Simultaneous Translation: Recent Advances and Remaining Challenges

Liang Huang
Baidu Research (USA) and Oregon State University
Consecutive vs. Simultaneous Interpretation

consecutive interpretation
- multiplicative latency (x2)

simultaneous interpretation
- additive latency (+3 secs)
Consecutive vs. Simultaneous Interpretation

Consecutive interpretation

*multiplicative latency (x2)*

Simultaneous interpretation

*additive latency (+3 secs)*

Simultaneous interpretation is extremely difficult

- only ~3,000 qualified simultaneous interpreters world-wide (AIIC)
- each interpreter can only sustain for at most 15-20 minutes
- the best interpreters can only cover ~60% of the source material
Simultaneous Interpreters: Strategies & Limitations

- anticipation, summarization, generalization, etc…
- and they inevitably make (quite a bit of) mistakes
- “human-level” quality: much lower than normal translation
- “human-level” latency: very short: 2~4 secs  (actually higher latency hurts quality…)

from United Nations Proceedings Speech Corpus (LDC2014S08, Chay et al, 2014)
Simultaneous Interpreters: Strategies & Limitations

- anticipation, summarization, generalization, etc…
- and they inevitably make (quite a bit of) mistakes
- “human-level” quality: much lower than normal translation
- “human-level” latency: very short: 2~4 secs (actually higher latency hurts quality…)

from United Nations Proceedings Speech Corpus (LDC2014S08, Chay et al, 2014)
Simultaneous Interpreters: Strategies & Limitations

- anticipation, summarization, generalization, etc…
- and they inevitably make (quite a bit of) mistakes
- “human-level” quality: much lower than normal translation
- “human-level” latency: very short: 2~4 secs (actually higher latency hurts quality…)

from United Nations Proceedings Speech Corpus (LDC2014S08, Chay et al, 2014)
Simultaneous Interpreters: Strategies & Limitations

- anticipation, summarization, generalization, etc…
- and they inevitably make (quite a bit of) mistakes
- “human-level” quality: much lower than normal translation
- “human-level” latency: very short: 2~4 secs (actually higher latency hurts quality…)

From United Nations Proceedings Speech Corpus (LDC2014S08, Chay et al, 2014)
Tradeoff between Latency and Quality

- high latency
  - word-by-word translation
  - consecutive interpretation
  - full-sentence machine translation
- low latency
  - ≤3 seconds
  - 1 sentence
- low quality
- high quality

Levels of quality and latency are inversely related:
- High quality with high latency
- Low quality with low latency
Tradeoff between Latency and Quality

- High quality
- Low quality
- Low latency: ~3 seconds
- High latency: 1 sentence

---

High Latency

- Word-by-word translation
- Simultaneous interpretation
- One of AI’s holy grails needs fundamentally new ideas!

Consecutive Interpretation

- Full-sentence machine translation
- Seq-to-seq is already very good

Previous work in simultaneous translation

Written translation

---

Consecutive Interpretation

- 1 sentence
- Simultaneous Interpretation
- Word-by-word translation
- One of AI’s holy grails needs fundamentally new ideas!
Tradeoff between Latency and Quality

- **High Quality**
  - One of AI's holy grails
  - Needs fundamentally new ideas!

- **Low Quality**
  - Word-by-word translation

- **Low Latency**
  - ~3 seconds

- **High Latency**
  - Simultaneous interpretation

- **Simultaneous Work in Simultaneous Translation**
  - Previous work in simultaneous translation
  - Full-sentence machine translation
  - Consecutive interpretation

- **Incremental Text-to-Speech**
  - Target text stream

- **Streaming Speech Recognition**
  - Source speech stream

- **Simultaneous Text-to-Text Translation**
  - Source text stream

- **Target Speech Stream**

---

President Bush ... (foundational work & Aims 1-3)

(Aims 4)
Outline

- Background on Simultaneous Interpretation

- Part I: Our Breakthrough in 2018
  - Prefix-to-Prefix Framework, Integrated Anticipation, Controllable Latency
  - New Latency Metric
  - Demos and Examples

- Part II: Towards Flexible (Adaptive) Translation Policies

- Part III: Remaining Challenges
Our Breakthrough in 2018

Baidu World Conference, Nov. 2017
full-sentence translation (latency: 10+ secs)

Baidu World Conference, Nov. 2018
low-latency simultaneous translation (latency: ~3 secs)

our work

Media coverage:

IEEE SPECTRUM, MIT Technology Review, CNBC, VentureBeat, siliconANGLE, Synced AI Technology & Industry Review, South China Morning Post, engadget, FORTUNE, The Register, Packt, lowyat.net, RED PULSE, FLIPBOARD, China Knowledge
Our Breakthrough in 2018

Baidu World Conference, Nov. 2017
full-sentence translation (latency: 10+ secs)

Baidu World Conference, Nov. 2018
low-latency simultaneous translation (latency: ~3 secs)

Media coverage:
Our Breakthrough in 2018

Baidu World Conference, Nov. 2017
full-sentence translation (latency: 10+ secs)

Baidu World Conference, Nov. 2018
low-latency simultaneous translation (latency: ~3 secs)

our work

Media coverage:
Our Breakthrough in 2018

Baidu World Conference, Nov. 2017
full-sentence translation (latency: 10+ secs)

Baidu World Conference, Nov. 2018
low-latency simultaneous translation (latency: ~3 secs)
Our Breakthrough in 2018

Baidu World Conference, Nov. 2017
full-sentence translation (latency: 10+ secs)

Baidu World Conference, Nov. 2018
low-latency simultaneous translation (latency: ~3 secs)

our work

Media coverage:
Our Breakthrough in 2018

**Baidu World Conference, Nov. 2017**
full-sentence translation (latency: 10+ secs)

**Baidu World Conference, Nov. 2018**
low-latency simultaneous translation (latency: ~3 secs)

---

*Our work*

---

*Request*

---

Haifeng Wang, Zhongjun He, Hao Xiong

Mingbo Ma, Kaibo Liu, Renjie Zheng
Our Breakthrough in 2018

Baidu World Conference, Nov. 2017
full-sentence translation (latency: 10+ secs)

Baidu World Conference, Nov. 2018
low-latency simultaneous translation (latency: ~3 secs)

Ken Church
I really need low-latency simultaneous translation!

Haifeng Wang
Zhongjun He
Hao Xiong
Ken Church
Mingbo Ma
Kaibo Liu
Renjie Zheng
Main Challenge: Word Order Difference

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

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

Grissom et al, 2014
Main Challenge: Word Order Difference

- e.g. translate from Subj-Obj-Verb (Japanese, German) to Subj-Verb-Obj (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
Main Challenge: Word Order Difference

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

\[
\frac{\text{ich bin mit dem Zug nach Ulm gefahren}}{\text{I am with the train to Ulm traveled}}
\]

\[
\frac{\text{I (...... waiting......)}}{\text{traveled by train to Ulm}}
\]

Grissom et al, 2014

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

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

- e.g. translate from Subj-Obj-Verb (Japanese, German) to Subj-Verb-Obj (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
Previous Solutions

• industrial systems
  • almost all “real-time” translation systems use full-sentence translation
  • some systems “repeatedly retranslate”, but constantly changing translations is annoying to the users and can’t be used for speech-to-speech translation

• academic papers (just to sample a few)
  • explicit prediction of German verbs (Grissom et al, 2014)
  • reinforcement learning (Gu et al, 2017) to decide READ or WRITE
  • segment-based (Bangalore et al, 2012; Fujita et al, 2013; Oda et al, 2014)
  • these efforts (a) use full-sentence translation model; (b) can’t ensure a given latency
Our Idea: Prefix-to-Prefix, not Seq-to-Seq

- standard seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix framework tailed to tasks with simultaneity
- special case: wait-k policy: translation is always k words behind source sentence
- decoding this way => controllable latency
- training this way => implicit anticipation on the target-side

\[
p(y_i | x_1 \ldots x_n, y_1 \ldots y_{i-1})
\]

seq-to-seq

source: 1 2 3 4 5

\[
p(y_i | x_1 \ldots x_{i+k-1}, y_1 \ldots y_{i-1})
\]

prefix-to-prefix (wait-k)

\[
\begin{align*}
1 & \rightarrow 2 \\
3 & \rightarrow 4 \\
5 & \rightarrow \ldots
\end{align*}
\]

source: 1 2 3 4 5

target: \ldots wait whole source sentence \ldots

source: 1 2 3 4 5

target: wait k words

\[
\begin{align*}
1 & \rightarrow 2 \\
3 & \rightarrow 4 \\
5 & \rightarrow \ldots
\end{align*}
\]
Our Idea: Prefix-to-Prefix, not Seq-to-Seq

- standard seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix framework tailed to tasks with simultaneity
- special case: wait-$k$ policy: translation is always $k$ words behind source sentence
- decoding this way => controllable latency
- training this way => implicit anticipation on the target-side

President Bush meets布什总统在莫斯科

$\text{wait } 2$ President Bush meets

deleting, $p(y_i | x_1 \ldots x_n, y_1 \ldots y_{i-1})$

denoting, $p(y_i | x_1 \ldots x_{i+k-1}, y_1 \ldots y_{i-1})$
Our Idea: Prefix-to-Prefix, not Seq-to-Seq

- standard seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix framework tailed to tasks with simultaneity
- special case: wait-k policy: translation is always k words behind source sentence
- decoding this way => controllable latency
- training this way => implicit anticipation on the target-side

```
p(y_i | x_1 \ldots x_n, y_1 \ldots y_{i-1})

source: 1 2 3 4 5
seq-to-seq

source: 1 2 3 4 5
prefix-to-prefix (wait-k)
```

```
Bùshí zǒngtǒng zài Mósīkè yǔ Éluōsī zǒngtǒng Pǔjīng hùiwù

布什 总统 在 莫斯科 与 俄罗斯 总统 普京 会晤

Bush President in Moscow with Russian President Putin meet

wait 2 President Bush meets with Russian President Putin in Moscow
More General Prefix-to-Prefix

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

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

\( g(\cdot) \) is a monotonic non-decreasing function
\( g(t) \): num. of source words used to predict \( y_t \)

This general framework can be used for other tasks such as incremental parsing and incremental text-to-speech.
This is just our research demo. Our production system is better (shorter ASR latency).
Research Demo

江泽民对法国总统的来华
jiang zemin expressed his appreciation

This is just our research demo. Our production system is better (shorter ASR latency).
This is just our research demo. Our production system is better (shorter ASR latency).

江泽民对法国总统的来华访问表示感谢。

jiang zemin expressed his appreciation for the visit by french president.
Research Demo

This is just our research demo. Our production system is better (shorter ASR latency).

江泽民对法国总统的来华
jiang zemin expressed his appreciation

江泽民对法国总统的来华访问表示感谢。
jiang zemin to French President’s to-China visit express gratitude

we support uh... Bolivia envoy & Russia envoy just-now made position

We support the position of Bolivia & Russia
Latency-Accuracy Tradeoff

Chinese input:

Pinyin:

Word-by-Word Translation:

Simultaneous Translation (wait 3):

Simultaneous Translation (wait 5):

Baseline Translation (greedy):

Baseline Translation (beam 5):
Latency-Accuracy Tradeoff

- Chinese input:
- Pinyin:
- Word-by-Word Translation:
- Simultaneous Translation (wait 3):
- Simultaneous Translation (wait 5):
- Baseline Translation (greedy):
- Baseline Translation (beam 5):
Deployment Demo

This is live recording from the Baidu World Conference on Nov 1, 2018.
Deployment Demo

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.

but while they self in congress not on one action agree can wait several states not longer

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

English translation (full-sentence baseline):
but, while congressional action can not be agreed, several states are no longer waiting.
New Latency Metric: Average Lagging

- previous metrics: CW (consecutive wait) and AP (average proportion)
- they do not directly measure the level of “lagging behind” (Gu et al ’17; Cho & Esipova ’16)
- our metric, *Average Lagging* (*AL*), measures on average how many source words the translation lags behind the source speech; ideally, $AL(\text{wait-k}) \approx k$
- closely related to “ear-voice span” (EVS) in the interpretation literature

<table>
<thead>
<tr>
<th>布什</th>
<th>总统</th>
<th>在</th>
<th>莫斯科</th>
<th>与</th>
<th>普京</th>
<th>会晤</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pres.</td>
<td>Bush</td>
<td>meets</td>
<td>with</td>
<td>Putin</td>
<td>in</td>
<td>Moscow</td>
</tr>
</tbody>
</table>

We support uh... Bolivia envoy & Russia envoy just-now made position of Bolivia & Russia

latency of “Bolivia” (+)
l latency of “position” (-)

latency

read

write
Experiments (de↔en & zh↔en)

RL: our adaptation of Gu et al (2017) on the same Transformer codebase, trained with CW=2, 5, 8.
Summary of Innovations in 2018

- prefix-to-prefix framework tailed to simultaneity (incremental on both sides)
- first genuinely simultaneous translation model (rather than full-sentence model)
- decoding like this => controllable latency
- training like this => implicit anticipation on the target side
- very easy to train and scalable — minor changes to most neural MT codebase
- prefix-to-prefix is very general; can be used in other tasks with simultaneity
- a new latency metric (AL) that resembles “ear-voice span” in interpretation
Part II: Towards Adaptive Translation Policies
### Part II: Towards Adaptive Translation Policies

<table>
<thead>
<tr>
<th></th>
<th>fixed-latency policies</th>
<th>adaptive policies</th>
</tr>
</thead>
<tbody>
<tr>
<td>full-sentence MT model</td>
<td>Dalvi et al. (2018); test-time wait-(k)</td>
<td>Grissom et al. (2014); Cho &amp; Esipova (2016); Satija &amp; Pineau (2016); Gu et al. (2017); Alinejad et al (2018); ...</td>
</tr>
<tr>
<td>simultaneous MT model</td>
<td>wait-(k) (Ma et al. 2018)</td>
<td>Arivazhagan et al. (ACL 2019)</td>
</tr>
<tr>
<td>(our invention)</td>
<td></td>
<td>Zheng et al. (ACL 2019)</td>
</tr>
</tbody>
</table>
Limitations of Fixed-Latency (wait-\(k\)) Policy

• can be too aggressive (anticipation errors) with small \(k\) (too fast)
Limitations of Fixed-Latency (wait-$k$) Policy

- can be too aggressive (anticipation errors) with small $k$ (too fast)

<table>
<thead>
<tr>
<th>input</th>
<th>wǒ</th>
<th>shàng</th>
<th>wèi</th>
<th>dédào</th>
<th>yǒuguàn</th>
<th>bùmén</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>我</td>
<td>尚</td>
<td>未</td>
<td>得到</td>
<td>有关</td>
<td>部门</td>
</tr>
<tr>
<td>wait-1</td>
<td>I</td>
<td>yet</td>
<td>not</td>
<td>receive</td>
<td>relevant</td>
<td>department</td>
</tr>
</tbody>
</table>

(wait-1 I have not received relevant (AL=1.4))
Limitations of Fixed-Latency (wait-$k$) Policy

- can be too aggressive (anticipation errors) with small $k$ (too fast)

<table>
<thead>
<tr>
<th>input</th>
<th>wǒ</th>
<th>shàng</th>
<th>wèi</th>
<th>dédào</th>
<th>yōuguan</th>
<th>bùmén</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>yet</td>
<td>not</td>
<td>receive</td>
<td>relevant</td>
<td>department</td>
</tr>
</tbody>
</table>

wait-1  I have not received relevant documents
(AL=1.4)
Limitations of Fixed-Latency (wait-\(k\)) Policy

- can be too aggressive (\textit{anticipation errors}) with small \(k\) (too fast)

\begin{itemize}
  \item I have not received relevant documents from relevant departments (\(\text{AL}=1.4\))
\end{itemize}
Limitations of Fixed-Latency (wait-$k$) Policy

- can be too aggressive (anticipation errors) with small $k$ (too fast)
- can also be too conservative with large $k$ (too slow)

<table>
<thead>
<tr>
<th>input</th>
<th>wǒ</th>
<th>shàng</th>
<th>wèi</th>
<th>dédào</th>
<th>yǒuɡuān</th>
<th>bùmén</th>
<th>de</th>
<th>huìyìnɡ</th>
<th>response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>我</td>
<td>尚</td>
<td>未</td>
<td>得到</td>
<td>有关</td>
<td>部门</td>
<td>的</td>
<td>回应</td>
<td>的</td>
</tr>
<tr>
<td>wait-1</td>
<td>I</td>
<td>not</td>
<td>receive</td>
<td>relevant</td>
<td>department</td>
<td>'s</td>
<td></td>
<td>response</td>
<td></td>
</tr>
<tr>
<td>(AL=1.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wait-4</td>
<td>I</td>
<td>have</td>
<td>not</td>
<td>received</td>
<td>relevant</td>
<td>documents</td>
<td>from</td>
<td>relevant departments</td>
<td></td>
</tr>
<tr>
<td>(AL=4.0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Limitations of Fixed-Latency (wait-k) Policy

- can be too aggressive (anticipation errors) with small $k$ (too fast)
- can also be too conservative with large $k$ (too slow)

**input**

<table>
<thead>
<tr>
<th>wait-1</th>
<th>wait-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>yet</td>
<td>have</td>
</tr>
<tr>
<td>not</td>
<td>not</td>
</tr>
<tr>
<td>receive</td>
<td>received</td>
</tr>
<tr>
<td>relevant</td>
<td>relevant</td>
</tr>
<tr>
<td>department</td>
<td>documents</td>
</tr>
<tr>
<td>from</td>
<td>from</td>
</tr>
<tr>
<td>relevant departments</td>
<td>relevant departments</td>
</tr>
</tbody>
</table>

**wait-1** *(AL=1.4)*

I have not received relevant documents from relevant departments

**wait-4** *(AL=4.0)*

I have not received response from relevant departments

**adaptive** *(AL=1.8)*

I have not received response from relevant departments
Previous Work on Adaptive Policy

- READ and WRITE actions
• READ and WRITE actions

• sequential decision making → reinforcement learning (Gu et al. 2017)

• unstable training (randomness in exploration)

• complicated (two models trained in two stages)

• worse performance (than wait-k model)
Previous Work on Adaptive Policy

- READ and WRITE actions
- sequential decision making → reinforcement learning (Gu et al. 2017)
  - unstable training (randomness in exploration)
  - complicated (two models trained in two stages)
  - worse performance (than wait-k model)
- can we learn a better model with adaptive policy via simpler methods?
Our Idea: Single Model, with READ as a Word

Vocabulary

{ the → WRITE
   learn → WRITE
   good → WRITE
   on → WRITE
   one → WRITE
   ...

WRITE
Our Idea: Single Model, with READ as a Word

Vocabulary

{ the \rightarrow WRITE \quad \text{learn} \rightarrow WRITE \quad \text{good} \rightarrow WRITE
\quad \text{on} \rightarrow WRITE \quad \text{one} \rightarrow WRITE
\ldots }

\rightarrow \quad R \quad \rightarrow \quad \text{READ}

Vocabulary

{ \text{the} \rightarrow \text{WRITE} \quad \text{learn} \rightarrow \text{WRITE} \quad \text{good} \rightarrow \text{WRITE}
\quad \text{on} \rightarrow \text{WRITE} \quad \text{one} \rightarrow \text{WRITE}
\ldots }
Our Idea: Single Model, with READ as a Word

Vocabulary

{ the \to WRITE, learn \to WRITE, good \to WRITE, on \to WRITE, one \to WRITE, ... }

NMT Model

{ R \to READ, the \to WRITE, learn \to WRITE, good \to WRITE, on \to WRITE, one \to WRITE, ... }

President Bush

Bùshí zōngtōng zài Mòsīkē
Bush President in Moscow

READ yǔ with

WRITE meets
Learn a Single Model via Imitation Learning

- imitation learning
- learn to imitate a given expert policy
Learn a Single Model via Imitation Learning

- imitation learning
- learn to imitate a given expert policy

basic ideas
- merge two models into one
- add read action into target vocabulary
- end-to-end training
- design an expert policy to use imitation learning
Learn a Single Model via Imitation Learning

- imitation learning
- learn to imitate a given expert policy

**basic ideas**
- merge two models into one
- add read action into target vocabulary
- end-to-end training
- design an expert policy to use imitation learning

_for more details_
_come to my short talk tomorrow_

(Target)

Source

- I have not received responses from relevant departments

(String)

- I
- have
- not
- received
- responses
- from
- relevant
- departments

(String)

- we
- have
- not
- received
- from
- relevant
- departments

(String/Response)
Another Much Simpler Idea

- on-the-fly decide READ or WRITE
- depending on $p(y_i|\cdots)$
- if not confident enough, READ
  - switch to wait-(k+1) (more conservative)
- otherwise WRITE
  - switch to wait-(k-1) (more aggressive)
Part III: Remaining Challenges
Part III: Remaining Challenges

- Speech Recognition-related
  - coping with ASR noise, esp. homophones
  - code switching
  - sentence breaking
  - prosody lost in translation
  - directly speech-to-speech without text-to-text?
Part III: Remaining Challenges

- Speech Recognition-related
  - coping with ASR noise, esp. homophones
  - code switching
  - sentence breaking
  - prosody lost in translation
  - directly speech-to-speech without text-to-text?
- Incremental Text-to-Speech Synthesis (TTS)
Part III: Remaining Challenges

- Speech Recognition-related
  - coping with ASR noise, esp. homophones
  - code switching
  - sentence breaking
  - prosody lost in translation
  - directly speech-to-speech without text-to-text?
- Incremental Text-to-Speech Synthesis (TTS)
- Better Dataset for Training
Part III: Remaining Challenges

- Speech Recognition-related
  - coping with ASR noise, esp. homophones
  - code switching
  - sentence breaking
  - prosody lost in translation
  - directly speech-to-speech without text-to-text?
- Incremental Text-to-Speech Synthesis (TTS)
- Better Dataset for Training
- Detecting and Fixing Mistakes (esp. anticipation errors)
Coping with ASR noise

- neural MT is fragile, and automatic speech recognition (ASR) output is noisy
- our work (Liu et al, ACL 2019): robust neural MT using phonetic information

<table>
<thead>
<tr>
<th>Clean Input</th>
<th>输出</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>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Noisy Input</th>
<th>输出</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>
<td></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>
<td></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.
Baidu ASR is awesome at code-switching (English terms in Chinese speech)
Baidu ASR’s Code-Switching Capabilities

- Baidu ASR is awesome at code-switching (English terms in Chinese speech)
Baidu ASR’s Code-Switching Capabilities

- Baidu ASR is awesome at code-switching (English terms in Chinese speech)
Better Dataset for Training Simultaneous Translation

- standard parallel text is not made for simultaneous translation
  - involves too many “unnecessary long-distance reorderings”
- simultaneous interpretation corpora is not ideal training data either
  - contains too many mistakes, speech repairs, and compressions
- again, our goal is short latency (like human simultaneous interpretation) and good quality (like human written translation)

uh... 我们认为 安理会... ah... 没有必要 介入
uh... we think sec. council uh... no need intervene

We believe that it is necessary for the security council to get involved
Better Dataset for Training Simultaneous Translation

- standard parallel text is not made for simultaneous translation
- involves too many “unnecessary long-distance reorderings”
- simultaneous interpretation corpora is not ideal training data either
- contains too many mistakes, speech repairs, and compressions
- again, our goal is short latency (like human simultaneous interpretation) and good quality (like human written translation)

We believe that it is necessary for the security council to get involved
Better Dataset for Training Simultaneous Translation

- idea: rephrase target side of parallel text to remove unnecessary reorderings

**mandatory reordering**

(Chinese) PP VP => (English) VP PP

(Chinese) PP S => (English) PP S or S PP

**optional reordering**

reference translation

- Xi Jinping in 2012 yr in Beijing elected

“Xi Jinping was elected in Beijing in 2012”

---

see also He et al (2015)

---

“About Clintonism, there is no accurate definition.”
Detecting and Fixing Mistakes

- idea: use a slower policy to verify the current policy’s output along the way
The point of this talk is to “抛砖引玉”, i.e., to stimulate interests in this long-standing problem.
非常感谢您来听我的演讲

Thank you very much for listening to my speech
非常感谢您来听我的演讲

Thank you very much for listening to my speech

Code (will be) available at https://nlp.baidu.com/paddlenlp using https://github.com/PaddlePaddle framework (it supports both static & dynamic graphs)
(the code for robust decoding with ASR noise is already available)

Two Posters after the coffee break (10:30), Session 4A (#4 & #6)
Short Talk tomorrow, Session 8D (17:13, CAVANIGLIA)