Learning From Wrong Predictions in Low-Resource Neural Machine Translation

Jia Cheng Hu, Roberto Cavicchioli, Giulia Berardinelli, Alessandro Capotondi

Speaker: Jia Cheng Hu

LREC-COLING 2024
Turin, Italy
May 2024

Outline

- Introduction
  - Problem
  - Proposal
- Related Works
- USKI
- Experimental Results
- Conclusion

Introduction

Neural Machine Translation

- Converting sentences in one language into another using Neural Networks

- Multiple Benefits of Neural Machine Translation

- Neural Networks in this application are Data Hungry...

Low Resourced and Endangered Langs

- 50% of the 7000 currently spoken languages are estimated to be severely endangered or dead in 2100

- 20 languages spoken by half of the global population

- Neural Machine Translation systems are valuable from a cultural, societal and economic perspective

Proposal

- Common language dataset size N -> Order of millions
- Low resourced languages dataset size M -> order of thousand (at best)

Our work:

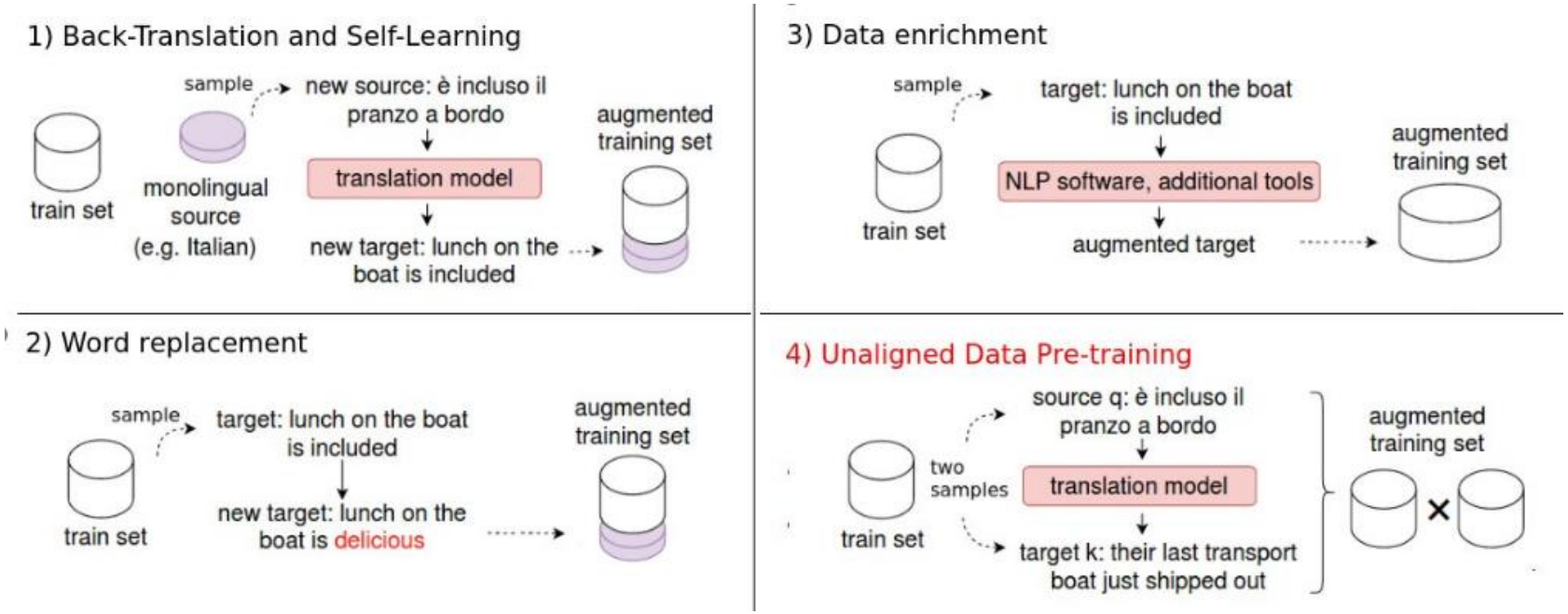
Unaligned Sentences KeyTokens Pre-traIning (USKI)

-> Learn from the corpus of unaligned sentences

New $M = M^2 \sim N$

Related Works

Related Works



USKI

KeyTokens

Initial dataset $D_{Lx, Ly}$.
Let D_{Lx} and D_{Ly} be the corpus of SRC and TRG sentences.
Construct $\tilde{D} = D_{L_X} \times D_{L_Y}$

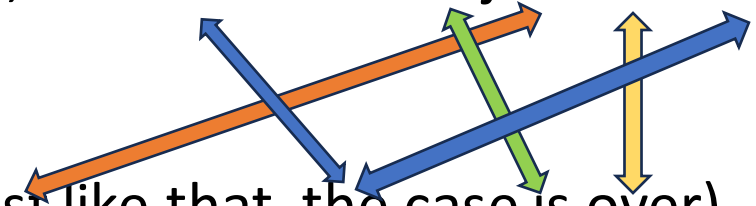
KeyTokens:
Given $X_i \in D_{Lx}$ and $Y_j \in D_{Ly}$ an arbitrary pair from $\tilde{D}$

The KeyTokens are defined as $K_{i,j} = Y_i \cap Y_j$

KeyTokens example (Italian – English)

Pair i:

(Xi = La casa è appena dietro la collina, Yi = The house is just over the hill)



Pair j:

(Xj =Ecco fatto, il caso è chiuso, Yj = Just like that, the case is over)

KeyTokens: ('just', 'the', 'is', 'over')

Unaligned Sentences KeyTokens Pre-traIning (USKI)

$$-\sum_{t}^{T} \mathbb{1}(g_t) \cdot log(p_\theta(y_t = g_t | y_{1:t-1}, X_i)) \cdot w(g_t)$$

Selector * CrossEntropy * IDF Weight function

Goal

- Increase quality of translation

  By increasing robustness and accuracy over a special set of tokens, the KeyTokens. The most frequent matching tokens between unaligned sentences.

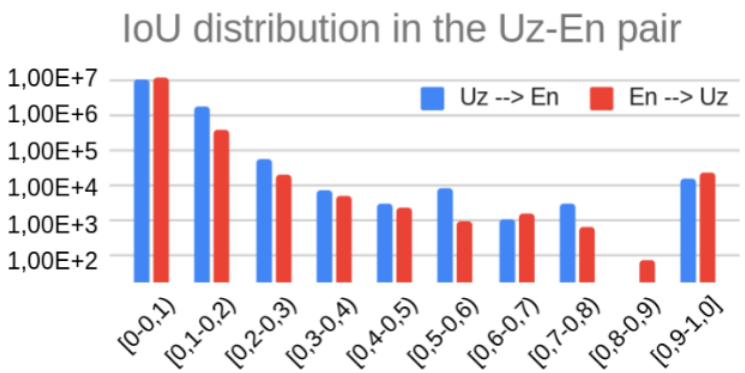
- Increase robustness in autoregressive decoding

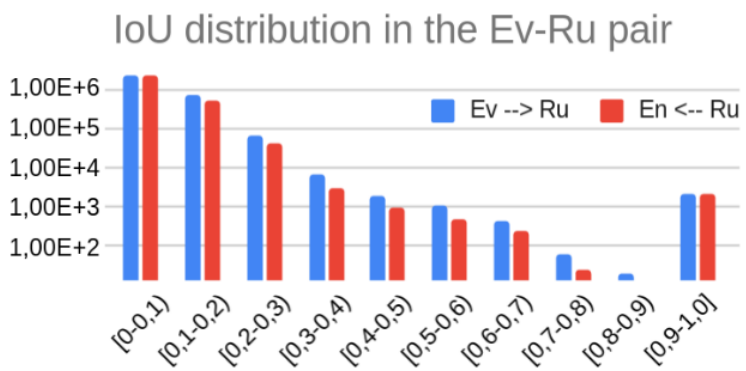
Experimental Results

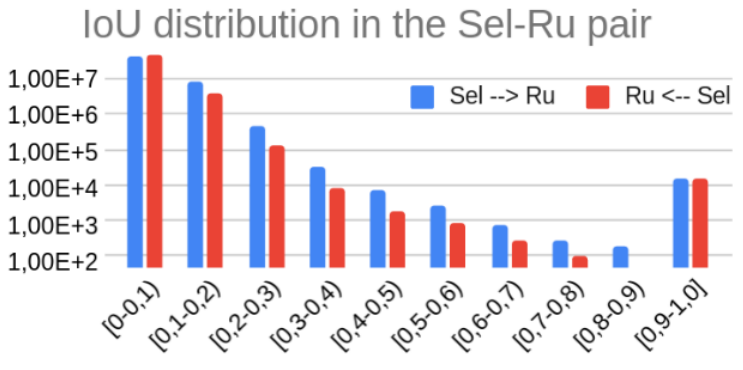
Datasets

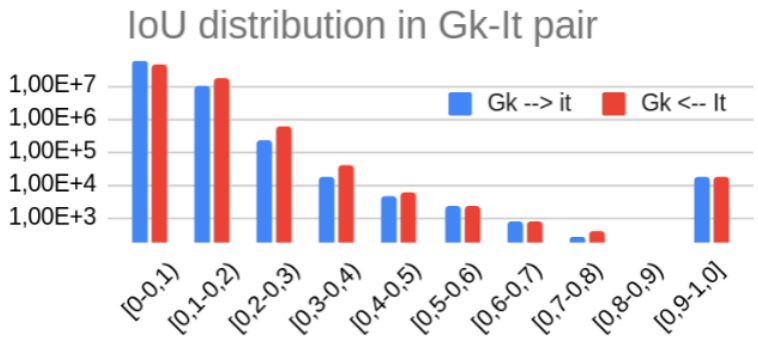
Language pair	# training	(# training)^2	# validation	# test
Sel-Ru	7251	52.57 $\cdot 10^6$	200	200
Ev-Ru	2136	4.56 $\cdot 10^6$	100	400
Gk-It	8136	66.19 $\cdot 10^6$	200	1000
Uz-En	3689	13.60 $\cdot 10^6$	99	199
Wol-It	5916	34.99 $\cdot 10^6$	200	1000

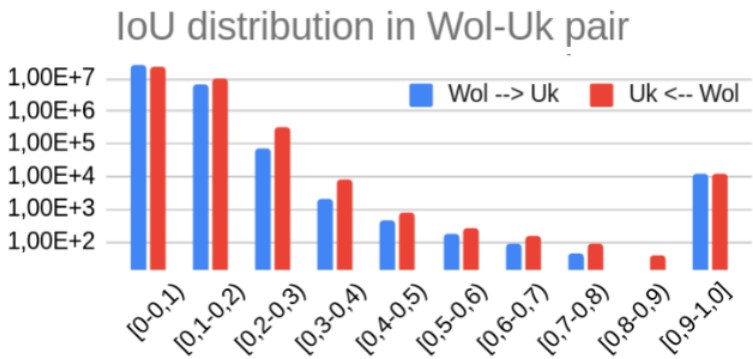
IoU filtering



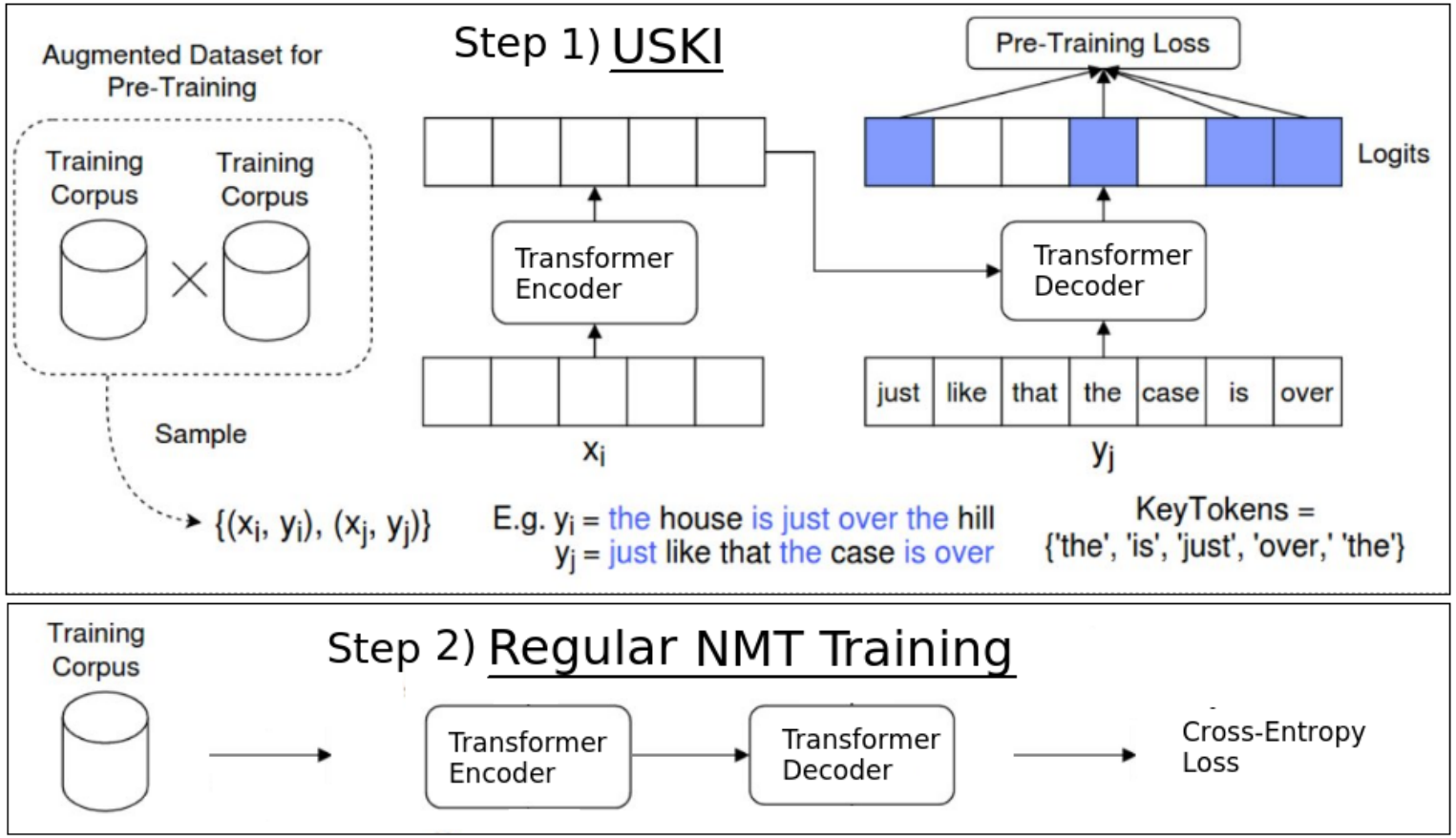








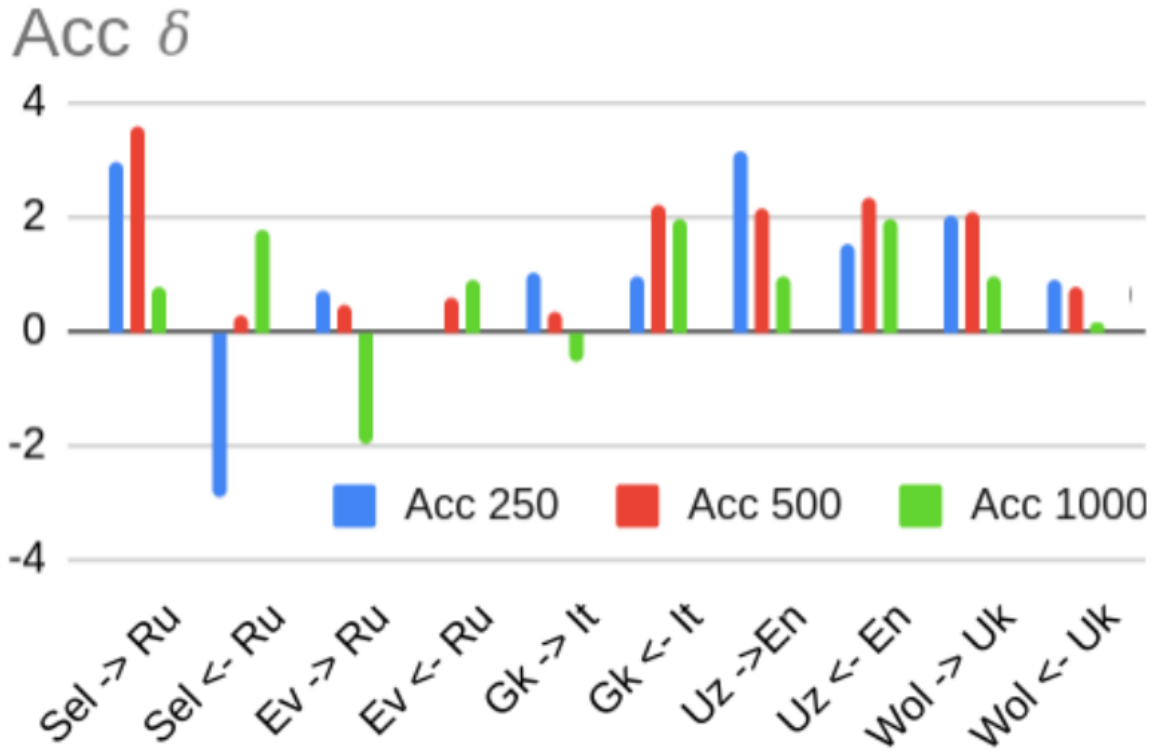
Training Overview



BLEU Improvements

Task	Baseline	w/ USKI	$\delta\uparrow$
Sel→Ru	7.05±0.50	8.19±0.38	1.14
Sel←Ru	4.15±0.45	5.98±0.34	1.83
Ev→Ru	6.58±0.48	7.26±0.39	0.68
Ev←Ru	7.36±0.85	8.65±0.90	1.29
Gk→It	5.91±0.10	6.23±0,.10	0.32
Gk←It	4.43±0.07	5.58±0.17	1.15
Uz→En	18.82±0.74	19.94±0.38	1.12
Uz←En	18.40±0.43	19.07±0.38	0.67
Wol→Uk	4.60±0.13	5.00±0.12	0.40
Wol←Uk	8.25±0.05	8.62±0.14	0.37

Accuracy on Single Tokens



Impact of Sub-word tokenization

Vocab size	Baseline	w/ USKI	$\delta \uparrow$
4356	19.07	19.64	0.57
7983	19.68	20.08	0.40
11385	19.49	19.95	0.46
13901	20.06	20.29	0.23

Results against State-of-the-Art

Task	Baseline	w/ USKI	mBART
Sel→Ru	7.05	8.19	18.50
Ev→Ru	6.58	7.26	26.03
Gk→It	5.91	6.23	9.12
Uz→En	18.82	19.94	20.88
Wol→Uk	4.69	5.00	6.24

Conclusion

Conclusion

- Addressing Data Shortage with incorrectly paired sentences
- Avg. 0.9 BLEU increase across all translation tasks with USKI
- Only 13 % of the entire corpus was used

Future works

-> develop more techniques to leverage the entire corpus

-> apply techniques on established SotA models.

Thank You

Contacts:
jiacheng.hu@unimore.it
roberto.cavicchioli@unimore.it
giulia.berardinelli@unimore.it
alessandro.capotondi@unimore.it

24