UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Streaming neural machine translation systems from European languages into English

Areg Mikael Sarvazyan

Jorge Civera Saiz

Javier Iranzo Sánchez

MLLP Machine Learning and Language Processing

VRAIN Valencian Research Institute for Artificial Intelligence

14 July, 2022

Contents

1 Introduction

Motivation

- Increase accessibility of multimedia content to non-speakers of the original language
- Remove language barriers in virtual meetings and video-conferences in real time

Framework

- Research internship at the VRAIN MLLP research group
- Technology-transfer contract between MLLP and CERN
  - Fr$\rightarrow$En machine translation (MT) systems for offline and real-time scenarios

Introduction

Goals

- To study the current state-of-the art approaches for offline and real-time MT, including domain adaptation techniques and automatic evaluation.
- To apply the most important tools employed in MT research for data processing, model training, inference and evaluation.
- To explore and refine data filtering and processing techniques to improve the quality of MT systems.
- To develop general domain and in-domain MT systems that provide accurate enough translations for the purposes of CERN.
- To develop simultaneous MT systems for the streaming MT scenario with low latency and good enough quality.

2 Preliminaries

Machine Translation

- In MT, we search for the best translation $\hat{\mathbf{y}}$ of $\mathbf{x}$ given by

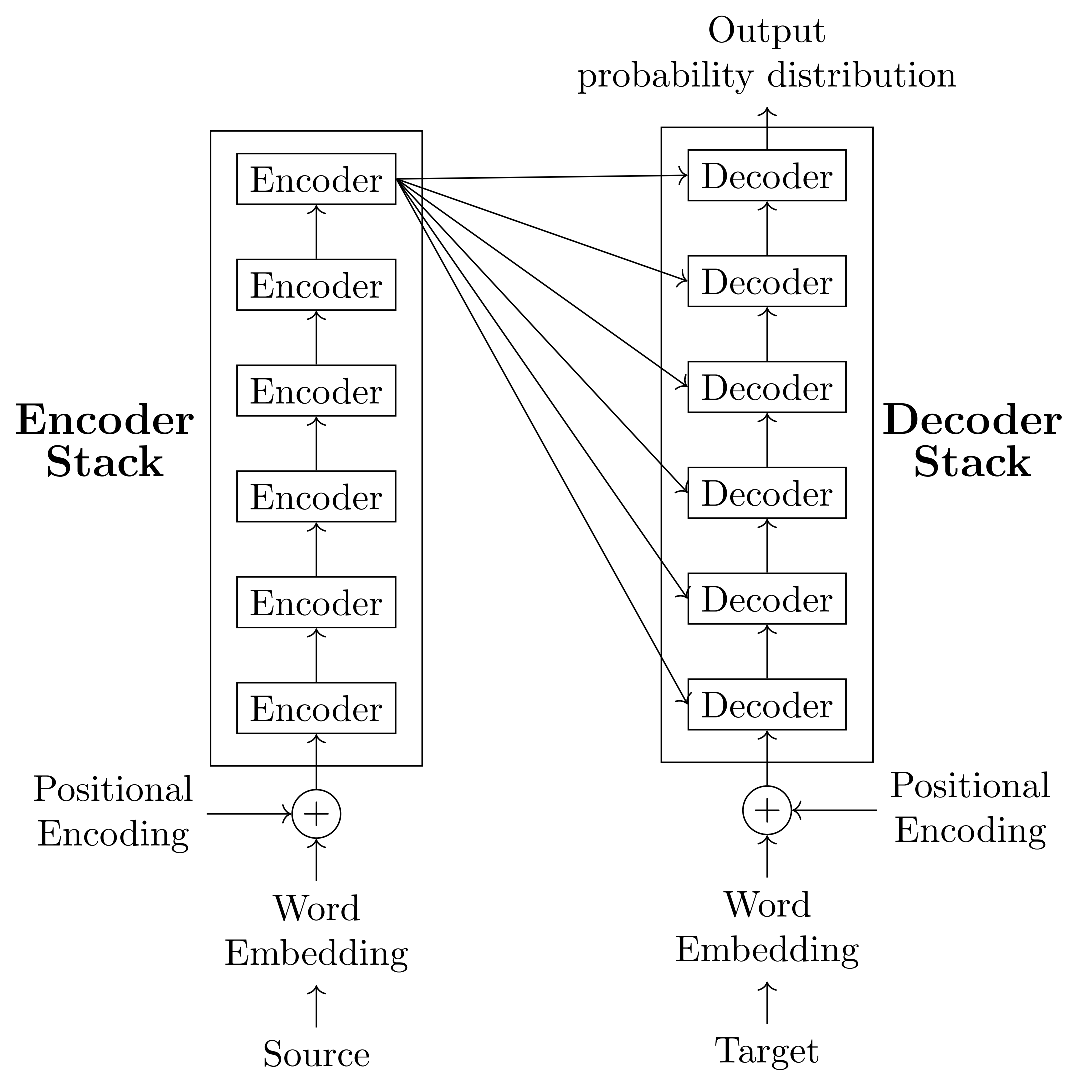
$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}^*} p(\mathbf{y}|\mathbf{x})$$

- $\mathbf{x}$ is the *source* sentence and $\mathbf{y}$ is the *target* sentence
- The neural models we will use approximate $p(\mathbf{y}|\mathbf{x})$ directly
- The *beam search* algorithm is used to instantiate the argmax

Preliminaries

Transformer

- Deep learning architecture based on the *attention* mechanism
- State-of-the-art in many tasks and modalities
- Considers the full source phrase when translating



Preliminaries

Evaluation

- Manual evaluation is very costly
- Automatic evaluation
  - ***BLEU: Bilingual Evaluation Understudy*** $\rightarrow$ higher is better
  - Others: chrF and TER

Tools

- Model training and inference: fairseq
- Data processing: Moses, subword-nmt, SentencePiece, etc.

3 Training Data

General Domain

Source	Corpus	Bilingual pairs	Words	
			English	French
Internet	WikiMatrix	2.7 M	57.8 M	63.1 M
	WikiMedia	1.0 M	24.1 M	25.8 M
	Giga Fr-En	22.5 M	575.8 M	672.2 M
	ParaCrawl	216.6 M	3.7 G	4.1 G
	CCAligned	15.6 M	156.7 M	171.1 M
	CommonCrawl	0.1 M	4.1 M	4.7 M
	EUBookshop	10.8 M	224.6 M	244.5 M
	UNPC	30.3 M	658.4 M	816.4 M
	News Commentary	3.2 M	70.7 M	76.6 M
Parliamentary Meetings	DGT-TM	4.9 M	86.3 M	95.4 M
	Europarl	1.2 M	28.6 M	29.9 M
	Europarl-ST	96.5 K	2.3 M	2.6 M
	Total	**309.0 M**	5.6 G	6.3 G

Training Data

Domain of CERN

Monolingual CDS Corpus

	Objects	Sentences	Words
Titles	519 K	519.0 K	4.6 M
Abstracts	130 K	652.0 K	15.6 M
Documents	296 K	48.9 M	1.1 G
Total	945 K	**50.0 M**	1.1 G

Bilingual CERN News Corpus

Dataset	Documents	Bilingual Pairs	Words	
			English	French
Training	3409	51.9 K	841.9 K	909.4 K
Validation	144	2.2 K	331.7 K	405.3 K
Test	128	1.8 K	274.0 K	333.3 K
Total	3681	**55.9 K**	1.5 M	1.7 M

Training Data

Data Processing

Raw text	`Thank you, Mr Segni, I shall do so gladly.`
Tokenization	`Thank you , Mr Segni , I shall do so gladly .`
Truecasing	`thank you , Mr Segni , I shall do so gladly .`
Byte-Pair Encoding	`thank you , Mr Se@@ gn@@ i , I shall do so gl@@ ad@@ ly .`
Raw text	`Thank you, Mr Segni, I shall do so gladly.`
Truecasing	`thank you, Mr Segni, I shall do so gladly.`
SentencePiece	`thank_ you ,_ Mr_ Seg ni ,_ I_ shall_ do_ so_ glad ly ._`

- Filtering → removes noise from data
  - Language identification
  - Sentence length
  - Source-to-Target length ratio
- Tokenization → divides text into *tokens*
- Truecasing → maintains the most frequent version of each token
- Subword Segmentation → represent the vocabulary with fewer tokens
  - Byte-Pair Encoding
  - SentencePiece

4 Domain Adaptation

Fine-Tuning

- Include *domain bias* in the modeling of $p(\mathbf{y} \mid \mathbf{x})$
- A model trained for a general task is used in a fine-grained task or domain
- We modify the model parameters to *adapt* it to the domain using in-domain data
- CERN's domain: particle physics

Backtranslations

- Translate monolingual text in the target language to the source language
  - Construct a synthetic bilingual corpus
- Leverage monolingual data in the target language and domain of interest
- Enhance the translation model's implicit language model (of the target language)

5 Streaming MT

- To simultaneously transalte $\mathbf{x}$ into the target $\hat{\mathbf{y}}$, we find the best translation by

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}^*} p_g(\mathbf{y} \mid \mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}^*} \prod_i p(y_i \mid \mathbf{x}_{\leq g(i)}, \mathbf{y}_{<i}).$$

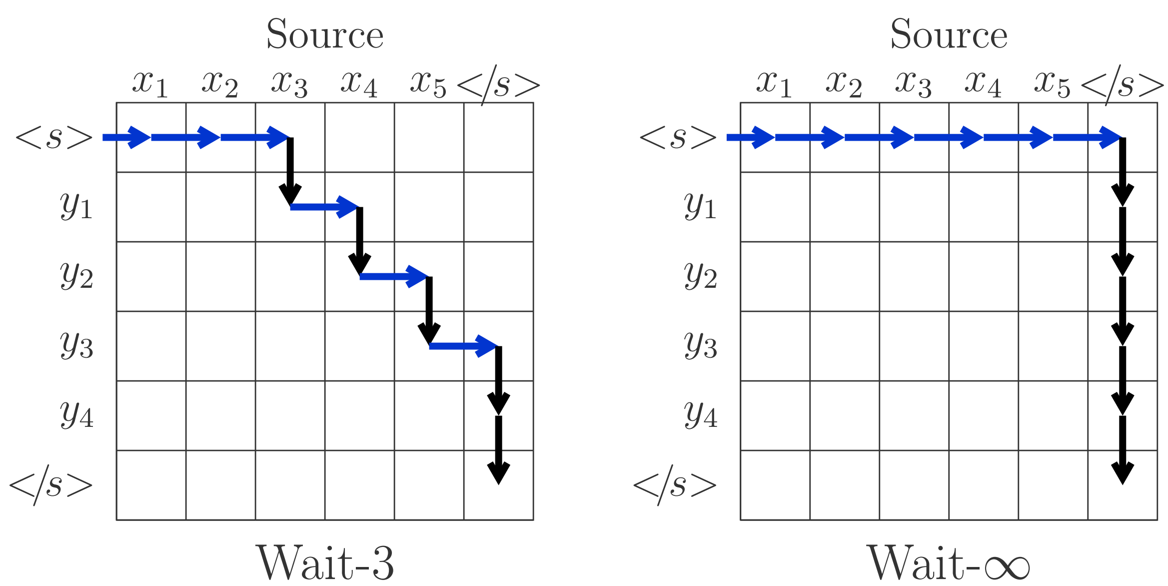
- The model only has access to a prefix of the full source to translate
- It needs a *policy* to decide when to perform a reading or writing action

Offline	Hay libros que valen la pena volver a leer.								
	wait whole sentence								There are books that are worth reading again.
Simultaneous	Hay libros	que	valen	la	pena	volver	a	leer.	
	wait 2 words There	are	books	that	are	worth	reading	again.	

Streaming MT

Wait-k Models

- Inspired by how human interpreters wait for enough context before translating
- The model reads k tokens before emitting translations
  - Afterwards, it alternates between writing and reading operations

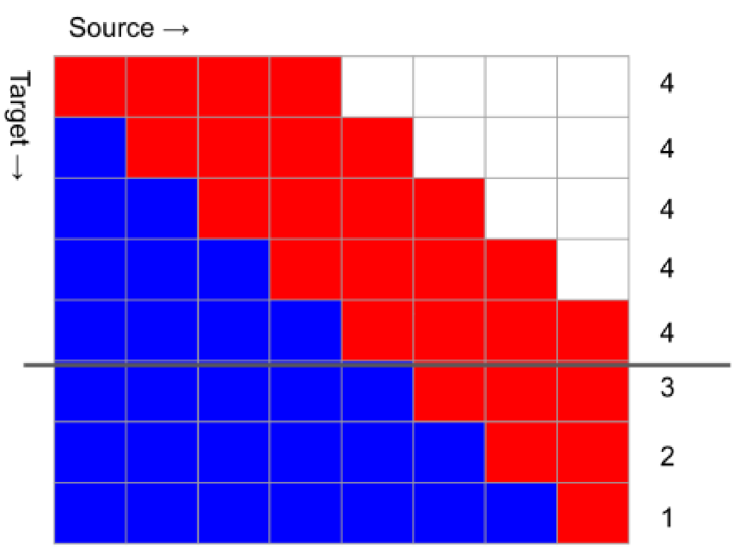


- ***Test-Time Wait-k*** introduces the wait-k policy to offline MT models
- ***Multi-Path Wait-k***: simultaneous MT training scheme that considers different values for k

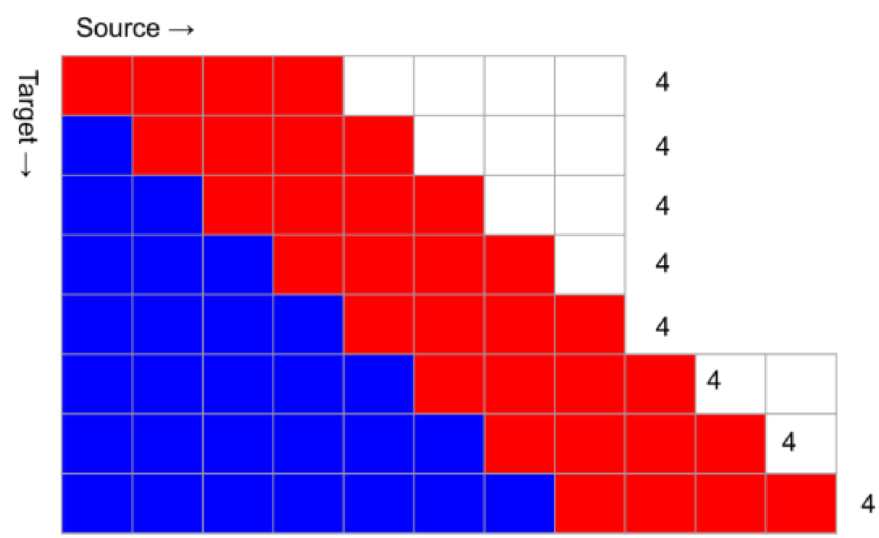
Streaming MT

Latency Evaluation

- Automatic evaluation of latency, independent of the hardware and environment
  - ***AL***: Average Lagging
  - ***DAL***: Differentiable Average Lagging



AL



DAL

6 Results

- *Baseline*: X5Gon system trained in 2019

Offline MT Systems

General Domain

Name	Filtering			Processing				
	Langid	Length	Ratio	Apostrophes	Tokenize	Truecase	BPE	SPM
V1	-	-	-	-	X	X	X	-
V2	X	<150	1.5	-	X	X	X	-
V3	X	<150	1.5	X	-	X	-	X

In-Domain

- V3-FT-CN: V3 fine-tuned on the CERN News corpus
- V3-FT-BT: V3 fine-tuned on 50K backtranslations of the CDS corpus
- V4: V3 + backtranslations

Results

General Domain Evaluation

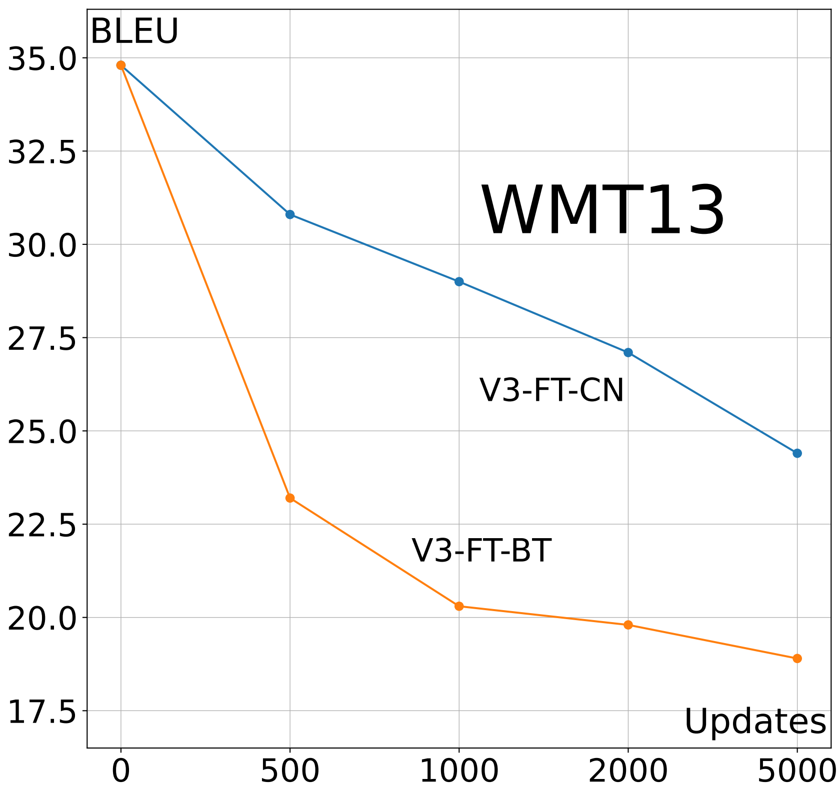
System	BLEU	
	WMT13	WMT14
X5Gon	34.7	39.4
V1	**35.1**	39.2
V2	**35.1**	**39.5**
V3	34.8	39.3
V4	34.6	39.1

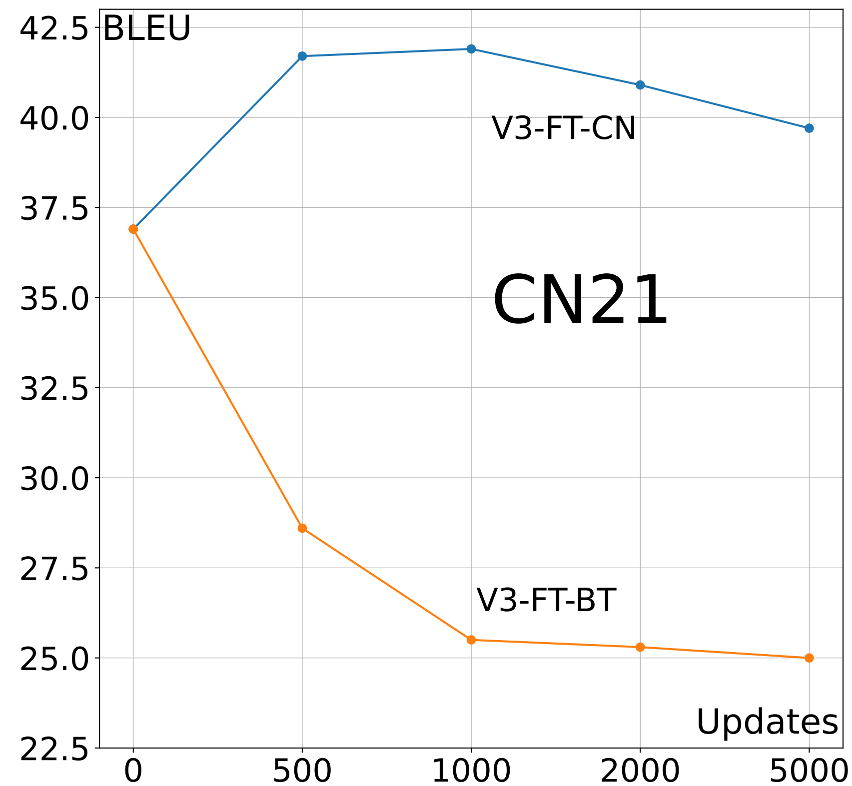
In-Domain Evaluation

System	BLEU	
	CN21	CN22
X5Gon	35.3	36.8
V3	**36.9**	**38.6**
V4	36.6	38.2

Results

Fine-Tuning



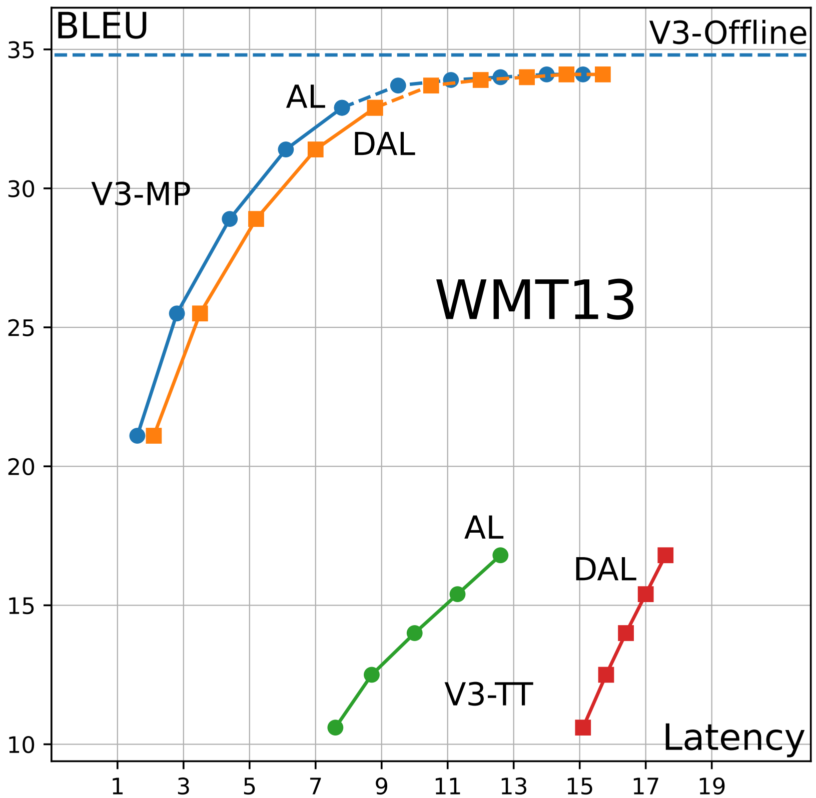


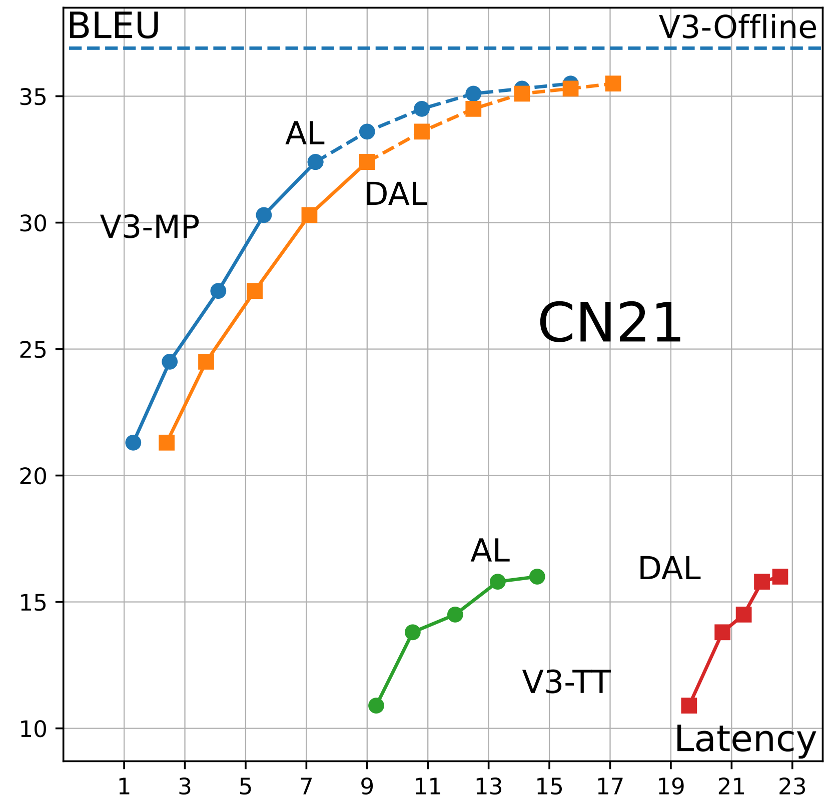
- Best model: V3 fine-tuned using CERN News with 1000 updates

Corpus	BLEU (WMT14)	BLEU (CN22)
X5Gon	39.4	36.8
V3-FT-CN	31.7	42.9

Results

Streaming MT





- Selected models: V3 Multi-Path Wait-k with $k \geq 7$

Corpus	BLEU	AL	DAL
WMT14	33.8	5.9	7.1
CN22	30.9	5.5	7.1

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

7 Conclusions

Achieved Goals

- Studied SOTA for offline and simultaneous MT, domain adaptation and automatic evaluation
- Leveraged current tools used for MT research to develop various MT systems
- Refined the data processing pipeline and improved MT system performance
- Developed in-domain MT systems, improving the baseline by a relative 11%
- Developed high-accuracy and low-latency simultaneous MT systems for streaming applications

Future work

- Deploy the MT systems in CERN's network
- Carry out domain adaptation for simultaneous MT systems
- Explore:
  - Domain adaptation beyond fine-tuning with *adapter layers*
  - Adaptive policies for simultaneous translation by identifying *meaningful units*

Streaming neural machine translation systems from European languages into English

Areg Mikael Sarvazyan

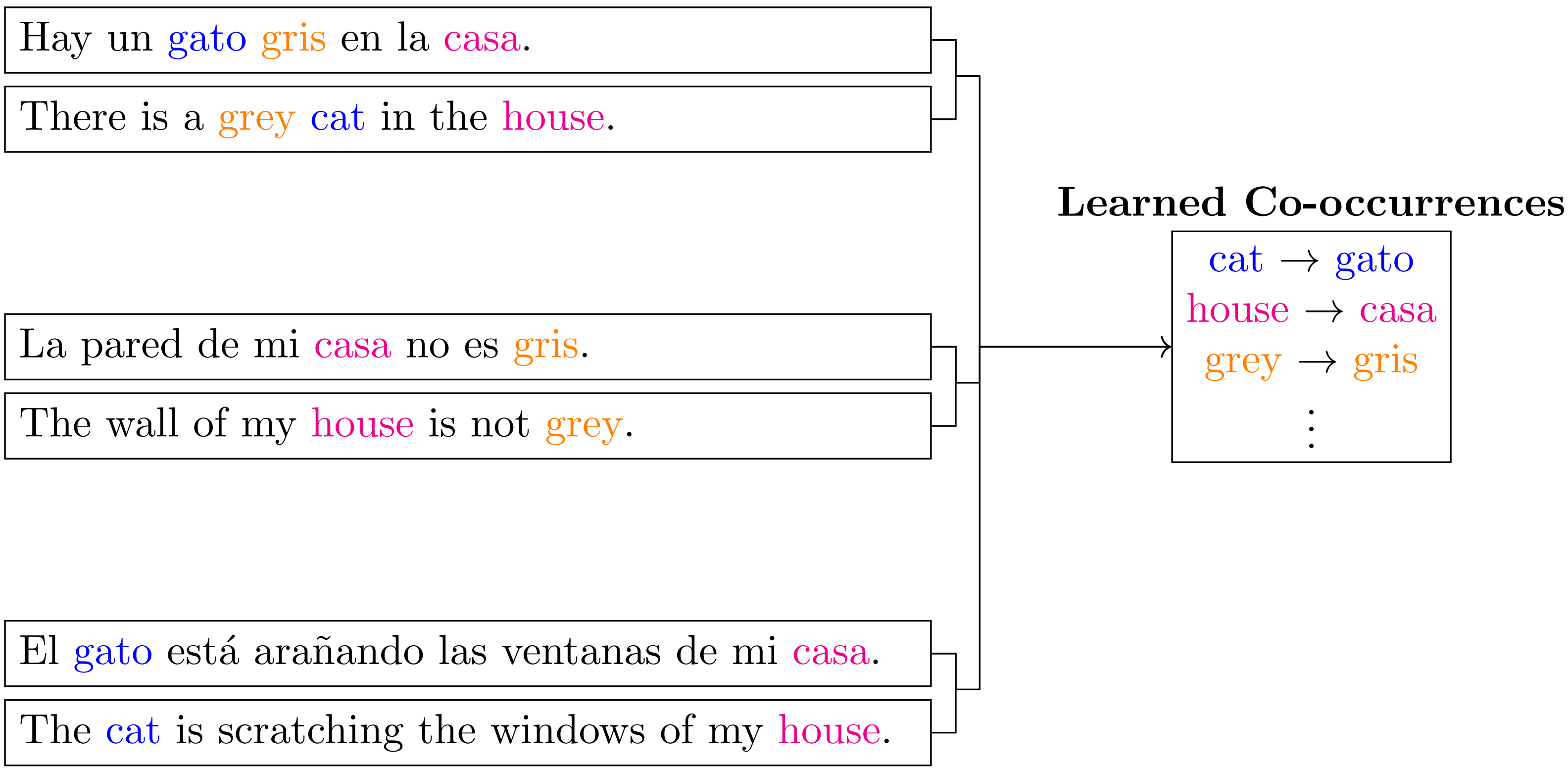
Jorge Civera Saiz

Javier Iranzo Sánchez

14 July, 2022

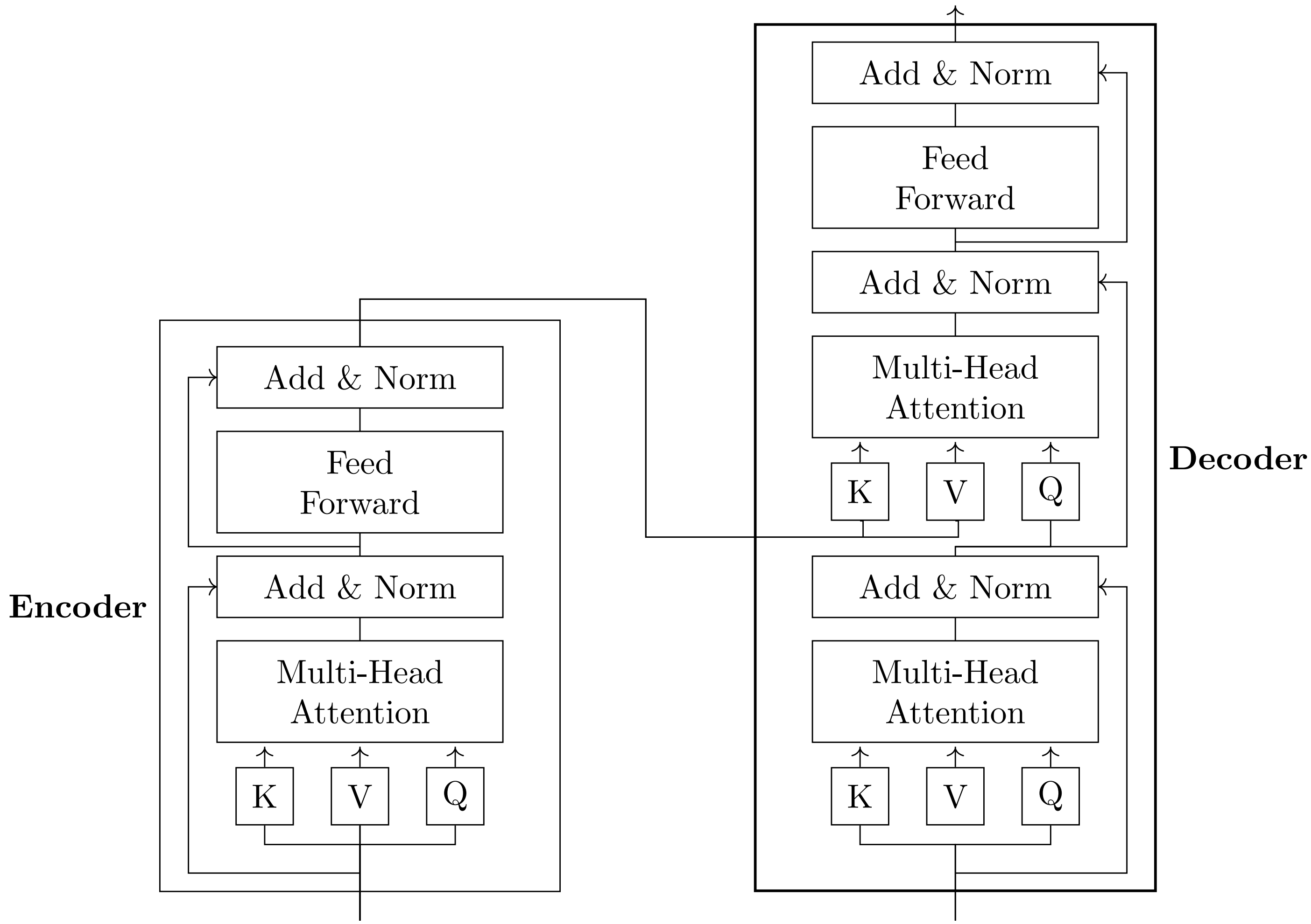
Appendix

IBM Model 1



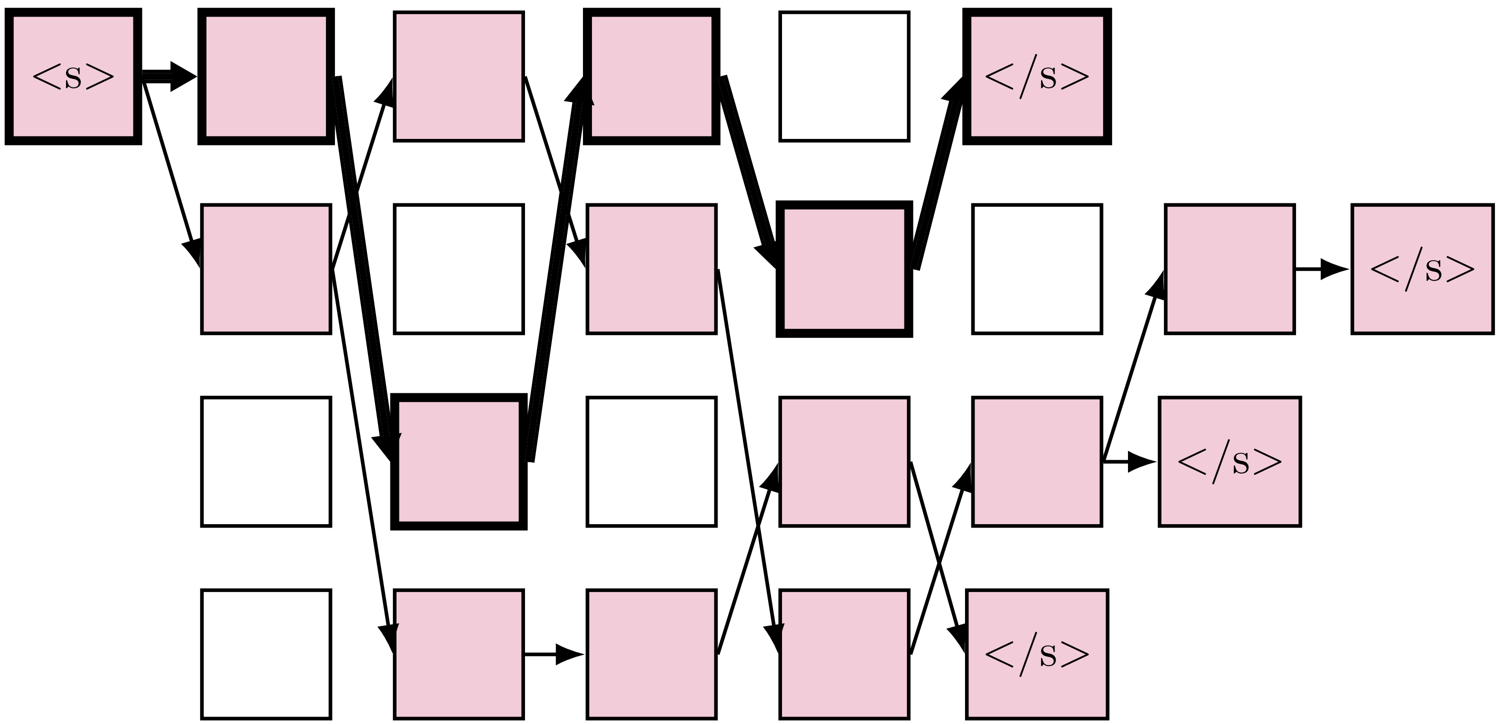
Appendix

Transformer Encoder-Decoder



Appendix

Beam Search



Appendix

Quality Evaluation

BLEU

$$BLEU(4) = BrevityPenalty \times AveragePrecision(4)$$

$$BrevityPenalty = \begin{cases} 1 & |output| > |reference| \\ \exp\left(1 - \frac{|output|}{|reference|}\right) & |output| \leq |reference| \end{cases}$$

$$AveragePrecision(N) = \frac{1}{N}\sum_{n=1}^{N} log p_n, \qquad p_n = \frac{\text{matching n-grams}}{\text{total n-grams in output}}$$

chrF

$$chrF\beta = (1+\beta)^2 \frac{chrP \times chrR}{\beta^2 \times chrP + chrR}$$

TER

$$TER = \frac{\text{word-level edit distance}}{|\text{reference}|}$$

Appendix

Simultaneous Translation

- $g(i)$ is the number of source tokens read when writing a translation at position i.

Wait-k

$$g_{\text{wait}-k}(i) = \left\lfloor k + \frac{i-1}{\gamma} \right\rfloor, \qquad \gamma = \mathbb{E}\left[\gamma_n\right], \qquad \gamma_n = \frac{|\mathbf{y}_n|}{|\mathbf{x}_n|}.$$

Multi-Path Wait-k

For one wait-k path $\mathbf{z}^k_{<i}$:

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}^k) = \prod_i p(y_i \mid \mathbf{x}_{\leq \mathbf{z}^k_i}, \mathbf{y}_{<i}, \mathbf{z}^k_{<i}).$$

We optimize over multiple wait-k paths:

$$\mathbb{E}_K[p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}^k)] \approx \prod_{k \sim \mathcal{U}(K)} p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}^k).$$

Appendix

Average Proportion

$$AP = \frac{1}{|\mathbf{x}||\mathbf{y}|} \sum_{i=1}^{\mathbf{y}} g(i)$$

Average Lagging

$$AL_g = \frac{1}{\tau} \sum_{i=1}^{\tau} \left(g(i) - \frac{i-1}{\gamma} \right), \qquad \tau = \tau_g(|\mathbf{x}|) = \min_{i:g(i)=|\mathbf{x}|} i$$

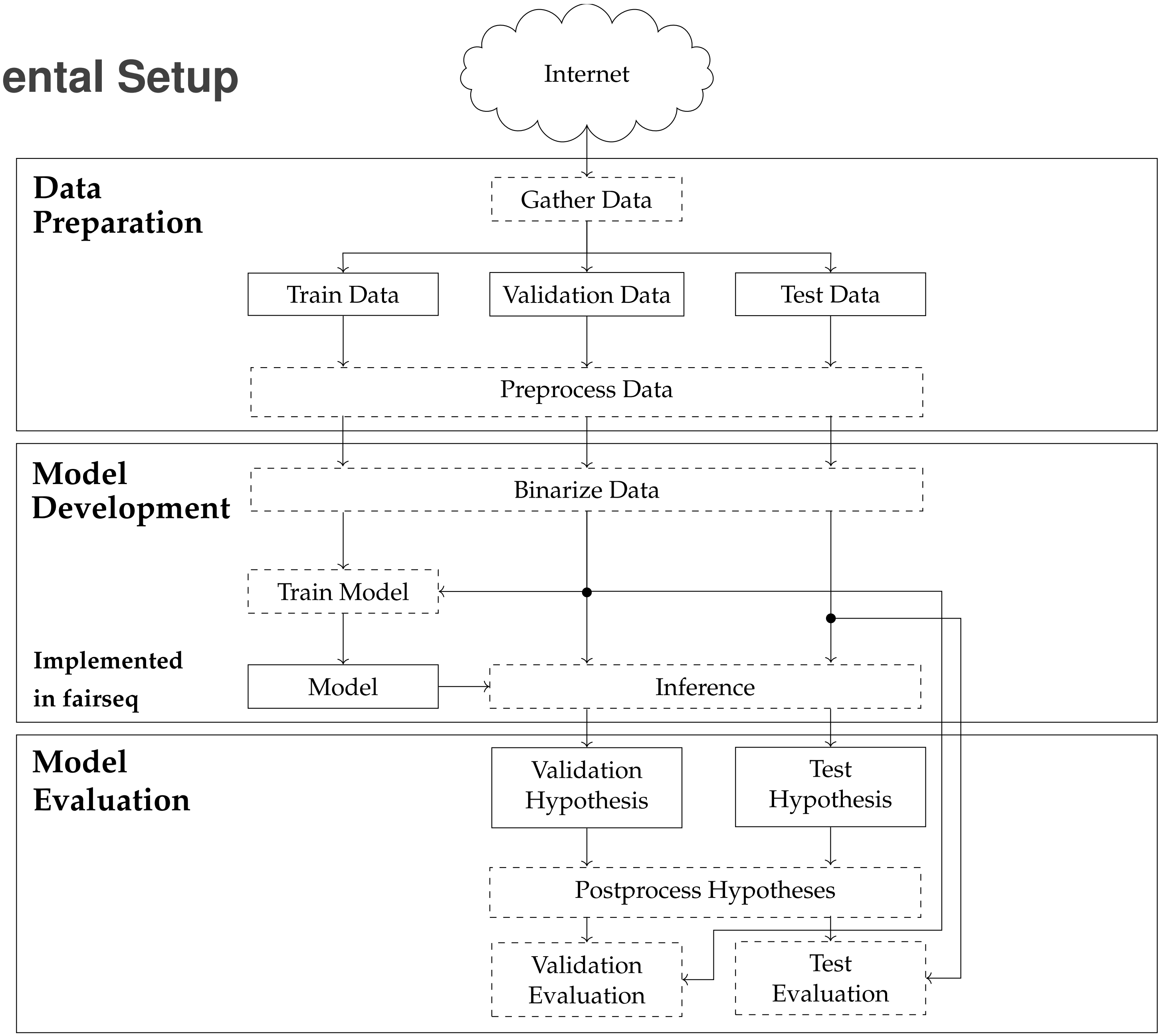
Differentiable Average Lagging

$$DAL_d = \frac{1}{|\mathbf{y}|} \sum_{i=1}^{|\mathbf{y}|} \left(g_{d'}(i) - (i-1)d \right), \qquad d = \frac{1}{\gamma} = \frac{|\mathbf{x}|}{|\mathbf{y}|}$$

$$g_{d'}(i) = \begin{cases} g(i) & i = 1 \\ \max\left(g(i), g_{d'}(i-1) + d \right) & i > 1 \end{cases}$$

Appendix

Experimental Setup



Appendix

- *Baseline*: X5Gon system trained in 2019

Offline MT Systems

General Domain

Name	Filtering			Processing				
	Langid	Length	Ratio	Apostrophes	Tokenize	Truecase	BPE	SPM
V1	-	-	-	-	X	X	X	-
V2	X	<150	1.5	-	X	X	X	-
V3	X	<150	1.5	X	-	X	-	X

In-Domain

- V3-FT-CN: V3 fine-tuned on the CERN News corpus
- V3-FT-BT: V3 fine-tuned on 50K backtranslations of the CDS corpus
- V4: V3 + backtranslations

Streaming MT Systems

- V3-TT: Test-Time Wait-k applied to V3
- V3-MP: Multi-Path Wait-k training with same data processing as V3