Breakthrough in Simultaneous Translation

full-sentence (non-simultaneous) translation

simultaneous translation, latency ~3 secs

Baidu World Conference, November 2017

Baidu World Conference, November 2018
Background: Consecutive vs. Simultaneous

**Consecutive interpretation**
- *multiplicative latency* (x2)

**Simultaneous interpretation**
- *additive latency* (+3 secs)
Background: Consecutive vs. Simultaneous

Consecutive interpretation

Multiplicative latency ($x2$)

Simultaneous interpretation

Additive latency (+3 secs)

Simultaneous interpretation is extremely difficult

Only ~3,000 qualified simultaneous interpreters world-wide

Each interpreter can only sustain for at most 10-30 minutes

The best interpreters can only cover ~60% of the source material
Tradeoff between Latency and Quality

- High quality
- Low quality

- Low latency: ~3 seconds, 1 sentence, high latency

- Word-by-word translation
- Simultaneous interpreters
- Machine translation

- Our goal: consecutive interpreters
Industrial Work in Simultaneous Translation

- almost all existing “real-time” translation systems use conventional full-sentence translation techniques, causing at least one-sentence delay
- some systems repeatedly retranslate, but constantly changing translations is annoying to the user and can’t be used for speech-to-speech translation

Baidu, Nov. 2017 (~12 seconds delay)  
Sougou, Oct. 2018 (~12 seconds delay)
Academic Work in Simultaneous Translation

- prediction of German verb (Grissom et al, 2014)
- reinforcement learning (Grissom et al, 2014; Gu et al, 2017)
- learning Read/Write sequences on top of a pretrained NMT model
- “encourages” latency requirements, but can’t force them in testing
- complicated, and slow to train

Grissom et al, 2014
Challenge: Word Order Difference

- e.g. translate from SOV language (Japanese, German) to SVO (English)
- German is underlyingly SOV, and Chinese is a mix of SVO and SOV
- human simultaneous interpreters routinely “anticipate” (e.g., predicting German verb)

Grissom et al, 2014

ich bin mit dem Zug nach Ulm gefahren
I am with the train to Ulm traveled
I (. . . . . waiting . . . . . ) traveled by train to Ulm
Challenge: Word Order Difference

- e.g. translate from SOV language (Japanese, German) to SVO (English)
- German is underlyingly SOV, and Chinese is a mix of SVO and SOV
- human simultaneous interpreters routinely “anticipate” (e.g., predicting German verb)

Grissom et al, 2014

President Bush meets with Russian President Putin in Moscow
Challenge: Word Order Difference

- e.g. translate from SOV language (Japanese, German) to SVO (English)
- German is underlyingly SOV, and Chinese is a mix of SVO and SOV
- human simultaneous interpreters routinely “anticipate” (e.g., predicting German verb)

\[
\begin{align*}
\text{ich bin mit dem Zug nach Ulm } & \text{ gefahren} \\
\text{I am with the train to Ulm } & \text{ traveled} \\
\text{I (…… waiting……) traveled by train to Ulm} 
\end{align*}
\]

Grissom et al, 2014

President Bush \text{ meets} with Russian \text{ President Putin} \text{ in Moscow}

\text{non-anticipative: President Bush (…… waiting …….)} \text{ meets with Russian …
Challenge: Word Order Difference

- e.g. translate from SOV language (Japanese, German) to SVO (English)
- German is underlyingly SOV, and Chinese is a mix of SVO and SOV
- human simultaneous interpreters routinely “anticipate” (e.g., predicting German verb)

Grissom et al., 2014

non-anticipative: President Bush (…… waiting ……) meets with Russian …

anticipative: President Bush meets with Russian President Putin in Moscow
Our Solution: Prefix-to-Prefix

- seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix, tailed to simultaneous MT
- special case: wait-\(k\) policy: translation is always \(k\) words behind source sentence
- training in this way enables anticipation
Our Solution: Prefix-to-Prefix

- seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix, tailed to simultaneous MT
- special case: wait-$k$ policy: translation is always $k$ words behind source sentence
- training in this way enables anticipation
Our Solution: Prefix-to-Prefix

- seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix, tailed to simultaneous MT
- special case: wait-k policy: translation is always $k$ words behind source sentence
- training in this way enables anticipation

- seq-to-seq

- prefix-to-prefix (wait-k)

**Example:**

- Source: Bush
- Target: in
- Translation: President Bush

- Source: Bùshí zòngtōng zài
- Target:布什 总统 在
- Translation: President in
Our Solution: Prefix-to-Prefix

- seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix, tailed to simultaneous MT
- special case: wait-k policy: translation is always k words behind source sentence
- training in this way enables anticipation
Our Solution: Prefix-to-Prefix

- seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix, tailed to simultaneous MT
- special case: wait-k policy: translation is always k words behind source sentence
- training in this way enables anticipation
Our Solution: Prefix-to-Prefix

- seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix, tailed to simultaneous MT
- special case: wait-k policy: translation is always k words behind source sentence
- training in this way enables anticipation

source: \[1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5\]

- seq-to-seq

target: \(\ldots \text{wait whole source sentence} \ldots\)

source: \[1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5\]

- prefix-to-prefix (wait-k)

target: \(\text{wait } k \text{ words}\)

- special case: wait-k policy: translation is always k words behind source sentence

- training in this way enables anticipation

President Bush meets with Russian
Our Solution: Prefix-to-Prefix

- seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix, tailed to simultaneous MT
- special case: wait-k policy: translation is always k words behind source sentence
- training in this way enables anticipation

\[
\begin{align*}
\text{source:} & \quad 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \\
\text{target:} & \quad \ldots \text{wait whole source sentence} \quad \ldots \\
\end{align*}
\]

\[
\begin{align*}
\text{source:} & \quad 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \\
\text{target:} & \quad \text{wait } k \text{ words} \\
\end{align*}
\]

President Bush meets with Russian President
Our Solution: Prefix-to-Prefix

- seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix, tailed to simultaneous MT
- special case: wait-k policy: translation is always k words behind source sentence
- training in this way enables anticipation

President Bush meets with Russian President Putin
Our Solution: Prefix-to-Prefix

- seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix, tailed to simultaneous MT
- special case: wait-k policy: translation is always k words behind source sentence
- training in this way enables anticipation

**President** Bush meets with Russian President Putin in Moscow
More General Prefix-to-Prefix

- seq-to-seq (given full source sent)
  \[ p(y_t \mid x_1 \ldots x_n, y_1 \ldots y_{t-1}) \]

- prefix-to-prefix (given source prefix)
  \[ p(y_t \mid x_1 \ldots x_{g(t)}, y_1 \ldots y_{t-1}) \]

\( g(\cdot) \) is a monotonic non-decreasing function

\( g(t) \): num. of source words used to predict \( y_t \)

<table>
<thead>
<tr>
<th></th>
<th>President</th>
<th>Bush</th>
<th>meets</th>
<th>with</th>
<th>Putin</th>
<th>in</th>
<th>Moscow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bush</td>
<td>布什</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pres.</td>
<td>总统</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>at</td>
<td>在</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moscow</td>
<td>莫斯科</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with</td>
<td>与</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Putin</td>
<td>普京</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>meet</td>
<td>会晤</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This is just our research demo. Our production system is better (shorter ASR latency).
### Demo 2 (Latency-Accuracy Tradeoff)

<table>
<thead>
<tr>
<th>Chinese input:</th>
<th>Word-by-Word Translation:</th>
</tr>
</thead>
<tbody>
<tr>
<td>江泽民对美国总统的发言表示遗憾。</td>
<td>jiang zemin to united states president of speak express regret 。</td>
</tr>
</tbody>
</table>

**Simultaneous Translation (wait 3):**
- jiang zemin expressed his welcome to the us president 's remarks .

**Simultaneous Translation (wait 5):**
- jiang zemin expressed his regret over the us president 's remarks .

**Baseline Translation (greedy):**
- jiang zemin expressed regret over the us president 's remarks .

**Baseline Translation (beam 5):**
- jiang zemin expressed regret over the us president 's remarks .
### Demo 2 (Latency-Accuracy Tradeoff)

<table>
<thead>
<tr>
<th>Chinese input:</th>
<th>Pinyin:</th>
<th>Word-by-Word Translation:</th>
</tr>
</thead>
<tbody>
<tr>
<td>江泽民对美国总统的发言表示遗憾。</td>
<td>jiāng zé mín dìu měiguó zǒngtǒng de fāyán biǎoshì yíhàn 。</td>
<td>jiang zemin to united states president of speak express regret 。</td>
</tr>
</tbody>
</table>

**Simultaneous Translation (wait 3):**
- jiang zemin expressed his welcome to the us president 's remarks.

**Simultaneous Translation (wait 5):**
- jiang zemin expressed his regret over the us president 's remarks.

**Baseline Translation (greedy):**
- jiang zemin expressed regret over the us president 's remarks.

**Baseline Translation (beam 5):**
- jiang zemin expressed regret over the us president 's remarks.
Demo 3 (Deployment)

This is live recording from the Baidu World Conference on Nov 1, 2018.
German source:
doch während man sich im kongress *nicht* auf ein *vorgehen* einigen *kann*, *warten* mehrere bundesstaaten *nicht* länger.

English translation (simultaneous wait 3 — training not converged yet):
but, while congress *does not* agree on a *course of action*, several states *no* longer *wait*.

English translation (full-sentence beam search):
but, while congressional *action* *can not* be *agreed*, several states are *no* longer *waiting*. 
Refinements: Wait-\(k\) with Catchup

- English translation length is often \(~1.25\times\) of the Chinese input length
- in a more or less “synchronized” policy like wait-\(k\), the English translation will be lagging behind more and more severely
- catchup: decode two English words in 1 out of 4 steps
New Latency Metric: Average Lagging

- previous latency metrics: CW (consecutive wait) and AP (average proportion)
- they’re good metrics but do not directly measure the level of “lagging behind”
- our metric, Average Lagging (AL), measures on average how many (source) words is the translation lagging behind; ideally, $AL$ (wait-$k$ with catchup) $\approx k$
Experiments: German <=> English

- trained on 4.5M sentence pairs (WMT 15); comparing with Gu et al 2017
Experiments: German<=>English

- trained on 4.5M sentence pairs (WMT 15); comparing with Gu et al 2017
Experiments: German<=>English

- trained on 4.5M sentence pairs (WMT 15); comparing with Gu et al 2017
Experiments: Chinese<=>English

- trained on 2M sentence pairs; evaluated on NIST 06 / 08; 1-ref and 4-ref BLEU
### Chinese=>English Examples From Recent News

<table>
<thead>
<tr>
<th>(a)</th>
<th>Mēiguó</th>
<th>dāngjú</th>
<th>duì</th>
<th>Shàtè</th>
<th>jīzhè</th>
<th>shīzōng</th>
<th>yī</th>
<th>àn</th>
<th>gǎndào</th>
<th>dānyōu</th>
</tr>
</thead>
<tbody>
<tr>
<td>US authorities</td>
<td>to</td>
<td>Saudi reporter</td>
<td>missing</td>
<td>a</td>
<td>case</td>
<td>feel</td>
<td>concern</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| \(k=3\) | the | US authorities | are | very | concerned | about | the | the | Saudi reporter’s missing case |
| \(k=3\) | the | US authorities | are | very | concerned | about | the | Saudi reporter’s missing case |

| \(k=\infty\) | US authorities concerned over Saudi journalists missing |
### Chinese=>English Examples From Recent News

#### (a)

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Měiguó</td>
<td>dāngjú</td>
<td>duì</td>
<td>Shātè</td>
<td>jīzhē</td>
<td>shīzōng</td>
<td>yī</td>
<td>àn</td>
<td>gǎndào</td>
<td>dānyōu</td>
</tr>
<tr>
<td>US authorities</td>
<td>to</td>
<td>Saudi reporter</td>
<td>missing</td>
<td>a</td>
<td>case</td>
<td>feel</td>
<td>concern</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$k=3$

$k=3$

the us authorities are very concerned about the the saudi reporter ’s missing case

$k=\infty$

us authorities concerned over saudi journalists missing

#### (b)

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Měiguó</td>
<td>dāngjú</td>
<td>duì</td>
<td>Shātè</td>
<td>jīzhē</td>
<td>shīzōng</td>
<td>yī</td>
<td>àn</td>
<td>gǎndào</td>
<td>dānyōu</td>
</tr>
<tr>
<td>US authorities</td>
<td>to</td>
<td>Saudi reporter</td>
<td>missing</td>
<td>a</td>
<td>case</td>
<td>feel</td>
<td>concern</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$k=3$

$k=3$

the us authorities are very concerned about the the saudi reporter ’s missing case

$k=\infty$

us authorities dis- satisfied with saudi reporters ’ missing case
This is another new development that has made foreign technology media so excited since the release of Baidu Deep Speech 2 in 2016.

— QbitAI (量子位)
Conclusions

• first simultaneous translation system with seamlessly integrated anticipation
  • human simultaneous interpreters also anticipate all the time
  • some previous works predict source language verbs
  • we don’t have a separate “anticipation” step, and only predict target side words
• first simultaneous translation system with arbitrary controllable latency
  • some previous works use reinforcement learning with latency as part of the reward, but can’t impose a hard constraint on latency at test time
• very easy to train and scalable — minor changes to any neural MT codebase
Conclusions

• first simultaneous translation system with seamlessly integrated anticipation
• human simultaneous interpreters also anticipate all the time
• some previous works predict source language verbs
• we don’t have a separate “anticipation” step, and only predict target side words

• first simultaneous translation system with arbitrary controllable latency
• some previous works use reinforcement learning with latency as part of the reward, but can’t impose a hard constraint on latency at test time

• very easy to train and scalable — minor changes to any neural MT codebase
Thank you very much for listening to my speech
Side Project: Translation with Noisy Input from ASR

- neural MT is fragile, and automatic speech recognition output is noisy
- Hairong Liu’s work (on arXiv): Robust Neural MT using phonetic information

<table>
<thead>
<tr>
<th>Clean Input</th>
<th>目前已发现有109人死亡, 另有57人获救</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output of Transformer</td>
<td>at present, 109 people have been found dead and 57 have been rescued</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Noisy Input</th>
<th>目前已发现又109人死亡, 另有57人获救</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output of Transformer</td>
<td>the hpv has been found dead so far and 57 have been saved</td>
</tr>
<tr>
<td>Output of Our Method</td>
<td>so far, 109 people have been found dead and 57 others have been rescued</td>
</tr>
</tbody>
</table>

Table 1: The translation results on Mandarin sentences without and with homophone noises. The word “有” (yǒu, “have”) in clean input is replaced by one of its homophone, “又” (yòu, “again”), to form a noisy input. This seemingly minor change completely fools the Transformer to generate something irrelevant (“hpv”). Our method, by contrast, is very robust to homophone noises thanks to phonetic information.