all three aspects. The conciseness score of ATS-A is comparable with that of CLSMS, but ATS can generate more informative and fluent summaries. In En2Zh task, ATS-A outperforms CLSMS from all three aspects as well. The informativeness score and conciseness score of CLSMT are comparable with those of ATS-A, but ATS-A can generate more fluent summaries. To sum up, ATS-A can outperform CLSMS and CLSMT in Zh2En task, and ATS-A can outperform CLSMS while performing comparably with CLSMT in En2Zh task.

4.5 Case Study

We show a case study of a sample from Zh2EnSum test set. The summaries generated by each model are presented in Figure 4.

Although the summary generated by the TNCLS captures the critical character "*the former director of zengcheng health*" and the crime of "*received bribes*" committed by the character, it mistakenly expresses "*sentenced*" as "*arrested*" and fails to identify the prison term. Both CLSMT-generated summary and CLSMS-generated summary are fluent and grammatically correct. However, the amount in the source article is an approximate value "nearly 340,000 yuan", while CLSMT-generated summary directly expresses the exact value, which is inappropriate. The downside to both CLSMT-generated summary and CLSMS-generated summary is that they contain redundant information, since they are relatively lengthy. The summary generated by our ATS-A method matches the reference best and nearly captures all the key points in the source article. In conclusion, our method can generate summaries with more accurate semantics than baselines.



Input (Chinese): 从广州市中级人民法院了解到，广东省增城市卫生局原局长郭铁军收受下属医疗单位20名负责人贿送的节日礼金近34万元一案，该院二审维持一审原判，驳回郭铁军上诉，以受贿罪判处其有期徒刑5年半。
According to the Guangzhou Intermediate People's Court, Guo Tiejun, former director of the Zengcheng Municipal Health Bureau in Guangdong Province, received bribes nearly 340,000 yuan in holiday gifts from 20 persons in charge of subordinate medical units. The court upheld the original judgment in the first instance, rejected Guo Tiejun's appeal and sentenced him to five and a half years in prison for bribery. (The English Translation of Source Text)

Reference: zengcheng 's former director of health received bribes and was sentenced to five and a half years' imprisonment

TNCLS: the former director of zengcheng health bureau was arrested on suspicion of accepting bribes

CLSMT: guo tiejun , former director of zengcheng health bureau , was sentenced to five and a half years 'imprisonment for accepting bribes of 340,000 yuan

CLSMS: the former director of zengcheng health bureau was sentenced to five and a half years 'imprisonment for accepting bribes of nearly 340,000 yuan

ATS-A: the former director of zengcheng health bureau was sentenced to five and a half years for bribery

Figure 4: Examples of generated summaries. The English translation of source text is also given for better reading. The blue shading intensity denotes the value of the translating probability p_{trans}.

5 Related Work

Cross-Lingual Summarization. The traditional cross-lingual summarization approaches are based on the pipelined paradigm and can be categorized into *translate-then-summarize* (Leuski et al., 2003; Ouyang et al., 2019) and *summarize-then-translate* (Orasan and Chiorean, 2008; Wan et al., 2010). Leuski et al. (2003) translate the Hindi document into English and then generate the English headline. Ouyang et al. (2019) train a robust abstractive summarization system on noisy English documents and clean English reference summaries. Then the system can learn to produce fluent summaries from disfluent inputs, which enables the system to summarize translated documents.

Orasan and Chiorean (2008) summarize the Romanian news and then translate the summary into English. Wan et al. (2010) apply the *summarize-then-translate* scheme to English-to-Chinese cross-lingual summarization, which extracts English sentences considering both the informativeness and translation quality of sentences and automatically translates the English summary into Chinese. They also argue that *summarize-then-translate* is better, since it can alleviate both the computational expense of translating sentences and sentence extraction errors caused by incorrect translations.

There have been some researches focusing on improving cross-lingual summarization with bilingual information. Wan (2011) translates the English document into Chinese and extracts sentences based on the original English sentences and Chinese translation. Yao et al. (2015) propose a compressive method which calculates the sentence scores based on the aligned bilingual phrases obtained by machine translation service and performs compression via deleting redundant or poorly translated phrases. Zhang et al. (2016) introduce an abstractive method that constructs a pool of bilingual concepts represented by the bilingual elements of the source-side predicate-argument structures and the target-side counterparts.

Recently, end-to-end methods have been applied to cross-lingual summarization. Due to the lack of supervised training data, Ayana et al. (2018) and Duan et al. (2019) focus on zero-shot training methods that use machine translation or monolingual summarization or both to teach the cross-lingual system. Zhu et al. (2019) propose to acquire large-scale datasets via a round-trip translation strategy. They incorporate monolingual summarization or machine translation into cross-lingual summarization training using multi-task learning.

Neural Abstractive Summarization. Rush et al. (2015) present the first neural abstractive summarization model, an attentive convolutional encoder and a neural network language model decoder, which learns to generate news headlines from the lead sentences of news articles. Their approach has been further improved with recurrent decoders (Chopra et al., 2016), abstractive meaning representations (Takase et al., 2016), hierarchical networks (Nallapati et al., 2016), variational autoencoders (Miao and Blunsom, 2016), hybrid strategy (Zhu et al., 2017), selective mechanism (Zhou et al., 2017), and entailment knowledge. See et al. (2017) propose a pointer-generator network, which allows copying words from the source text with the copying mechanism (Gu et al., 2016). Li et al. (2018) incorporate entailment knowledge into summarization model to improve the correctness of the generated summaries. Li et al. (2020) apply guidance signals of keywords to both the encoder and decoder in the abstractive summarization model.

Inspired by the pointer-generator network and the translation pattern in obtaining cross-lingual summaries, we introduce a novel model in this paper, which integrates the operation of attending, translating, and summarizing.

6 Conclusion and Future Work

In this paper, we present a novel method consistent with the translation pattern in the process of obtaining a cross-lingual summary. This method first attends to the source words, then obtains the translation candidates, and incorporates them into the generation of the final summary. Experimental results have shown that our method can significantly outperform the baseline and achieve comparable performance with the state-of-the-art. Furthermore, our method has two advantages over the state-of-the-art: (1) Our model requires only an additional probabilistic bilingual lexicon rather than large-scale parallel datasets of other tasks, thus reducing the model's dependence on data and making it easier for the model to migrate to other domains or other language pairs. (2) Our model has a much smaller size and a much faster training efficiency.

In our future work, we consider incorporating our method into the multi-task method. Besides, we will also explore the influence of probabilistic bilingual lexicon obtained by learning only from monolingual data on our method.

7 Acknowledgments

The research work described in this paper has been supported by the National Key Research and Development Program of China under Grant No. 2017YFC0820700. We thank the three anonymous reviewers for their insightful comments and suggestions. We would like to thank Haitao Lin, Yu Lu, Yining Wang, Lu Xiang, and Yang Zhao for their invaluable contributions in shaping the early stage of this work. We thank Jinliang Lu, Xina Fu, and Meng Zhang for conducting human evaluation. We make our code publicly available here: `https://github.com/ZNLP/ATSum`.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2289–2294. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 789–798. Association for Computational Linguistics.

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

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 1853–1861. Curran Associates, Inc.

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

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

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1723–1732. Association for Computational Linguistics.

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 (ACL)*, pages 3162–3172. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 644–648. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 22:457–479.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10). Society for Artificial Intelligence and Statistics*, pages 249–256.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1631–1640. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1430–1441. Association for Computational Linguistics.

Haoran Li, Junnan Zhu, Jiajun Zhang, Chengqing Zong, and Xiaodong He. 2020. Keywords-guided abstractive sentence summarization. In *Proceedings of the Thirty-Forth AAAI Conference on Artificial Intelligence (AAAI)*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Yishu Miao and Phil Blunsom. 2016. Language as a latent variable: Discrete generative models for sentence compression. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 319–328. Association for Computational Linguistics.

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

Constantin Orasan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual Romanian-English multi-document summariser. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA).

Jessica Ouyang, Boya Song, and Kathy McKeown. 2019. A robust abstractive system for cross-lingual summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2025–2031. Association for Computational Linguistics.

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

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 (ACL)*, pages 1073–1083. Association for Computational Linguistics.

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

Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural headline generation on abstract meaning representation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1054–1059. Association for Computational Linguistics.

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

Xiaojun Wan. 2011. Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1546–1555. Association for Computational Linguistics.

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 (ACL)*, pages 917–926. Association for Computational Linguistics.

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 (EMNLP)*, pages 118–127. Association for Computational Linguistics.

Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2016. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 24(10):1842–1853.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1959–1970. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

Junnan Zhu, Long Zhou, Haoran Li, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2017. Augmenting neural sentence summarization through extractive summarization. In *Proceedings of the 6th Conference on Natural Language Processing and Chinese Computing (NLPCC)*, pages 16–28. Springer.

A Supplemental Material

Zh2EnSum	train	valid	test
#Documents	1,693,713	3,000	3,000
#AvgChars (S)	103.59	103.56	140.06
#AvgWords (R)	13.70	13.74	13.84
#AvgSentChars	52.73	52.41	53.38
#AvgSents	2.32	2.33	2.30

Table 8: Corpus statistics of Zh2EnSum. **#AvgChars (S)** is the average number of Chinese characters in the source document. **#AvgWords (R)** means the average number of English words in the reference. **#AvgSentChars** refers to the average number of characters in a sentence in the source document. **#AvgSents** denotes the average number of sentences in the source document.

En2ZhSum	train	valid	test
#Documents	364,687	3,000	3,000
#AvgWords (S)	755.09	759.55	744.84
#AvgChars (R)	55.21	55.28	54.76
#AvgSentWords	19.62	19.63	19.61
#AvgSents	40.62	41.08	40.25

Table 9: Corpus statistics of En2ZhSum. **#AvgWords (S)** is the average number of English words in the source document. **#AvgChars (R)** means the average number of Chinese characters in the reference. **#AvgSentWords** refers to the average number of Words in a sentence in the source document. **#AvgSents** denotes the average number of sentences in the source document.

Datasets. Table 8 and Table 9 show the statistics of Zh2EnSum dataset and En2ZhSum dataset, respectively.

Task	Unit	Source	Target
Zh2En	sw-sw	100,000	40,000
Zh2En	w-w	100,000	40,000
En2Zh	w-c	100,000	18,000

Table 10: The vocabulary size of models with different segmentation granularities.

Vocabulary Size. Table 10 gives the vocabulary size of models with different segmentation granularities. We employ the Urheen[5] tool to segment the Chinese text into words.

ROUGE Evaluation Details. In En2Zh task, we first delete the tags "<t>" and "</t>" generated by models. Then, we convert the text units in the reference and the system output into English IDs, such as "word1", "word2", etc. Each text unit has a unique English ID. Finally, we report the ROUGE scores based on these English IDs. The ROUGE scores reported in this paper can also be obtained by files2rouge[6] tool.

[5] http://www.nlpr.ia.ac.cn/cip/software.htm

[6] https://github.com/pltrdy/files2rouge

Input (English): ed miliband 's plan to cut university tuition fees is facing internal opposition with predictions it could cause a civil war within the party . ed miliband 's plan to cut university tuition fees was yesterday facing mounting opposition - with even a former labour no10 aide joining the attack . there were predictions last night that the party could descend into civil war over the controversial proposals after ex-tony blair aide huw evans was joined by the leader of britain 's nurses in challenging the plans . mr miliband has said his pledge to slash the fees from £ 9,000 a year to £ 6,000 is 'cast-iron ' , adding the plan will be a 'red line ' in any possible future coalition talks . but the plan – to be paid for by cutting middle-class pension pots – has been condemned as 'financial illiteracy ' by some critics , while university chiefs warn it could jeopardise the scrutiny of their long-term funding . the policy has also led to more than four years of rows within the shadow cabinet , with claims that ed balls repeatedly warned mr miliband that the £ 2.9billion fees cut was difficult to fund . mr evans , speaking in his capacity as director general of the association of british insurers (abi) , joined a growing number of pensions experts to challenge labour 's plans . mr evans , who worked for mr blair from 2005 to 2006 and is also a former adviser to ex-home secretary david blunkett , said : ' the pensions and long-term savings industry supports reform of tax relief but this is not the way to do it . ' we need a focus on reforming the pension tax relief system as a whole to make it fairer , better value and encourage saving from middle earners , rather than piecemeal cutting back the existing system to pay for other policy objectives . ' under the labour plan , tax relief for pensioners with incomes more than £ 150,000 would be cut from 45p to 20p while the tax-free lifetime allowance on a pension would drop from £ 1.25million to £ 1million . but the proposals could also hit people due to retire with a pension pot worth just £ 26,000 a year from an annuity while young people saving just £ 400 a month may also be affected . the plans have led to fears nurses , teachers and firefighters could also be hit . dr peter carter , of the royal college of nursing , said : ' helping students financially is important . however , this must not be at the expense of hard-working nurses . we will examine these proposals to ensure their pensions will not be affected . ' last night the comments were seized on by health secretary jeremy hunt . he wrote to his labour opposite number , andy burnham , saying : ' i wanted to ensure you are fully aware of the impact of this announcement on nhs staff . for example , if a nurse team leader earning around £ 35,500 , who is in a final salary , defined benefit pension scheme , achieves the promotion to matron they have been working 25 years to achieve , they will face a tax charge of £ 5,000 on their pension pot . this is what happens when policies are not properly thought through . ' but mr miliband has claimed the pensions raid would hit only the very wealthy . he said : ' the scourge of debt from tuition fees is not only holding back our young people , it is a burden on our country . ' mr miliband pictured at leeds college of music yesterday , where he announced his plan to slash tuition fees . .

Reference: 埃德米利班德在削减学费的计划上面临内部反对。但是费用削减是通过削减中产阶级养老金来支付的。现在预计该党可能会因为他的政策陷入内战。(ed miliband is facing internal opposition over plans to slash tuition fees . but the fee reductions are to paid for by cutting middle-class pension pots . it is now predicted the party could descend into civil war over his policy .)

TNCLS: 埃德米利班德面临内部学费削减的威胁。前助手休•埃文斯助手加入了反对党的行列。此举可能会导致米利班德削减成本。(ed miliband is under threat of internal tuition cuts. former aide hugh Evans' aide joined the opposition. the move could lead miliband to cut costs.)

CLSMT: 埃德米利班德削减大学学费的计划面临越来越多的反对。预计该党可能会因为这些提议陷入内战。但是削减中产阶级养老金的计划被指责“金融文化”(ed miliband's plan to cut college tuition is facing increasing opposition. it is expected that the party may fall into civil war because of these proposals. but plans to cut middle-class pensions have been accused of "financial culture.")

CLSMS: 埃德米利班德削减大学学费的计划面临内部反对。预测该党可能会陷入内战。工党领袖表示，这将是一个“红线”，任何未来的联合会谈。(ed miliband's plan to cut college tuition faces internal opposition. it is predicted that the party may fall into civil war. labor leaders said that this would be a "red line" for any future joint talks.)

ATS-A: 埃德米利班德削减学费的计划面临越来越大的反对。费用削减是通过削减中产阶级养老金来支付的。但是现在预计该党可能会陷入内战。(ed miliband's plan to cut tuition fees faces growing opposition. the cost reduction was paid by cutting middle-class pensions. but now it is expected that the party may fall into civil war.)

Figure 5: Examples of generated En2Zh summaries. The English translation of target-side text is also given for better reading. The blue shading intensity denotes the value of the translating probability p_{trans}.