Interactive Decoding for NMT

Jiajun Zhang

NLPR, CASIA

jjzhang@nlpr.ia.ac.cn
www.nlpr.ia.ac.cn/cip/jjzhang.htm
Outline

- Background
- Bidirectional Interactive Inference
- Interactive Inference for Multilingual Translation
- Interactive Inference for Speech Translation
- Summary
Machine Translation

- **Machine Translation**: a technology that automatically translates one language into another

The 2022 Winter Olympics will be held in Beijing

Transformer (Vaswani et al., 2017)
there are ten students

Bidirectional Encoder

Unidirectional Decoder

有十位学生。
Transformer

Bidirectional Encoder

Unidirectional Decoder

students  ten  are  there

<s> students  ten  are

有  十位  学生  。
Problems for Unidirectional Inference

Left-to-Right Decoding

$x_0 \ x_1 \ \cdots \ x_j \ \cdots \ x_m$

$y_{0}^{l2r} \rightarrow y_{1}^{l2r} \rightarrow \cdots \ y_{i}^{l2r} \ \cdots \ \cdots$

Cannot Access the **Right** Contexts

Right-to-Left Decoding

$y_{n'}^{-1}^{r2l} \rightarrow y_{n' - 2}^{r2l} \rightarrow y_{i'}^{r2l} \ \cdots \ \cdots$

Cannot Access the **Left** Contexts
## Problems: Unbalanced Outputs

<table>
<thead>
<tr>
<th>Source</th>
<th>捷克总统哈维卸任新总统仍未确定</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>czech president havel steps down while new president still not chosen</td>
</tr>
<tr>
<td>L2R</td>
<td>czech president leaves office</td>
</tr>
<tr>
<td>R2L</td>
<td>the outgoing president of the czech republic is still uncertain</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>他们正在研制一种超大型的叫做炸弹之母。</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>they are developing a kind of superhuge bomb called the mother of bombs .</td>
</tr>
<tr>
<td>L2R</td>
<td>they are developing a super , big , mother , called the bomb .</td>
</tr>
<tr>
<td>R2L</td>
<td>they are working on a much larger mother called the mother of a bomb .</td>
</tr>
</tbody>
</table>
Problems: Unbalanced Outputs

• Statistical Analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>The first 4 tokens</th>
<th>The last 4 tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2R</td>
<td>40.21%</td>
<td>35.10%</td>
</tr>
<tr>
<td>R2L</td>
<td>35.67%</td>
<td>39.47%</td>
</tr>
</tbody>
</table>

Table: Translation accuracy of the first 4 tokens and last 4 tokens in NIST Chinese-English translation tasks.

How to effectively utilize bidirectional decoding?
Outline

- Background
- Bidirectional Interactive Inference
- Interactive Inference for Multilingual Translation
- Interactive Inference for Speech Translation
- Summary
Solution: Asynchronous Bidirectional Decoding from the Perspective of Model Integration

• ABD-NMT

Problems:
(1) This work still requires two NMT models or decoders.

Question: How to utilize bidirectional decoding more effectively and efficiently?
Bidirectional Interactive Inference

Synchronously Bidirectional Neural Machine Translation

Long Zhou, Jiajun Zhang and Chengqing Zong.
Transactions on ACL 2019.

Synchronous Bidirectional Inference for Neural Sequence Generation

Jiajun Zhang, Long Zhou, Yang Zhao and Chengqing Zong.
Artificial Intelligence, 281(2020)103234, pp.1-19
Synchronous
Bidirectional Neural Machine Translation

Let \( y \) and \( x \) denote the translation and the source sentence, respectively. The probability of \( y \) given \( x \) is defined as:

\[
P(y|x) = \begin{cases} 
\sum_{i=0}^{n-1} p(\tilde{y}_i|\tilde{y}_0 \ldots \tilde{y}_{i-1}, x, \bar{y}_0 \ldots \bar{y}_{i-1}) & \text{if L2R} \\
\sum_{i=0}^{n'-1} p(\tilde{y}_i|\tilde{y}_0 \ldots \tilde{y}_{i-1}, x, \bar{y}_0 \ldots \bar{y}_{i-1}) & \text{if R2L}
\end{cases}
\]

L2R (R2L) inference not only uses its previously generated outputs, but also uses future contexts predicted by R2L (L2R) decoding.
Synchronous Bidirectional Attention

\[
\begin{align*}
\vec{y}_0 &\quad \vec{y}_1 & \ldots & \vec{y}_{i-1} & \vec{y}_i & \vec{y}_{i+1} & \ldots & \vec{y}_1 & \vec{y}_0 \\
\vec{z}_0 &\quad \vec{z}_1 & \ldots & \vec{z}_{i-1} & \vec{z}_i & \vec{z}_{i+1} & \ldots & \vec{z}_1 & \vec{z}_0 \\
\end{align*}
\]

\[
\vec{z}_i = fuse(\vec{H}_i, \vec{H}_i)
\]

\[
\begin{align*}
\vec{y}_0 &\quad \vec{y}_1 & \ldots & \vec{y}_{i-1} & \vec{y}_i & \vec{y}_{i+1} & \ldots & \vec{y}_1 & \vec{y}_0 \\
\vec{z}_0 &\quad \vec{z}_1 & \ldots & \vec{z}_{i-1} & \vec{z}_i & \vec{z}_{i+1} & \ldots & \vec{z}_1 & \vec{z}_0 \\
\end{align*}
\]

\[
\begin{align*}
\vec{z}_i = fuse(\vec{H}_i, \vec{H}_i)
\end{align*}
\]
Synchronous Bidirectional Attention

- Synchronous Bidirectional Dot-Product Attention

\[
\overrightarrow{H}_i = \text{Fusion}(\overrightarrow{H}^f_i, \overrightarrow{H}^b_i)
\]

- Linear Interpolation

\[
\overrightarrow{H}_i = \overrightarrow{H}^f_i + \lambda \ast \overrightarrow{H}^b_i
\]

- Nonlinear Interpolation

\[
\overrightarrow{H}_i = \overrightarrow{H}^f_i + \lambda \ast AF(\overrightarrow{H}^b_i) \begin{cases} \text{tanh} \\ \text{Relu} \end{cases}
\]

- Gate Mechanism

\[
r_t, z_t = \sigma(W^g[\overrightarrow{H}^f_i; \overrightarrow{H}^b_i])
\]

\[
\overrightarrow{H}_i = r_t \odot \overrightarrow{H}^f_i + z_t \odot \overrightarrow{H}^b_i
\]
Synchronous Bidirectional Beam Search Algorithm
Training Objective Function

\[ J(\theta) = \frac{1}{Z} \sum_{z=1}^{Z} \sum_{j=1}^{M} \left\{ \log p\left( \overrightarrow{y}^{(z)}_j \mid \overrightarrow{y}^{(z)}_{<j}, \overleftarrow{y}^{(z)}_{<j}, x^{(z)}(z), \theta \right) + \log p\left( \overleftarrow{y}^{(z)}_j \mid \overrightarrow{y}^{(z)}_{<j}, \overleftarrow{y}^{(z)}_{<j}, x^{(z)}(z), \theta \right) \right\} \]
Experiments: Machine Translation

• Setup
  - Dataset:
    1. NIST Chinese-English translation (2M, 30K tokens, MT03-06 as test set)
    2. WMT14 English-German translation (4.5M, 37K shared tokens, newstest2014 as test set)
  - Train details:
    1. Transformer_big setting
    2. Chinese-English: 1 GPUs, single model, case-insensitive BLEU.
    3. English-German: 3 GPUs, model averaging, case-sensitive BLEU.
Experiments: Machine Translation

• Baselines

  - **Moses**: an Open source phrase-based SMT system.
  - **RNMT**: RNN-based NMT with default setting.
  - **Transformer**: Predict target sentence from left to right.
  - **Transformer(R2L)**: Predict sentence from right to left.
  - **Rerank-NMT**: (1) first run beam search to obtain two k-best lists; (2) then re-score and get the best candidate.
  - **ABD-NMT**: (1) use backward decoder to generate reverse sequence states; (2) perform beam search on the forward decoder to find the best translation.
Experiments: Machine Translation

• Results on Chinese-English Translation
  ➢ Translation Quality

<table>
<thead>
<tr>
<th>Model</th>
<th>DEV</th>
<th>MT03</th>
<th>MT04</th>
<th>M05</th>
<th>MT06</th>
<th>AVE</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses</td>
<td>37.85</td>
<td>37.47</td>
<td>41.20</td>
<td>36.41</td>
<td>36.03</td>
<td>37.78</td>
<td>-9.41</td>
</tr>
<tr>
<td>RNMT</td>
<td>42.43</td>
<td>42.43</td>
<td>44.56</td>
<td>41.94</td>
<td>40.95</td>
<td>42.47</td>
<td>-4.72</td>
</tr>
<tr>
<td>Transformer</td>
<td>48.12</td>
<td>47.63</td>
<td>48.32</td>
<td>47.51</td>
<td>45.31</td>
<td>47.19</td>
<td>-</td>
</tr>
<tr>
<td>Transformer(R2L)</td>
<td>47.81</td>
<td>46.79</td>
<td>47.01</td>
<td>46.50</td>
<td>44.13</td>
<td>46.11</td>
<td>-1.08</td>
</tr>
<tr>
<td>Rerank-NMT</td>
<td>49.18</td>
<td>48.23</td>
<td>48.91</td>
<td>48.73</td>
<td>46.51</td>
<td>48.10</td>
<td>+0.91</td>
</tr>
<tr>
<td>ABD-NMT</td>
<td>48.28</td>
<td>49.47</td>
<td>48.01</td>
<td>48.19</td>
<td>47.09</td>
<td>48.19</td>
<td>+1.00</td>
</tr>
<tr>
<td><strong>Our Model</strong></td>
<td><strong>50.99</strong></td>
<td><strong>51.61</strong></td>
<td><strong>51.41</strong></td>
<td><strong>51.19</strong></td>
<td><strong>49.84</strong></td>
<td><strong>51.01</strong></td>
<td><strong>+3.82</strong></td>
</tr>
</tbody>
</table>

Table: Evaluation of translation quality for Chinese-English translation tasks with case-insensitive BLEU scores.
Experiments: Machine Translation

• Results on English-German Translation

Table: Results of English-German translation using case-sensitive BLEU.

<table>
<thead>
<tr>
<th>Model</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNMT (Wu et al., 2016)</td>
<td>24.61</td>
</tr>
<tr>
<td>Conv (Gehring et al., 2017)</td>
<td>25.16</td>
</tr>
<tr>
<td>AttIsAll (Vaswani et al., 2017)</td>
<td>28.40</td>
</tr>
<tr>
<td>Transformer</td>
<td>27.72</td>
</tr>
<tr>
<td>Transformer(R2L)</td>
<td>27.13</td>
</tr>
<tr>
<td>Rerank-NMT</td>
<td>27.81</td>
</tr>
<tr>
<td>ABD-NMT</td>
<td>28.22</td>
</tr>
<tr>
<td><strong>Our Model</strong></td>
<td><strong>29.21</strong></td>
</tr>
</tbody>
</table>

Baselines (+1.49)
MT Analysis: Unbalanced Outputs

Get the best translation accuracy no matter for the first 4 words and the last 4 words.

Figure: Translation accuracy of the first 4 tokens and last 4 tokens for L2R, R2L, Rerank-NMT, ABD-NMT and our proposed model.
### MT Analysis: Case Study

<table>
<thead>
<tr>
<th>Source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>捷克总统哈维卸任 新总统仍未确定</td>
<td>czech president havel steps down while new president still not chosen</td>
</tr>
<tr>
<td>Reference</td>
<td>czech president leaves office</td>
</tr>
<tr>
<td>L2R</td>
<td>the outgoing president of the czech republic is still uncertain</td>
</tr>
<tr>
<td>R2L</td>
<td>czech president havel leaves office , new president yet to be determined</td>
</tr>
<tr>
<td>Ours</td>
<td>czech president havel leaves office , new president yet to be determined</td>
</tr>
</tbody>
</table>

L2R produces **good prefix**, whereas R2L generates **better suffixes**.

**Our approach** can make full use of bidirectional decoding and produce balanced outputs in these cases.
MT Analysis: Parameters and Speeds

Table: Statistics of parameters, training and testing speeds. Train denotes the number of global training steps processed per second; Test indicates the amount of translated sentences in one second.

<table>
<thead>
<tr>
<th>Model</th>
<th>Param</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Train</td>
</tr>
<tr>
<td>Transformer</td>
<td>207.8M</td>
<td>2.07</td>
</tr>
<tr>
<td>Transformer(R2L)</td>
<td>207.8M</td>
<td>2.07</td>
</tr>
<tr>
<td>Rerank-NMT</td>
<td>415.6M</td>
<td>1.03</td>
</tr>
<tr>
<td>ABD-NMT</td>
<td>333.8M</td>
<td>1.18</td>
</tr>
<tr>
<td>Our Model</td>
<td>207.8M</td>
<td>1.26</td>
</tr>
</tbody>
</table>

No additional parameters except for lambda

Slightly Slower than baseline Transformer
Outline

- Background
- Bidirectional Interactive Inference
- Interactive Inference for Multilingual Translation
- Interactive Inference for Speech Translation
- Summary
Interactive Inference for Multilingual Translation

\[
x_0 \ x_1 \ \ldots \ x_j \ \ldots \ x_m
\]

**Left-to-Right Decoding**

\[
y_0^{l2r} \rightarrow y_1^{l2r} \rightarrow \ldots \ y_i^{l2r} \ \ldots \ \ldots
\]

**Right-to-Left Decoding**

\[
y_{n'-1}^{r2l} \rightarrow y_{n'-2}^{r2l} \rightarrow \ldots \ y_i^{r2l} \ \ldots \ \ldots
\]

**English-to-Chinese Decoding**

\[
y_0 \rightarrow y_1 \rightarrow \ldots \ y_i \ \ldots \ \ldots
\]

**English-to-Japanese Decoding**

\[
z_0 \rightarrow z_1 \rightarrow \ldots \ z_j \ \ldots \ \ldots
\]
Interactive Inference for Multilingual Translation

Synchronously Generating Two Languages with Interactive Decoding

Yining Wang, Jiajun Zhang, Long Zhou, Yuchen Liu and Chengqing Zong.


Interactive Decoding for Multilingual Neural Machine Translation

Hao He, Qian Wang, Zhipeng Yu, Yang Zhao, Jiajun Zhang and Chengqing Zong.

Conventional Multilingual Translation

- **Separate Encoder or Decoder network**
- **Shared Encoder or Decoder network**
- **Shared with partial parameter**

Diagram:
- **Shared Encoder**
  - Source Language: "En"
  - Target Language 1: "De"
  - Decoder 1
  - Target Language 2: "Nl"
  - Decoder 2
- **Shared Encoder**
  - Source Language: "En"
  - Target Language 1: "De"
  - Shared Decoder
  - Target Language 2: "Nl"
- **Shared Encoder**
  - Source Language: "En"
  - Target Language 1: "De"
  - Shareable Parameters
  - Decoder 1
  - Decoder 2
  - Target Language 2: "Nl"
Interactive Decoding for Multilingual Translation

I bought a book today.

English

Chinese

Japanese
Interactive Decoding for Multilingual Translation

Synchronous Self-Attention Model:

\[ H'_1 = \text{SyncAtt} (Q_1, [K_1; K_2], [V_1; V_2]) \]
\[ H'_2 = \text{SyncAtt} (Q_2, [K_1; K_2], [V_1; V_2]) \]
Synchronous Bi-language Attention
Interactive Decoding for Multilingual Translation

I cook for my parents today

Chinese Forward (L2R)

我 (I) → 今天 (today) → 为 (for) → 我的 (my) → 父母 (parents) → 做饭 (cook) → pad → pad

Chinese Backward (R2L)

做饭 (cook) → 爸妈 (parents) → 我的 (my) → 给 (for) → 今天 (today) → 我 (I) → pad → pad

Japanese Forward (L2R)

今日 (today) → は (for) → 両親 (parents) → の (of) → ために (for) → 料理 (cook) → を (to) → します (does)

Japanese Backward (R2L)

作った (cooked) → を (for) → 夕食 (dinner) → ために (for) → の (of) → 両親 (parents) → は (for) → 今日 (today)

T=1  T=2  T=3  T=4  T=5  T=6  T=7  T=8
Interactive Decoding for Multilingual Translation

(a) The synchronous multilingual decoder
(b) The cross lingual attention (CLA)
(c) The attentional fusion of CLA
Training Data Construction

Training Instance Format Requirement: trilingual translation example \((x, y, z)\) in which \((x, y)\) and \((x, z)\) are parallel sentence pairs.

\[ x_0 \ x_1 \ \cdots \ x_j \ \cdots \ x_m \]

\[ \begin{align*}
y_0 &\rightarrow y_1 \rightarrow \cdots \rightarrow y_i \rightarrow \cdots \rightarrow \cdots \\
\text{English-Chinese Decoding} & \\
\begin{align*}
z_0 &\rightarrow z_1 \rightarrow \cdots \rightarrow z_j \rightarrow \cdots \rightarrow \cdots \\
\text{English-to-Japanese Decoding} &
\end{align*}
\]
Experiments

1. Large Scale

<table>
<thead>
<tr>
<th></th>
<th>WMT14 subset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>En-De</td>
</tr>
<tr>
<td>Train</td>
<td>2.43M</td>
</tr>
<tr>
<td>Test</td>
<td>3003</td>
</tr>
</tbody>
</table>

2. Small Scale

<table>
<thead>
<tr>
<th></th>
<th>IWSLT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>En-Ja</td>
</tr>
<tr>
<td>Train</td>
<td>223K</td>
</tr>
<tr>
<td>Test</td>
<td>3003</td>
</tr>
</tbody>
</table>
Main Results

- **English-Chinese/Japanese and English-German/French**

<table>
<thead>
<tr>
<th>Method</th>
<th>En-Zh/Ja</th>
<th>En-De/Fr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>En-Zh</td>
<td>En-Ja</td>
</tr>
<tr>
<td><strong>Indiv</strong></td>
<td>15.68</td>
<td>16.56</td>
</tr>
<tr>
<td><strong>Indiv + pseudo</strong></td>
<td>16.72</td>
<td>18.02</td>
</tr>
<tr>
<td><strong>Multi</strong></td>
<td>17.06</td>
<td>18.31</td>
</tr>
<tr>
<td><strong>Multi + pseudo</strong></td>
<td>17.10</td>
<td>18.40</td>
</tr>
<tr>
<td><strong>Sync-Trans</strong></td>
<td>17.97</td>
<td>19.31</td>
</tr>
</tbody>
</table>

- **Indiv**: System learned with bilingual training
- **Multi**: shared encoder-decoder networks
- **Sync-Trans** significantly outperforms Indiv and Multi
Main Results

- Large-scale WMT Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>WMT14 subset</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>En-De</td>
<td>En-Fr</td>
<td>En-De</td>
</tr>
<tr>
<td>Indiv</td>
<td>24.33</td>
<td>37.12</td>
<td>26.53</td>
</tr>
<tr>
<td>Multi</td>
<td>23.46</td>
<td>36.33</td>
<td>25.81</td>
</tr>
<tr>
<td>Sync-Trans</td>
<td>24.84†*</td>
<td>37.66†*</td>
<td>27.01†*</td>
</tr>
</tbody>
</table>

- Sync-Trans significantly outperforms Indiv and Multi
Outline

- Background
- Bidirectional Interactive Inference
- Interactive Inference for Multilingual Translation
- Interactive Inference for Speech Translation
- Summary
Interactive Inference for other Two Tasks

\[ x_0 \rightarrow x_1 \rightarrow \cdots \rightarrow x_j \rightarrow \cdots \rightarrow x_m \]

- **Left-to-Right Decoding**
  \[ y_0^{l2r} \rightarrow y_1^{l2r} \rightarrow \cdots \rightarrow y_i^{l2r} \rightarrow \cdots \rightarrow \cdots \]

- **Right-to-Left Decoding**
  \[ y_{n'-1}^{r2l} \rightarrow y_{n'-2}^{r2l} \rightarrow \cdots \rightarrow y_{i'}^{r2l} \rightarrow \cdots \rightarrow \cdots \]

**Speech Recognition**

\[ x_0 \rightarrow x_1 \rightarrow \cdots \rightarrow x_i \rightarrow \cdots \rightarrow \cdots \]

**Speech-to-Text Translation**

\[ y_0 \rightarrow y_1 \rightarrow \cdots \rightarrow y_j \rightarrow \cdots \rightarrow \cdots \]
Interactive Inference for Speech Translation

Synchronous Speech Recognition and Speech-to-Text Translation with Interactive Decoding

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang and Chengqing Zong.

Speech Translation - Pipeline Model

- Resegment
- Punctuation
- Text Normalization
- Error Correction

Speech Recognition → Text Processing → Machine Translation

Speech Signal Processing

Noise reduction
Audio Converter
Audio Segmentation
Language identification

Time delay
Error propagation
Parameter redundancy

More robust NMT model
Speech Translation – End-to-End

Speech Translation

Encoder → Decoder

Speech Signal Processing

Speech Synthesis
Interactive Inference for Speech Recognition and Speech-to-Text Translation

• Speech Features

Original signal

Mel filter

Filter bank

Discrete Fourier transform

MFCC
Interactive Inference for Speech Recognition and Speech-to-Text Translation

• Overall Architecture
Interactive Inference for Speech Recognition and Speech-to-Text Translation

- **Interactive Attention**

  \[
  H_1 = \text{Attention}(Q_1, K_1, V_1)
  \]

  \[
  H_2 = \text{Attention}(Q_1, K_2, V_2)
  \]

  \[
  H_{\text{final}} = \text{Fusion}(H_1, H_2)
  \]

- **Fusion**
  - Linear Interpolation
    \[
    H_{\text{final}} = \lambda_1 * H_1 + \lambda_2 * H_2
    \]
  - Nonlinear Interpolation
    \[
    H_{\text{final}} = \lambda_1 * H_1 + \lambda_2 * \text{tanh}(H_2)
    \]
  - Gate Interpolation
    \[
    r, z = \sigma(\mathcal{W}[H_1; H_2])
    \]
    \[
    H_{\text{final}} = r \odot H_1 + z \odot H_2
    \]
Interactive Inference for Speech Recognition and Speech-to-Text Translation

• Training

\[
J(\theta) = \sum_{z=1}^{Z} \sum_{j=1}^{n} \{ \log p(x_j^z|x_{<j}^z, y_{<j}^z, s^z, \theta_{Rec}) \\
+ \log p(y_j^z|y_{<j}^z, x_{<j}^z, s^z, \theta_{Tran}) \}
\]
Interactive Inference for Speech Recognition and Speech-to-Text Translation

• Experimental Setup

  ➢ Dataset:
    TED En-Fr, En-Zh
  
  ➢ Train details:
    (1) Transformer_big setting
    (2) English-Chinese: 2 GPUs, character BLEU
    (3) English-French: 2 GPUs, tokenizer BLEU
## Interactive Inference for Speech Recognition and Speech-to-Text Translation

### Data Size

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Total</th>
<th>Source (per segment)</th>
<th>Target (per segment)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>segments</td>
<td>hours</td>
<td>frames</td>
</tr>
<tr>
<td>Fisher/Callhome (En-Es)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>train</td>
<td>138,819</td>
<td>138:00</td>
<td>762</td>
</tr>
<tr>
<td>dev</td>
<td>3,961</td>
<td>2:00</td>
<td>673</td>
</tr>
<tr>
<td>test</td>
<td>3,641</td>
<td>3:44</td>
<td>657</td>
</tr>
<tr>
<td>TED (En-Zh/En-Fr)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>train</td>
<td>305,971</td>
<td>527:00</td>
<td>662</td>
</tr>
<tr>
<td>dev</td>
<td>1,148</td>
<td>2:23</td>
<td>659</td>
</tr>
<tr>
<td>test</td>
<td>1,223</td>
<td>2:37</td>
<td>624</td>
</tr>
</tbody>
</table>
Interactive Inference for Speech Recognition and Speech-to-Text Translation

• Baselines

  - **Pipeline**: Transformer ASR + Transformer MT
  - **Pre-trained E2E**: Pretrain on ASR, fintune on ST
  - **Multi-task**: ASR + ST with encoder shared
  - **Two-stage**: (1) use the first decoder to generate transcription sequence; (2) use the output of first decoder on the second decoder
Interactive Inference for Speech Recognition and Speech-to-Text Translation

- Overall Results

<table>
<thead>
<tr>
<th>Model</th>
<th>En-De</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER(↓)</td>
<td>BLEU(↑)</td>
<td>WER(↓)</td>
<td>BLEU(↑)</td>
<td>WER(↓)</td>
<td>BLEU(↑)</td>
<td>WER(↓)</td>
</tr>
<tr>
<td>MT</td>
<td>/</td>
<td>22.19</td>
<td>/</td>
<td>30.68</td>
<td>/</td>
<td>25.01</td>
<td>/</td>
</tr>
<tr>
<td>Multi-task</td>
<td>14.20</td>
<td>19.08</td>
<td>13.04</td>
<td>28.71</td>
<td>13.43</td>
<td>20.60</td>
<td>14.01</td>
</tr>
<tr>
<td>Two-stage</td>
<td>14.27</td>
<td>20.08</td>
<td>13.34</td>
<td><strong>30.08</strong></td>
<td>13.55</td>
<td>20.29</td>
<td>13.85</td>
</tr>
<tr>
<td>Interactive</td>
<td><strong>14.16</strong></td>
<td><strong>21.11</strong></td>
<td><strong>12.58</strong></td>
<td>29.79</td>
<td><strong>13.38</strong></td>
<td><strong>21.68</strong></td>
<td><strong>13.52</strong></td>
</tr>
</tbody>
</table>
Outline

- Background
- Bidirectional Interactive Inference
- Interactive Inference for Multilingual Translation
- Interactive Inference for Speech Translation
- Summary
Summary

• The synchronous bidirectional Inference model that can take full advantage of both history and future information provided by bidirectional decoding states, achieving promising results.

• The bidirectional inference model can be generalized to generate two languages synchronously and interactively.

• The bidirectional inference model can be generalized to speech translation that performs ASR and ST synchronously and interactively.
Unified Interactive Text Generation from Text, Speech and Image

(a) Text, Speech or Image encoding:

I had no idea what was coming

or

Encoder

(b) Conventional text generation:

or

\( T=1 \) \( T=2 \) \( T=3 \) \( T=4 \) \( \ldots \)

\( y_1 \) \( y_2 \) \( y_3 \) \( y_4 \) \( \ldots \) \( y_m \)

(c) Synchronous Interactive text generation:

or

\( T=1 \) \( T=2 \) \( T=3 \) \( T=4 \) \( \ldots \)

\( z_1 \) \( z_2 \) \( z_3 \) \( z_4 \) \( \ldots \) \( z_n \)
谢谢！

Thanks!