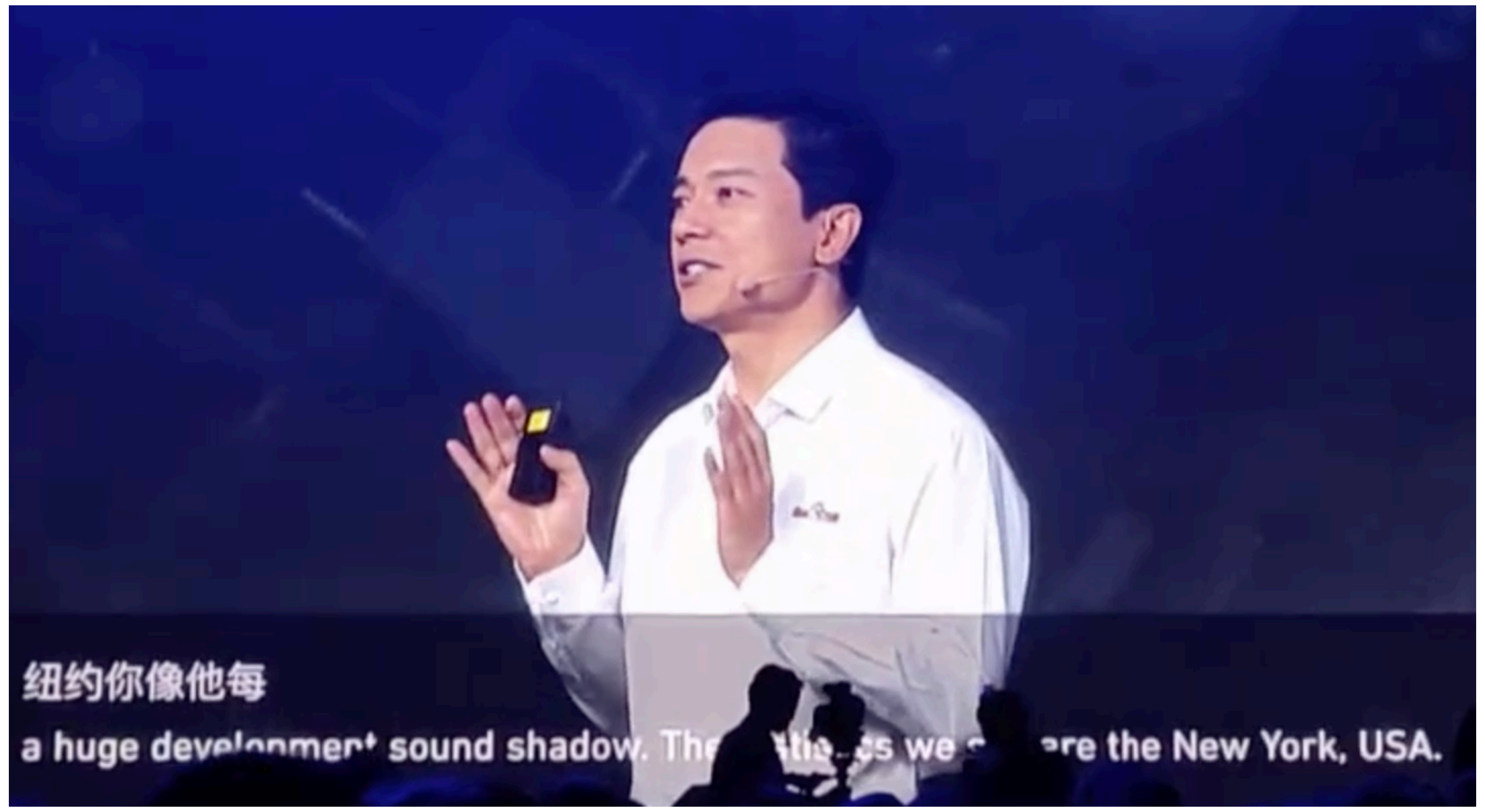
STACL: Simultaneous Translation with Integrated Anticipation & Controllable Latency



Liang Huang

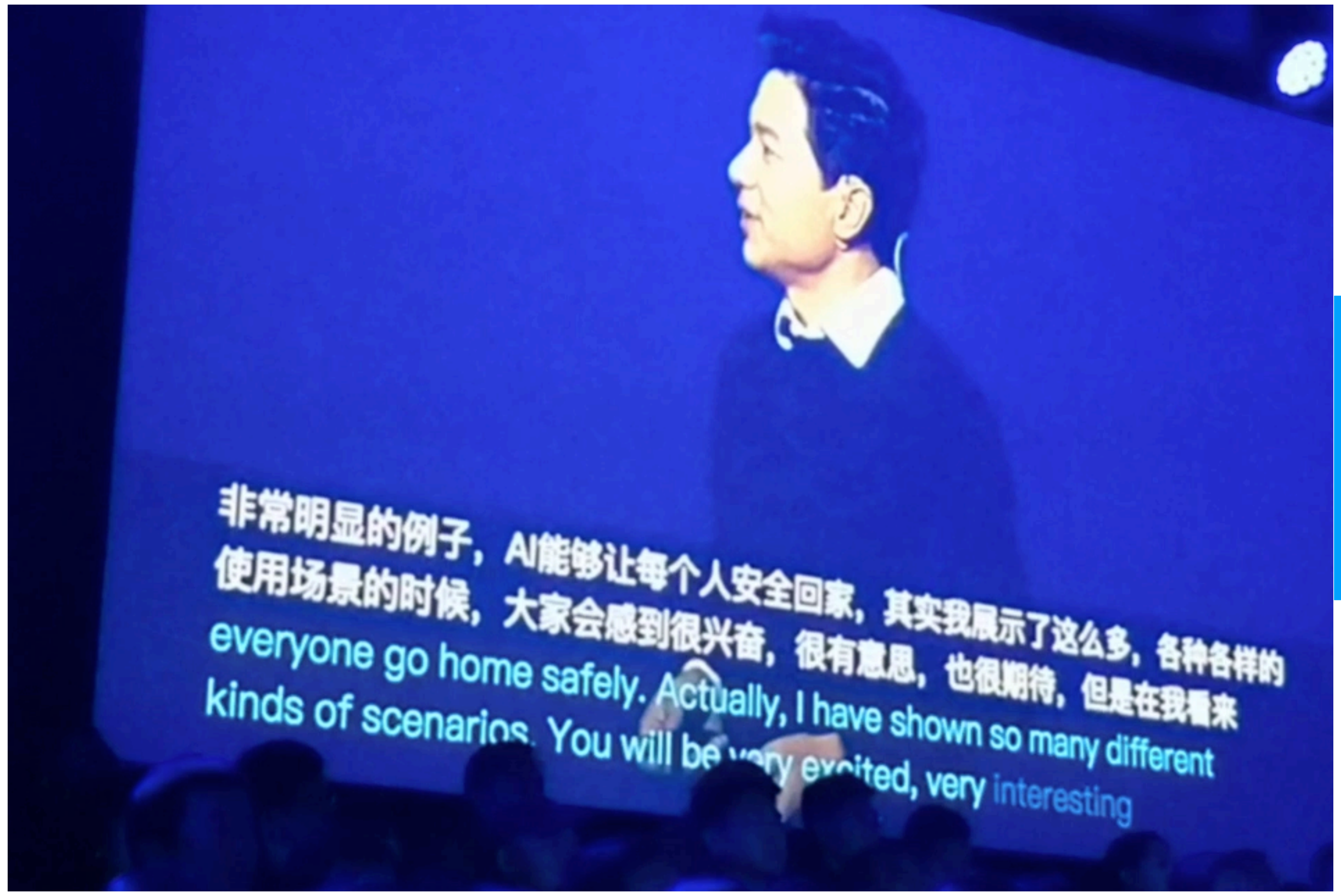
Principal Scientist, Baidu Research

Assistant Professor (on-leave), Oregon State University

Joint work between Baidu Research (Sunnyvale) and Baidu NLP (Beijing)

Breakthrough in Simultaneous Translation

full-sentence (non-simultaneous) translation



Baidu World Conference, November 2017

STACL

simultaneous translation, latency ~3 secs

Baidu World Conference, November 2018

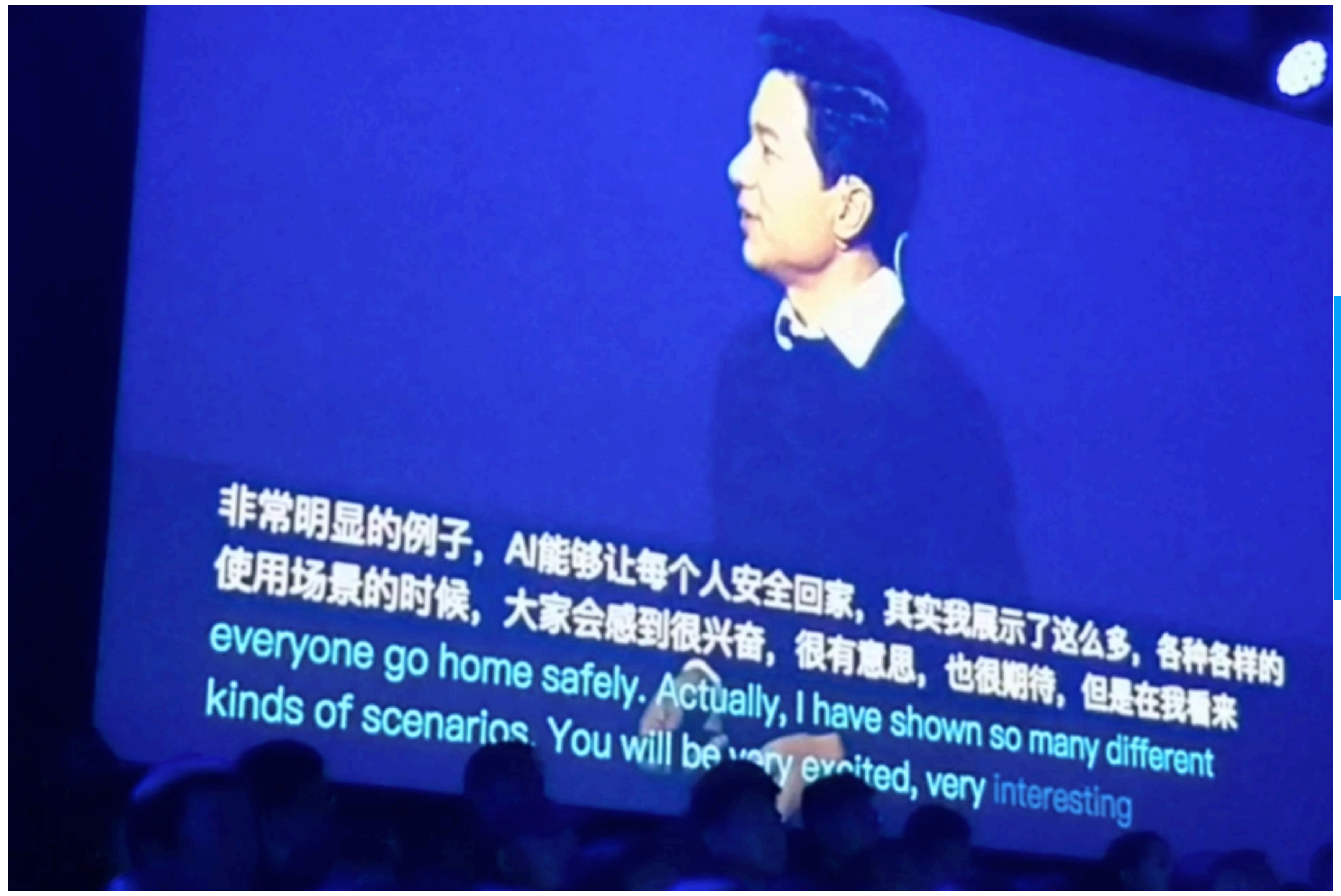
Media coverage:

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

量子位, 机器之心, 中国新闻網, 鳳凰網科技, sina 新浪财经, 环球网智能, 雷锋网, 博客园, AiTechYun, 前瞻网, 动点科技, IT经理网

Breakthrough in Simultaneous Translation

full-sentence (non-simultaneous) translation



Baidu World Conference, November 2017

STACL

simultaneous translation, latency ~3 secs

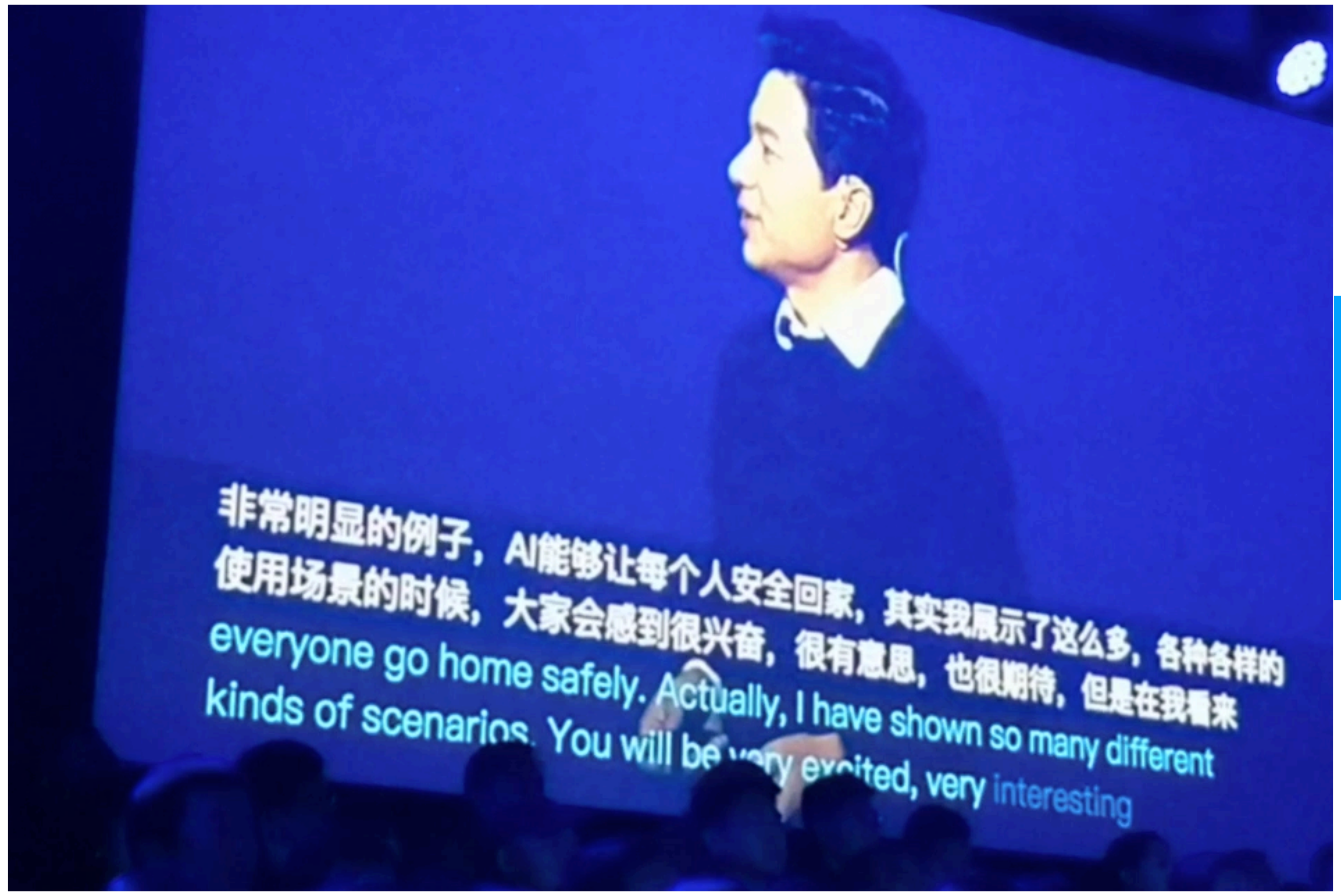
Baidu World Conference, November 2018

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 · 量子位 · 机器之心 · 中国新闻网 · 鳳凰網科技 · sina 新浪财经 · 环球网智能 · 雷锋网 · 博客园 · AiTechYun · 前瞻网 · 动点科技 · IT经理网

Breakthrough in Simultaneous Translation

full-sentence (non-simultaneous) translation



Baidu World Conference, November 2017

STACL

simultaneous translation, latency ~3 secs

Baidu World Conference, November 2018

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

量子位 · 机器之心 · 中国新闻网 WWW.CHINANEWS.COM · 鳳凰網科技 · sina 新浪财经 finance.sina.com.cn · 环球网 智能 smart.huanqiu.com · 雷锋网 · 博客园 cnblogs.com · AiTechYun · 前瞻网 qianzhan.com · 动点科技 · IT经理网 CTOCIO.com

Background: Consecutive vs. Simultaneous

consecutive interpretation
multiplicative latency (x2)

simultaneous interpretation
additive latency (+3 secs)

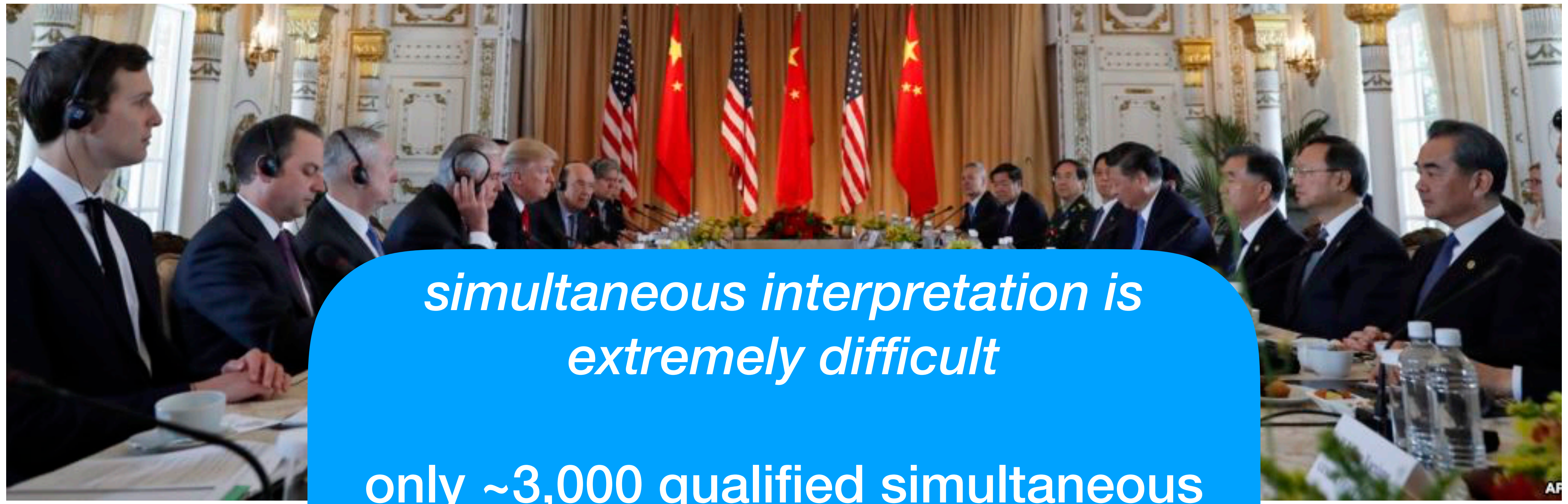
Background: Consecutive vs. Simultaneous

consecutive interpretation

multiplicative latency (x2)

simultaneous interpretation

additive latency (+3 secs)

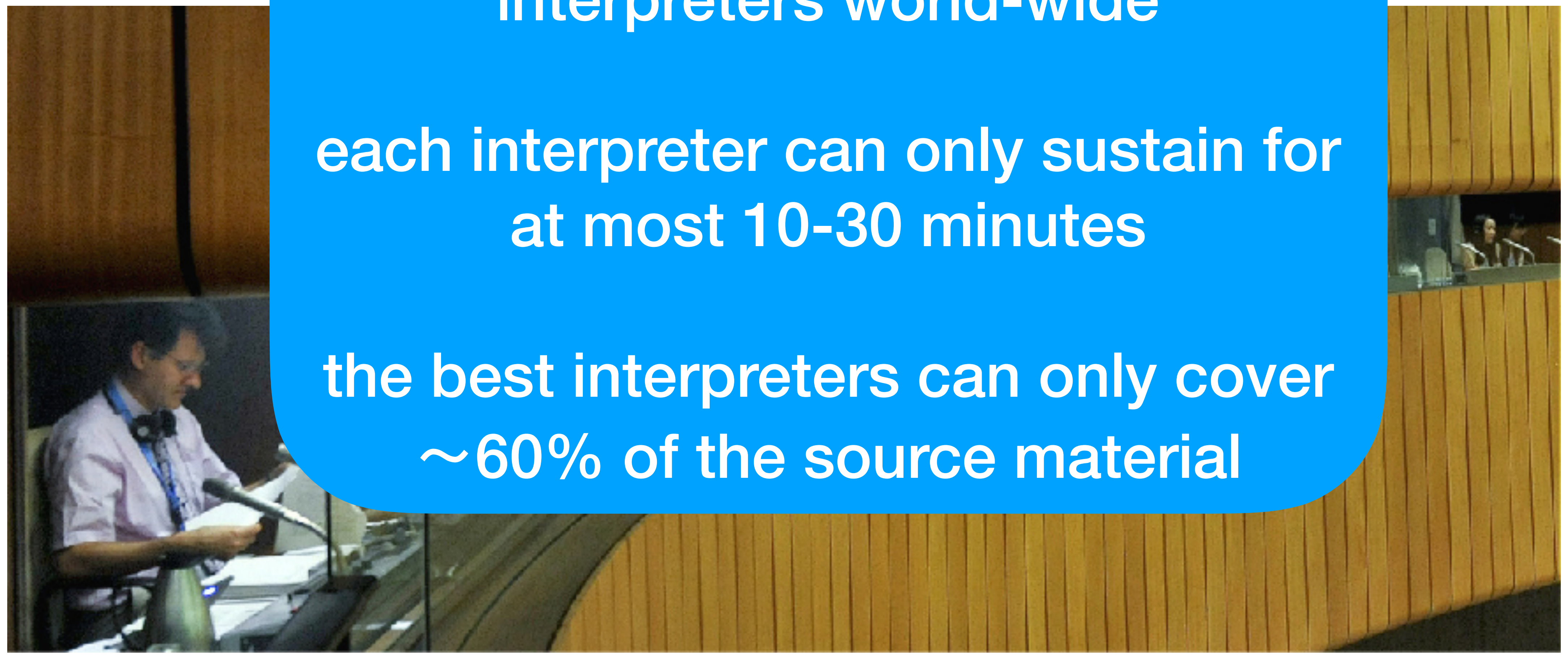


simultaneous interpretation is extremely difficult

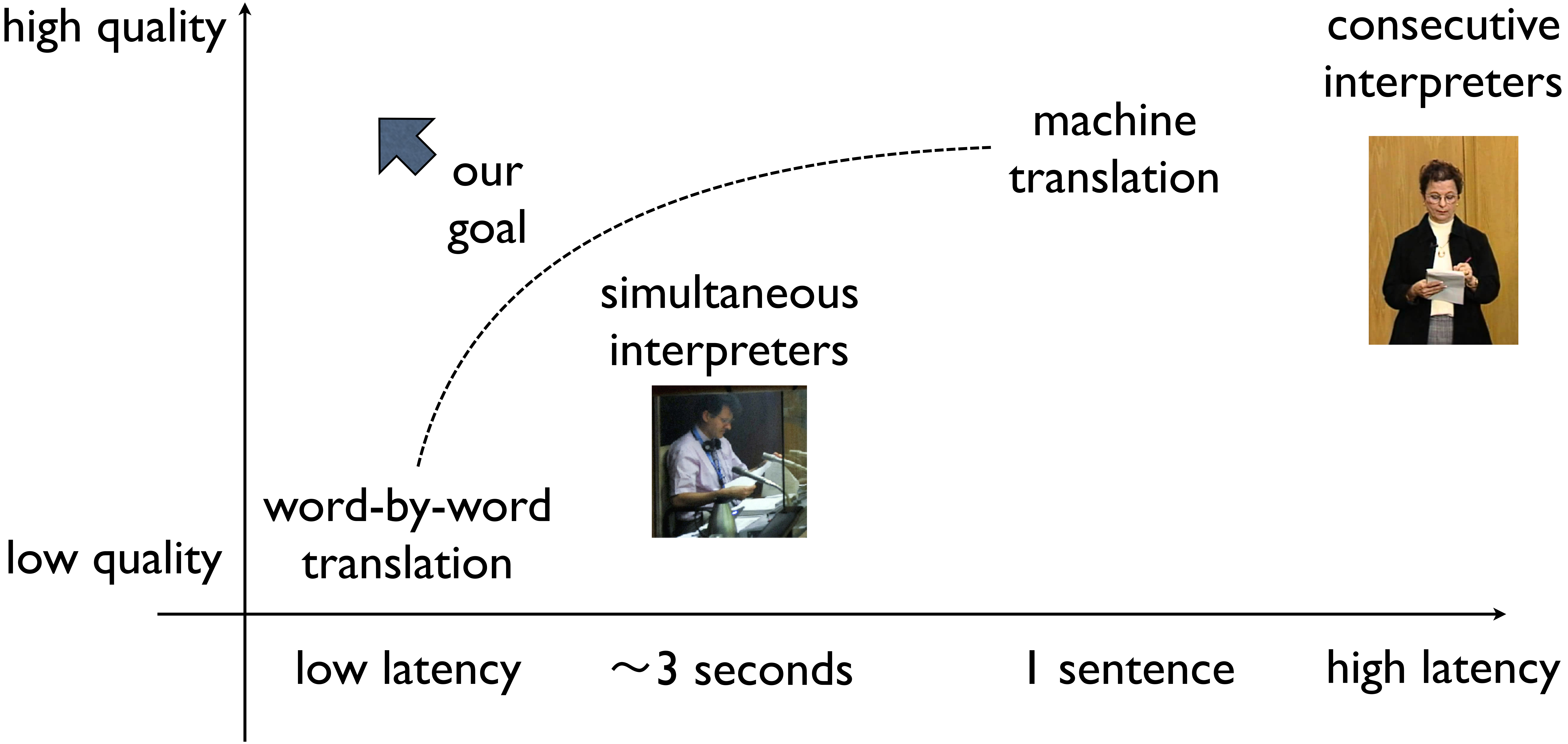
only ~3,000 qualified simultaneous interpreters world-wide

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

the best interpreters can only cover ~60% of the source material

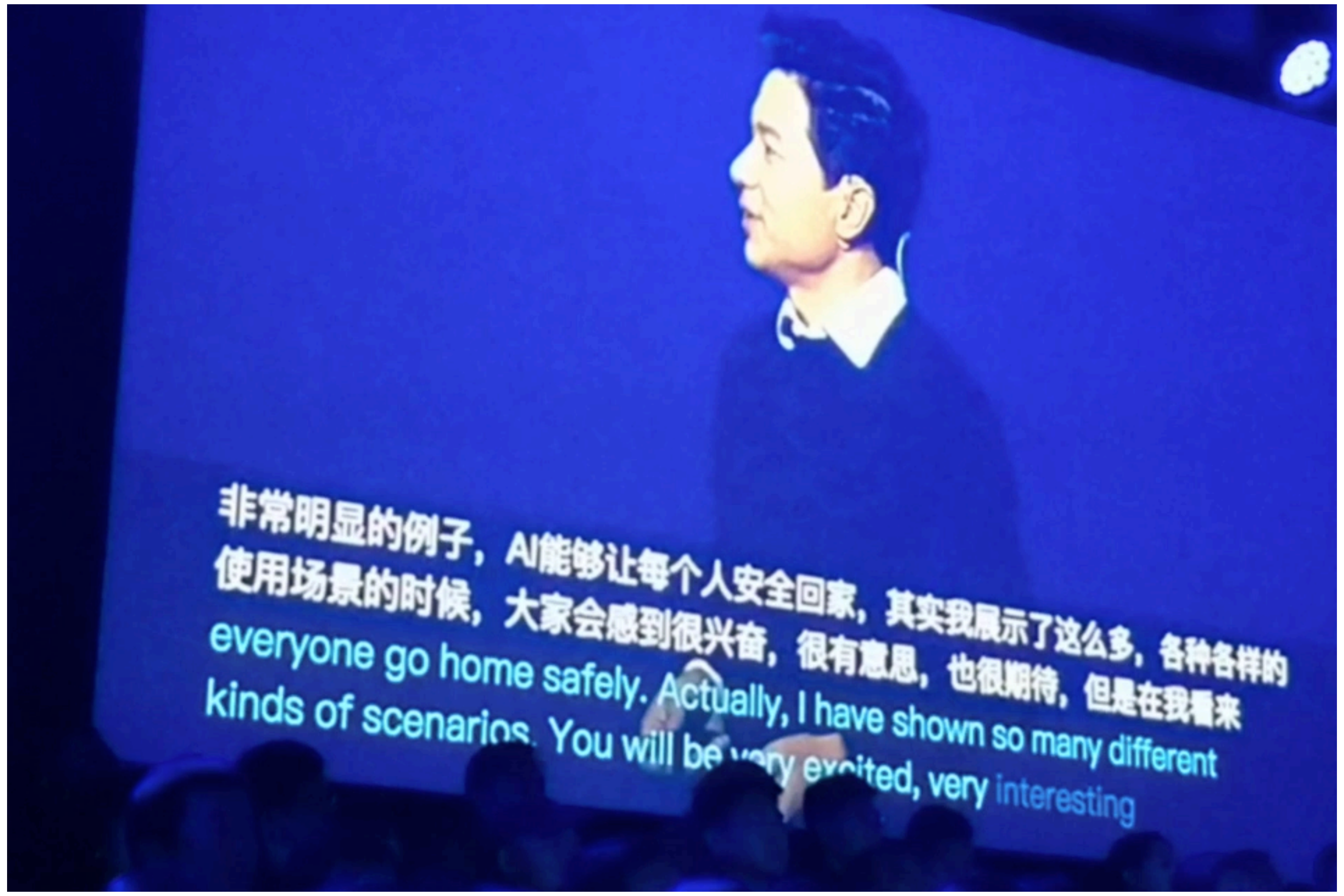


Tradeoff between Latency and Quality



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

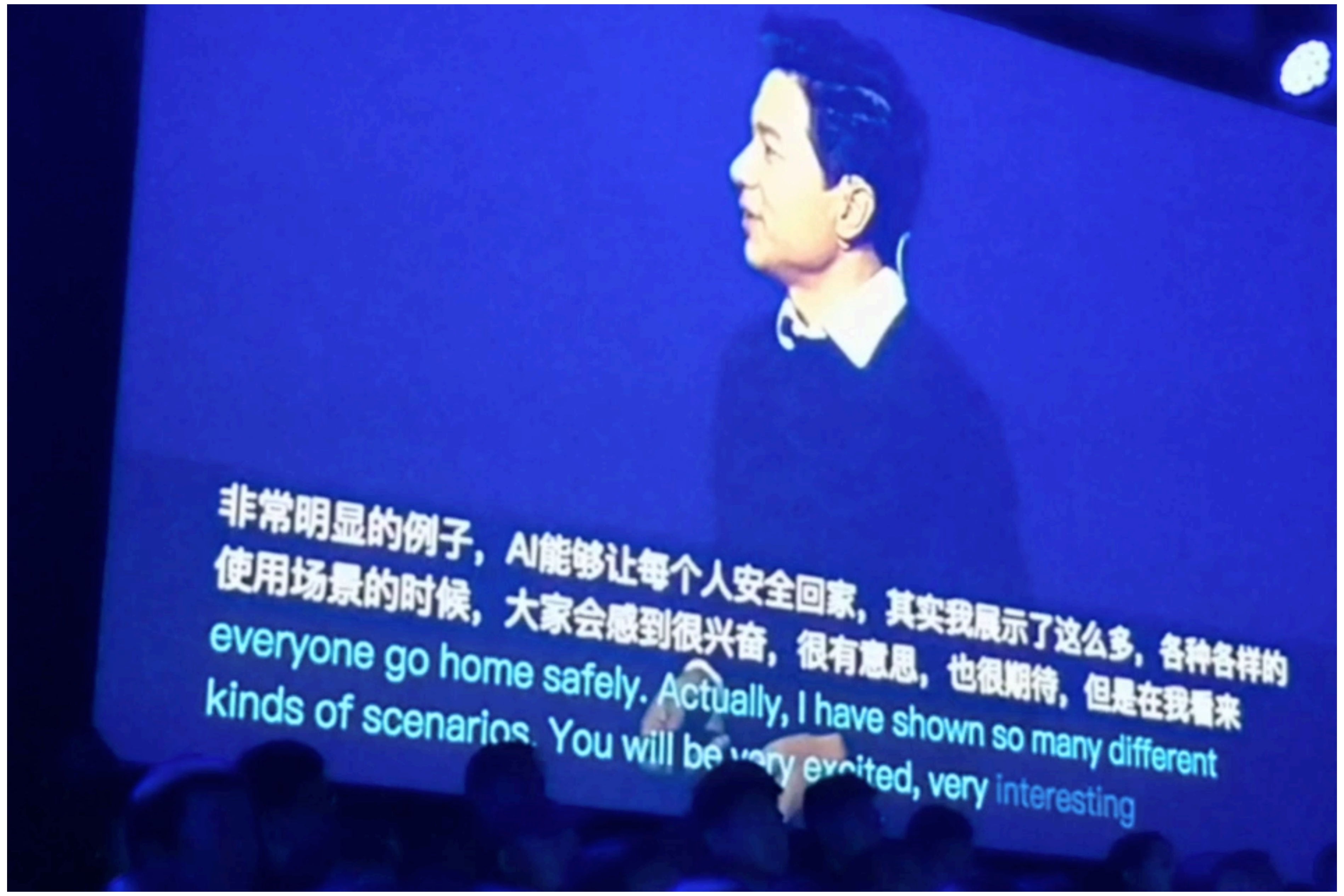


Baidu, Nov. 2017 (~12 seconds delay)

Sougou, Oct. 2018 (~12 seconds delay)

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

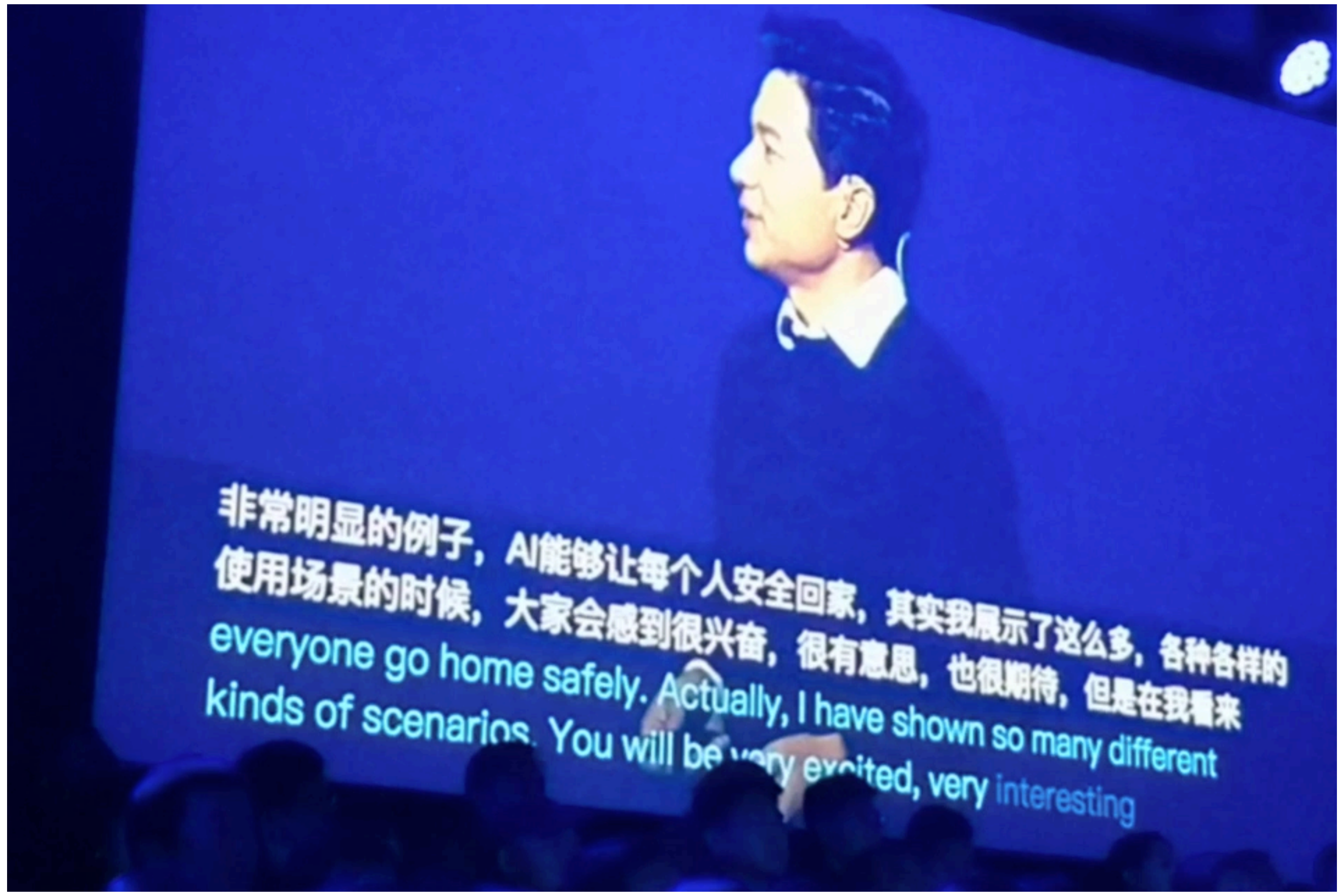


Baidu, Nov. 2017 (~12 seconds delay)

Sougou, Oct. 2018 (~12 seconds delay)

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 pretained NMT model
  - “encourages” latency requirements, but can’t force them in testing
  - complicated, and slow to train

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

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)

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

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)

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

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)

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 …

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)

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*

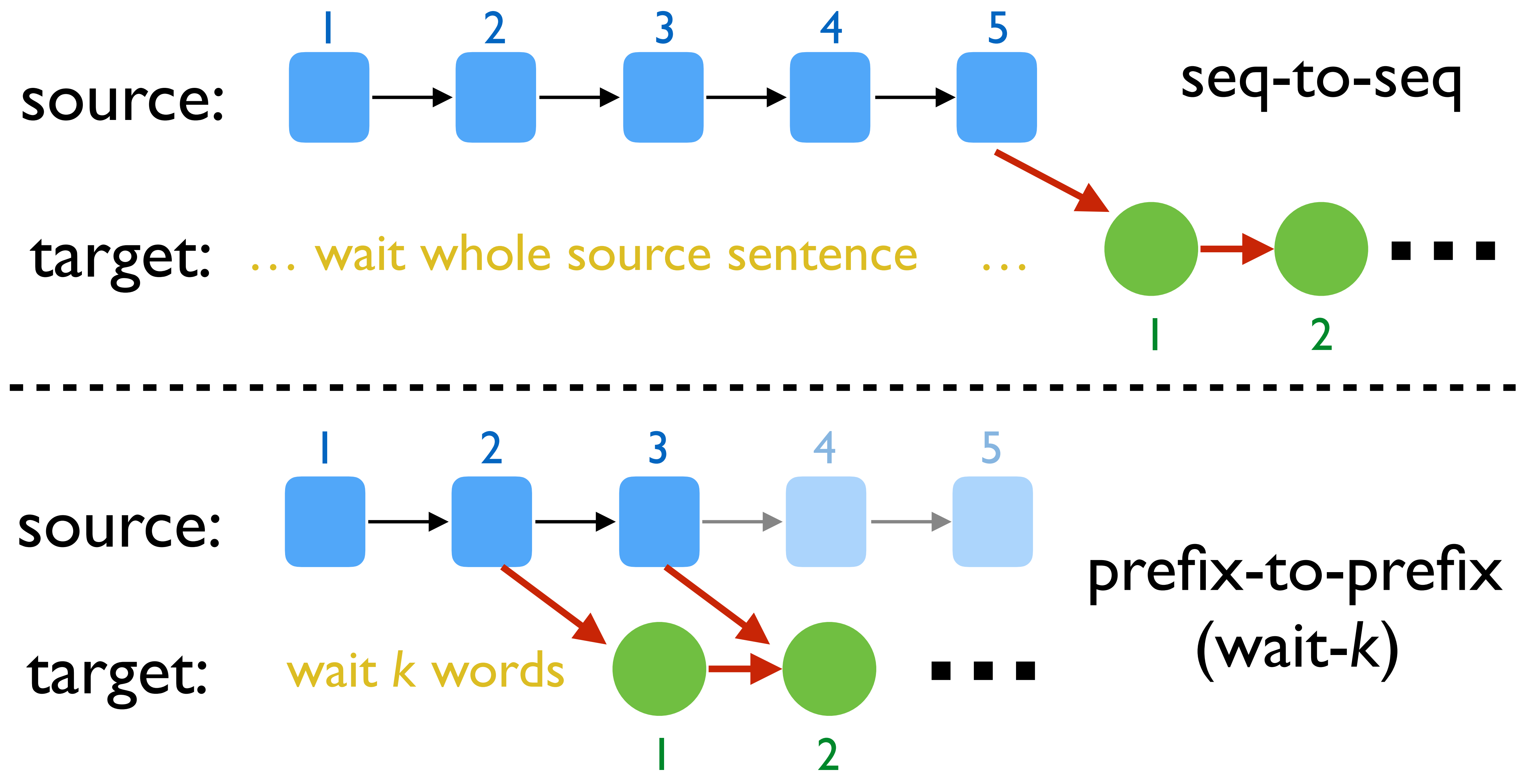
President Bush meets with Russian President Putin in Moscow

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

anticipative: President Bush meets with Russian President Putin in Moscow

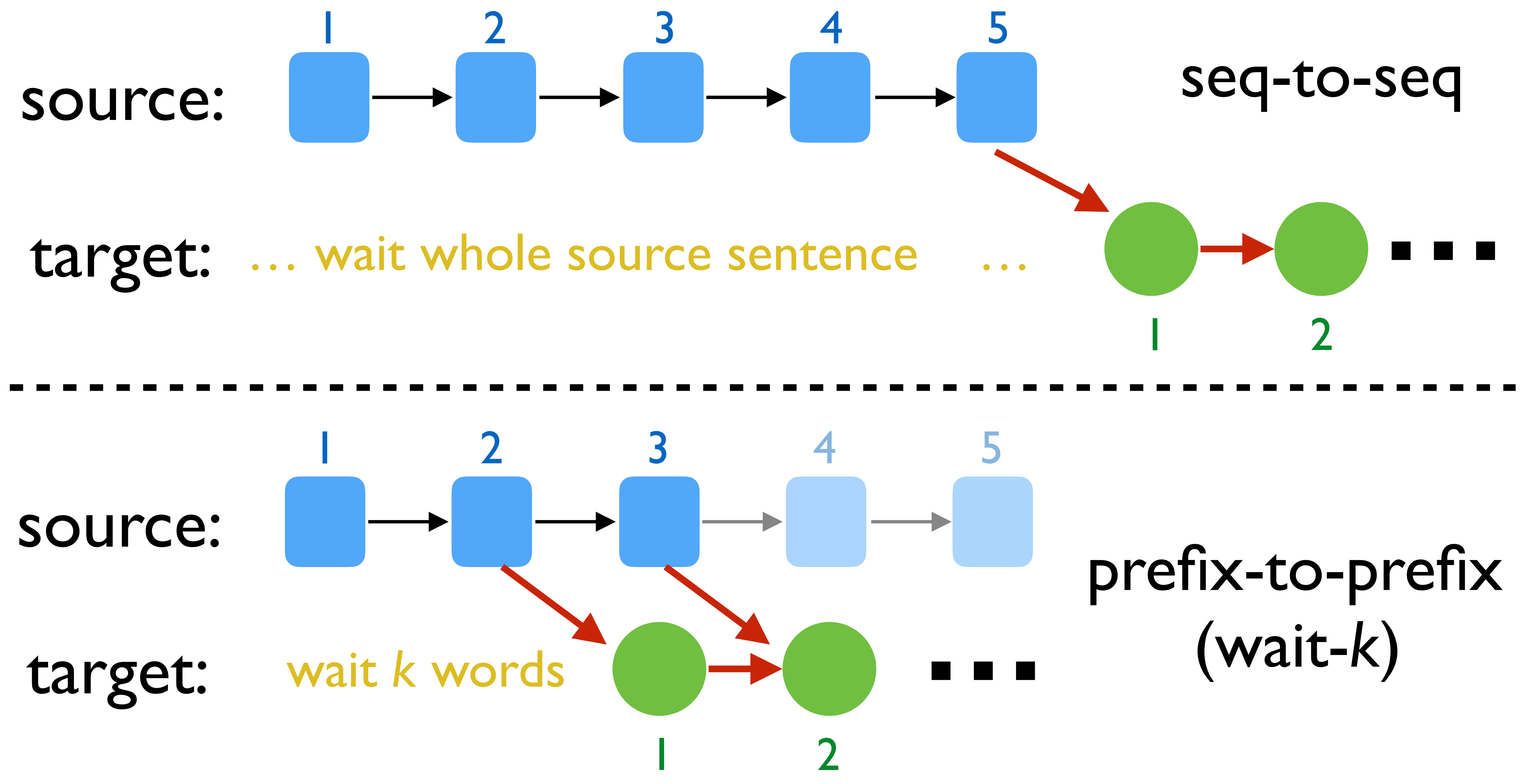
Our Solution: Prefix-to-Prefix

- 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

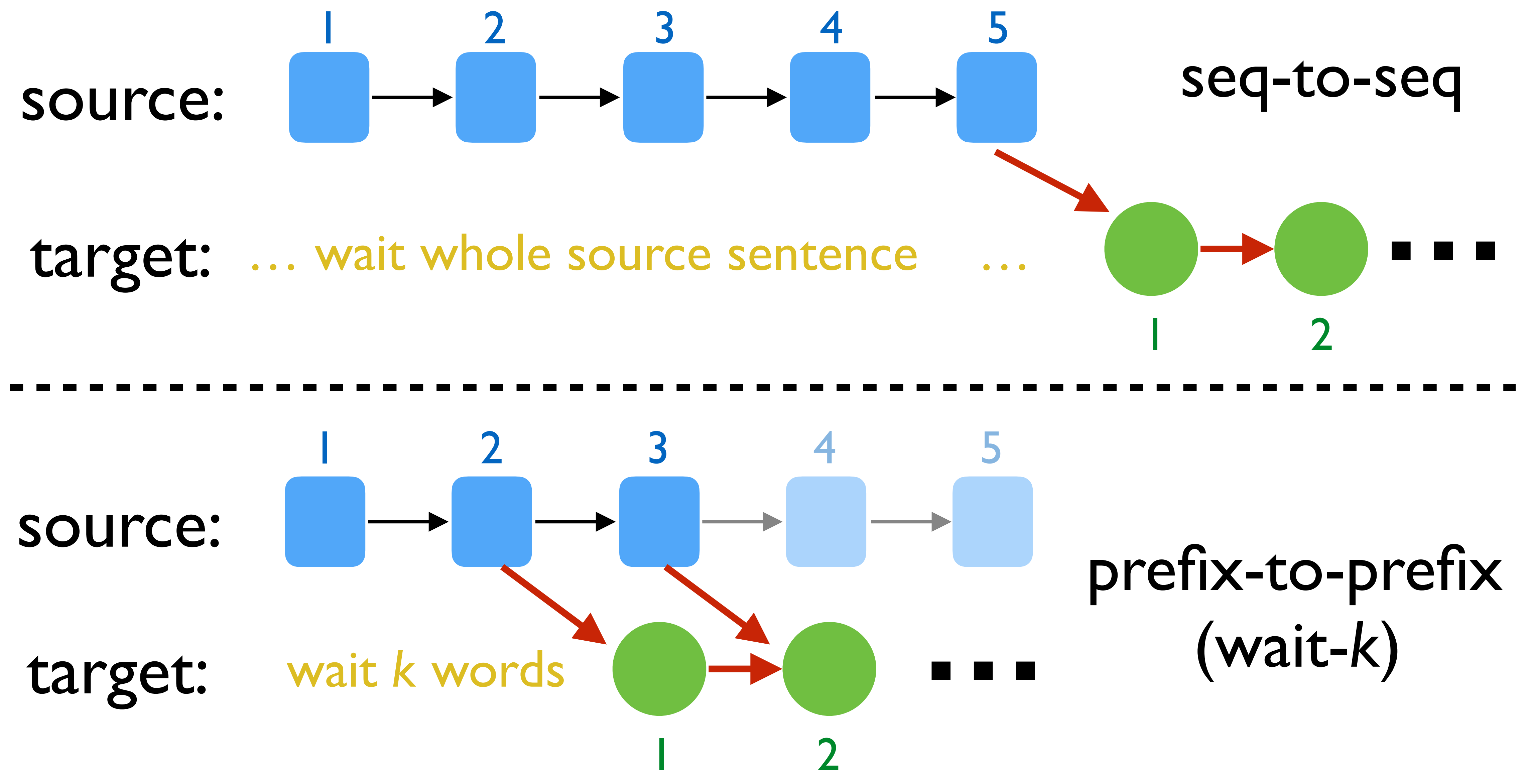


Bùshí	*zǒngtǒng*
布什	总统
Bush	*President*

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

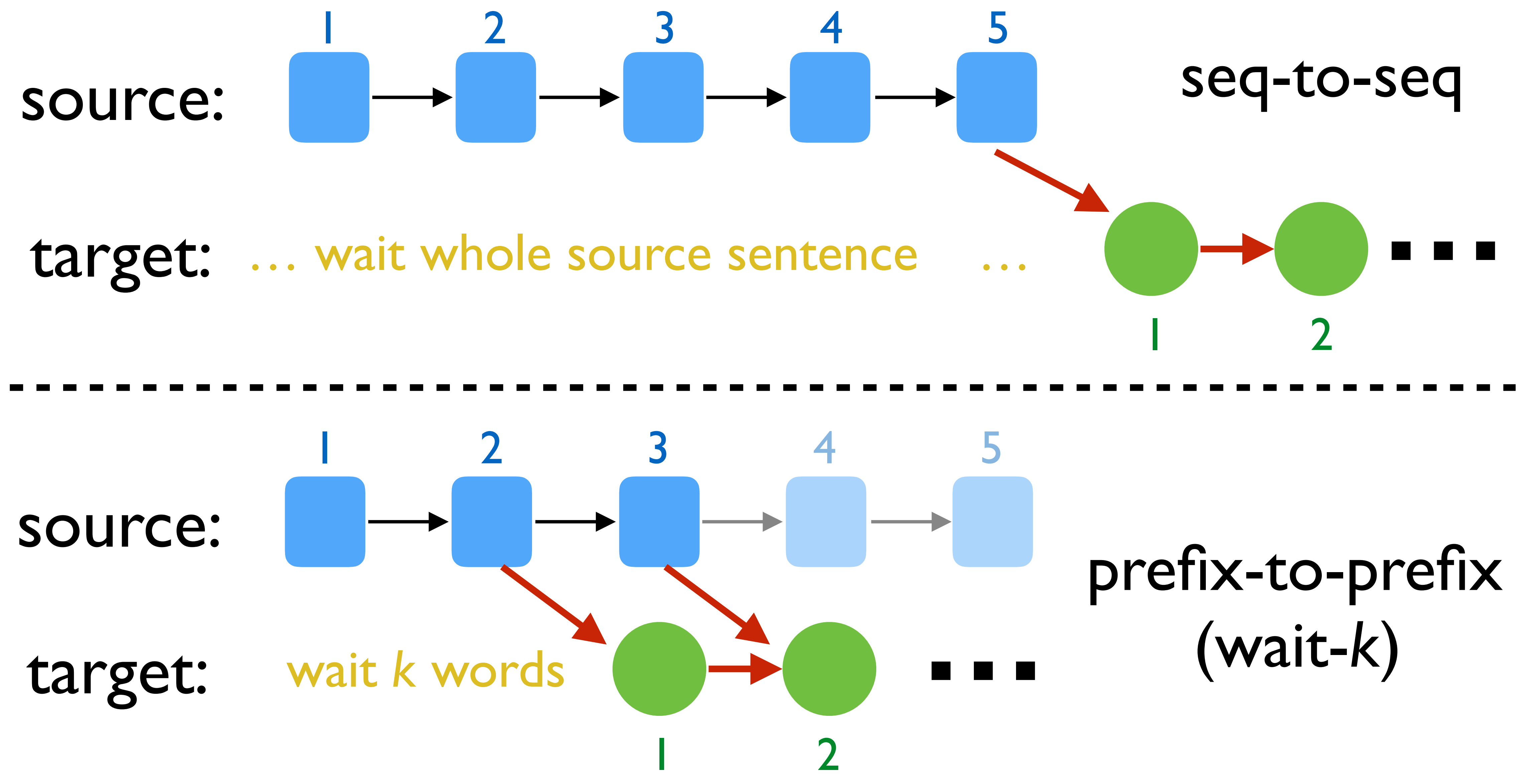


Bùshí	*zǒngtǒng*	*zài*
布什	总统	在
Bush	*President*	*in*

President Bush

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

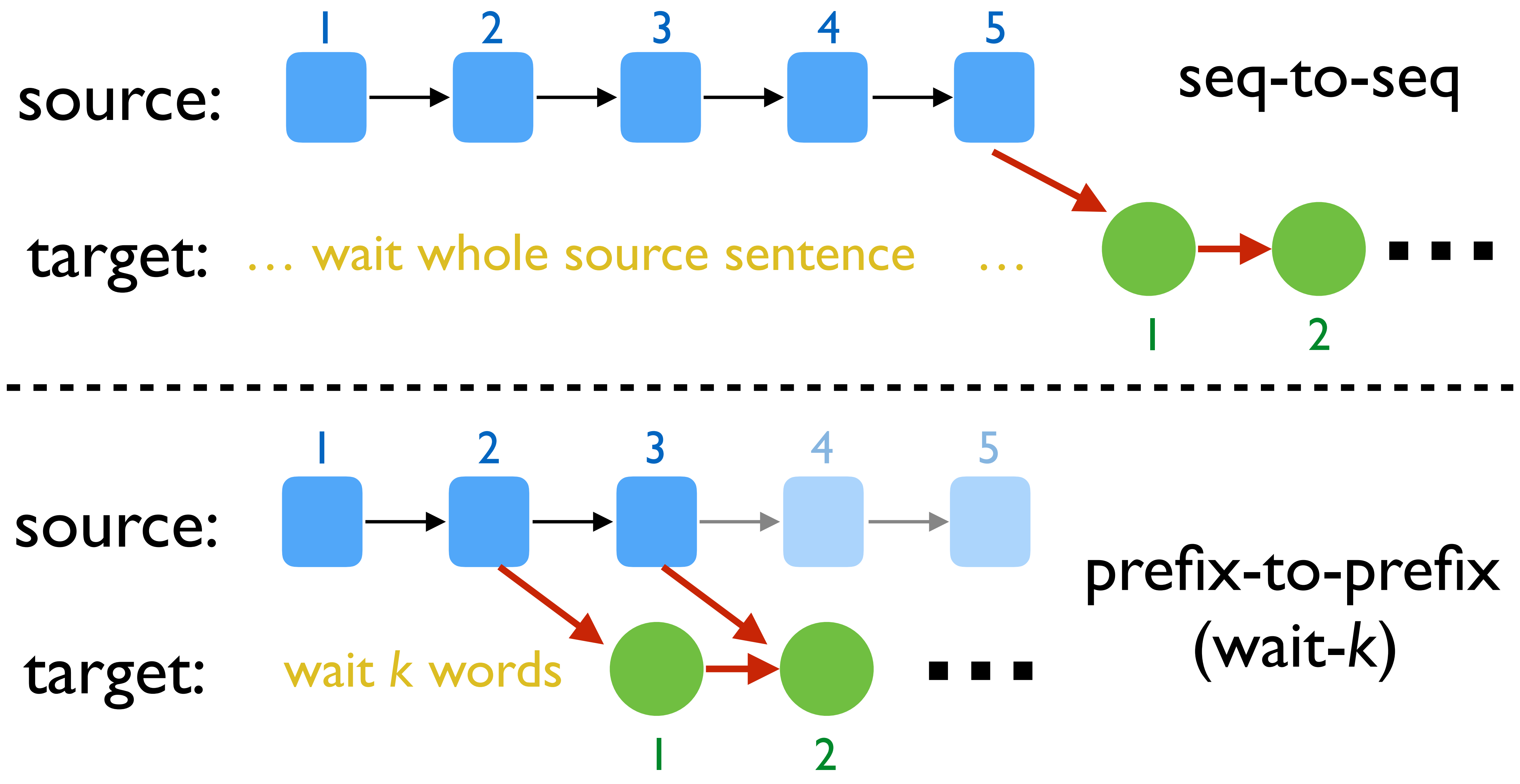


Bùshí	*zǒngtǒng*	*zài*	*Mòsīkē*
布什	总统	在	莫斯科
Bush	*President*	*in*	*Moscow*

President Bush meets

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

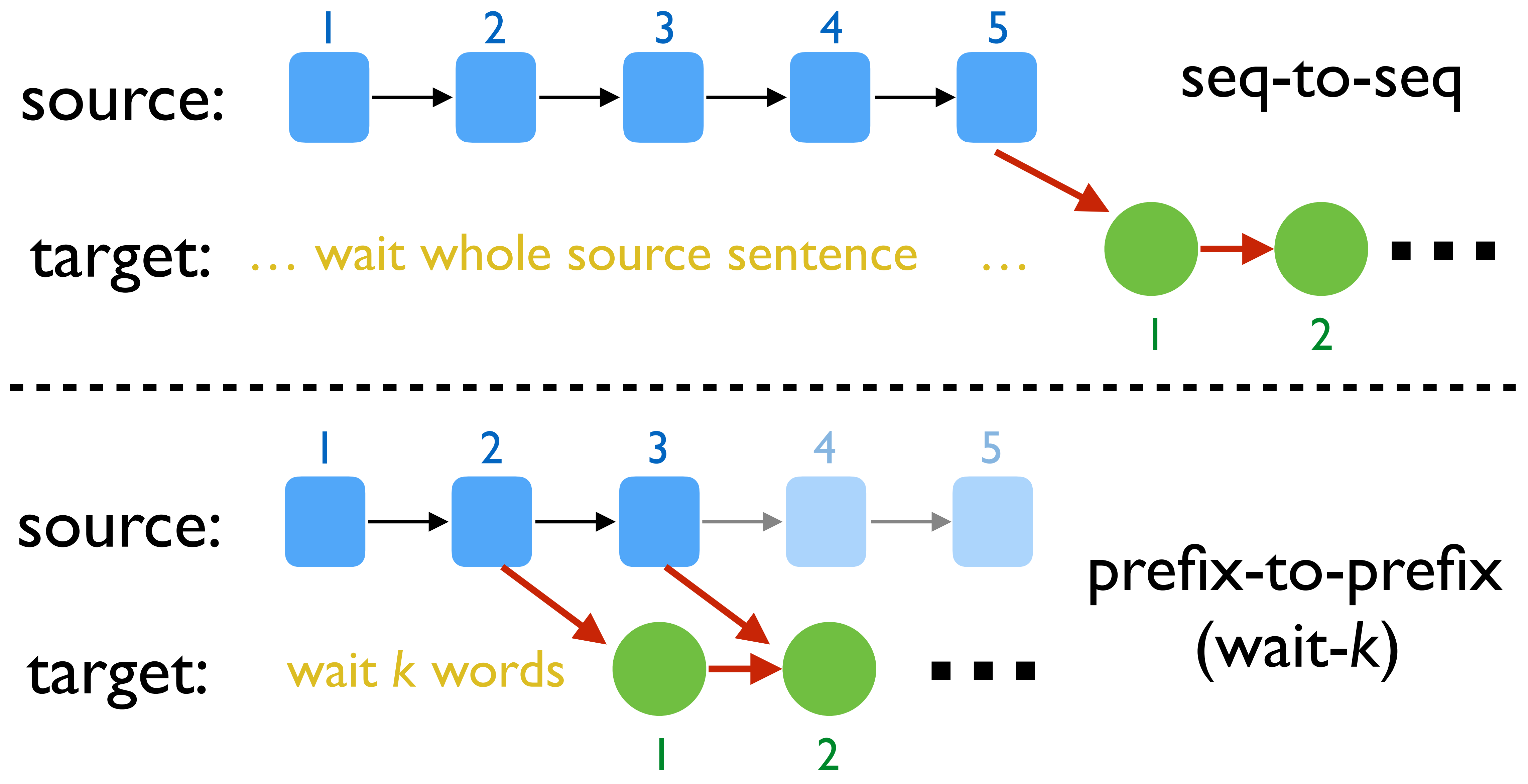


Bùshí	*zǒngtǒng*	*zài*	*Mòsīkē*	*yǔ*
布什	总统	在	莫斯科	与
Bush	*President*	*in*	*Moscow*	*with*

President Bush meets with

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

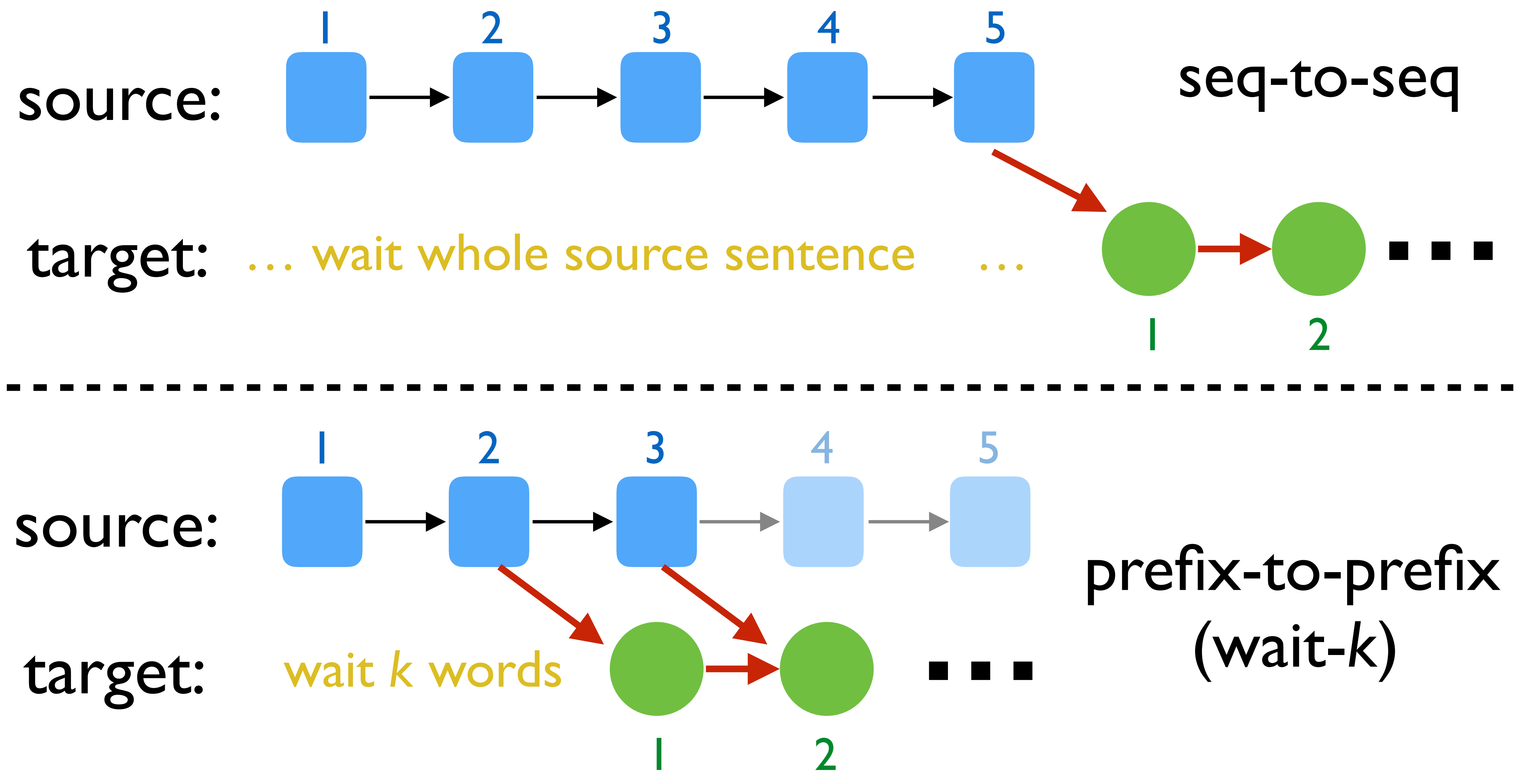


Bùshí	*zǒngtǒng*	*zài*	*Mòsīkē*	*yǔ*	*Éluósī*
布什	总统	在	莫斯科	与	俄罗斯
Bush	*President*	*in*	*Moscow*	*with*	*Russian*

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

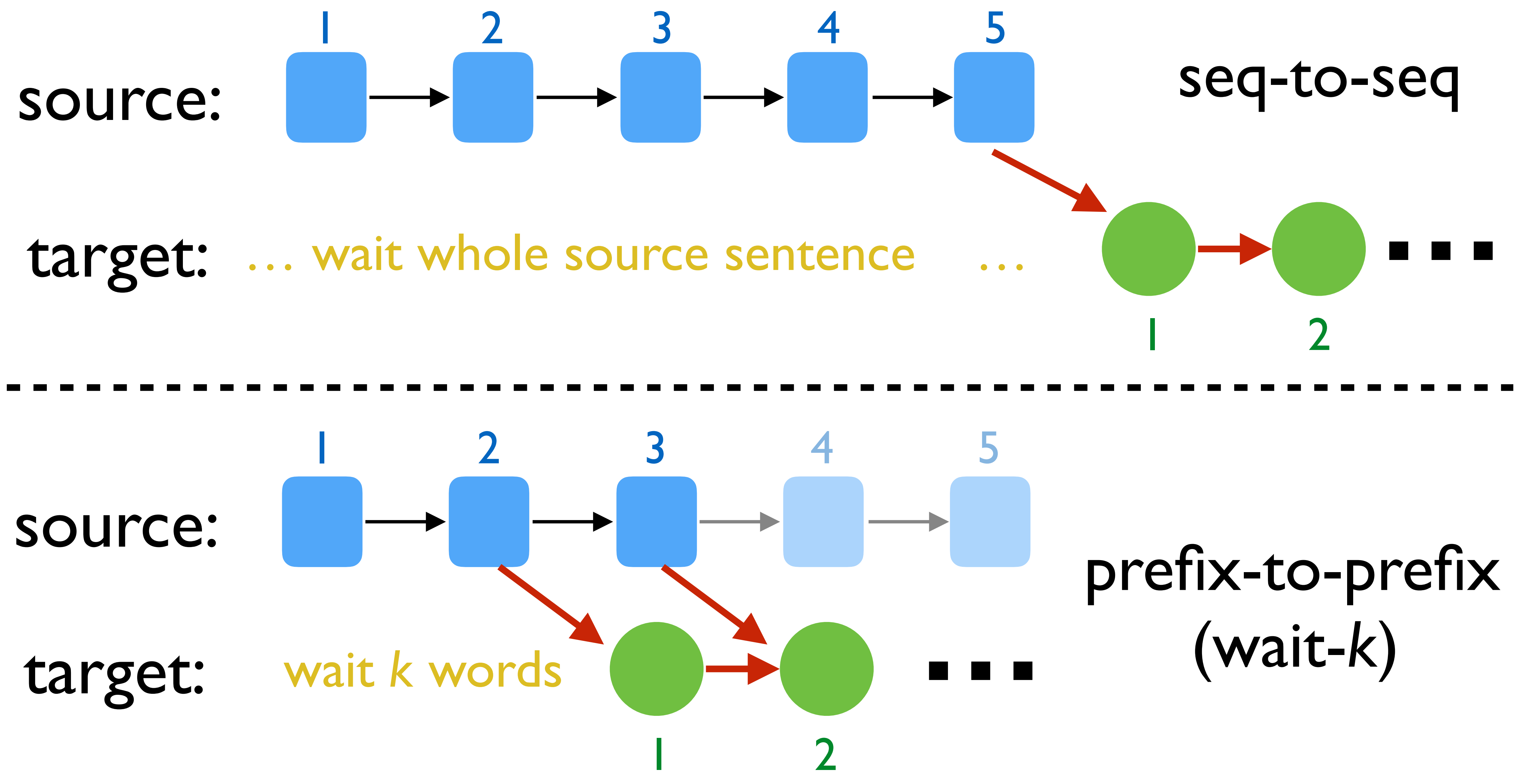


Bùshí	*zǒngtǒng*	*zài*	*Mòsīkē*	*yǔ*	*Éluósī*	*zǒngtǒng*
布什	总统	在	莫斯科	与	俄罗斯	总统
Bush	*President*	*in*	*Moscow*	*with*	*Russian*	*President*

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

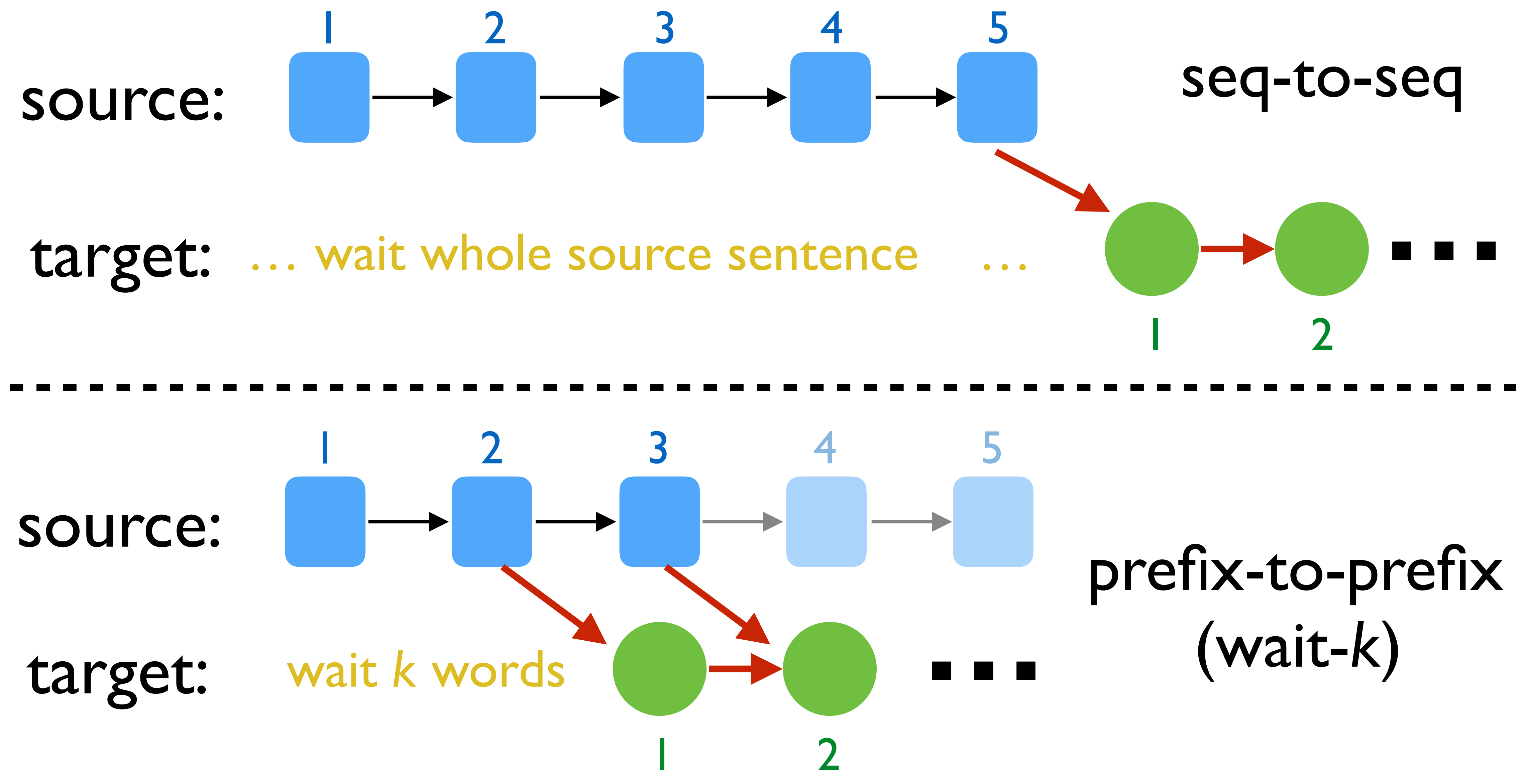


Bùshí	*zǒngtǒng*	*zài*	*Mòsīkē*	*yǔ*	*Éluósī*	*zǒngtǒng*	*Pǔjīng*
布什	总统	在	莫斯科	与	俄罗斯	总统	普京
Bush	*President*	*in*	*Moscow*	*with*	*Russian*	*President*	*Putin*

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

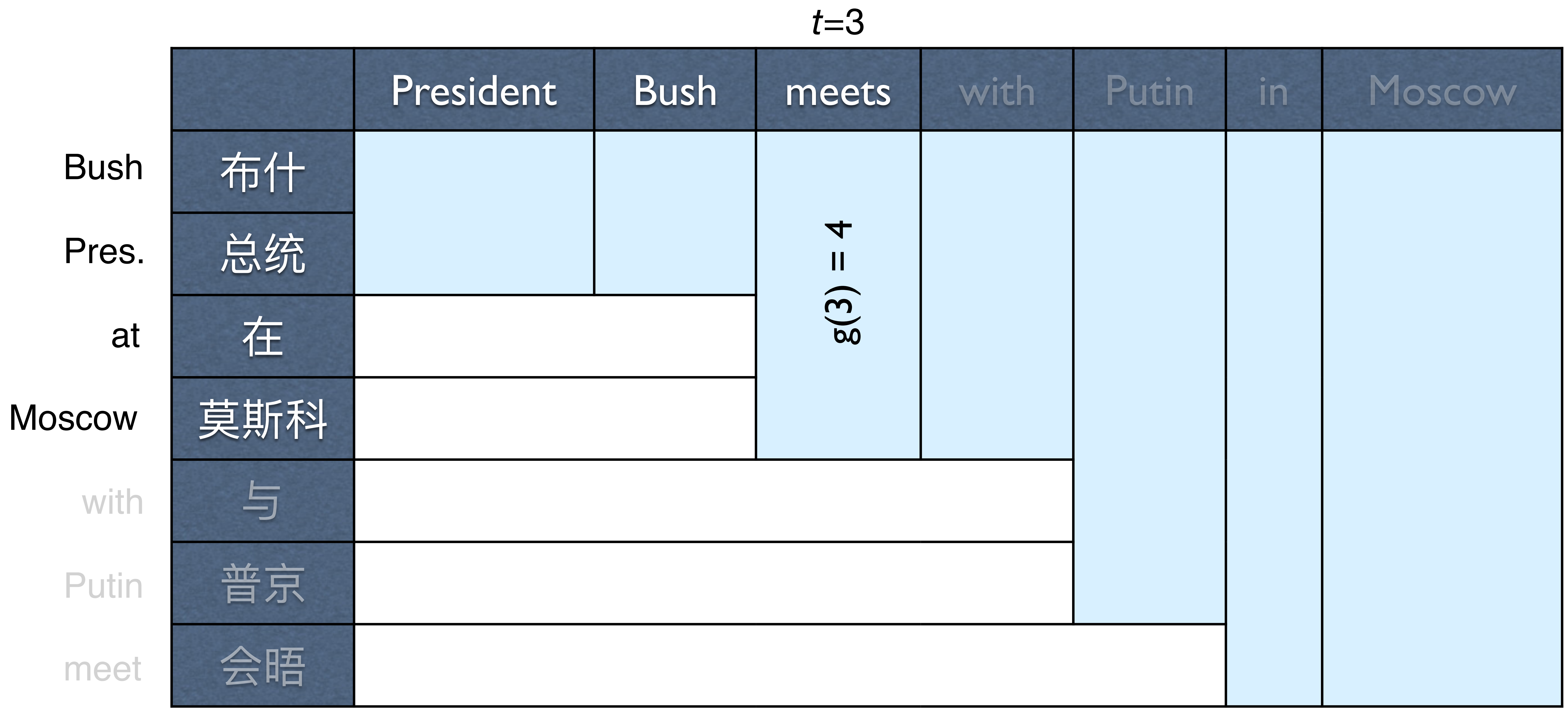


Bùshí	*zǒngtǒng*	*zài*	*Mòsīkē*	*yǔ*	*Éluósī*	*zǒngtǒng*	*Pǔjīng*	*huìwù*
布什	总统	在	莫斯科	与	俄罗斯	总统	普京	会晤
Bush	*President*	*in*	*Moscow*	*with*	*Russian*	*President*	*Putin*	*meet*

President Bush meets with Russian President Putin in Moscow

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



Demo 1 (Research)

美国总统布什在莫斯科与
us president bush met

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

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

Demo 1 (Research)

美国总统布什在莫斯科与
us president bush met

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

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

Demo 1 (Research)

美国总统布什在莫斯科与
us president bush met

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).

Demo 1 (Research)

美国总统布什在莫斯科与
us president bush met

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 .

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

Demo 2 (Latency-Accuracy Tradeoff)

Chinese input:	江 泽民 对 美国 总统 的 发言 表示 遗憾 。
Pinyin:	jiāng zémín dùi měiguó zǒngtǒng de fāyán biǎoshì yíhàn 。
Word-by-Word Translation:	jiang zemin to united states president of speak express regret 。
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 Tranlation (gready):	jiang zemin expressed regret over the us president 's remarks .
Baseline Tranlation (beam 5):	jiang zemin expressed regret over the us president 's remarks .

Demo 2 (Latency-Accuracy Tradeoff)

Chinese input:	江 泽民 对 美国 总统 的 发言 表示 遗憾 。
Pinyin:	jiāng zémín dùi měiguó zǒngtǒng de fāyán biǎoshì yíhàn 。
Word-by-Word Translation:	jiang zemin to united states president of speak express regret 。
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 Tranlation (gready):	jiang zemin expressed regret over the us president 's remarks .
Baseline Tranlation (beam 5):	jiang zemin expressed regret over the us president 's remarks .

Demo 3 (Deployment)

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

Demo 3 (Deployment)

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

German => English Example

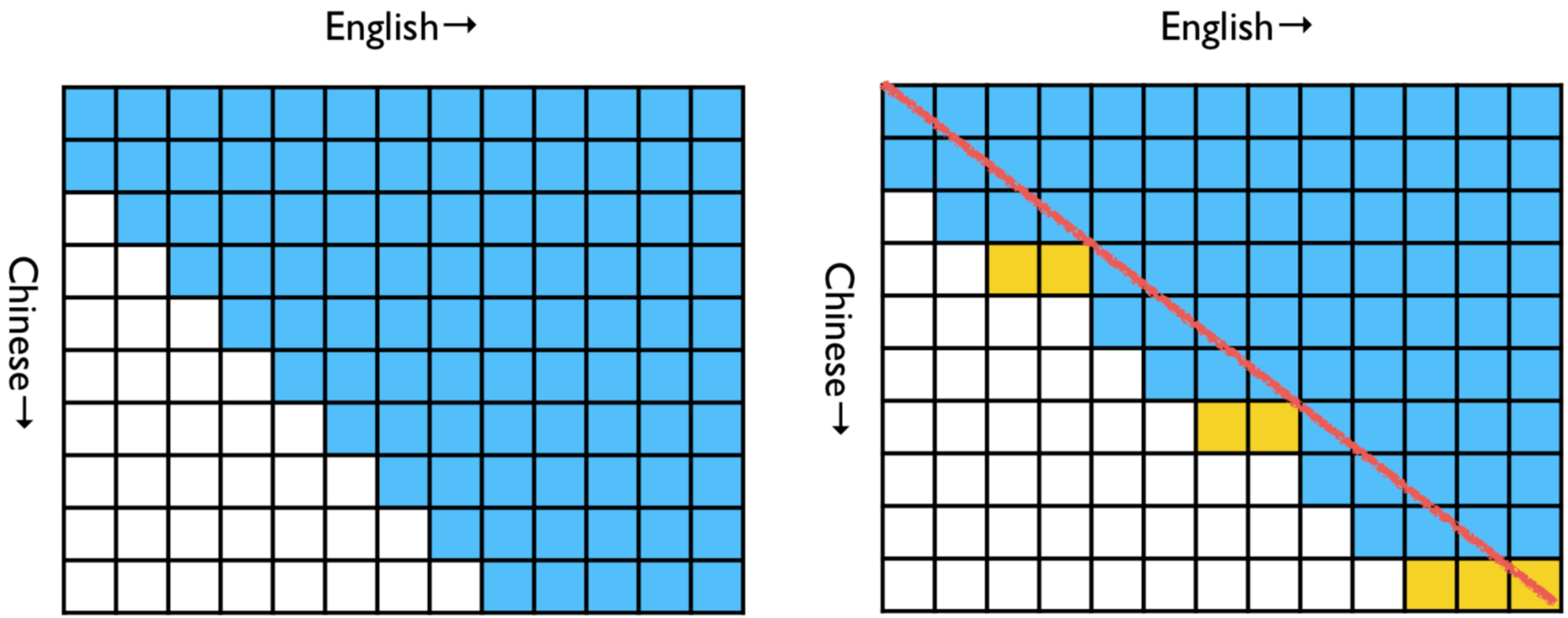
German source:
doch während man sich im kongress nicht auf ein vorgehen einigen kann , warten mehrere bundesstaaten nicht länger .

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 .

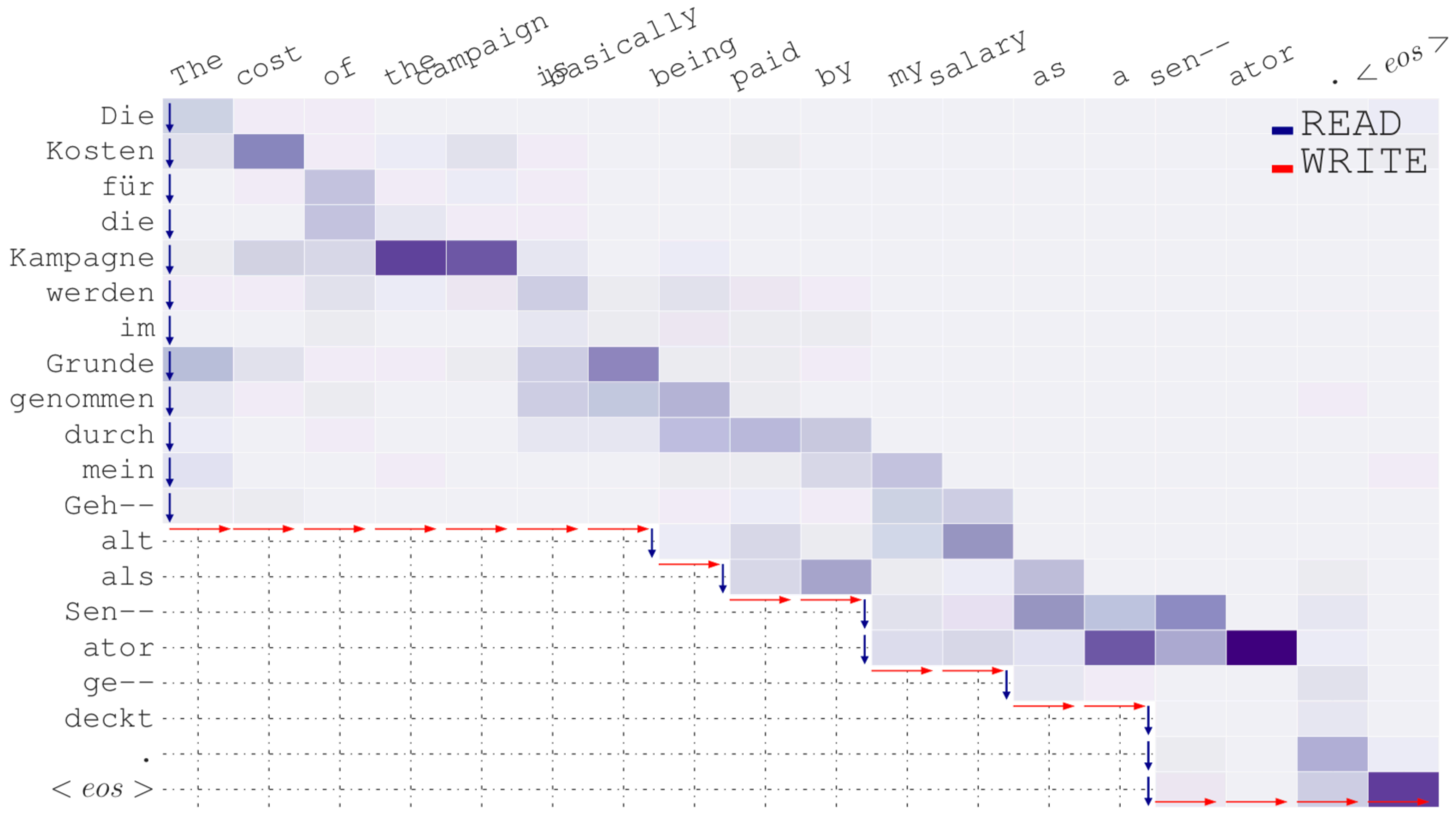
Refinements: Wait-*k* with Catchup

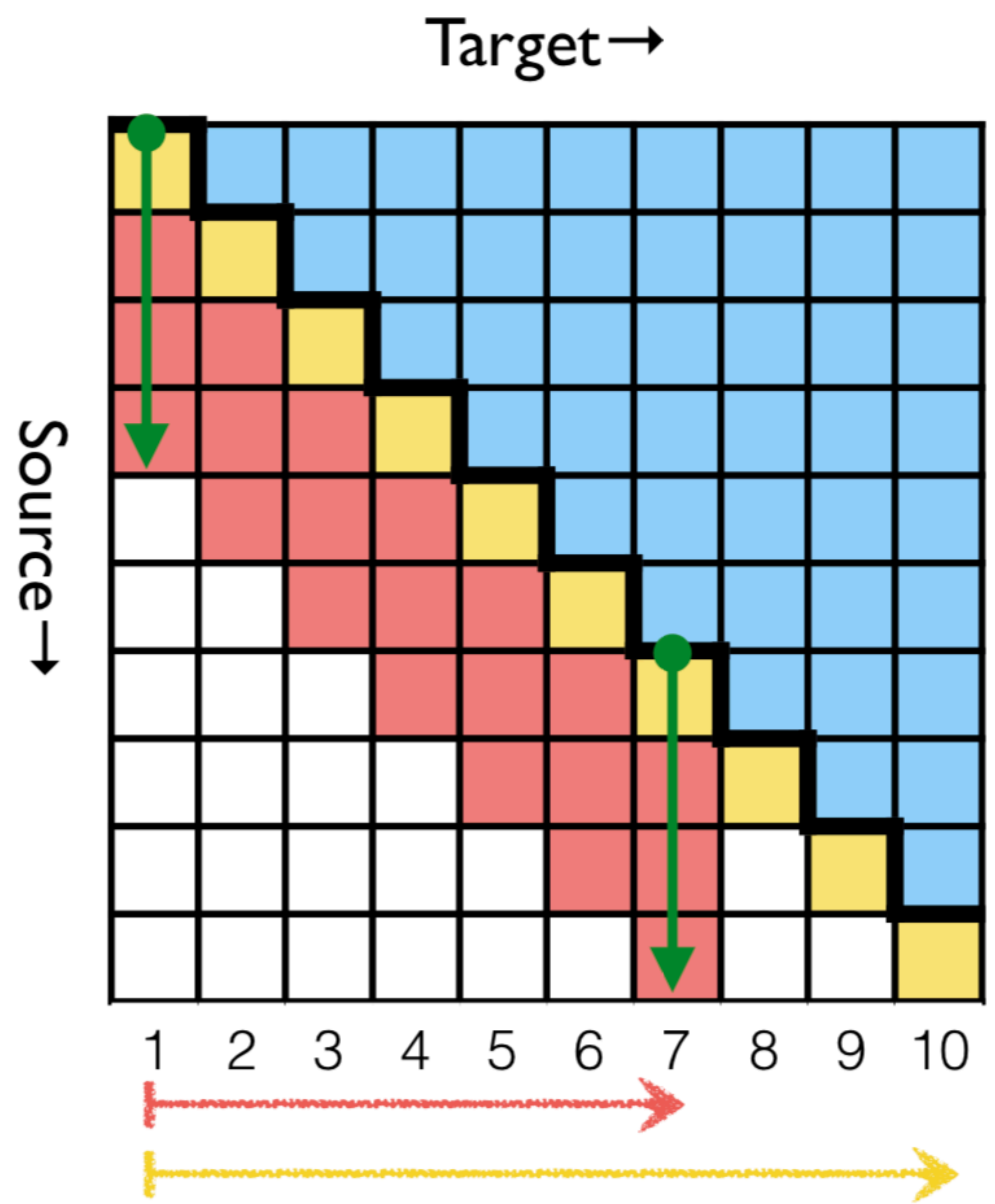
- English translation length is often ~1.25x of the Chinese input length
  - in a more or less "synchronized" policy like wait-*k*, the English translation will be lagging behind more and more severely
  - catchup: decode two English words in 1 out of 4 steps

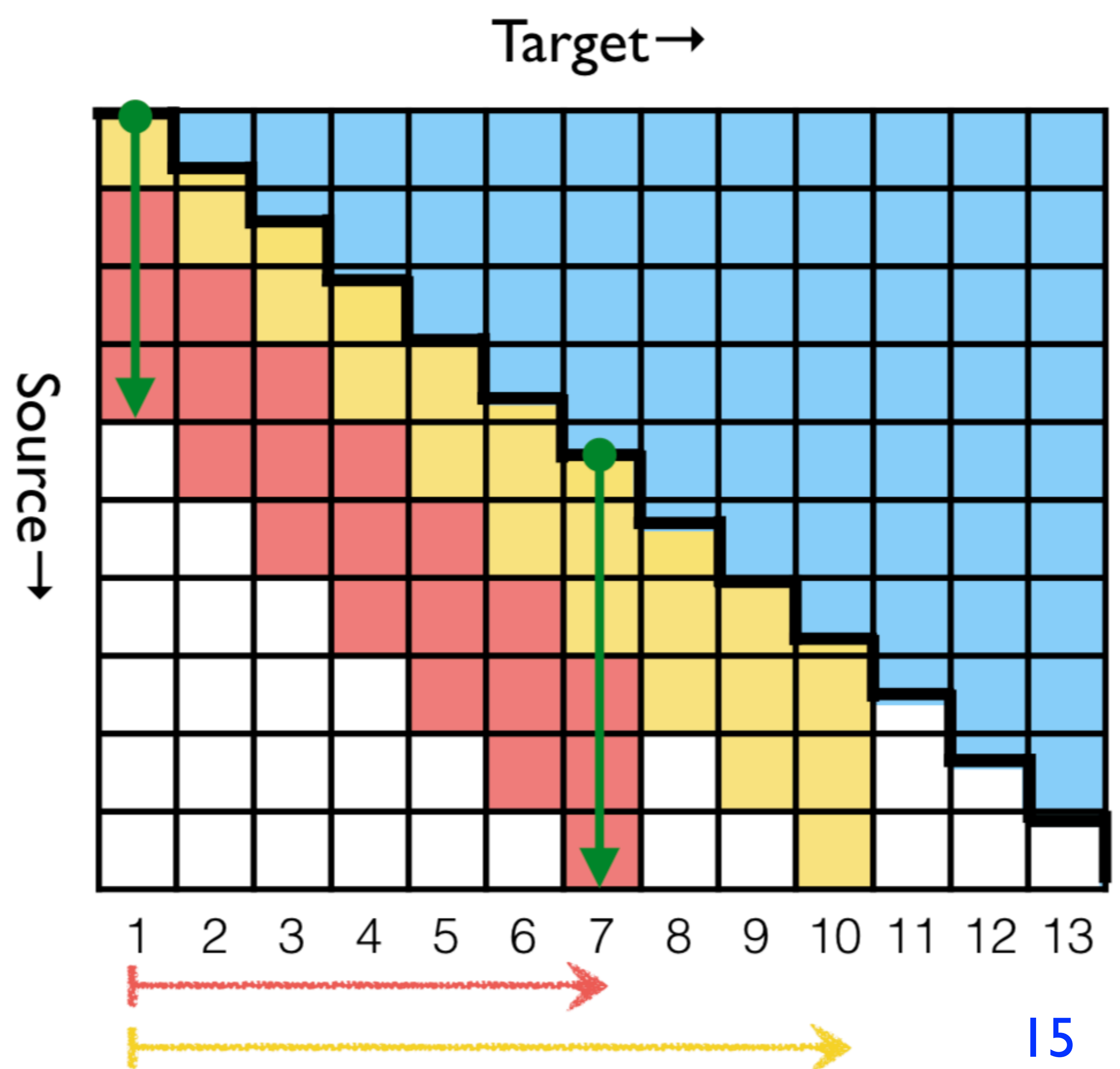


New Latency Metric: Average Lagging

- 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) ≈ *k*

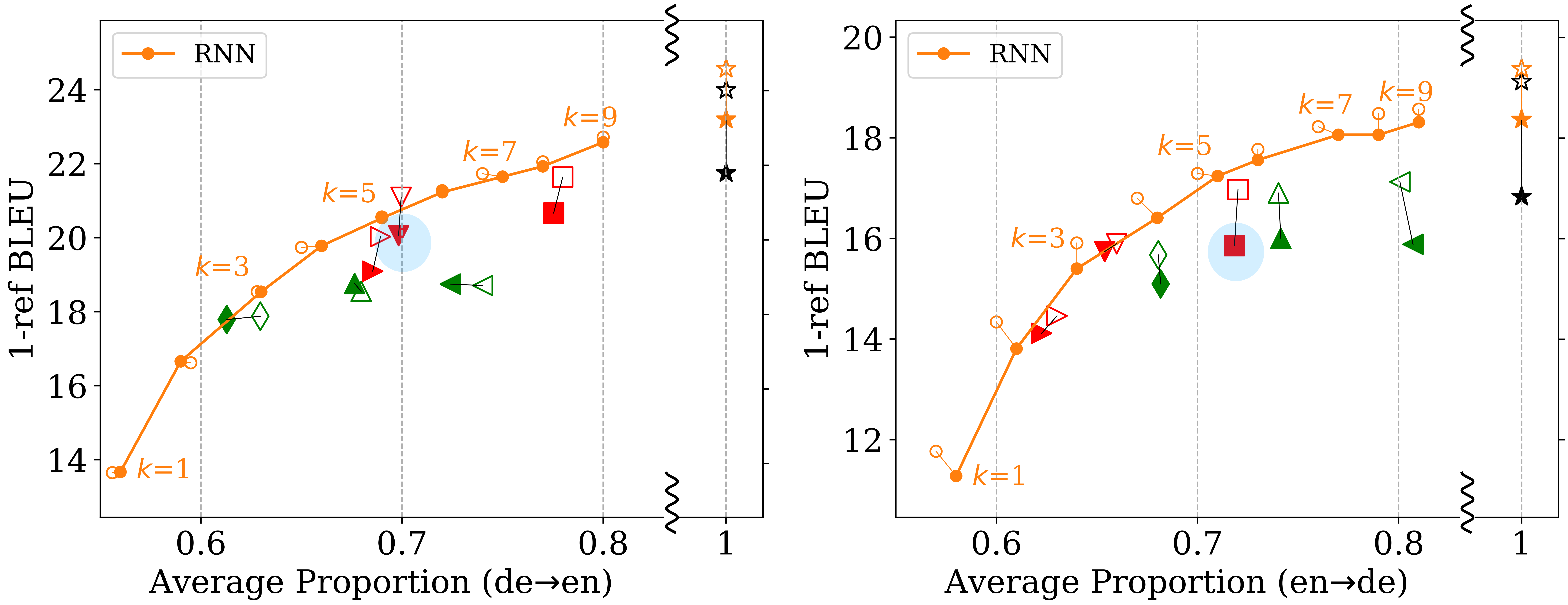






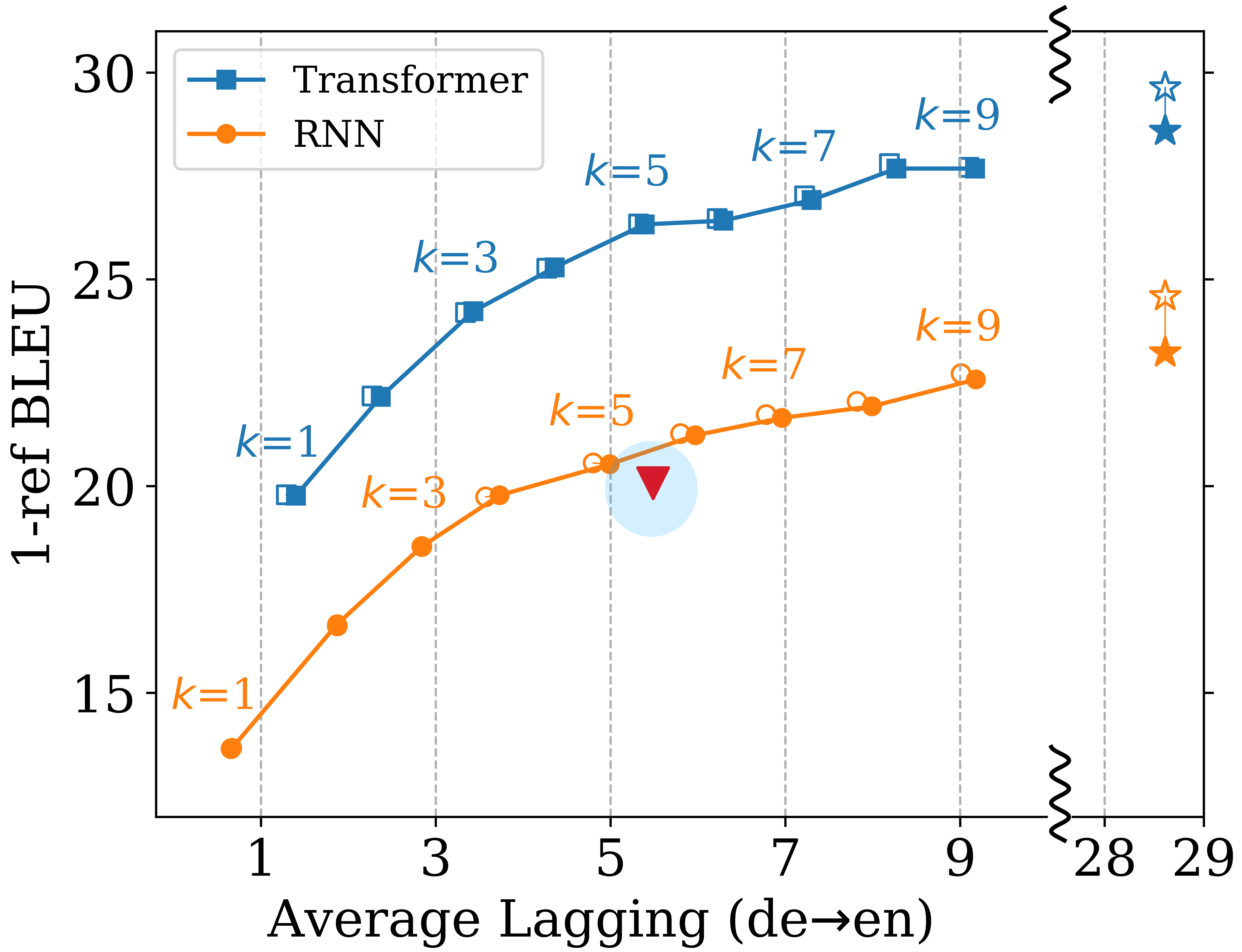
Experiments: German<=>English

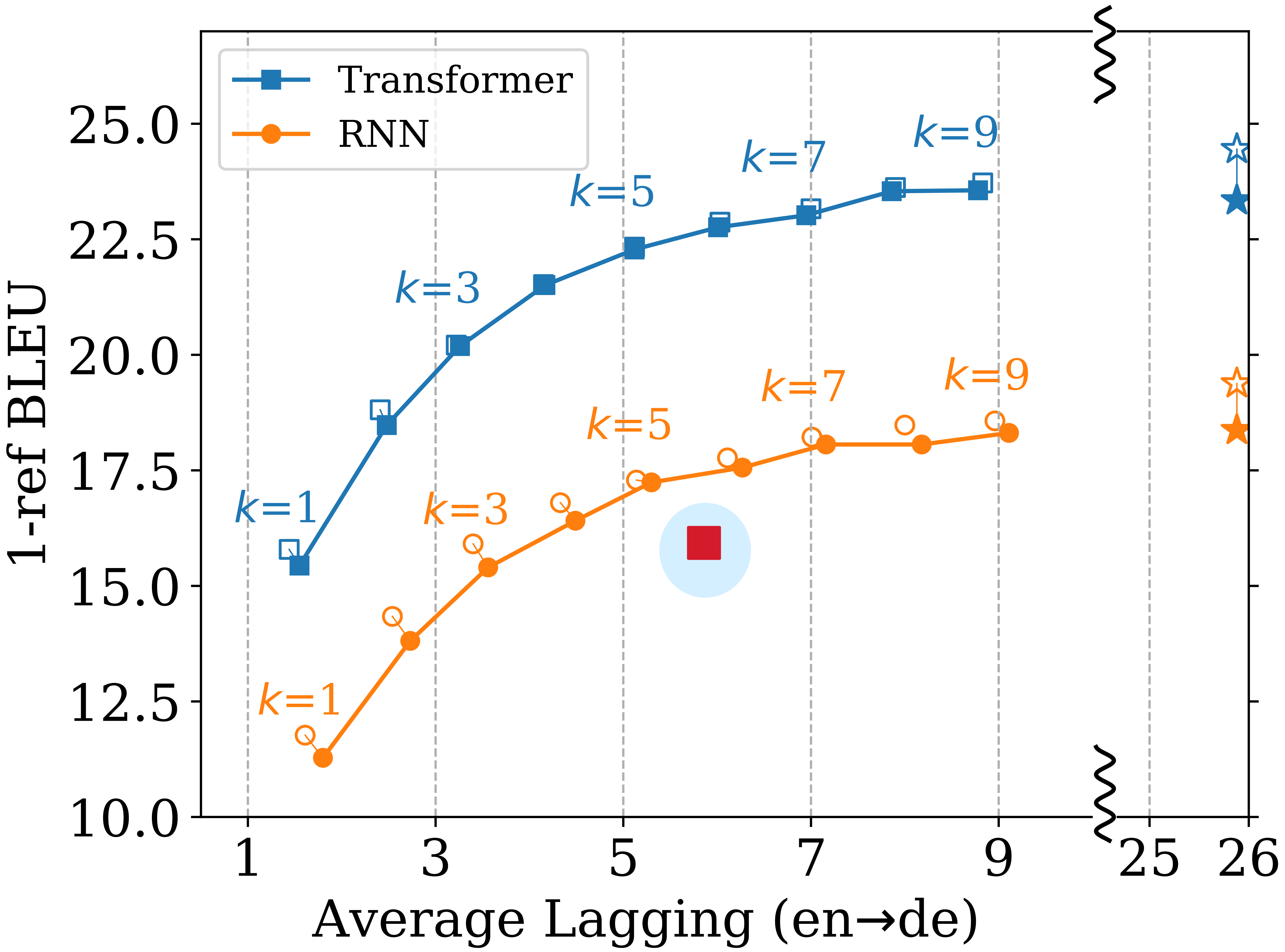
- trained on 4.5M sentence pairs (WMT 15); comparing with Gu et al 2017



Experiments: German<=>English

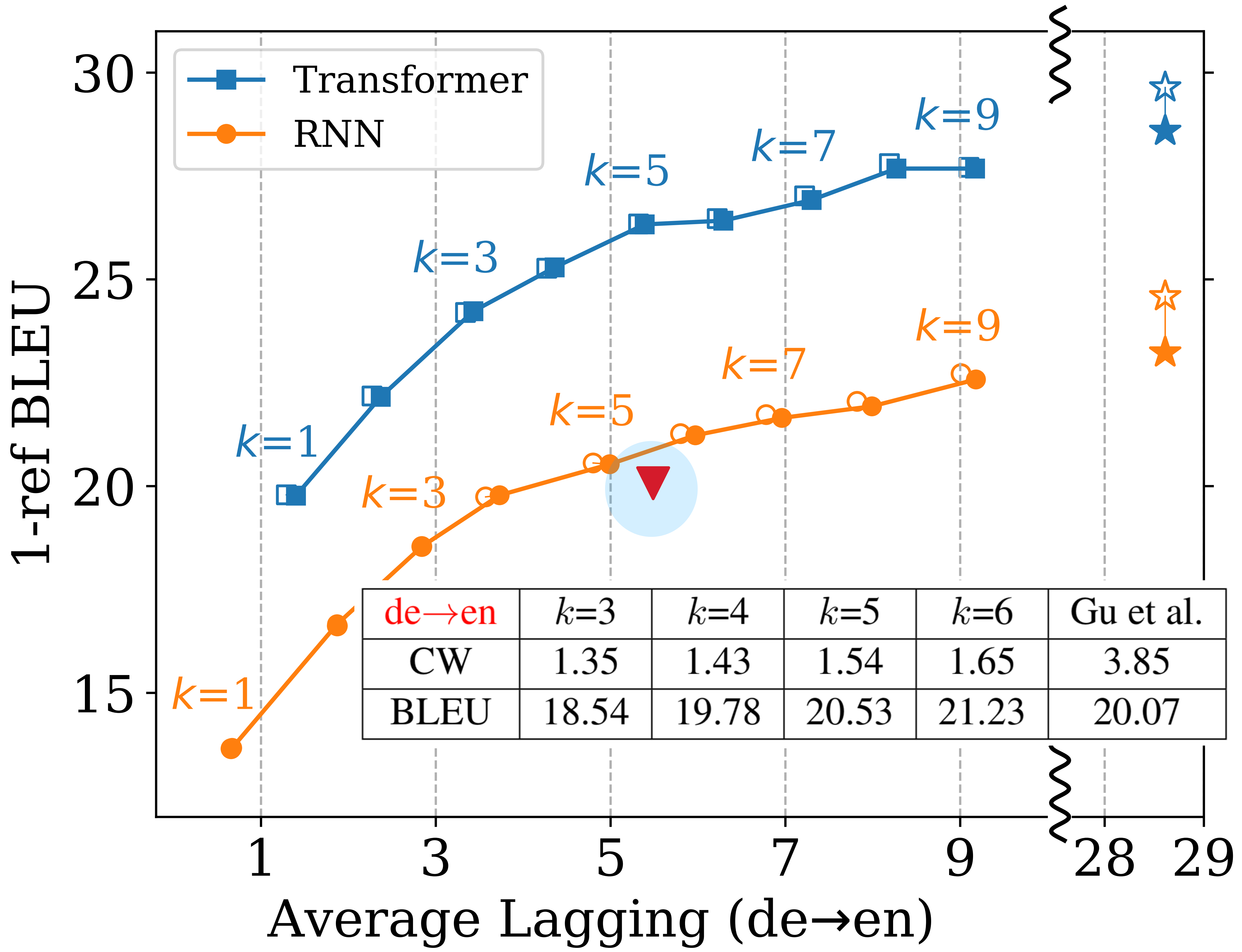
- trained on 4.5M sentence pairs (WMT 15); comparing with Gu et al 2017



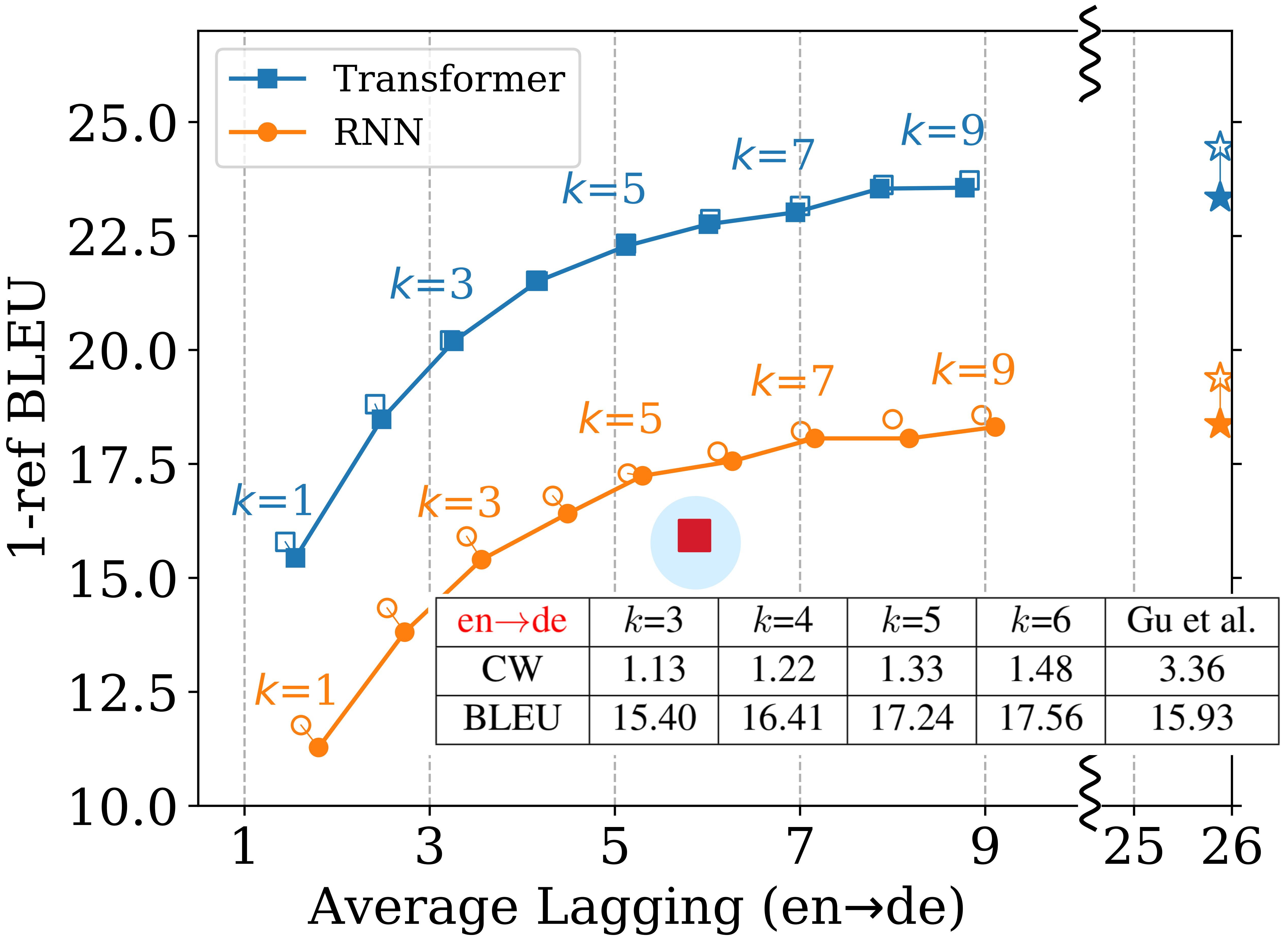


Experiments: German<=>English

- trained on 4.5M sentence pairs (WMT 15); comparing with Gu et al 2017



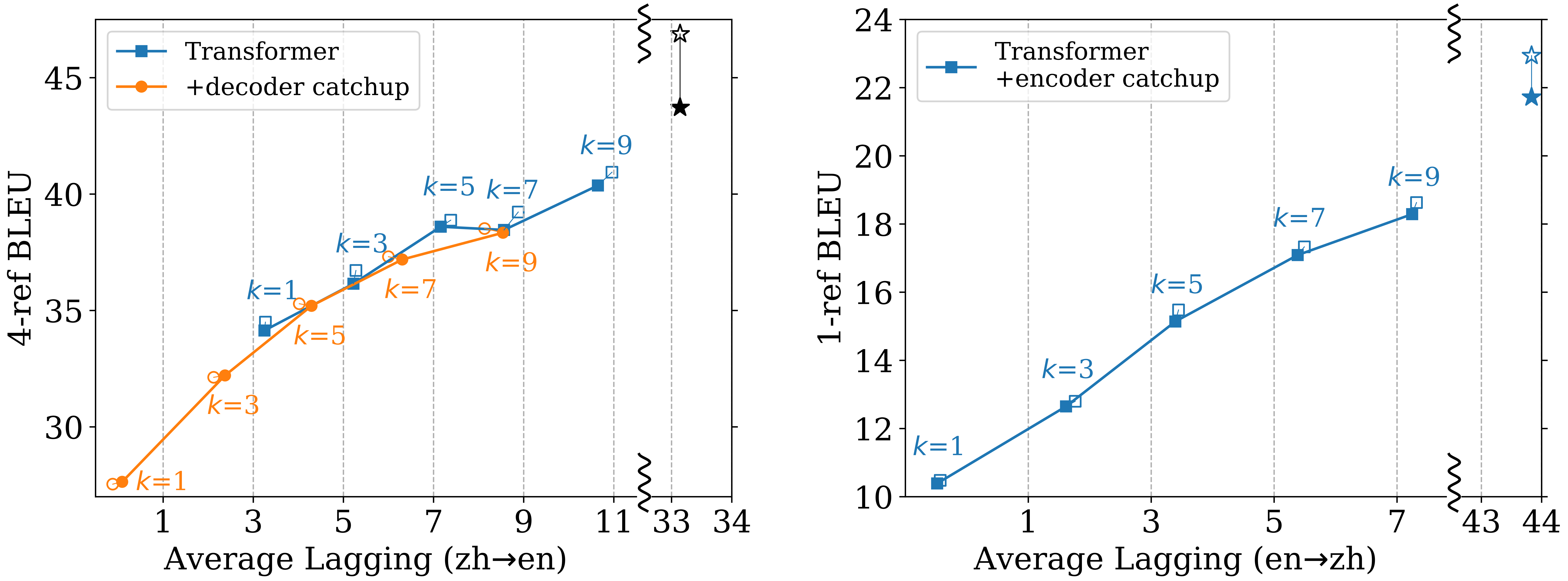
de→en	k=3	k=4	k=5	k=6	Gu et al.
CW	1.35	1.43	1.54	1.65	3.85
BLEU	18.54	19.78	20.53	21.23	20.07



en→de	k=3	k=4	k=5	k=6	Gu et al.
CW	1.13	1.22	1.33	1.48	3.36
BLEU	15.40	16.41	17.24	17.56	15.93

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

	1	2	3	4	5	6	7	8	9	10	
	Měiguó	*dāngjú*	*duì*	*Shātè*	*jìzhě*	*shīzōng*	*yī*	*àn*	*gǎndào*	*dānyōu*	
(a)	美国	当局	对	沙特	记者	失踪	一	案	感到	担忧	
	US	authorities	to	Saudi	reporter	missing	a	case	feel	concern	
k=3				the	us	authorities	are	very	concerned	about	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

	1	2	3	4	5	6	7	8	9	10	
	Měiguó	*dāngjú*	*duì*	*Shātè*	*jìzhě*	*shīzōng*	*yī*	*àn*	*gǎndào*	*dānyōu*	
(a)	美国	当局	对	沙特	记者	失踪	一	案	感到	担忧	
	US	authorities	to	Saudi	reporter	missing	a	case	feel	concern	
k=3				the	us	authorities	are	very	concerned	about	the saudi reporter 's missing case
k=3†				the	us	authorities	are very	concerned	about	the	saudi reporter 's missing case
k=∞											us authorities concerned over saudi journalists missing

										bùmǎn	
(b)	美国	当局	对	沙特	记者	失踪	一	案	感到	不满	
k=3				the	us	authorities	are	very	concerned	about	the saudi reporter 's missing case
k=5						the	us	authorities	have	expressed	**dissatisfaction** with the incident of saudi arabia 's missing reporters
k=∞											us authorities dis- satisfied with saudi reporters ' missing case

Media Reports

Media coverage:

Media Reports

Media coverage:

同传AI，刚刚在国内掀起过暴风骤雨。

但现在，百度于硅谷宣布了最新重大突破——一个名为**STACL**的同传AI，论文结果优异，Demo效果惊人。

MIT科技评论、IEEE Spectrum等一众外媒，还纷纷给出好评，这是2016年百度Deep Speech 2发布以来，又一项让技术外媒们如此激动的新进展。

百度自己披露：与现在大多数AI“实时”翻译系统不同，STACL的特点是**能预测**和**延时可控**，能够在演讲者讲话后几秒钟开始翻译，并在句子结束后几秒钟内完成。

STACL不走“整句说完再翻译”的路线，甚至还会预测发言者未来几秒的内容，于是延时更短，更接近人类同传。

究竟能达到什么程度？IEEE Spectrum采访后给出类比：跟联合国会议里的人类同传相媲美。

This is another new development that has made foreign technology media so excited since the release of Baidu Deep Speech 2 in 2016.

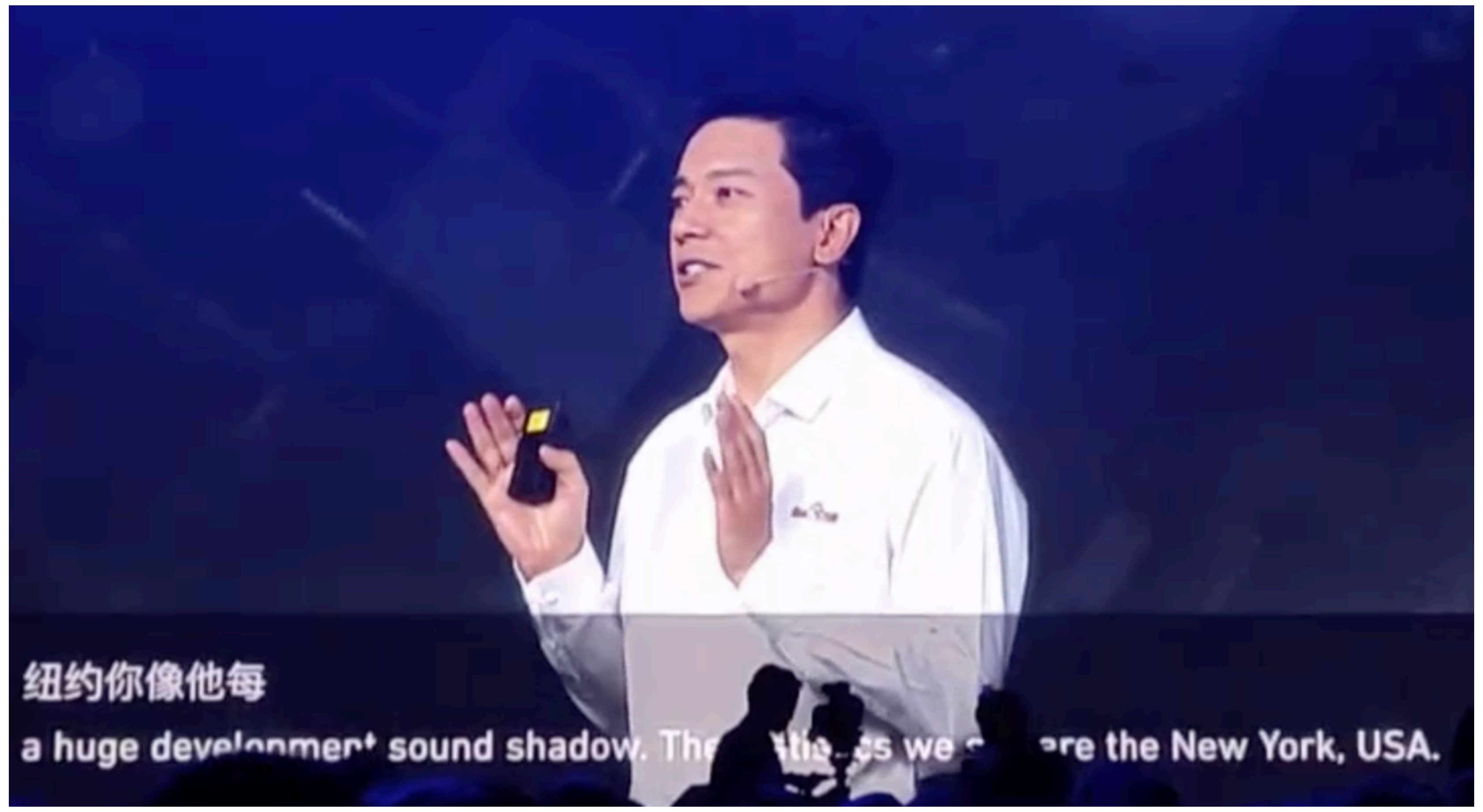
— QbitAI (量子位)

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

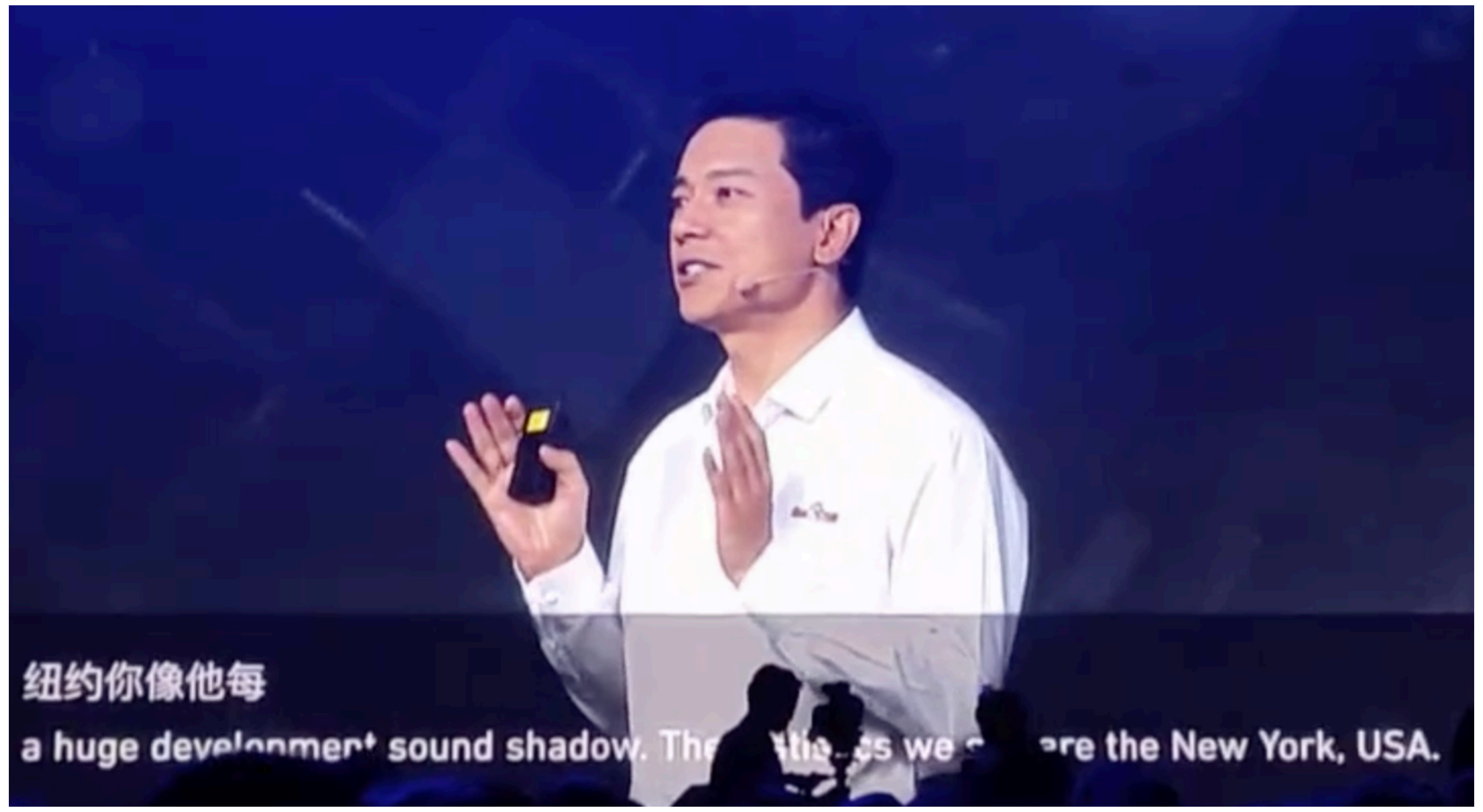
Conclusions

- 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


纽约你像他每
a huge development sound shadow. The
we
are the New York, USA.

非常 感谢 您 来 听 我 的 演讲

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

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.