How to get the model to do what we want:

Simultaneous Speech Translation

Dominik Macháček
11/4/2024
Speech Translation

Cascaded ST:

was the uh time

Direct ST

¿Qué hora es?
Simultaneous Speech Translation

Source Audio →

How
is
the
weather
today

Listen
Write
Simultaneous speech translation

- **Simultaneous** = Live = Real-Time = Low-latency = Incremental
  - Source available continuously, one **chunk** at a time
  - The **chunk** can be:
    - audio segment ... in the direct speech-to-text translation or transcription = **ASR**
    - or word (text) produced by incremental ASR ... in a cascaded system = **ASR + MT**
  - Provide the target “at the same time” as the source is being produced
    - = **simultaneously** = with a small additive delay

In the European Parliament:

-> English original source

-> English-to-Czech Sim. Interpreting

-> English-to-German Sim. Intp.
Simultaneous speech translation

- **Simultaneous** = Live = Real-Time = Low-latency = Incremental
  - Source available continuously, one **chunk** at a time
  - The **chunk** can be:
    - audio segment ... in the direct speech-to-text translation or transcription = ASR
    - or word (text) produced by incremental ASR ... in a cascaded system = ASR + MT
  - Provide the target “at the same time” as the source is being produced
    = simultaneously = with a small additive **delay**

In the European Parliament:

- -> English original source
- -> English-to-Czech Sim. Interpreting
- -> English-to-German Sim. Intp.
Challenges:

... tell me
Next week, police will accuse some of the people involved in the case.
Challenges: Most of parallel data are for translation.

Chinese source:

jingfang  xiazhou  jiang  dui  bu fen  she an  ren yuan  ti qi gong su
警方  下周  将  对  部分  涉案  人员  提起  公诉
police  next week  will  for  part  involved  people  accuse

English “offline” translation:  ... no problem with re-ordering

Police will accuse some of the people involved in the case next week.

But in simultaneous:  ... face the word order diff. + wait / or guess and risk being wrong

Next week, police will ...[long waiting]... accuse some of the people involved in the case.

Challenges:

- Word orders: wait or translate / re-translate with every new chunk
- Quality
- Latency
- Stability
- Model
  - + Training
  - + Data
  - + Decoding
- Practical
Simultaneous approaches
Re-Translation vs. Wait and append

- Re-translate from beginning of sentence each time: **rewrite + append**
- Latency vs stability. **Top quality.**

- Alternates between reading from ASR and translating: **no rewrites, only append**
- Latency vs. quality. **Top stability.**

<table>
<thead>
<tr>
<th>Source</th>
<th>Output</th>
<th>Erasure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Neue</td>
<td>New</td>
<td>-</td>
</tr>
<tr>
<td>2: Arzneimittel</td>
<td>New Medicines</td>
<td>0</td>
</tr>
<tr>
<td>3: könnten</td>
<td>New Medicines</td>
<td>0</td>
</tr>
<tr>
<td>4: Lungen-</td>
<td>New drugs may be lung</td>
<td>1</td>
</tr>
<tr>
<td>5: und</td>
<td>New drugs could be lung and</td>
<td>3</td>
</tr>
<tr>
<td>6: Eierstockkrebs</td>
<td>New drugs may be lung and ovarian cancer</td>
<td>4</td>
</tr>
<tr>
<td>7: verlangsamen</td>
<td>New drugs may slow lung and ovarian cancer</td>
<td>5</td>
</tr>
</tbody>
</table>

Content Delay: 1 4 6 7 7 7 7 7

Source: [Arivazhagan et al., 2020]

Source: [Ren et al., 2020]
Stability in Re-Translation
How to make re-translation more stable?

**Baseline** ("standard" offline MT)

<table>
<thead>
<tr>
<th>Source</th>
<th>Output</th>
<th>Erasure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Neue</td>
<td>New</td>
<td>-</td>
</tr>
<tr>
<td>2: Arzneimittel</td>
<td>New Medicines</td>
<td>0</td>
</tr>
<tr>
<td>3: könnten</td>
<td>New Medicines</td>
<td>0</td>
</tr>
<tr>
<td>4: Lungen-</td>
<td>New drugs may be lung</td>
<td>1</td>
</tr>
<tr>
<td>5: und</td>
<td>New drugs could be lung and</td>
<td>3</td>
</tr>
<tr>
<td>6: Eierstockkrebs</td>
<td>New drugs may be lung and ovarian cancer</td>
<td>4</td>
</tr>
<tr>
<td>7: verlangsamen</td>
<td>New drugs may slow lung and ovarian cancer</td>
<td>5</td>
</tr>
</tbody>
</table>

**Improvement**

<table>
<thead>
<tr>
<th>Source</th>
<th>New</th>
<th>Erasure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Neue</td>
<td>New drugs</td>
<td>0</td>
</tr>
<tr>
<td>2: Arzneimittel</td>
<td>New drugs may</td>
<td>0</td>
</tr>
<tr>
<td>3: könnten</td>
<td>New drugs may lung</td>
<td>0</td>
</tr>
<tr>
<td>4: Lungen-</td>
<td>New drugs may lung and</td>
<td>0</td>
</tr>
<tr>
<td>5: und</td>
<td>New drugs may lung and ovarian cancer</td>
<td>0</td>
</tr>
<tr>
<td>6: Eierstockkrebs</td>
<td>New drugs may slow lung and ovarian cancer</td>
<td>4</td>
</tr>
</tbody>
</table>

**Stability measure**: 13 erasures for 8 generated tokens = 1.625

4 erasures for 8 generated tokens = 0.5

Example: [Arivazhagan et al., 2020]. The method first proposed by …TBD.
How to make re-translation more stable?

### Baseline ("standard" offline MT)

<table>
<thead>
<tr>
<th>Source</th>
<th>Output</th>
<th>Erasure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Neue</td>
<td>New</td>
<td>-</td>
</tr>
<tr>
<td>2: Arzneimittel</td>
<td>New Medicines</td>
<td>0</td>
</tr>
<tr>
<td>3: können</td>
<td>New Medicines</td>
<td>0</td>
</tr>
<tr>
<td>4: Lungen-</td>
<td>New drugs may be lung</td>
<td>1</td>
</tr>
<tr>
<td>5: und</td>
<td>New drugs could be lung and</td>
<td>3</td>
</tr>
<tr>
<td>6: Eierstokkrebse</td>
<td>New drugs may be lung and ovarian cancer</td>
<td>4</td>
</tr>
<tr>
<td>7: verlangsamen</td>
<td>New drugs may slow lung and ovarian cancer</td>
<td>5</td>
</tr>
</tbody>
</table>

**Stability measure:** 13 erasures for 8 generated tokens = 1.625

### Improvement

<table>
<thead>
<tr>
<th>Source</th>
<th>Erasure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Neue</td>
<td>-</td>
</tr>
<tr>
<td>2: Arzneimittel</td>
<td>0</td>
</tr>
<tr>
<td>3: können</td>
<td>0</td>
</tr>
<tr>
<td>4: Lungen-</td>
<td>0</td>
</tr>
<tr>
<td>5: und</td>
<td>0</td>
</tr>
<tr>
<td>6: Eierstokkrebse</td>
<td>0</td>
</tr>
<tr>
<td>7: verlangsamen</td>
<td>5</td>
</tr>
</tbody>
</table>

4 erasures for 8 generated tokens = 0.5

**Learn this:** Proportional prefix training

1. Train a standard offline MT
2. Finetune on 1:1 mix of full sent. pairs and src-target prefixes
3. Create the prefixes from the length proportion,
4. do not care about the parallel words in the truncated suffix => anticipation
5. Measure the MT quality and erasures
6. Select a suitable trade-off

---

Example: [Arivazhagan et al., 2020]. The method first proposed by Niehues et al., 2018.
My results: learning curves

**BLEU vs. steps**

Colors are src-tgt variants, from the top: En->Cs, En+De->Cs (multi-sourcing), De->Cs

**Stability vs. steps**

De->Cs, En->Cs

Section 6.4.1 in Multi-Source Simultaneous Speech Translation, Macháček D., 2024, dissertation thesis
Select a checkpoint: quality-stability trade-off

Stability vs. MT quality (checkpoints for De-Cs)

E.g. this one

<table>
<thead>
<tr>
<th>checkpoint</th>
<th>En</th>
<th></th>
<th>De</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>starting</td>
<td>33.2</td>
<td>1.77</td>
<td>25.9</td>
<td>3.15</td>
</tr>
<tr>
<td>selected</td>
<td>33.0</td>
<td>1.21</td>
<td>25.0</td>
<td>1.52</td>
</tr>
<tr>
<td>diff</td>
<td>-0.2</td>
<td>-40%</td>
<td>-0.9</td>
<td>-52%</td>
</tr>
</tbody>
</table>

Table 6.13: The results of fine-tuning for stability, on ESIC dev. NE stands for "Normalized Erasure" (Arivazhagan et al., 2020b), measure of stability of re- translating simultaneous MT.

Results ... for En->Cs, -0.2 BLEU, 40% higher stability
... for De->Cs, -0.9 BLEU, 52% higher stability
Other simultaneous problems and solutions
Other SST problems and solutions (briefly)

- How to learn when to wait and translate?
Other SST problems and solutions (briefly)

- How to learn when to wait and translate?
  => RL agent (outdated),
  => or simultaneous streaming policies (in the next lessons)
• How to train simultaneous encoder-decoder effectively? On all the prefixes at once?
  => Encoder with monotonic look-back attention
Other SST problems and solutions (briefly)

- How to continue translation with the previous prefix?
Other SST problems and solutions (briefly)

- How to **continue translation** with the previous prefix?
  => **autoregressive decoding** can start with any tgt. prefix
Other SST problems and solutions (briefly)

- How to suggest the *target terminology*?
Other SST problems and solutions (briefly)

- How to suggest the **target terminology**?
  Prompts.
  OpenAI Whisper documentation:

```python
# baseline transcript with no prompt
transcribe(bbq_plans_filepath, prompt="")

"Hello, my name is Bristen Tuggle. I'm based in New York City. This weekend I have really exciting plans with some friends of mine. Amy and Sean. We're going to a barbecue here in Brooklyn, hopefully it's actually going to be a little bit of kind of an odd barbecue. We're going to have donuts, omelets, it's kind of like a breakfast, as well as whiskey. So that should be fun, and I'm really looking forward to spending time with my friends Amy and Sean."

While Whisper's transcription was accurate, it had to guess at various spellings. For example, it assumed the friends' names were spelled Amy and Sean rather than Aimee and Shawn. Let's see if we can steer the spelling with a prompt.

```python
# spelling prompt
transcribe(bbq_plans_filepath, prompt="Friends: Aimee, Shawn")

"Hello, my name is Bristen Tuggle. I'm based in New York City. This weekend I have really exciting plans with some friends of mine. **Aimee and Shawn**. We're going to a barbecue here in Brooklyn. Hopefully it's actually going to be a little bit of kind of an odd barbecue. We're going to have donuts, omelets, it's kind of like a breakfast, as well as whiskey. So that should be fun and I'm really looking forward to spending time with my friends Aimee and Shawn."

Success!
```
Other SST problems and solutions (briefly)

- How to learn when to wait and translate?
  => RL (outdated), or simultaneous streaming policies (some other time)

- How to train simultaneous encoder-decoder effectively?
  => monotonic look-back attention in the encoder

- How to continue translation with the previous prefix?
  => autoregressive decoding can start with any tgt. prefix

- MT gives too long targets, the users in real-time need shorter synonyms.
  => filter a parallel corpus for the shorter src-tgt pairs, train on them

- MT is verbose and literal, but too complicated to perceive
  => style transfer, learn e.g. on the simultaneous interpreting data
  ... or synthesize them

- How to suggest the target terminology? Whisper model with prompting.

- Some other time: speech-to-text tutorial, interactive demo,
  Discussion in 99 languages
Some other time

- Live interactive **demo** – ELITR, Whisper-Streaming
  Live speech src. in 99 languages, translation into 43 langs.
- Speech-to-text models (= like LLMs with speech input)
- Simultaneous streaming policies
Summary

- You learned what is the Simultaneous Speech Translation
- What are its challenges
- You learned how to stabilize re-translation: fine-tune on prefixes

See demo next time!