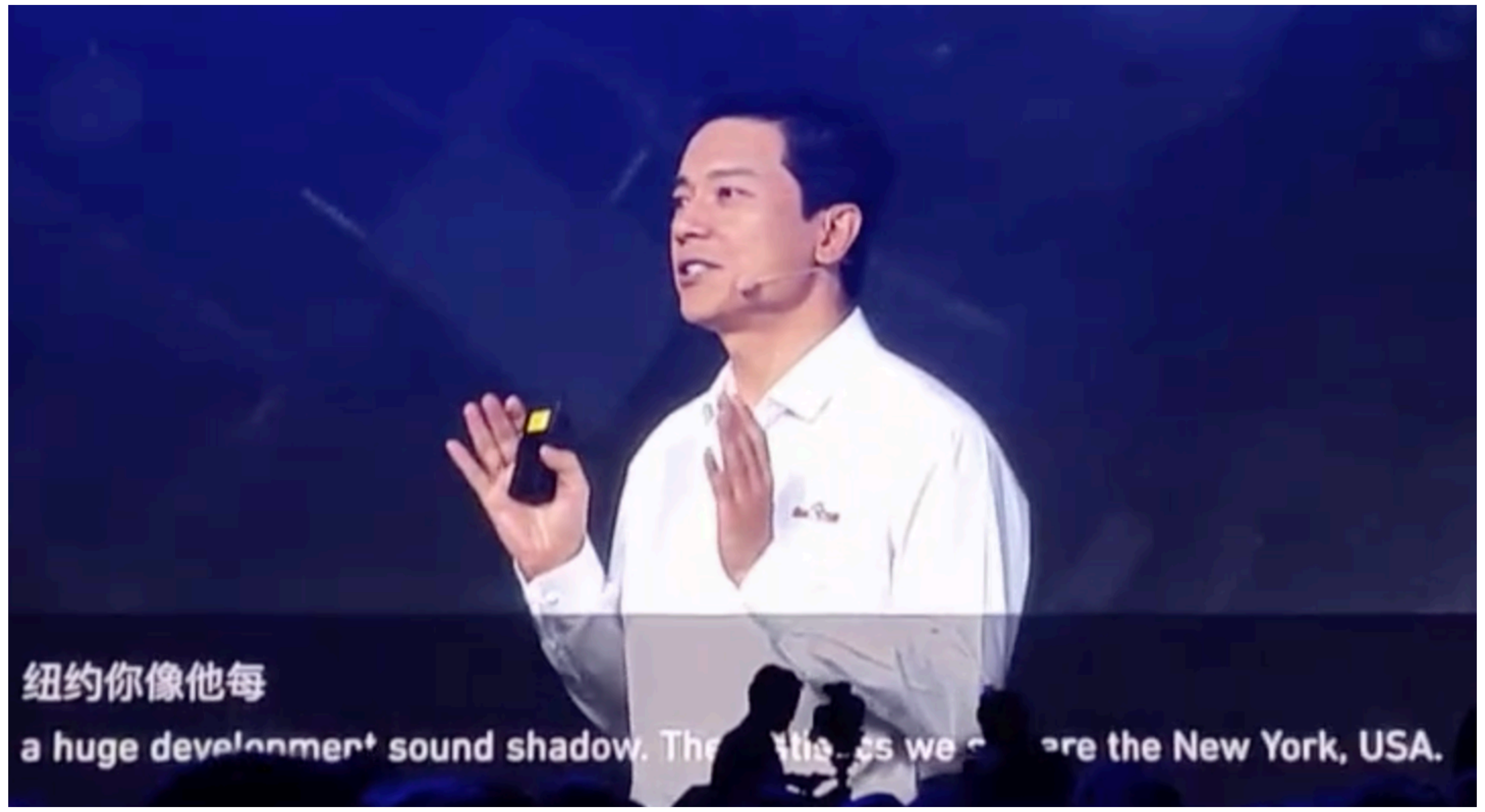
Simultaneous Translation:
Recent Advances and Remaining Challenges



Liang Huang

Baidu Research (USA) and Oregon State University

Baidu Research

Association for Computational Linguistics
57th ACL 2019 Florence
ANNUAL MEETING July 28th - August 2nd
of the Association for Computational Linguistics

Baidu 百度

Consecutive vs. Simultaneous Interpretation

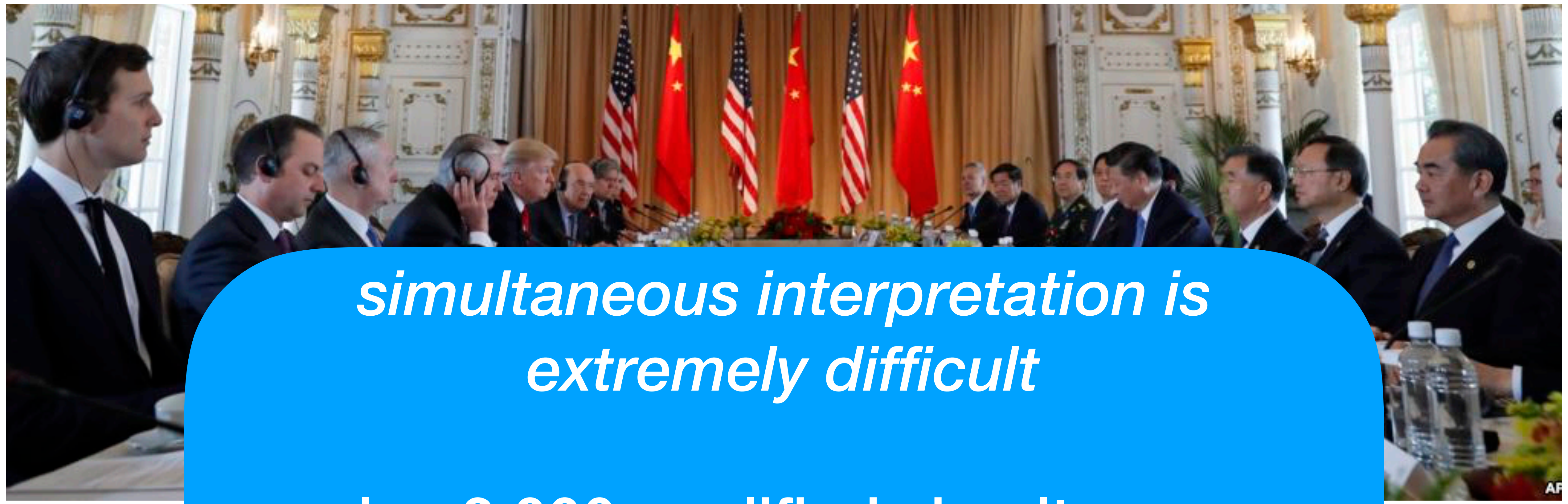
consecutive interpretation
multiplicative latency (x2)

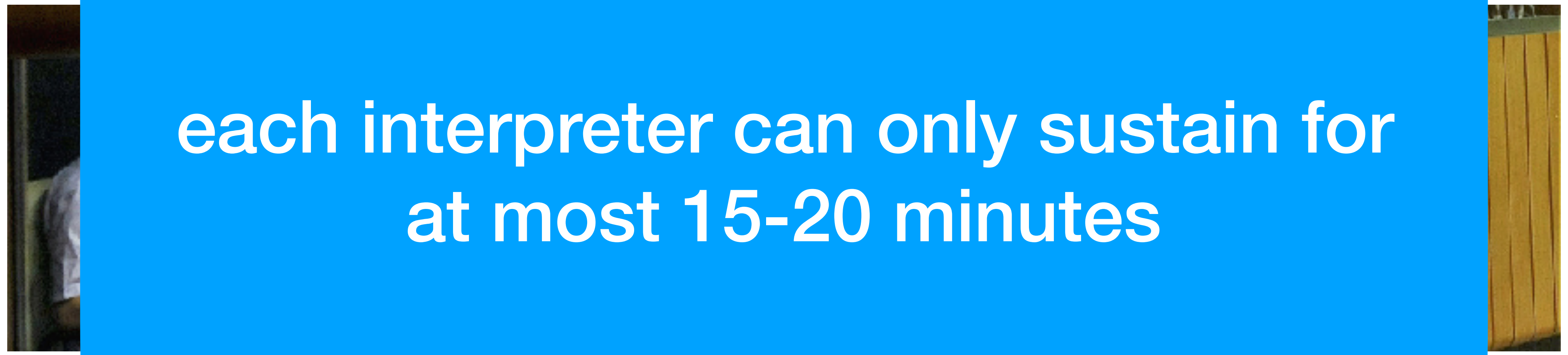
simultaneous interpretation
additive latency (+3 secs)

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…)

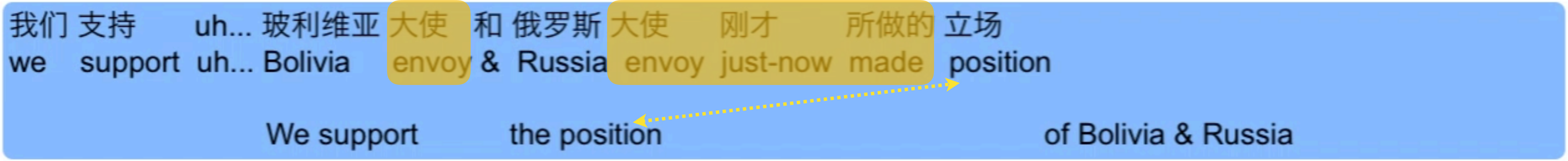
我们 支持 uh... 玻利维亚 大使 和 俄罗斯 大使 刚才 所做的 立场
we support uh... Bolivia envoy & Russia envoy just-now made position

We support the position of Bolivia & Russia

from United Nations Proceedings Speech Corpus (LDC2014S08, Chay et al, 2014)

Simultaneous Interpreters: Strategies & Limitations

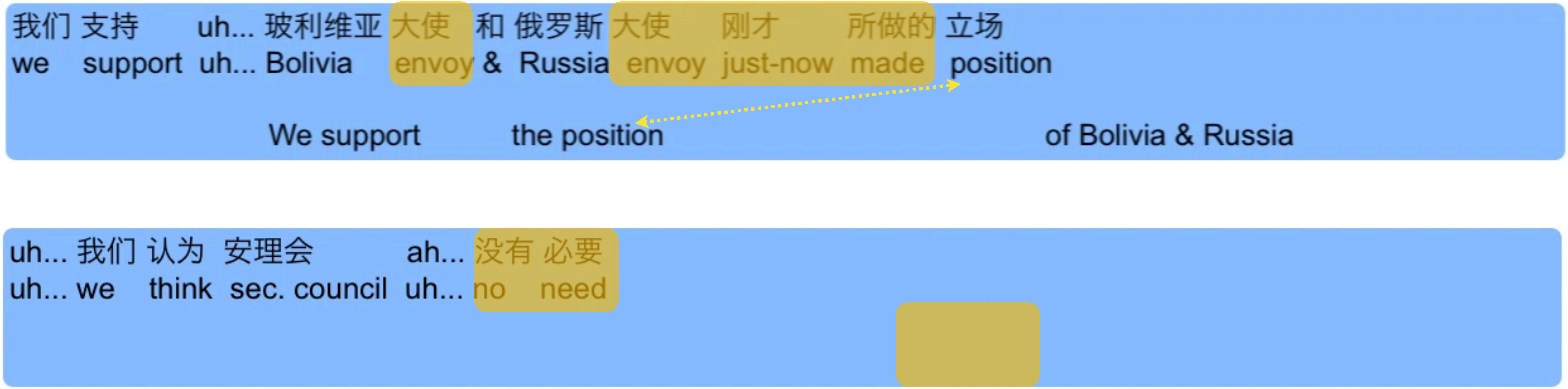
- anticipation, summarization, generalization, etc…
- and they inevitably make (quite a bit of) mistakes
- “human-level” *quality*: much lower than normal translation
- “human-level” *latency*: very short: 2~4 secs (actually higher latency *hurts* quality…)



from United Nations Proceedings Speech Corpus (LDC2014S08, Chay et al, 2014)

Simultaneous Interpreters: Strategies & Limitations

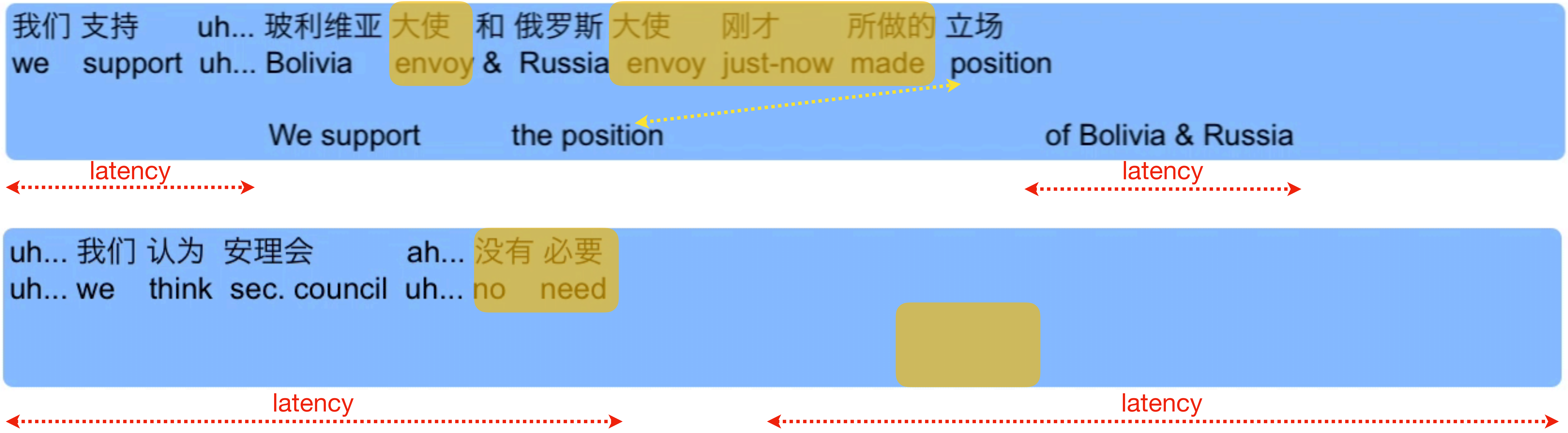
- anticipation, summarization, generalization, etc…
- and they inevitably make (quite a bit of) mistakes
- “human-level” *quality*: much lower than normal translation
- “human-level” *latency*: very short: 2~4 secs   (actually higher latency *hurts* quality…)



from United Nations Proceedings Speech Corpus (LDC2014S08, Chay et al, 2014)

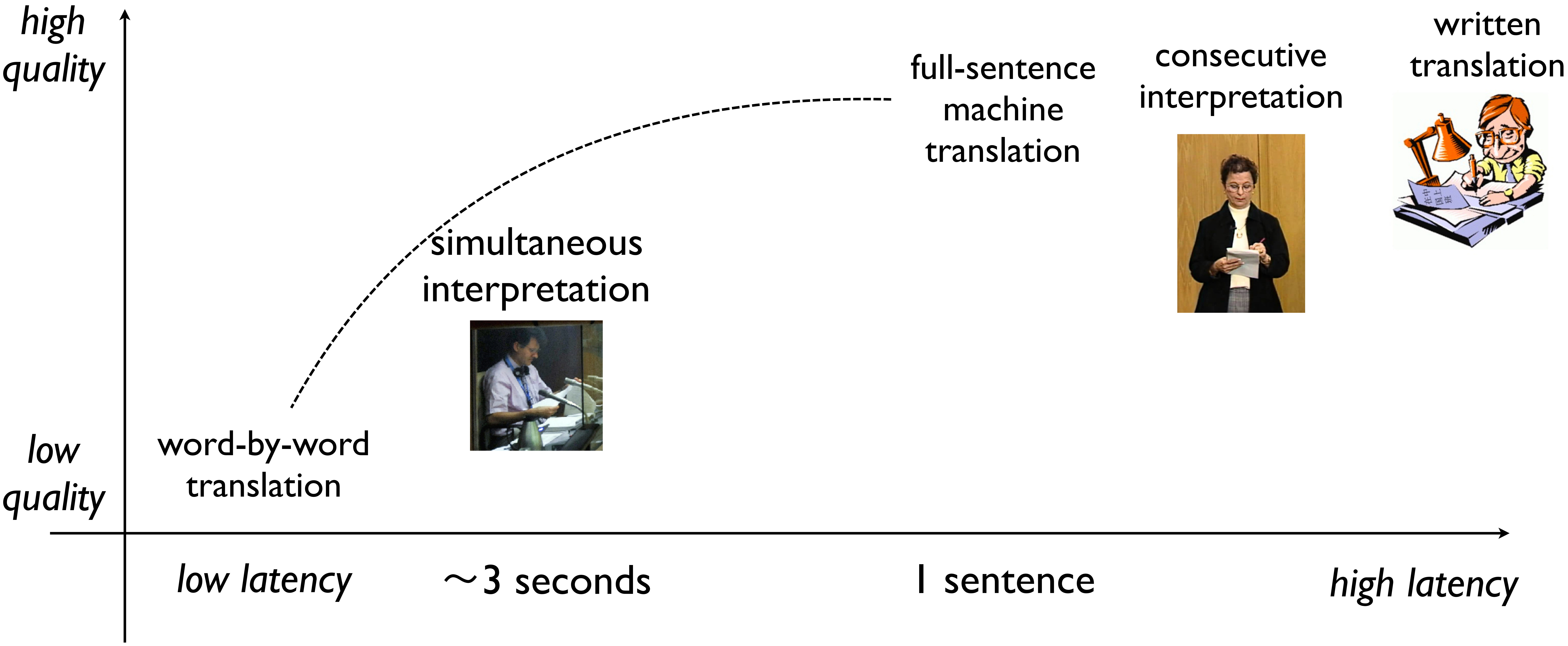
Simultaneous Interpreters: Strategies & Limitations

- 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…)

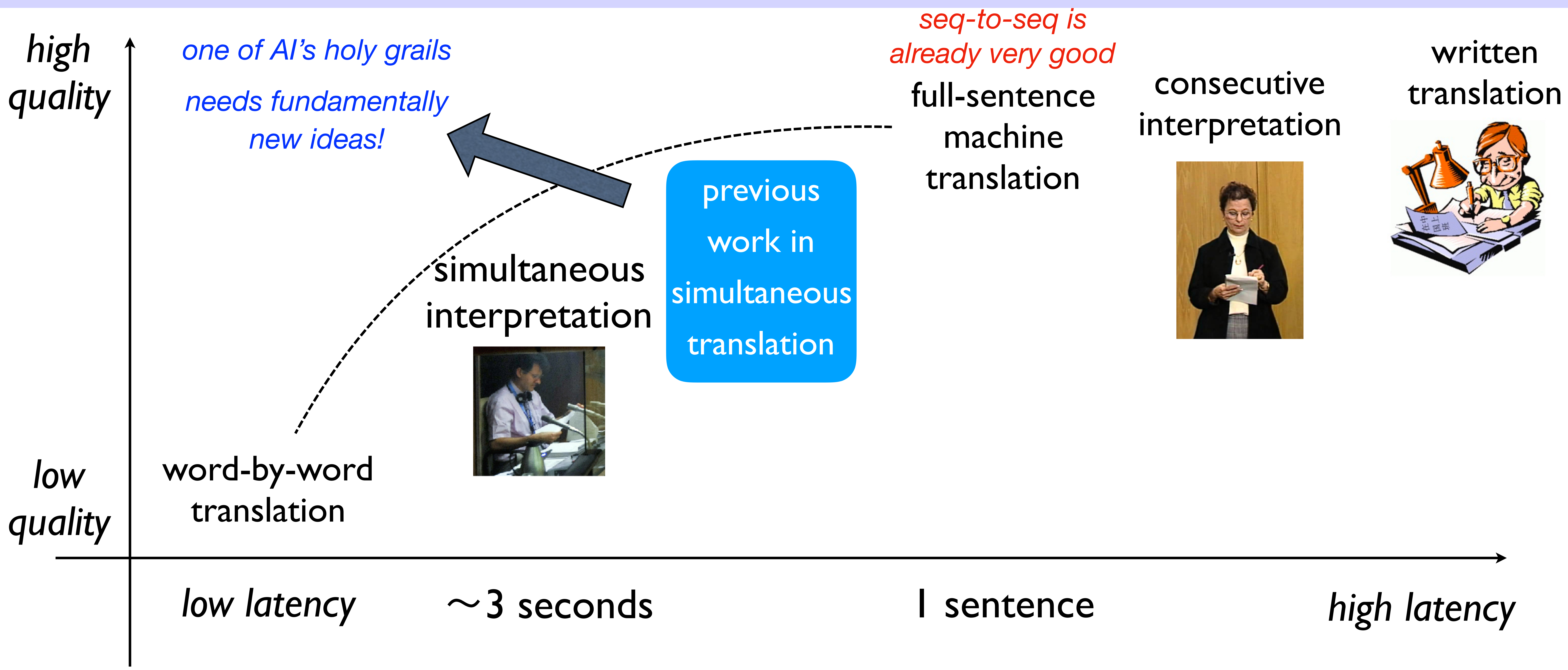


from United Nations Proceedings Speech Corpus (LDC2014S08, Chay et al, 2014)

Tradeoff between Latency and Quality



Tradeoff between Latency and Quality



Tradeoff between Latency and Quality

high quality

one of AI's holy grails

needs fundamentally new ideas!

seq-to-seq is already very good

full-sentence machine translation

consecutive interpretation

written translation

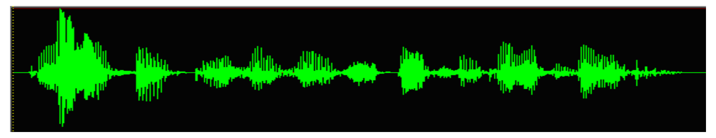
previous work in simultaneous translation

simultaneous interpretation

low quality

word-by-word translation

low latency ~3 seconds 1 sentence *high latency*



source speech stream … → *streaming* speech recognition → 布什 总统 在 … → source text stream → *simultaneous text-to-text translation* → **President Bush …** → target text stream → *incremental text-to-speech* →

…

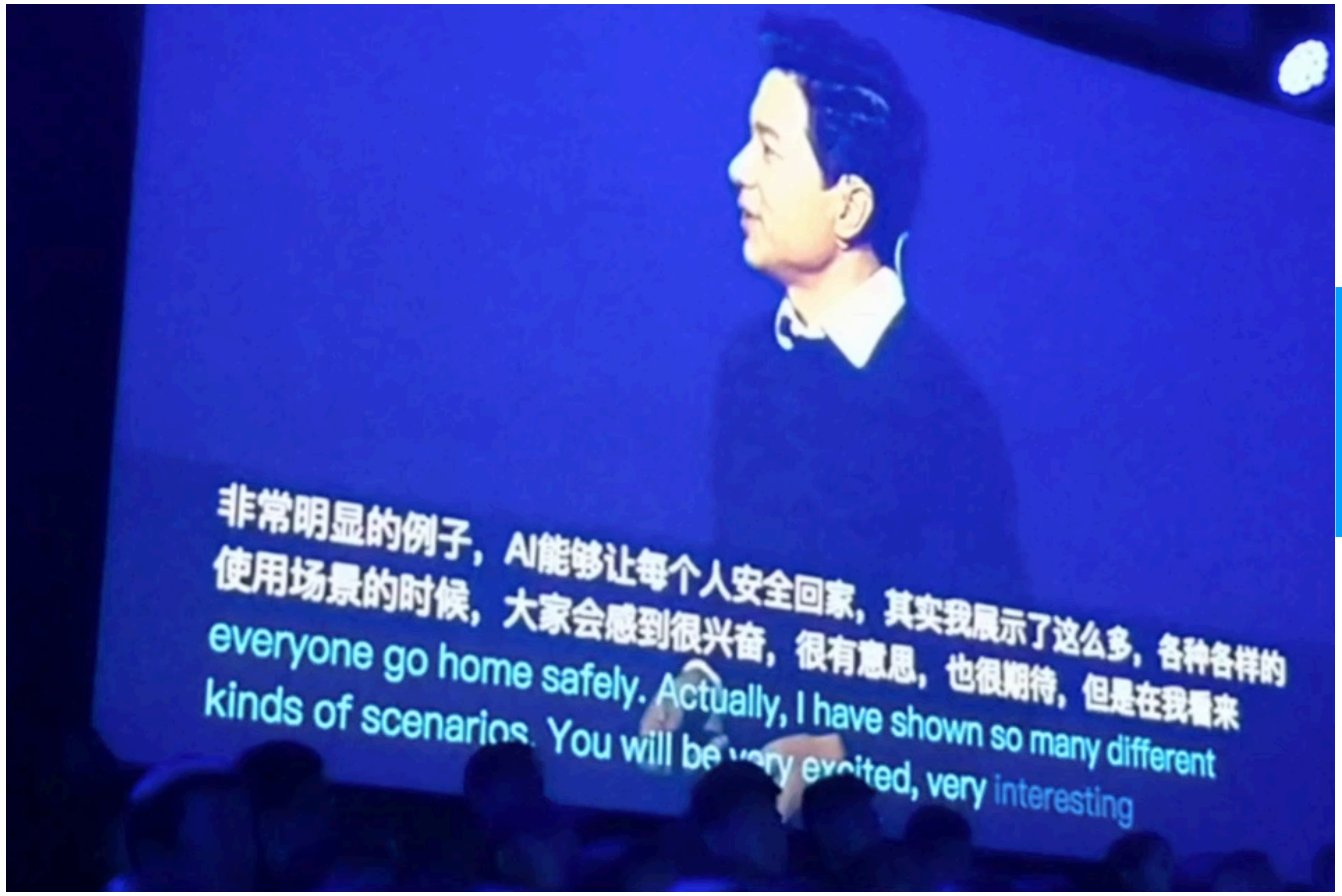
target speech stream

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)



our work

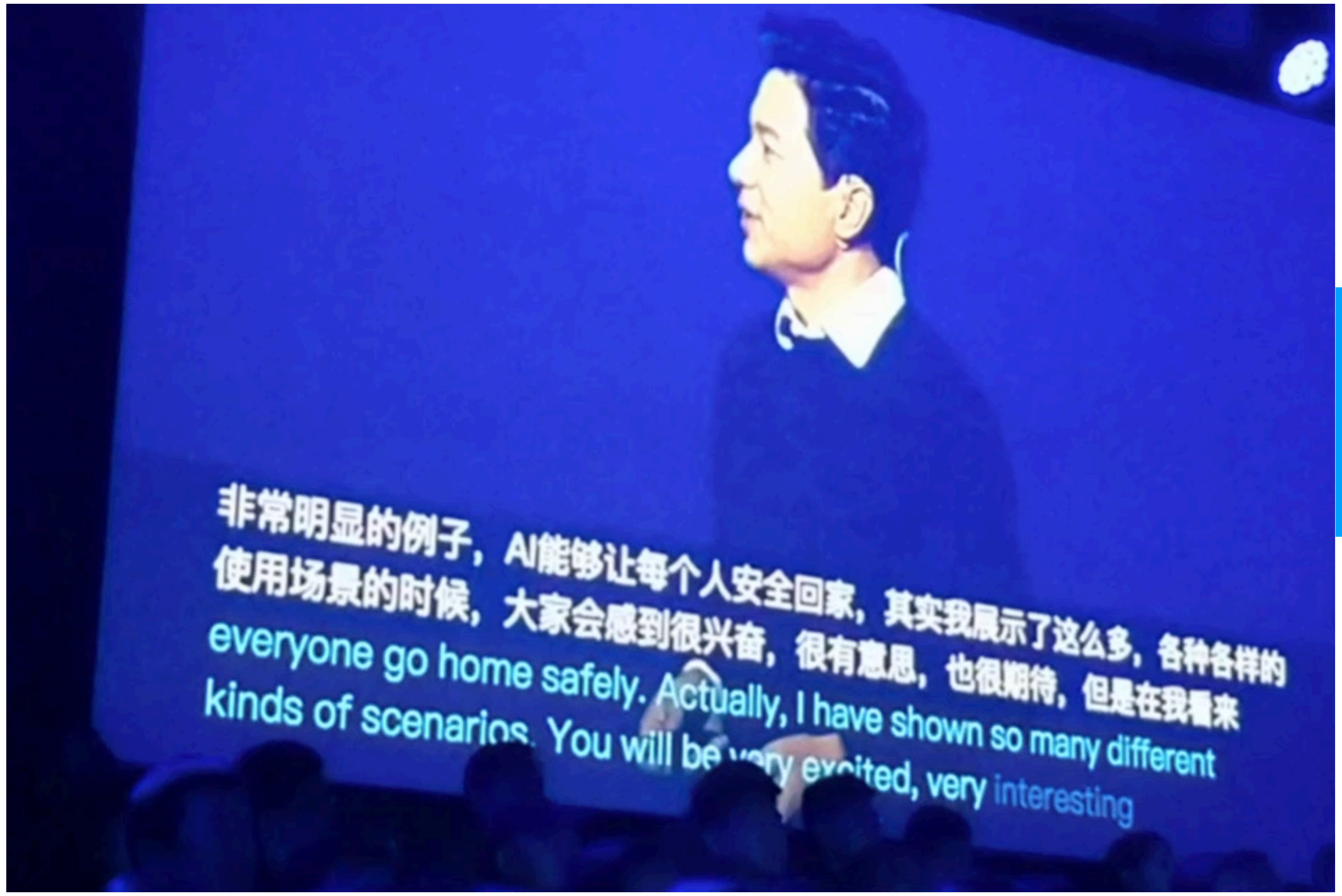
Baidu World Conference, Nov. 2018
low-latency simultaneous translation (latency: ~3 secs)

Media coverage:

IEEE SPECTRUM, MIT Technology Review, CNBC, VentureBeat, siliconANGLE, Synced AI Technology & Industry Review, South China Morning Post, engadget, FORTUNE, PROGRAMMER, The Register (Biting the hand that feeds IT), lowyat.net (malaysia's largest online community), Packt, RED PULSE, FLIPBOARD, China Knowledge

Our Breakthrough in 2018

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



our work

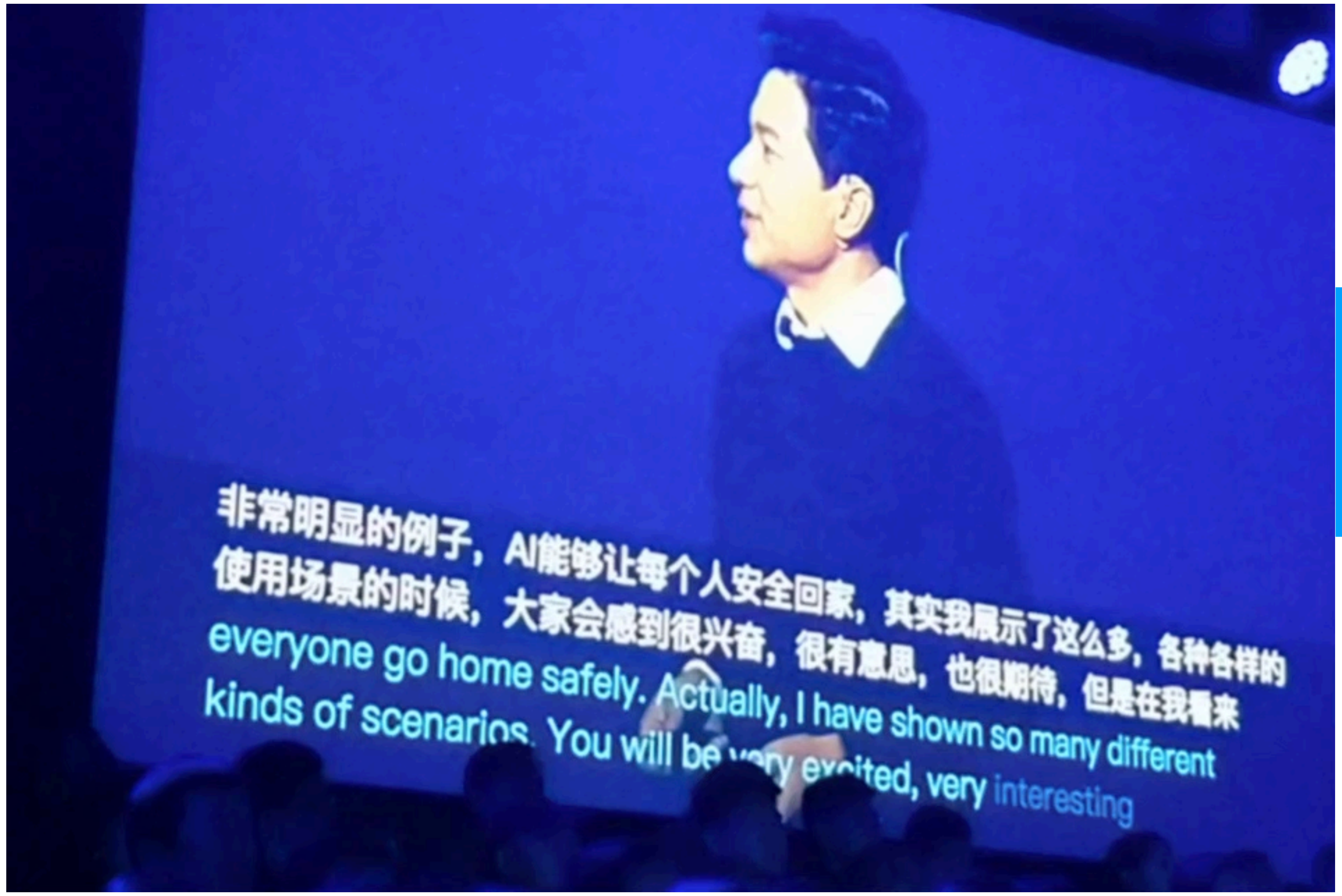
Baidu World Conference, Nov. 2018
low-latency simultaneous translation (latency: ~3 secs)

Media coverage:

IEEE SPECTRUM, MIT Technology Review, CNBC, VentureBeat, siliconANGLE, Synced AI Technology & Industry Review, South China Morning Post, engadget, FORTUNE, I PROGRAMMER, The Register Biting the hand that feeds IT, lowyat.net malaysia's largest online community, Packt, RED PULSE, FLIPBOARD, China Knowledge

Our Breakthrough in 2018

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



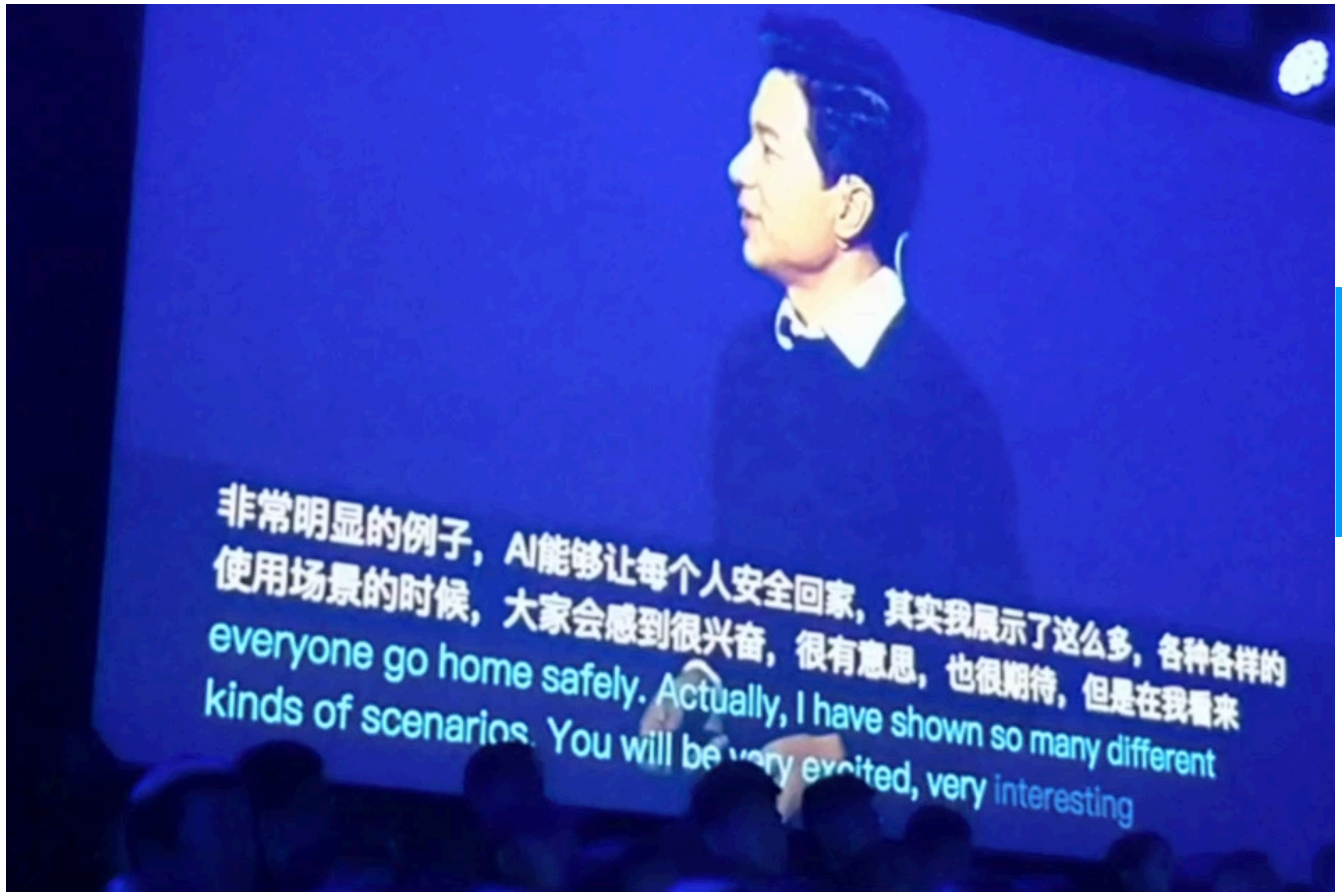
our work

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

Media coverage:
IEEE SPECTRUM, MIT Technology Review, CNBC, VentureBeat, silicon ANGLE, Synced AI Technology & Industry Review, South China Morning Post, engadget, FORTUNE, PROGRAMMER, The Register (Biting the hand that feeds IT), lowyat.net (malaysia's largest online community), Packt, RED PULSE, FLIPBOARD, China Knowledge

Our Breakthrough in 2018

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



our work

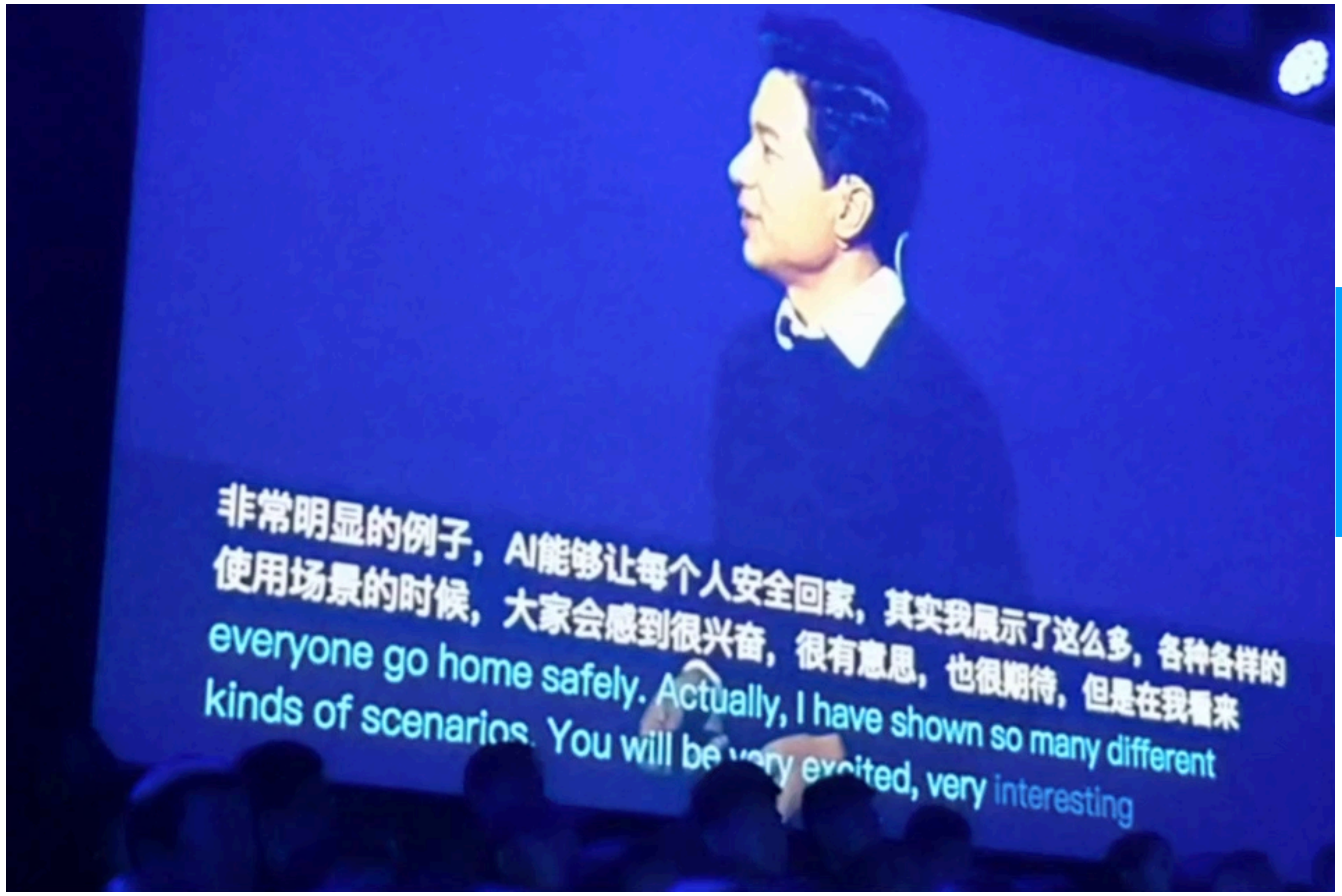
Baidu World Conference, Nov. 2018
low-latency simultaneous translation (latency: ~3 secs)

Media coverage:

IEEE Spectrum, MIT Technology Review, CNBC, VentureBeat, siliconANGLE, Synced, South China Morning Post, engadget, FORTUNE, Programmer, The Register, lowyat.net, Packt, Red Pulse, Flipboard, China Knowledge

Our Breakthrough in 2018

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



our work

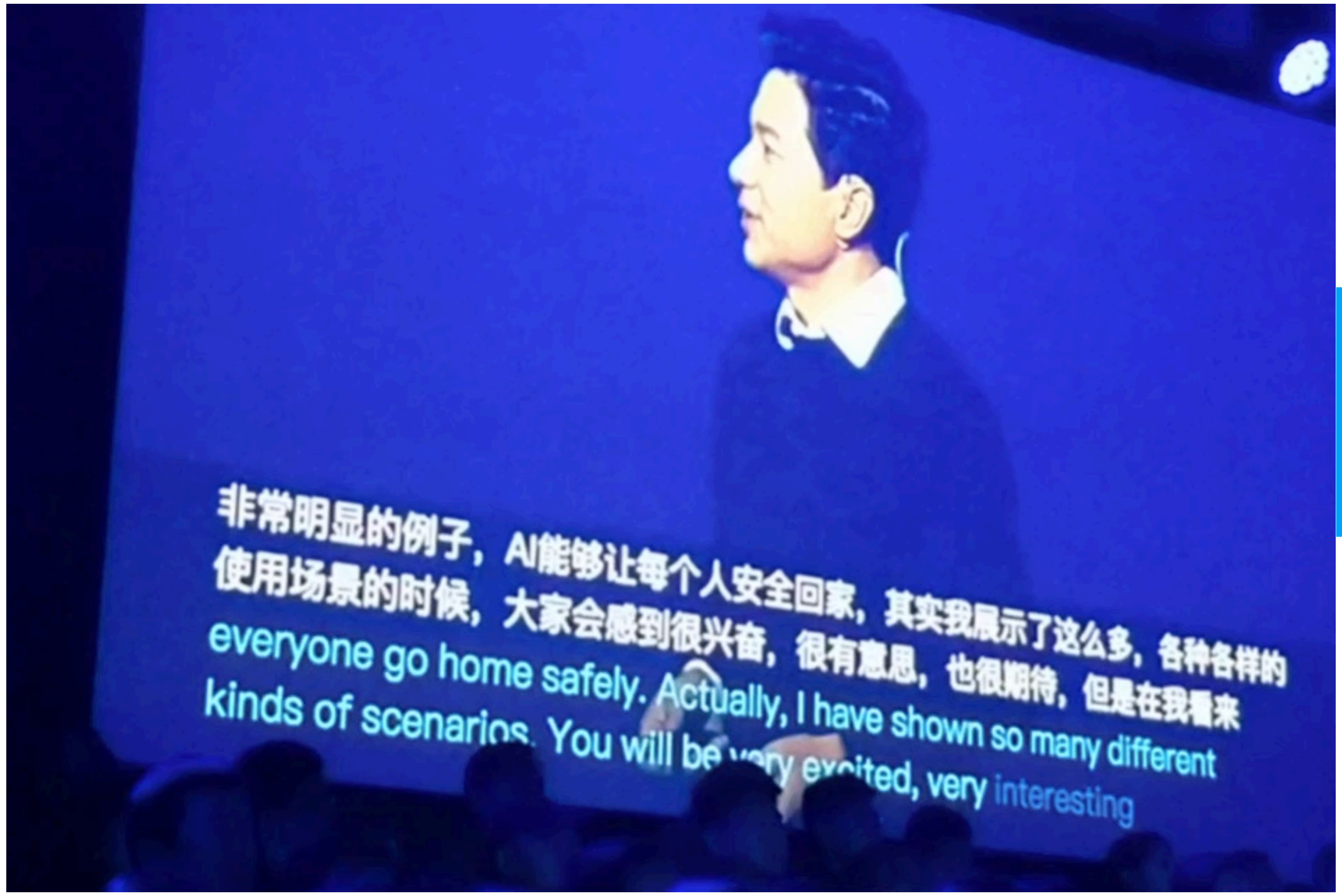
Baidu World Conference, Nov. 2018
low-latency simultaneous translation (latency: ~3 secs)

Media coverage:

IEEE SPECTRUM · MIT Technology Review · CNBC · VentureBeat · siliconANGLE · Synced AI TECHNOLOGY & INDUSTRY REVIEW · South China Morning Post · engadget · FORTUNE · PROGRAMMER · The Register Biting the hand that feeds IT · lowyat.net malaysia's largest online community · Packt · RED PULSE · FLIPBOARD · China Knowledge

Our Breakthrough in 2018

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



our work

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

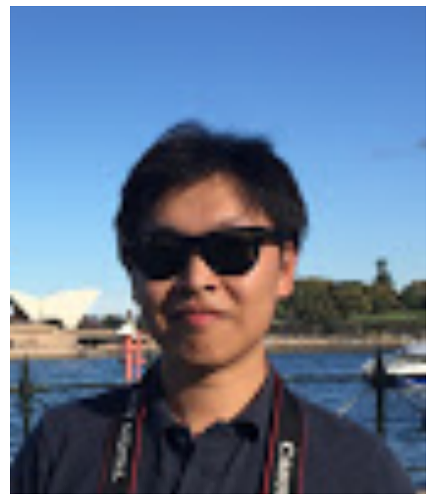

Haifeng Wang


Zhongjun He


Hao Xiong

request


Mingbo Ma


Kaibo Liu

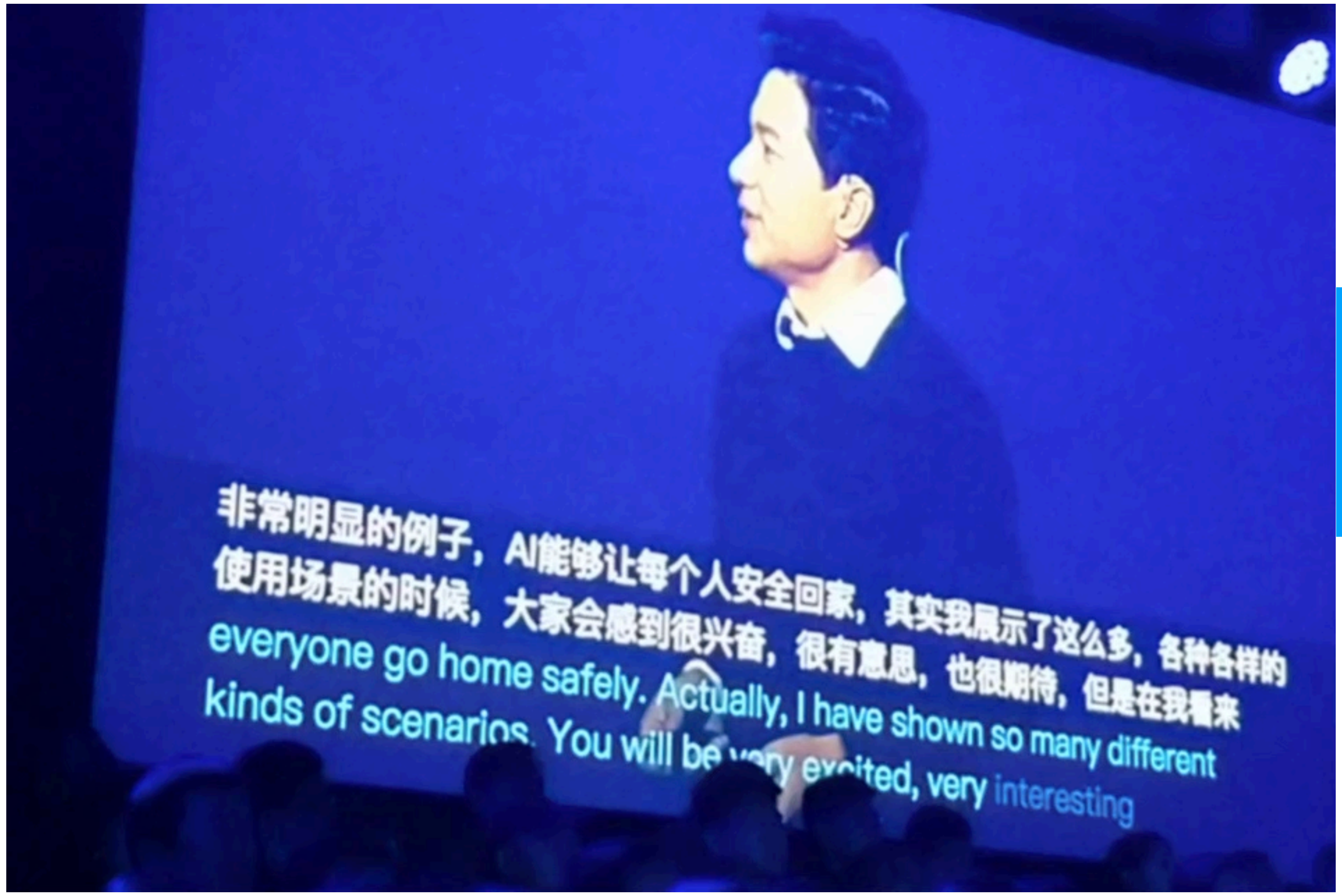

Renjie Zheng

Baidu 百度

Baidu Research

Our Breakthrough in 2018

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



our work

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


Haifeng Wang


Zhongjun He

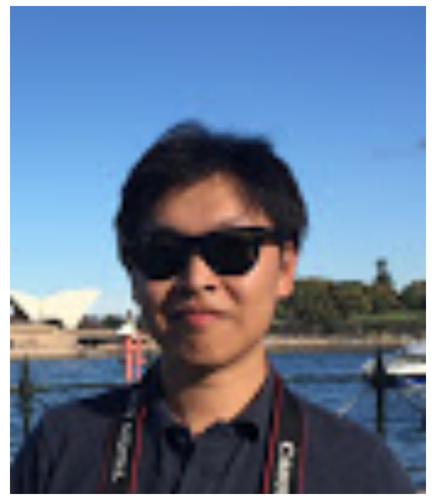

Hao Xiong


Ken Church

request

I really need low-latency simultaneous translation!


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)

ich	bin	mit	dem	Zug	nach	Ulm	**gefahren**
I	am	with	the	train	to	Ulm	**traveled**
I	(. . . . . . *waiting* . . . . . .)						**traveled** by train to Ulm

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)

ich	bin	mit	dem	Zug	nach	Ulm	**gefahren**	
I	am	with	the	train	to	Ulm	**traveled**	
I		*(. . . . . . waiting. . . . . .)*					**traveled**	by train to Ulm

Grissom et al, 2014

Bùshí	*zǒngtǒng*	*zài*	*Mòsīkē*	*yǔ*	*Éluósī*	*zǒngtǒng*	*Pǔjīng*	*huìwù*
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*

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)

ich	bin	mit	dem	Zug	nach	Ulm	**gefahren**
I	am	with	the	train	to	Ulm	**traveled**
I		*(. . . . . . waiting. . . . . .)*					**traveled** by train to Ulm

Grissom et al, 2014

Bùshí	zǒngtǒng	zài	Mòsīkē	yǔ	Éluósī	zǒngtǒng	Pǔjīng	huìwù
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

President Bush meets with Russian President Putin in Moscow

non-anticipative: President Bush (…… *waiting* ……) 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)

ich	bin	mit	dem	Zug	nach	Ulm	**gefahren**
I	am	with	the	train	to	Ulm	**traveled**
I	*(...... waiting......)*						**traveled** by train to Ulm

Grissom et al, 2014

Bùshí	*zǒngtǒng*	*zài*	*Mòsīkē*	*yǔ*	*Éluósī*	*zǒngtǒng*	*Pǔjīng*	*huìwù*
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*

President Bush meets with Russian President Putin in Moscow

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

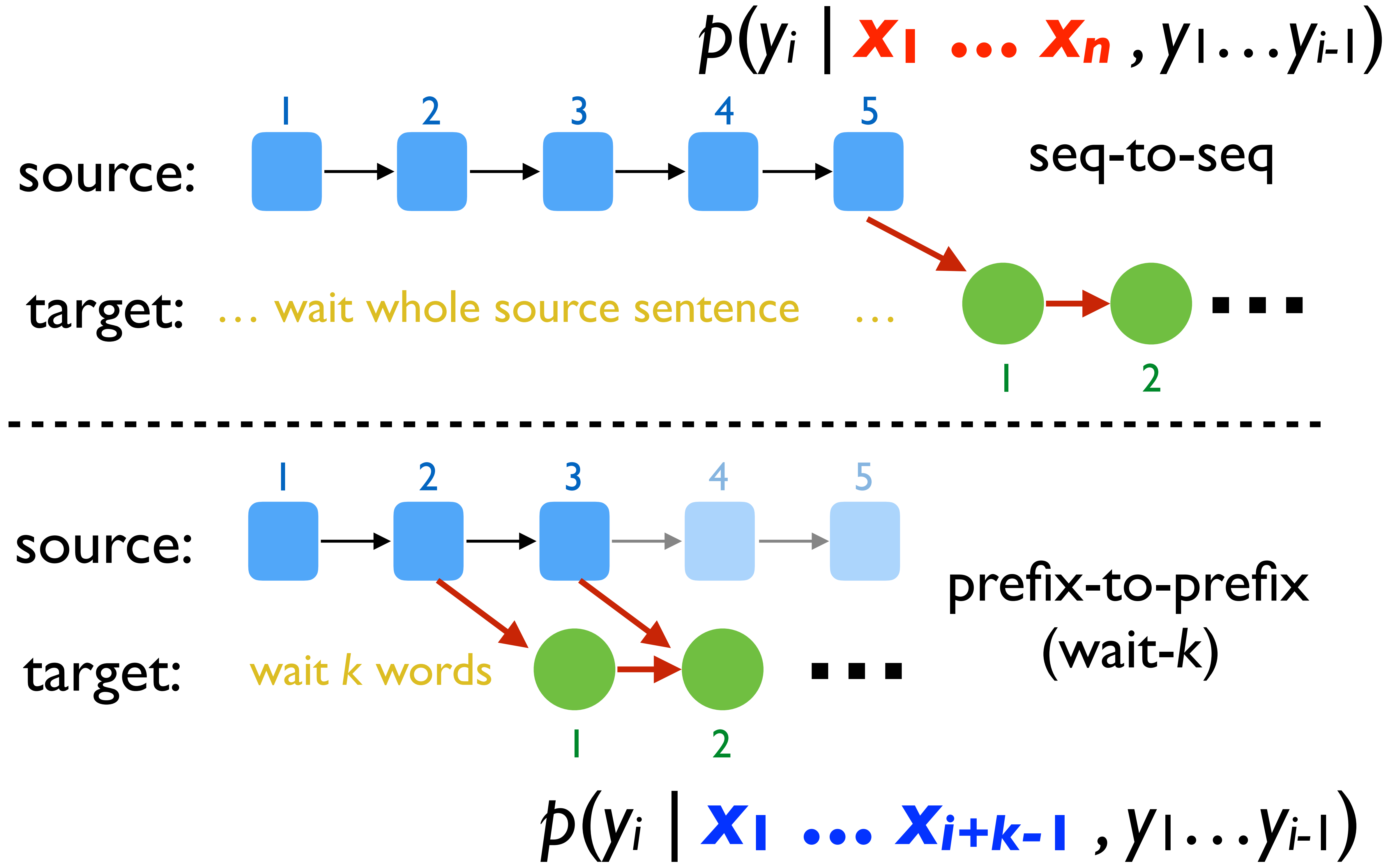
anticipative: President Bush meets with Russian President Putin in Moscow

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

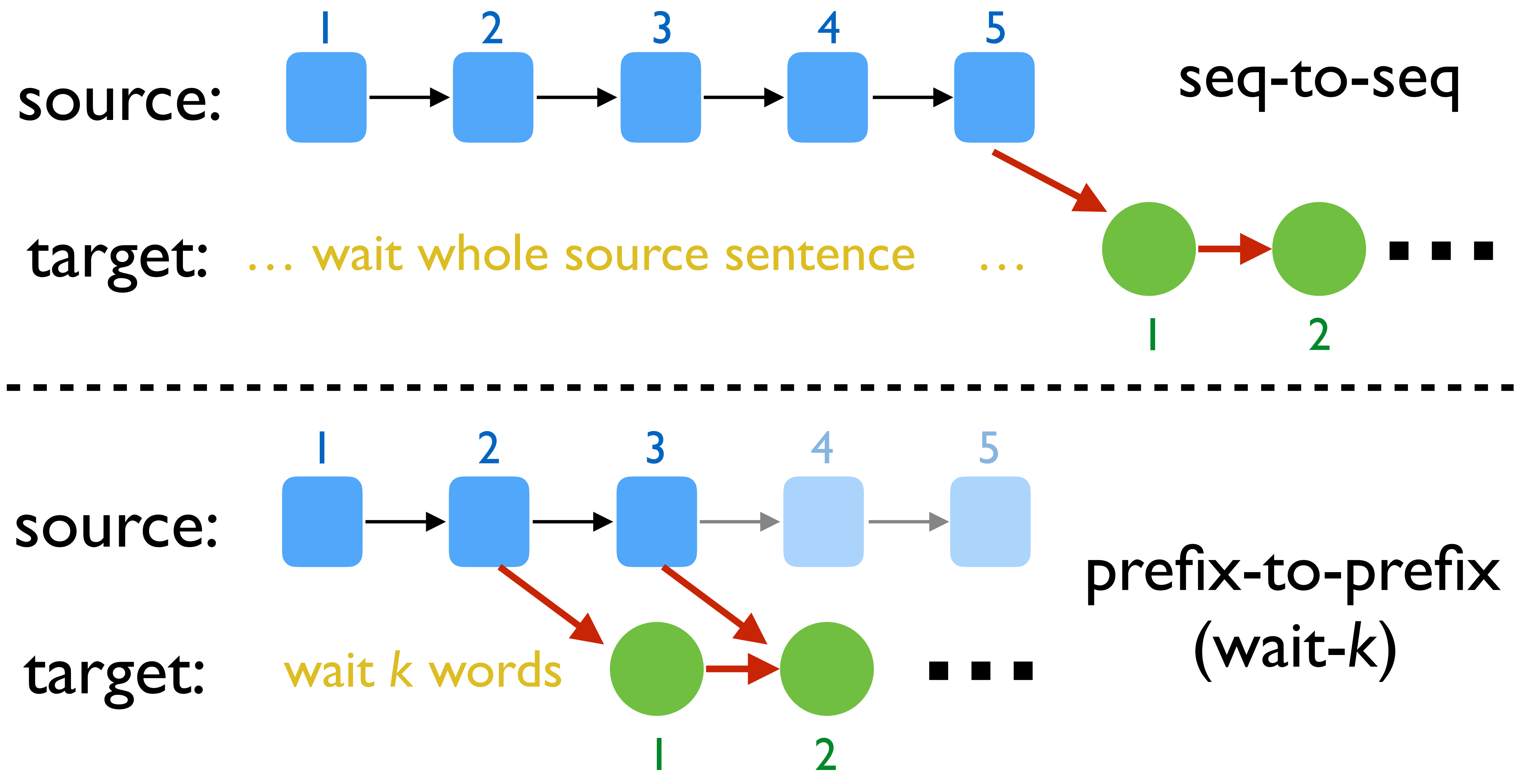
- standard seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix framework tailed to tasks with simultaneity
  - special case: wait-*k* policy: translation is always *k* words behind source sentence
  - decoding this way => controllable latency
  - training this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



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*

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



$$p(y_i \mid x_1 \ldots x_{i+k-1}, y_1 \ldots y_{i-1})$$

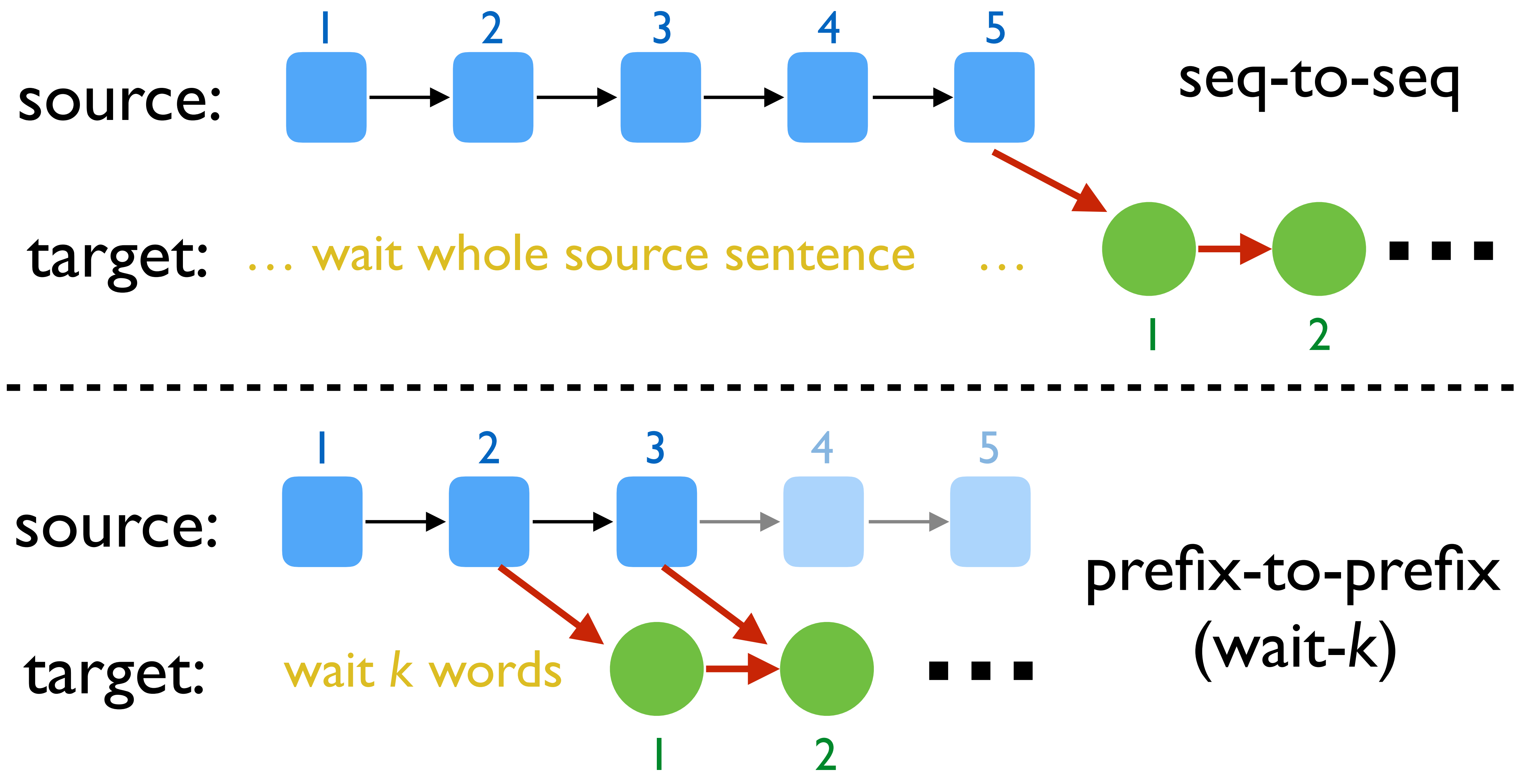
Bùshí	*zǒngtǒng*	*zài*	*Mòsīkē*
布什	总统	在	莫斯科
Bush	*President*	*in*	*Moscow*

wait 2 President Bush meets

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*

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



$$p(y_i \mid \mathbf{x_1} \ldots \mathbf{x_{i+k-1}}, y_1 \ldots y_{i-1})$$

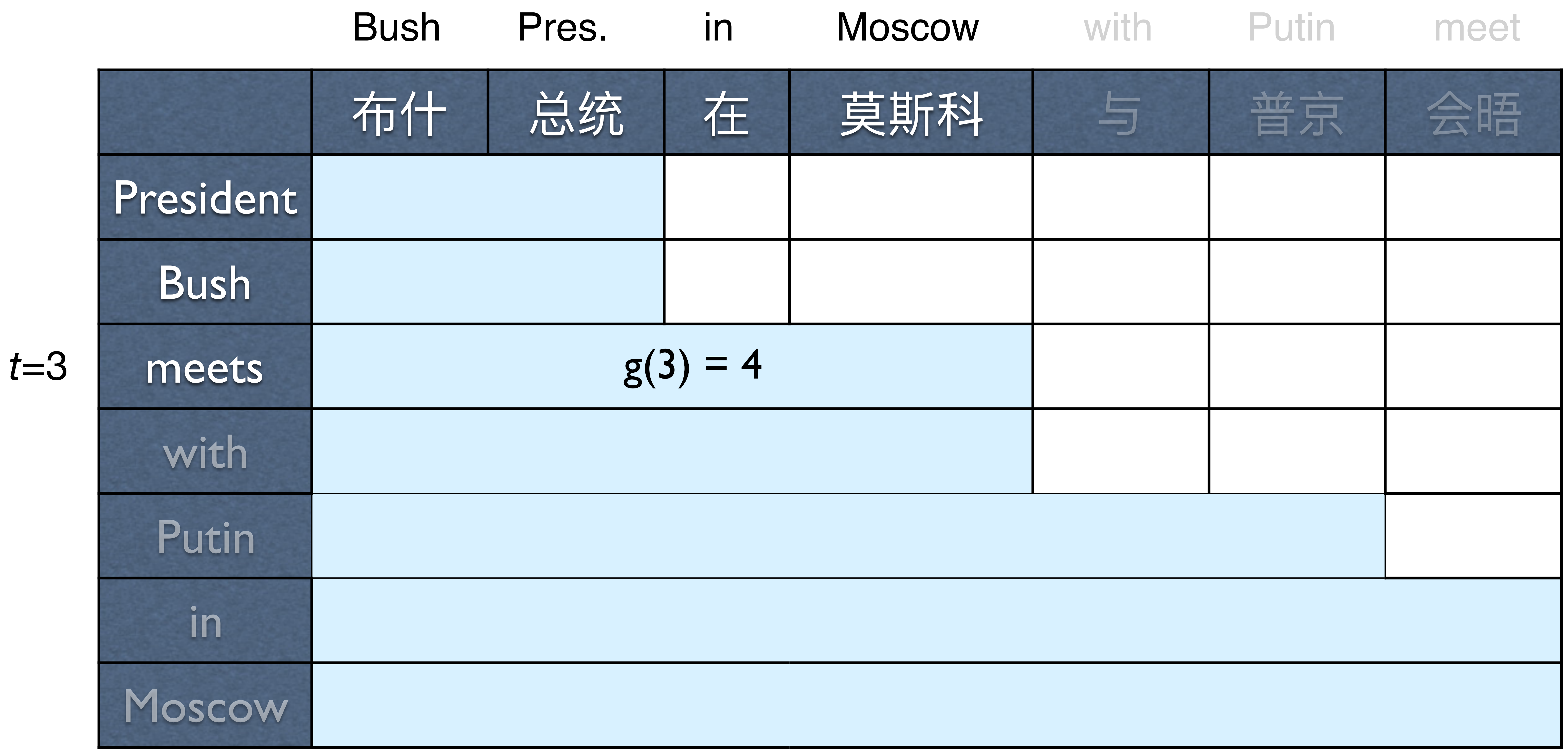
Bùshí	zǒngtǒng	zài	Mòsīkē	yǔ	Éluósī	zǒngtǒng	Pǔjīng	huìwù
布什	总统	在	莫斯科	与	俄罗斯	总统	普京	会晤
Bush	President	in	Moscow	with	Russian	President	Putin	meet

wait 2 President Bush meets with Russian President Putin in Moscow

More General Prefix-to-Prefix

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

- 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



this general framework can be used for other tasks such as incremental parsing and incremental text-to-speech

Research Demo

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

Research Demo

Baidu Research 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

jiāng zémín	duì	fǎ guó	zǒng tǒng	de	lái huá	fǎng wèn	biǎo shì	gǎn xiè	
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	

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

Baidu Research 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

jiāng zémín duì fǎ guó zǒng tǒng de lái huá fǎng wèn biǎo shì gǎn xiè

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

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

我们 支持 uh... 玻利维亚 大使 和 俄罗斯 大使 刚才 所做的 立场

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

We support the position of Bolivia & Russia

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

Latency-Accuracy Tradeoff

Latency-Accuracy Tradeoff

Deployment Demo

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

Deployment Demo

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

German=>English Anticipation Example

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):

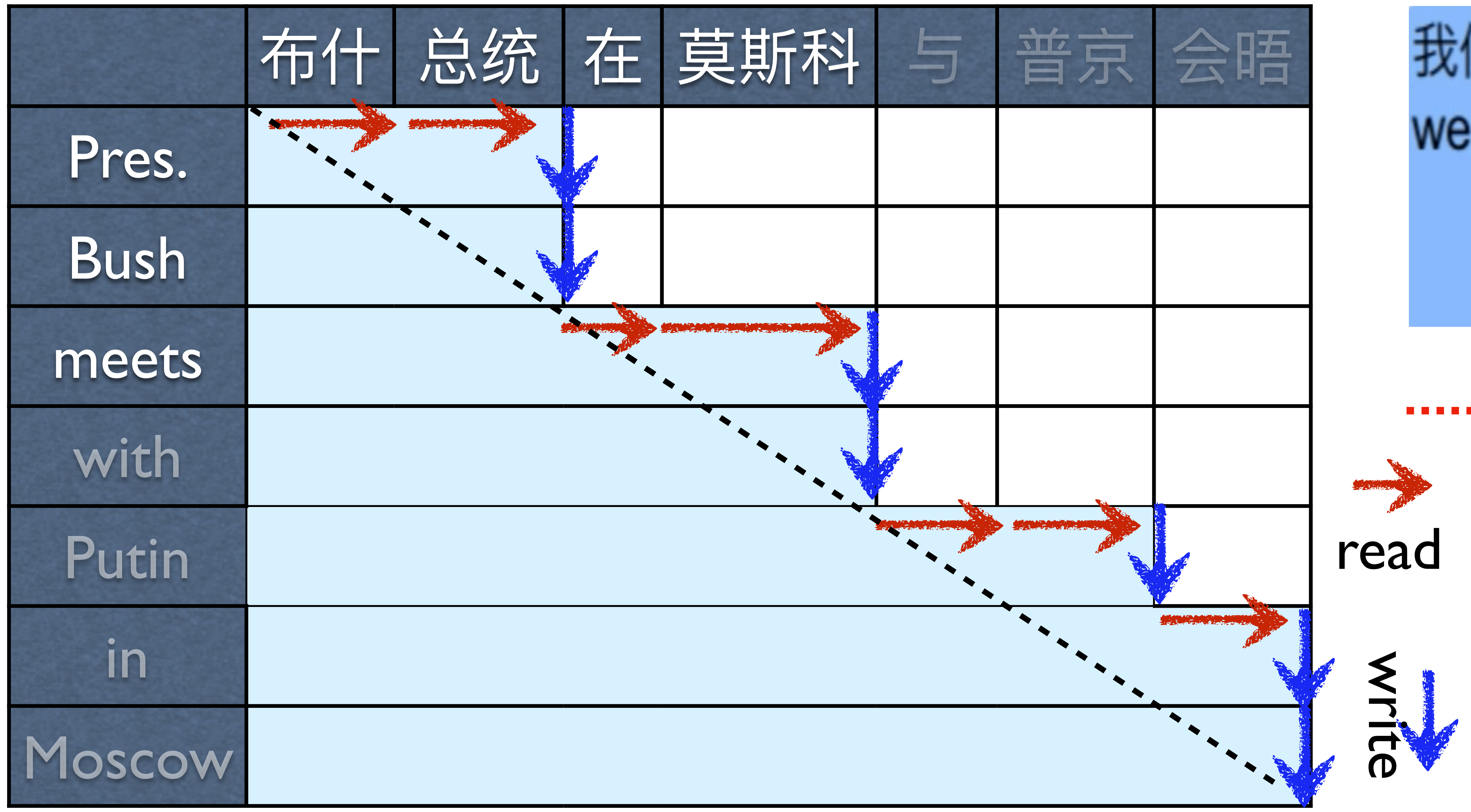
but , while congress does not agree on a course of action , several states no longer wait .

English translation (full-sentence baseline):

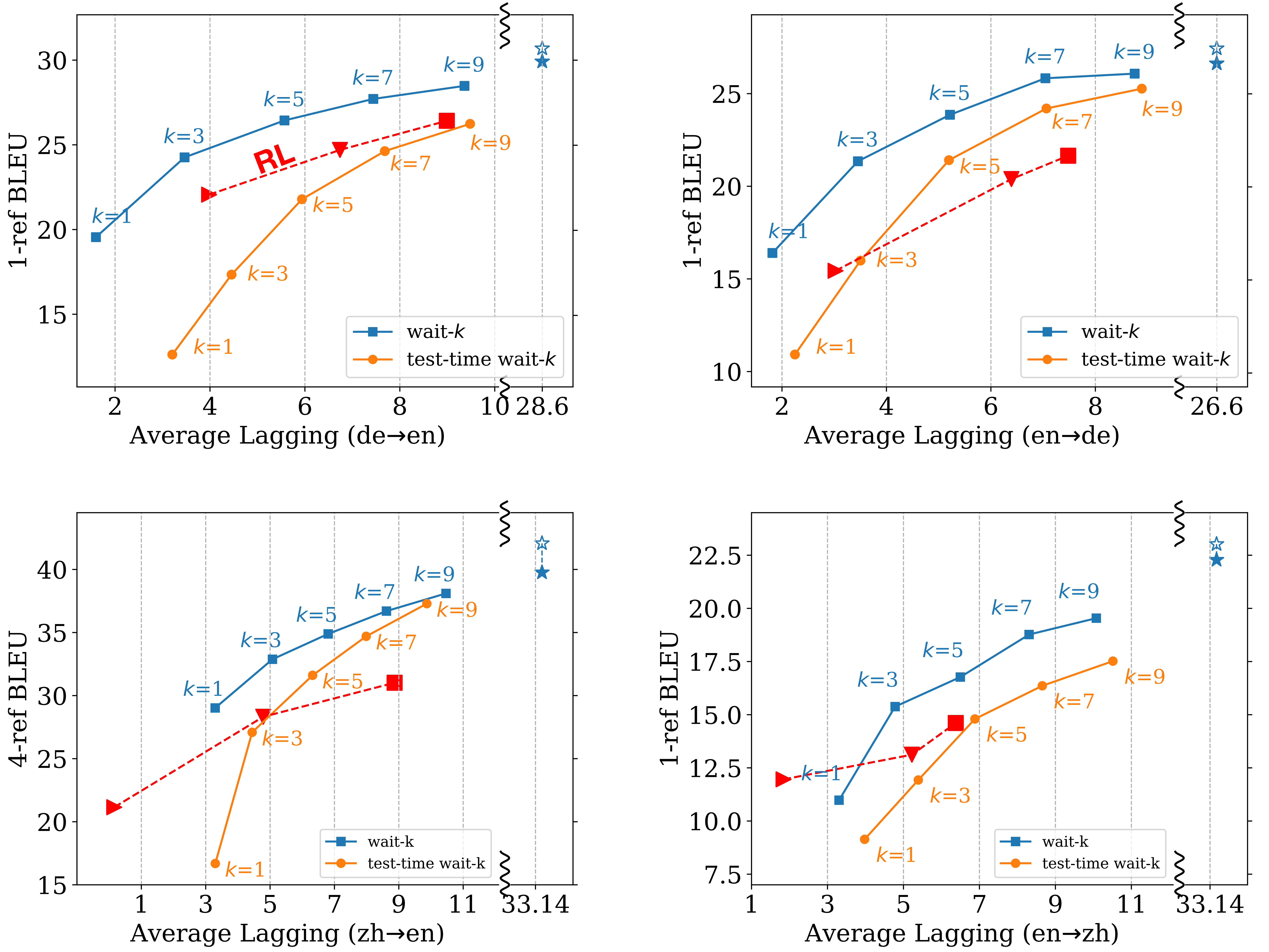
but , while congressional action can not be agreed , several states are no longer waiting .

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* (wait-*k*) ≈ *k*
- closely related to "ear-voice span" (EVS) in the interpretation literature



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

Media coverage:

IEEE SPECTRUM, MIT Technology Review, CNBC, VentureBeat, siliconANGLE, Synced (AI Technology & Industry Review), South China Morning Post, engadget, FORTUNE, PROGRAMMER, The Register (Biting the hand that feeds IT), lowyat.net (malaysia's largest online community), Packt, RED PULSE, FLIPBOARD, China Knowledge

Part II: Towards Adaptive Translation Policies

Part II: Towards Adaptive Translation Policies

	fixed-latency policies	*adaptive policies*
full-sentence MT model	Dalvi et al. (2018); **test-time wait-*k* (Ma et al. 2018)**	Grissom et al. (2014); Cho & Esipova (2016); Satija & Pineau (2016); Gu et al. (2017); Alinejad et al (2018); …
simultaneous MT model (our invention)	**wait-*k* (Ma et al. 2018)**	Arivazhagan et al. (ACL 2019) **Zheng et al. (ACL 2019)**

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)

	wǒ	*shàng*	*wèi*	*dédào*	*yǒuguān*	*bùmén*
input	我	尚	未	得到	有关	部门
	I	*yet*	*not*	*receive*	*relevant*	*department*

wait-1 (AL=1.4): **I have not received relevant**

Limitations of Fixed-Latency (wait-*k*) Policy

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

	wǒ	*shàng*	*wèi*	*dédào*	*yǒuguān*	*bùmén*
input	我	尚	未	得到	有关	部门
	I	*yet*	*not*	*receive*	*relevant*	*department*
wait-1 (AL=1.4)	**I**	**have**	**not**	**received**	**relevant**	~~**documents**~~

Limitations of Fixed-Latency (wait-*k*) Policy

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

	wǒ	*shàng*	*wèi*	*dédào*	*yǒuguān*	*bùmén*	*de*	*huíyìng*
input	我	尚	未	得到	有关	部门	的	回应
	I	*yet*	*not*	*receive*	*relevant*	*department*	*'s*	*response*

wait-1 (AL=1.4) **I have not received relevant ~~documents~~ from relevant departments**

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	*wǒ* 我 *I*	*shàng* 尚 *yet*	*wèi* 未 *not*	*dédào* 得到 *receive*	*yǒuguān* 有关 *relevant*	*bùmén* 部门 *department*	*de* 的 *'s*	*huíyìng* 回应 *response*
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**

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	*wǒ* 我 *I*	*shàng* 尚 *yet*	*wèi* 未 *not*	*dédào* 得到 *receive*	*yǒuguān* 有关 *relevant*	*bùmén* 部门 *department*	*de* 的 *'s*	*huíyìng* 回应 *response*	
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**	

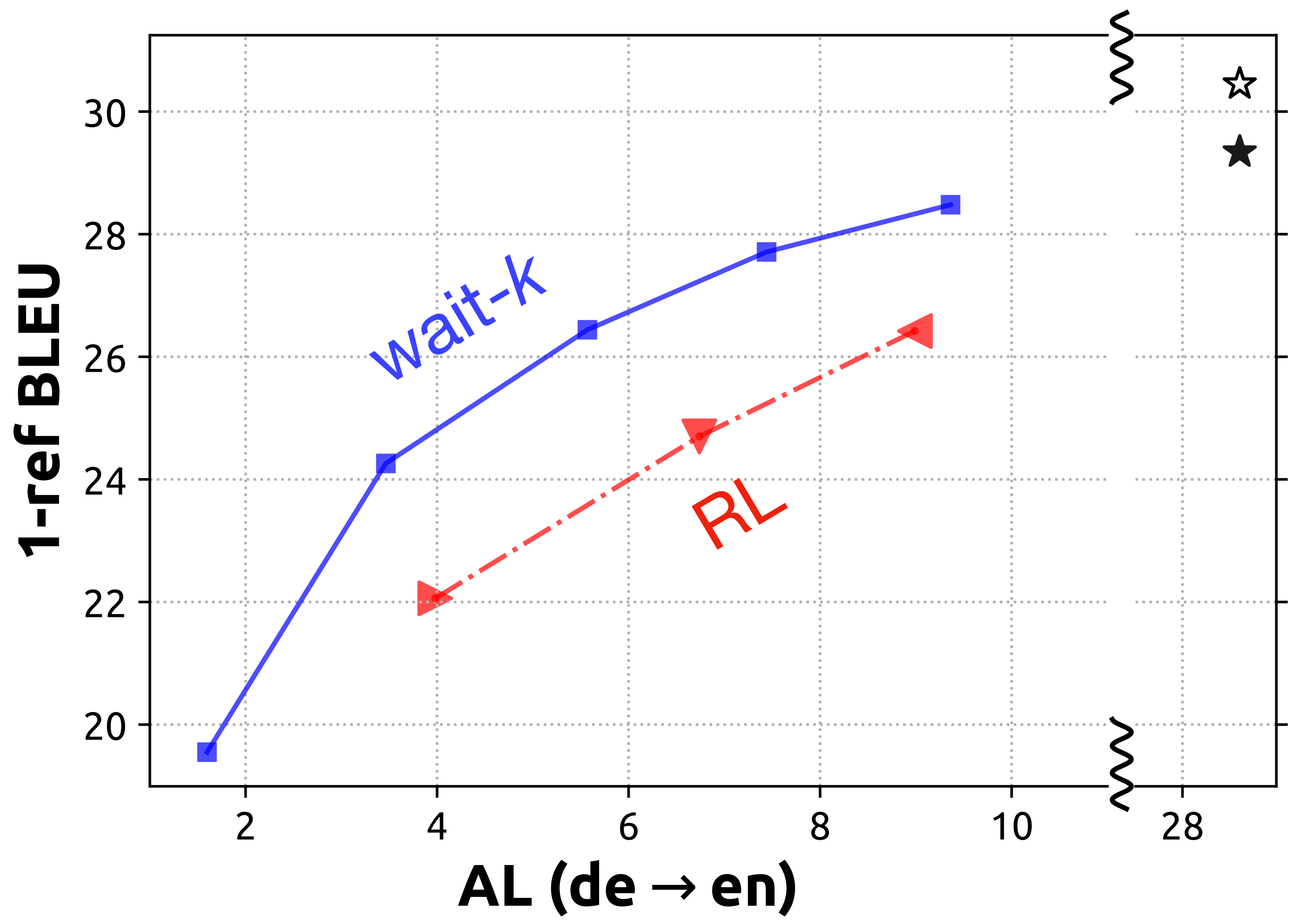
Previous Work on Adaptive Policy

- READ and WRITE actions

Previous Work on Adaptive Policy

- READ and WRITE actions

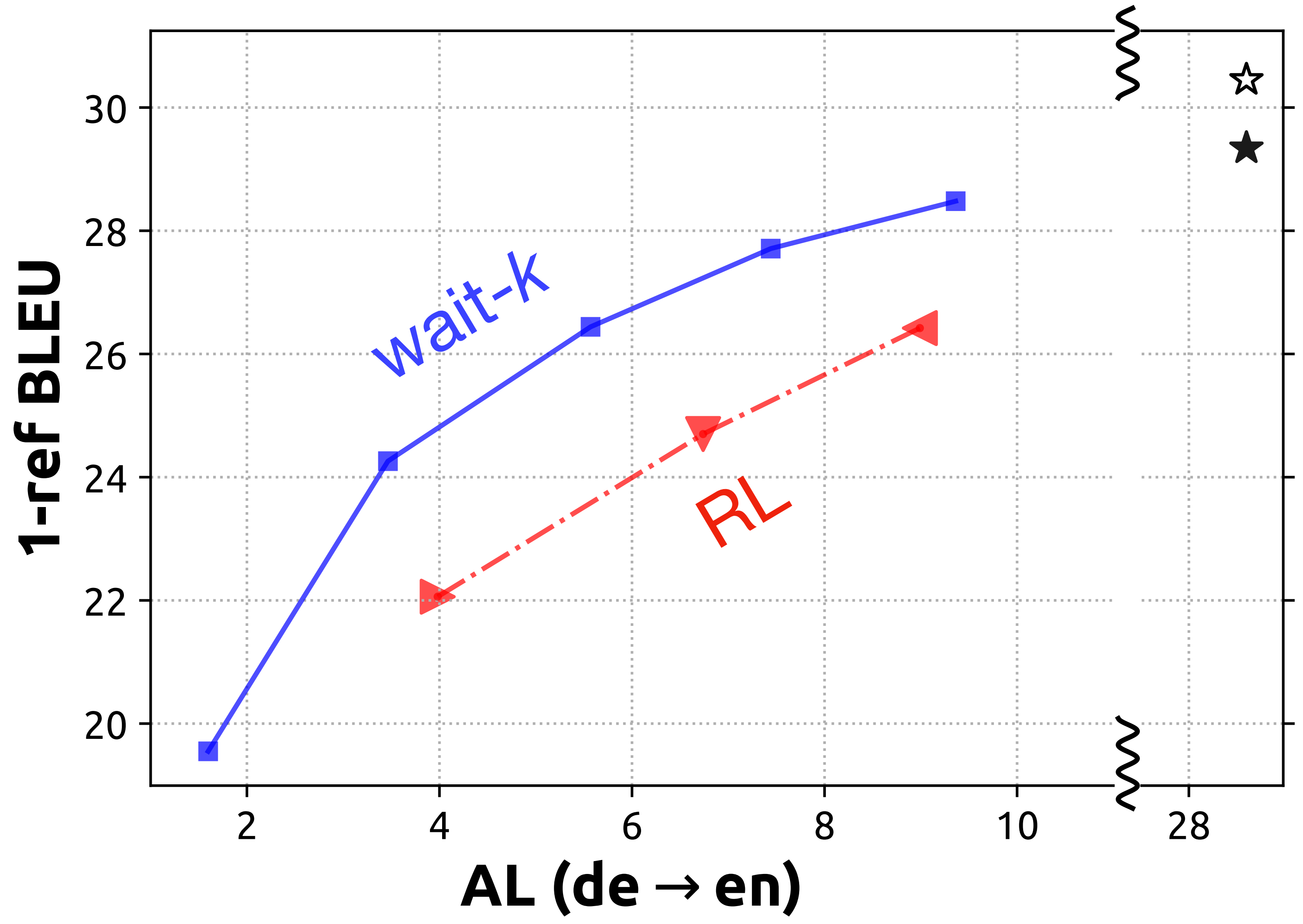
- sequential decision making ⟶ reinforcement learning (Gu et al. 2017)
  - unstable training (randomness in exploration)
  - complicated (two models trained in two stages)
  - worse performance (than wait-k model)



Previous Work on Adaptive Policy

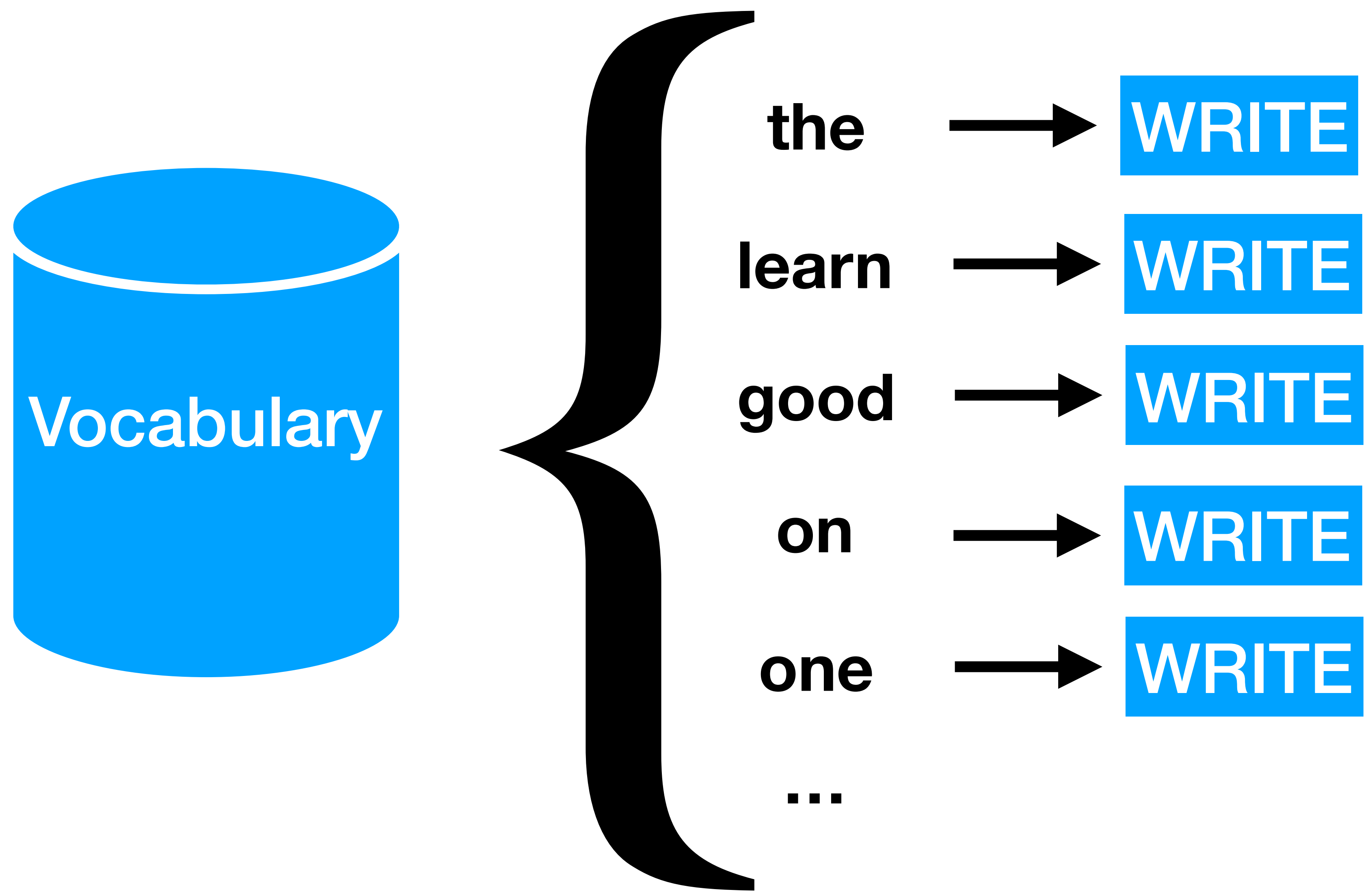
- READ and WRITE actions

- 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)

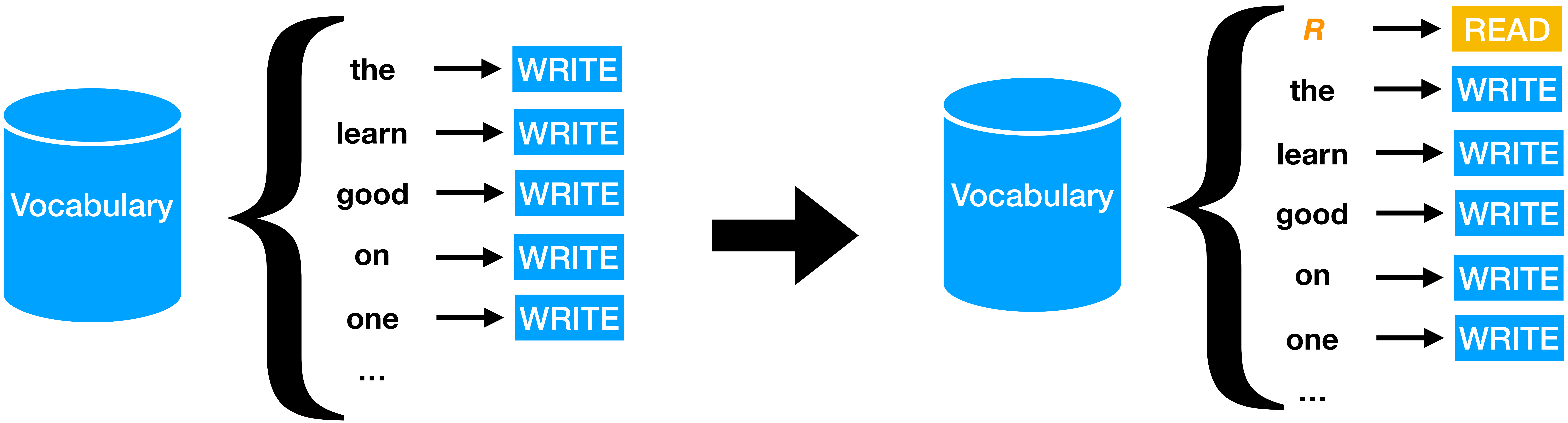


- can we learn a better model with adaptive policy via simpler methods ?

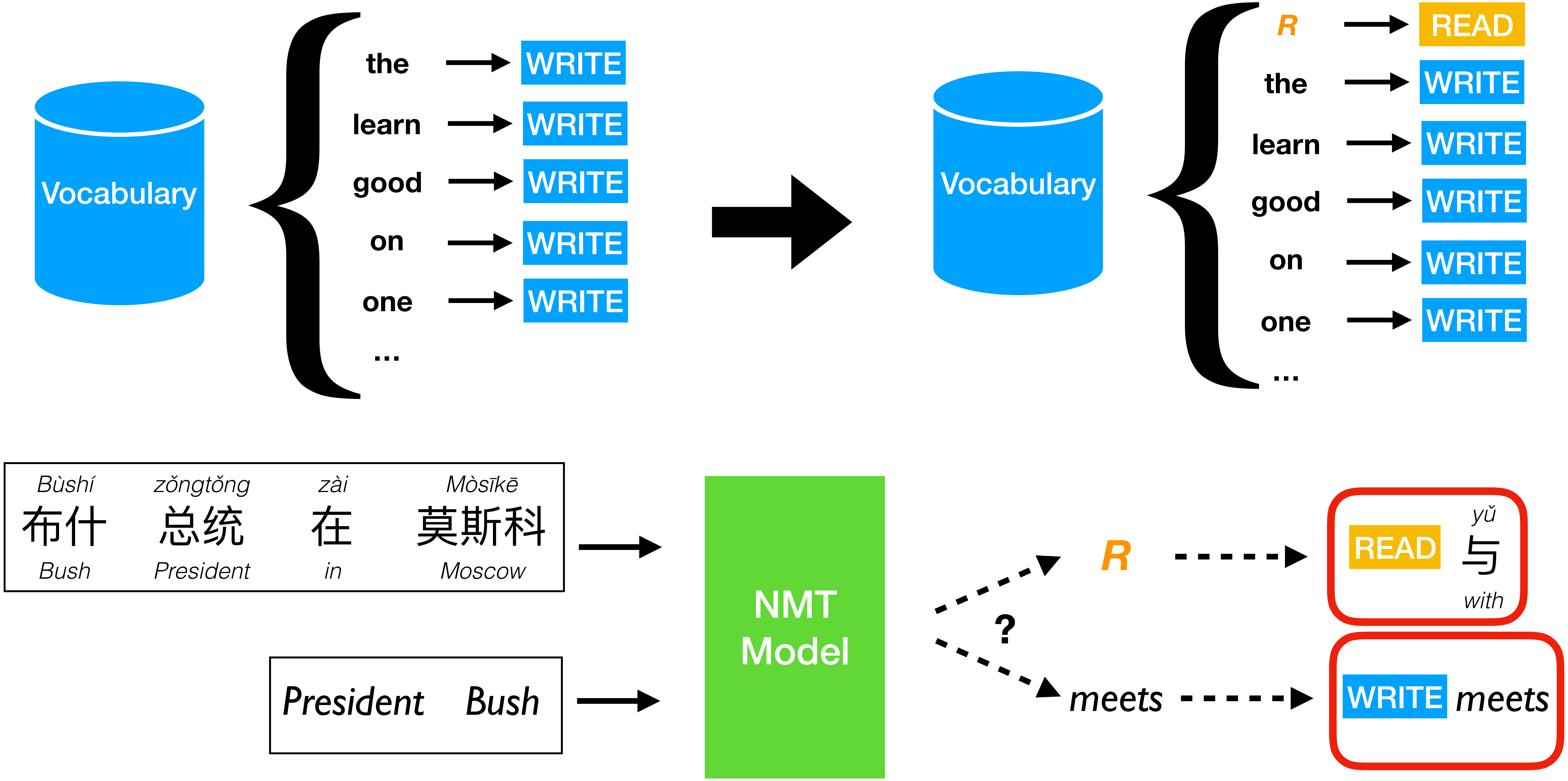
Our Idea: Single Model, with READ as a Word



Our Idea: Single Model, with READ as a Word



Our Idea: Single Model, with READ as a Word

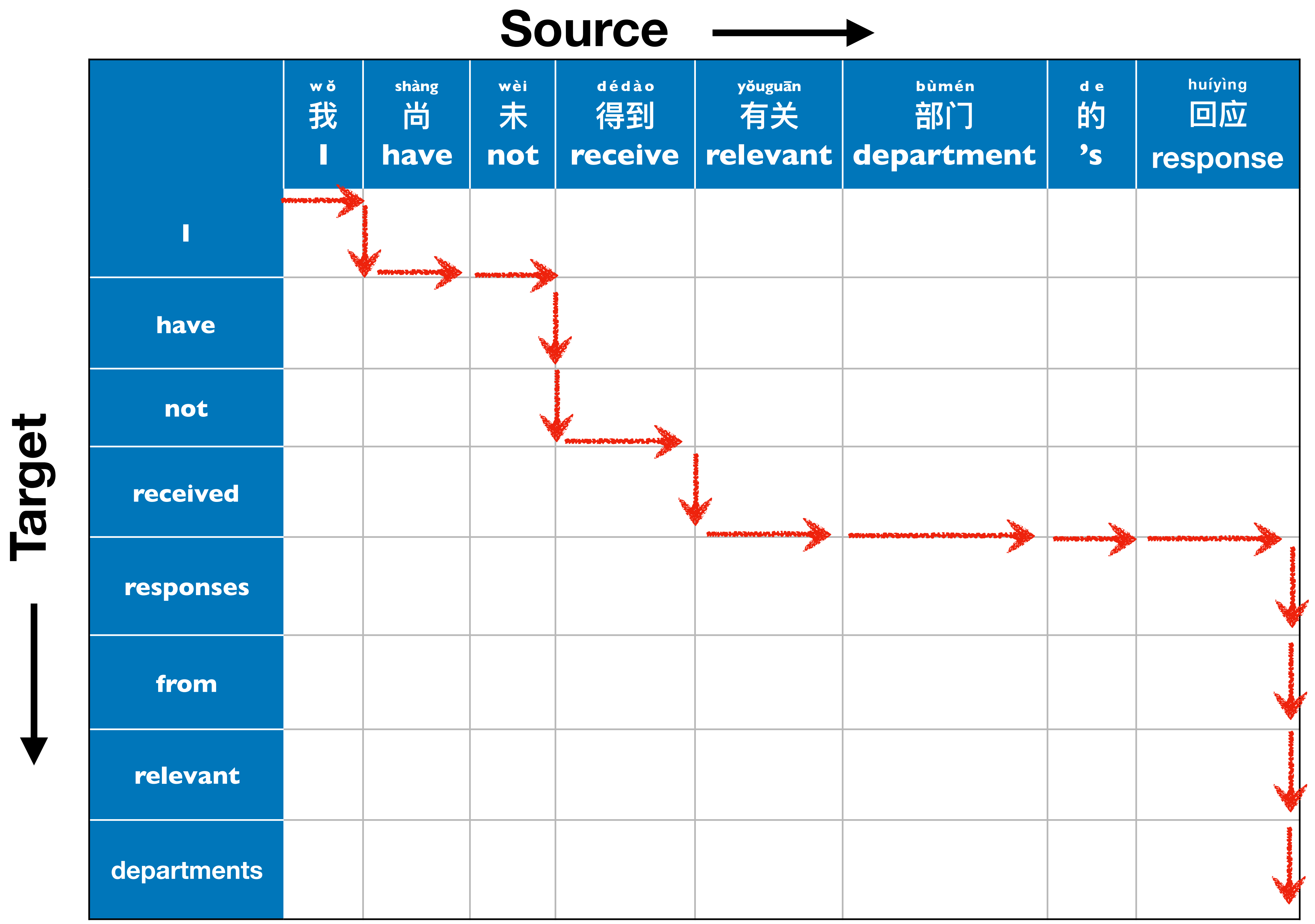


Learn a Single Model via Imitation Learning

- 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



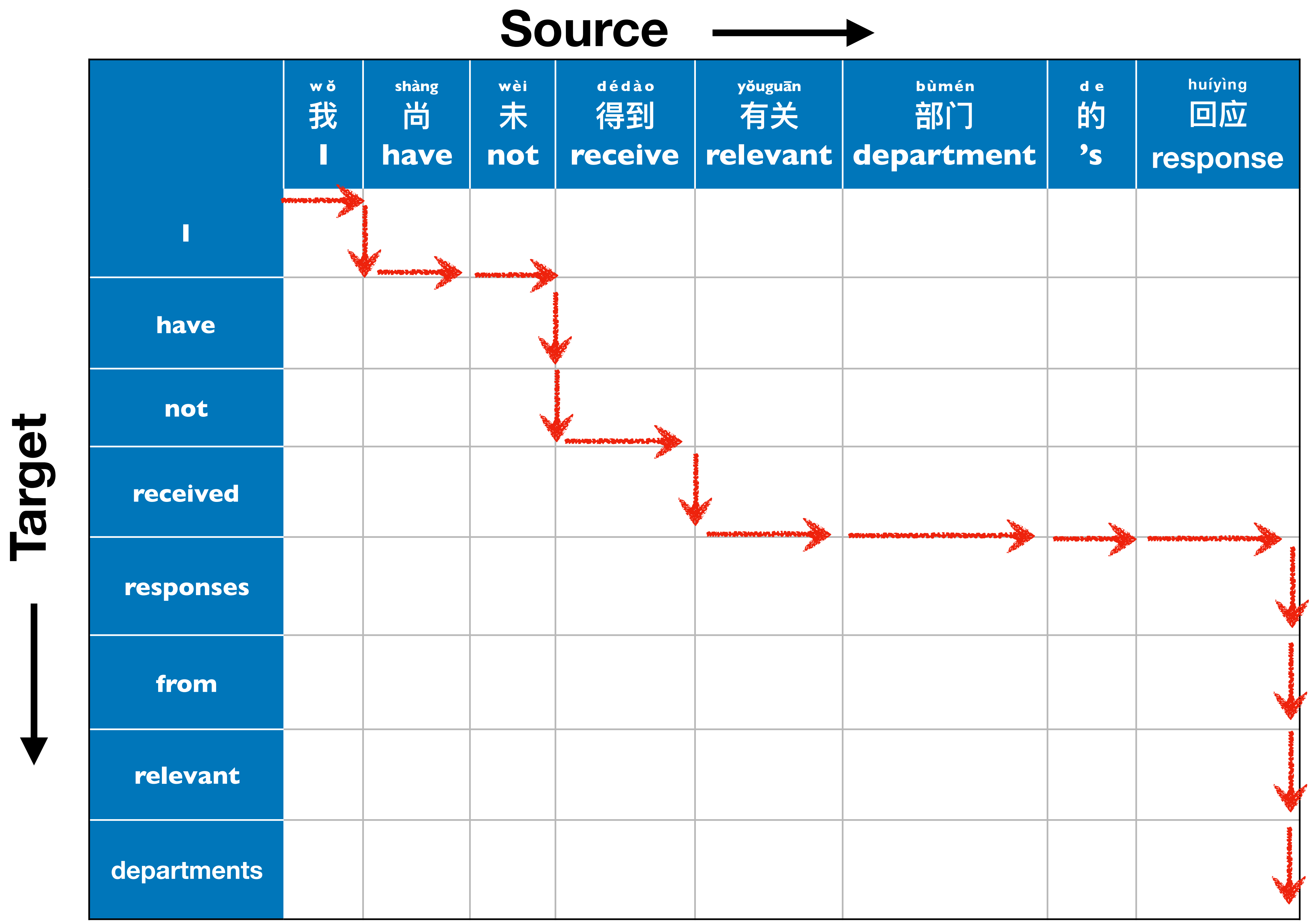
Learn a Single Model via Imitation Learning

- imitation learning
  - learn to imitate a given expert policy

- basic ideas

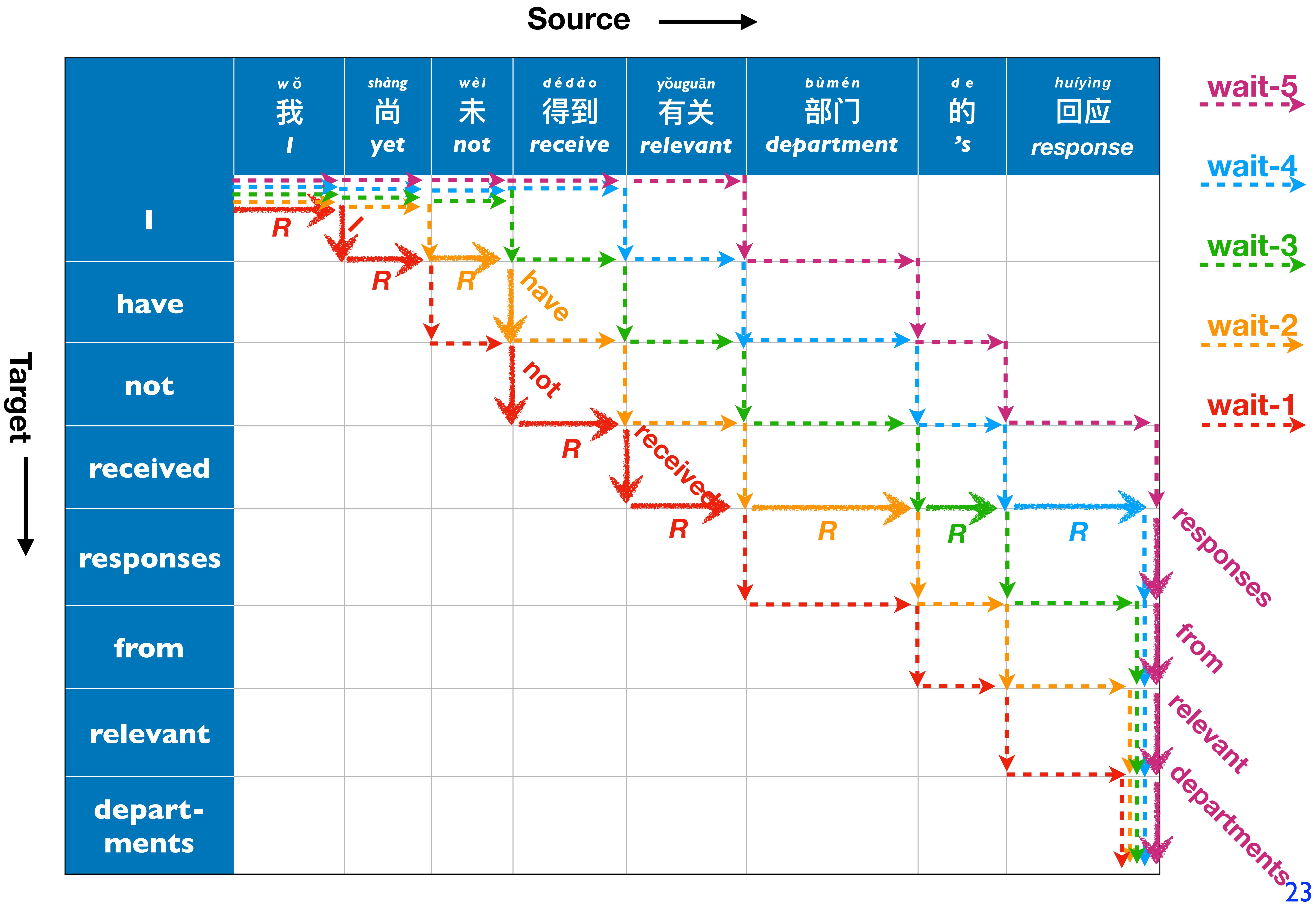
for more details
come to my short talk tomorrow

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



Another Much Simpler Idea

- on-the-fly decide READ or WRITE
  - depending on $p(y_i| \ldots)$
  - 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

Clean Input	目前已发现有109人死亡, 另有57人获救
Output of Transformer	at present, 109 people have been found dead and 57 have been rescued
Noisy Input	目前已发现又109人死亡, 另有57人获救
Output of Transformer	the hpv has been found dead so far and 57 have been saved
Output of Our Method	so far, 109 people have been found dead and 57 others have been rescued

yǒu
有
have

yòu
又
again

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 irrelvant ("hpv"). Our method, by contrast, is very robust to homophone noises thanks to phonetic information.

Baidu ASR's Code-Switching Capabilities

Baidu AI Create, July 2019

- Baidu ASR is awesome at code-switching (English terms in Chinese speech)

Baidu ASR's Code-Switching Capabilities

Baidu AI Create, July 2019

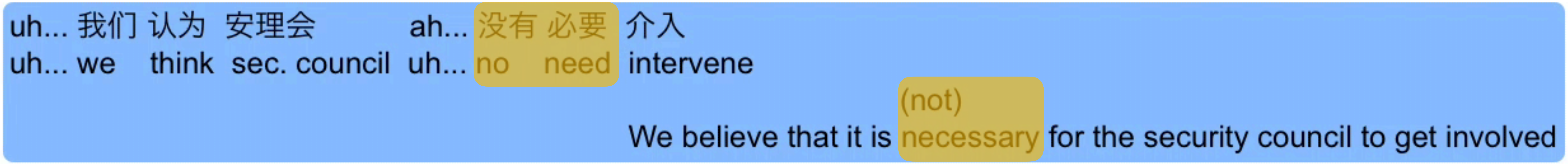
- Baidu ASR is awesome at code-switching (English terms in Chinese speech)

Baidu ASR's Code-Switching Capabilities

- Baidu ASR is awesome at code-switching (English terms in Chinese speech)

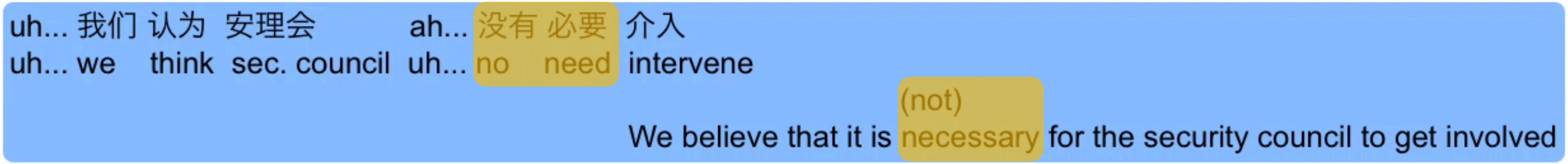
Better Dataset for Training Simultaneous Translation

- 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)



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)

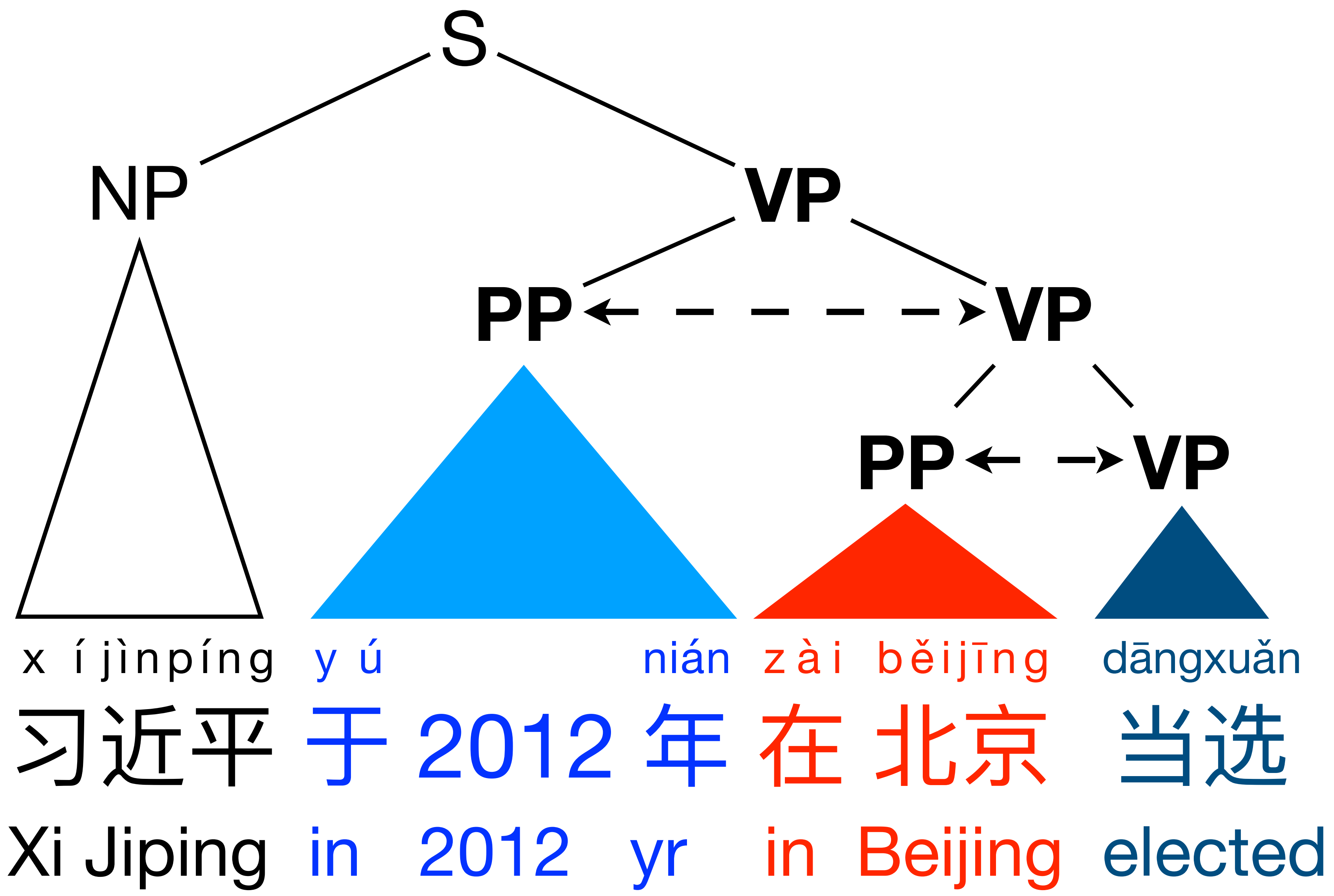


Better Dataset for Training Simultaneous Translation

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

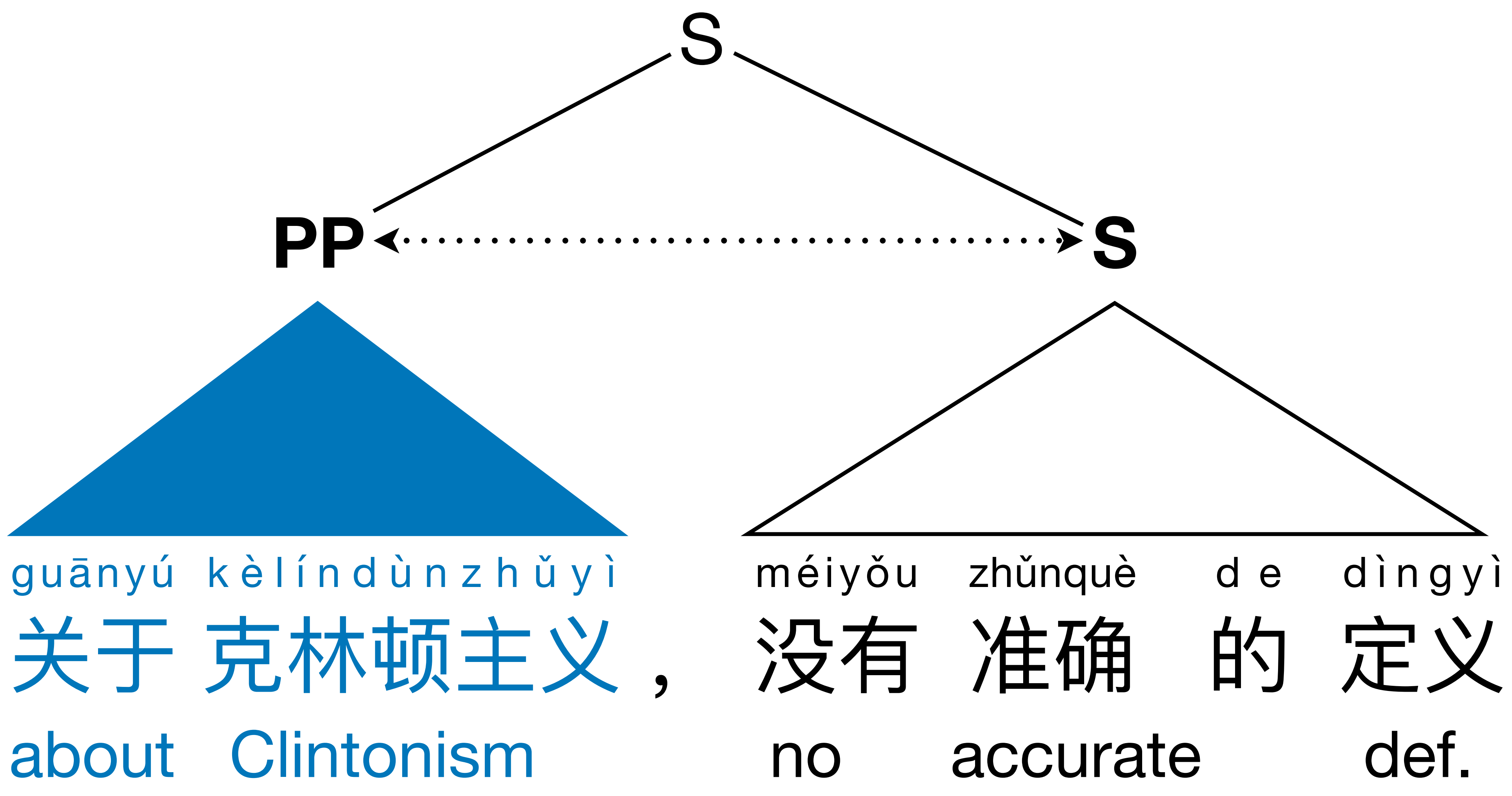
mandatory reordering

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



optional reordering

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



reference translation “Xi Jinping was elected in Beijing in 2012”

“There is no accurate definition of Clintonism.”

ideal => About Clintonism, there is no accurate definition.

see also He et al (2015)

Detecting and Fixing Mistakes

- idea: use a slower policy to verify the current policy's output along the way

	1	2	3	4	5	6	7	
	Bùshí	***zǒngtǒng***	***zài***	Mòsīkē	*cānjiā*	*zhòngyào*	*fēnghuì*	
	布什	总统	在	莫斯科	参加	重要	峰会	
	Bush	**president**	**in**	**Moscow**	**attend**	important	summit	
wait-2 decode			president	bush	met			
wait-3 check				president	bush	met		
				0.9	0.8	0.0		
revision						**I mean** *attended*		
wait-3 continue							an	important summit in moscow

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)