From Simultaneous Translation to Simultaneous Interpretation

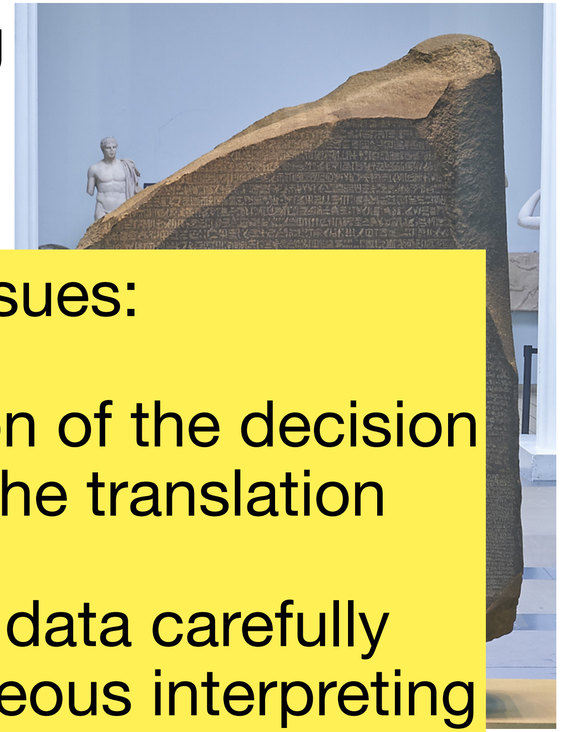
Trevor Cohn

THE UNIVERSITY OF
MELBOURNE

Simul-MT workshop @ NAACL, July 2022

Simultaneous MT vs. simultaneous interpretation

- Inspired by simultaneous interpretation, but rather different
- Main research focus on adapting **consecutive translation methods** to real-time interpreting setting
  - Developed based on **parallel translation corpora**, evaluating
  - Key algorithmic problems:
    - **Decision process** of whe
      versus generating outputs
    - **Generating outputs** give
    - A **speech to speech** prob



- This framing encounters key issues:
  - **Unsupervised**: No annotation of the decision sequence used to generate the translation
  - **Data mismatch**: Translation data carefully edited, not result of simultaneous interpreting

Talk outline

- Two parts, addressing the two issues
  - **Unsupervised**: How can we extract good “oracle” decision sequences, and use these to learn SimulMT policies? Framed as imitation learning, with twin policies.
  - **Data mismatch**: Can we obtain interpretation data? How can this be used in development and evaluation of SimulMT methods.

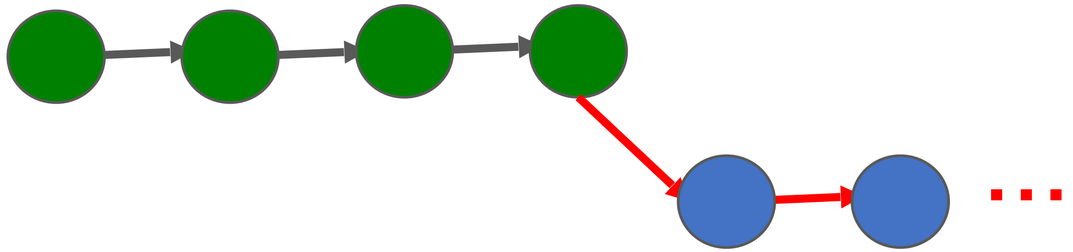
Learning Coupled Policies for Simultaneous Machine Translation using Imitation Learning

Arthur, Cohn & Haffari, **EACL 2021**

Prefix-to-Prefix and Wait-k (Ma et al 2018)

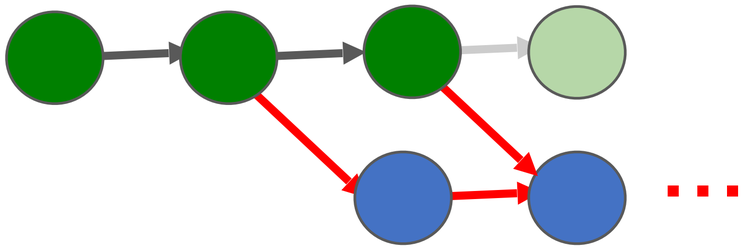
- Standard seq-to-seq is only suitable for conventional full-sentence MT

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

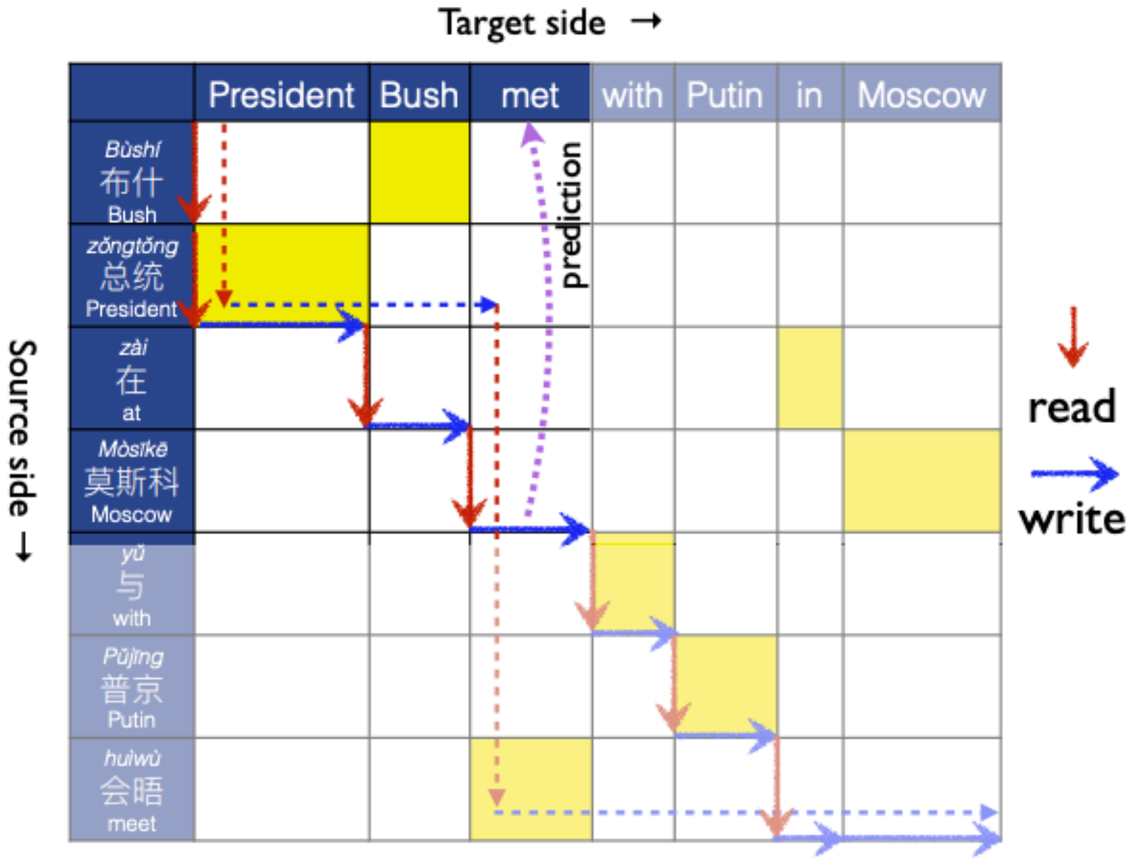


- Prefix-to-prefix framework for SI
  - Special case: wait-k policy where translation is always k words behind the source
  - Decoding this way → controllable latency
  - Training this way → implicit anticipation on target side

$$p(y_i | x_1, .., x_{i-k+1}, y_1, .., y_{i-1})$$

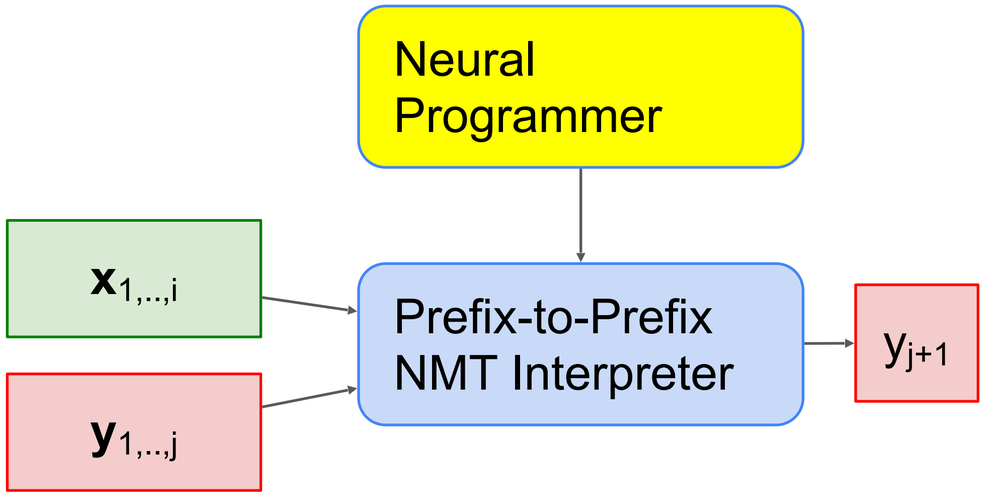


Wait-K (Ma+, 2018)



- K is predefined before training the system.
- Easy to implement and parallelize
- Weakness:
  - Hard to handle long distance dependencies
  - Setting of K critical to balance delay vs quality

Our Work: Neural Programmer-Interpreter (NPI) (Arthur et al, EACL 2021)

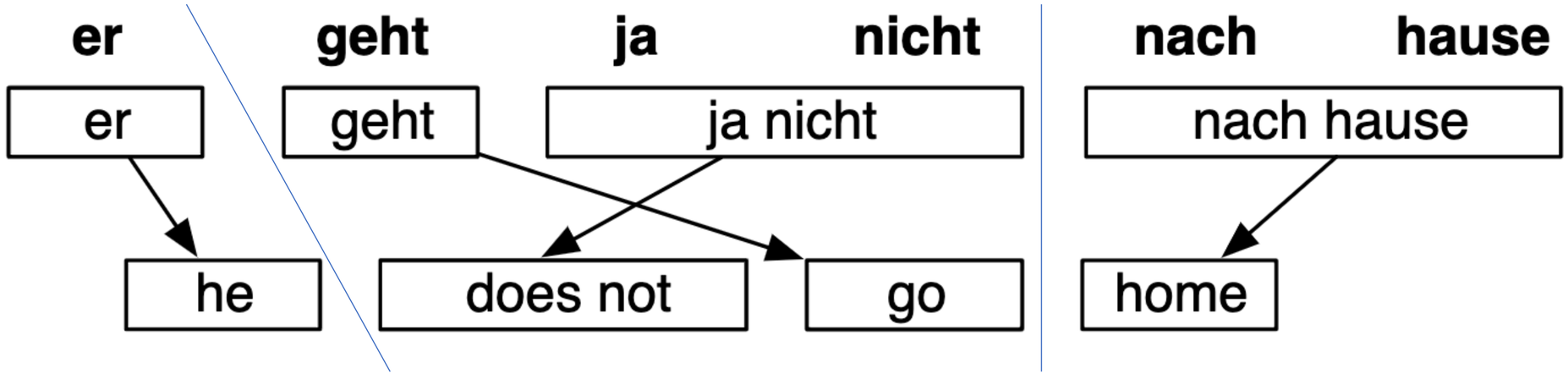


if NextAction($x_1,..,x_i,y_1,..,y_j$) is WRITE **then**
    $y_{j+1} \sim p(y|x_1,..,x_i,y_1,..,y_j)$
else
    extend the input by x_{i+1}

- Programmer controls the underlying prefix-to-prefix NMT interpreter
  - In the next time step, whether to produce a translation word or extend the input
- Our framework is based on **neural programmer-interpreter**
  - Learning the programmer and interpreter **policies** jointly
  - Coupled imitation learning with scheduled sampling

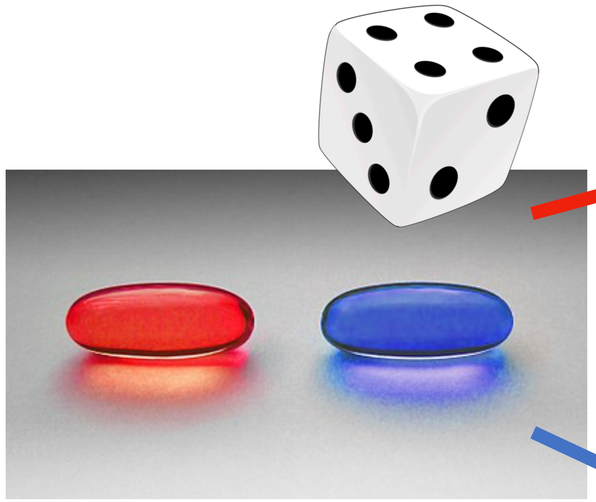
Inducing the oracle from word alignments

- For each target token y_t:
  - Repeatedly READ until aligned phrase is revealed
  - WRITE y_t
- In other words: *capture all crossing alignments in one large read*



Scheduled sampling (Bengio+, 2015)

- A form of regulariser, used with teacher forcing: exposes learner to mistakes during training



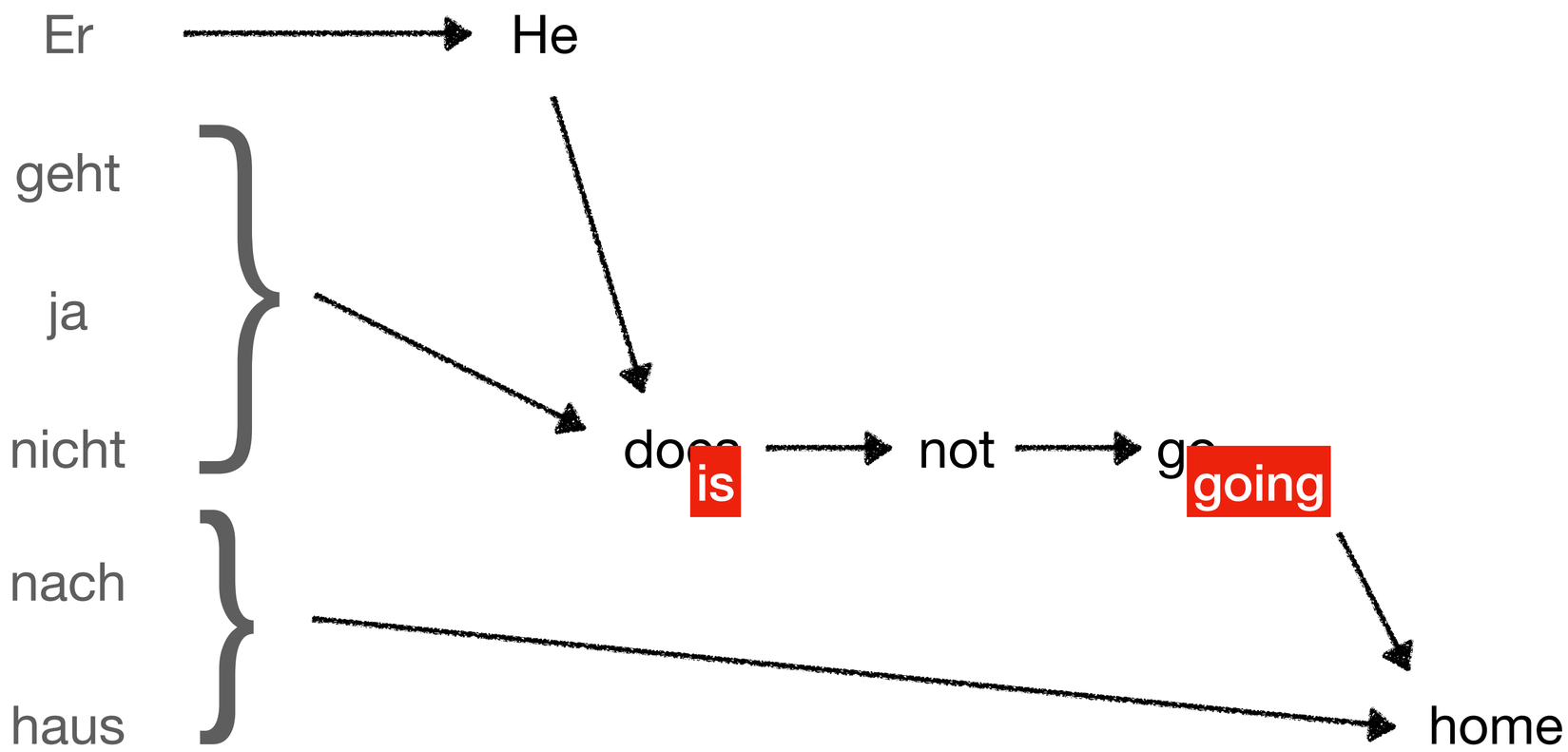
$h_t = \text{RNN}(h_{t-1}, y_t^*)$
Update with reference

Teacher forcing:
generate y_{t+1}^* given h_t

$\hat{y}_t \sim \text{softmax}(Wh_t + b)$
$h_t = \text{RNN}(h_{t-1}, \hat{y}_t)$
Update with model prediction

Here, assumes a given programme,
i.e., sequence of read and write actions

Scheduled sampling for the Interpreter



Given fixed programmer actions

Next, use these perturbed interpreter actions as context for learning programmer

Learning Coupled Policies

Algorithm 2 Training NPI-SIMT

Require: $\mathcal{D}$: Sentence pairs with oracle actions, $\beta_1, \beta_2, \beta_3$: scheduled sampling probabilities for $\boldsymbol{y}', \boldsymbol{a}', \boldsymbol{a}''$.

while a stopping condition is not met **do**
  randomly pick $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a}) \in \mathcal{D}$
  $\boldsymbol{y}' \leftarrow \text{perturbSeq}(\boldsymbol{y}, \beta_1, \theta_{\text{intp}})$
  $\boldsymbol{a}' \leftarrow \text{perturbSeq}(\boldsymbol{a}, \beta_2, \theta_{\text{prog}})$
  $\boldsymbol{a}'' \leftarrow \text{perturbProgValid}(\boldsymbol{a}, \beta_3)$
  $\hat{\boldsymbol{y}}, \boldsymbol{X}, \boldsymbol{Y} \leftarrow \text{forward\_intp}(\theta_{\text{intp}}, \boldsymbol{x}, \boldsymbol{y}', \boldsymbol{a}'')$
  $\hat{\boldsymbol{a}} \leftarrow \text{forward\_prog}(\theta_{\text{prog}}, \boldsymbol{a}', \hat{\boldsymbol{X}}, \hat{\boldsymbol{Y}})$
  $\theta_{\text{intp}} \leftarrow \theta_{\text{intp}} - \alpha_1 \nabla \delta(\hat{\boldsymbol{y}}, \boldsymbol{y})$
  $\theta_{\text{prog}} \leftarrow \theta_{\text{prog}} - \alpha_2 \nabla \delta(\hat{\boldsymbol{a}}, \boldsymbol{a})$
end while

Include mistakes predicted by the interpreter & programmer

Randomly generate a **valid** programme for (x, y)

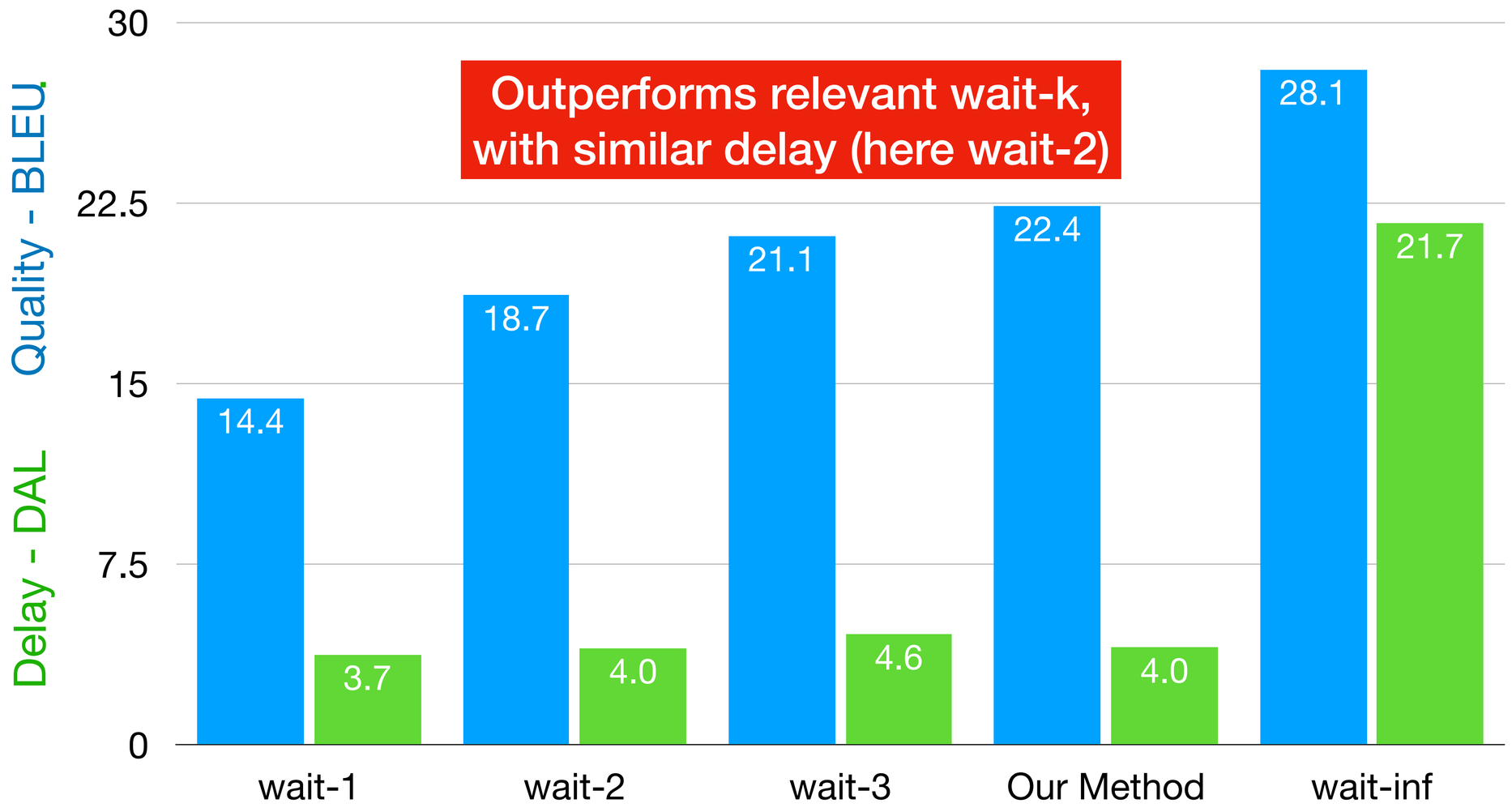
Evaluate interpreter and programmer, with coupling

Teacher forcing gradient updates

Experiments

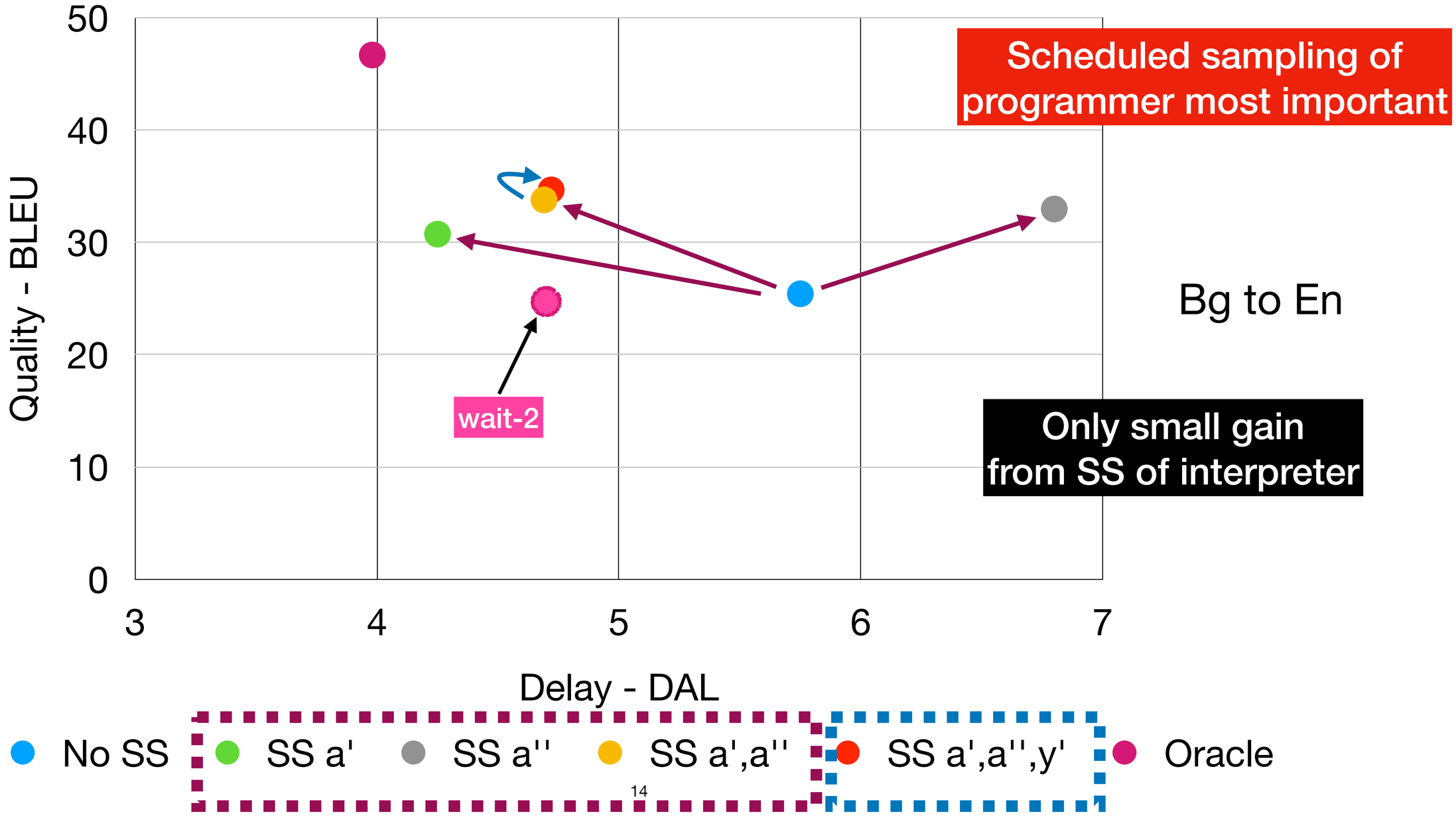
- Datasets: IWSLT+SETIMES
  - AR, DE, CZ, RO, HG, BG into EN
- Evaluation
  - Quality: **BLEU**
  - Delay: AP, AL, **DAL**
- Baseline
  - Wait-*k*
- NPI using architecture
  - LSTM

Comparison to wait-*k*

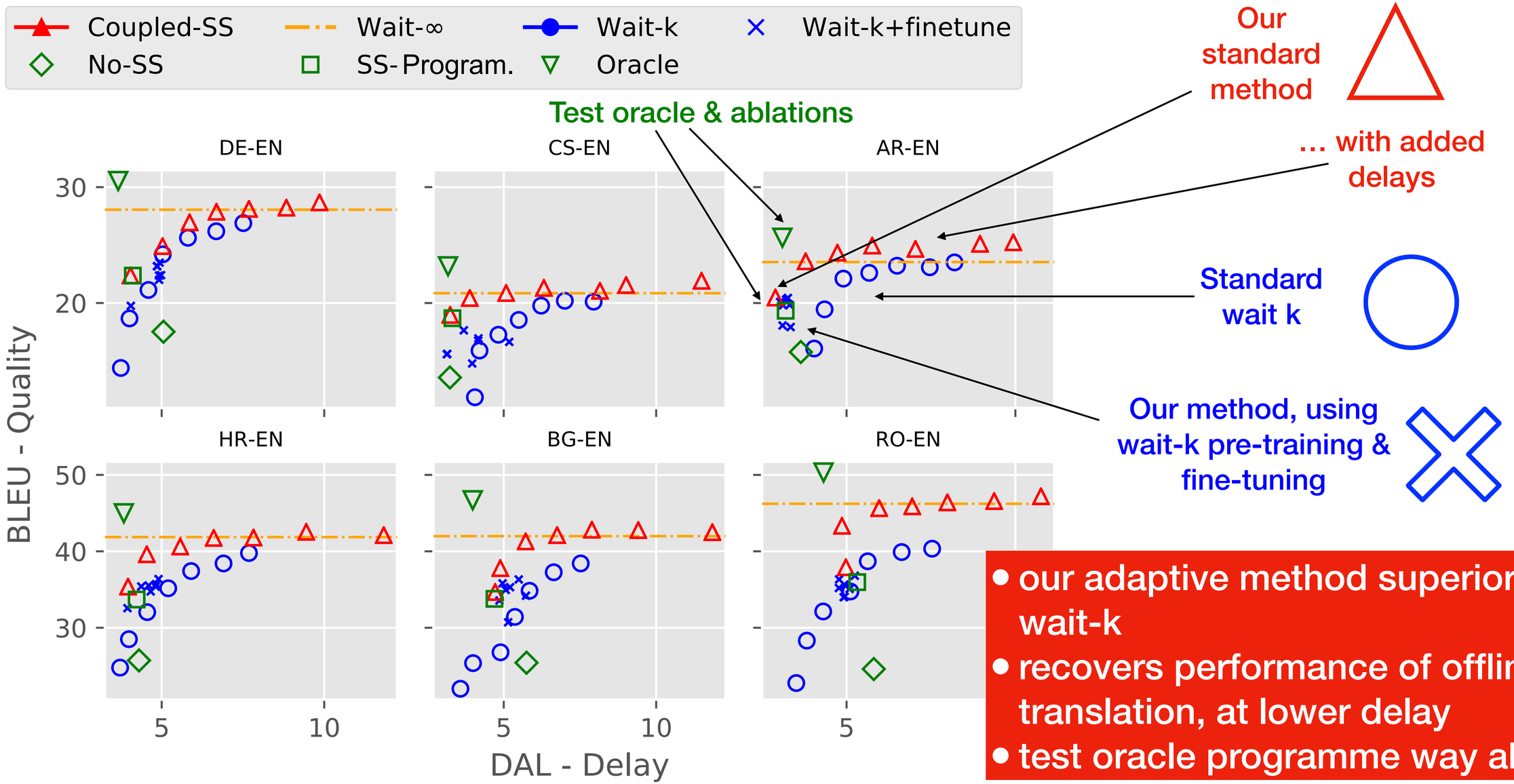


De to En

Utility of scheduled sampling



NPI vs benchmark methods



- our adaptive method superior to wait-k
- recovers performance of offline translation, at lower delay
- test oracle programme way ahead

Conclusions

- Method based on finding ***sufficient input*** to translate each target token
- Achieved by our algorithmic oracle, derived from word-alignments
- Scheduled sampling critical to learning accurate and low delay system

Open questions:

- How to improve automatic oracle, e.g., alignment errors, easily anticipated tokens
- Applying the method to speech data, benchmark against human interpretation

Talk outline

- Two parts, addressing the two issues
  - **Unsupervised**: How can we extract good “oracle” decision sequences, and use these to learn SimulMT policies? Framed as imitation learning, with twin policies.
  - **Data mismatch**: Can we obtain interpretation data? How can this be used in development and evaluation of SimulMT methods.

It is Not as Good as You Think! Evaluating Simultaneous Machine Translation on Interpretation Data

Zhou, Arthur, Haffari, Cohn & Shareghi, **EMNLP 2021**

How do humans interpret?

- Requires listening, analysis of structure, and production, *all done simultaneously*
  - ? Prioritise primary information
  - ★ Wait to start speaking
  - ★ Anticipate what speaker might say next
- Must be robust to interference between input & output languages
  - ★ syntactic differences, “reordering”
  - ★ breaking up long sentences
  - ? hesitations, corrections, etc

A learned skill, takes long and careful training, and few people can do it well

Current Status of Simultaneous Translation

- Models are trained and evaluated on offline translations.

- But is the performance observed a reality or a hallucination?
  - In real-life interpreting scenarios, interpretations are very different.
  - There is a clear mismatch between training & interpreting evaluation.

Can we learn from the data from human interpreting?

Acquiring SI corpus

- Very few existing resources, mostly very small

En↔Jp	Toyama ea., 2004; Shimuzu ea., 2014; Doi ea., 2021
En↔Es	Paulik & Waibul, 2009
Zh→En	Zhang ea., 2021
En↔It, En↔Fr, En→Pl	Bernardini ea., 2016

- Developed a pipeline for collecting SI data from Europarl archives, **De→En**
  - Confounded by issues of ASR errors (mixed audio), language mismatch, time alignment, procedural matters, interpreter failures…
  - Built dataset of <*source, interpretation, translation*> sentences

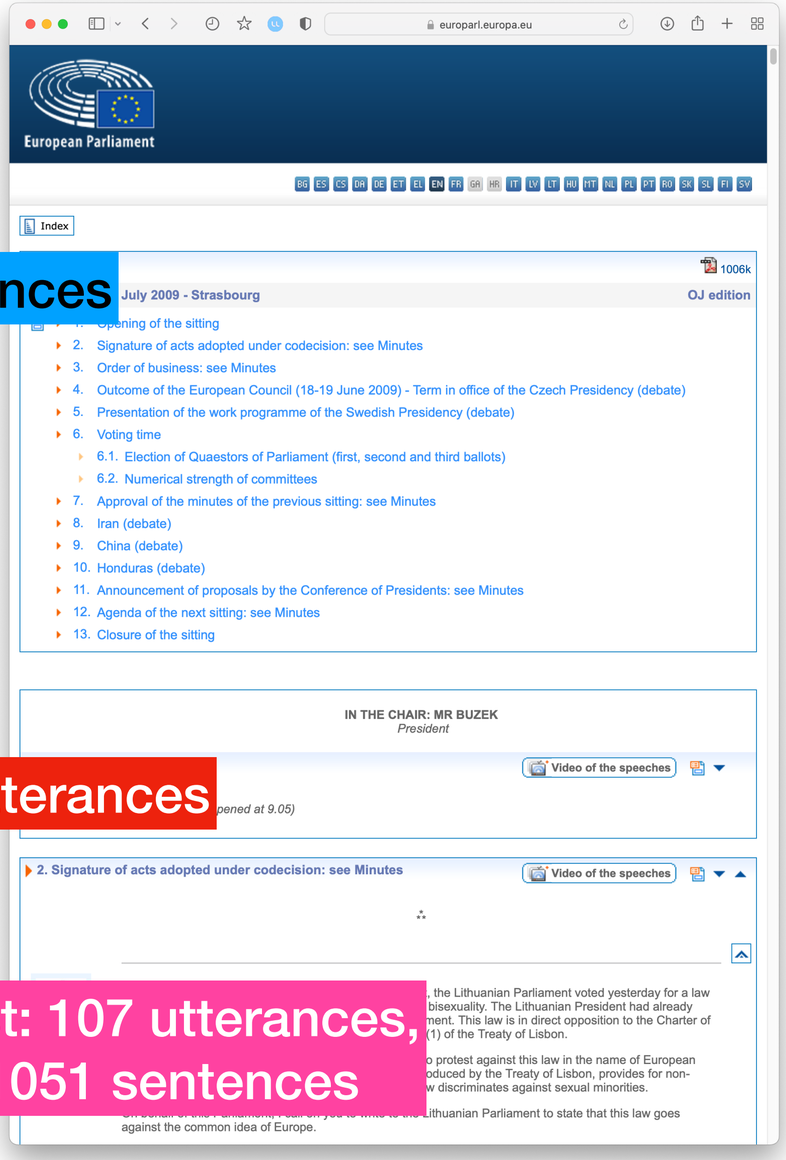
Pipeline for building SI multiparallel corpus

- Crawled Europarl archive 2008-2012
  - 323 hours of video for 238 debates, plus official transcription & translation
- Applied Google speech API for ASR
- Filtered aggressively:
  - wrong language input
  - too short, different lengths, or entirely procedural
- Sentence segmentation & alignment
- Manual correction for segmentation and ASR errors

Raw: 5239 utterances

Clean: 987 utterances

Test: 107 utterances, 1051 sentences



SI in Europarl

Offline Translation

German (original)

feststellen

Ich kann eigentlich nur zusammenfassend, dass die Europäer in den letzten Monaten auf den internationalen Bühnen in Sachen Klimaschutz *geglänzt haben* durch ihre neue Zögerlichkeit– wie weit wollen wir wirklich gehen mit den Reduktionszielen? – und, das gilt leider auch schon für die Schweden, durch neuen Geiz.

English SI

I can only summarise that Europeans .... have said ... have been rather hesitant in the international arena.
How far do we really want to go with climate change goals?
And this applies also to Sweden.

In summary, I can only say that, with regard to climate protection, Europeans have been conspicuous on the international stage in recent months as a result of their new-found hesitation – how far do we really want to go with the reduction targets? – and new tight-fistedness, and that, unfortunately, also applies to Sweden

Gamut of SI strategies in action

Generalisation/paraphrase

- De (original): *Wie ernst meinen wir es mit unserer Selbstverpflichtung zur Unterstützung der erforderlichen Öffentlichkeit?*
- En (offline)*: How serious are we about our commitment to support the necessary public?*
- En SI: *Now I think we need to backup what we've said.*

Passive/active alternation

- De (original): *Ihre Antworten haben hier nicht wirklich überzeugt.*
- En (offline): *Your answers here are not really convincing.*
- En SI: *I wasn't entirely persuaded by your answers.*

Investigate and Establish Benchmarks

- Evaluate a SOTA model's performance on translations vs. interpretations
  - Train models on offline datasets over 4 language pairs; test the models on translations and interpretations.

The gap is up to 14 BLEU.

- Bridge the gap
  - Issue: no large-scale parallel dataset
  - Solution: create pseudo-interpretations via style transfer from offline translations.

The gap can be reduced by up to 3 BLEU.

Evaluating the Performance Gap

- Evaluation on translation vs. interpretation

wait-k

	# of pairs			Evaluation					
	Europarl Offline		⋆	Translation Test			Interpretation Test		
Lang.	Train	Dev	Test	AP	AL	Bleu	AP	AL	Bleu
DE	1,666,904	3,587	1,051+	0.61	2.84	22.78	0.61	2.84	12.34
FR	1,929,486	9,736	675	0.58	2.41	21.24	0.58	2.41	9.28
PL	601,021	2,035	463	0.61	2.94	24.24	0.61	2.94	13.71
IT	1,832,809	9,256	480	0.56	2.45	24.47	0.56	2.45	10.64

There is a huge evaluation gap between translations and interpretations.

Bridging the Gap by style transfer

- Create a pseudo-interpretation corpus via style transfer

- A form of paraphrasing, starting with edited translation
  - round-trip translate, to produce "*translationese*"
  - train a HPBMT model to paraphrase into *interpretationese*

See also Chen et al, 2021 also creating synthetic pseudo-interpretations

There has therefore been enough time for the Commission to prepare its programme and for us to become familiar with it and explain it to our citizens.

style transfer

So there has been enough time for the Commission to draw up the program and for us to be aware that and explain it to our citizens.

Bridging the Gap — style transfer improves BLEU

- Create a pseudo-interpretation corpus via style transfer

- Reduced gap

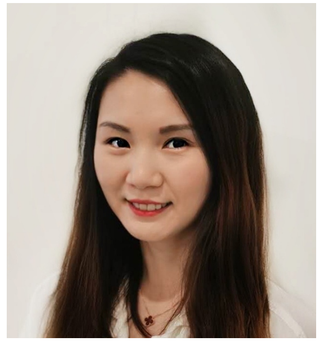
Model	AL	AP	BLEU	
			Translation	Interpretation
train on <Source, Translation>	0.61	2.84	22.78	11.47
train on <Source, Pseudo-Interpretation>	0.62	3.00	18.55	**14.26**

Conclusion

- Added to the small collection of open SI datasets, for learning and evaluation.
  - See also ***BSTC*** (Zhang et al, 2021) and ***Voxpopuli*** (Wang et al, 2021)
- Quantifyied the evaluation gap between translations and interpretations.
- Proposed a style transfer technique to construct a pseudo-interpretation corpus.
- Gap still noticeable, calling for constructing large-scale interpretation corpora, and *cleaner* evaluation corpora.

Collaborators

- Joint work with collaborators:
  - Philip Arthur
  - Jinming Zhao
  - Reza Haffari
  - Ehsan Shareghi
- Funded by ARC & Amazon



Australian Government
Australian Research Council

amazon research awards