UM4: Unified Multilingual Multiple Teacher-Student Model for Zero-Resource Neural Machine Translation

Jian Yang[1*], Yuwei Yin[2*], Shuming Ma[2], Dongdong Zhang[2], Shuangzhi Wu[3], Hongcheng Guo[2], Zhoujun Li[1], Furu Wei[2]

[1]State Key Lab of Software Development Environment,

Beihang University, Beijing, China

[2]Microsoft Research Asia

[3]Tencent Cloud Xiaowei

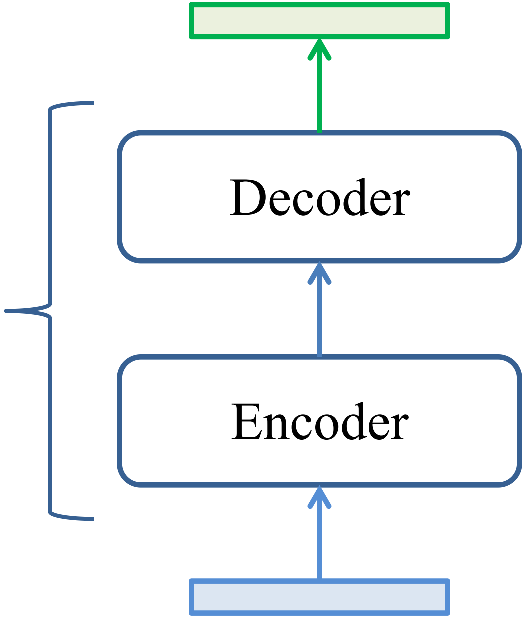
IJCAI ECAI VIENNA 22

Outline

Neural Machine Translation (NMT)

Translate source language X to target language Y if we have

sentence $F(X_k)$ of target language Y



sentence X_k of source language X

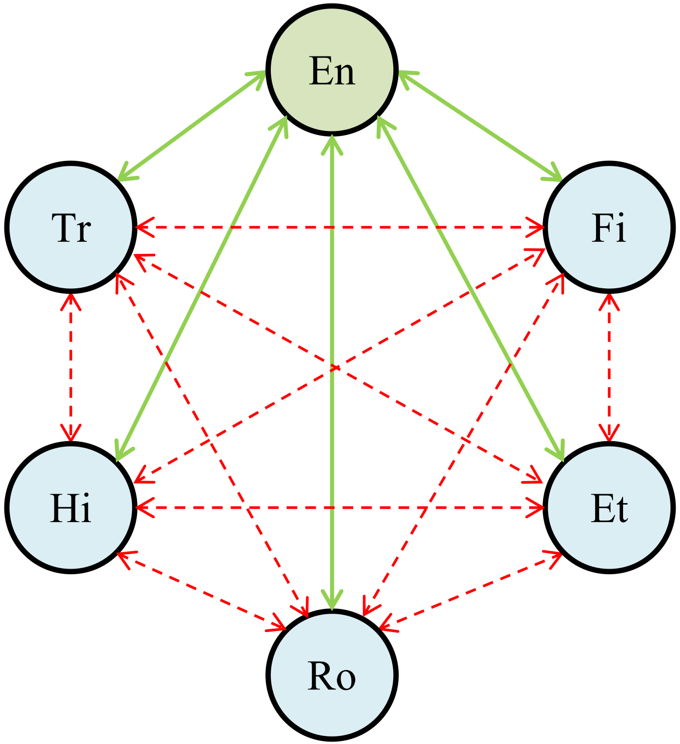
NMT model F
Enc-Dec framework

e.g., Transformer

Objective

minimize the difference between the model output $F(X_k)$ and the reference translation Y_k

Background: zero-resource translation



English(En)-centric parallel corpora

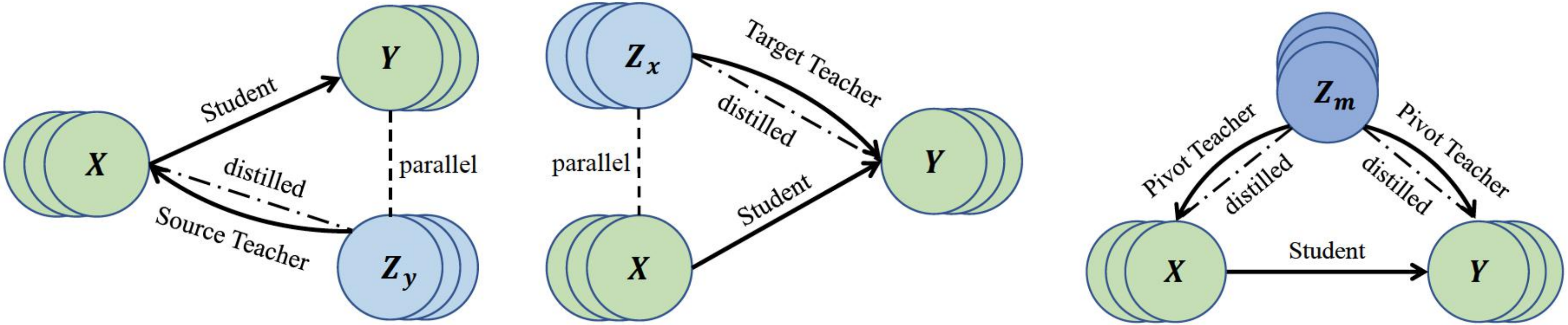
6 languages, $6 \times (6 - 1) = 30$ directions

√ 5 En↔X parallel corpora

× 25 parallel corpora of all other directions

Outline

Method: UM4

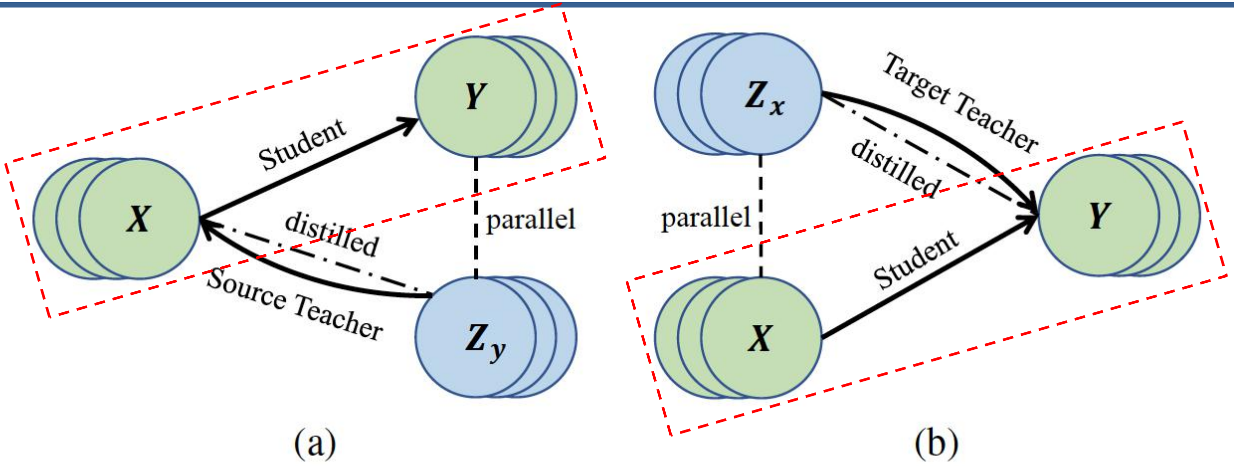


(a) Source Teacher

(b) Target Teacher

(c) Pivot Teacher

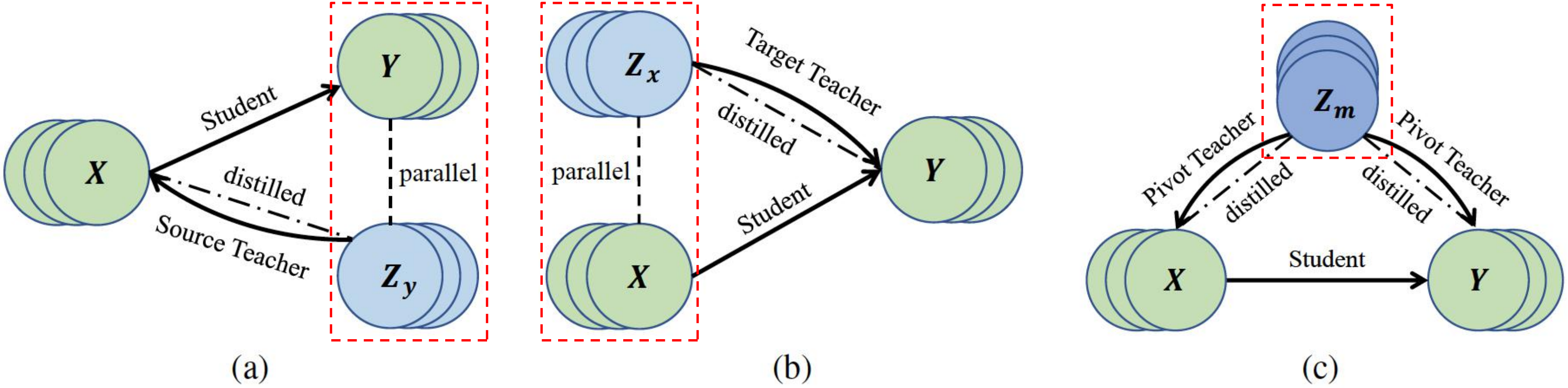
Method: UM4



Source Teacher Target Teacher Pivot Teacher

Goal: tranlate X→Y. (problem: lack of X-Y parallel data)

Method: UM4

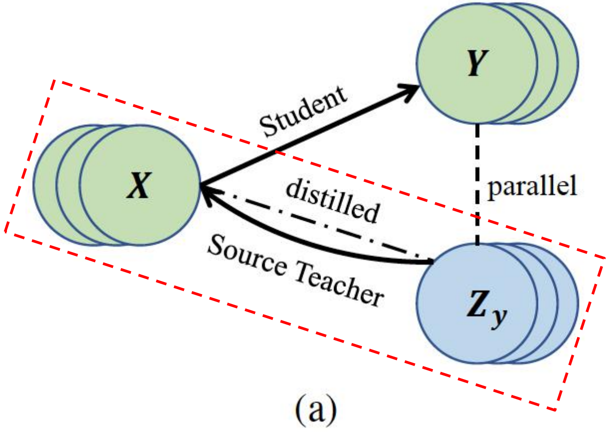


Source Teacher Target Teacher Pivot Teacher

Goal: tranlate X→Y. (problem: lack of X-Y parallel data)

Available parallel data: Y-Z_y & X-Z_x ; Available monolingual data: Z_m

Method: UM4 - Source Teacher

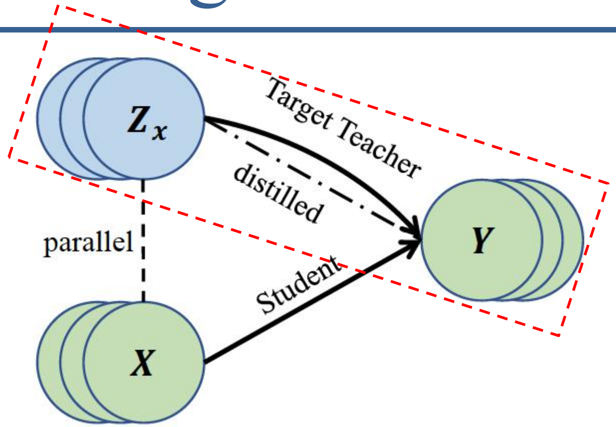


Source Teacher

Goal: tranlate X→Y.   (problem: lack of X-Y parallel data)
Available parallel data: Y-Z_y & X-Z_x ; Available monolingual data: Z_m

Source Teacher: translate Z_y→pseudo X. (pseudo X, real Y) pair

Method: UM4 - Target Teacher



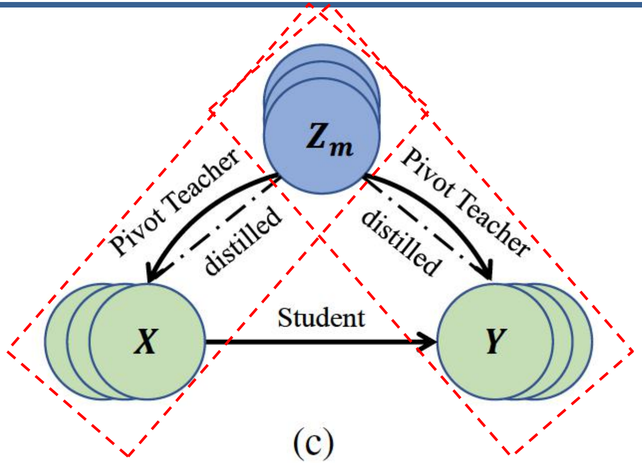
(b)

Target Teacher

Goal: tranlate X→Y. (problem: lack of X-Y parallel data)
Available parallel data: Y-Z_y & X-Z_x ; Available monolingual data: Z_m

Target Teacher: translate Z_x→pseudo Y. (real X, pseudo Y) pair

Method: UM4 - Pivot Teacher

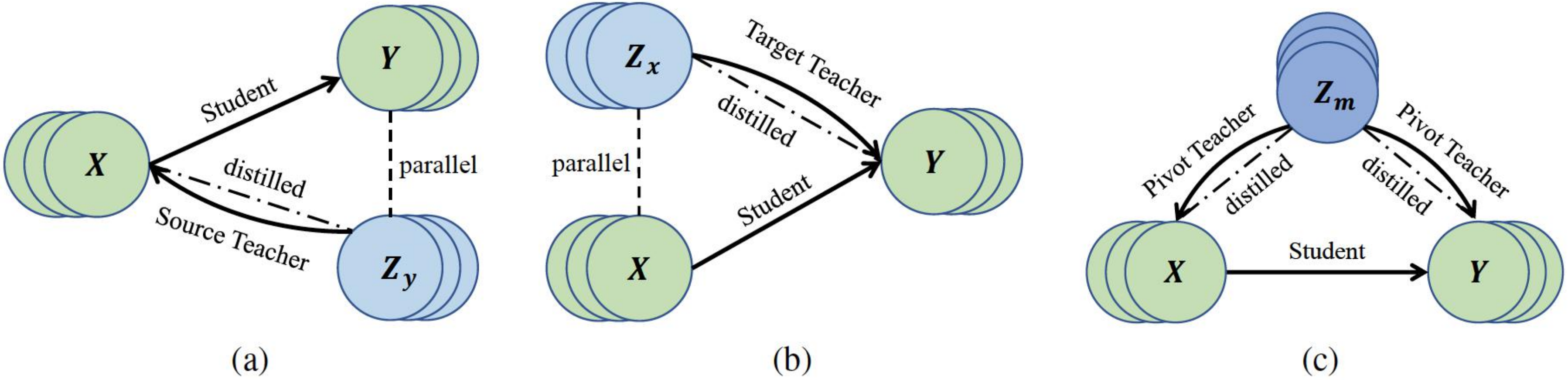


Pivot Teacher

Goal: tranlate X→Y. (problem: lack of X-Y parallel data)
Available parallel data: Y-Z_y & X-Z_x ; Available monolingual data: Z_m

Pivot Teacher: translate Z_m→pseudo X/Y. (pseudo X, pseudo Y) pair

Our Approach: UM4



Source Teacher

Target Teacher

Pivot Teacher

$$\mathcal{L}_S^{src} = -\mathbb{E}_{y,z_y \in D_{B_m}} \left[P(x|z_y;\theta_{z_y \to x}) \log P_\theta(y|x) \right]$$

$$\mathcal{L}_S^{tgt} = -\mathbb{E}_{x,z_x \in D_{B_i}} \left[P(y|z_x;\theta_{z_x \to x}) \log P_\theta(y|x) \right]$$

$$\mathcal{L}_S^{pivot} = -\mathbb{E}_{z_m \in D_M} \left[w_{x,y} \log P_\theta(y|x) \right]$$

$$\boxed{\mathcal{L}_S = \mathcal{L}_S^{src} + \mathcal{L}_S^{tgt} + \mathcal{L}_S^{pivot}}$$

Outline

Dataset

Bitext: En-Fr, En-Cs, En-De, En-Fi, En-Et, En-Ro, En-Hi, En-Tr

Training set: WMT benchmark

Valid and Test sets: TED Talks

	Language	#Bitext of Training	Training
Fr	French	10.0M	WMT15
Cs	Czech	10.0M	WMT19
De	German	4.6M	WMT19
Fi	Finnish	4.8M	WMT19
Et	Estonian	0.7M	WMT18
Ro	Romanian	0.5M	WMT16
Hi	Hindi	0.26M	WMT14
Tr	Turkish	0.18M	WMT18

Monolingual Data (English):

randomly sampled 1 million English sentences form NewsCrawl dataset

Baseline: Only Bitext Data

1. Pivot method
   1) Pivot method (Bilingual)
   2) Pivot method (Multilingual)
2. Direct Multilingual Method
   1) basic multilingual model
   2) MTL (multitask learning) multilingual model
3. Monolingual Adapter Method
4. Teacher-Student Method

UM4 (our method): Source Teacher + Target Teacher

Baseline: with extra Monolingual En

1. Pivot method + BT (Back Translation)
   1) Pivot method (Bilingual) + BT
   2) Pivot method (Multilingual) + BT
2. Direct Multilingual Method + BT
   1) basic multilingual model + BT
   2) MTL (multitask learning) multilingual model + BT
3. Monolingual Adapter Method + BT
4. Teacher-Student Method + BT

UM4 (our method): Source Teacher + Target Teacher + Pivot Teacher

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

Beam Search: beam size = 5; length penalty = 1.0

Model Ensemble: the parameters of last 5 checkpoints are averaged

Implementation

Architecture of all experiments: **Transformer big**

encoder & decoder: 6 layers with 16 heads per layer
word embedding size: 1024
FFN (feed-forward network) size: 4096
learning rate: 3e-4
warmup steps: 4000
optimizer: Adam
mini-batch size: 4096 tokens
loss: label smoothing cross-entropy (smoothing ratio = 0.1)

training device: 64 Tesla V100 GPUs

Outline

Test Data Size

$9 \times (9 - 1) = 72$ all directions, including 16 original parallel pairs En→X & X→En
$8 \times (8 - 1) = 56$ **zero-resource** translation directions

	En	Fr	Cs	De	Fi	Et	Ro	Hi	Tr
En	-	10.2K	7.8K	9.6K	2.7K	2.0K	9.4K	2.2K	9.7K
Fr	10.2K	-	7.7K	9.0K	2.5K	2.0K	8.8K	2.1K	8.8K
Cs	7.8K	7.7K	-	7.0K	2.0K	1.2K	7.0K	1.7K	7.0K
De	9.6K	9.0K	7.0K	-	2.6K	1.9K	8.6K	2.0K	8.7K
Fi	2.7K	2.5K	2.0K	2.6K	-	0.7K	2.4K	0.7K	2.3K
Et	2.0K	2.0K	1.2K	1.9K	0.7K	-	1.9K	0.4K	1.7K
Ro	9.4K	8.8K	7.0K	8.6K	2.4K	1.9K	-	2.0K	8.4K
Hi	2.2K	2.1K	1.7K	2.0K	0.7K	0.4K	2.0K	-	1.9K
Tr	9.7K	8.8K	7.0K	8.7K	2.3K	1.7K	8.4K	1.9K	-

Experimental Results

Directions: X→Y (richness of language X > language Y)
Data: only bitext data

consistent improvement

X (High) → Y (Low)	Fr→Fi	Cs→Fi	Cs→Ro	Cs→Hi	De→Et	Fi→Et	Fi→Ro	Fi→Tr	Avg_8	$Avg_{28}^{\gtrless}$
Train on Parallel Data (Bitext).										
Bilingual Pivot [Cheng *et al.*, 2017]	13.5	13.4	15.2	2.6	13.4	12.7	13.1	3.2	10.9	9.5
Multilingual Pivot [Lakew *et al.*, 2019]	12.5	11.9	16.1	6.9	14.8	13.3	14.0	5.3	11.9	11.2
Multilingual [Johnson *et al.*, 2017]	3.8	10.2	12.6	5.1	12.5	12.0	10.7	4.0	8.9	8.1
Teacher-Student [Chen *et al.*, 2017]	13.0	13.6	16.4	7.1	15.6	14.6	14.6	5.0	12.5	10.9
Monolingual Adapter [Philip *et al.*, 2020]	8.2	10.7	14.3	5.9	12.1	12.6	12.4	4.8	10.1	9.2
MTL [Wang *et al.*, 2020]	6.0	9.0	13.0	6.0	14.3	12.0	11.7	4.6	9.6	8.9
UM4 w/o pivot-teacher model (our method)	**13.8**	**13.9**	**16.8**	**7.3**	**16.3**	**14.9**	**15.1**	**5.4**	**12.9**	**11.8**
Train on Parallel and Monolingual Data (Bitext + MonoData).										
Bilingual Pivot + BT [Cheng *et al.*, 2017]	13.9	13.4	16.3	6.9	15.3	13.7	13.6	4.8	12.2	11.0
Multilingual Pivot + BT [Lakew *et al.*, 2019]	13.5	12.6	16.0	6.7	14.8	13.3	14.0	5.6	12.1	11.2
Multilingual + BT [Johnson *et al.*, 2017]	7.5	10.2	14.4	5.7	12.5	12.9	10.7	5.3	9.9	9.4
Teacher-Student + BT [Chen *et al.*, 2017]	13.6	13.0	16.6	6.8	15.2	14.8	15.2	5.5	12.6	11.6
Monolingual Adapter + BT [Philip *et al.*, 2020]	10.8	7.6	15.1	5.0	15.4	14.1	14.1	5.4	10.9	10.0
MTL + BT [Wang *et al.*, 2020]	10.6	9.0	13.5	5.4	12.7	12.8	12.8	5.2	10.3	8.0
UM4 (our method)	**14.1**	**14.1**	**17.1**	**7.4**	**16.2**	**15.0**	**15.8**	**5.9**	**13.2**	**12.4**

Experimental Results

Directions: X→Y (richness of language X > language Y)
Data: bitext data + monolingual English data

consistent improvement

X (High) → Y (Low)	Fr→Fi	Cs→Fi	Cs→Ro	Cs→Hi	De→Et	Fi→Et	Fi→Ro	Fi→Tr	Avg_8	$\text{Avg}^{\gtrsim}_{28}$
Train on Parallel Data (Bitext).										
Bilingual Pivot [Cheng *et al.*, 2017]	13.5	13.4	15.2	2.6	13.4	12.7	13.1	3.2	10.9	9.5
Multilingual Pivot [Lakew *et al.*, 2019]	12.5	11.9	16.1	6.9	14.8	13.3	14.0	5.3	11.9	11.2
Multilingual [Johnson *et al.*, 2017]	3.8	10.2	12.6	5.1	12.5	12.0	10.7	4.0	8.9	8.1
Teacher-Student [Chen *et al.*, 2017]	13.0	13.6	16.4	7.1	15.6	14.6	14.6	5.0	12.5	10.9
Monolingual Adapter [Philip *et al.*, 2020]	8.2	10.7	14.3	5.9	12.1	12.6	12.4	4.8	10.1	9.2
MTL [Wang *et al.*, 2020]	6.0	9.0	13.0	6.0	14.3	12.0	11.7	4.6	9.6	8.9
UM4 w/o pivot-teacher model (our method)	**13.8**	**13.9**	**16.8**	**7.3**	**16.3**	**14.9**	**15.1**	**5.4**	**12.9**	**11.8**
Train on Parallel and Monolingual Data (Bitext + MonoData).										
Bilingual Pivot + BT [Cheng *et al.*, 2017]	13.9	13.4	16.3	6.9	15.3	13.7	13.6	4.8	12.2	11.0
Multilingual Pivot + BT [Lakew *et al.*, 2019]	13.5	12.6	16.0	6.7	14.8	13.3	14.0	5.6	12.1	11.2
Multilingual + BT [Johnson *et al.*, 2017]	7.5	10.2	14.4	5.7	12.5	12.9	10.7	5.3	9.9	9.4
Teacher-Student + BT [Chen *et al.*, 2017]	13.6	13.0	16.6	6.8	15.2	14.8	15.2	5.5	12.6	11.6
Monolingual Adapter + BT [Philip *et al.*, 2020]	10.8	7.6	15.1	5.0	15.4	14.1	14.1	5.4	10.9	10.0
MTL + BT [Wang *et al.*, 2020]	10.6	9.0	13.5	5.4	12.7	12.8	12.8	5.2	10.3	8.0
UM4 (our method)	**14.1**	**14.1**	**17.1**	**7.4**	**16.2**	**15.0**	**15.8**	**5.9**	**13.2**	**12.4**

Experimental Results

Directions: X→Y (richness of language X < language Y)
Data: bitext data w/ or w/o monolingual English data

consistent improvement

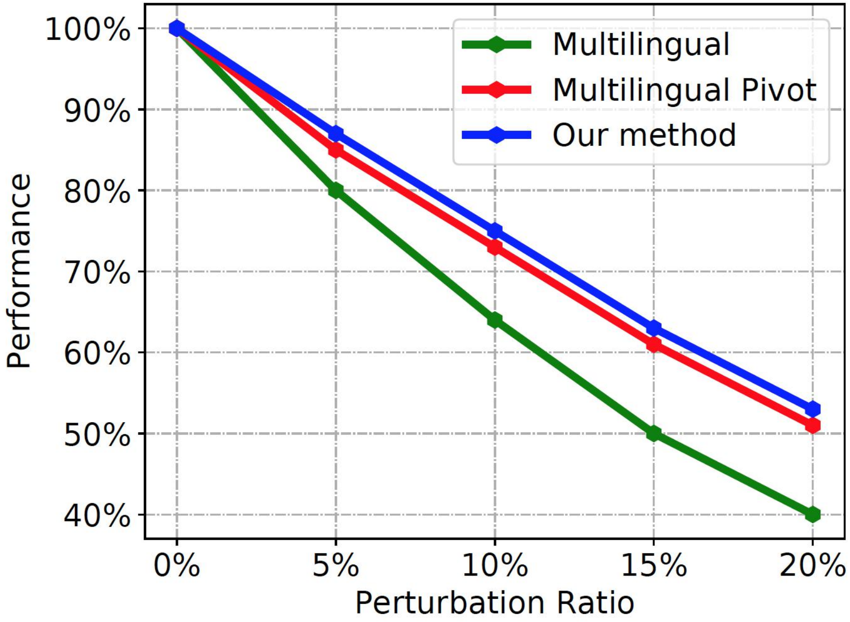
X (Low) → Y (High)	Fi→De	Et→De	Et→Fi	Ro→Cs	Ro→De	Ro→Et	Tr→Fr	Tr→Et	Avg_8	$\text{Avg}_{28}^{<}$
Train on Parallel Data (Bitext).										
Bilingual Pivot [Cheng *et al.*, 2017]	15.5	15.3	11.0	14.6	16.8	11.8	10.0	5.8	12.6	11.1
Multilingual Pivot [Lakew *et al.*, 2019]	14.6	16.3	12.9	15.1	18.2	14.0	15.7	9.9	14.6	13.6
Multilingual [Johnson *et al.*, 2017]	11.4	12.5	10.1	12.1	15.6	10.7	7.2	5.2	10.6	9.2
Teacher-Student [Chen *et al.*, 2017]	16.0	17.9	14.1	16.0	19.1	15.1	16.4	11.0	15.7	13.6
Monolingual Adapter [Philip *et al.*, 2020]	11.8	14.7	11.5	13.1	16.4	12.2	11.7	7.8	12.4	10.4
MTL [Wang *et al.*, 2020]	11.7	15.1	10.1	13.0	16.1	12.5	10.4	7.0	12.0	10.4
UM4 w/o pivot-teacher model (our method)	**16.6**	**18.5**	**14.2**	**16.3**	**19.9**	**15.4**	**17.1**	**11.3**	**16.2**	**14.7**
Train on Parallel and Monolingual Data (Bitext + MonoData).										
Bilingual Pivot + BT [Cheng *et al.*, 2017]	15.0	17.0	12.3	16.0	18.6	13.9	14.6	9.0	14.6	13.8
Multilingual Pivot + BT [Lakew *et al.*, 2019]	16.2	17.4	12.8	15.8	19.4	14.2	16.7	10.4	15.4	14.1
Multilingual + BT [Johnson *et al.*, 2017]	13.6	16.3	12.3	14.9	16.1	12.7	12.1	8.6	13.3	11.3
Teacher-Student + BT [Chen *et al.*, 2017]	16.6	19.0	13.8	16.5	20.0	15.0	16.8	10.9	16.1	14.3
Monolingual Adapter + BT [Philip *et al.*, 2020]	13.8	13.8	11.6	15.6	11.7	13.7	13.4	9.6	12.9	10.8
MTL + BT [Wang *et al.*, 2020]	12.8	16.6	11.5	13.9	17.0	13.0	14.2	8.7	13.5	11.7
UM4 (our method)	**17.6**	**19.6**	**14.3**	**17.2**	**20.7**	**15.6**	**17.5**	**11.5**	**16.8**	**15.1**

Ablation Study

Source	Target	Mono	Fr→De	De→Ro	Et→Ro	Avg_{56}
✓			21.3	17.0	14.5	12.3
	✓		21.4	16.2	15.2	13.0
		✓	22.5	17.2	15.4	12.7
	✓	✓	22.4	17.5	15.8	13.4
✓		✓	22.3	16.5	14.6	12.6
✓	✓		21.7	17.5	15.6	13.3
✓	✓	✓	**22.8**	**17.7**	**16.4**	**13.7**

- Every single Teacher model is beneficial.
- The more Teacher models, the better.

Robustness against Input Errors



Corrupt the input sentences by

- deletion (drop words),
- masking (replace words with "[unk]"),
- swap (swap words), and
- substitution (replace words with random words in the vocab)

➢ Our multilingual student is more robust than baseline models.

Number of Training Language Pairs

#Pairs	Fr→De	Ro→De	Tr→Cs	Avg_{16}	Avg_{56}
Supervised	11.7	16.1	9.6	22.8	8.7
Zero-resource	22.0	19.8	11.5	-	13.0
Both	**22.8**	**20.7**	**12.3**	**23.1**	**13.7**

Training data choices:

- Supervised: 8 original English-centric parallel coropora (16 directions)
- Zero-resource: $8 \times (8 - 1) = 56$ pairs distilled by UM4 teachers
- Both: $16 + 56 = 72$ all pairs

Number of Training Language Pairs

#Pairs	Fr→De	Ro→De	Tr→Cs	Avg_{16}	Avg_{56}
Supervised	11.7	16.1	9.6	22.8	8.7
Zero-resource	22.0	19.8	11.5	-	13.0
Both	**22.8**	**20.7**	**12.3**	**23.1**	**13.7**

Training data choices:

- Supervised: $8 \times 2 = 16$ original English-centric parallel coropora
- Zero-resource: $8 \times (8 - 1) = 56$ pairs distilled by UM4 teachers
- Both: $16 + 56 = 72$ all pairs

Test on 16 En→X & X→En dirctions:

➢ Our UM4 method (23.1) further enhance the supervised English-centric translation directions (+ 0.3)

Number of Training Language Pairs

#Pairs	Fr→De	Ro→De	Tr→Cs	Avg_{16}	Avg_{56}
Supervised	11.7	16.1	9.6	22.8	8.7
Zero-resource	22.0	19.8	11.5	-	13.0
Both	**22.8**	**20.7**	**12.3**	**23.1**	**13.7**

Training data choices:

- Supervised: 8 × 2 = 16 original English-centric parallel coropora
- Zero-resource: 8 × (8 - 1) = 56 pairs distilled by UM4 teachers
- Both: 16 + 56 = 72 all pairs

Test on 56 zero-resource dirctions:

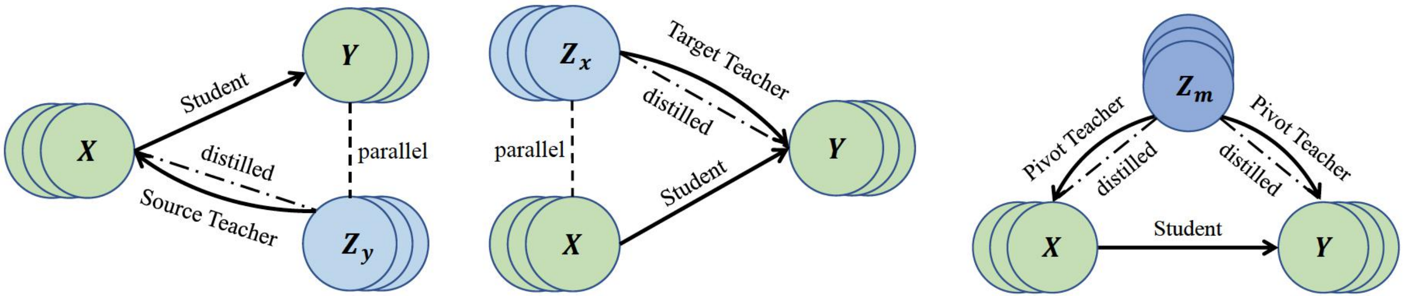
➢ Significant imporvement (+ 4.3) using the distilled data (13.0)

➢ Further enhancement (+0.7) with the original parallel data (13.7)

Outline

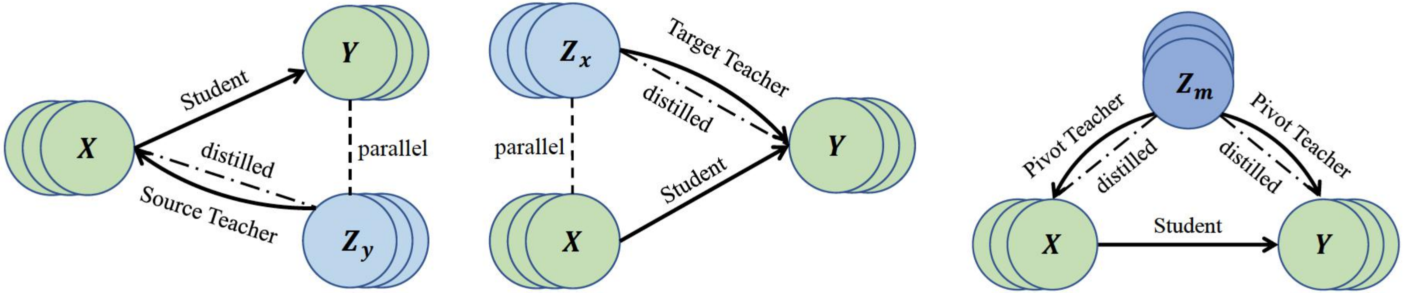
Conclusion

✓ In this work, we propose **a novel method** called Unified Multilingual Multiple teacher-student Model for NMT (**UM4**) to ameliorate the translation of **zero-resource directions**.



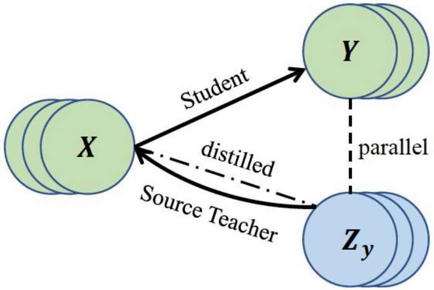
Conclusion

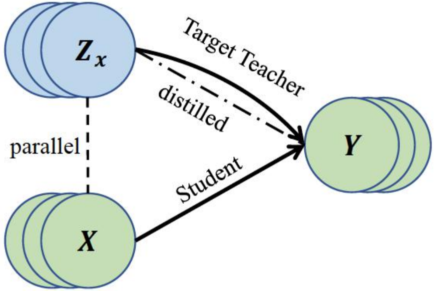
- ✓ In this work, we propose **a novel method** called Unified Multilingual Multiple teacher-student Model for NMT (**UM4**) to ameliorate the translation of **zero-resource directions**.
- ✓ Our method **unifies** the **source-teacher** model, **target-teacher** model, and **pivot-teacher** model to guide the multilingual source→target student model, alleviating the **error propagation** problem caused by **two-pass translation**.

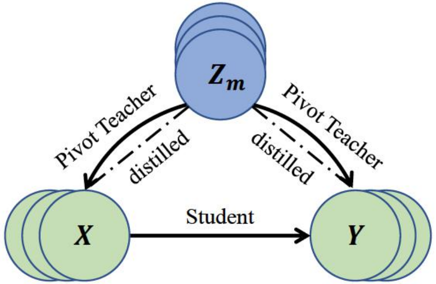


Conclusion

- ✓ In this work, we propose **a novel method** called Unified Multilingual Multiple teacher-student Model for NMT (**UM4**) to ameliorate the translation of **zero-resource directions**.
- ✓ Our method **unifies** the **source-teacher** model, **target-teacher** model, and **pivot-teacher** model to guide the multilingual source→target student model, alleviating the **error propagation** problem caused by **two-pass translation**.
- ✓ Experimental results on the **multilingual dataset** of the WMT benchmark corroborate the effectiveness of our method in **leveraging the distilled knowledge from the unified teachers**.







Conclusion

- ✓ In this work, we propose **a novel method** called Unified Multilingual Multiple teacher-student Model for NMT (**UM4**) to ameliorate the translation of **zero-resource directions**.
- ✓ Our method **unifies** the **source-teacher** model, **target-teacher** model, and **pivot-teacher** model to guide the multilingual source→target student model, alleviating the **error propagation** problem caused by **two-pass translation**.
- ✓ Experimental results on the **multilingual dataset** of the WMT benchmark corroborate the effectiveness of our method in **leveraging the distilled knowledge from the unified teachers**.

- ✓ Our **code** and **data** have been released
  - ➢ https://github.com/YuweiYin/UM4

Conclusion

- ✓ In this work, we propose **a novel method** called Unified Multilingual Multiple teacher-student Model for NMT (**UM4**) to ameliorate the translation of **zero-resource directions**.
- ✓ Our method **unifies** the **source-teacher** model, **target-teacher** model, and **pivot-teacher** model to guide the multilingual source→target student model, alleviating the **error propagation** problem caused by **two-pass translation**.
- ✓ Experimental results on the **multilingual dataset** of the WMT benchmark corroborate the effectiveness of our method in **leveraging the distilled knowledge from the unified teachers**.

- ✓ Our **code** and **data** have been released
  - ➢ https://github.com/YuweiYin/UM4

Thanks!