Mitigating Data Imbalance and Representation Degeneration in Multilingual Machine Translation

Wen Lai[1,2], Alexandra Chronopoulou[1,2] and Alexander Fraser[1,2]

[1] Center for Information and Language Processing, LMU Munich, Germany
[2] Munich Center for Machine Learning, Germany
{lavine, achron, fraser}@cis.lmu.de

Introduction

- We consider two major challenging in multilingual neural machine translation (MNMT):
  - **Data Imbalance**: The imbalance in the amount of parallel corpora for all language pairs, especially for long-tail languages (i.e., very low-resource languages).
  - **Representation Degeneration**: The problem of encoded tokens tending to appear only in a small subspace of the full space available to the MNMT model.
- As shown in Figure 1, the two challenging mentioned above pose a serious problem, i.e, the translation performance of different language pairs in the MNMT model varies significantly.
- We propose a novel approach **Bi-ACL** (bidirectional autoencoder and bidirectional contrastive learning) in a challenge scenario: only uses target-side monolingual data and a bilingual dictionary.

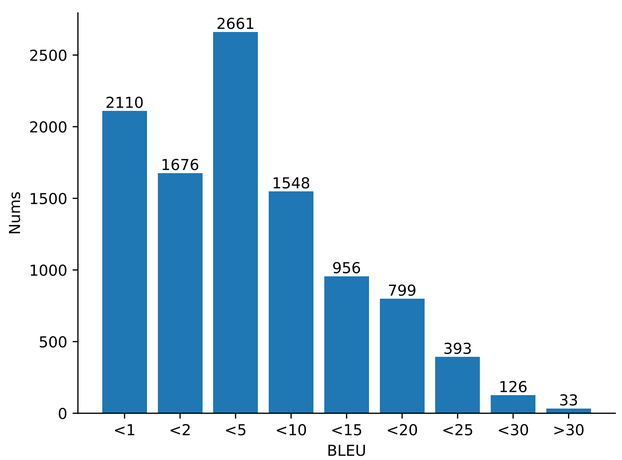

Figure 1: BLEU score statistics of the *m2m_100* model on Flores101 dataset for $102 \times 101 = 10302$ language pairs. Each bar denotes the number of language pairs in the interval of the BLEU score.

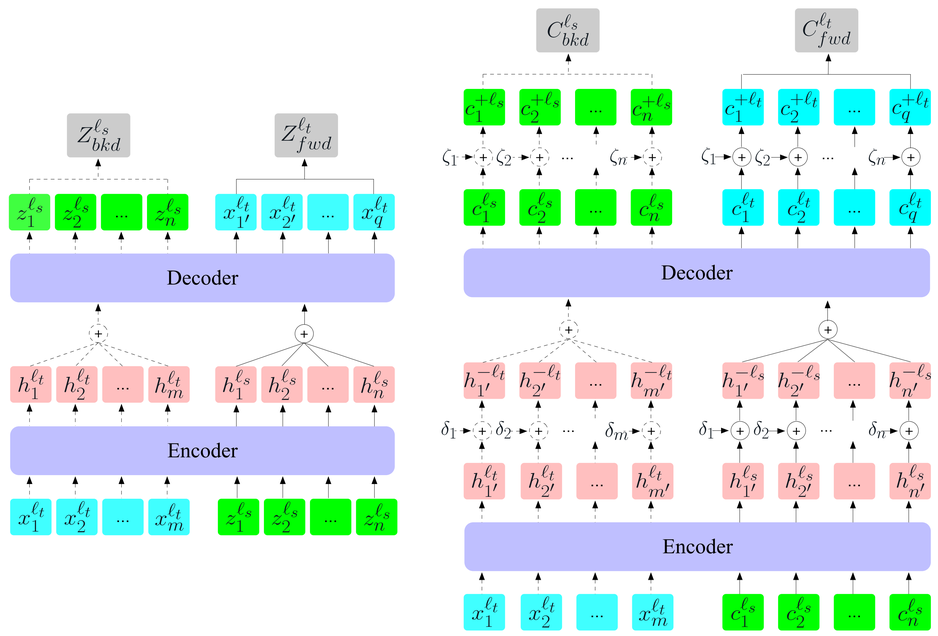

Figure 2: Model architecture: bidirectional autoencoder (left) and bidirectional contrastive learning (right)

Method

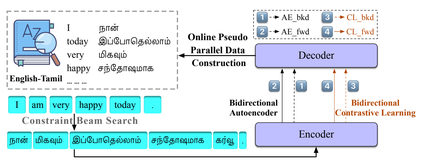

Figure 3: Method Overview.

Online Pseudo-Parallel Data Construction

- Construct a pseudo-parallel data with constrained beam search (i.e., constrained beam search) using target-side monolingual data and a bilingual dictionary.

Bidirectional Autoencoder and Bidirectional Contrastive Learning

- Bidirectional autoencoder (BAE) designed for mitigating data imbalance problem. Loss functions for forward autoencoder and backward autoencoder are denoted as $\mathcal{L}_{AE\_fwd}$ and $\mathcal{L}_{AE\_bkd}$ separately.
- Bidirectional contrastive learning (BCL) designed for mitigating representation degeneration problem. Loss functions for forward contrastive learning and backward contrastive learning are denoted as $\mathcal{L}_{CL\_fwd}$ and $\mathcal{L}_{CL\_bkd}$ separately.

Results

Models	en→ta	en→kk	ar→ta	ca→ta	ga→bs	kk→ko	ka→ar	ta→tr	af→ta	hi→kk
	Bilingual Setting									
m2m	2.12	0.26	0.34	1.75	0.51	0.85	2.14	1.41	1.46	0.84
pivot_en	-	-	0.30	0.74	0.00	0.27	0.15	1.38	1.00	0.22
BT	0.76	0.67	0.60	1.13	0.63	0.97	0.06	2.05†	0.72	0.43
wbw_lm	2.76†	0.36	0.87	0.68	0.36	0.07	2.86†	2.26†	1.47	0.04
syn_lexicon	1.33	0.14	0.72	2.07†	0.93	1.10†	0.85	0.57	2.07†	0.89
Bi-ACL w/o Curriculum	4.57‡	1.35‡	1.76‡	3.14‡	1.81‡	3.07‡	3.92‡	4.18‡	3.15‡	1.53‡
Bi-ACL (*ours*)	**5.14**‡	**2.59**‡	**2.32**‡	**3.50**‡	**2.37**‡	**3.61**‡	**4.76**‡	**4.97**‡	**3.68**‡	**2.47**‡
Δ	+3.02	+2.33	+1.98	+1.75	+1.86	+3.03	+2.62	+3.56	+2.22	+1.63
	Multilingual Setting									
	ta	hy	ka	be	kk	az	mn	gu	my	ga
m2m	1.46	1.69	0.52	1.95	0.67	2.32	1.12	0.26	0.24	0.09
Bi-ACL	**2.54**‡	**3.17**‡	**2.38**‡	**3.12**‡	**1.44**‡	**3.28**‡	**1.95**‡	**1.18**‡	**1.94**‡	**1.37**‡
Δ	+1.08	+1.48	+1.86	+1.17	+0.77	+0.96	+0.83	+0.92	+1.70	+1.28
	Multilingual Setting (specific language pair)									
	en→ta*	ar→ta*	ca→ta*	af→ta*	el→ta	en→kk*	hi→kk*	fa→kk	jv→kk	ml→kk
m2m	2.12	0.34	1.75	1.46	1.21	0.26	0.84	0.54	1.77	0.69
Bi-ACL	**5.37**‡	**2.81**‡	**3.82**‡	**4.16**‡	**3.24**‡	**2.94**‡	**2.91**‡	**2.87**‡	**3.73**‡	**3.29**‡
Δ	+3.25	+2.47	+2.07	+2.70	+2.03	+2.68	+2.07	+2.33	+1.96	+2.60
Φ	+0.23	+0.49	+0.32	+0.48	-	+0.35	+0.44	-	-	-

Table 1: **Main Results**: BLEU scores for low-resource language pairs in the bilingual, multilingual setting.

	ar→ta				ta→tr				de→fr			
	Encoder		Decoder		Encoder		Decoder		Encoder		Decoder	
	$I_1\uparrow$	$I_2\downarrow$	$I_1\uparrow$	$I_2\downarrow$	$I_1\uparrow$	$I_2\downarrow$	$I_1\uparrow$	$I_2\downarrow$	$I_1\uparrow$	$I_2\downarrow$	$I_1\uparrow$	$I_2\downarrow$
m2m	0.042	20.017	0.012	26.639	0.036	20.408	0.006	26.901	0.058	16.521	0.016	24.695
pivot_en	0.034	22.852	0.008	24.472	0.019	22.889	0.007	25.977	0.056	16.843	0.016	24.763
BT	0.011	25.825	0.007	25.797	0.028	22.009	0.009	27.492	0.074	14.774	0.015	24.878
wbw_lm	0.023	23.485	0.015	24.746	0.038	19.389	0.010	26.320	0.037	19.099	0.015	24.935
syn_lexicon	0.059	17.513	0.015	25.694	0.028	20.640	0.013	26.475	0.020	23.859	0.014	24.137
Bi-ACL w/o Curriculum	0.074	16.174	0.017	24.176	0.039	19.139	0.018	24.712	0.078	14.165	0.017	24.128
Bi-ACL (*ours*)	**0.086**	**15.714**	**0.020**	**23.251**	**0.043**	**18.672**	**0.021**	**22.716**	**0.086**	**13.666**	**0.017**	**24.067**

Table 2: **Main Results**: Isotropic embedding space analysis in ar→ta, ta→tr and en→de translation task.

Ablation Study

	$\mathcal{L}_{AE\_bkd}$	$\mathcal{L}_{AE\_fwd}$	$\mathcal{L}_{CL\_bkd}$	$\mathcal{L}_{CL\_fwd}$	en→ta			ta→tr			en→de		
					BLEU	$I_1\uparrow$	$I_2\downarrow$	BLEU	$I_1\uparrow$	$I_2\downarrow$	BLEU	$I_1\uparrow$	$I_2\downarrow$
#1	√	×	×	×	2.51	0.005	30.737	3.34	0.004	32.378	23.14	0.011	24.876
	×	√	×	×	**3.27**	0.006	29.299	**3.96**	0.006	29.373	**23.82**	0.011	24.651
	×	×	√	×	2.39	0.008	26.562	2.69	0.007	28.663	22.57	0.013	24.872
	×	×	×	√	2.36	**0.009**	**26.541**	2.65	**0.007**	**27.155**	22.89	**0.013**	**24.367**
#2	√	√	×	×	**4.03**	0.009	27.147	**4.12**	0.011	27.039	**25.62**	0.014	24.075
	√	×	√	×	2.36	0.012	26.782	3.64	0.010	27.636	24.84	0.013	24.513
	√	×	×	√	2.50	0.014	26.007	3.37	0.012	26.881	24.36	0.012	24.841
	×	√	√	×	3.54	0.012	26.964	3.89	0.011	27.175	24.59	0.012	24.764
	×	√	×	√	3.81	**0.019**	**25.597**	4.03	**0.015**	**26.460**	25.17	**0.014**	**24.025**
	×	×	√	√	2.53	0.013	28.459	3.61	0.012	27.639	24.73	0.012	24.723
#3	√	√	√	×	3.85	0.020	24.732	3.73	0.016	26.197	25.43	0.014	24.137
	√	√	×	√	**4.31**	**0.028**	**23.861**	**4.29**	**0.019**	**25.573**	**26.44**	**0.015**	**24.019**
	√	×	√	√	2.82	0.023	24.352	3.77	0.018	25.852	25.63	0.014	24.257
	×	√	√	√	3.83	0.025	24.173	4.05	0.015	26.447	26.17	0.015	14.192
#4	√	√	√	√	**5.14**	**0.031**	**22.392**	**4.97**	**0.022**	**24.175**	**27.76**	**0.016**	**23.951**

Table 3: Ablation study of four loss functions on en→ta and ta→ar translation task.

Conclusion

- We propose **Bi-ACL**, which prove to be effective (performance and representation) both on long-tail languages and low-resource languages.
- Bi-ACL provides a paradigm that an inexpensive bilingual lexicon and monolingual data should be fully exploited when there are no bilingual parallel corpora, which we believe more researchers in the community should be aware of.

Paper

Code