Controllable and Explainable End-to-End Speech Translation

Shinji Watanabe and Brian Yan

Language Technologies Institute
Carnegie Mellon University

SIG SLT Seminar, November 18, 2022
Breaking the End-to-End Black Box

“I really need it”

Translate

Ich wirklich das brauche

Ich wirklich brauche

What went wrong and how can it be fixed?
Breaking the End-to-End Black Box

What went wrong and how can it be fixed?

- Generating translations that are too short? 

Part 1
Breaking the End-to-End Black Box

What went wrong and how can it be fixed?

- Generating translations that are too short?  
- Recognizing what was said?
Breaking the End-to-End Black Box

Still, I don’t know why this error happens

What went wrong and how can it be fixed?

- Generating translations that are too short?
- Recognizing what was said?

Part 1

Part 2
Breaking the End-to-End Black Box

"I really need it" Translate Ich wirklich das brauche

"I really need [it]" Ich wirklich [das] brauche

Still, I don’t know why this error happens because I don’t know that “it” corresponds to “das”

What went wrong and how can it be fixed?

- Generating translations that are too short? Part 1
- Recognizing what was said? Part 2
- Explaining why this is mistranslated? Part 3
Today’s Talk

- CMU’s IWSLT 2022 Dialect Speech Translation System
  - **Part 1:** Controlling ST output lengths via joint CTC/attention
  - **Part 2:** Controlling/explaining ST via searchable ASR intermediates

- Explainable E2E Speech Translation via Operation Sequence Generation
  - **Part 3:** Explaining ST via word-level ASR alignments
Today’s Talk

- CMU’s IWSLT 2022 Dialect Speech Translation System
  - **Part 1**: Controlling ST output lengths via joint CTC/attention
  - **Part 2**: Controlling/explaining ST via searchable ASR intermediates

- Explainable E2E Speech Translation via Operation Sequence Generation
  - **Part 3**: Explaining ST via word-level ASR alignments

---

CMU’s IWSLT 2022 Dialect Speech Translation System

Brian Yan, Patrick Fernandes, Siddharth Dalmia, Jiatong Shi, Yifan Peng, Dan Berrebbi, Xinyi Wang, Graham Neubig, Shinji Watanabe

1Language Technologies Institute, Carnegie Mellon University, USA
2Instituto Superior Técnico & LUMILS (Lisbon ELLIS Unit), Portugal
3Electrical and Computer Engineering, Carnegie Mellon University, USA
4Human Language Technology Center of Excellence, Johns Hopkins University, USA
(byan, pfernand, sdalmia, jiatongs@cs.cmu.edu)

Abstract

In particular, our contributions are the following:

1. Dialectal transfer from large paired MSA corpora to improve ASR and MT systems ($\S$3.1)

ALIGN, WRITE, RE-ORDER: EXPLAINABLE END-TO-END SPEECH TRANSLATION VIA OPERATION SEQUENCE GENERATION

Motoki Omachi, Brian Yan, Siddharth Dalmia, Yuya Fujita, Shinji Watanabe

1Yahoo Japan Corporation, Tokyo, JAPAN; 2Carnegie Mellon University, PA, USA

ABSTRACT

The black-box nature of end-to-end speech translation (E2E ST) systems makes it difficult to understand how source language inputs are being mapped to the target language. To solve this problem, we would like to introduce a new automatic speech assessment (ASA) framework.
Length Control in Speech Translation

What is a good translation?

- Correct meaning
- Correct **length**
  - e.g. isometric ST for subtitling

<table>
<thead>
<tr>
<th>Source</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline MT</td>
<td>Es ist tatsächlich die wahre Integration von Mensch und Maschine.</td>
</tr>
<tr>
<td>Isometric MT</td>
<td>Es ist die wirkliche Integration von Mensch und Maschine.</td>
</tr>
</tbody>
</table>

*Example from IWSLT 2022 Isometric ST Track*
Length Control in Speech Translation

What is a good translation?

- Correct meaning
- Correct **length**
  - e.g. isometric ST for subtitling

**Problem:** autoregressive decoders do not have robust end-detection

- Reliant on length penalty/bonus hyperparameter; not robust across domains/datasets

<table>
<thead>
<tr>
<th>Source</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline MT</td>
<td>Es ist tatsächlich die wahre Integration von Mensch und Maschine.</td>
</tr>
<tr>
<td>Isometric MT</td>
<td>Es ist die wirkliche Integration von Mensch und Maschine.</td>
</tr>
</tbody>
</table>

*Example from IWSLT 2022 Isometric ST Track*

---

Degenerating quality due to incorrect length penalty leading to overly long outputs.
Length Control in Speech Translation

**Problem:** autoregressive decoders do not have robust end-detection
- Reliant on length penalty/bonus hyperparameter; not robust across domains/datasets

Over-tuning easily happens! This was our experience in IWSLT 2021

---

### Results on “original” blind test set; similar lengths to dev data

<table>
<thead>
<tr>
<th>System</th>
<th>segm.</th>
<th>data condition</th>
<th>BLEU.TEDRef</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESPNet-ST</td>
<td>Own</td>
<td>Constrained</td>
<td>26.0</td>
</tr>
<tr>
<td>HW-TSC</td>
<td>Own</td>
<td>Constrained</td>
<td>25.4</td>
</tr>
<tr>
<td>KIT</td>
<td>Own</td>
<td>Constrained</td>
<td>25.4</td>
</tr>
<tr>
<td>ESPNet-ST</td>
<td>Own</td>
<td>Constrained</td>
<td>24.7</td>
</tr>
<tr>
<td>FBK</td>
<td>Own</td>
<td>Constrained</td>
<td>24.7</td>
</tr>
<tr>
<td>UPC†</td>
<td>Own</td>
<td>Unconstrained</td>
<td>24.6</td>
</tr>
<tr>
<td>AppTeK</td>
<td>Own</td>
<td>Constrained</td>
<td>24.5</td>
</tr>
<tr>
<td>VOLCTRANS</td>
<td>Given</td>
<td>Constrained</td>
<td>24.3</td>
</tr>
<tr>
<td>KIT</td>
<td>Own</td>
<td>Constrained</td>
<td>23.2</td>
</tr>
<tr>
<td>AppTeK</td>
<td>Own</td>
<td>Constrained</td>
<td>23.1</td>
</tr>
<tr>
<td>NIUTRANS</td>
<td>Own</td>
<td>Constrained</td>
<td>22.8</td>
</tr>
<tr>
<td>OPPO</td>
<td>Given</td>
<td>Constrained</td>
<td>22.6</td>
</tr>
<tr>
<td>VOLCTRANS</td>
<td>Given</td>
<td>Constrained</td>
<td>22.2</td>
</tr>
<tr>
<td>VUS</td>
<td>Given</td>
<td>Constrained</td>
<td>13.7</td>
</tr>
<tr>
<td>BUT</td>
<td>Given</td>
<td>Unconstrained</td>
<td>11.4</td>
</tr>
<tr>
<td>Li</td>
<td>Given</td>
<td>Constrained</td>
<td>0.2</td>
</tr>
</tbody>
</table>

---

### Results on new blind test set w/ shorter references (different annotation guidelines)

<table>
<thead>
<tr>
<th>System</th>
<th>segm.</th>
<th>data condition</th>
<th>BLEU.NewRef</th>
<th>BLEU.TEDRef</th>
<th>BLEU.MultiRef</th>
</tr>
</thead>
<tbody>
<tr>
<td>HW-TSC</td>
<td>Own</td>
<td>Constrained</td>
<td>24.6</td>
<td>20.3</td>
<td>34.0</td>
</tr>
<tr>
<td>KIT</td>
<td>Own</td>
<td>Constrained</td>
<td>23.4</td>
<td>19.0</td>
<td>32.0</td>
</tr>
<tr>
<td>AppTeK</td>
<td>Own</td>
<td>Constrained</td>
<td>22.6</td>
<td>18.3</td>
<td>31.0</td>
</tr>
<tr>
<td>KIT</td>
<td>Own</td>
<td>Constrained</td>
<td>22.0</td>
<td>18.1</td>
<td>30.3</td>
</tr>
<tr>
<td>AppTeK</td>
<td>Own</td>
<td>Constrained</td>
<td>21.9</td>
<td>18.1</td>
<td>30.4</td>
</tr>
<tr>
<td>VOLCTRANS</td>
<td>Given</td>
<td>Constrained</td>
<td>21.8</td>
<td>17.1</td>
<td>29.5</td>
</tr>
<tr>
<td>UPC†</td>
<td>Own</td>
<td>Constrained</td>
<td>21.8</td>
<td>18.3</td>
<td>30.6</td>
</tr>
<tr>
<td>VOLCTRANS</td>
<td>Given</td>
<td>Constrained</td>
<td>21.7</td>
<td>18.7</td>
<td>31.3</td>
</tr>
<tr>
<td>ESPNet-ST</td>
<td>Own</td>
<td>Constrained</td>
<td>21.7</td>
<td>18.2</td>
<td>30.6</td>
</tr>
<tr>
<td>FBK</td>
<td>Own</td>
<td>Constrained</td>
<td>21.6</td>
<td>18.4</td>
<td>30.6</td>
</tr>
<tr>
<td>OPPO</td>
<td>Given</td>
<td>Constrained</td>
<td>21.5</td>
<td>17.8</td>
<td>30.2</td>
</tr>
<tr>
<td>ESPNet-ST</td>
<td>Own</td>
<td>Constrained</td>
<td>21.2</td>
<td>19.3</td>
<td>31.4</td>
</tr>
<tr>
<td>NIUTRANS</td>
<td>Own</td>
<td>Constrained</td>
<td>20.6</td>
<td>19.6</td>
<td>30.3</td>
</tr>
<tr>
<td>VUS</td>
<td>Given</td>
<td>Constrained</td>
<td>15.3</td>
<td>12.4</td>
<td>20.9</td>
</tr>
<tr>
<td>BUT</td>
<td>Given</td>
<td>Unconstrained</td>
<td>11.7</td>
<td>9.8</td>
<td>16.1</td>
</tr>
<tr>
<td>Li</td>
<td>Given</td>
<td>Constrained</td>
<td>3.6</td>
<td>2.7</td>
<td>4.8</td>
</tr>
</tbody>
</table>
Hmm. We had the *same experience* before…
Joint CTC/Attention for ASR [Kim+ (2017), Hori+ (2017), Watanabe+ (2017)]
Joint CTC/Attention for ASR [Kim+ (2017), Hori+ (2017), Watanabe+ (2017)]
Joint CTC/Attention for **ASR** [Kim+ (2017), Hori+ (2017), Watanabe+ (2017)]

Use **CTC** for decoding together with the attention decoder

CTC explicitly eliminates non-monotonic alignment
More robust input/output alignment of attention

- Alignment of one selected utterance from CHiME4 ASR task

**Attention Model**

Epoch 1  
Epoch 3  
Epoch 5  
Epoch 7  
Epoch 9

**Joint CTC/attention model**

Correctly control the length!
Let’s Apply Joint CTC/Attention Architecture to Speech Translation

Hierarchical Encoding (ASR→ST)

\[ \mathcal{L} = \mathcal{L}_{SRCCTC} + \lambda_1 \mathcal{L}_{TGTCTC} + \lambda_2 \mathcal{L}_{ATTN} \]
Let’s Apply Joint CTC/Attention Architecture to Speech Translation

Hierarchical Encoding (ASR→ST)

\[ L = L_{SRCCTC} + \lambda_1 L_{TGTCTC} + \lambda_2 L_{ATTN} \]

But wait … CTC is **monotonic** and ST requires re-ordering
Joint CTC/Attention Architecture

Hierarchical Encoding (ASR→ST)

![Diagram of Joint CTC/Attention Architecture]

Self-attentional encoder learn to re-order
- Final encoder representations become **monotonic** w.r.t. target translations
- Decoder source attention patterns:

\[ \mathcal{L} = \mathcal{L}_{\text{SRCCTC}} + \lambda_1 \mathcal{L}_{\text{TGTCTC}} + \lambda_2 \mathcal{L}_{\text{ATTN}} \]

But wait … CTC is **monotonic** and ST requires re-ordering.
Joint CTC/Attention Architecture

Hierarchical Encoding (ASR→ST)

Self-attentional encoder learn to re-order

- Final encoder representations become **monotonic** w.r.t. target translations
- Decoder source attention patterns:

\[ \mathcal{L} = \mathcal{L}_{\text{SRCCTC}} + \lambda_1 \mathcal{L}_{\text{TGTCTC}} + \lambda_2 \mathcal{L}_{\text{ATTN}} \]

But wait … CTC is **monotonic** and ST requires re-ordering
Joint CTC/Attention Architecture

Part 1: Controlling ST output lengths via joint CTC/attention

Hierarchical Encoding (ASR→ST)

- Final encoder representations become **monotonic** w.r.t. target translations
- Decoder source attention patterns:

\[ \mathcal{L} = \mathcal{L}_{\text{SRCCTC}} + \lambda_1 \mathcal{L}_{\text{TGTCTC}} + \lambda_2 \mathcal{L}_{\text{ATTN}} \]

But wait … CTC is **monotonic** and ST requires re-ordering
Joint CTC/Attention Decoding: 2 Synchronous Methods

CTC prefix scores indirectly help end-detection by penalizing hypotheses of incorrect length.

**Algorithm 2 Output-Synchronous Step Function:**
attentional decoder proposes candidates to expand hypotheses which are all of \( l \)-length at step \( l \).

1. procedure \( \text{OUTPUTSTEP}(\text{prtHs}, X, l, p, \text{maxL}) \)
2. newPrtHs = \{\}; endHs = \{
3. for \( y_{l-1} \in \text{prtHs} \) do
4. \( \text{attnCnds} = \text{top-k}(P_{\text{Attn}}(y_l|X, y_{l-1}), k = p) \)
5. for \( c \in \text{attnCnds} \) do
6. \( y_{1:l} = y_{1:l-1} \oplus c \)
7. \( \alpha_{\text{CTC}} = \text{CTCScore}(y_{1:l}, X_{1:T}) \)
8. \( \alpha_{\text{Attn}} = \text{AttnScore}(y_{1:l}, X_{1:T}) \)
9. \( \beta = \text{LengthPen}(y_{1:l}) \)
10. \( P_{\text{Beam}}(y_{1:l}|X) = \alpha_{\text{CTC}} + \alpha_{\text{Attn}} + \beta \)
11. if \((\text{c is <eos>}) \) or \((l \text{ is maxL})\) then
12. \( \text{endHs}[y_{1:l}] = P_{\text{Beam}}(\cdot) \)
13. else
14. \( \text{newPrtHs}[y_{1:l}] = P_{\text{Beam}}(\cdot) \)
15. end if
16. end for
17. return newPrtHs, endHs
19. end procedure
Joint CTC/Attention Decoding: 2 Synchronous Methods

CTC prefix scores \textbf{indirectly} help end-detection by penalizing hypotheses of incorrect length

CTC \textbf{directly} handles end-detection by consuming all input frames
Joint CTC/Attention: Robust End-Detection

Carefully tuning length penalty may not be necessary!
Part 1: Controlling ST output lengths via joint CTC/attention

CMU's IWSLT 2022 Dialect Speech Translation System, Yan et al., IWSLT 2022

Joint CTC/Attention: Results

Attention-Only vs. CTC/Attention (BLEU)

- **IWSLT22 (Ta-En):** +1.1
- **MuST-C (En-De):** +1.4
- **MuST-C (En-Ja):** +1.0
- **MTedx (All-En):** +2.4
- **IWSLT14 (De-En):** +1.3
- **IWSLT14 (Es-En):** +0.9
- **MTedx (All-En):** +2.5

ST

MT
Goodbye overturning!

<table>
<thead>
<tr>
<th>ID</th>
<th>Type</th>
<th>Model Name</th>
<th>Child System(s)</th>
<th>Dialect Transfer</th>
<th>test1 BLEU↑</th>
<th>test2 BLEU↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Cascade</td>
<td>ASR Mixing Cascade</td>
<td>A1, B1</td>
<td>✗</td>
<td>16.4</td>
<td>-</td>
</tr>
<tr>
<td>C2</td>
<td>Cascade</td>
<td>+ ASR Rover Comb.</td>
<td>A2, B1</td>
<td>✗</td>
<td>16.7</td>
<td>-</td>
</tr>
<tr>
<td>C3</td>
<td>Cascade</td>
<td>+ MT Posterior Comb.</td>
<td>A2, B2</td>
<td>✗</td>
<td>17.5</td>
<td>18.6</td>
</tr>
<tr>
<td>C4</td>
<td>Cascade</td>
<td>ASR Mixing Cascade</td>
<td>A3, B3</td>
<td>✓</td>
<td>17.3</td>
<td>-</td>
</tr>
<tr>
<td>C5</td>
<td>Cascade</td>
<td>+ ASR Rover Comb.</td>
<td>A4, B3</td>
<td>✓</td>
<td>17.4</td>
<td>-</td>
</tr>
<tr>
<td>C6</td>
<td>Cascade</td>
<td>+ MT Posterior Comb.</td>
<td>A4, B4</td>
<td>✓</td>
<td>17.9</td>
<td>19.4</td>
</tr>
<tr>
<td>D1</td>
<td>E2E ST</td>
<td>Hybrid Multi-Decoder</td>
<td>-</td>
<td>✗</td>
<td>17.7</td>
<td>-</td>
</tr>
<tr>
<td>D2</td>
<td>Mix</td>
<td>+ ROVER Intermediates</td>
<td>A2</td>
<td>✗</td>
<td>18.1</td>
<td>19.1</td>
</tr>
<tr>
<td>D3</td>
<td>Mix</td>
<td>+ ST/MT Posterior Comb.</td>
<td>A2, B5</td>
<td>✗</td>
<td>18.7</td>
<td>19.7</td>
</tr>
<tr>
<td>D4</td>
<td>E2E ST</td>
<td>Hybrid Multi-Decoder</td>
<td>-</td>
<td>✓</td>
<td>18.2</td>
<td>-</td>
</tr>
<tr>
<td>D5</td>
<td>Mix</td>
<td>+ ROVER Intermediates</td>
<td>A4</td>
<td>✓</td>
<td>18.3</td>
<td>19.5</td>
</tr>
<tr>
<td>D6</td>
<td>Mix</td>
<td>+ ST/MT Posterior Comb.</td>
<td>A4, B5</td>
<td>✓</td>
<td>18.9</td>
<td>19.8</td>
</tr>
<tr>
<td>E1</td>
<td>Mix</td>
<td>Min. Bayes-Risk Ensemble</td>
<td>C3, D3</td>
<td>✗</td>
<td>19.2</td>
<td>20.4</td>
</tr>
<tr>
<td>E2</td>
<td>Mix</td>
<td>Min. Bayes-Risk Ensemble</td>
<td>C6, D6</td>
<td>✓</td>
<td>19.5</td>
<td>20.8</td>
</tr>
</tbody>
</table>

Table 3: Results of our cascaded, E2E, and integrated cascaded/E2E systems as measured by BLEU score on the blind test2 and provided test1 sets. Dialect Transfer indicates the use of either MGB2 or OPUS data. Rover, posterior combinations, and minimum bayes-risk ensembling were applied to both cascaded and E2E systems, with Child System(s) indicating the inputs to the resultant systems combinations.

Our tuning efforts have high correlation with the blind test set (test2)
We got the nice result this time 😊

Table 6: Summary of select systems for Dialect Shared Task (BLEU on test2). We highlight the BLEU improvements ($\Delta$) obtained when training with additional MSA/English data compared with just the Tunisian/English (TA/EN) in the basic condition.

<table>
<thead>
<tr>
<th>Team / Condition / System</th>
<th>Architecture</th>
<th>Training Data</th>
<th>BLEU</th>
<th>$\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMU / basic / E1</td>
<td>Mix</td>
<td>TA/EN</td>
<td>20.4</td>
<td>-</td>
</tr>
<tr>
<td>CMU / dialect adapt / E2</td>
<td>Mix</td>
<td>TA/EN + MSA/EN</td>
<td>20.8</td>
<td>0.4</td>
</tr>
<tr>
<td>JHU / basic / primary</td>
<td>Cascaded</td>
<td>TA/EN</td>
<td>17.1</td>
<td>-</td>
</tr>
<tr>
<td>JHU / dialect adapt / primary</td>
<td>Cascaded</td>
<td>TA/EN + MSA/EN</td>
<td>18.9</td>
<td>1.8</td>
</tr>
<tr>
<td>ON-TRAC / basic / primary</td>
<td>End-to-End</td>
<td>TA/EN</td>
<td>12.4</td>
<td>-</td>
</tr>
<tr>
<td>ON-TRAC / unconstrained / post-eval</td>
<td>Cascaded</td>
<td>TA/EN + MSA/EN</td>
<td>14.4</td>
<td>2.0</td>
</tr>
</tbody>
</table>
Today’s Talk

- CMU’s IWSLT 2022 Dialect Speech Translation System
  - Part 1: Controlling ST output lengths via joint CTC/attention
  - Part 2: Controlling/explaining ST via searchable ASR intermediates

- Explainable E2E Speech Translation via Operation Sequence Generation
  - Part 3: Explaining ST via word-level ASR alignments
Better Speech Translation via Better Speech Recognition

Vanilla E2E (w/ ASR Multi-Task)

- ASR Enc
- ASR Dec
- ST Dec

Fully Cascaded

- MT Enc
- MT Dec
- ASR Enc
- ASR Dec

ASR and ST decodings are independent / parallel

Better transcription likely to yield better translation
Better Speech Translation via Better Speech Recognition

Vanilla E2E (w/ ASR Multi-Task)

- Transcript
- Translation

ASR Enc

ASR Dec

ST Dec

Better transcription likely to yield better translation

ASR and ST decodings are independent / parallel

Fully Cascaded

- Transcript
- Translation

ASR Enc

ASR Dec

ASR Enc

ASR Dec

MT Enc

MT Dec

Can we make an E2E differentiable cascade?

E2E Multi-Decoder (ASR Searchable Hidden Intermediates)

- Transcript
- Translation

ASR Enc

ASR Dec

MT Enc

MT Dec
ASR Decoder State

E2E ASR based on attention

- Transcript is obtained by the conditional likelihood

\[
\text{argmax}_W \ p(W|O) = \text{argmax}_W \ \prod_j p(w_j|h_j)
\]

- ASR decoder state

\[
h_j = \text{Decoder}(h_{j-1}, w_{j-1}, \text{Encoder}(O))
\]

  - ASR decoder state is **differentiable** (no argmax)
  - ASR decoder state is **searchable**
    - During inference, \( w_{j-1} \) is obtained by search or fusion (beam search with a language model etc.)
    - We can incorporate various information with the decoder state \( h_j \)
    - We call it **Searchable Intermediates**
  - Note that the ASR encoder state does not have this property
    - \( z_t = \text{Encoder}(O) \) does not have the token dependency
Multi-Decoder with Searchable Hidden Intermediates

**Searchable ASR Hidden Intermediates:**
During inference, ASR decoder representations are retrieved (e.g. via beam search) and passed to the subsequent ST Encoder.
Multi-Decoder with Searchable Hidden Intermediates

**Searchable ASR Hidden Intermediates:**
During inference, ASR decoder representations are retrieved (e.g. via beam search) and passed to the subsequent ST Encoder.

\[ L = \lambda_1 L^{ASR}_{CE} + \lambda_2 L^{ASR}_{CTC} + \lambda_3 L^{ST}_{CE} \]
Multi-Decoder with Searchable Hidden Intermediates

Searchable ASR Hidden Intermediates:
During inference, ASR decoder representations are retrieved (e.g. via beam search) and passed to the subsequent ST Encoder

Better ASR → Better ST
Multi-Decoder with Searchable Hidden Intermediates

Searchable ASR Hidden Intermediates:
During inference, ASR decoder representations are retrieved (e.g. via beam search) and passed to the subsequent ST Encoder.

We can guide ASR hidden intermediate retrieval using external models!

Strategy 1:
Beam search over ASR output w/ use of external models (e.g. LM, CTC)

Strategy 2:
Use of post-processing to improve ASR output (e.g. ROVER ensembling)
Multi-Decoder with Searchable Hidden Intermediates

We can guide ASR hidden intermediate retrieval using external models!

Strategy 1:
Beam search over ASR output w/ use of external models (e.g. LM, CTC)

Strategy 2:
Use of post-processing to improve ASR output (e.g. ROVER ensembling)
Today’s Talk

- CMU’s IWSLT 2022 Dialect Speech Translation System
  - Part 1: Controlling ST output lengths via joint CTC/attention
  - Part 2: Controlling/explaining ST via searchable ASR intermediates

- Explainable E2E Speech Translation via Operation Sequence Generation
  - Part 3: Explaining ST via word-level ASR alignments
Which one is more explainable?

E2E ST System A

E2E ST System B

I really need it

ich brauche das wirklich
Word-Level Speech Translation Explanations

Vanilla E2E (w/ ASR Multi-Task)

ASR Dec  ->  ST Dec

ASR Enc

I really need it  ->  Ich brauche das wirklich

ASR and ST decodings are independent / parallel

Explainable E2E

Dec

Enc

Can we simultaneously generate ASR/ST predictions + word-level alignments?
Align, Write, Re-order

Our goal is to get the speech translation result via this aligned information.
Part 3: Explaining ST via word-level ASR alignments

Align, Write, Re-Order: Explainable E2E Speech Translation via Operation Sequence Generation, Pre-print

Align, Write, Re-order

I really need it

Ich brauche das wirklich

(I, [Position A], Ich) (really, [Position B], wirklich) (need, [Position C], brauche) (it, [Position D], das)

Align to serialize
Align, Write, Re-order

(I, [Position A], Ich) (really, [Position B], wirklich) (need, [Position C], brauche) (it, [Position D], das)

Ich wirklich brauche das

Write the aligned (synchronized) translation result
Align, Write, Re-order

I really need it

Ich brauche das wirklich

Part 3: Explaining ST via word-level ASR alignments

Reorder to get the final translation result
Align, Write, Re-order

(I, [Position A], Ich) (really, [Position B], wirklich) (need, [Position C], brauche) (it, [Position D], das)

Ich wirklich brauche das

Ich *brauche* das wirklich

- Explainable
  - Easy to analyze
- Streamable
  - Get the translation result as soon as we get the ASR result
Operation Sequence Generation: Absolute Position

**Objective:** represent ASR/ST + word-alignment information as a single sequence

**Strategy 1:**
- Obtain word-level alignments on training data using statistical aligner (e.g. GIZA++)
- Define sequence of tuples of (source word, target word **absolute position**, target word)
- Insert target words into correct order in post-processing

(I, [0], Ich) (really, [3], wirklich) (need, [1], brauche) (it, [2], das)

**Absolute position operation sequence**
- By predicting target word absolute positions, we can generating translations **out-of-order**
  (while generating transcriptions **in-order**)
Operation Sequence Generation: Relative Shift

Strategy 2:
- Obtain word-level alignments on training data using statistical aligner (e.g. GIZA++)
- Define sequence of tuples of (source word, target word relative shift, target word)
- Insert target words into correct order in post-processing

Shifting Write-Head Operations
Based on prior MT work (Stahlberg et al., 2018)

[NO_OPS] - no operation
[SET_MARKER] - place write-head marker
[JMP_FWD] - jump right to next write-head marker
[JMP_BWD] - jump left to prev write-head marker
[NO_SRC] - no aligned source word
[NO_TGT] - no aligned target word
[EOP] - end of tuple (not displayed for space)

Special thanks to Prof. Graham Neubig
Operation Sequence Generation: Training and Inference

- **Training:**
  - After the data preparation, we *just throw it* to E2E ST Training (only data preparation)

- **Inference:**
  - *just throw it* to E2E ST beam search
  - Extract translation results with re-ordering

Demo: [https://i.imgur.com/9MT5NoH.gifv](https://i.imgur.com/9MT5NoH.gifv)
Operation Sequence Generation: Results

MuST-C En-De ST BLEU

- **Absolute Position:**
  Difficult to generalize; performance lags behind direct ST models

- **Relative Shift:**
  On-par with direct ST models → achieves explainability without sacrificing performance!
Takeaways

End-to-end systems do not have to be black-boxes:

- **Part 1**: CTC alignments stabilize the length problem of autoregressive decoders
- **Part 2**: External model correction of searchable ASR intermediates improves ST
- **Part 3**: Word-level explainability does not sacrifice translation quality

We are putting them to ESPnet

Let’s work together on Controllable and Explainable E2E Speech Translation!
Thank You!