High-resource Language-specific Training for Multilingual Neural Machine Translation

Jian Yang[1], Yuwei Yin[2], Shuming Ma[2], Dongdong Zhang[2], Zhoujun Li[1], Furu Wei[2]

[1]State Key Lab of Software Development Environment,

Beihang University, Beijing, China

[2]Microsoft Research Asia

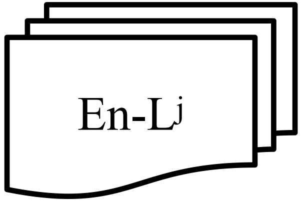
IJCAI ECAI VIENNA 22

Outline

Bilingual vs. Mulilngual

Given English-centric parallel corpora En-L^i ($L^i \neq$ En, $1 \leq i \leq N$)

multiple parallel corpora
En-L^i (En$\neq L^i$, $1 \leq i \leq N$)



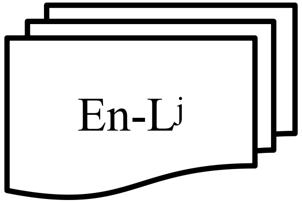
En_k and L^i_k ($1 \leq k \leq m$) are semantically equivalent

Bilingual vs. Mulilngual

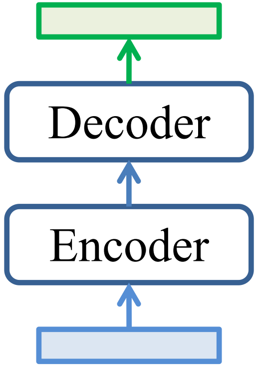
Given English-centric parallel corpora En-L^i ($L^i \neq$ En, $1 \leq i \leq N$)

Typically, we can train **2N bilingual** models

multiple parallel corpora
En-L^i (En$\neq L^i$, $1 \leq i \leq N$)



En_k and L^i_k ($1 \leq k \leq m$) are semantically equivalent



En→L^i

N × bilingual model F_{ei}

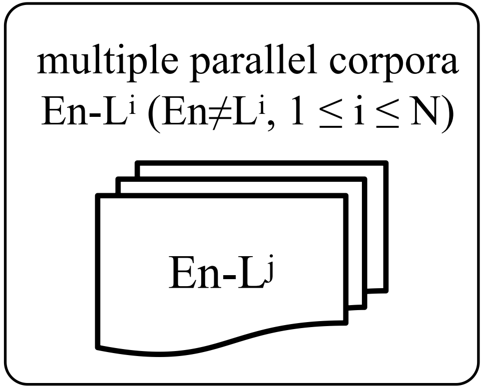
L^i→En

N × bilingual model F_{ie}

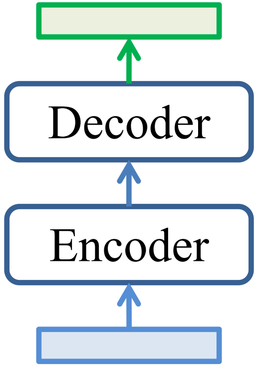
Bilingual vs. Mulilngual

Given English-centric parallel corpora En-L^i ($L^i \neq$ En, $1 \leq i \leq N$)

Typically, we can train 2N bilingual models, or **2 multilingual** models.



En_k and L^i_k ($1 \leq k \leq m$) are semantically equivalent

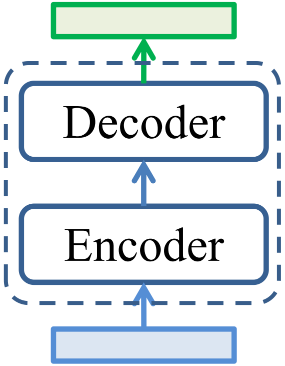


En$\rightarrow L^i$

N × bilingual model F_{ei}

$L^i \rightarrow$En

N × bilingual model F_{ie}



shared model

En$\rightarrow L^i$

1 × multilingual model F_{ei}

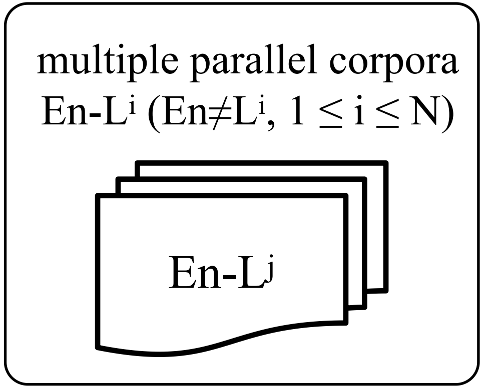
$L^i \rightarrow$En

1 × multilingual model F_{ie}

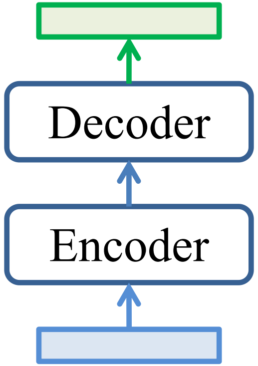
Bilingual vs. Mulilngual

Given English-centric parallel corpora En-L^i ($L^i \neq$ En, $1 \leq i \leq N$)

Typically, we can train 2N bilingual models, or **1 multilingual** models.



En_k and L^i_k ($1 \leq k \leq m$) are semantically equivalent

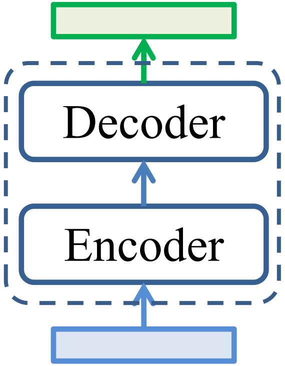


En→L^i

N × bilingual model F_{ei}

L^i→En

N × bilingual model F_{ie}



shared model

both En→L^i and L^i→En

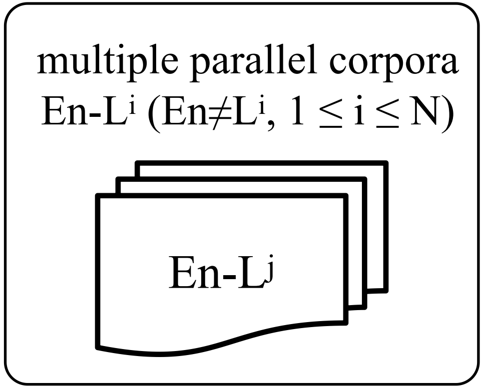
1 × multilingual model F_e

Bilingual vs. Mulilngual

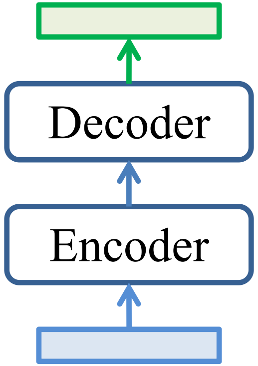
Given English-centric parallel corpora En-L^i ($L^i \neq$ En, $1 \leq i \leq N$)

Typically, we can train 2N bilingual models, or 1 multilingual models.

Which model is better on each En→L^i and L^i→En direction?



En_k and L^i_k ($1 \leq k \leq m$) are semantically equivalent



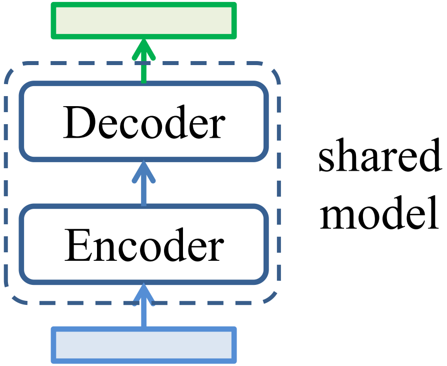
En→L^i

N × bilingual model F_{ei}

L^i→En

N × bilingual model F_{ie}

VS.



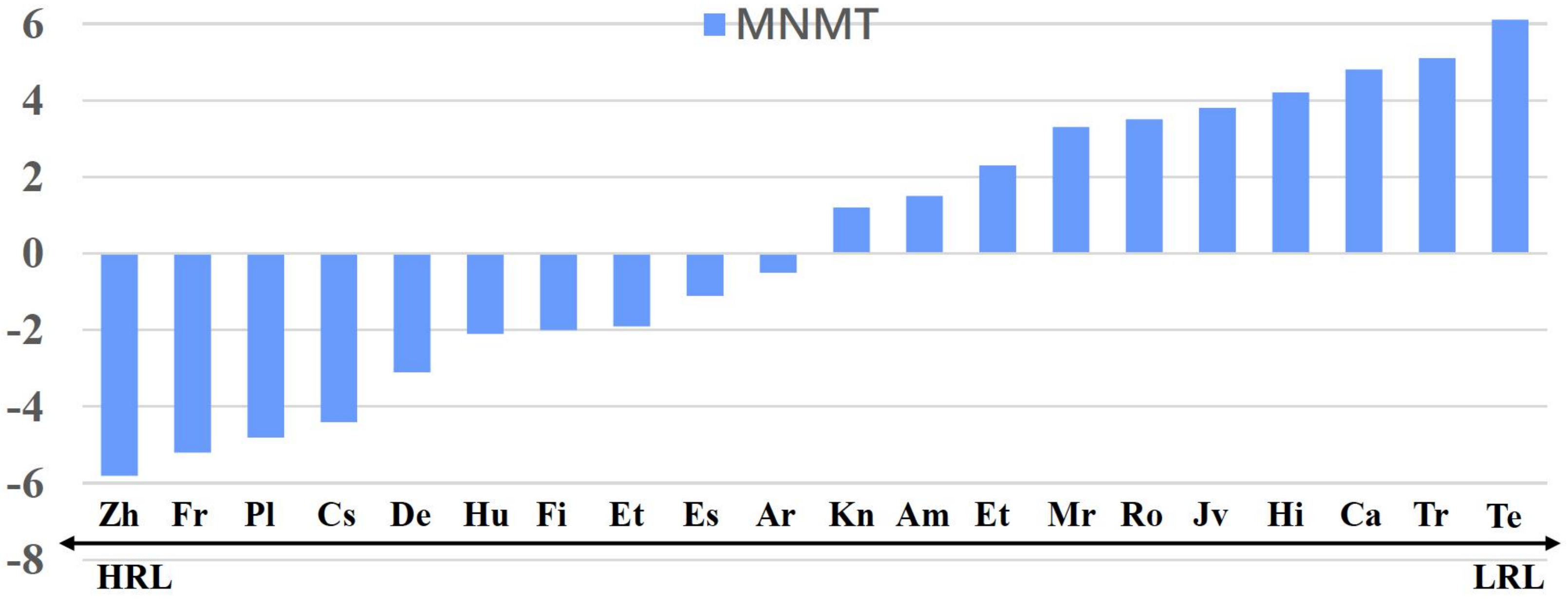
both En→L^i and L^i→En

1 × multilingual model F_e

Bilingual vs. Mulilngual

It depends on the **richness** of language L^i

HRL: High-Resource Language; **LRL**: Low-Resource Language

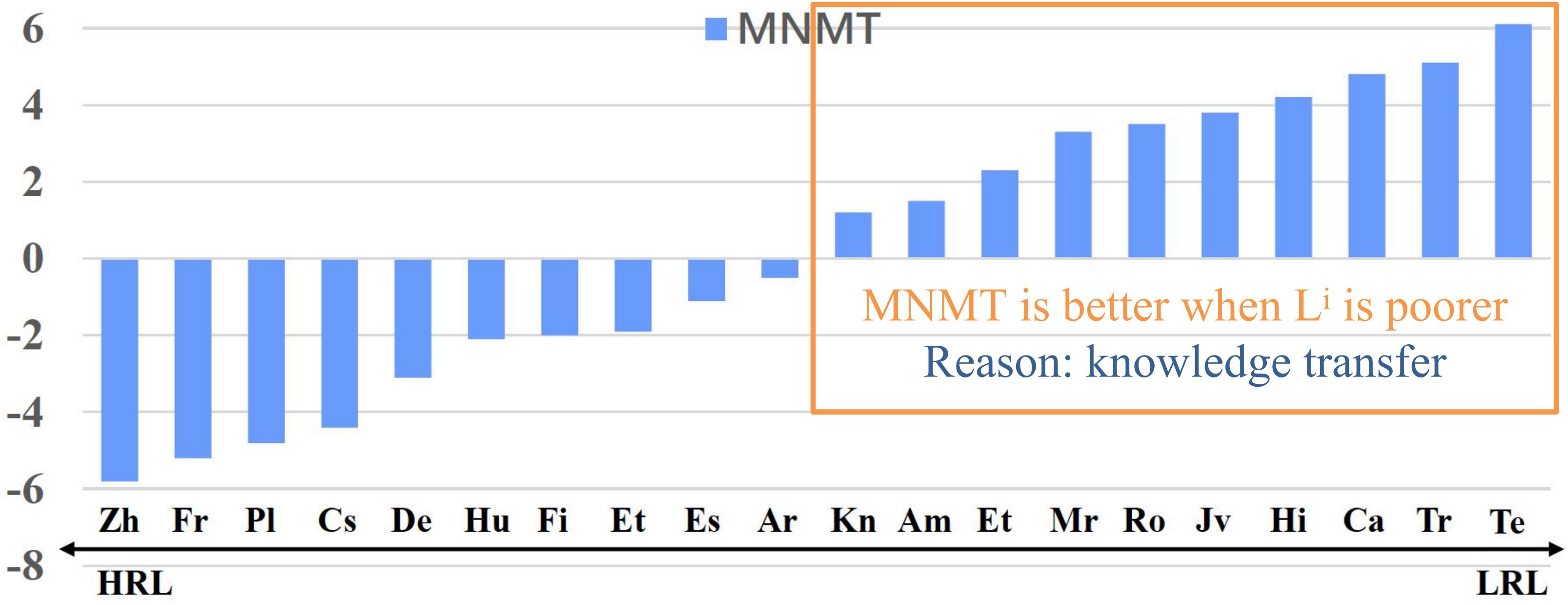


ΔBLEU socre between MNMT and BiNMT on En→L^i and L^i→En (averaged)

Bilingual vs. Mulilngual

It depends on the **richness** of language L^i

HRL: High-Resource Language; **LRL**: Low-Resource Language

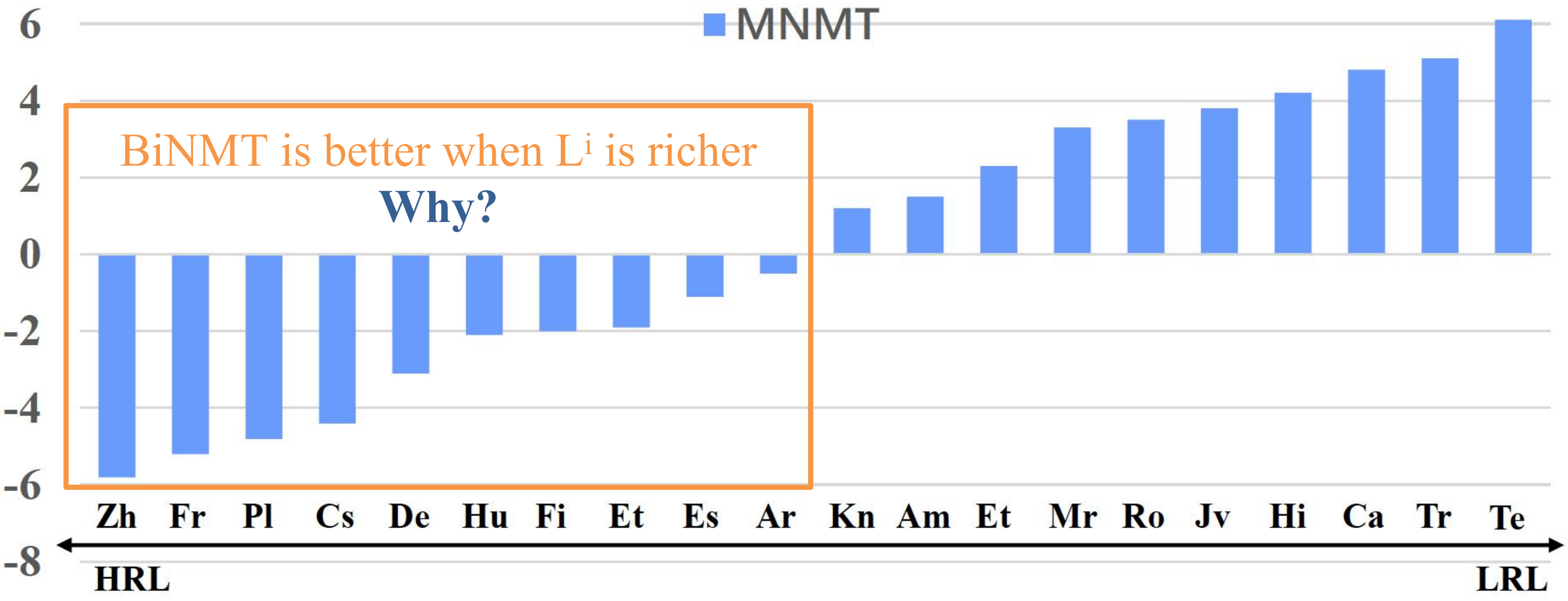


ΔBLEU socre between MNMT and BiNMT on En→L^i and L^i→En (averaged)

Bilingual vs. Mulilngual

It depends on the **richness** of language L^i

HRL: High-Resource Language; **LRL**: Low-Resource Language

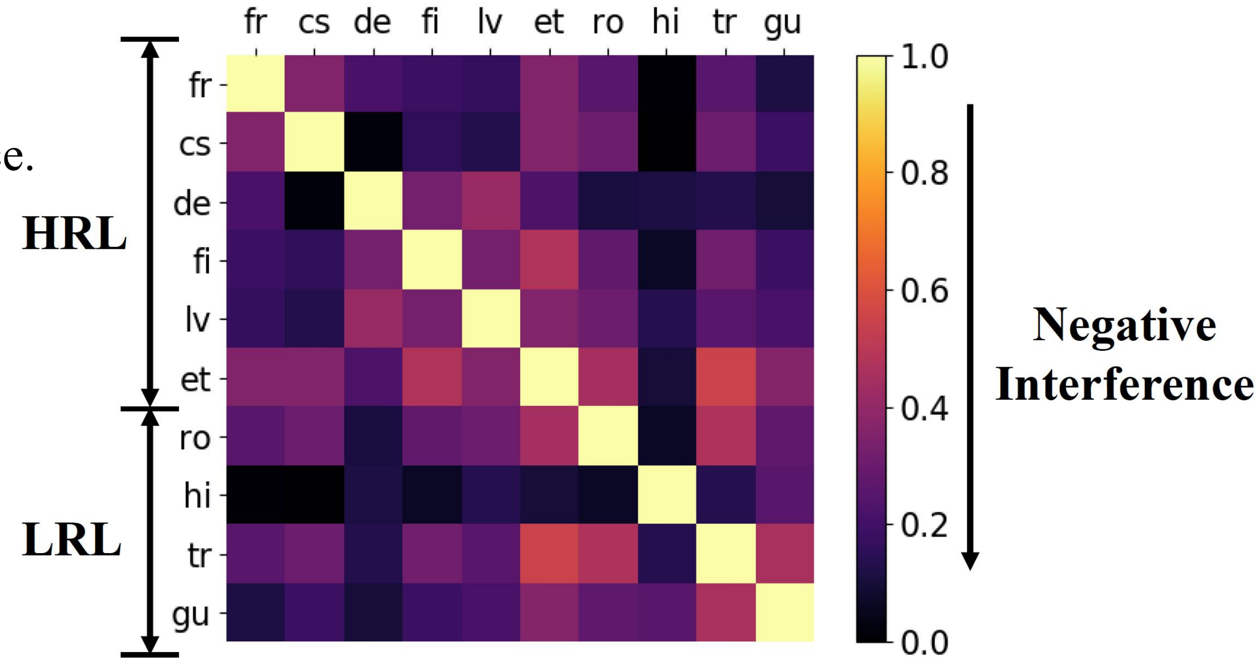


ΔBLEU socre between MNMT and BiNMT on En→L^i and L^i→En (averaged)

Negative Language Interference

Different directions conflict with each other to various extents.

The less gradient similarity,
the darker the color,
the more negative interference.



cosine similarities between
gradients of two translation directions

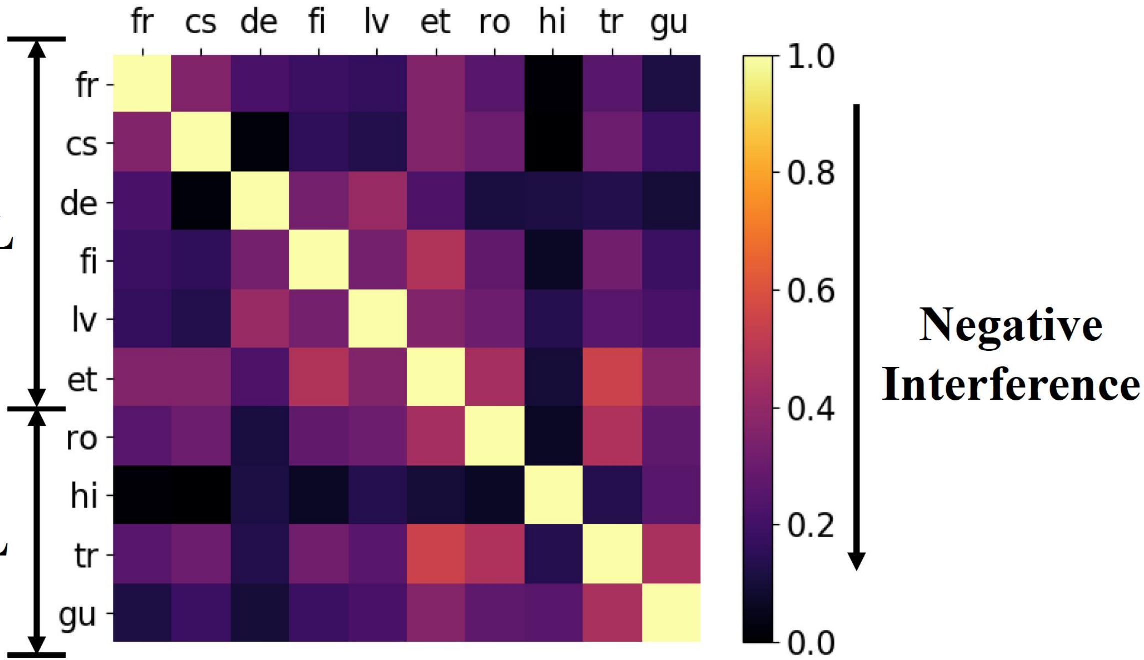
Negative Language Interference

Different directions conflict with each other to various extents.

The less gradient similarity,
the darker the color,
the more negative interference.

Our Goals

I. Mitigate the negative interference among languages.
II. Prevent the HRL from negative interference introduced by LRL.
III. Retain high translation quality of all directions.

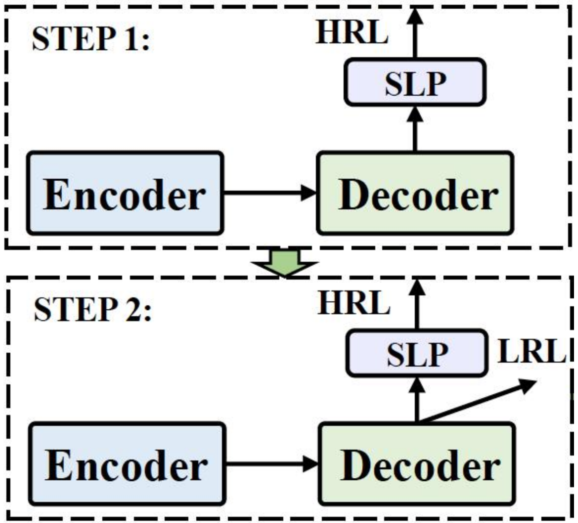


cosine similarities between
gradients of two translation directions

Outline

Method Overview

Two-Stage Traning



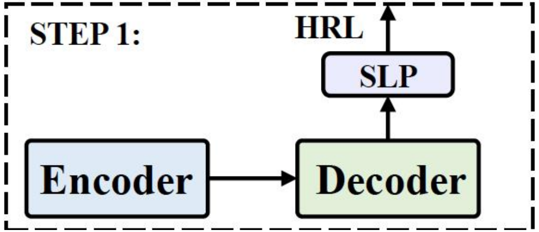
(a) Two-Stage Training

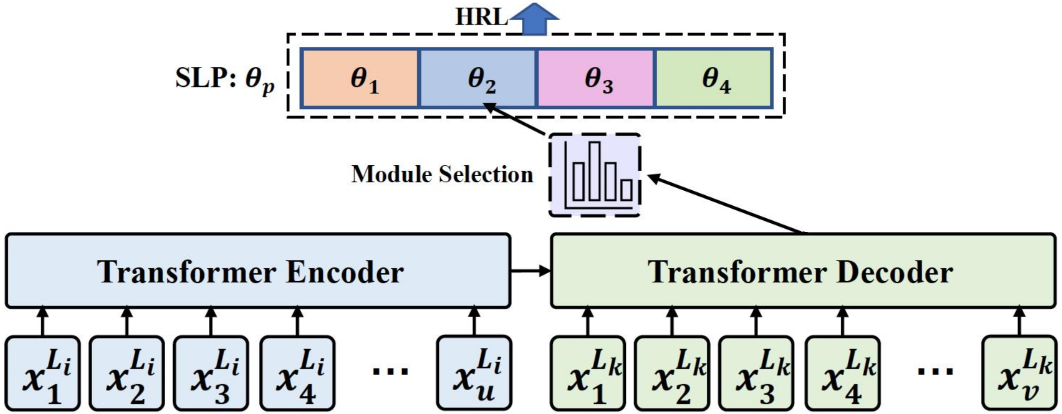
Step 1: train a MNMT model on HRLs

Step 2: continue training the model on all pairs

Method Overview

Step 1: train a MNMT model on HRLs



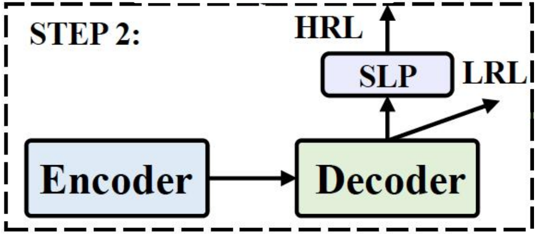


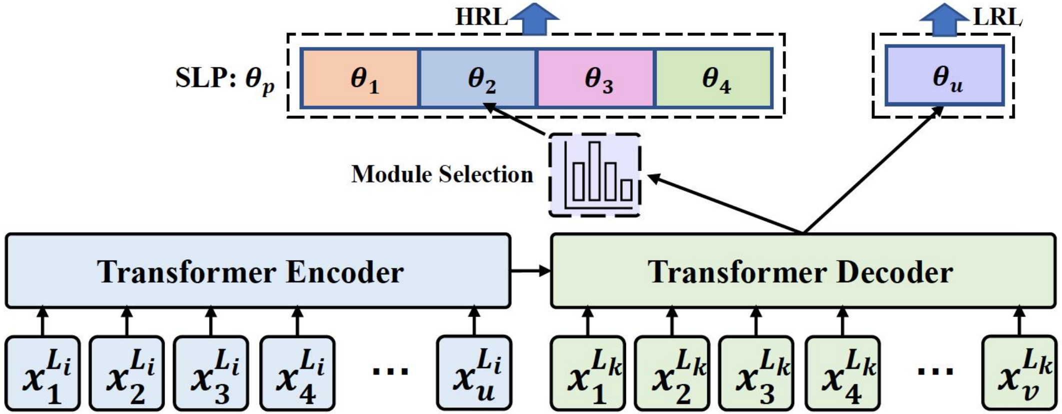
(b) Model Architecture

- ➢ no negative interference from LRLs
- ➢ mitigate negative interference among HRLs
  - ❖ **SLP**: Selective Language-specific Pool

Method Overview

Step 2: continue training on all pairs (HRLs & LRLs)



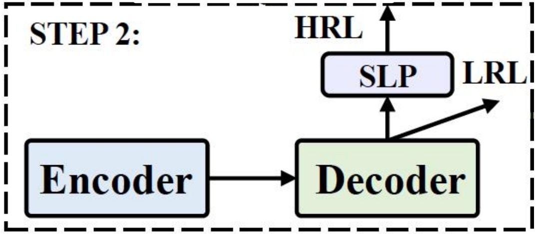


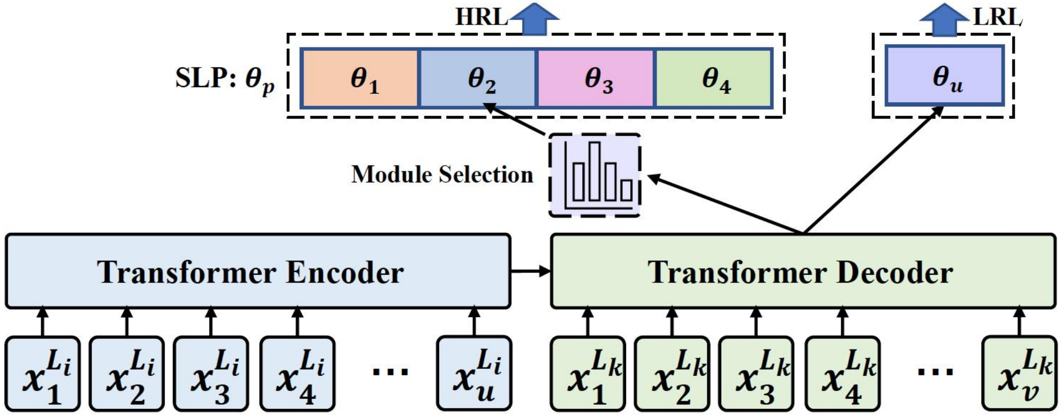
(b) Model Architecture

- HRLs still use SLP selection mechanism
- LRLs utilize the trained MNMT model

Method Overview

Step 2: continue training on all pairs (HRLs & LRLs)





(b) Model Architecture

- ➢ HRLs still use SLP selection mechanism
- ➢ LRLs utilize the trained MNMT model
  - ✓ share the same MNMT → Knowledge Transfer
  - ✓ less training batches on LRLs → avoid overfitting

Outline

Dataset

WMT-10: En-X (X in {Fr, Cs, De, Fi, Lv, Et, Ro, Hi, Tr, Gu})

HRLs: Fr, Cs, De, Fi, Lv, and Et; LRLs: Ro, Hi, Tr, and Gu

Code	Language	#Bitext	Training	Valid	Test
Fr	French	10M	WMT15	Newstest13	Newstest15
Cs	Czech	10M	WMT19	Newstest16	Newstest18
De	German	4.6M	WMT19	Newstest16	Newstest18
Fi	Finnish	4.8M	WMT19	Newstest16	Newstest18
Lv	Latvian	1.4M	WMT17	Newsdev17	Newstest17
Et	Estonian	0.7M	WMT18	Newsdev18	Newstest18
Ro	Romanian	0.5M	WMT16	Newsdev16	Newstest16
Hi	Hindi	0.26M	WMT14	Newsdev14	Newstest14
Tr	Turkish	0.18M	WMT18	Newstest16	Newstest18
Gu	Gujarati	0.08M	WMT19	Newsdev19	Newstest19

OPUS-100: 94 En-X pairs: 95 langs including En, except 5 langs w/o valid/test sets

High-resource: 45 pairs; Medium-resource: 21 pairs; Low-resource: 28 pairs.

Baseline

1. BiNMT: bilingual Transformer model
2. MNMT: multilingual Transformer model trained on all directions
3. mBART: multilingual BART (denoising autoencoder for pretraining seq-to-seq models) model, fine-tuned on all directions
4. XLM-R: pretained Transformer-based masked language model on 100 languages
5. LS-MNMT: language-specific many-to-many multilingual model trained on 100 languages

Implementation

Architecture of all experiments: **Transformer**

learning rate: 5e-4
warmup steps: 4000
optimizer: Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$)
mini-batch size: 4096 tokens
loss: label smoothing cross-entropy (smoothing ratio = 0.1)

training device: 64 Tesla V100 GPUs

Evaluation Metrics: the case-sensitive detokenized BLEU using sacreBLEU

BLEU+case.mixed+lang.{src}-{tgt}+numrefs.1+smooth.exp+tok.13a+version.1.3.1

Outline

Experimental Results: WMT-10

En→X on WMT-10: 1→1 (bilingual), 1→N (one-to-many), N→N (many-to-many) models

En→X test sets		#Params	Fr	Cs	De	Fi	Lv	Et	Ro	Hi	Tr	Gu	Avg_{all}
			HRLs						LRLs				
1→1	BiNMT [Vaswani *et al.*, 2017]	242M/10M	36.3	22.3	40.2	15.2	16.5	15.0	23.0	12.2	13.3	7.9	20.2
1→N	MNMT [Johnson *et al.*, 2017]	242M	34.2	20.9	40.0	15.0	18.1	20.9	26.0	14.5	17.3	13.2	22.0
	mBART [Liu *et al.*, 2020]	611M	33.7	20.8	38.9	14.5	18.2	20.5	26.0	15.3	16.8	12.9	21.8
	XLM-R [Conneau *et al.*, 2020]	362M	34.7	21.5	40.1	15.2	18.6	20.8	26.4	15.6	**17.4**	**14.9**	22.5
	LS-MNMT [Fan *et al.*, 2020]	409M	35.0	21.7	40.6	15.5	18.9	21.0	26.2	14.8	16.5	12.8	22.3
	HLT-MT (Our method)	381M	**36.2**	**22.2**	**41.8**	**16.6**	**19.5**	**21.1**	**26.6**	**15.8**	17.1	14.6	**23.2**
N→N	MNMT [Johnson *et al.*, 2017]	242M	34.2	21.0	39.4	15.2	18.6	20.4	26.1	15.1	17.2	13.1	22.0
	mBART [Liu *et al.*, 2020]	611M	32.4	19.0	37.0	13.2	17.0	19.5	25.1	**15.7**	16.7	14.2	21.0
	XLM-R [Conneau *et al.*, 2020]	362M	34.2	21.4	39.7	15.3	18.9	20.6	26.5	15.6	17.5	14.5	22.4
	LS-MNMT [Fan *et al.*, 2020]	409M	34.8	21.1	39.3	15.2	18.7	20.5	26.3	14.9	17.3	12.3	22.0
	HLT-MT (Our method)	381M	**35.8**	**22.4**	**41.5**	**16.3**	**19.6**	**21.0**	**26.6**	**15.7**	**17.6**	**14.7**	**23.1**

- ➢ significantly outperform BiNMT on LRLs, yet retain high perfomance on HRLs
- ➢ clear improvement over previous multilingual baselines on HRLs and LRLs
- ➢ the extra model parameters for our SLP pool and Universal layer are modest

Experimental Results: WMT-10

En→X on WMT-10: 1→1 (bilingual), 1→N (one-to-many), N→N (many-to-many) models

En→X test sets		#Params	Fr	Cs	De	Fi	Lv	Et	Ro	Hi	Tr	Gu	Avg_{all}
			HRLs						LRLs				
1→1	BiNMT [Vaswani *et al.*, 2017]	242M/10M	36.3	22.3	40.2	15.2	16.5	15.0	23.0	12.2	13.3	7.9	20.2
1→N	MNMT [Johnson *et al.*, 2017]	242M	34.2	20.9	40.0	15.0	18.1	20.9	26.0	14.5	17.3	13.2	22.0
	mBART [Liu *et al.*, 2020]	611M	33.7	20.8	38.9	14.5	18.2	20.5	26.0	15.3	16.8	12.9	21.8
	XLM-R [Conneau *et al.*, 2020]	362M	34.7	21.5	40.1	15.2	18.6	20.8	26.4	15.6	**17.4**	**14.9**	22.5
	LS-MNMT [Fan *et al.*, 2020]	409M	35.0	21.7	40.6	15.5	18.9	21.0	26.2	14.8	16.5	12.8	22.3
	HLT-MT (Our method)	381M	**36.2**	**22.2**	**41.8**	**16.6**	**19.5**	**21.1**	**26.6**	**15.8**	17.1	14.6	**23.2**
N→N	MNMT [Johnson *et al.*, 2017]	242M	34.2	21.0	39.4	15.2	18.6	20.4	26.1	15.1	17.2	13.1	22.0
	mBART [Liu *et al.*, 2020]	611M	32.4	19.0	37.0	13.2	17.0	19.5	25.1	**15.7**	16.7	14.2	21.0
	XLM-R [Conneau *et al.*, 2020]	362M	34.2	21.4	39.7	15.3	18.9	20.6	26.5	15.6	17.5	14.5	22.4
	LS-MNMT [Fan *et al.*, 2020]	409M	34.8	21.1	39.3	15.2	18.7	20.5	26.3	14.9	17.3	12.3	22.0
	HLT-MT (Our method)	381M	**35.8**	**22.4**	**41.5**	**16.3**	**19.6**	**21.0**	**26.6**	**15.7**	**17.6**	**14.7**	**23.1**

- ➢ significantly outperform BiNMT on LRLs, yet retain high perfomance on HRLs
- ➢ clear improvement over previous multilingual baselines on HRLs and LRLs
- ➢ the extra model parameters for our SLP pool and Universal layer are modest

Experimental Results: WMT-10

En→X on WMT-10: 1→1 (bilingual), 1→N (one-to-many), N→N (many-to-many) models

			HRLs						LRLs				
En→X test sets		#Params	Fr	Cs	De	Fi	Lv	Et	Ro	Hi	Tr	Gu	Avg_{all}
1→1	BiNMT [Vaswani *et al.*, 2017]	242M/10M	36.3	22.3	40.2	15.2	16.5	15.0	23.0	12.2	13.3	7.9	20.2
1→N	MNMT [Johnson *et al.*, 2017]	242M	34.2	20.9	40.0	15.0	18.1	20.9	26.0	14.5	17.3	13.2	22.0
	mBART [Liu *et al.*, 2020]	611M	33.7	20.8	38.9	14.5	18.2	20.5	26.0	15.3	16.8	12.9	21.8
	XLM-R [Conneau *et al.*, 2020]	362M	34.7	21.5	40.1	15.2	18.6	20.8	26.4	15.6	**17.4**	**14.9**	22.5
	LS-MNMT [Fan *et al.*, 2020]	409M	35.0	21.7	40.6	15.5	18.9	21.0	26.2	14.8	16.5	12.8	22.3
	HLT-MT (Our method)	381M	**36.2**	**22.2**	**41.8**	**16.6**	**19.5**	**21.1**	**26.6**	**15.8**	17.1	14.6	**23.2**
N→N	MNMT [Johnson *et al.*, 2017]	242M	34.2	21.0	39.4	15.2	18.6	20.4	26.1	15.1	17.2	13.1	22.0
	mBART [Liu *et al.*, 2020]	611M	32.4	19.0	37.0	13.2	17.0	19.5	25.1	**15.7**	16.7	14.2	21.0
	XLM-R [Conneau *et al.*, 2020]	362M	34.2	21.4	39.7	15.3	18.9	20.6	26.5	15.6	17.5	14.5	22.4
	LS-MNMT [Fan *et al.*, 2020]	409M	34.8	21.1	39.3	15.2	18.7	20.5	26.3	14.9	17.3	12.3	22.0
	HLT-MT (Our method)	381M	**35.8**	**22.4**	**41.5**	**16.3**	**19.6**	**21.0**	**26.6**	**15.7**	**17.6**	**14.7**	**23.1**

- significantly outperform BiNMT on LRLs, yet retain high perfomance on HRLs
- clear improvement over previous multilingual baselines on HRLs and LRLs
- the extra model parameters for our SLP pool and Universal layer are modest

Experimental Results: OPUS-100

X→En and En→X on OPUS-100: N→N (many-to-many) models

Models (N→N)	#Params	X→En					En→X				
		$High_{45}$	Med_{21}	Low_{28}	Avg_{94}	WR	$High_{45}$	Med_{21}	Low_{28}	Avg_{94}	WR
Previous Best System [Zhang *et al.*, 2020]	254M	30.3	32.6	31.9	31.4	-	23.7	25.6	22.2	24.0	-
MNMT [Johnson *et al.*, 2017]	242M	32.3	35.1	35.8	33.9	*ref*	26.3	31.4	31.2	28.9	*ref*
XLM-R [Conneau *et al.*, 2020]	362M	33.1	35.7	36.1	34.6	-	26.9	31.9	31.7	29.4	-
LS-MNMT [Fan *et al.*, 2020]	456M	33.4	35.8	35.9	34.7	-	27.5	31.6	31.5	29.6	-
HLT-MT (Our method)	381M	**34.2**	**36.7**	**36.1**	**35.3**	75.5	**27.6**	**33.3**	**31.8**	**30.1**	78.7

- consistently outperform previous multilingual baselines on high/medium/low resource language pairs (both X→En and En→X directions)

Ablation Study

XLM-R	Two-stage Training	SLP	Avg_{high}	Avg_{low}	Avg_{all}
			24.9	17.8	22.0
	✓		25.4	18.0	22.4
	✓	✓	26.0	18.1	22.8
✓			25.2	18.5	22.5
✓	✓		26.0	17.9	22.8
✓	✓	✓	**26.2**	**18.5**	**23.2**

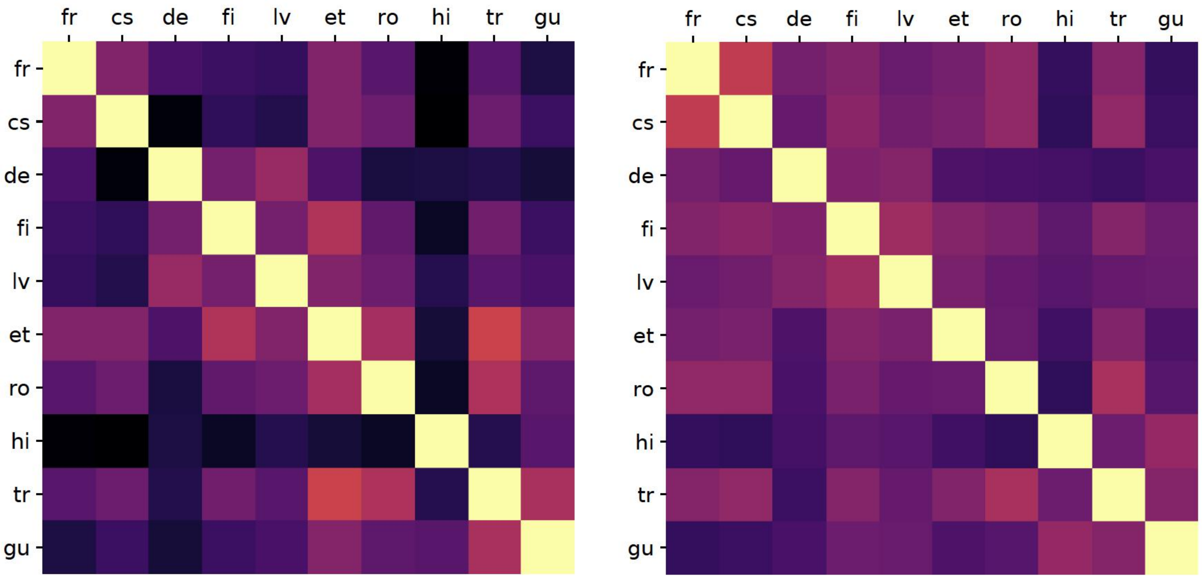
- XLM-R initialization, Two-stage Training strategy, and SLP selective mechanism are all beneficial.

Conflicting Gradient

gradiant similarity (cosine)

$$\Phi(L_a, L_b) = \frac{g_{L_a} \cdot g_{L_b}}{\|g_{L_a}\| \|g_{L_b}\|}$$

The less gradient similarity,
the darker the color,
the more negative interference.

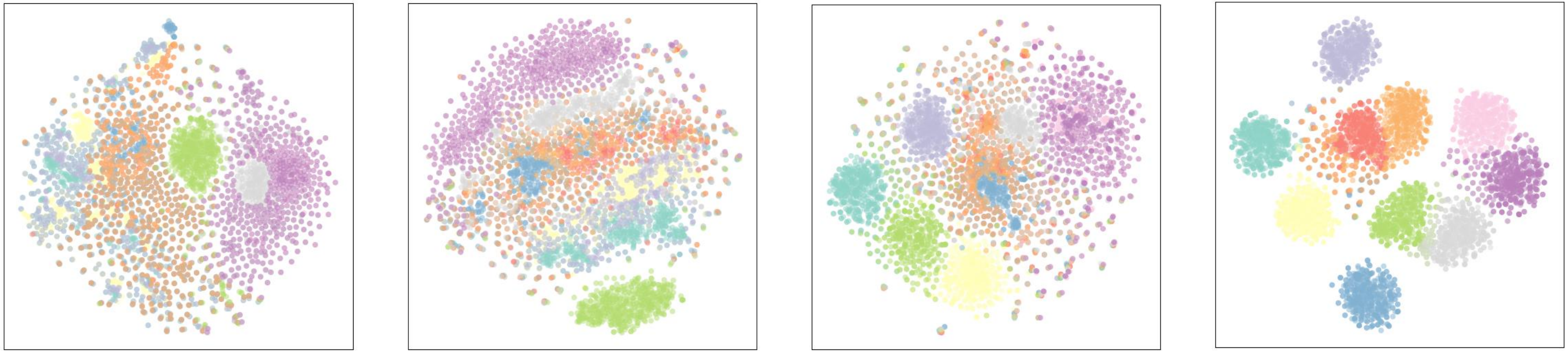


(a) Baseline (b) Our method

➢ clearly mitigate the negative interference among most directions

Decoder Representation Visualization

The t-SNE visualization of 500 random English sentences (hidden states of Decoder), ordered from the bottom decoder layer to the top layer. (a, b, c in Decoder, d in SLP)



(a) 2-th (b) 3-th (c) 6-th (d) 7-th

- ➢ different languages become more distinct and less likely to overlap with each other
- ➢ SLP effectively projects the language-shared representations into language-distinct ones for better target language generation.

Outline

01 Introduction

02 Method

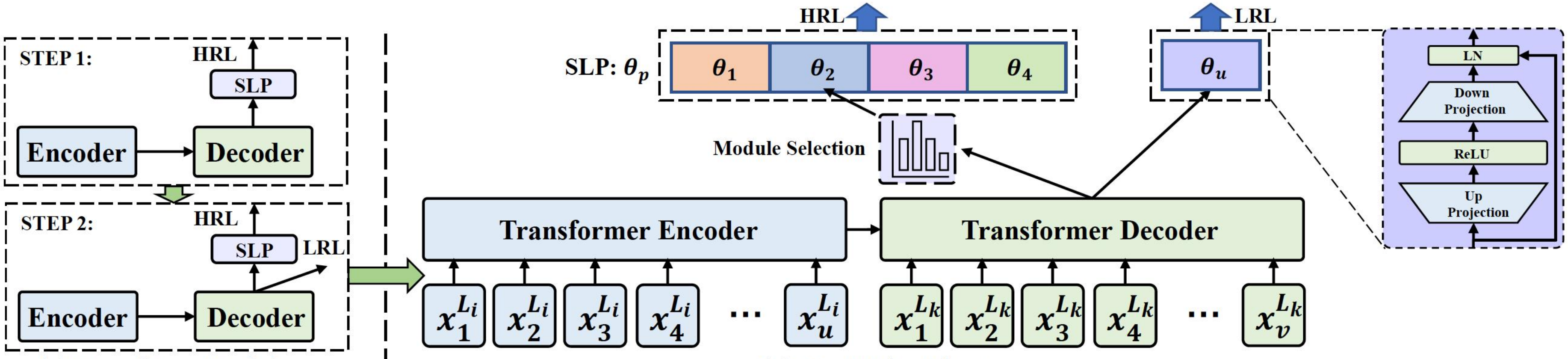
03 Experiment

04 Analysis

05 Conclusion

Conclusion

✓ In this work, we propose **a novel multilingual translation model** with the high-resource language-specific training called **HLT-MT**.

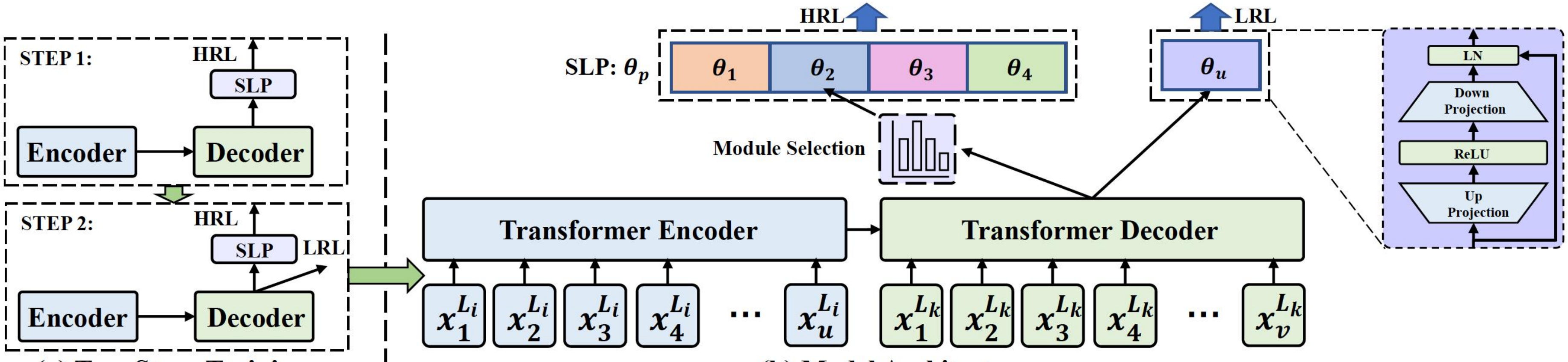


(a) Two-Stage Training (b) Model Architecture

Conclusion

✓ In this work, we propose **a novel multilingual translation model** with the high-resource language-specific training called **HLT-MT**.

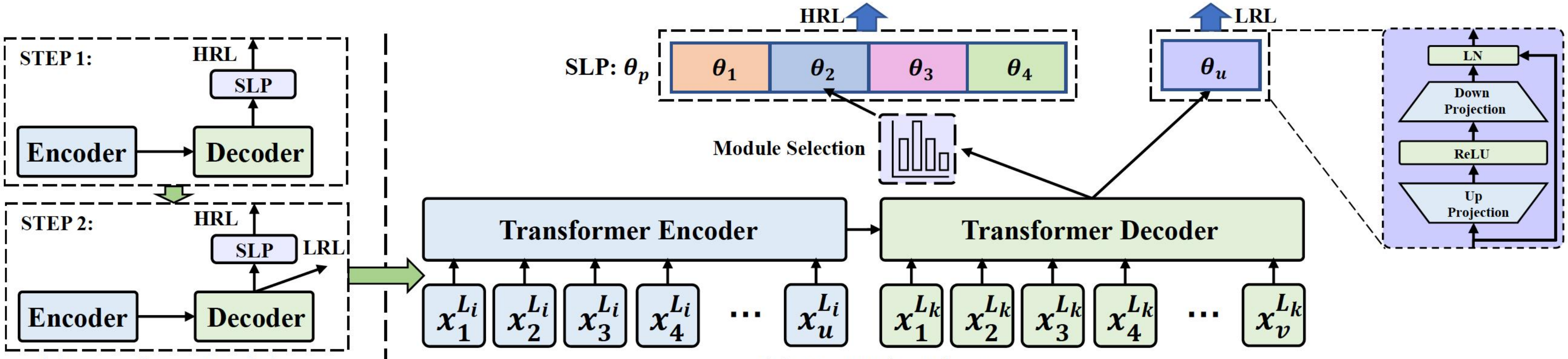
✓ The proposed **two-stage training strategy** and **selective language-specific pool** (**SLP**) mitigate the **negative inference** among different directions.



(a) Two-Stage Training (b) Model Architecture

Conclusion

- ✓ In this work, we propose **a novel multilingual translation model** with the high-resource language-specific training called **HLT-MT**.
- ✓ The proposed **two-stage training strategy** and **selective language-specific pool** (**SLP**) mitigate the **negative inference** among different directions.
- ✓ Experimental results evaluated on **WMT-10** and **OPUS-100** benchmarks demonstrate that HLT-MT **significantly outperforms all previous baselines**.



(a) Two-Stage Training (b) Model Architecture

Conclusion

- ✓ In this work, we propose **a novel multilingual translation model** with the high-resource language-specific training called **HLT-MT**.
- ✓ The proposed **two-stage training strategy** and **selective language-specific pool** (**SLP**) mitigate the **negative inference** among different directions.
- ✓ Experimental results evaluated on **WMT-10** and **OPUS-100** benchmarks demonstrate that HLT-MT **significantly outperforms all previous baselines**.

- ✓ Our **code** has been released
  - ➢ https://github.com/YuweiYin/HLT-MT

Conclusion

IJCAI ECAI VIENNA 22

- ✓ In this work, we propose **a novel multilingual translation model** with the high-resource language-specific training called **HLT-MT**.
- ✓ The proposed **two-stage training strategy** and **selective language-specific pool** (**SLP**) mitigate the **negative inference** among different directions.
- ✓ Experimental results evaluated on **WMT-10** and **OPUS-100** benchmarks demonstrate that HLT-MT **significantly outperforms all previous baselines**.

- ✓ Our **code** has been released
  - ➢ https://github.com/YuweiYin/HLT-MT

Thanks!