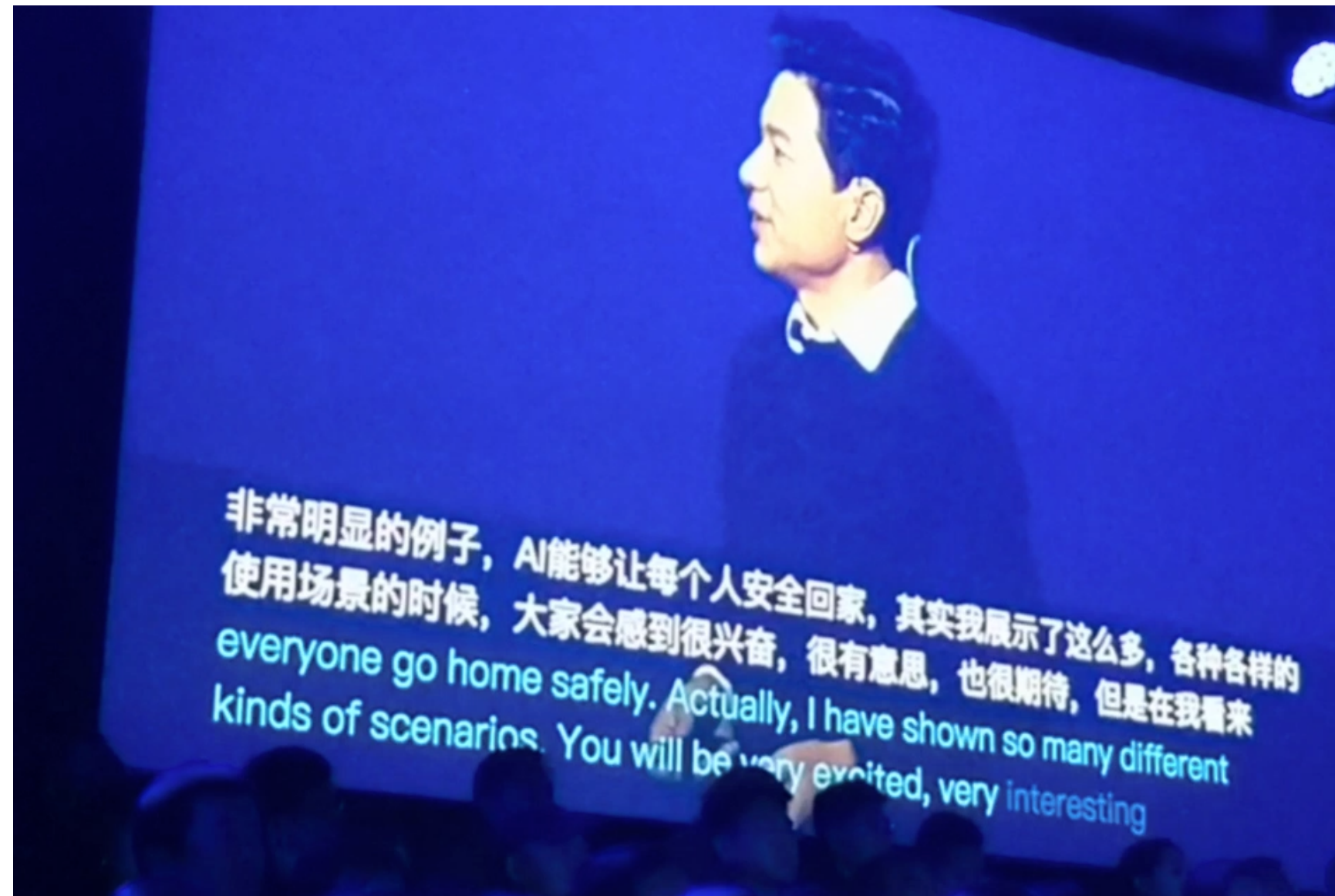


Breakthrough in Simultaneous Translation

full-sentence (non-simultaneous) translation



Baidu World Conference, November 2017

simultaneous translation, latency ~3 secs



Baidu World Conference, November 2018

STACL

Media coverage:

IEEE
SPECTRUM

MIT
Technology
Review

CNBC

**Venture
Beat**

SILICON
ANGLE

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

量子位

机器之心

新浪科技

中国新闻网
WWW.CHINANEWS.COM

鳳凰網 科技

新浪财经
finance.sina.com.cn

环球网 智能
smart.huanqiu.com

雷锋网

博客园
cnblogs.com

AI TechYun

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)

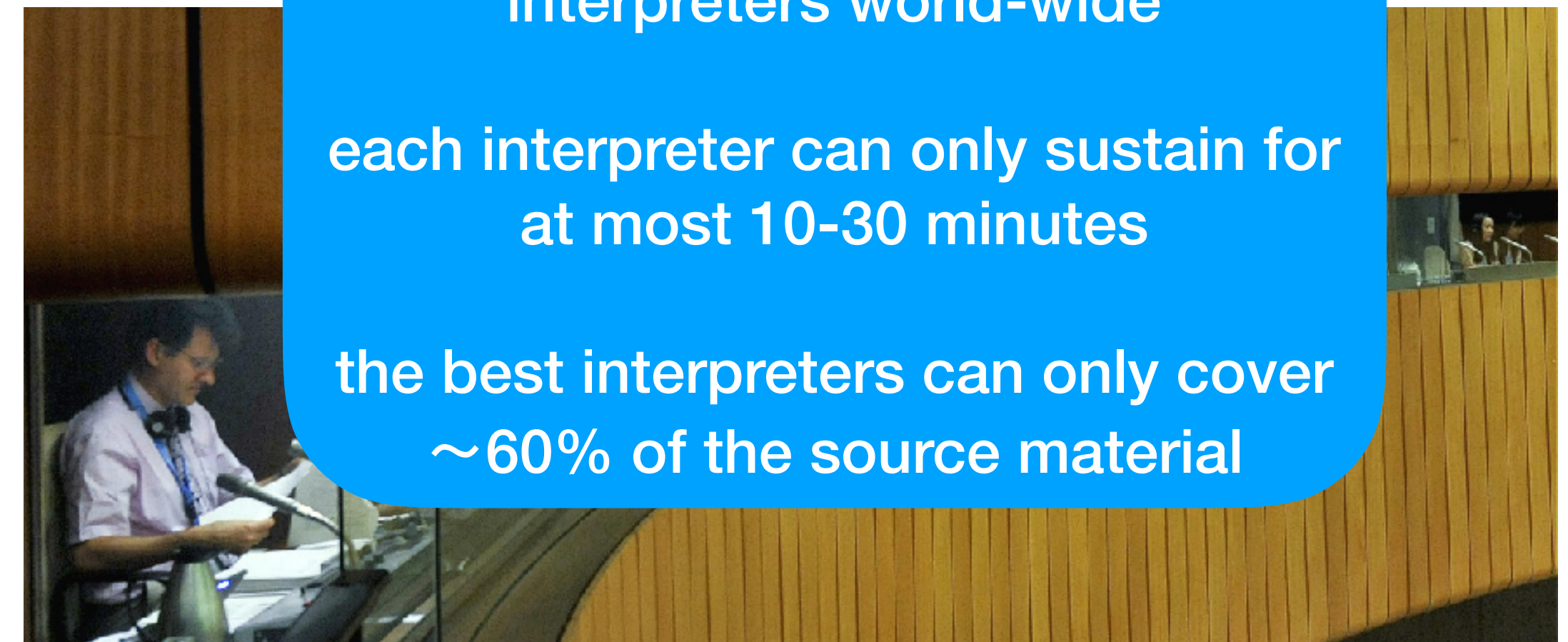
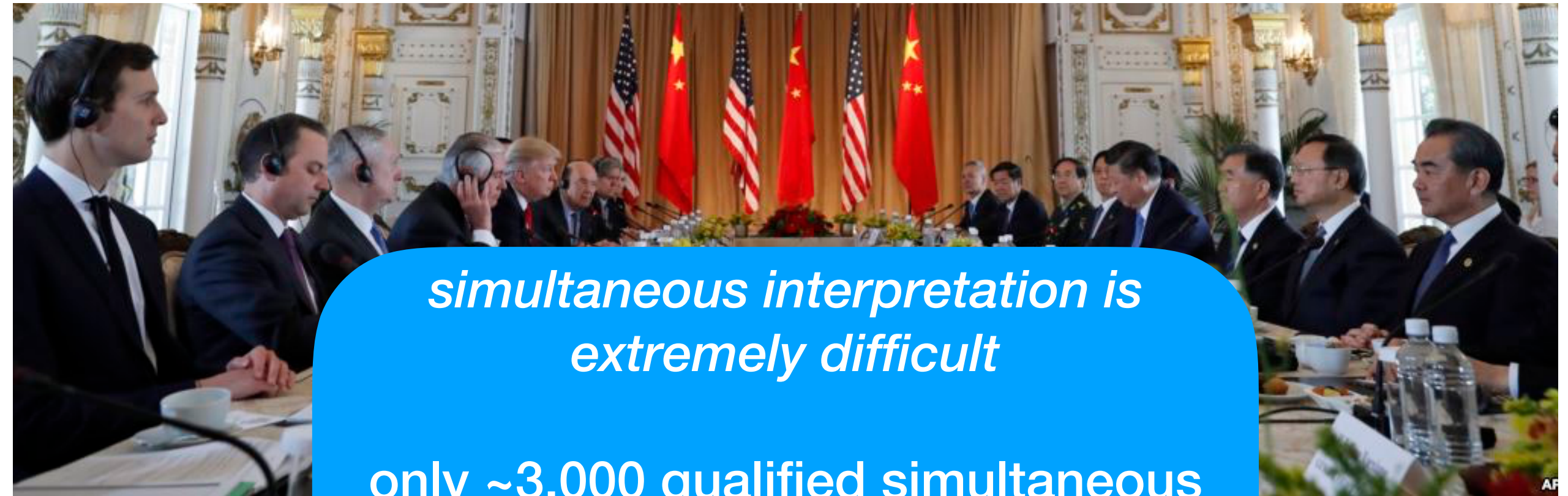


Background: Consecutive vs. Simultaneous

consecutive interpretation
multiplicative latency (x2)



simultaneous interpretation
additive latency (+3 secs)



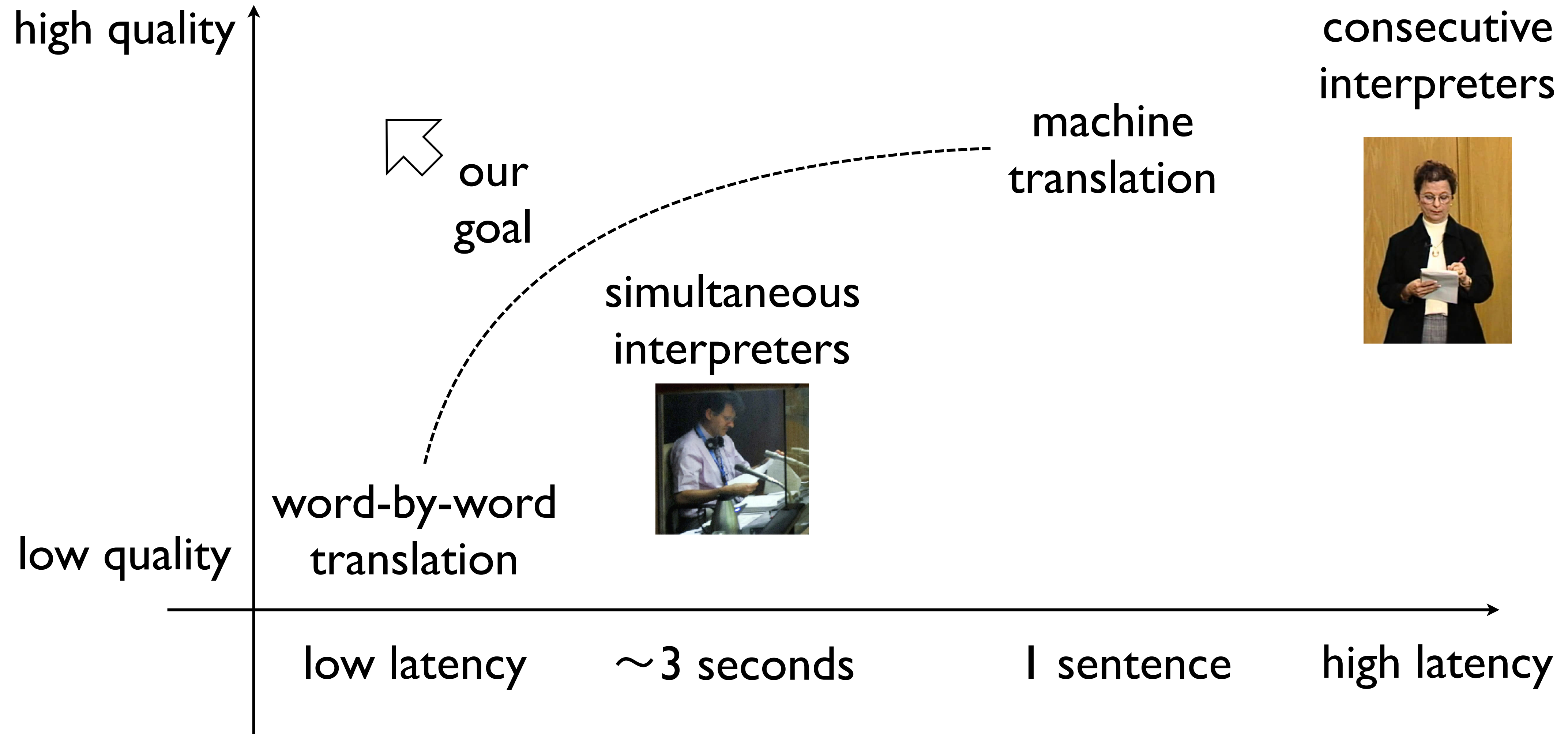
simultaneous interpretation is extremely difficult

only ~3,000 qualified simultaneous interpreters world-wide

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

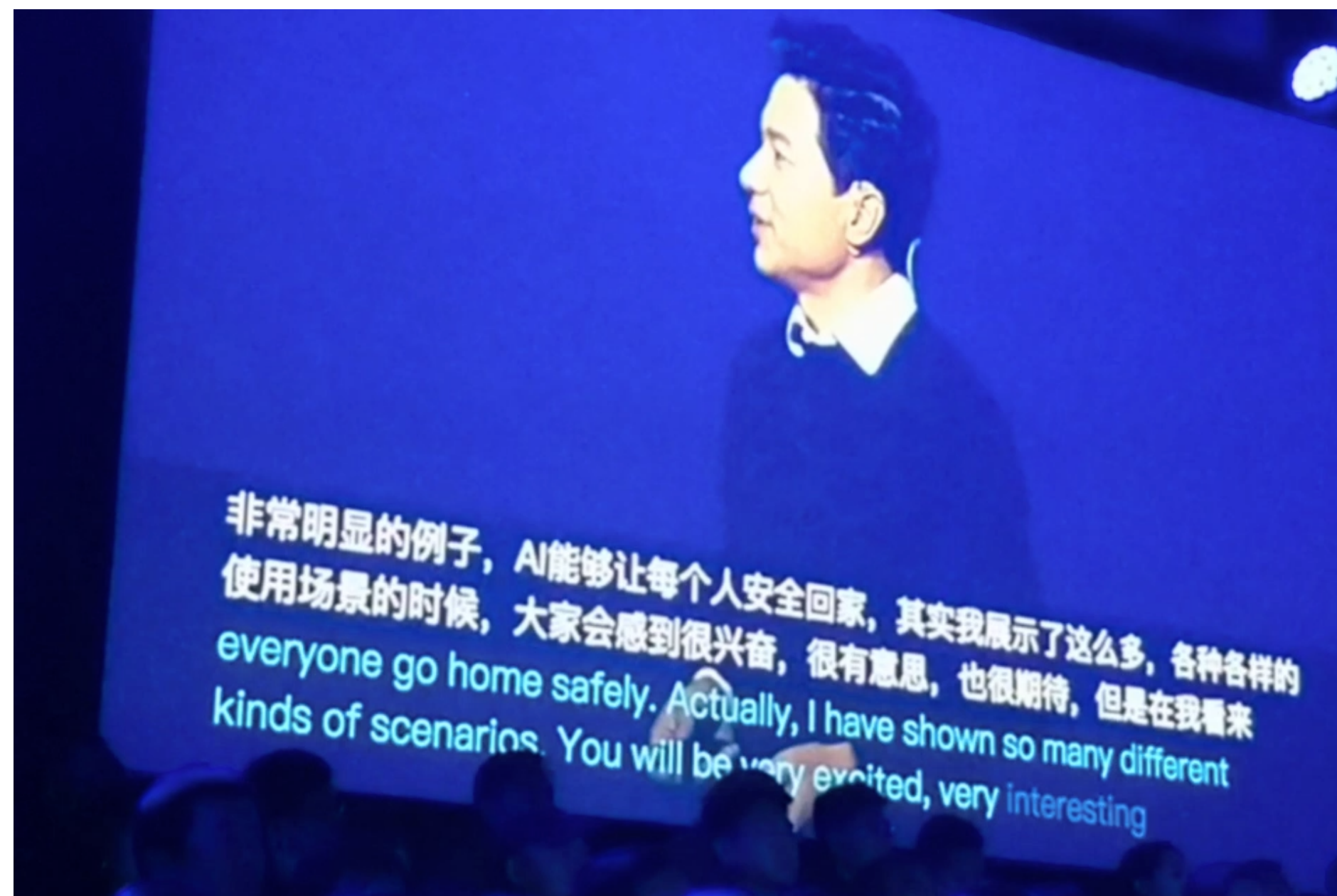
the best interpreters can only cover ~60% of the source material

Tradeoff between Latency and Quality

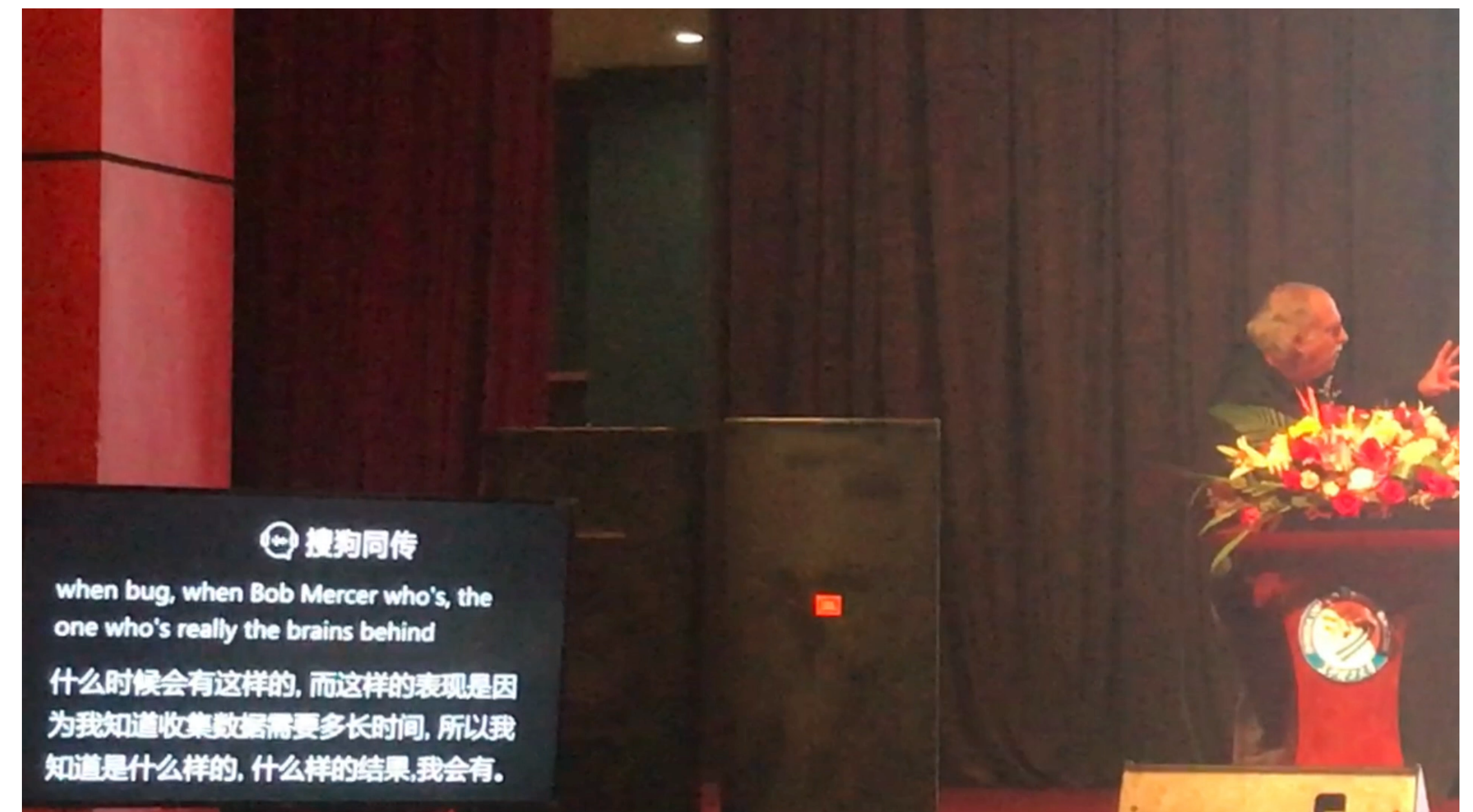


Industrial Work in Simultaneous Translation

- 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)

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

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**

Grissom et al, 2014

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)

ich bin mit dem Zug nach Ulm **gefahren**

I am with the train to Ulm **traveled**

Grissom et al, 2014

I (..... *waiting*.....) **traveled** by train to Ulm

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

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**

Grissom et al, 2014

I (..... *waiting*) **traveled** by train to Ulm

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**

Grissom et al, 2014

I (..... *waiting*) **traveled** by train to Ulm

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

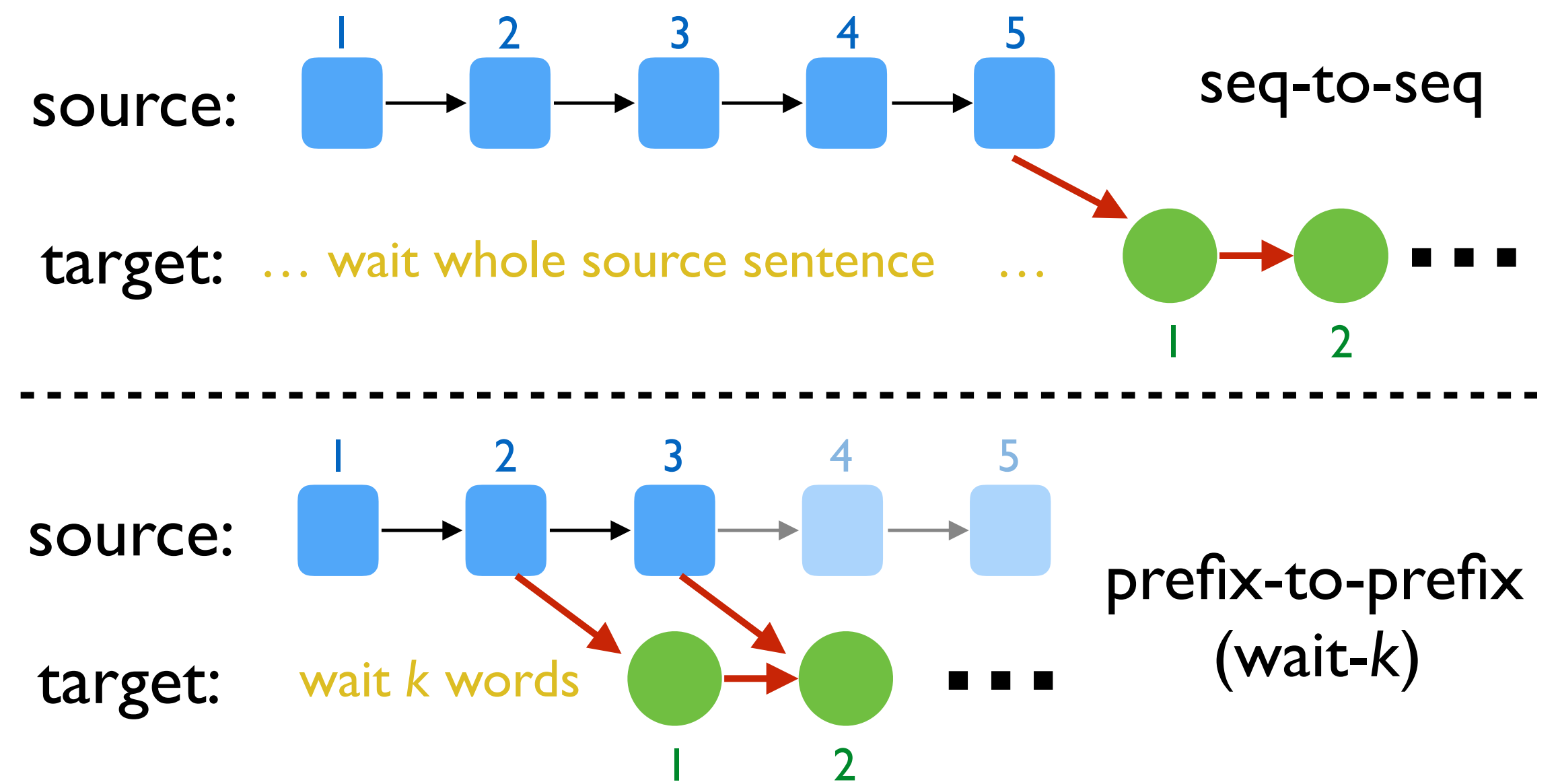
President Bush **meets** with Russian President Putin in Moscow

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

anticipative: President Bush **meets** with Russian President Putin in Moscow

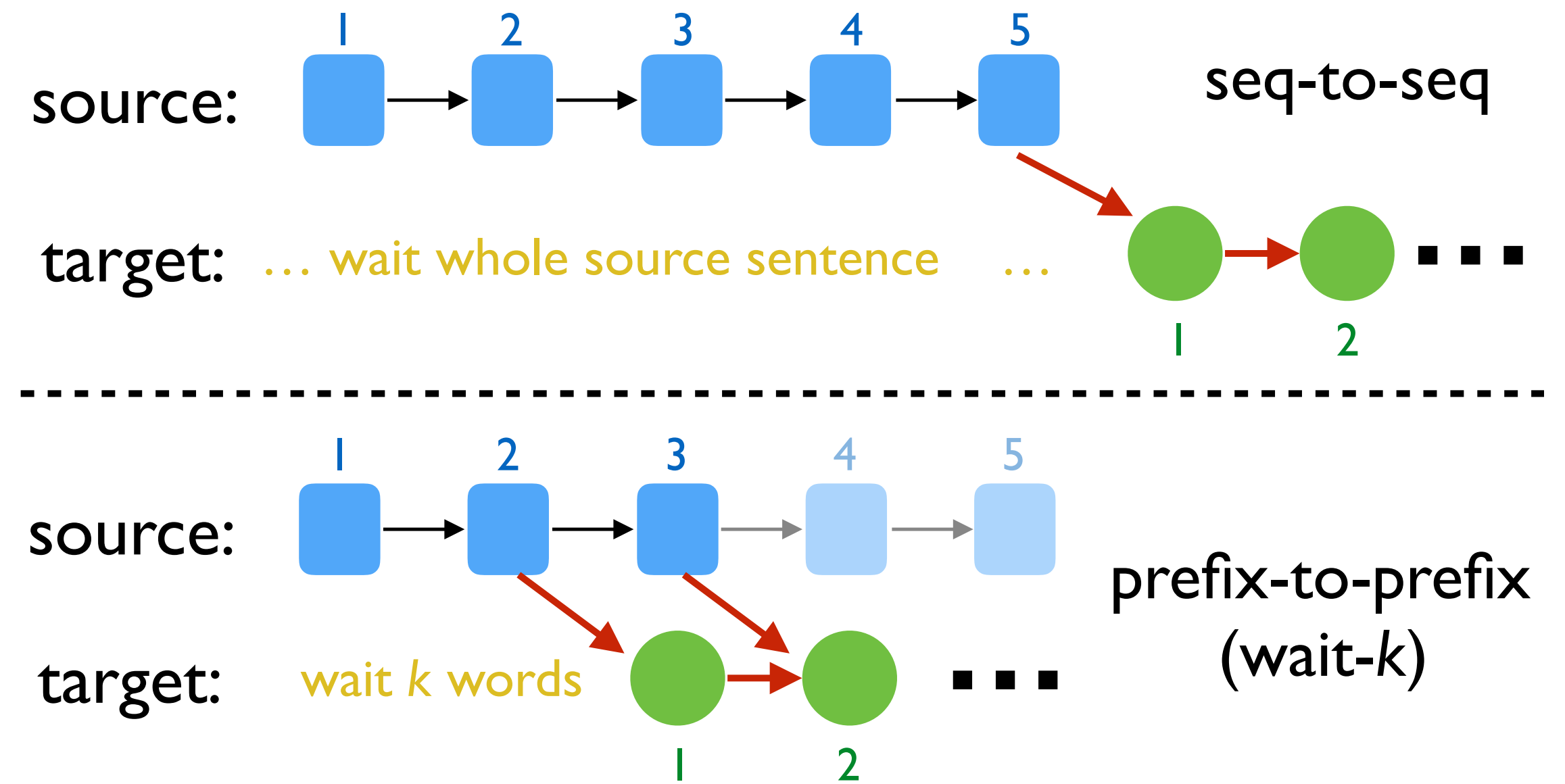
Our Solution: Prefix-to-Prefix

- 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

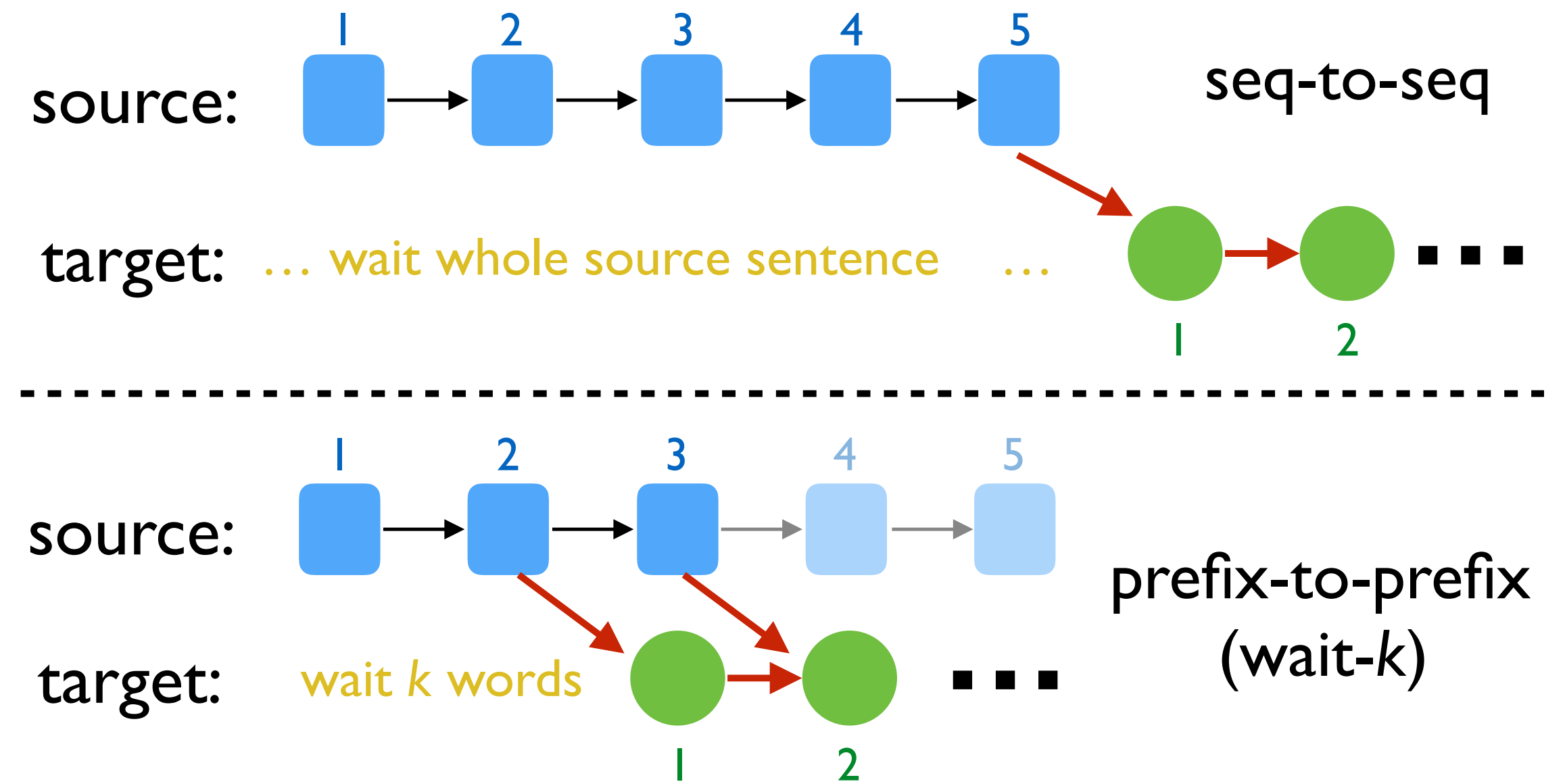


Bùshí	zǒngtǒng
布什	总统
Bush	President

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

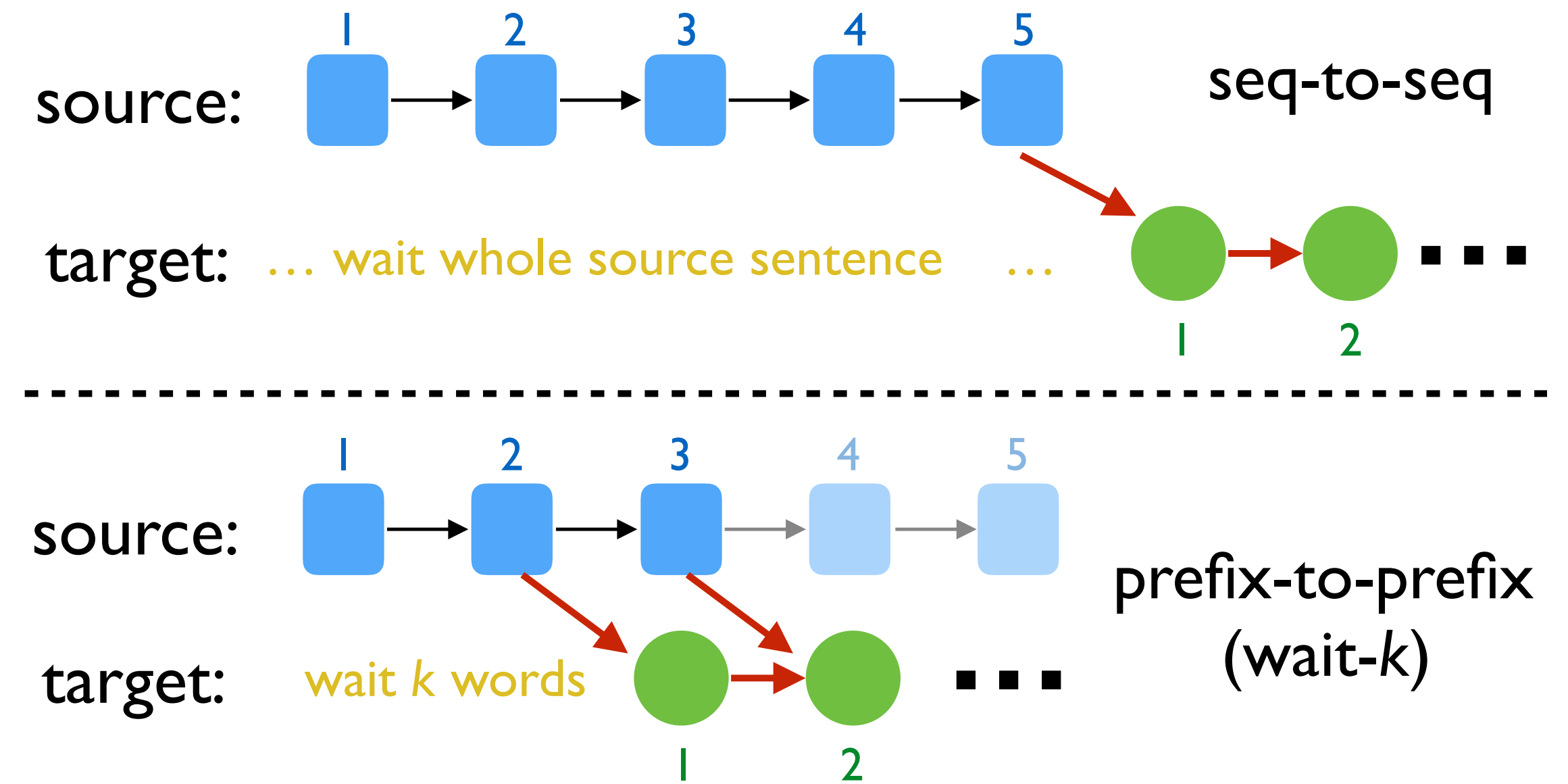


Bùshí	zǒngtǒng	zài
布什	总统	在
Bush	President	in

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

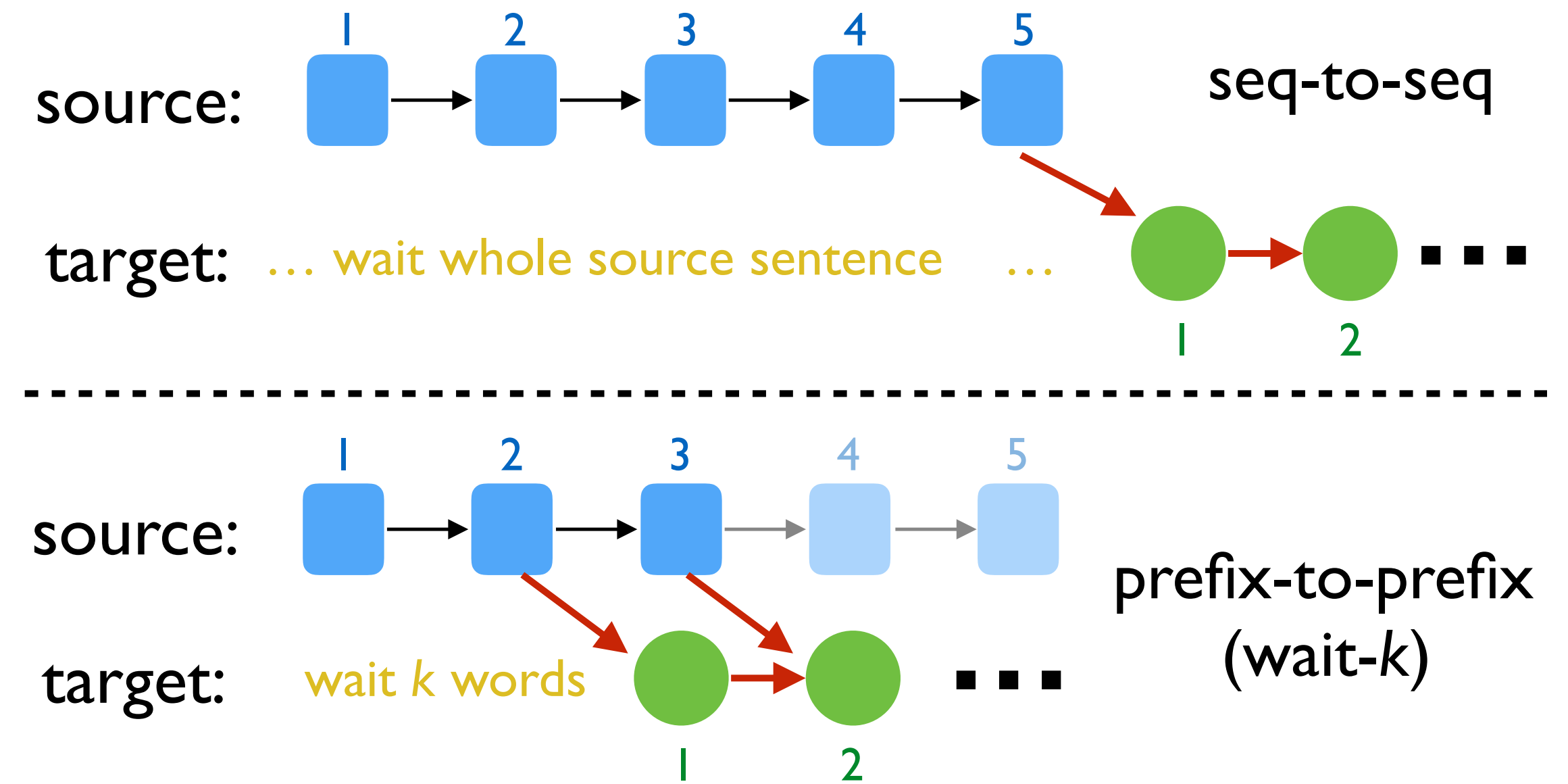


Bùshí	zǒngtǒng	zài	Mòsīkē
布什	总统	在	莫斯科
Bush	President	in	Moscow

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

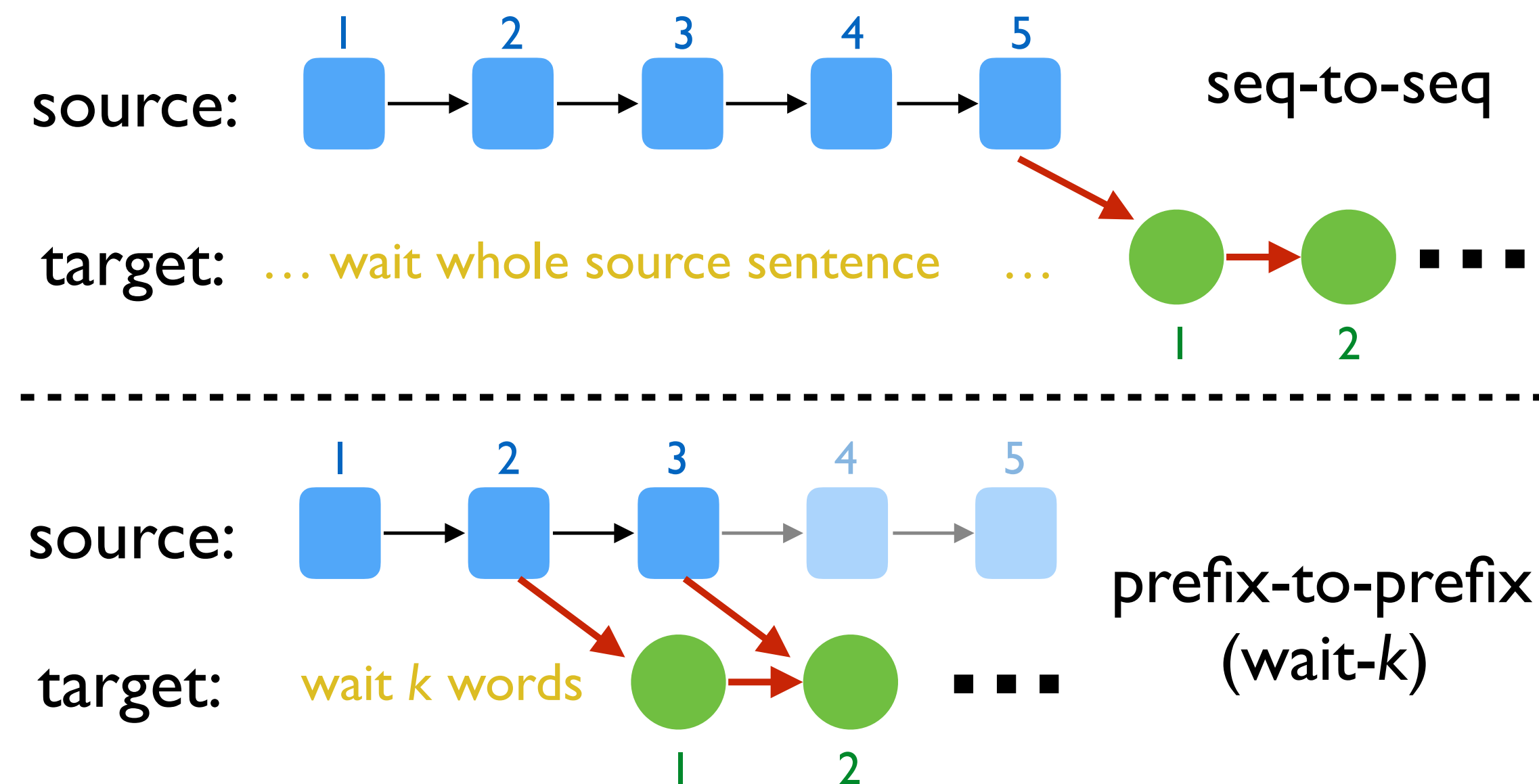


Bùshí	zǒngtǒng	zài	Mòsīkē	yǔ
布什	总统	在	莫斯科	与
Bush	President	in	Moscow	with

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

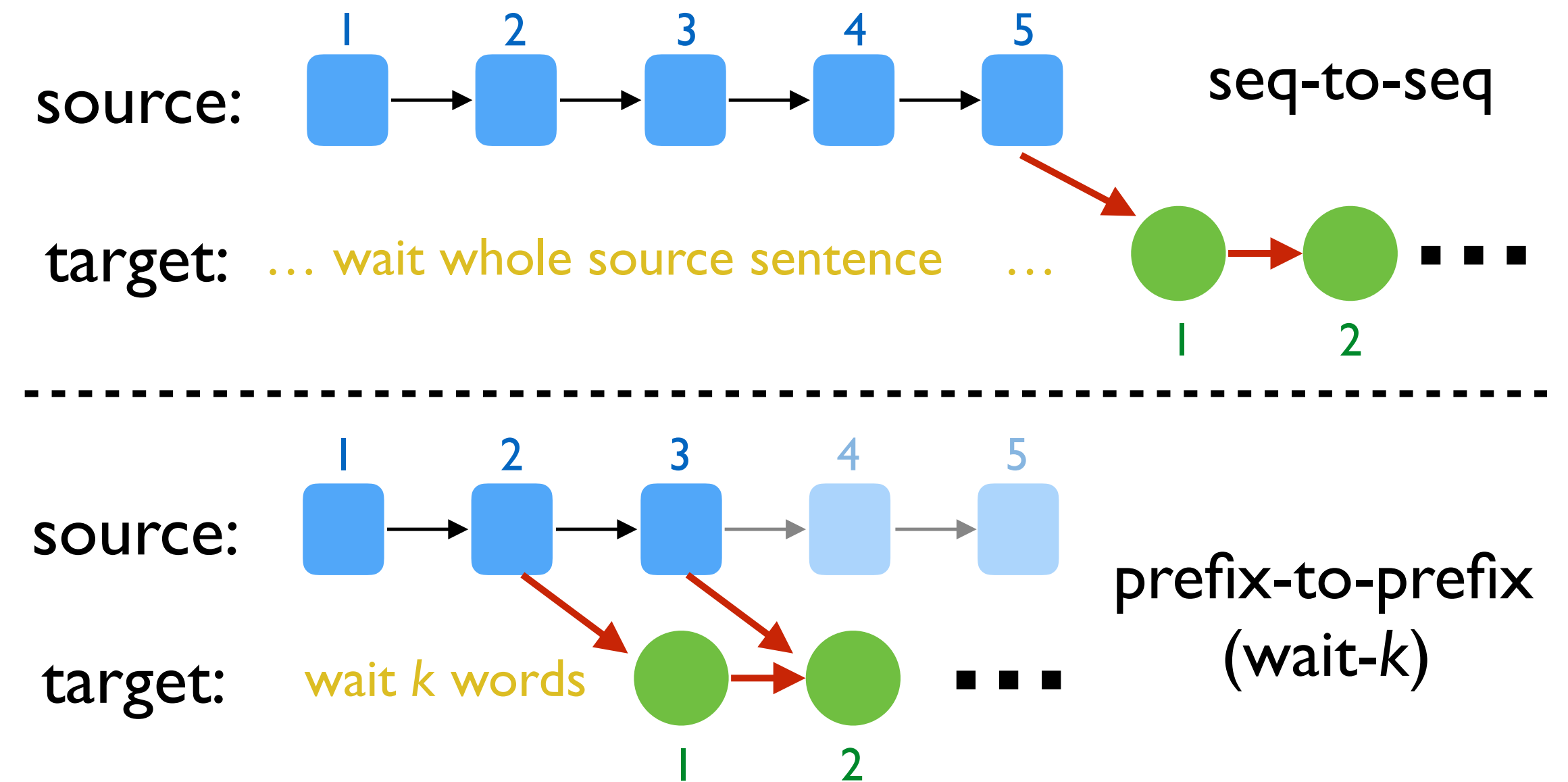


Bùshí	zǒngtǒng	zài	Mòsīkē	yǔ	Éluósī
布什	总统	在	莫斯科	与	俄罗斯
Bush	President	in	Moscow	with	Russian

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

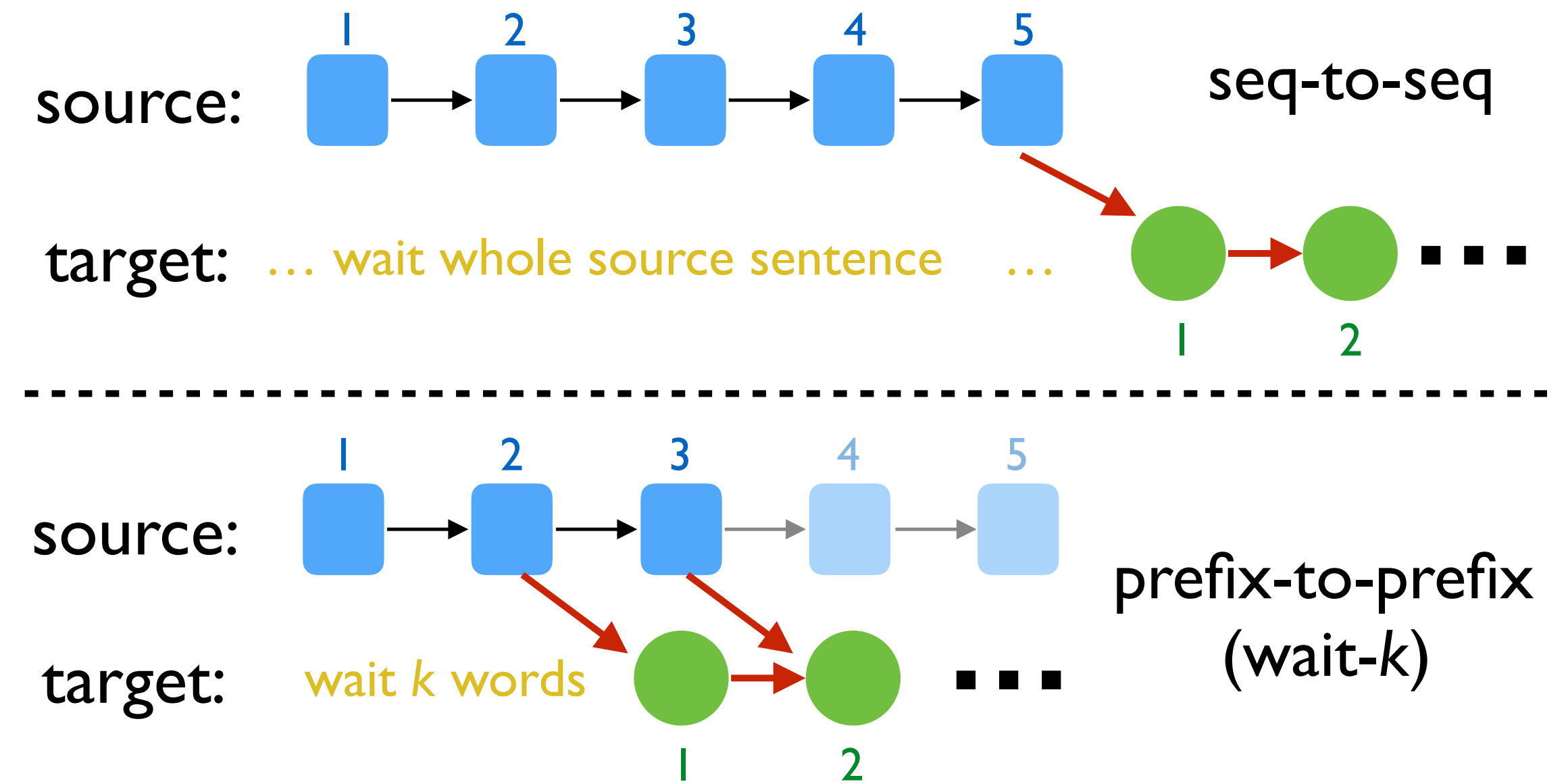


Bùshí	zǒngtǒng	zài	Mòsīkē	yǔ	Éluósī	zǒngtǒng
布什	总统	在	莫斯科	与	俄罗斯	总统
Bush	President	in	Moscow	with	Russian	President

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

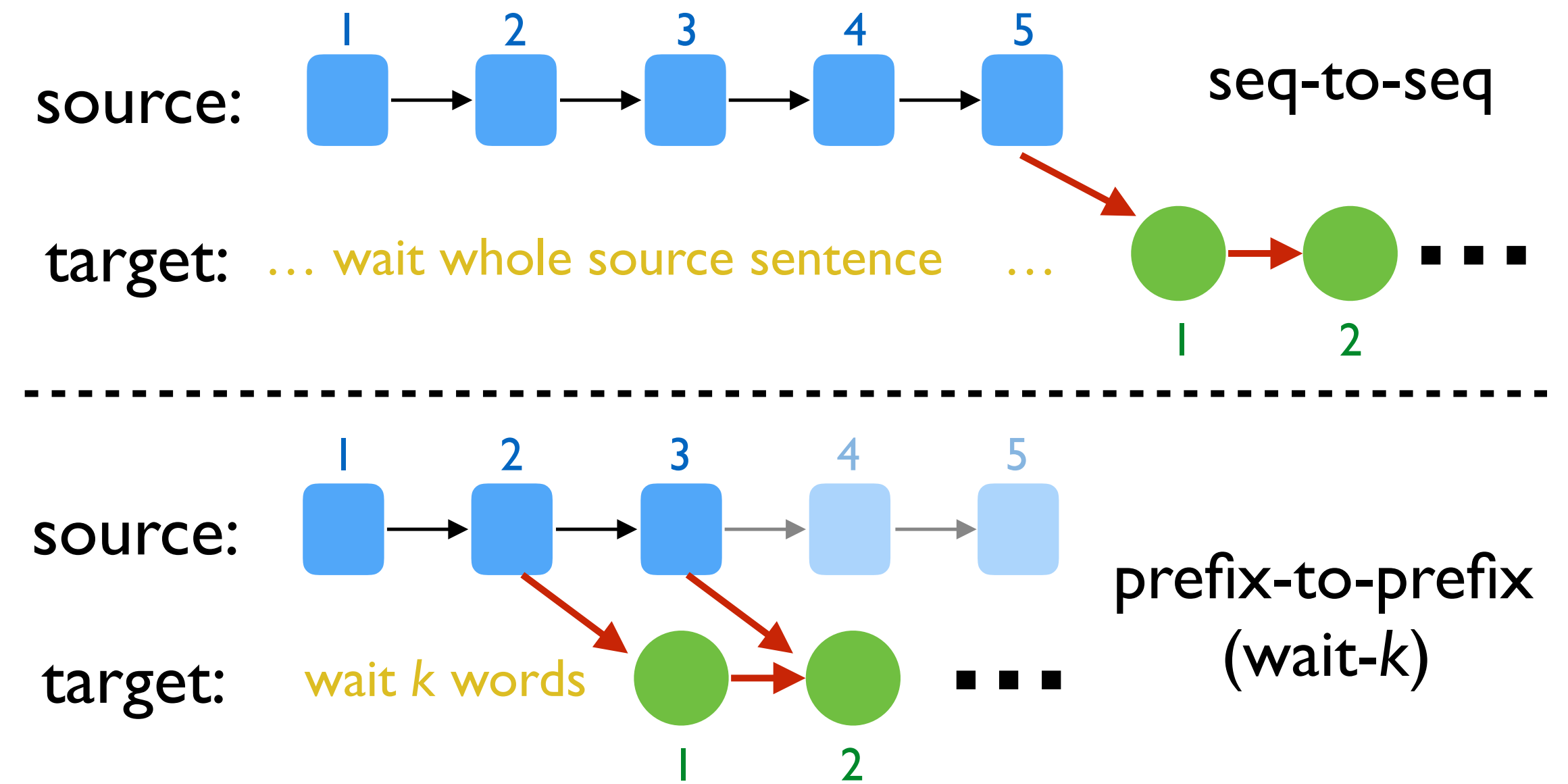


Bùshí	zǒngtǒng	zài	Mòsīkē	yǔ	Éluósī	zǒngtǒng	Pǔjīng
布什	总统	在	莫斯科	与	俄罗斯	总统	普京
Bush	President	in	Moscow	with	Russian	President	Putin

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)
 $p(y_t \mid \mathbf{x}_1 \dots \mathbf{x}_n, y_1 \dots y_{t-1})$
- prefix-to-prefix (given source prefix)
 $p(y_t \mid \mathbf{x}_1 \dots \mathbf{x}_{g(t)}, y_1 \dots y_{t-1})$
 $g(\cdot)$ is a monotonic non-decreasing function
 $g(t)$: num. of source words used to predict y_t

t=3							
	President	Bush	meets	with	Putin	in	Moscow
Bush	布什总统		g(3) = 4				
Pres.							
at	在						
Moscow	莫斯科						
with	与						
Putin	普京						
meet	会晤						

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 zé mǐn duì fǎ guó zǒng tǒng de
江 泽 民 对 法 国 总 统 的
jiang zemin to French President's

lái huá fǎng wèn
来 华 访 问
to-China visit

biǎo shì gǎn xiè
表 示 感 谢 。
express gratitude

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

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 duì 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 (greedy):	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 duì 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 (greedy):	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)



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):

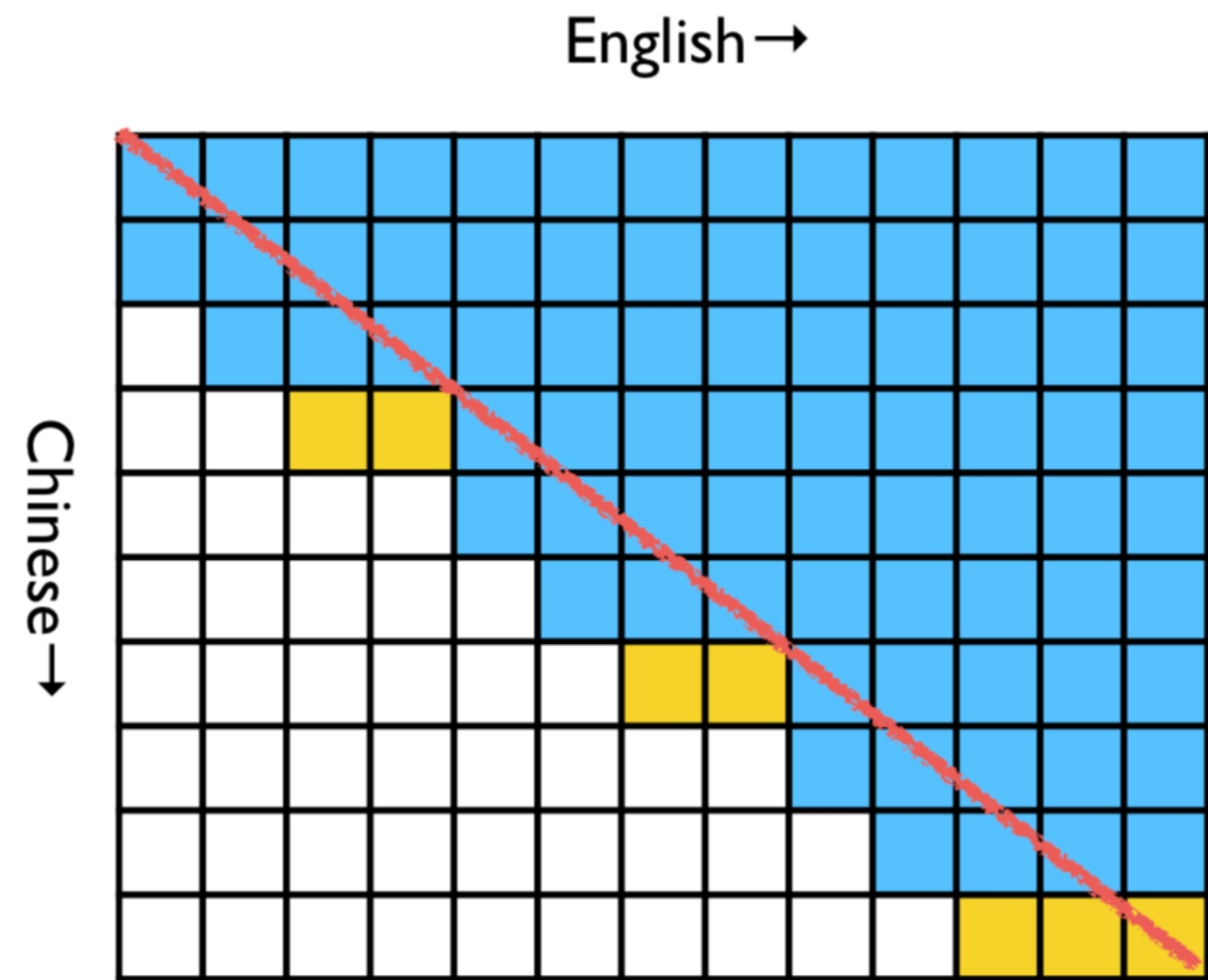
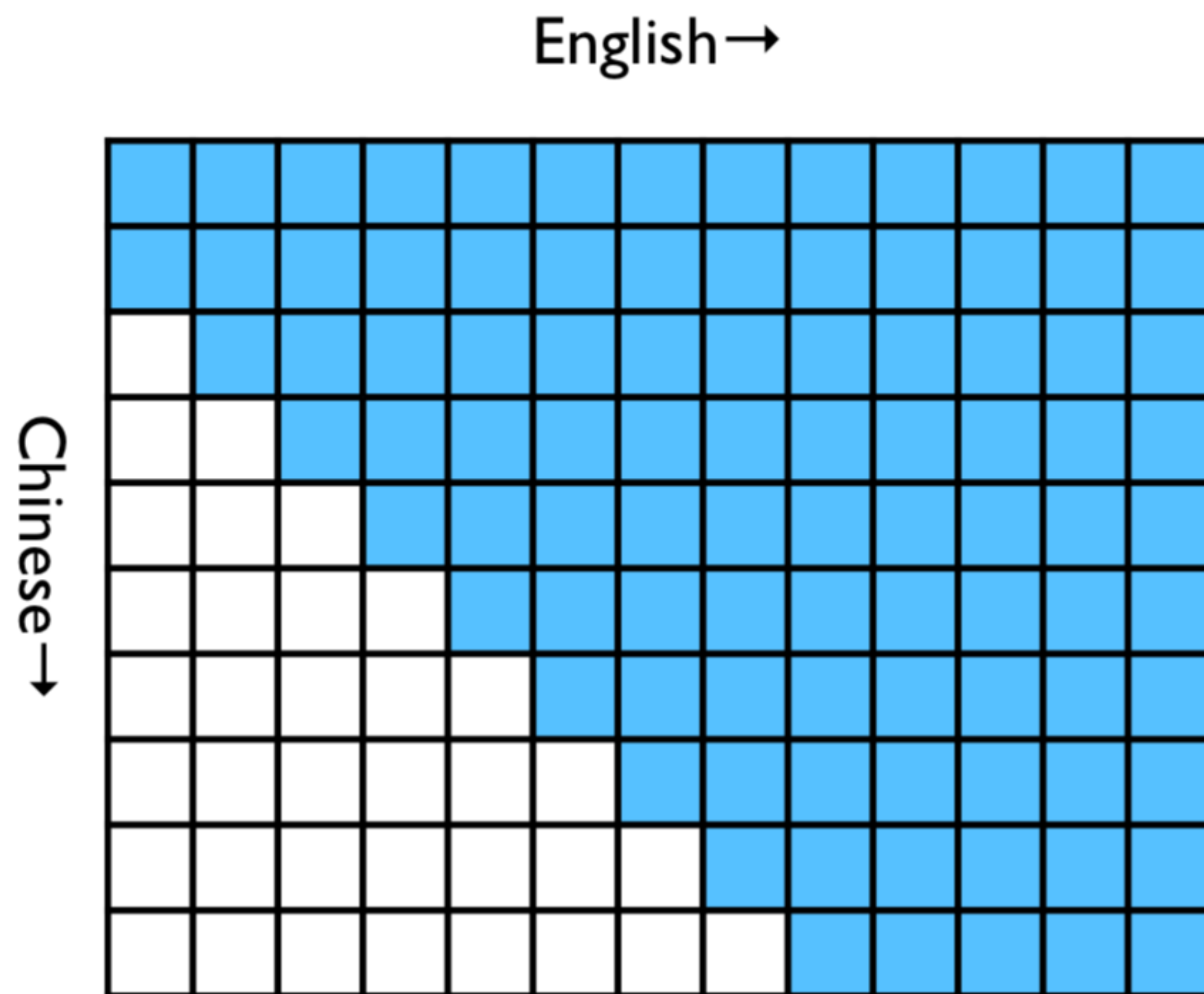
but , while congress **does not agree** on a **course of action** , several states **no** longer **wait** .

English translation (full-sentence beam search):

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

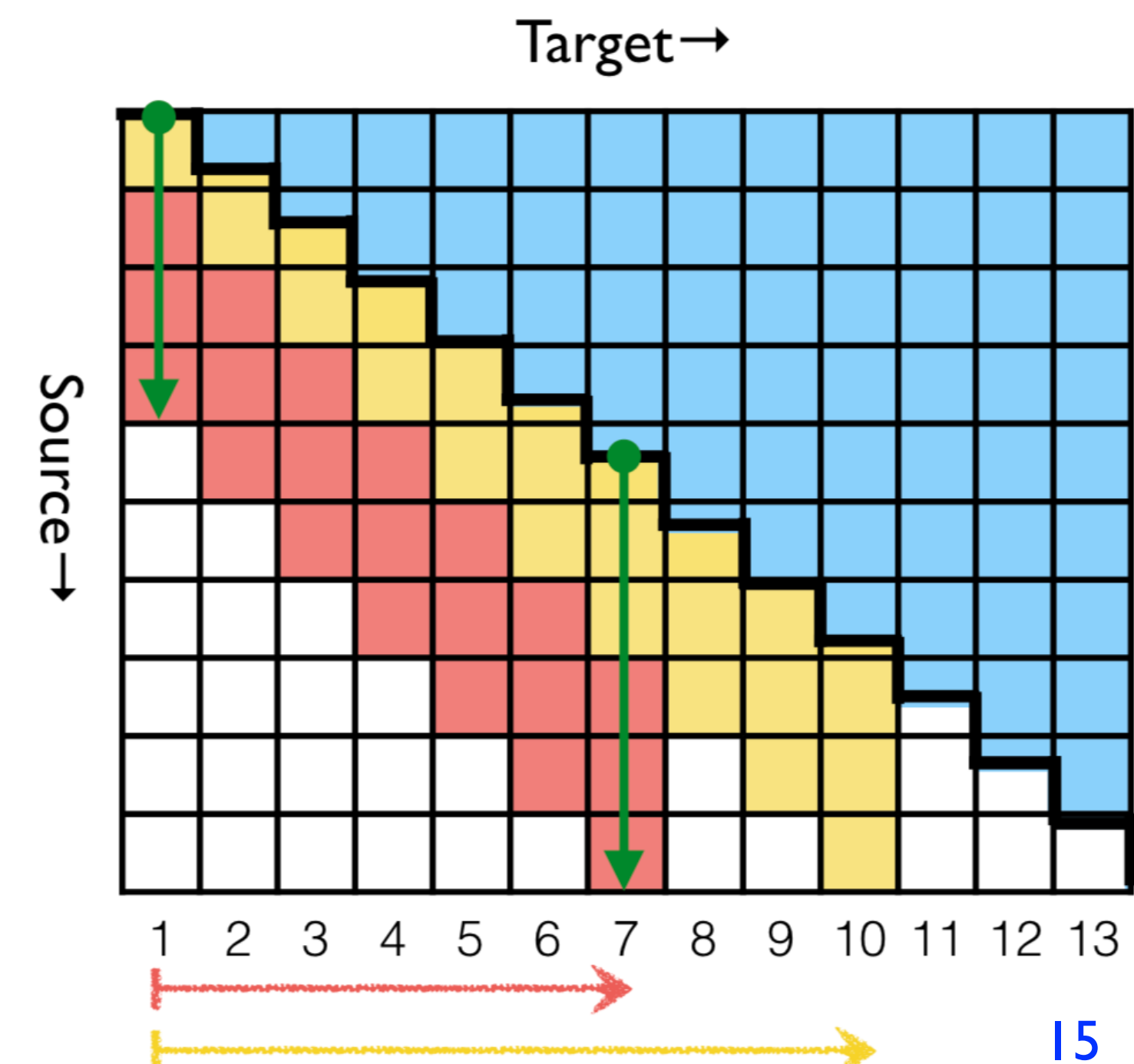
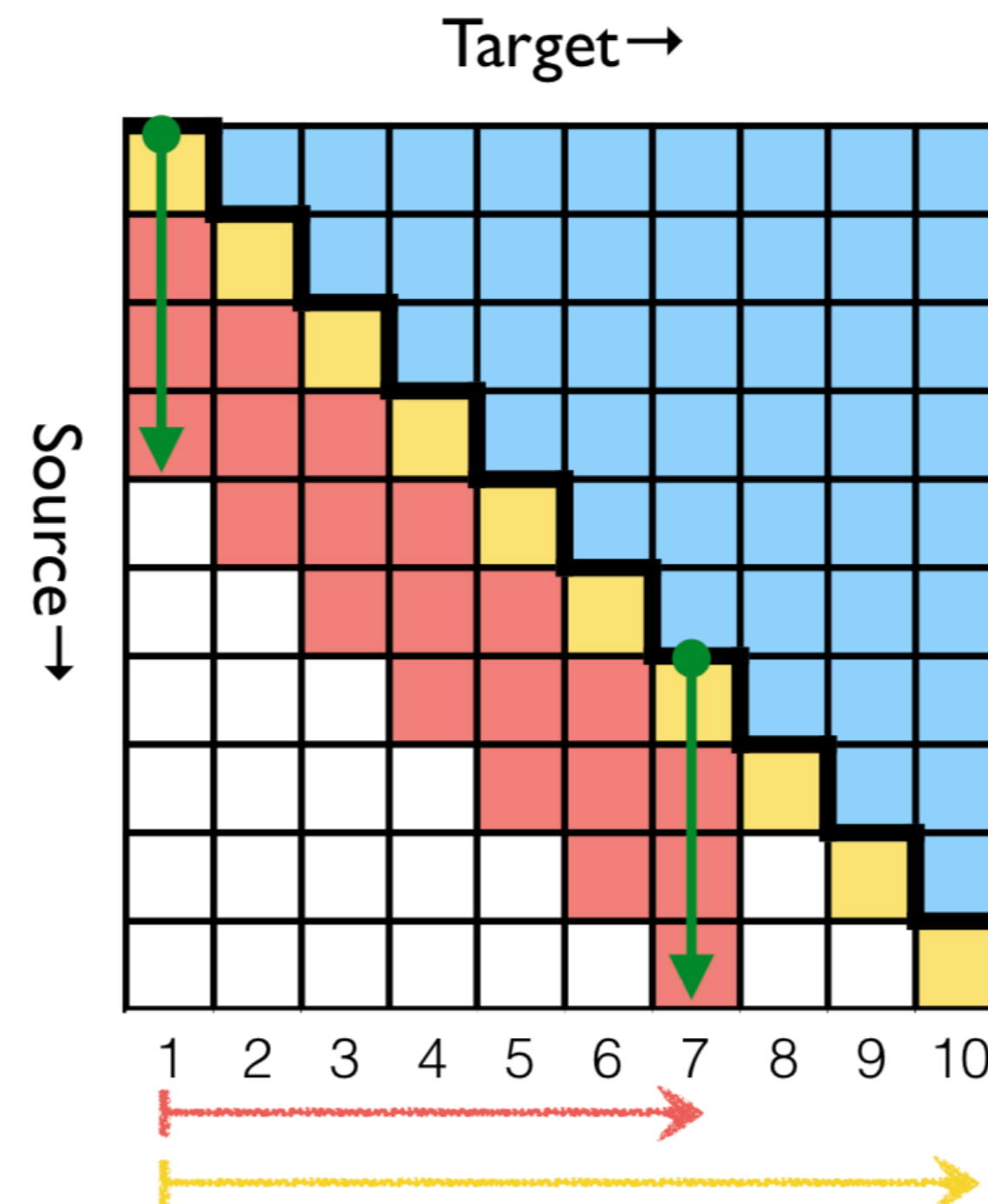
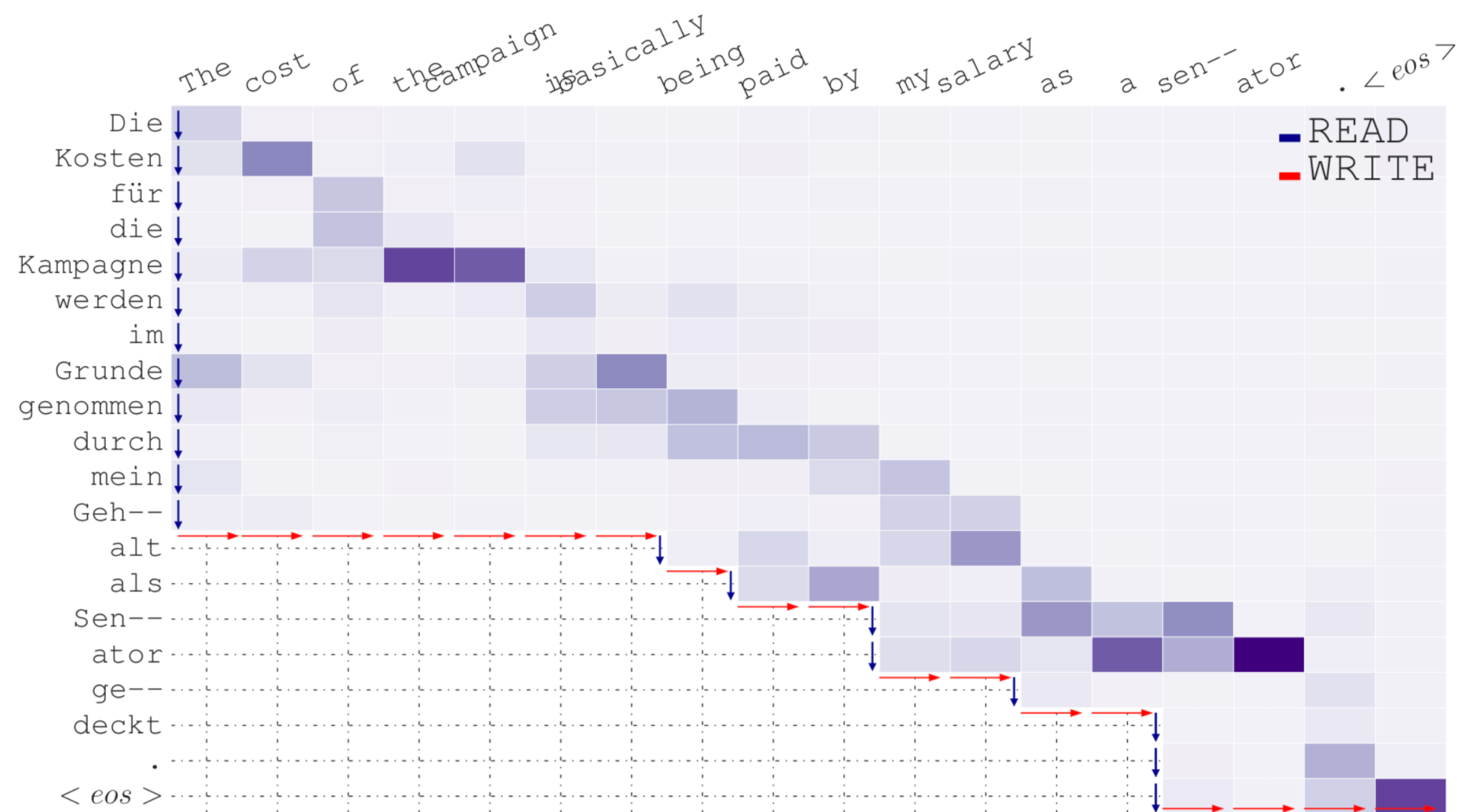
Refinements: Wait- k with Catchup

- English translation length is often $\sim 1.25\times$ 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



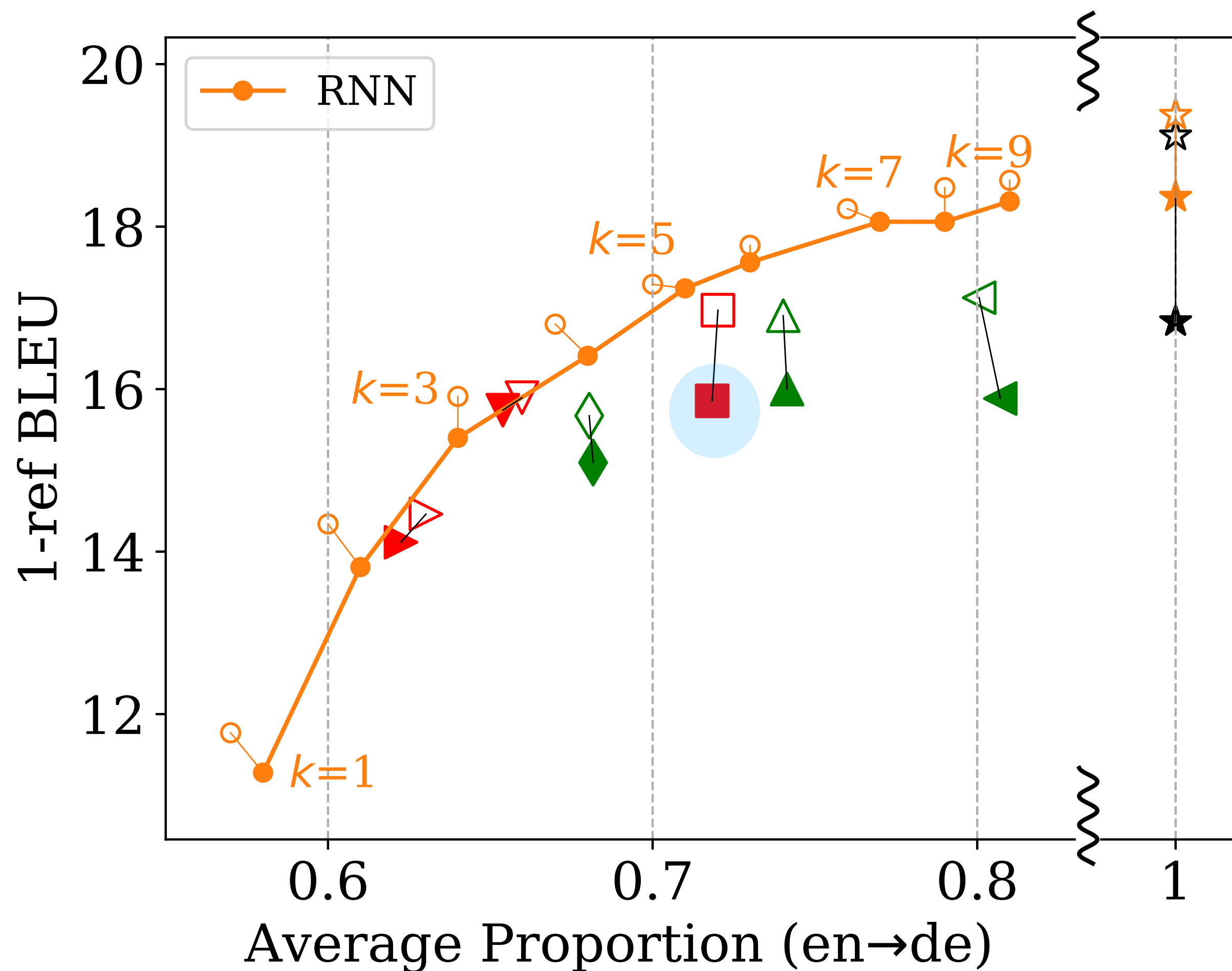
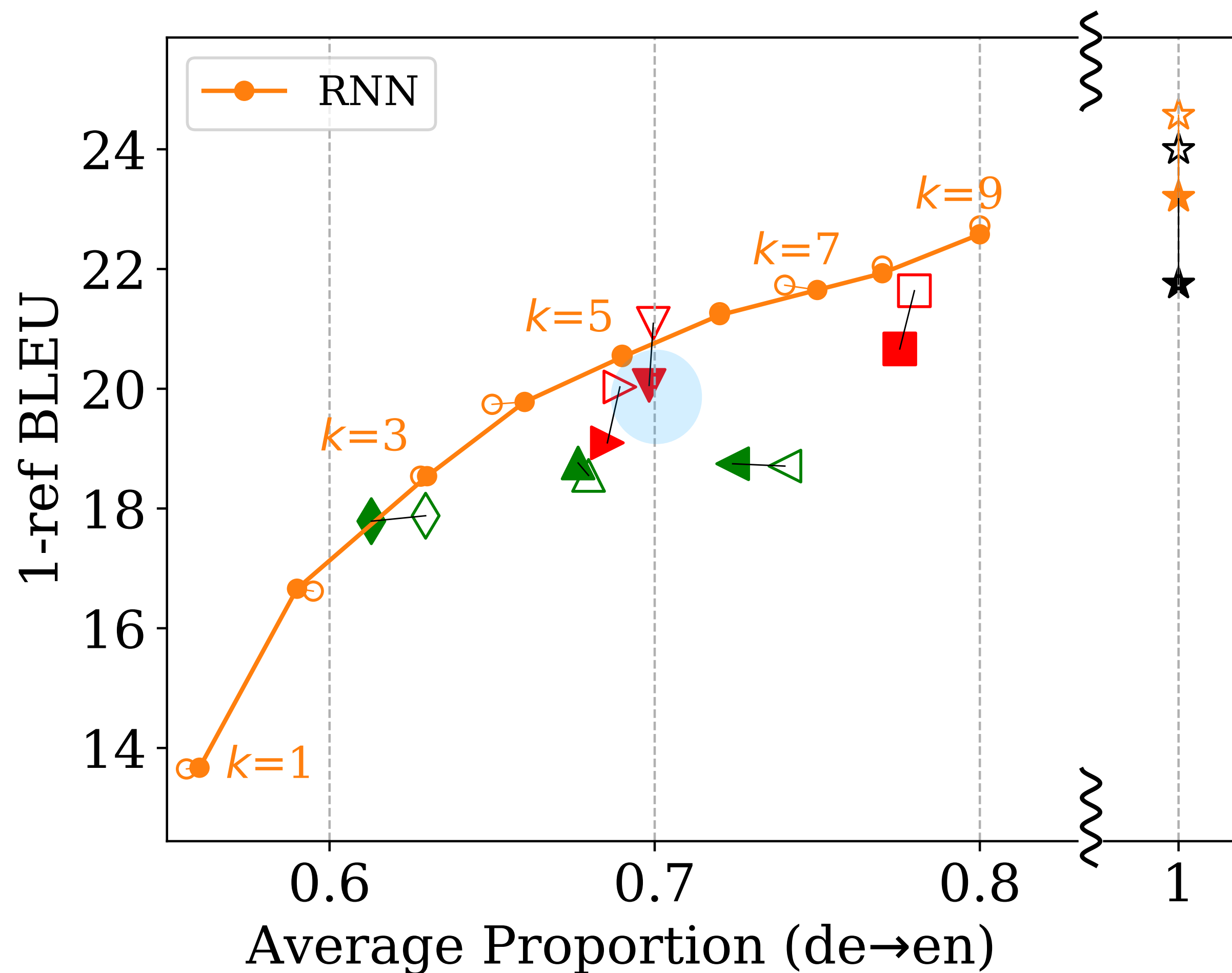
New Latency Metric: Average Lagging

- 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, $AL(\text{wait-}k \text{ with catchup}) \approx k$



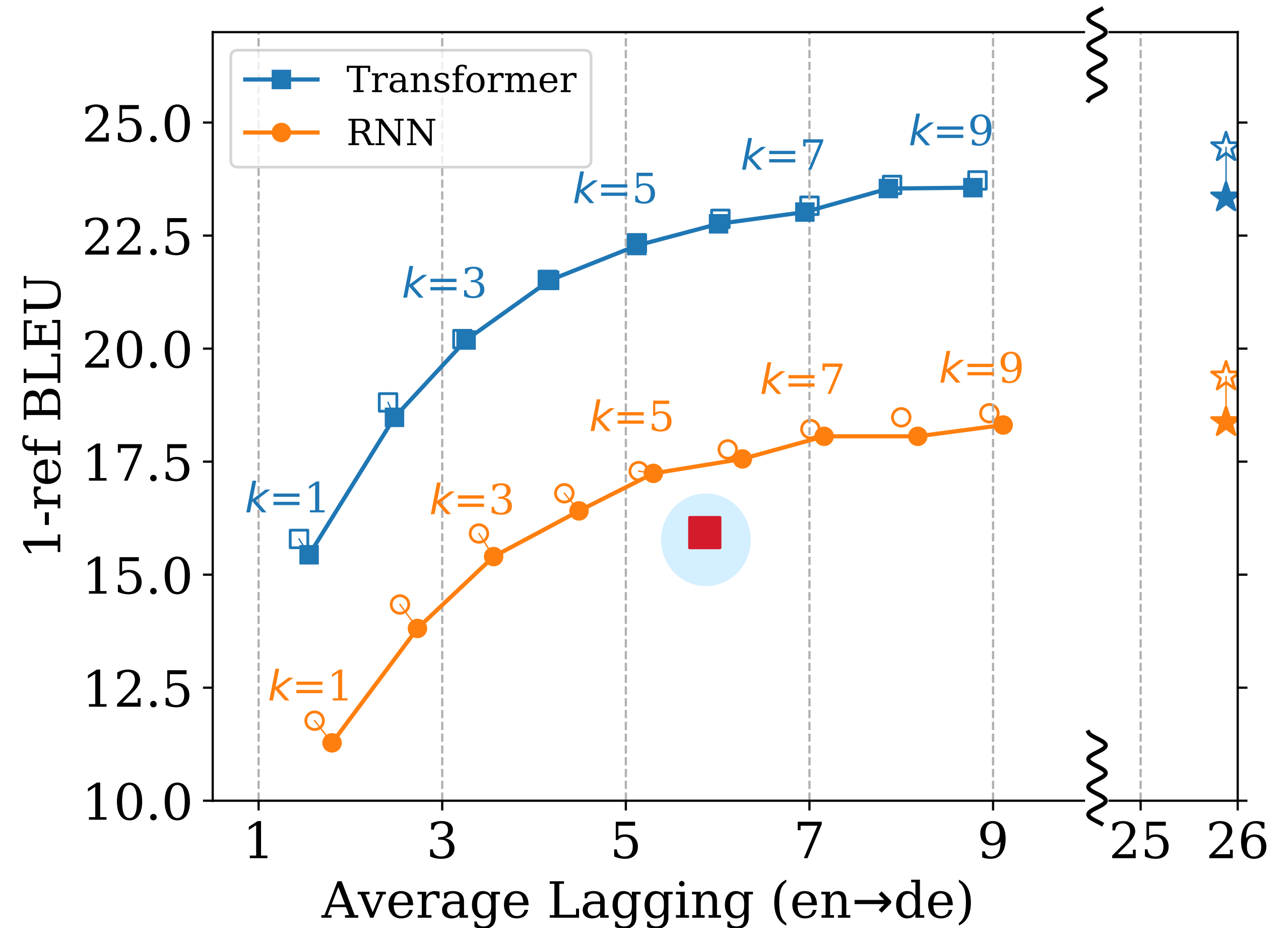
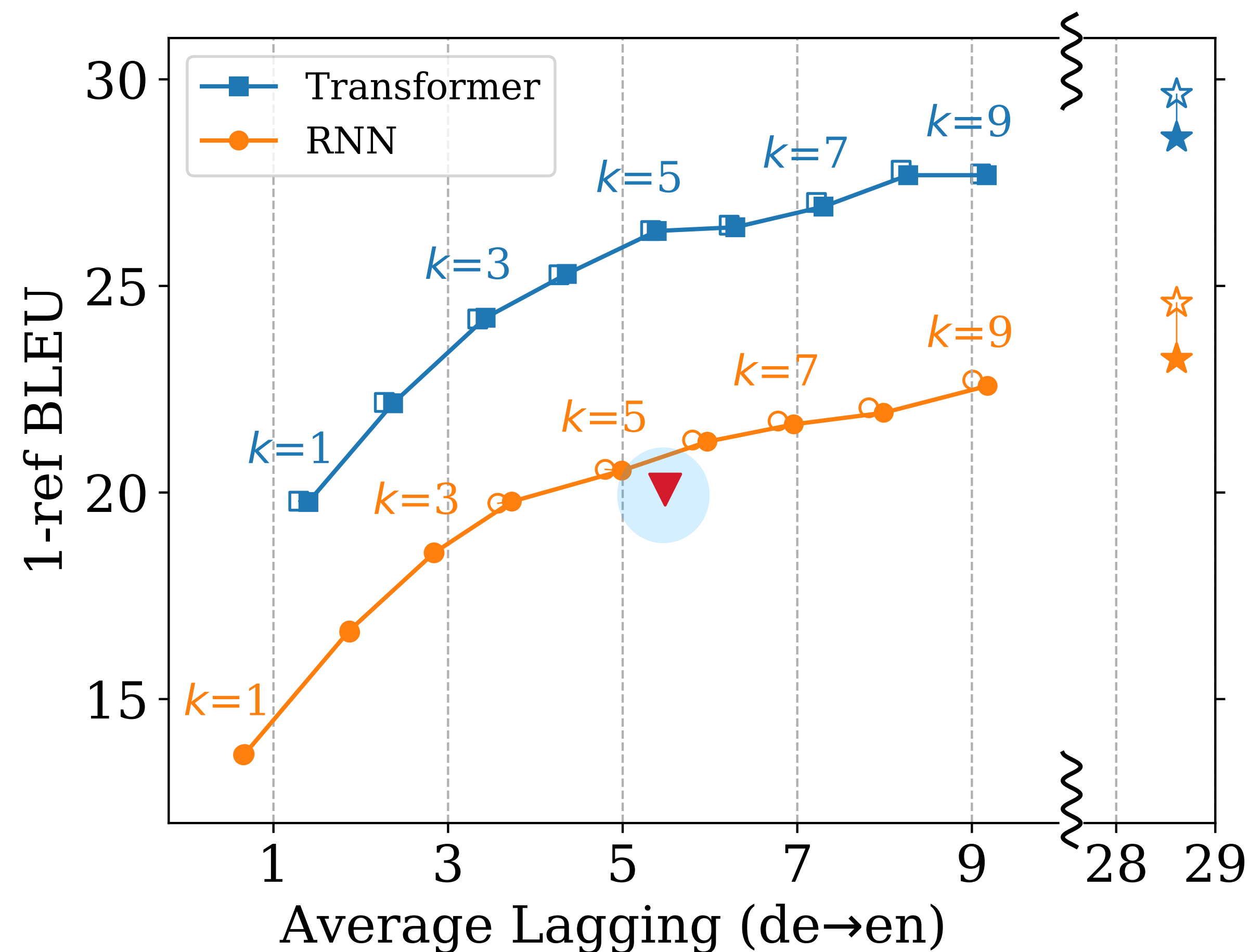
Experiments: German $\leftrightarrow$ English

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



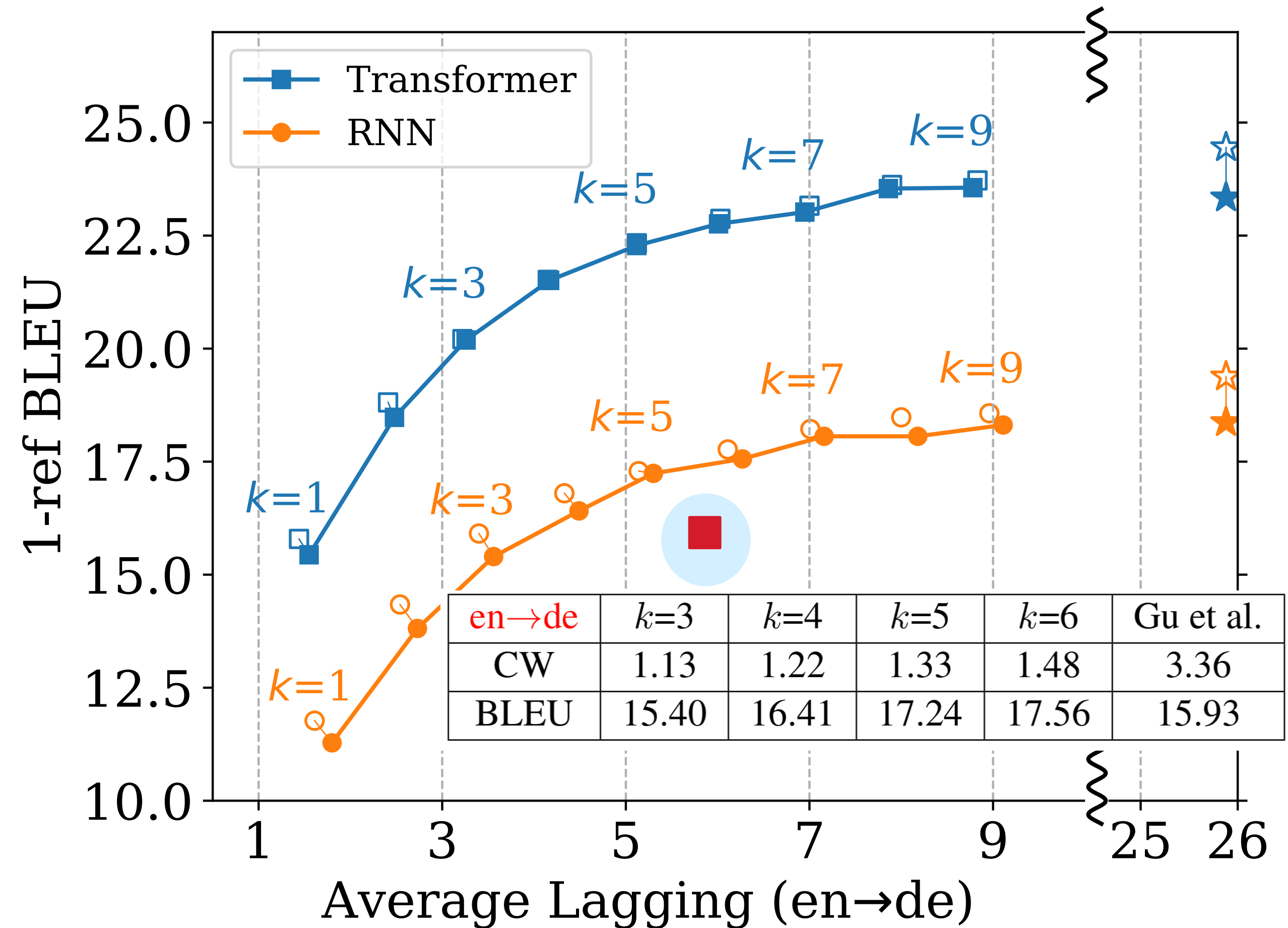
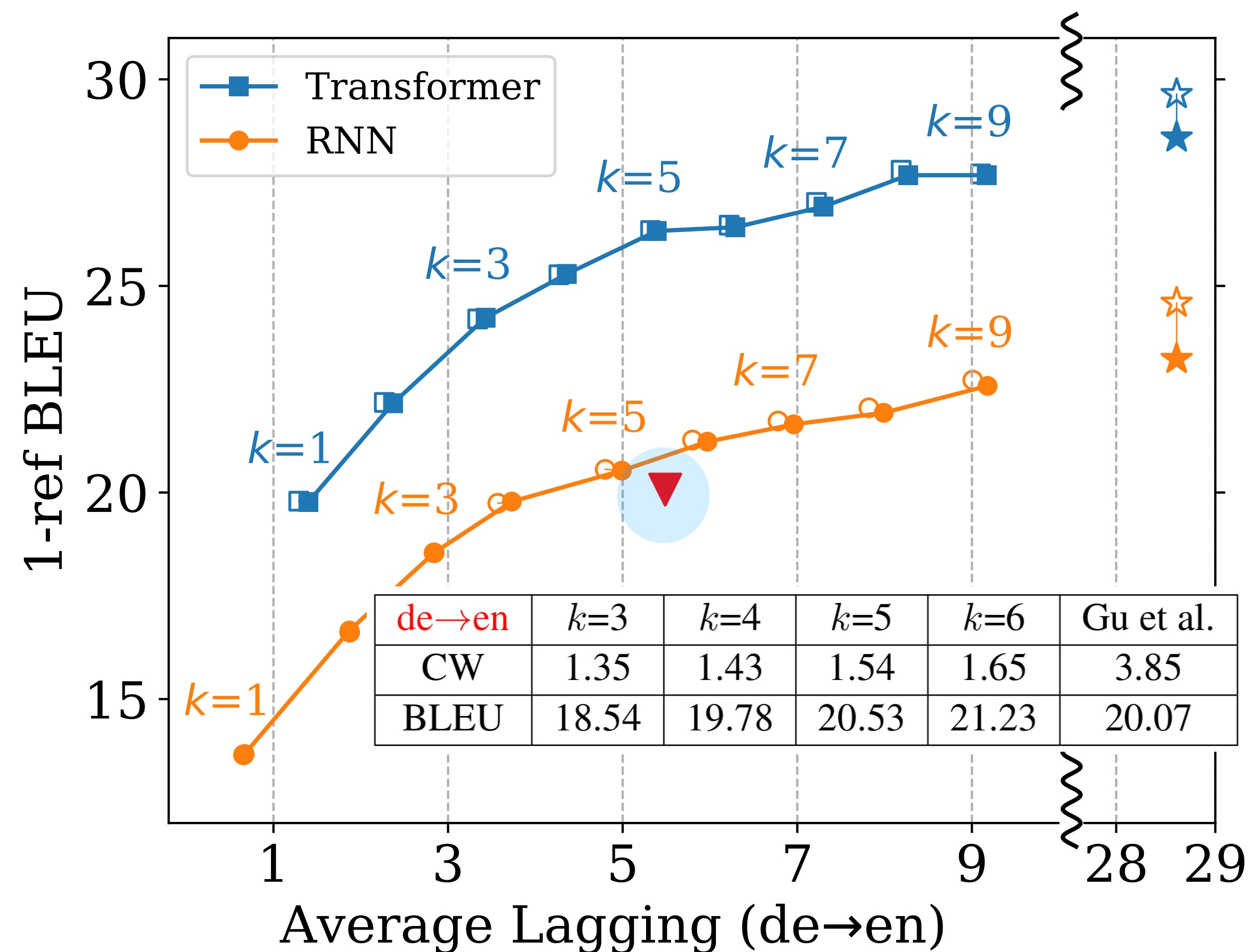
Experiments: German $\leftrightarrow$ English

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



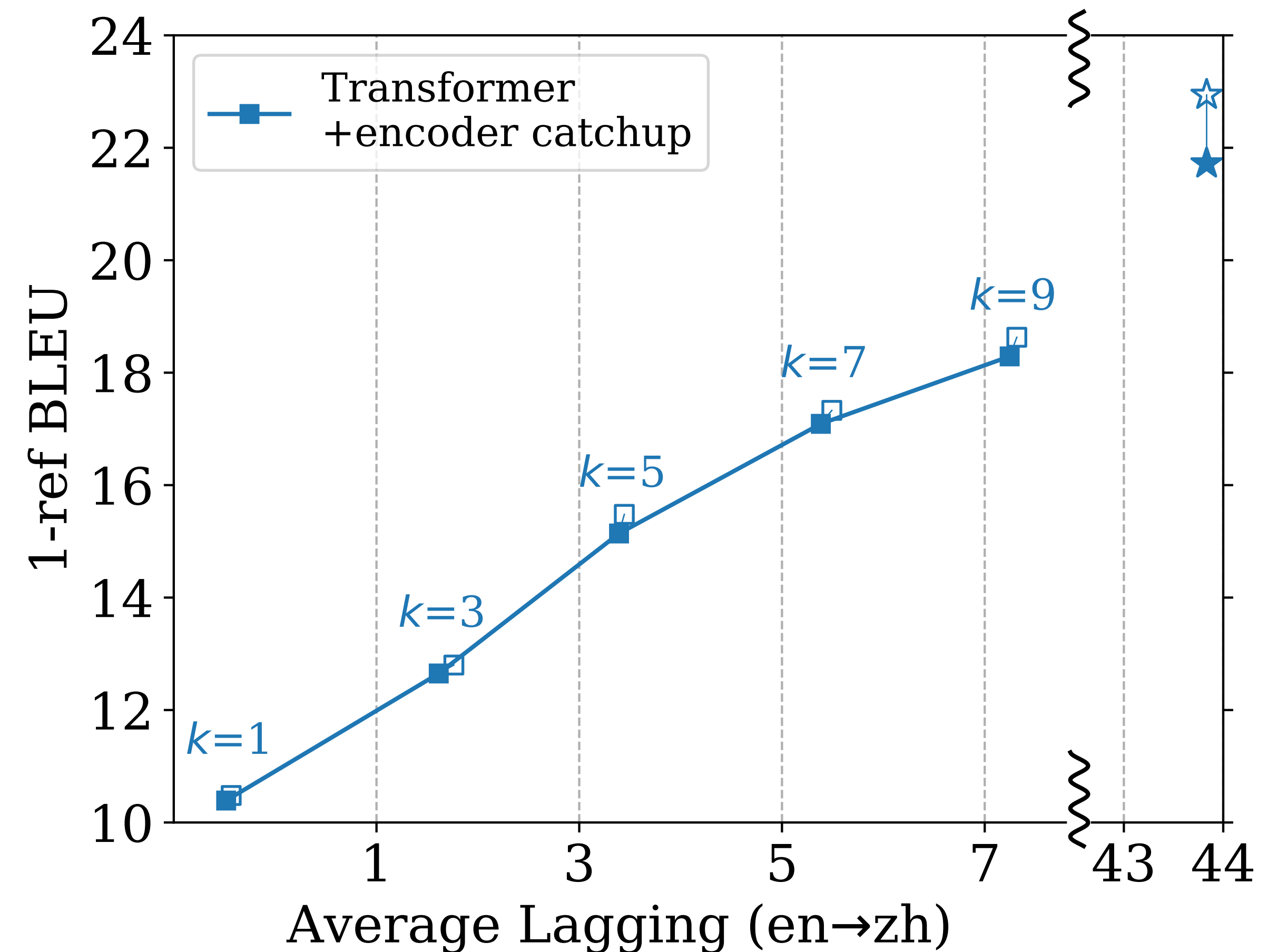
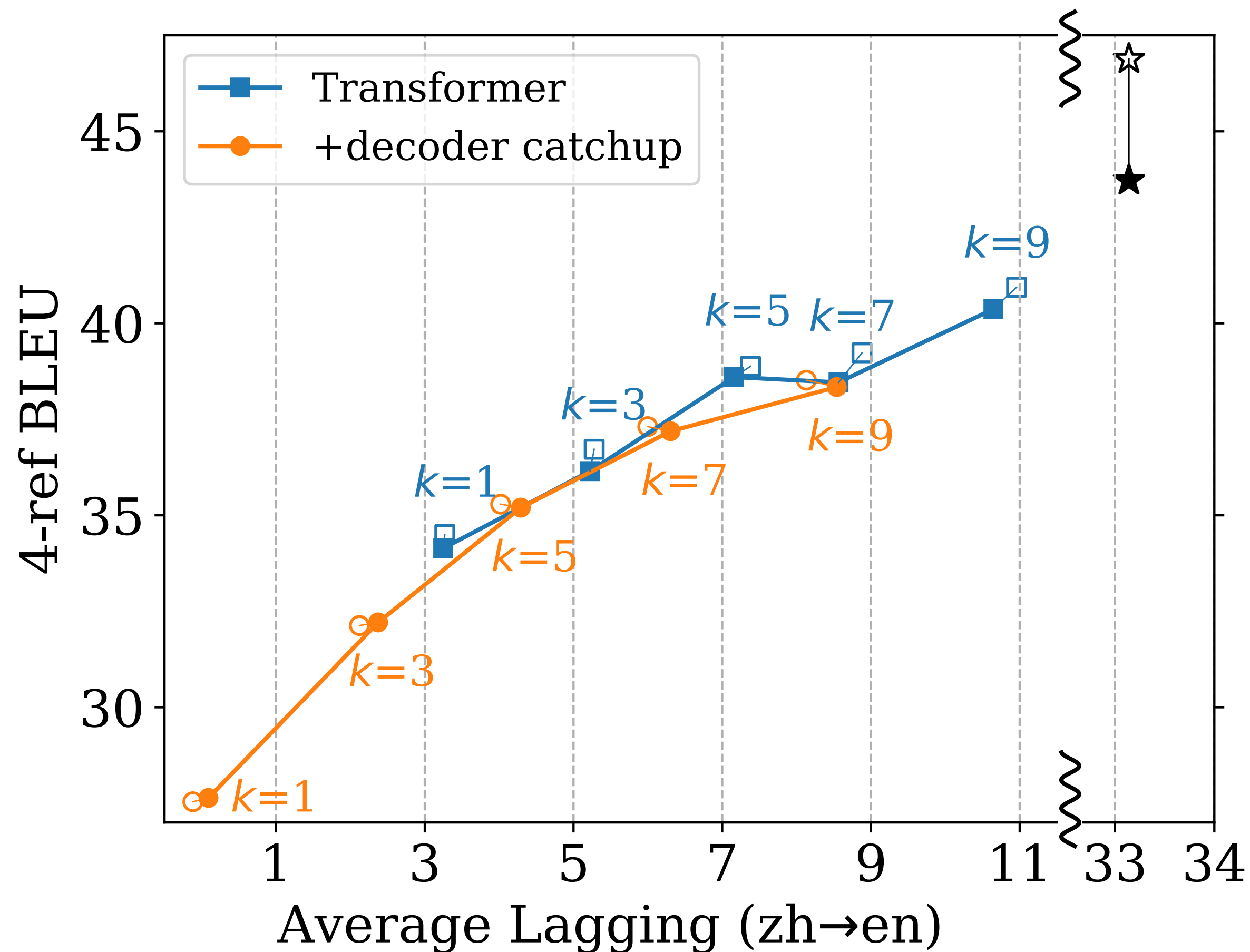
Experiments: German $\leftrightarrow$ English

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



Experiments: Chinese $\leftrightarrow$ 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	
(a)	<i>Měiguó</i>	<i>dāngjú</i>	<i>duì</i>	<i>Shātè</i>	<i>jìzhě</i>	<i>shīzōng</i>	<i>yī</i>	<i>àn</i>	<i>gǎndào</i>	<i>dānyōu</i>	
	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
$k=3^\dagger$				the	us	authorities	<u>are very</u>	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	
(a)	<i>Měiguó</i>	<i>dāngjú</i>	<i>duì</i>	<i>Shātè</i>	<i>jìzhě</i>	<i>shīzōng</i>	<i>yī</i>	<i>àn</i>	<i>gǎndào</i>	<i>dānyōu</i>	
	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
$k=3^\dagger$				the	us	authorities	are very	concerned	about	the	saudi reporter 's missing case
$k=\infty$											us authorities concerned over saudi journalists missing
(b)	美国	当局	对	沙特	记者	失踪	一	案	感到	<i>bùmǎn</i> 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
$k=\infty$											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

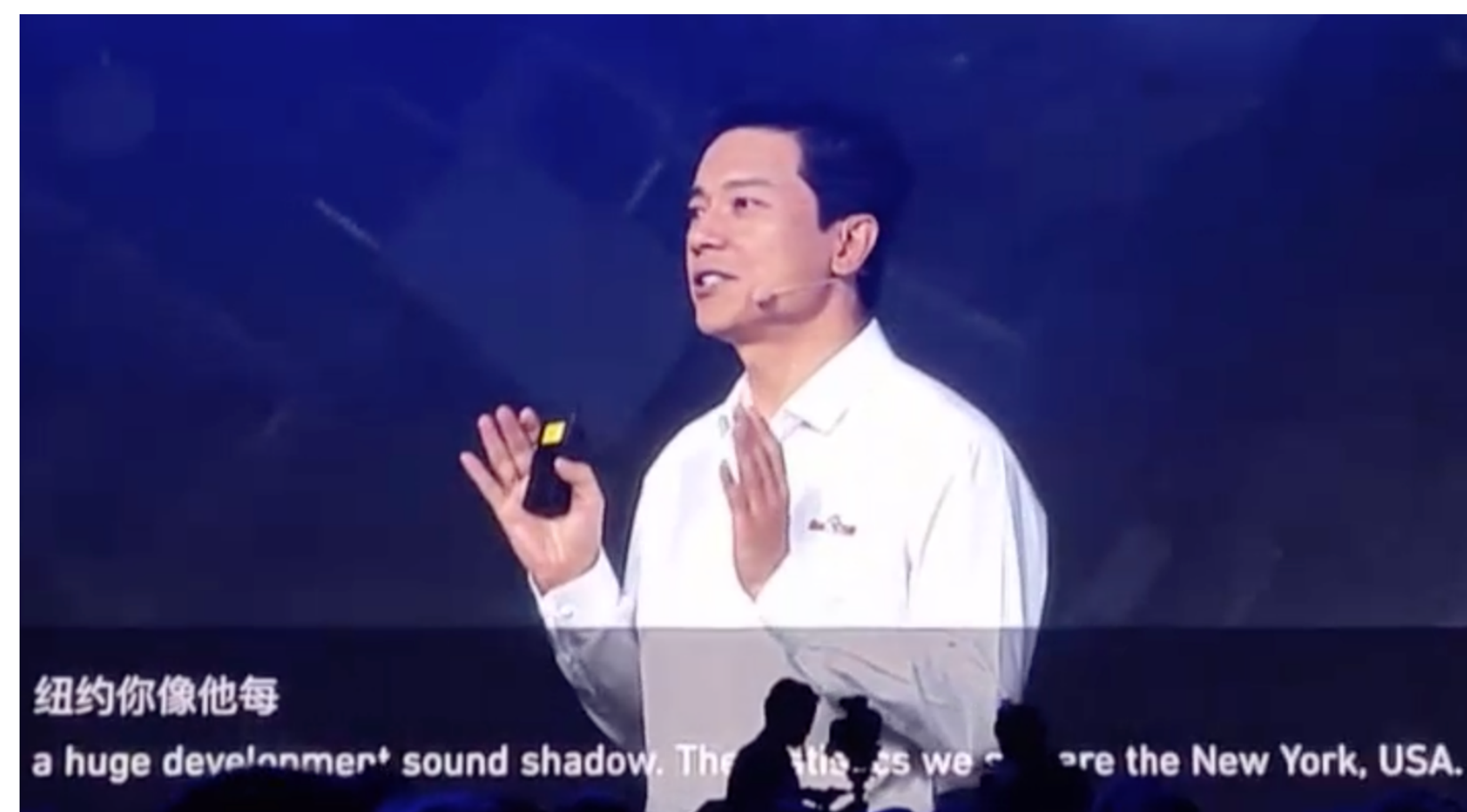
- first simultaneous translation system with seamlessly integrated anticipation
 - human simultaneous interpreters also anticipate all the time
 - some previous works predict source language verbs
 - we don't have a separate “anticipation” step, and only predict target side words
 - first simultaneous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

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

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.