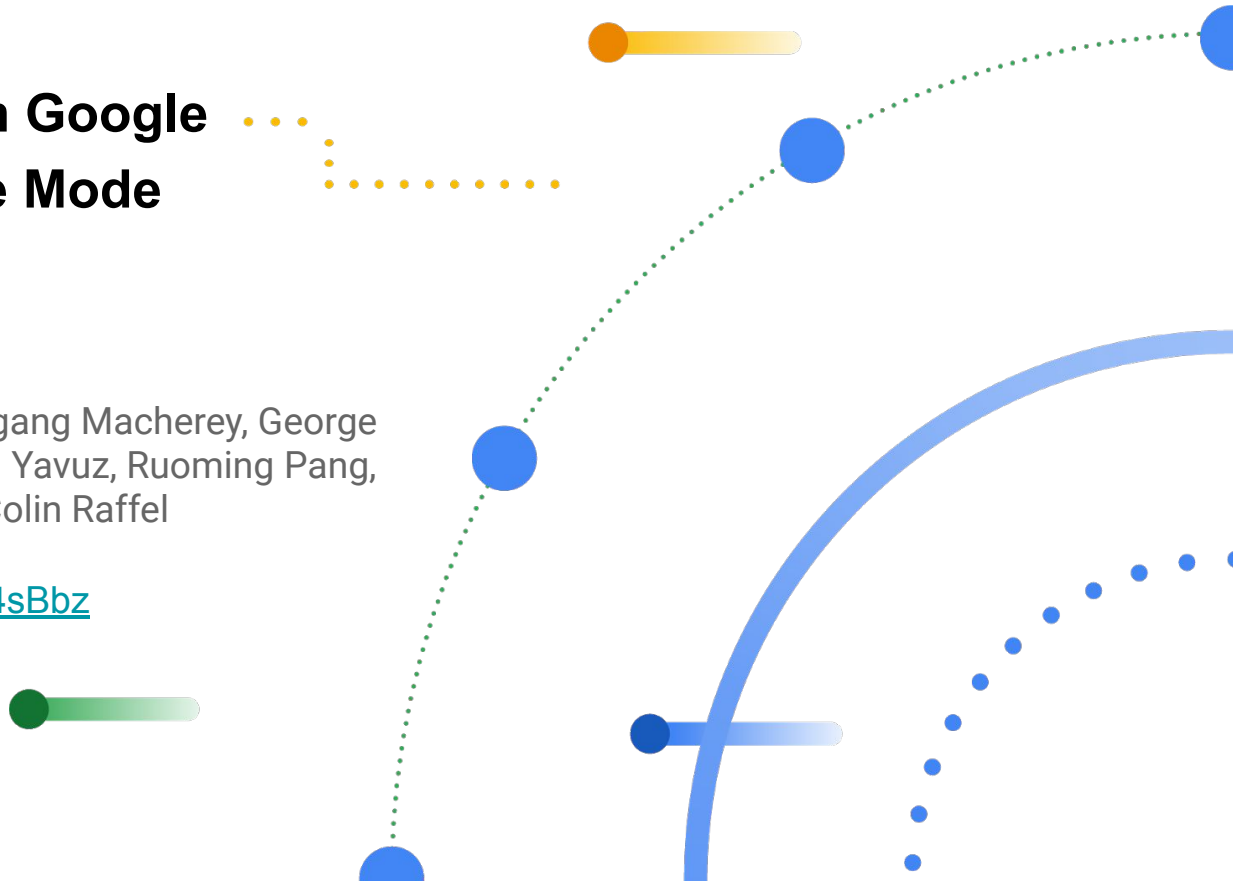
Google Research

Research stories from Google Translate's Transcribe Mode

Colin Cherry

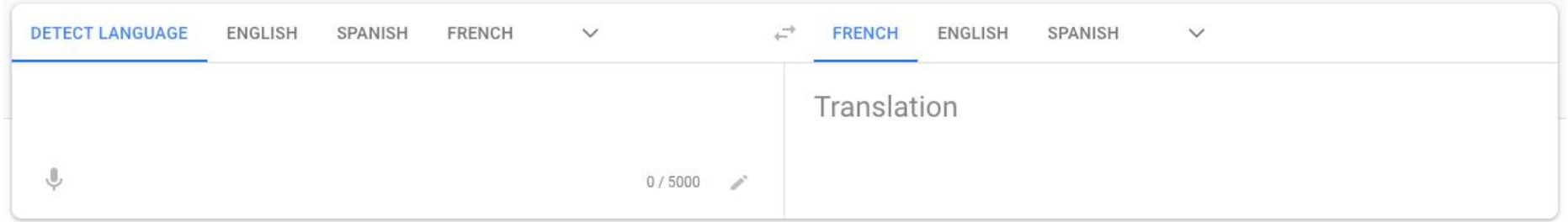
Joint work with Naveen Ari, Wolfgang Macherey, George Foster, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, Te I, Pallavi Baljekar and Colin Raffel

Google slides: https://t.co/9dOZj4sBbz



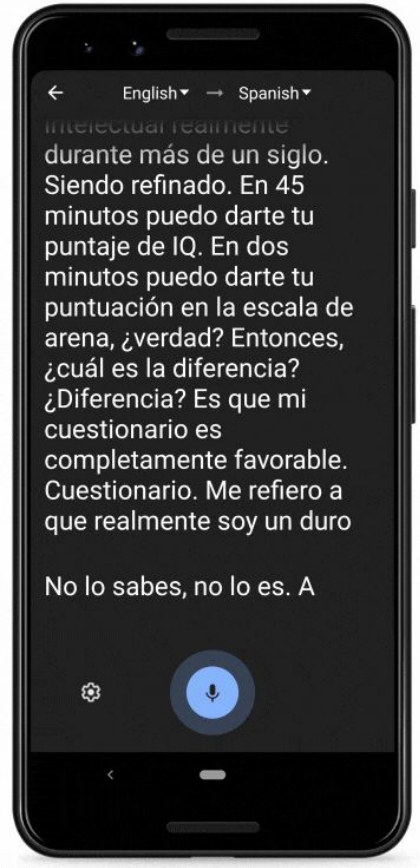
Google Translate

Google Research

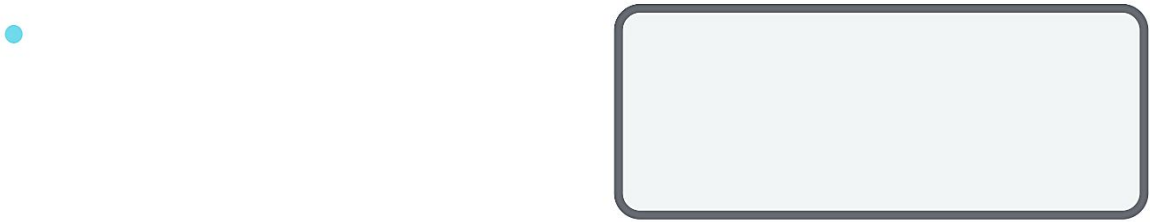


Transcribe Mode

- The user is listening to someone speaking in a language they do not understand
  - A lecture they’re attending
  - Their grandmother telling a story
- The mic is left on, and they can read the translation in real time.



Translating sentence by sentence is too slow



We want to understand how to translate a sentence in progress.

Abstraction: Simultaneous Translation as Control

- In the beginning, there is nothing, just an empty source sequence

- A streaming agent has two possible actions:
  - **Read** a source token (waiting if one is not available)
  - **Write** a target token to extend the translation

- Joint goal: minimize **latency**, maximize **quality**
  - That is: **read** just enough to **write** accurately

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

The

Read

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

The big

Read

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

The big red

Read

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

The big red dog

Read

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

The big red dog

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

Write

Le

The big red dog

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

Write

Le gros

The big red dog

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

Write

Le gros chien

The big red dog

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

Write

Le gros chien rouge

The big red dog

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

Le gros chien rouge

The big red dog loves

Read

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

Write

Le	gros	chien	rouge	aime
The	big	red	dog	loves

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

Le gros chien rouge aime

The big red dog loves Emily

Read

Idealized example

Text in Black
Operations in Blue

Target on Top
Source on Bottom

					Write
Le	gros	chien	rouge	aime	Emily
The	big	red	dog	loves	Emily

Previous work

2006-2016
Phrase-based techniques (often segmentation-driven) from KIT, AT&T, NAIST and others

[Don't Until the Final Verb Wait: Reinforcement Learning for Simultaneous Machine Translation]
Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, Hal Daumé III; 2014

[Can neural machine translation do simultaneous translation?]
Kyunghyun Cho, Masha Esipova; 2016

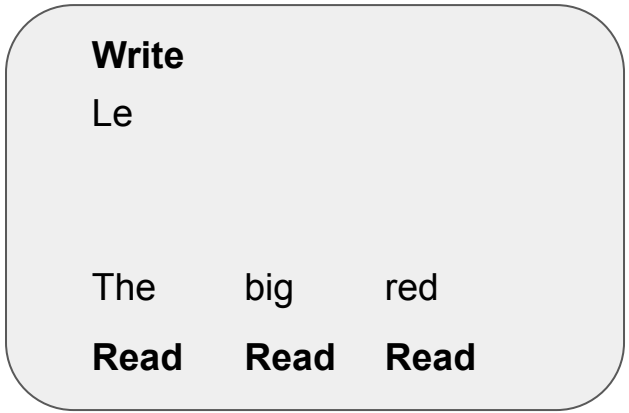
[Learning to Translate in Real-time with Neural Machine Translation Strategies]
Jiatao Gu, Graham Neubig, Kyunghyun Cho, Victor O.K. Li; 2017

Enter STACL and wait-k (Ma et al., 2019 https://arxiv.org/abs/1810.08398)

(See also: Dalvi et al., 2018 https://www.aclweb.org/anthology/N18-2079/)

- Extremely simple agent:
  - Read k source tokens, then alternate write and read until end of source

Example:
k=3

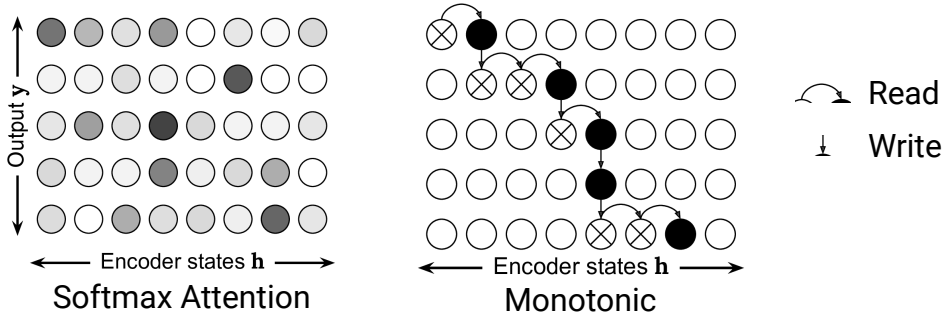


- Integrates nicely into training:
  - Leads to ability to **anticipate**!

- Cleaner evaluation metric for latency:
  - **Average lagging:** how far do we lag behind an ideal word-for-word system?

Monotonic Attention: (Raffel et al., 2017 https://arxiv.org/abs/1704.00784)

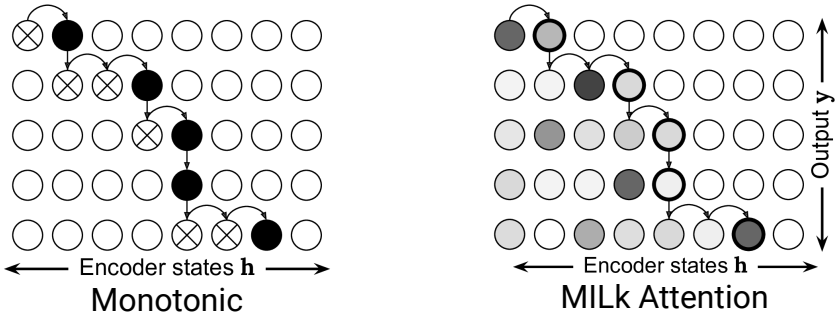
- Attention links the decoder to the encoder by a softmax over encoder positions
- Monotonic attention models a series of read-write decisions



- Trainable in expectation with dynamic programming
  - NMT learns to **anticipate**; agent learns to **adapt** to context

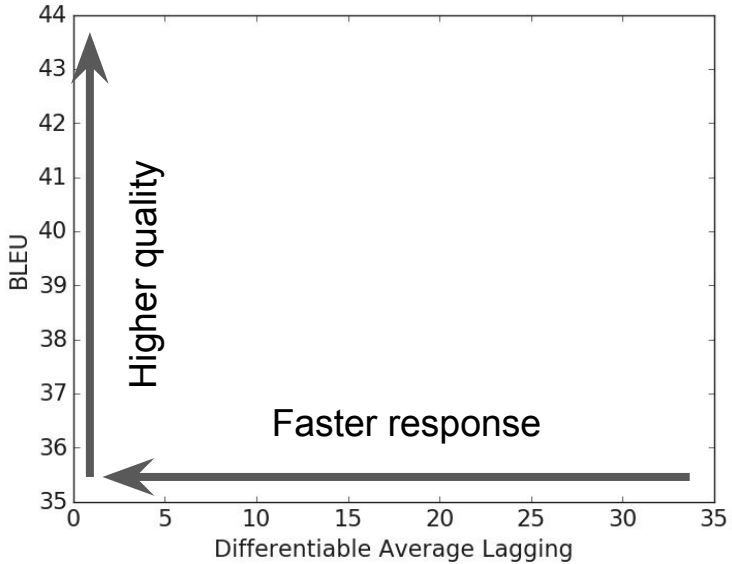
MILk: Monotonic Infinite Lookback (Arivazhagan et al., 2019: https://arxiv.org/abs/1906.05218)

- Extend Monotonic to not attend to the last word read, but with a softmax over all words read thus far
  - Much better fit for MT; dynamic program remains intact



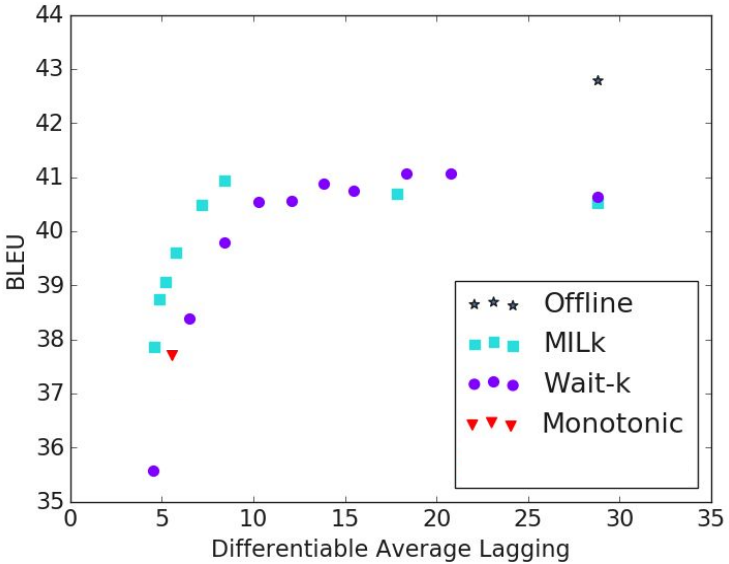
- Need a new training criterion that optimizes latency + quality
  - Therefore also need a differentiable metric for latency
  - Trade-off controlled by a hyper-parameter (analogous to k)

Quality-Latency Trade-off Graphs

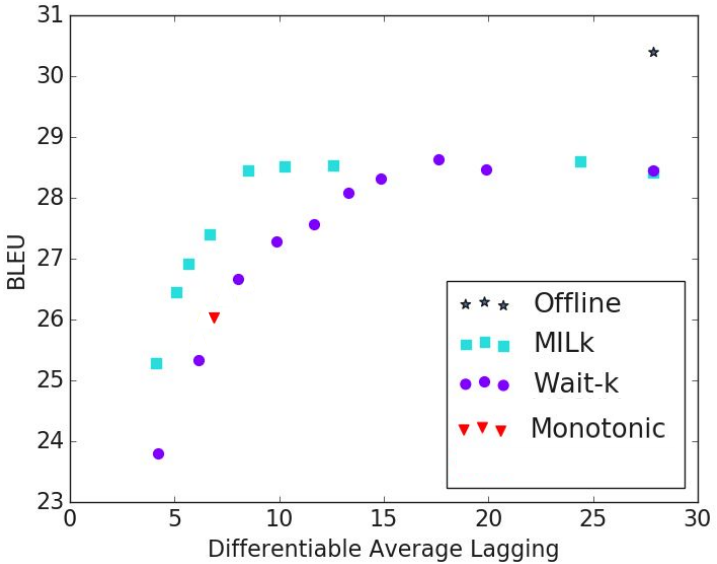


MILk vs Wait-k: The value of adaptivity

WMT 14 English to French



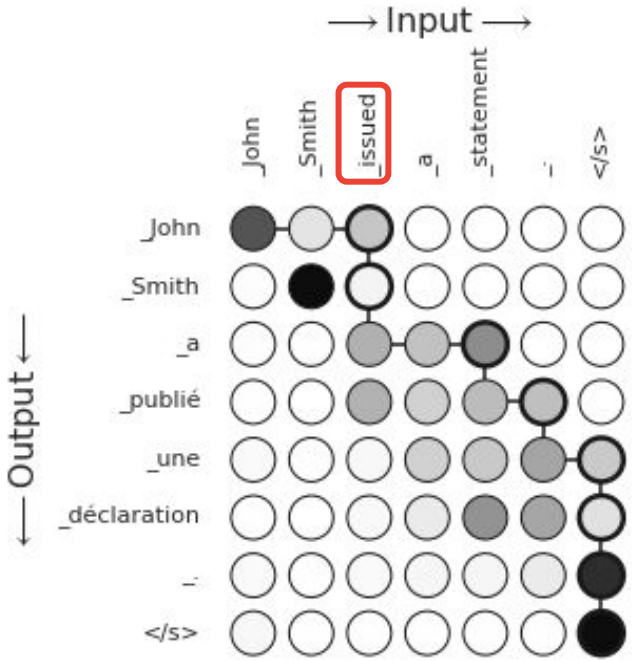
WMT 15 German to English

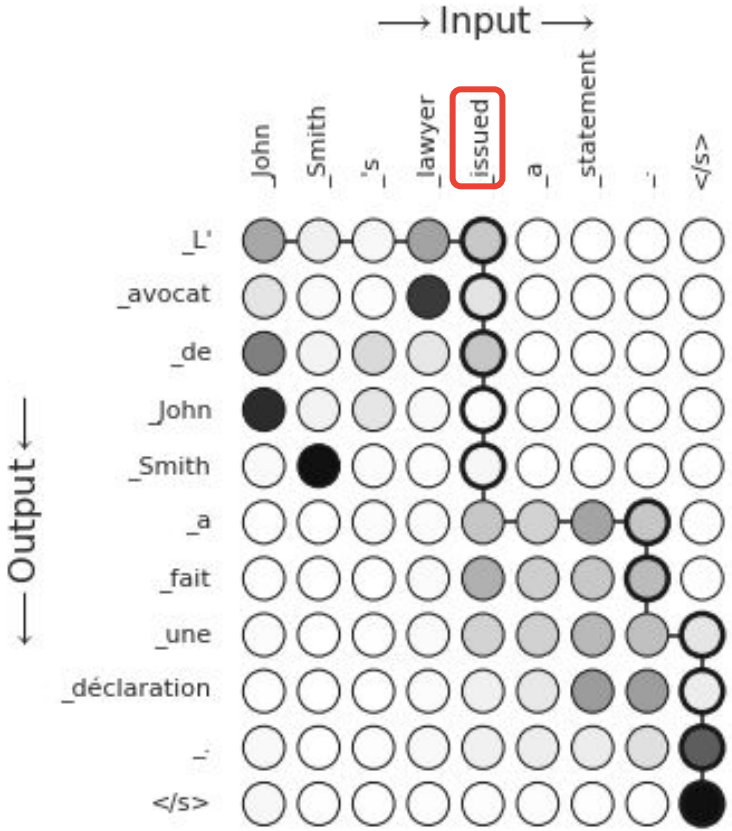


Google Research

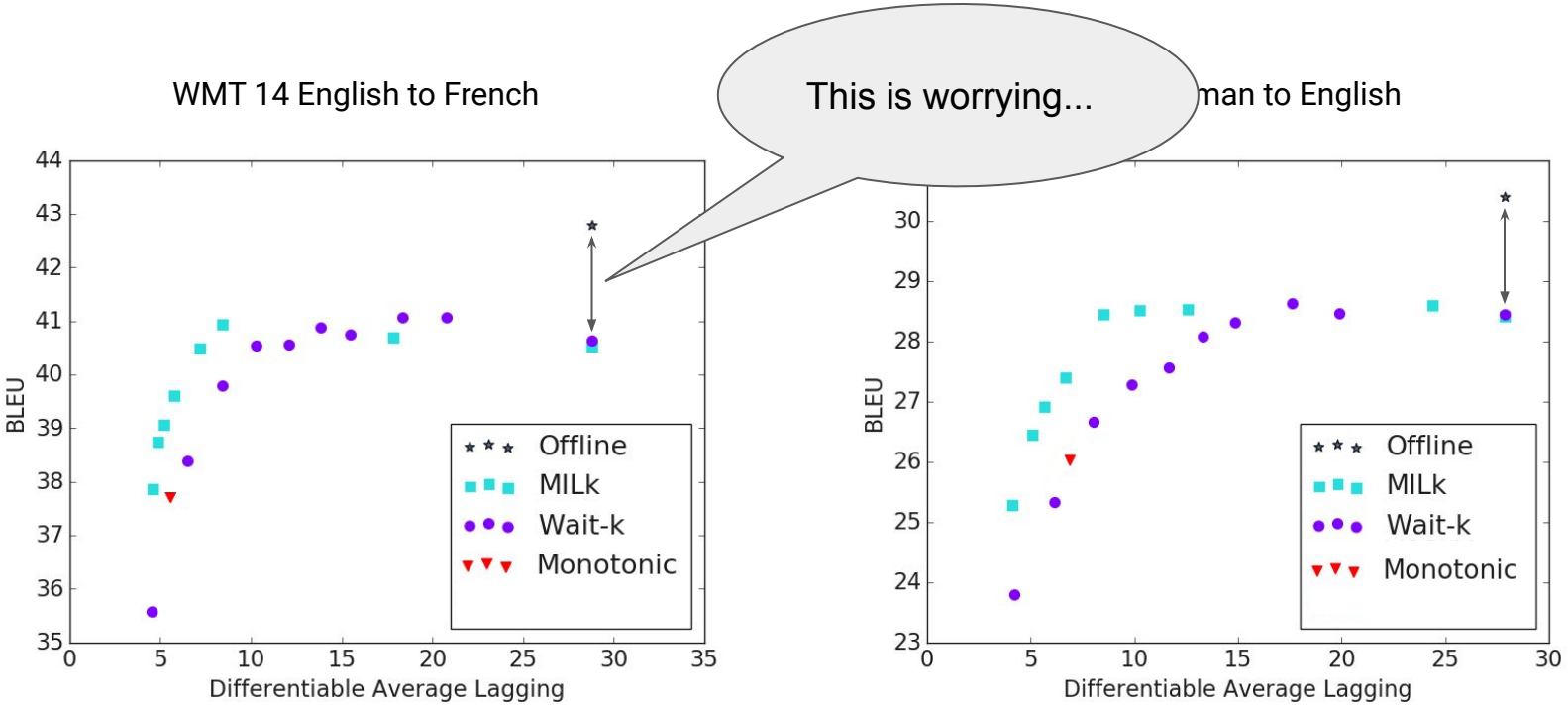
What is the MILk attention doing?

#1 Noun-phrase chunking





Simultaneous Versus Offline



Addressing the simultaneous versus offline gap

- Bidirectional encoding
  - STACL already has an (expensive) recipe for this for wait-k
  - Ma et al., 2019 (Baidu) https://arxiv.org/abs/1810.08398

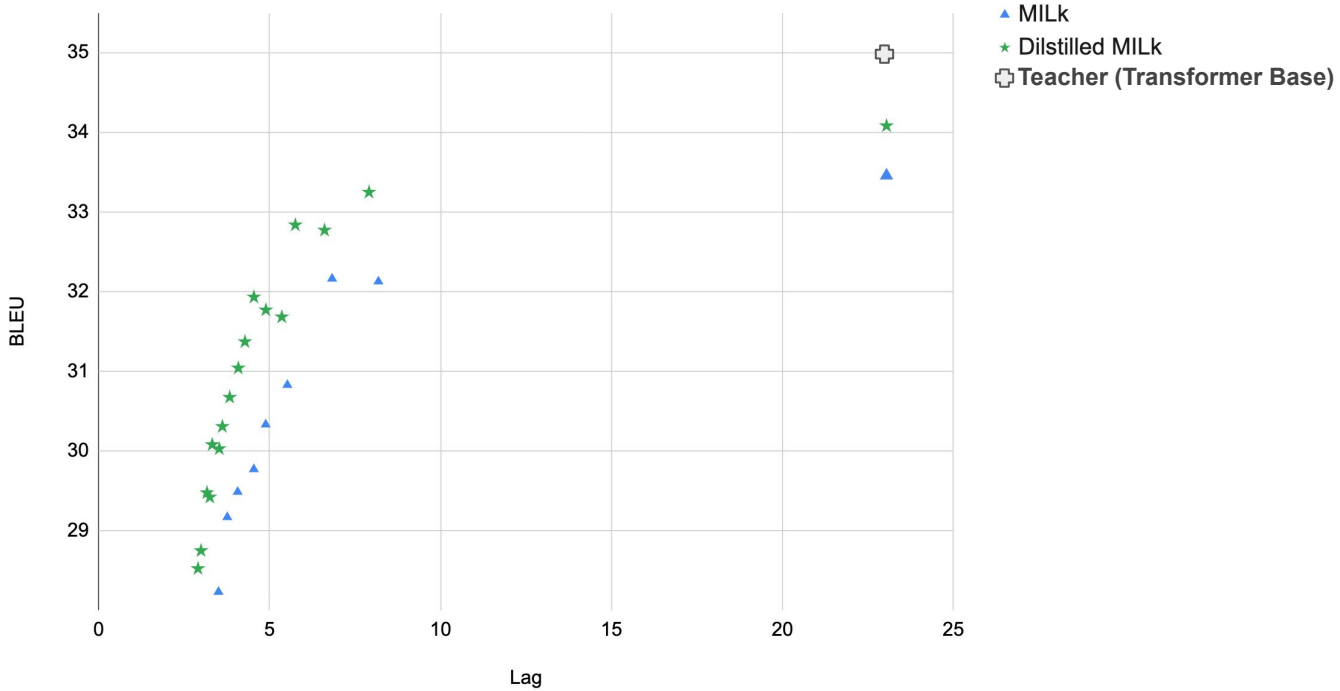
- Speculative Beam Search
  - Search a little further at each step to improve score estimates
  - Zheng et al., 2019 (Baidu) https://arxiv.org/abs/1909.05421

- Monotonic Multihead Attention
  - Extend monotonic attention (+MILk) to Transformers
  - Ma et al., 2020 (Facebook) https://arxiv.org/abs/1909.12406

Google Research

Distilling a full-sentence model into simultaneous

(DeEn IWLST 2015, from ongoing work)



Check in on the product: they're using a different abstraction

- Simultaneous translation isn't control - it's event handling
  - Each update to the source arrives as an event, triggers a new translation of the source prefix up to this point

    The ⟶ La
    The big ⟶ Le grand
    The big red ⟶ Le grand rouge
    The big red dog ⟶ Le gros chien rouge
    ...

- Retranslation!

Retranslation Pros and Cons

- Simple, portable - can immediately apply to our strongest models
- We can translate content as soon as it becomes available, and revise it later as we get more context
  - Potentially very responsive, with high final quality

Retranslation Pros and Cons

- Simple, portable - can immediately apply to our strongest models
- We can translate content as soon as it becomes available, and revise it later as we get more context
    - Potentially very responsive, with high final quality
- Problem: The output is unstable!

Back to the literature

- Niehus et al., 2018 - https://arxiv.org/abs/1808.00491
- Provides us a metric for instability
  - Average number of tokens deleted to transition from one event to another (I'll call this **Erasure**)
- Provides a way to improve stability:
  - Prefix training: Train with a 50-50 mix of full pairs and pairs where both sides are truncated to a random prefix length.
  - Reduces erasure by 50%!

Le	gros	chien	rouge	aime	Emily
The	big	red	dog	loves	Emily

Le	gros
The	big

Improving stability further (Arivazhagan et al., 2019 https://arxiv.org/abs/1912.03393, https://arxiv.org/abs/2004.03643)

- Mask-k: Simply truncate the last k tokens
  - Trades speed for stability

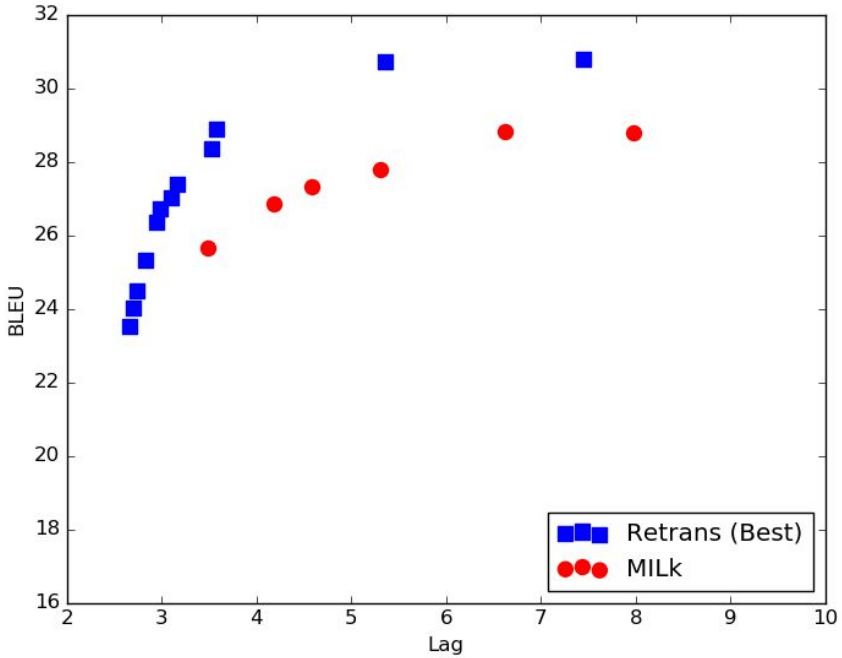
- Biased search:
  - Some revisions are necessary, others are spurious, such as alternating between “may” and “could”
  - Bias the model / search to prefer outputs it committed to earlier
  - Trades quality for stability

$$p'(y_j|y_{<j}, x_{\leq i}) = (1-\beta) \cdot p(y_j|y_{<j}, x_{\leq i}) + \beta \cdot \delta(y_j, y'_j)$$

Improved stability, visually

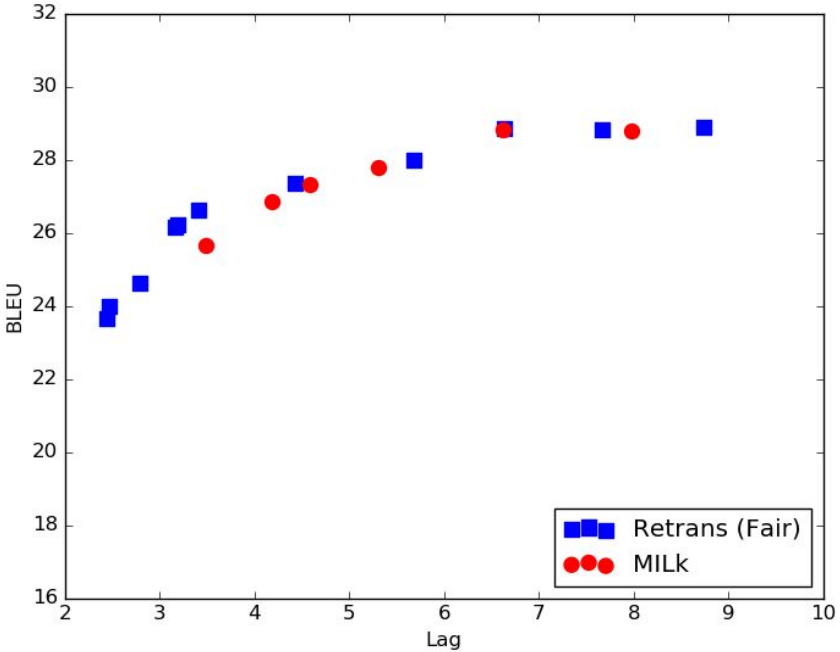
Comparing Retranslation (Best possible model)

- German-to-English WMT
- All retranslation points selected to be highly stable
  (<1 erasure per 5 final target words)
- Calculating lag in terms of time to stabilization: a token doesn't count until it and all tokens before it stop changing.



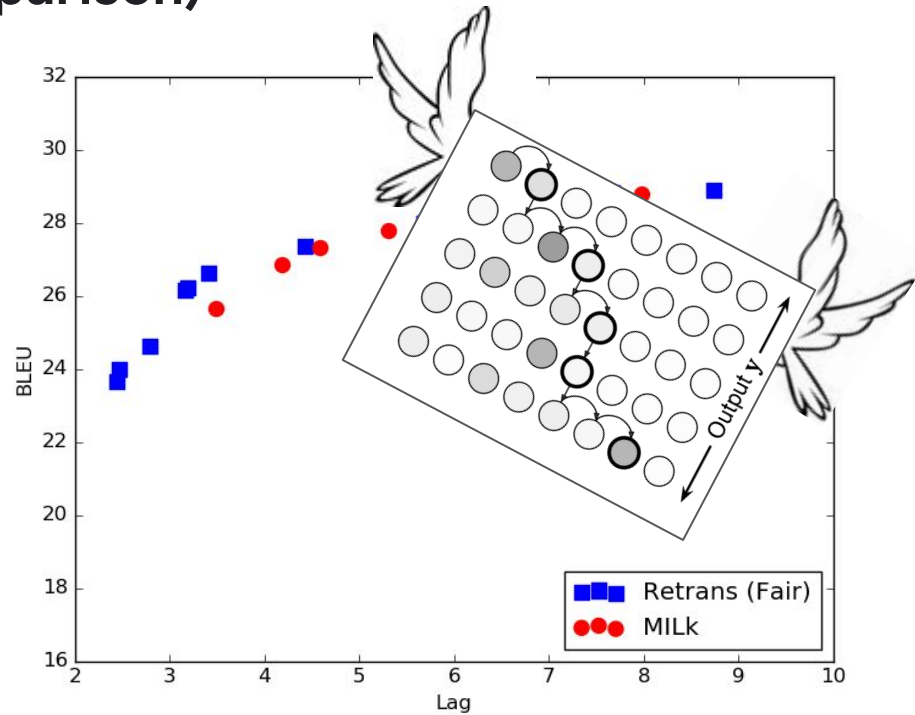
Comparing Retranslation (Fair comparison)

- German-to-English WMT
- All retranslation points selected to be highly stable
  (<1 erasure per 5 final target words)
- Calculating lag in terms of time to stabilization: a token doesn't count until it and all tokens before it stop changing.



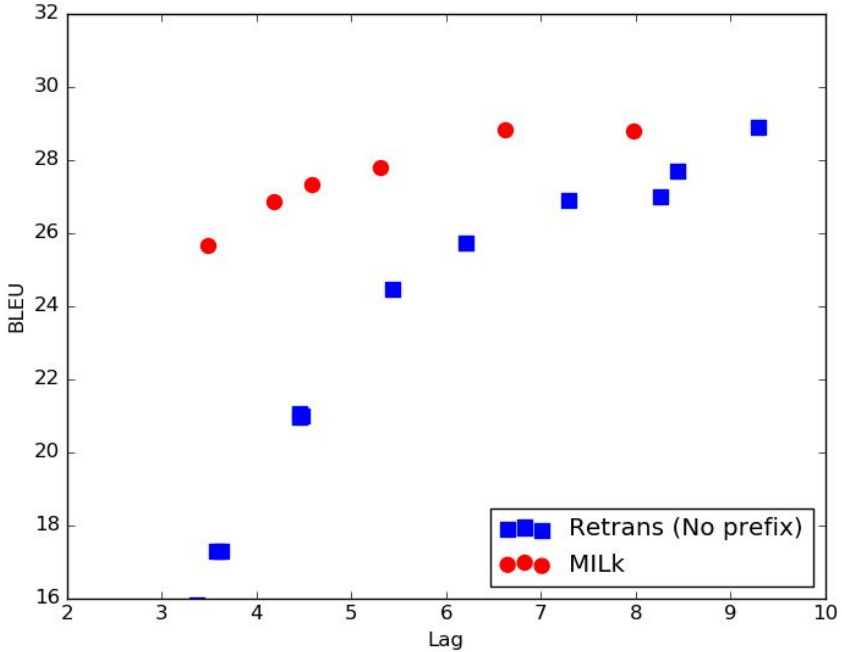
Comparing Retranslation (Fair comparison)

- German-to-English WMT
- All retranslation points selected to be highly stable
  (<1 erasure per 5 final target words)
- Calculating lag in terms of time to stabilization: a token doesn't count until it and all tokens before it stop changing.



Comparing Retranslation (Importance of Prefix Training)

- German-to-English WMT
- All retranslation points selected to be highly stable
  (<1 erasure per 5 final target words)
- Calculating lag in terms of time to stabilization: a token doesn't count until it and all tokens before it stop changing.



Moving to long-form and speech recognition (ASR) output

- Given a long-form streaming speech recognizer this is pretty straightforward.
  - Also assuming stable unspoken punctuation prediction, or endpointing.
  - Run retranslation on the latest sentence in the ASR output

Evaluating long-form latency

- Our evaluation metrics are overfit to the sentence level
  - Working on ASR sentence boundaries makes life very difficult
- BLEU trick (Matusov et al, 2005)
  - Levenshtein alignment between NMT output and reference to project reference sentence boundaries onto NMT
- Latency:
  - How many seconds does machine translation lag behind the source speaker?
  - Same trick to align NMT to reference target, and from there to reference source
  - In-sentence length ratios to relate each target word to a source word
  - Average the difference between NMT timestamps and source timestamps

Confirming our stability heuristics

- Tested on TED talks, En→XX
- “Improved” includes mask-k and biased search; configuration tuned on En→De
- Downloaded captions directly from TED for maximum language coverage
- Consistently achieve:
  - Near-perfect stability
  - Minimal BLEU degradation
  - Acceptable latency degradation

Language	Variant	BLEU	Lag (s)	Erasure
Es	Base	24.9	2.9	4.34
	Improved	25.3	4.2	0.05
Fr	Base	21.4	2.5	5.25
	Improved	21.0	3.8	0.11
It	Base	25.3	2.2	2.08
	Improved	25.2	3.9	0.01
Nl	Base	22.0	2.3	2.14
	Improved	21.9	4.0	0.01
Pt	Base	25.1	2.3	2.31
	Improved	24.4	4.0	0.01
Ru	Base	17.4	2.4	2.76
	Improved	17.0	4.1	0.01

Surprise! Not all research findings transfer!

- Notice that prefix training wasn't on the list of improvements.
- Needs more verification, but preliminary experiments with prefix-trained production-scale models indicate that the improvement from prefix training is marginal.
    - Unlike WMT data, perhaps our production data already contains prefix-like phenomena through extraction errors and other forms of data noise.

Looking back

- Monotonic Attention with Infinite Lookback (MILk)
- Improving and validating Retranslation
- An evaluation framework to compare the two

On research

- Started on this topic partly because of a product, partly because of a cool technology, and partly because of great work outside
- Checking back in with the product changed our direction, and then changed the product's direction
  - Algorithmic refinements and metrics

Looking forward

- Modeling of retranslation
- Return to As-Control, or other abstractions (As-Segmentation?)
- Document-level modeling
- More focus on interaction with speech
  - Speech-adapted NMT
  - End-to-end models
  - Simultaneous text-to-speech

You might also like

- **Simultaneous Speech Translation in Google Translate**: Jeff Pitman
  - Happened at 1:00 PM Eastern today.
  - Product-oriented, broader view of Transcribe Mode and Speech Translation.

- **Dynamic Masking for Improved Stability in Online Spoken Language Translation**: Yuekun Yao and Barry Haddow from U. Edinburgh
  - Happening at 4:00 PM Eastern today (right after this).
  - Great work on improving Retranslation.

Thanks!