SEGMENTATION-FREE STREAMING MACHINE TRANSLATION

Javier Iranzo-Sánchez Jorge Iranzo-Sánchez Adrià Giménez

Jorge Civera Alfons Juan

www.mllp.upv.es

MLLP | Machine Learning and Language Processing

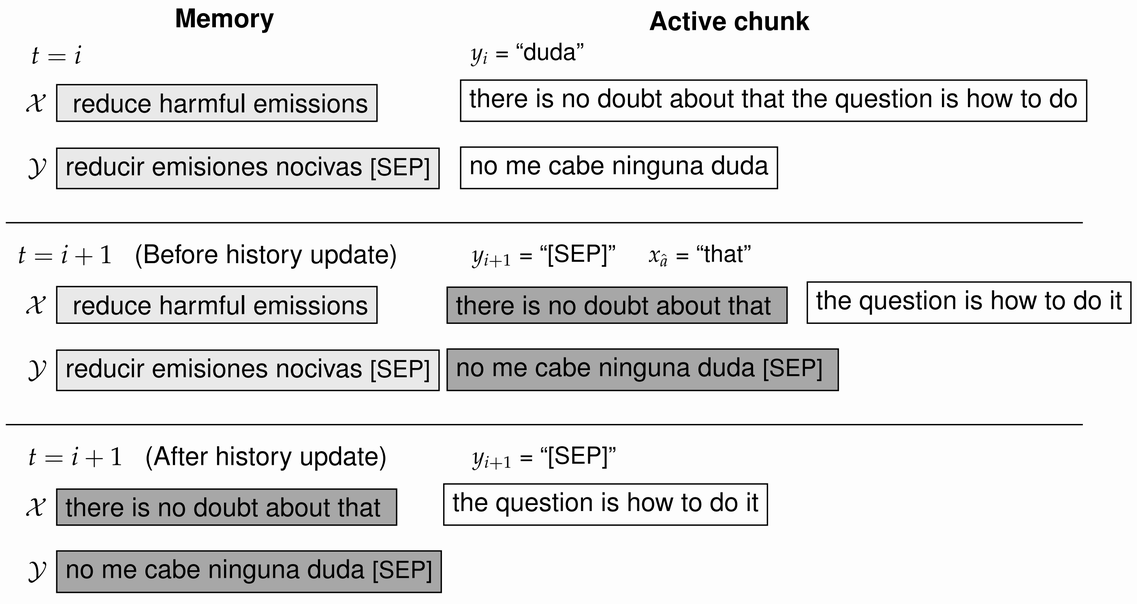
VRAIN Valencian Research Institute for Artificial Intelligence

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

INTRODUCTION

- Streaming Machine Translation: Translate unbounded input text stream in real-time
- Mismatch between MT model training (sentence-level) and streaming input
- Previous systems use external segmentation models
- Our approach: **Let the translation model learn to segment**

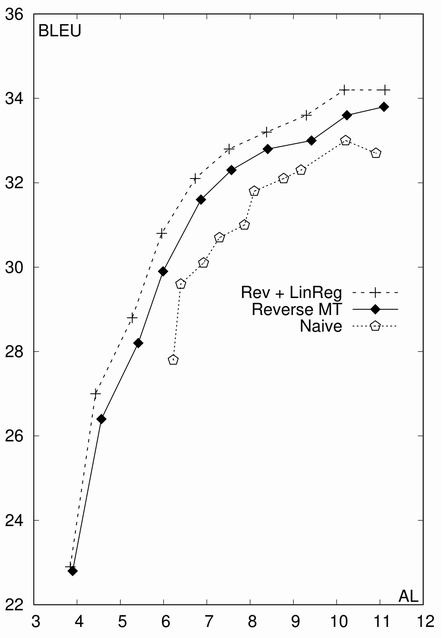
MEMORY MECHANISM

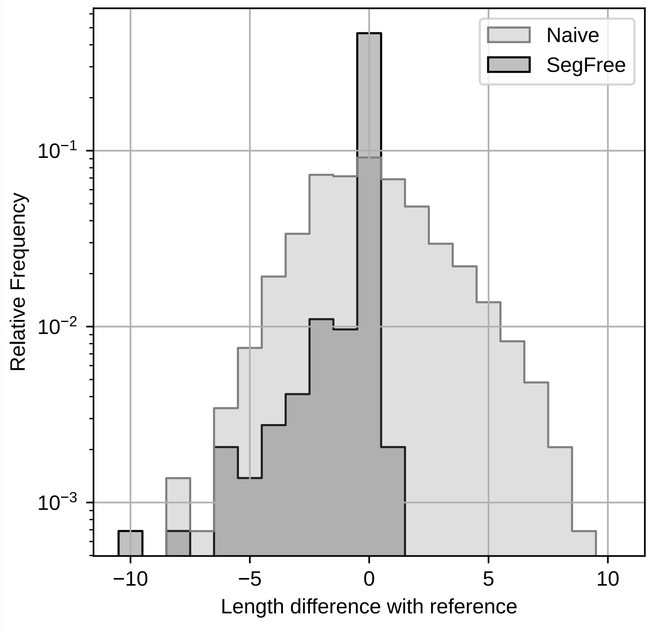


- Model learns to emit special [SEP] token during training
- Memory mechanism tracks positions of src-tgt segments in streams
- Log-linear model detects a positions when generating [SEP]

$$\hat{a} = \arg\max_{a} \sum_{f} \lambda_f \log h_f(a, \boldsymbol{x}, \hat{\boldsymbol{y}}).$$

- Lot of flexibility when defining features $h(\cdot)$





EXPERIMENTAL SETUP

Original	Prefix-augmented
I'm going to talk today about energy and climate.	I'm going to talk today
Heute spreche ich zu Ihnen über Energie und Klima.	Heute spreche ich zu Ihnen
Think about it. [SEP] The PC is a miracle.	Think about it. [SEP] The PC is
Denk darüber nach. [SEP] Der PC ist ein Wunder.	Denk darüber nach. [SEP] Der PC ist

Data:

- Training: Corpus from OPUS MT + IWSLT ST corpora (∼95-320M)
- Augmented document-level corpora and regular bitext corpora
- Lowercase source → Punctuated and truecased target

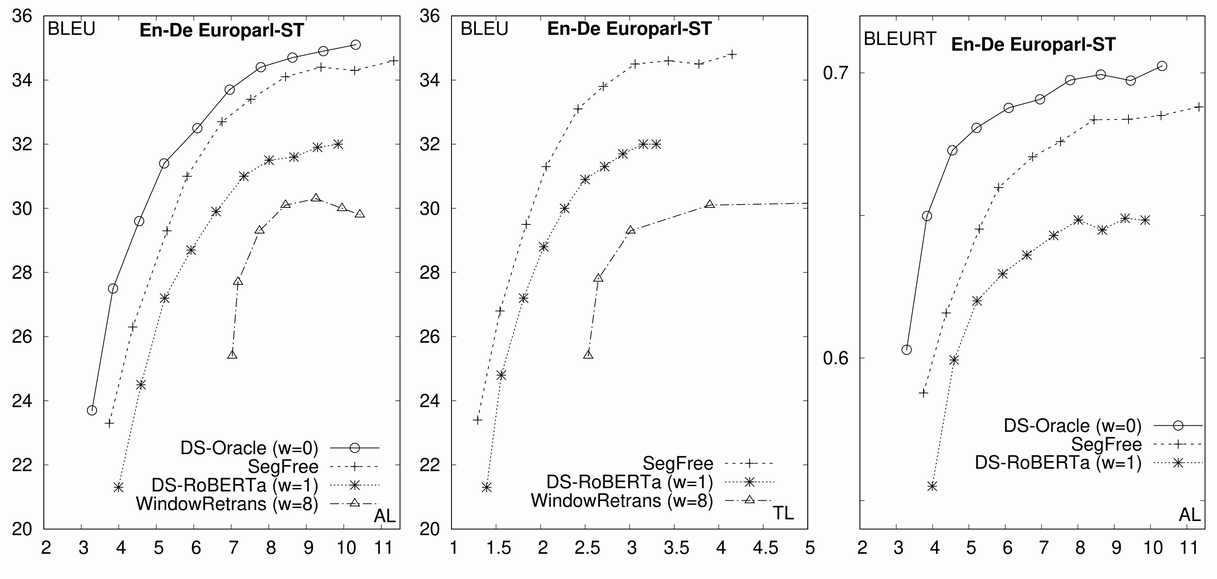
Models:

- Enc-Dec Transformers with bidirectional encoder and wait-k
- Can be extended to adaptive policies and other architectures

Evaluation:

- Stream-level latency metrics and traditional/neural quality metrics
- Eval: **MuST-C** and **Europarl-ST** datasets
- Langs: En ↔ De and En → Fr, Es

SYSTEM EVALUATION



- Results are consistent with compute aware and neural metrics
- En → Fr, Es and MusT-C show similar results

CONCLUSIONS

- Flexible framework to create true streaming MT models.
- Improved performance and reduced latency compared to baselines
- Eliminates need for external segmentation models, simplifying the translation pipeline.

Acknowledgments

The research leading to these results has received funding from European Union's Horizon 2020 research and innovation program under grant agreement no. 952215 and EU4Health Programme 2021-2027 as part of Europe's Beating Cancer Plan under Grant Agreements nos. 101056995 and 101129375; and from the Government of Spain's grant PID2021-122443OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", grant PDC2022-133049-I00 funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR" and FPU scholarship FPU18/04135. The authors gratefully acknowledge the financial support of the Generalitat Valenciana under project IDIFEDER/2021/059.