Mitigating Data Imbalance and Representation Degeneration in Multilingual Machine Translation

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

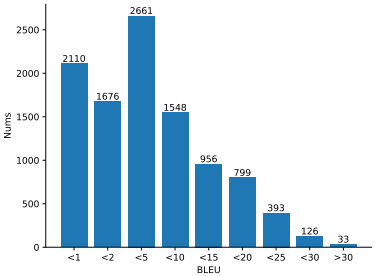
[1]Center for Information and Language Processing, LMU Munich, Germany
[2]Munich Center for Machine Learning, Germany

6th December, 2023

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Background

- BLEU score statistics of the m2m_100 model (Fan et al., 2021)[a] on Flores-101 dataset.
- Findings
  - 21% of the language pairs in the m2m_100 model have a BLEU score of less than 1 and more than 50% have a BLEU score of less than 5.
  - Only 13% have a BLEU score over 20.



[a] Beyond english-centric multilingual machine translation (Fan et al., JMLR 22.1 (2021): 4839-4886)

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Motivation

- We consider two major challenges 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.

[1] Unsupervised Machine Translation Using Monolingual Corpora Only (Lample et al., ICLR 2018)

[2] Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation (Wang et al., ACL 2022)

Motivation

- We consider two major challenges 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.
- Common Practice
  - Improve the performance of a machine translation model without using any parallel data to overcome the data imbalance problem.
    - Unsupervised Machine Translation (Lample et al., 2018)[1]
    - Using bilingual dictionary (Wang et al., 2022)[2]
  - Contrastive learning to overcome the representation degeneration problem.

[1] Unsupervised Machine Translation Using Monolingual Corpora Only (Lample et al., ICLR 2018)

[2] Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation (Wang et al., ACL 2022)

Challenges

- UNMT uses large-scale monolingual data from both source-side and target-side of the language pair.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Challenges

- UNMT uses large-scale monolingual data from both source-side and target-side of the language pair.
- Using bilingual dictionary without parallel data is not explored well in previous work.

Challenges

- UNMT uses large-scale monolingual data from both source-side and target-side of the language pair.
- Using bilingual dictionary without parallel data is not explored well in previous work.
- The naïve contrastive learning framework that utilizes random non-target sequences as negative examples is suboptimal, because they are easily distinguishable from the correct output.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Challenges

- UNMT uses large-scale monolingual data from both source-side and target-side of the language pair.
- Using bilingual dictionary without parallel data is not explored well in previous work.
- The naïve contrastive learning framework that utilizes random non-target sequences as negative examples is suboptimal, because they are easily distinguishable from the correct output.
- Current methods consider the two problems (data imbalance and representation degeneration) separately.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Research Question

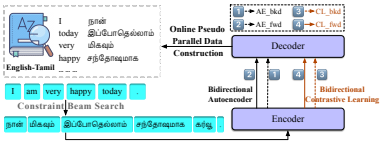
- Can we consider both of the problems at the same time?

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Overview

- We consider a challenging scenario: using readily available data sources, which contains target-side monolingual data and a bilingual dictionary.
- we propose a 3-step approach:
  - **Online Pseudo-Parallel Data Construction**
  - **Bidirectional Autoencoder**
  - **Bidirectional Contrastive learning**



[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Online Pseudo-Parallel Data Construction

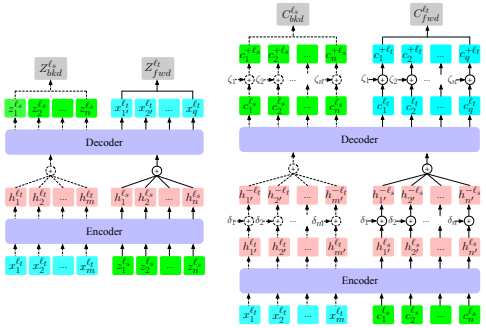
- Given target-side monolingual data, we use lexically constrained decoding (i.e., constrained beam search; Post and Vilar, 2018[3]) to generate the pseudo source language sentence in an online mode.

[3] Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation (Post et al., NAACL 2018)

[1] Center for Information and Language Processing, LMU Munich, Germany [2] Munich Center for Machine Learning, Germany

Bidirectional Autoencoder

- Different with UNMT, which uses both source and target side monolingual data. we uses only target-side monolingual data and the pseudo parallel data from previous step.
- **Backforward Autoencoder**
- **Forward Autoencoder**

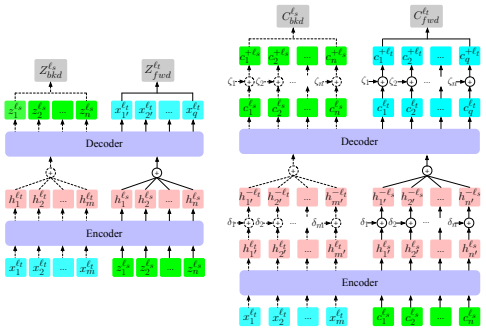


Please refer to Section 3.2 of the paper for more details.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Bidirectional Contrastive Learning

- Different with naive contrastive learning framework, which uses ground-truth target sentences as positive examples and random non-target sentences as negative examples.
- We automatically generate negative and positive examples by adding perturbations, such that both kinds of examples are difficult for the model to classify correctly.
- **Backward contrastive learning**
- **Forward contrastive learning**



Please refer to Section 3.3 and Appendix A of the paper for more details.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Curriculum Learning

- We compute the token coverage ratio in each sentence and use this score as a curriculum to determine the order of the sentences sampled during training.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Training Objective

- The model can be trained by minimizing a composite loss from four loss functions.

$$\begin{aligned}\mathcal{L}^* =& \lambda(\mathcal{L}_{AE\_bkd} + \mathcal{L}_{AE\_fwd}) + \\ & (1-\lambda)(\mathcal{L}_{CL\_bkd} + \mathcal{L}_{CL\_fwd})\end{aligned} \tag{1}$$

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Datasets

- Monolingual Data
  - News-Crawl[4]
  - CCAligned[5]
- Bilingual Dictionary
  - Extract from Wikitionary[6].
  - For pairs not involving English, we pivot through English.

[4] https://data.statmt.org/news-crawl
[5] https://opus.nlpl.eu/CCAligned.php
[6] https://www.wiktionary.org

Baselines

- **m2m**: Using the original m2m model to generate the translations.
- **pivot_en**: Using English as a pivot language.
- **BT**[7]: Back-Translate target-side monolingual data using m2m_100 model to generate the pseudo source-target parallel dataset.
- **wbw_lm**[8]: Use a bilingual dictionary, cross-lingual word embeddings and a target-side language model to translate word-by-word.
- **syn_lexicon**[9]: Replace the words in the target monolingual sentence with the corresponding source language words in a bilingual dictionary.

[7] Improving Neural Machine Translation Models with Monolingual Data (Sennrich et al., ACL 2016)
[8] Improving Unsupervised Word-by-Word Translation with Language Model and Denoising Autoencoder (Kim et al., EMNLP 2018)
[9] Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation (Wang et al., ACL 2022)

Evaluate Metrics

- **BLEU**: Using the original m2m model to generate the translations;
- **Isotropy**: I_1 and I_2 isotropy measures from Wang et al., (2019).[10] Larger I_1 and smaller I_2 indicate a more isotropic embedding space in the MNMT model;

[10] Improving neural language generation with spectrum control (Wang et al., ICLR 2019)

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Main Results (low-resource languages)

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	-	-	-

- Our approach performs well on low-resource languages in the bilingual setting, multilingual setting, and 10 randomly selected language pairs in the multilingual setting.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Main Results (high-resource languages)

Models	en→de	en→fr	en→cs	de→fr	de→cs	fr→cs
m2m	22.79	32.50	21.65	28.53	20.73	20.30
pivot_en	-	-	-	11.68	7.09	6.60
BT	24.08	27.71	21.52	19.45	17.41	17.00
wbw_lm	7.52	11.53	9.04	8.38	9.44	10.15
syn_lexicon	5.35	12.56	10.90	5.90	8.39	8.89
Bi-ACL w/o Curriculum	25.17	35.52	23.91	29.43	22.71	22.04
Bi-ACL (*ours*)	**27.76**	**37.84**	**25.89**	**30.66**	**23.80**	**23.56**
Δ	+4.97	+5.34	+4.24	+2.13	+3.07	+3.26

- Our approach performs well in high-resource languages.
- Curriculum learning takes full advantage of the original model in the high-resource setting, with stronger gains in performance than in the low-resource setting.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Main Results (Isotropy Analysis)

	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**

- The embedding space on the encoder side is more isotropic than on the decoder side.
- Compared to other baseline systems, we get a higher I_1 and lower I_2 score, which shows a more isotropic embedding space in our methods.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Ablation Study

	$\mathcal{L}_{AE\_bkd}$	$\mathcal{L}_{AE\_fwd}$	$\mathcal{L}_{CL\_bkd}$	$\mathcal{L}_{CL\_fwd}$	en→ta BLEU	en→ta $l_1\uparrow$	en→ta $l_2\downarrow$	ta→tr BLEU	ta→tr $l_1\uparrow$	ta→tr $l_2\downarrow$	en→de BLEU	en→de $l_1\uparrow$	en→de $l_2\downarrow$
#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**

- The bidirectional autoencoder losses play a more critical role than the bidirectional contrastive learning losses in terms of BLEU score.
- Using forward direction losses results in a better translation quality compared to backward direction losses.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Further Analysis

- Our approach show better domain transfer and language transfer ability than previous method (more details are shown in Appendix E.3 and E.4).

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Conclusion

- We present a framework named **Bi-ACL** which improves the performance of MNMT models using only target-side monolingual data and a bilingual dictionary;
- We employ a bidirectional autoencoder and bidirectional contrastive learning, which prove to be effective both on long-tail languages and high-resource languages;
- We also find that Bi-ACL shows language transfer and domain transfer ability in zero-shot scenarios;
- In addition, 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.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany

Thank You!

Email: lavine@cis.lmu.de
Homepage: `https://lavine-lmu.github.io`
Address: Oettingenstraße 67, 80538 Munich, Germany

Paper Code

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning, Germany