# Cross-Lingual Summarization

（AntNLP）

Lei Zhengyi

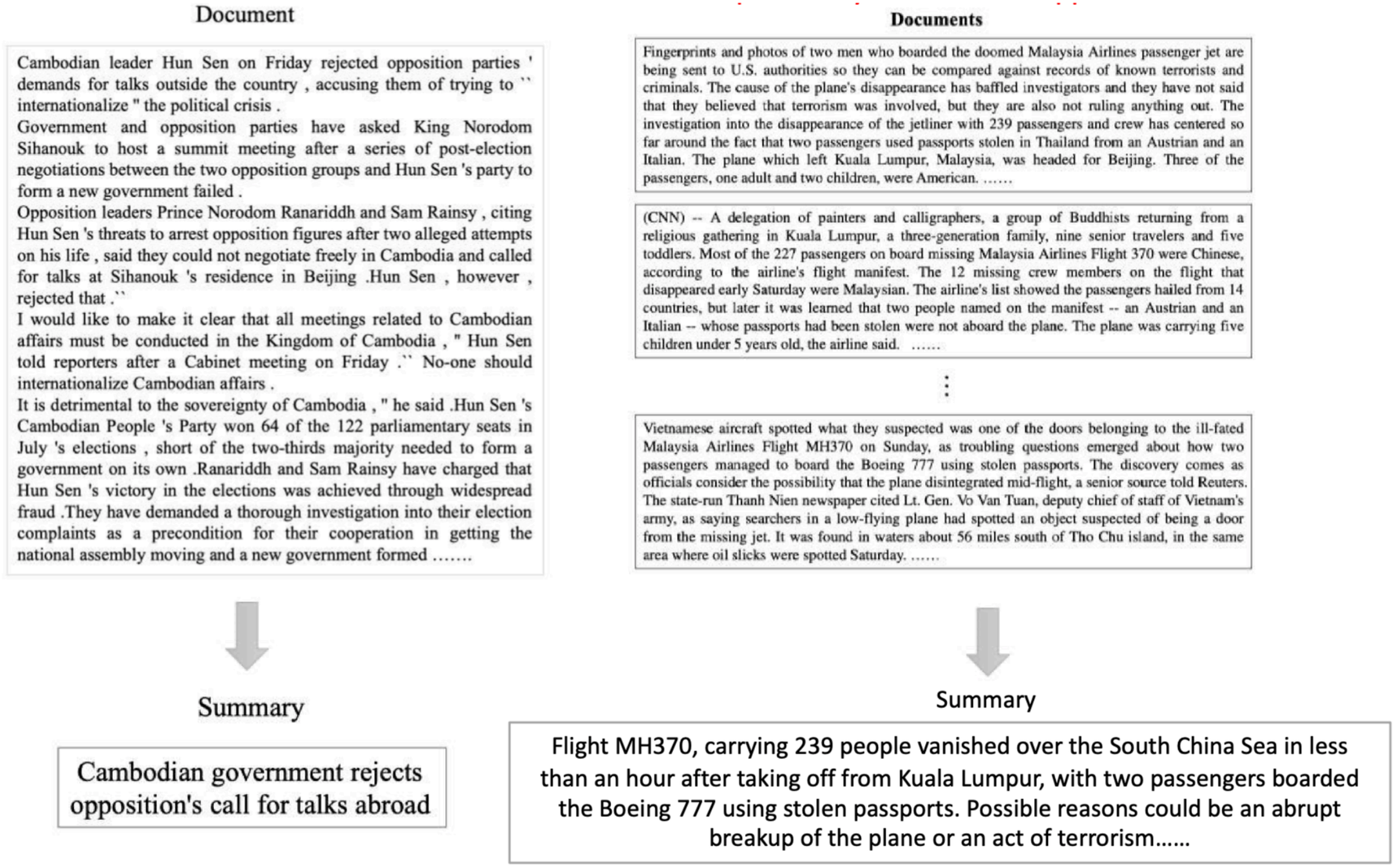
2020-11-27

# Overview

- Summarization
- Cross-Lingual Summarization（CLS）
- Datasets
- Evaluation
- Resent work：
  - 【EMNLP19】NCLS: Neural Cross-Lingual Summarization
  - 【ACL20】Jointly Learning to Align and Summarize for Neural Cross-Lingual Summarization
  - 【ACL20】Attend, Translate and Summarize: An Efficient Method for Neural Cross-Lingual Summarization

# Summarization

- Input：
  - Single-Document
  - Multi-Document
- Output：
  - Extractive Sum
  - Abstractive Sum



# Cross-Lingual Summarization（CLS）

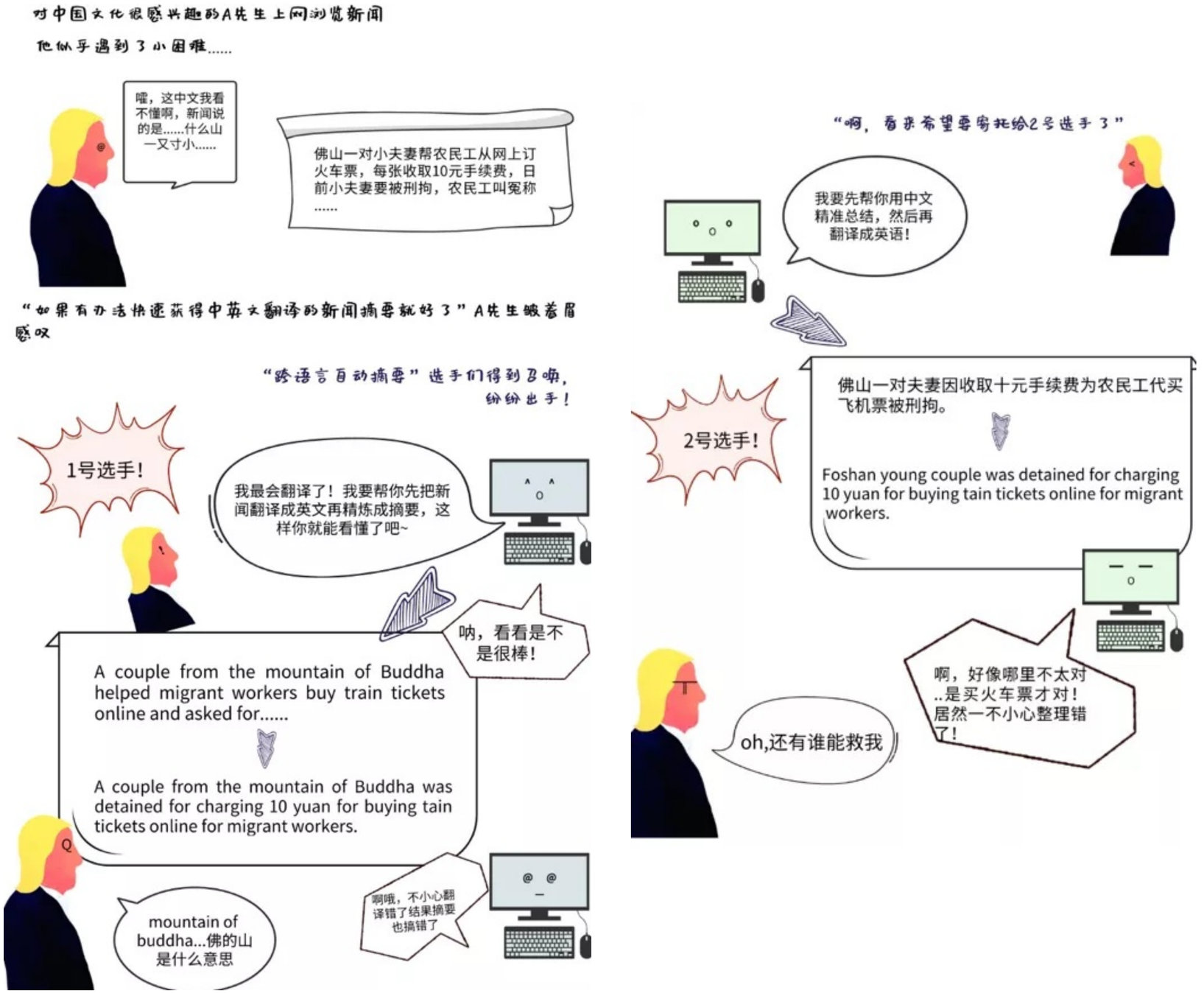
- An example of CLS

| |
|---|
| Input (Chinese): 在成本压力加大的情况下，流通企业不仅没有缩减IT投资反而继续增加。2011年，中国流通行业的IT投资规模由2010年的96.6亿元增加至2011年的109.2亿元。预计2012年流通行业的IT投资增速将达14.1%，规模超120亿元。<br>Under the circumstance of increasing cost pressure, circulation enterprises not only did not reduce IT investment but also continued to increase. In 2011, the scale of IT investment in China 's circulation industry increased from 9.66 billion yuan in 2010 to 10.92 billion yuan in 2011. It is estimated that the IT investment in the circulation industry will grow by 14.1% in 2012, with a scale exceeding 12 billion yuan. |
| Gold Summary: in 2012 , the scale of it investment in china 's circulation industry will exceed 12 billion yuan . |
| GETran: in the case of increased cost pressures, distribution companies have not only reduced it investment but continued to increase. |
| GLTran: it investment in china 's circulation industry will increase by 14.1 % in 2011 |
| TNCLS: it investment in circulation industry continues to increase |
| CLS+MS: it investment in china 's circulation industry will exceed 12 billion yuan in 2012 . |
| CLS+MT: china 's circulation industry is expected to increase it investment by 14.1 % in 2012 . |

Figure 4: Examples of generated summaries.

# Cross-Lingual Summarization （CLS）

Why do we need it?



# Cross-Lingual Summarization（CLS）

Motivation & Applications:

Why do we need it?

- 内容推荐（为用户推荐外语新闻）
- 跨语言自动摘要方法研究对于跨境电商（辅助用户进行决策）
- 舆情分析（帮助分析人员过滤冗余信息）

# Cross-Lingual Summarization （CLS）

**XiaoKe** :

A cross-language news generation system for generating scientific news stories about the latest discoveries from the world's leading science journals.
The system has been deployed in China Science Daily .

# Timeline

pipeline

- translate-summarizes（early translation ）
- summarizes-translate（late translation）

improve CLS with bilingual information

- Wan (2011)：proposes two graph-based summarization methods to leverage both the English-side and Chinese-side information.（ACL）
- Yao et al. (2015)：calculate the sentence scores based on the aligned bilingual phrases.（EMNLP）
- Zhang et al. (2016)：propose an abstractive CLS which constructs a pool of bilingual concepts represented by the bilingual elements.（TASLP）

NCLS

- End-to-end deep learning
- encoder-decoder（Transformer-based et al.）

Now

# Timeline

pipeline

- translate-summarizes（early translation ）
- summarizes-translate（late translation）

Relies heavily on hand-crafted features

Causes error propagation ～

improve CLS with bilingual information

- Wan (2011)：proposes two graph-based summarization methods to leverage both the English-side and Chinese-side information.（ACL）
- Yao et al. (2015)：calculate the sentence scores based on the aligned bilingual phrases.（EMNLP）
- Zhang et al. (2016)：propose an abstractive CLS which constructs a pool of bilingual concepts represented by the bilingual elements.（TASLP）

NCLS

- End-to-end deep learning
- encoder-decoder（Transformer-based et al.）

Now

# Timeline

pipeline

improve CLS with bilingual information

NCLS

- translate-summarizes（early translation ）
- summarizes-translate（late translation）

Relies heavily on hand-crafted features

Causes error propagation ～

- Wan (2011)：proposes two graph-based summarization methods to leverage both the English-side and Chinese-side information.（ACL）
- Yao et al. (2015)：calculate the sentence scores based on the aligned bilingual phrases.（EMNLP）
- Zhang et al. (2016)：propose an abstractive CLS which constructs a pool of bilingual concepts represented by the bilingual elements.（TASLP）

- End-to-end deep learning
- encoder-decoder（Transformer-based et al.）

lack of large-scale training data

# Datasets

- Monolingual
  - 1，CNN/Daily Mail：多句摘要数据集，分匿名和非匿名版本
  - 2，DUC：数据规模小，常用于测试集
  - 3，Gigaword：单句摘要数据集
  - 4，New York Times：96-07年的文章
  - 5，Newsroom：来自社交媒体：新闻、体育、娱乐、金融等
  - 6，LCSTS：中文短文本数据集
- Cross-lingual
  - 1，ZH2ENSUM and EN2ZHSUM（EMNLP19）
  - 2，WikiLingua（EMNLP20 Findings）
  - 3，MLSUM（EMNLP20）

# Evaluation

- Automative Evaluation
  - ROUGE-1、ROUGE-2、ROUGE-N、ROUGE-L、MoverScore…
- Human Evaluation
  - IF（informative）、CC（concise）、FL（fluent）

# NCLS: Neural Cross-Lingual Summarization (EMNLP-19)

**Junnan Zhu[1,2], Qian Wang[1,2], Yining Wang[1,2],**

**Yu Zhou[1,2*], Jiajun Zhang[1,2], Shaonan Wang[1,2], and Chengqing Zong[1,2,3]**

[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China

[2] University of Chinese Academy of Sciences, Beijing, China

[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China

{junnan.zhu, yzhou, jjzhang, cqzong}@nlpr.ia.ac.cn

# Introduction

- Pipeline
- Lack of data
- Propose a novel round-trip translation（RTT） strategy
- Train the CLS systems in an end-to-end manner, we present NCLS
- Multi-task（CLS+MS/MT）

# Dataset Construction

- Monolingual Datasets used：
  - **ENSUM**：the union（并集） set of **CNN/Daily Mail** and **MSMO**
  - **LCSTS**：constructed from the Chinese microblogging website *Sina Weibo*

# Dataset Construction

- **RTT**：Round-trip translation strategy.

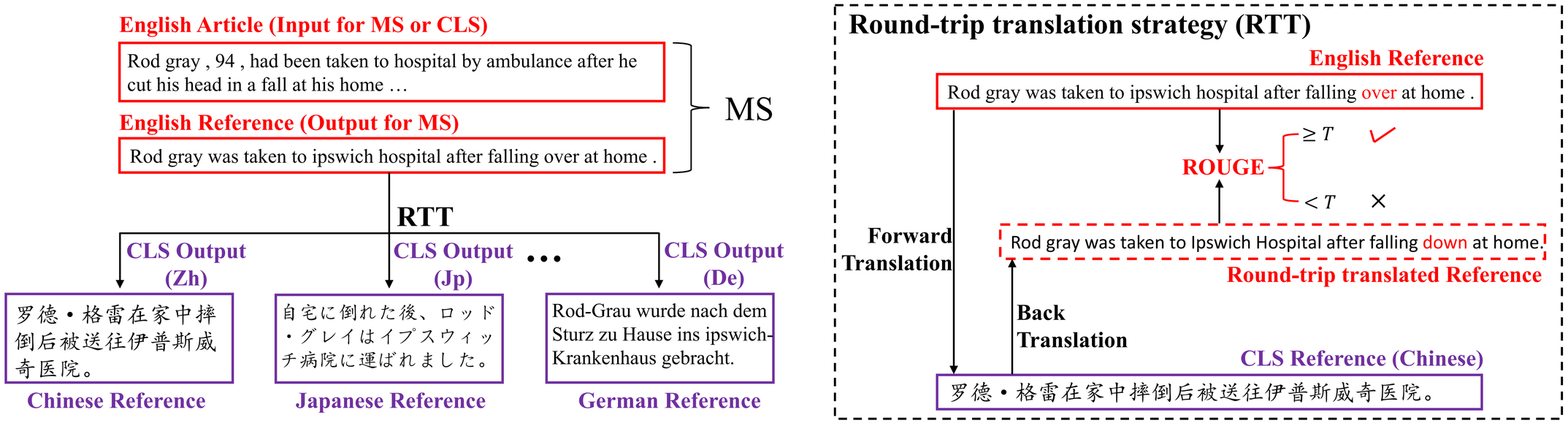



Figure 1: Overview of CLS corpora construction. Our method can be extended to many other language pairs and we focus on En2Zh and Zh2En in this paper. During RTT, we filter the sample in which ROUGE F1 score between the original reference and the round-trip translated reference is below a preset threshold $T$.

# Dataset Construction

- Corpus Statistics.

| En2ZhSum | train | valid | test | Zh2EnSum | train | valid | test |
|---|---|---|---|---|---|---|---|
| #Documents | 364,687 | 3,000 | 3,000 | #Documents | 1,693,713 | 3,000 | 3,000 |
| #AvgWords (S) | 755.09 | 759.55 | 744.84 | #AvgChars (S) | 103.59 | 103.56 | 140.06 |
| #AvgEnWords (R) | 55.21 | 55.28 | 54.76 | #AvgZhChars (R) | 17.94 | 18.00 | 18.08 |
| #AvgZhChars (R) | 95.96 | 96.05 | 95.33 | #AvgEnWords (R) | 13.70 | 13.74 | 13.84 |
| #AvgSentsWords | 19.62 | 19.63 | 19.61 | #AvgSentsChars | 52.73 | 52.41 | 53.38 |
| #AvgSents | 40.62 | 41.08 | 40.25 | #AvgSents | 2.32 | 2.33 | 2.30 |

Table 1: Corpus statistics. **#AvgWords (S)** is the average number of English words in the source document. Each reference has a bilingual version since each reference in CLS corpus is translated from the corresponding reference in the MS corpus. **#AvgEnWords (R)** means the average number of words in English reference and **#AvgZhChars (R)** denotes the average number of characters in Chinese reference. **#AvgSentsWords** (**#AvgSentsChars**) indicates the average number of words (characters) in a sentence in the source document. **#AvgSents** refers to the average number of sentences in the source document.

# Approach

- Baseline Pipeline Methods(trans-Sum / Sum-trans)
- Neural Cross-Lingual Summarization based on transformer encoder-decoder network

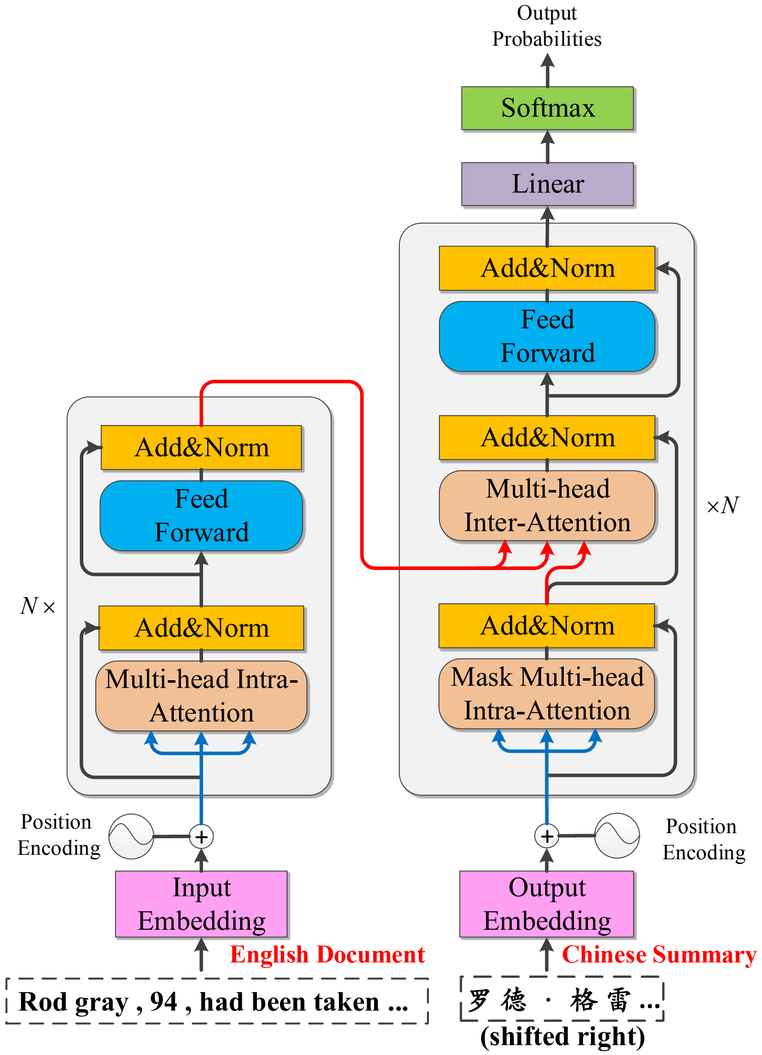



Figure 2: Transformer-based NCLS models (TNCLS).

# Approach

- Improving NCLS with MS and MT (
  - 1,**CLS+MS**:

$$L_\theta = \sum_{t=1}^{N^{(1)}} \log P(y_t^{(1)} | y_{<t}^{(1)}, x; \theta) + \sum_{t=1}^{N^{(2)}} \log P(y_t^{(2)} | y_{<t}^{(2)}, x; \theta) \tag{4}$$

  - 2,**CLS+MT**: alternating training strategy (Dong et al., 2015), which optimizes each task for a fixed number of mini-batches before switching to the next task. )

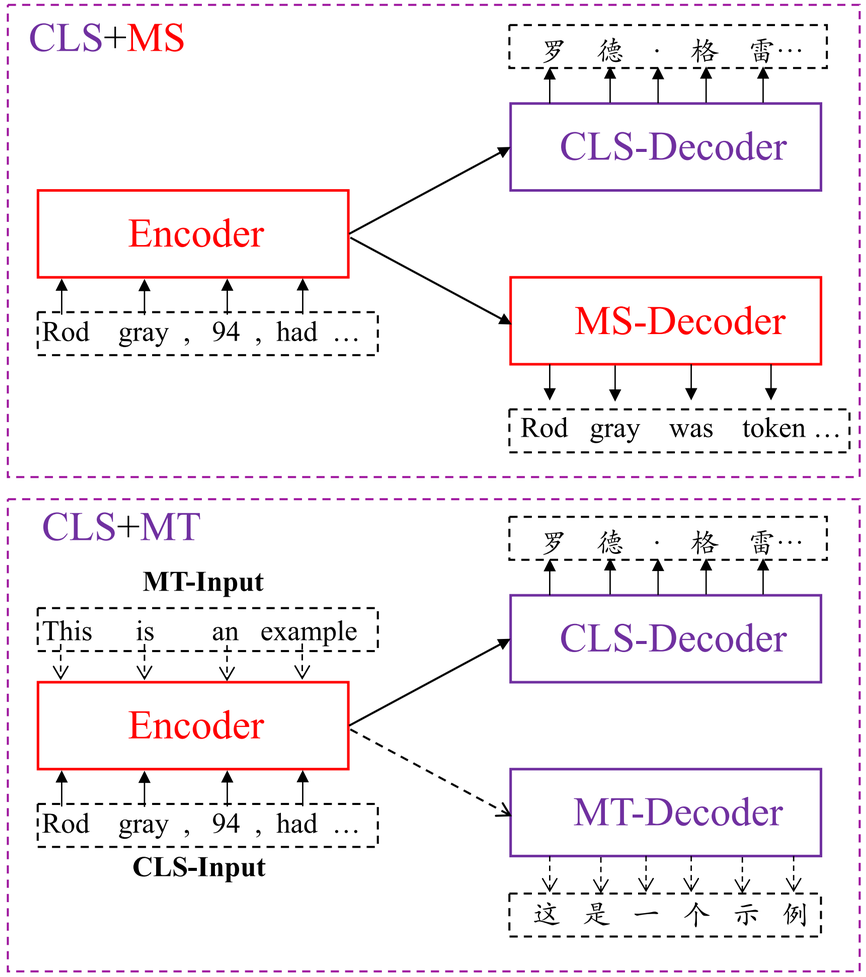



Figure 3: Overview of multi-task NCLS. The lower half is CLS+MT using alternating training strategy. Different colors represent different languages.

# Experiments

- The performance of our transformer-based MS models

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| Gu et al. (2016) | 35.00 | 22.30 | 32.00 |
| Li et al. (2017) | 36.99 | 24.15 | 34.21 |
| Transformer | 39.71 | 27.45 | 37.13 |

Table 2: Performance of our implemented transformer-based monolingual summarization model on LCSTS.

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| See et al. (2017) | 39.53 | 17.28 | 36.38 |
| Transformer | 39.24 | 16.67 | 36.42 |

Table 3: Performance of our implemented transformer-based MS model on CNN/Dailymail.

# Experiments

- Comparison between NCLS with baselines.

| Model | Unit | En2ZhSum | En2ZhSum* | Zh2EnSum | Zh2EnSum* |
|---|---|---|---|---|---|
| | | RG1-RG2-RGL(↑) | RG1-RG2-RGL(↑) | RG1-RG2-RGL(↑) | RG1-RG2-RGL(↑) |
| TETran | – | **26.12-10.59-23.21** | **26.15-10.60-23.24** | **22.81- 7.17-18.55** | **23.09- 7.33-18.74** |
| GETran | – | **28.17-11.38-25.75** | **28.19-11.40-25.77** | **24.03- 8.91-19.92** | **24.34- 9.14-20.13** |
| TLTran | c-c | – | – | 32.85-15.34-29.21 | 33.01-15.43-29.32 |
| | w-w | **30.20-12.20-27.02** | **30.22-12.20-27.04** | 31.11-13.23-27.55 | 31.38-13.42-27.69 |
| | sw-sw | – | – | **33.64-15.58-29.74** | **33.92-15.81-29.86** |
| GLTran | c-c | – | – | 34.44-15.71-30.13 | 34.58-16.01-30.25 |
| | w-w | **32.15-13.84-29.42** | **32.17-13.85-29.43** | 32.42-15.19-28.75 | 32.52-15.39-28.88 |
| | sw-sw | – | – | **35.28-16.59-31.08** | **35.45-16.86-31.28** |
| TNCLS | c-w | – | – | 36.36-19.74-32.66 | 35.82-19.04-32.06 |
| | w-c | **36.83-18.76-33.22** | **36.82-18.72-33.20** | – | – |
| | w-w | 33.09-14.85-29.82 | 33.10-14.83-29.82 | 38.54-22.34-35.05 | 37.70-21.15-34.05 |
| | sw-sw | – | – | **39.80-23.15-36.11** | **38.85-21.93-35.05** |

Table 4: ROUGE F1 scores (%) on En2ZhSum and Zh2EnSum test sets. En2ZhSum* and Zh2EnSum* are the corresponding human-corrected test sets. *Unit* denotes the granularity combination of text units, where *c* means character, *w* means word, and *sw* means subword. RG refers to ROUGE for short. ↑ indicates that the larger values, the better the results are. Our NCLS models perform significantly better than baseline models by the 95% confidence interval measured by the official ROUGE script[8].

# Experiments

- Why Back Translation?
- Results of Multi-task NCLS.

| DataVersion | BT? | En2ZhSum | En2ZhSum* | Zh2EnSum | Zh2EnSum* |
|---|---|---|---|---|---|
| | | RG1-RG2-RGL(↑) | RG1-RG2-RGL(↑) | RG1-RG2-RGL(↑) | RG1-RG2-RGL(↑) |
| Filter | YES | 36.83-18.76-33.22 | 36.82-18.72-33.20 | 39.80-23.15-36.11 | 38.85-21.93-35.05 |
| Pseudo-Filter | NO | 36.04-17.80-32.49 | 36.03-17.78-32.48 | 35.58-17.93-31.71 | 35.00-17.37-31.10 |
| Non-Filter | NO | 37.62-19.88-33.99 | 37.62-19.85-33.99 | 36.51-19.23-32.77 | 36.03-18.63-32.19 |

Table 5: Experimental results on different versions of datasets. *Filter* refers to the version of dataset for which we employ RTT strategy to filter. *Non-Filter* denotes the version of the dataset obtained by simply forward translation without filtering process including back translation. *Pseudo-Filter* is the dataset randomly sampled from *Non-Filter* version and is of the same size as *Filter* version. BT refers to back translation in RTT. For En2Zh task, we train the TNCLS (*w-c*). For Zh2En task, we train the TNCLS (*sw-sw*).

| Model | En2ZhSum | En2ZhSum* | Zh2EnSum | Zh2EnSum* |
|---|---|---|---|---|
| | RG1-RG2-RGL(↑) | RG1-RG2-RGL(↑) | RG1-RG2-RGL(↑) | RG1-RG2-RGL(↑) |
| TNCLS | **36.83-18.76-33.22** | **36.82-18.72-33.20** | **39.80-23.15-36.11** | **38.85-21.93-35.05** |
| CLS+MS | 38.23-20.21-34.76 | 38.25-20.20-34.76 | 41.08-23.67-37.19 | **40.34-22.65-36.39** |
| CLS+MT | **40.24-22.36-36.61** | **40.23-22.32-36.59** | **41.09-23.70-37.17** | 40.25-22.58-36.21 |

Table 6: Results of multi-task NCLS. The granularity combination of input and output in En2Zh task is "word to character" (*w-c*), and that in Zh2En task is "subword to subword" (*sw-sw*).

# Experiments

- Human Evaluation
- Case Study

| Model | En2Zh | | | Zh2En | | |
|---|---|---|---|---|---|---|
| | IF | CC | FL | IF | CC | FL |
| GLTran | 3.06 | 3.37 | 3.13 | 3.53 | 4.21 | 4.25 |
| TNCLS | 3.25 | 3.33 | 3.17 | 3.67 | 4.25 | 4.24 |
| CLS+MS | 3.53 | 3.58 | 3.53 | 3.72 | 4.31 | 4.28 |
| CLS+MT | **3.58** | **3.76** | **3.63** | **3.78** | **4.43** | **4.35** |

Table 7: Human evaluation results. IF, CC and FL denote informative, concise, and fluent respectively.

| |
|---|
| Input (Chinese): 在成本压力加大的情况下，流通企业不仅没有缩减IT投资反而继续增加。2011年，中国流通行业的IT投资规模由2010年的96. 6亿元增加至2011年的109. 2亿元。预计2012年流通行业的IT投资增速将达14. 1%，规模超120亿元。<br>Under the circumstance of increasing cost pressure, circulation enterprises not only did not reduce IT investment but also continued to increase. In 2011, the scale of IT investment in China 's circulation industry increased from 9.66 billion yuan in 2010 to 10.92 billion yuan in 2011. It is estimated that the IT investment in the circulation industry will grow by 14.1% in 2012, with a scale exceeding 12 billion yuan. |
| Gold Summary: in 2012 , the scale of it investment in china 's circulation industry will exceed 12 billion yuan . |
| GETran: in the case of increased cost pressures, distribution companies have not only reduced it investment but continued to increase. |
| GLTran: it investment in china 's circulation industry will increase by 14.1 % in 2011 |
| TNCLS: it investment in circulation industry continues to increase |
| CLS+MS: it investment in china 's circulation industry will exceed 12 billion yuan in 2012 . |
| CLS+MT: china 's circulation industry is expected to increase it investment by 14.1 % in 2012 . |

Figure 4: Examples of generated summaries.

# Conclusion and Future Work

1. present neural cross-lingual summarization for the first time;

2. propose to acquire large-scale supervised data from existing monolingual summarization datasets via round-trip translation strategy;

3. apply end-to-end methods on our constructed datasets;

4. utilizing MT and MS to further improve NCLS.

- In our future work, we will adopt our RTT strategy to obtain CLS datasets of other language pairs, such as English-to-Japanese, English- to-German, Chinese-to-Japanese, and Chinese-to-German, etc.

# Jointly Learning to Align and Summarize for Neural Cross-Lingual Summarization（ACL-20）

**Yue Cao, Hui Liu, Xiaojun Wan**

Wangxuan Institute of Computer Technology, Peking University

Center for Data Science, Peking University

The MOE Key Laboratory of Computational Linguistics, Peking University

{yuecao,xinkeliuhui,wanxiaojun}@pku.edu.cn

# Introduction

- **Early researches**: summarization-translation or translation-summarization pipeline paradigm and adopt different strategies to incorporate bilingual features into the pipeline model.

- **Recently**: teacher-student framework、multi-task learning framework、Learning cross-lingual representations…

- Problem：lack of parallel corpora；learning to understand different languages and learning how to summarize at the same time is a big challenge.

- We propose a multi-task framework that jointly learns to **summarize** and **align** context-level representations:
  - design relevant loss functions to train this framework
  - propose several methods to enhance the isomorphism and cross-lingual transfer between languages.

# Overview

- Our model consists of two encoders, two decoders, two linear mappers, and two discriminators.

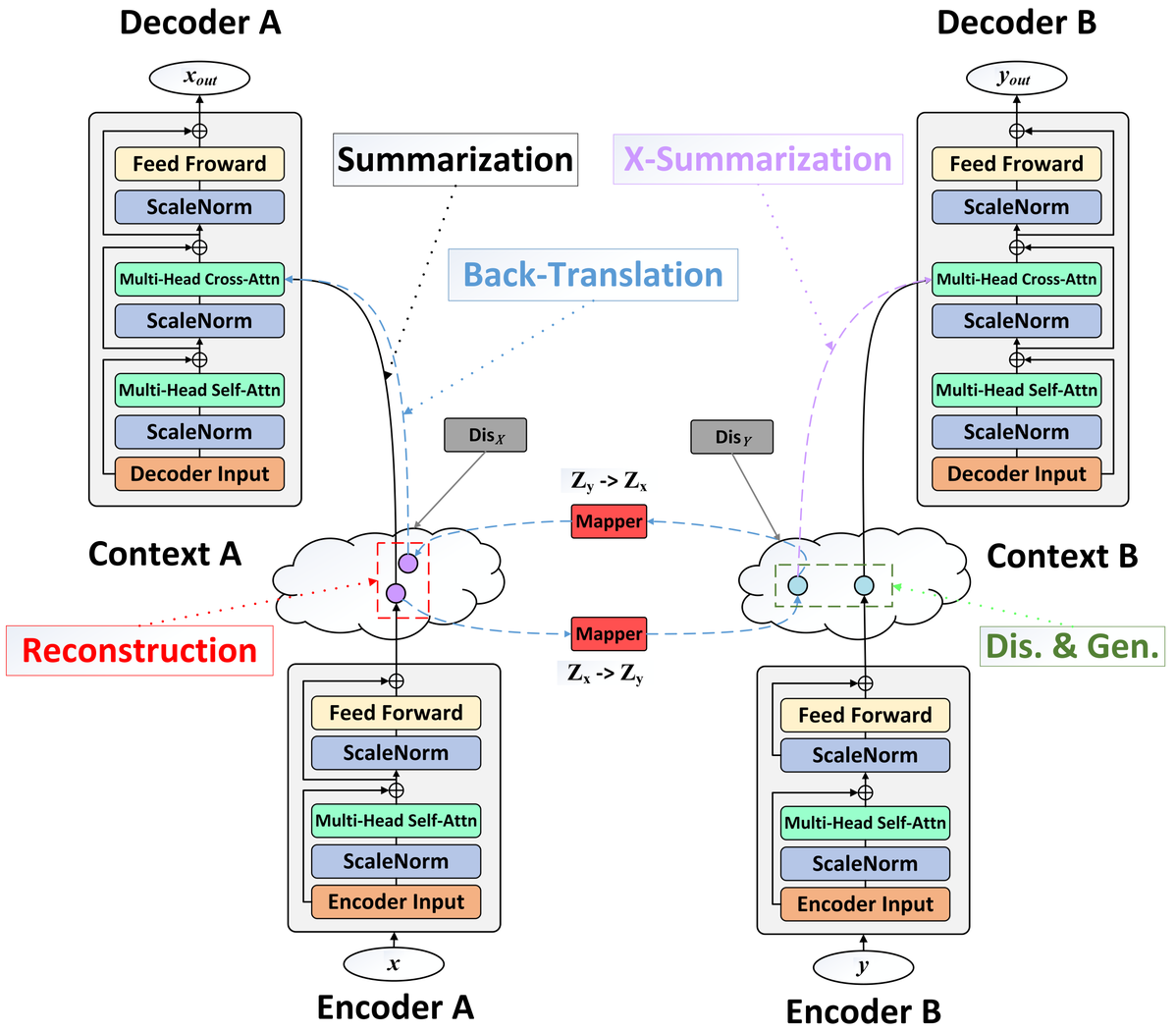



Figure 1: The overall framework of our proposed model.

# Model Adjustment for Cross-Lingual Transfer

- **Normalizing the Representations**

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

$$\Downarrow$$

$$o_{\ell+1} = o_\ell + F_\ell(\text{ScaleNorm}\left(o_\ell\right)) \longrightarrow \text{ScaleNorm}(x; g) = g \cdot x/\|x\|$$

Advantage：after being normalized, the dot-product of two vectors u⊤v is equivalent to their cosine distance $\frac{u^\top v}{\|u\|\|v\|}$

- **Enhancing the Isomorphism**
  - First, we combine the English and Chinese summarization corpora and build a unified vocabulary.
  - Second, we share encoders and decoders in our model.
  - Third, we train several mono-lingual summarization steps before cross-lingual training.

# Unsupervised Training Objective

- 1，Summarization Loss

$$\boldsymbol{z_x} = \phi_{E_\mathcal{X}}(\boldsymbol{x}), \quad \tilde{\boldsymbol{x}} = \phi_{D_\mathcal{X}}(\boldsymbol{z_x})$$

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

- 2，Generative and Discriminative Loss

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

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

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

- 3，Cycle Reconstruction Loss

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

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

- 4，Back-Translation Loss

$$\hat{\boldsymbol{x}} = \phi_{D_\mathcal{X}}(\hat{\boldsymbol{z}}_{\boldsymbol{x}})$$

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

- 3，Total Loss

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

$$\mathcal{L} = \mathcal{L}_\mathcal{X} + \mathcal{L}_\mathcal{Y} \qquad \mathcal{L}_{dis} = \mathcal{L}_{dis_\mathcal{X}} + \mathcal{L}_{dis_\mathcal{Y}}$$

# Supervised Training Objective

- 1，X-Summarization Loss

$$\boldsymbol{z_x}=\phi_{E_{\mathcal{X}}}(\boldsymbol{x}),\ \boldsymbol{z_{x\to y}}=M_{\mathcal{X}}(\boldsymbol{z_x}),\ \tilde{\boldsymbol{y}}=\phi_{D_{\mathcal{Y}}}(\boldsymbol{z_{x\to y}})$$

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

- 2，Reconstruction Loss

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

$$\mathcal{L}_{\text{rec}_{\mathcal{X}}}\left(\tilde{z}_x,\tilde{z}_{y\to x}\right)=\left\|\tilde{z}_x-\tilde{z}_{y\to x}\right\| \tag{14}$$

- 3，Total Loss

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

# Overview

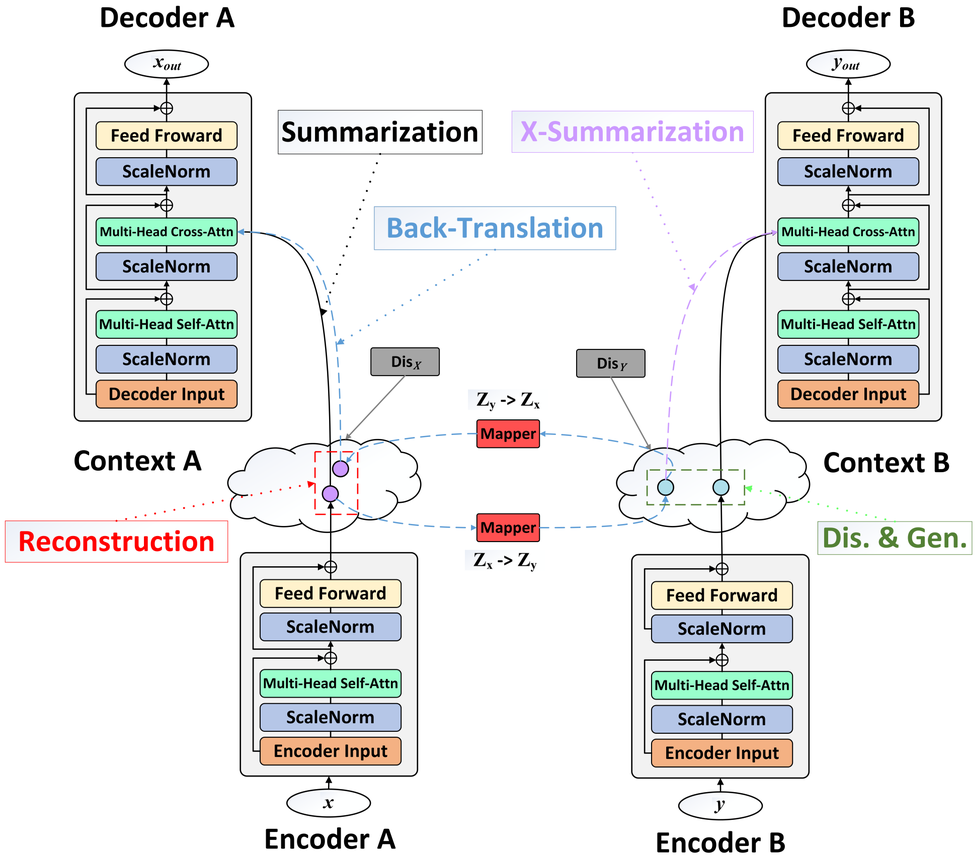



Figure 1: The overall framework of our proposed model.

**Algorithm 1** Cross-lingual summarization

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

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

# Experiments

- Unsupervised Cross-Lingual Summarization（Table 2）
- Supervised Cross-Lingual Summarization（Table 1）

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

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

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

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

# Experiments

- Human Evaluation

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

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

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

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

# Experiments

- Ablation Tests

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

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

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

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

# Conclusions

- we propose a framework that jointly learns to align and summarize for neural cross-lingual summarization
- We design training objectives for supervised and unsupervised cross-lingual summarizations, respectively.
- We also propose methods to enhance the isomorphism and cross-lingual transfer between languages.

# Attend, Translate and Summarize: An Efficient Method for Neural Cross-Lingual Summarization (ACL-20)

**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

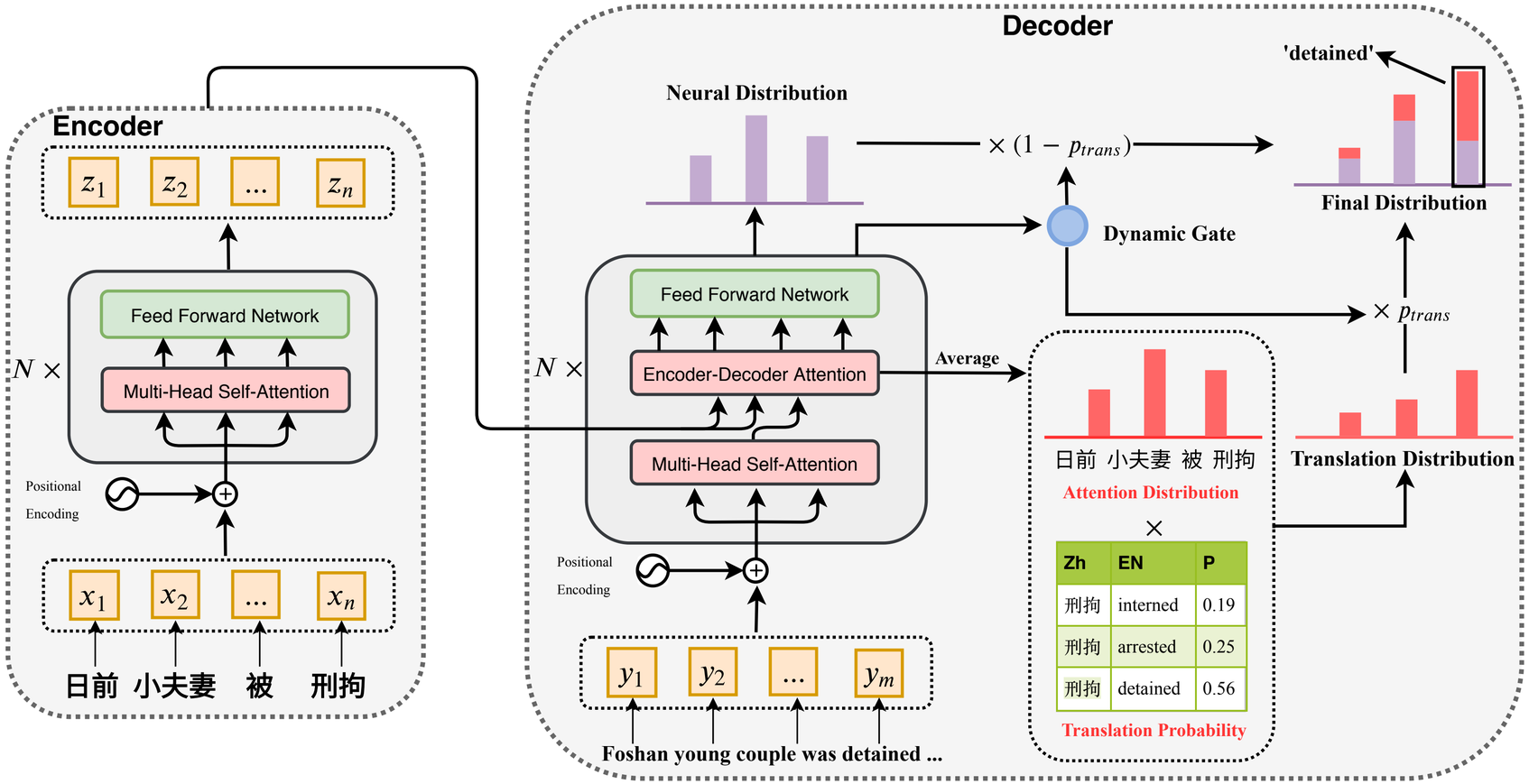

- inspired by the translation pattern in the process of obtaining a cross-lingual summary.
- **first** attend to some words in the source text, **then** translate them into the target language, and **summarize** to get the final summary.
- **motivation**： Inspired by the pointer-generator network and 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.

# Introduction

- pipeline-based approach suffers from error propagation
- some previous researches focus on zero-shot methods to acquire cross-lingual summarization dataset、round-trip translation strategy to obtain large-scale cross-lingual summarization datasets，**But exist the following problems:** **(1)** difficult to migrate to languages with low resources；**(2)** The multi-task methods time-consuming training process.
- **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 final distribution is the weighted sum(weighed by the translating probability) of the translation distribution and the neural distribution.

# Our Model

- Overview




| 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* $p_{\text{trans}}$ 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 $p_{\text{trans}}$) of the neural distribution $P_{\text{N}}$ and the translation distribution $P_{\text{T}}$. Best viewed in color.

# Our Model

- Our proposed method is a hybrid between Transformer and an additional translation layer

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

- **Attend.**

- **Translate.**To achieve that, we obtain a probabilistic bilingual lexicon $P_L(w1 \Rightarrow w2)$ from existing machine translation corpora and then acquire the translation probability $P_T$ based on $P_L(w1 \Rightarrow w2)$.

$$P(w) = p_{\text{trans}} \sum_{i:w_i=w_{\text{src}}} \alpha_{t,i} P_{\text{T}}(w_{\text{src}} \Rightarrow w) + (1 - p_{\text{trans}}) P_{\text{N}}(w)$$

- **Summarize.**

# Our Model

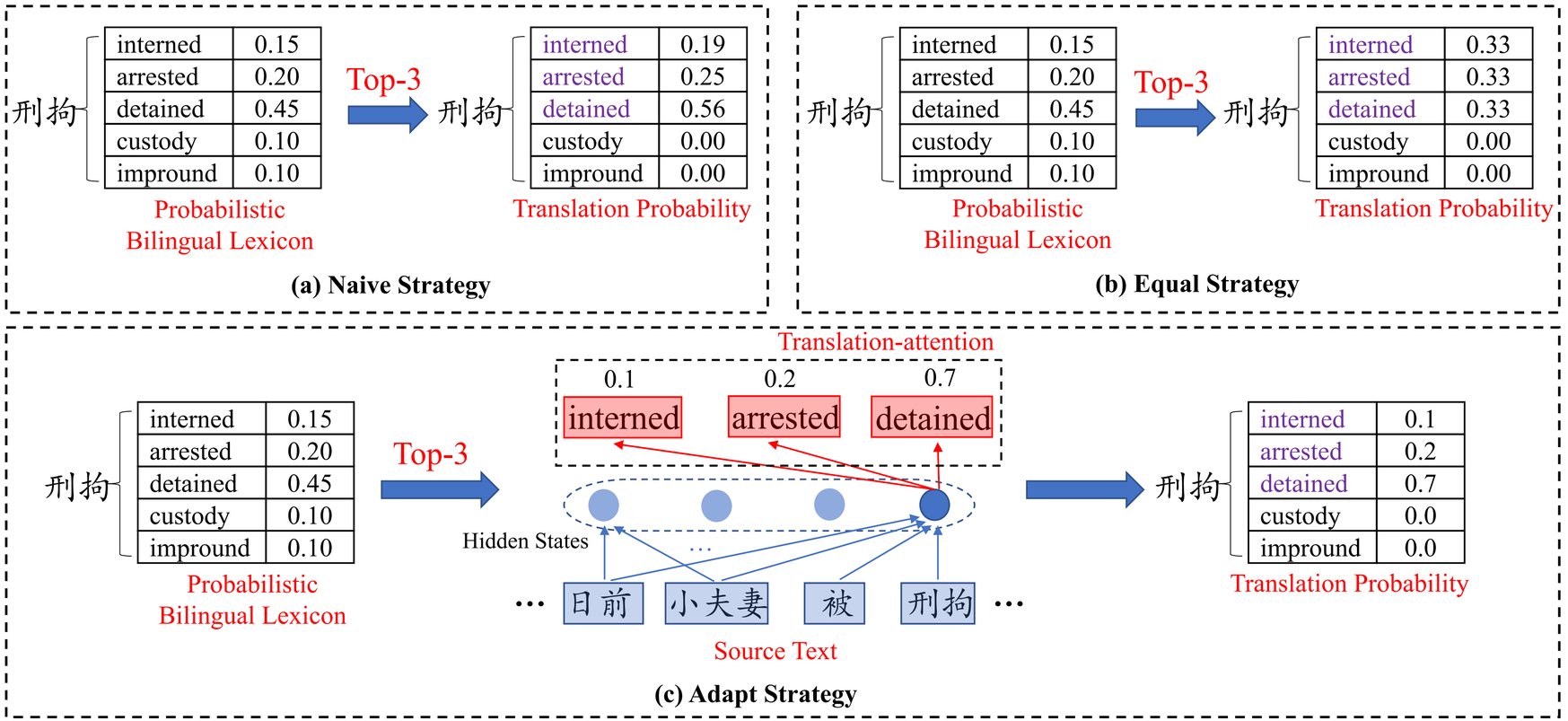



Figure 3: Overview of our three strategies to obtain the translation probability from the probabilistic bilingual lexicon. We take *m*=3 for example.

- Acquisition of the probabilistic bilingual lexicon.
- (1) **Naive.**

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

- (2) **Equal**： This strategy can be considered to achieve the goal of small vocabulary with the help of translation knowledge.
- (3) **Adapt**:

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

# Experiments

- Datasets：En2ZhSum and Zh2EnSum
- Comparative Methods：GETran、GLTran、TNCLS、CLSMS、CLSMT
- We denote our method as ATS

# Experiments

- Results on Zh2EnSum and En2ZhSum

| | 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 ($p < 0.01$).

# Experiments

| 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 impact of $m$.

| 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.

# Experiments

- The impact of segmentation granularity.

| 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 ($p < 0.01$).

- Translating Probability.

| Task | Unit | $p_{\text{trans}}^{\text{macro}}$ | $p_{\text{trans}}^{\text{micro}}$ | $\text{r}^{\text{macro}}$ | $\text{r}^{\text{micro}}$ |
|---|---|---|---|---|---|
| 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 $p_{\text{trans}}$ in ATS-A models. $p_{\text{trans}}^{\text{macro}}$ (%) and $p_{\text{trans}}^{\text{micro}}$ (%) respectively represent the macro-average and micro-average translating probability during decoding. $\text{r}^{\text{macro}}$ (%) and $\text{r}^{\text{micro}}$ (%) respectively represent the ratio of words where $p_{\text{trans}} > 0.5$ during decoding.

# Experiments

- Human Evaluation
- Examples of generated summaries

| 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.

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 $p_{\text{trans}}$.

# Conclusion and Future Work

- 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.