TAL-SJTU Low-Resource Translation System for the WMT22 Translation Task

Zhiwei He[1] Xing Wang[2] Rui Wang[1] Shuming Shi[2] Zhaopeng Tu[2]

[1]Shanghai Jiao Tong University [2]Tencent AI Lab

SUMMARY

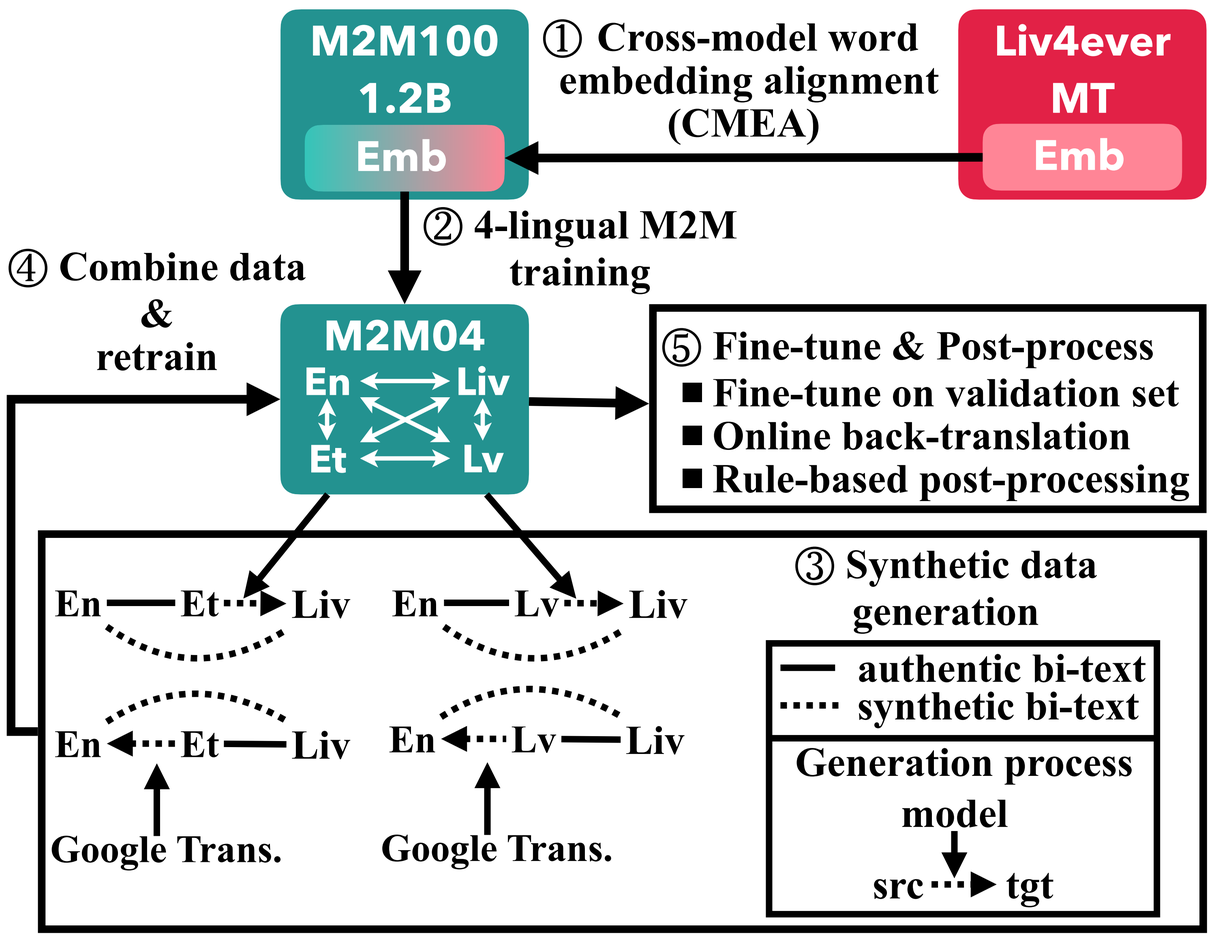
- We participate in the general translation task on low-resource **English⇔Livonian**.
- **Model**
  - **M2M100 1.2B** as a **multilingual** pre-trained model
  - Novel techniques that **adapt** it to the target language pairs
- **Evaluation**
  - We point out and correct the **inconsistent Unicode normalization** problem of Rikters et al. (2022).
  - We employ monolingual English and **round-trip** BLEU to evaluate the models, easing bi-text scarcity and achieving more accurate evaluation.

UNICODE INCONSISTENCY PROBLEM

- Rikters et al. (2022) has released the Liv4ever-MT model and liv4ever benchmark.
- We find that the liv4ever data have **different encodings** for the same character. For example, õ (00F5) and õ (006F 0303).
- After **normalization** and re-evaluation, the overall performance of Liv4ever-MT has been greatly **improved**.
- Check the **Github repo** to reproduce the full results

	En-Liv ⇒	En-Liv ⇐	Et-Liv ⇒	Et-Liv ⇐	Lv-Liv ⇒	Lv-Liv ⇐
Liv4ever-MT (Rikters et al.)	11.0	19.0	16.5	23.1	17.7	25.2
+ Norm. Ref.	**14.3**	**19.3**	**20.5**	**24.4**	**22.3**	**29.3**

APPROACH



- **Cross-model word embedding alignment (CMEA):** transfer the word embeddings of Liv4ever-MT to M2M100, enabling it to support Livonian
- **4-lingual M2M training:** many-to-many translation training for all language pairs in {En, Liv, Et, Lv}, using only parallel data
- **Synthetic data generation:** generate synthetic bi-text for En-Liv, using Et and Lv as pivot languages
- **Combine data and retrain:** combine all the authentic and synthetic bi-text and retrain the model
- **Fine-tune & post-process:** fine-tune the model on En⇔Liv using the validation set and perform online back-translation using monolingual data. Finally, apply rule-based post-processing to the model output.

EXPERIMENT

CMEA

	En-Liv ⇒	En-Liv ⇐	Et-Liv ⇒	Et-Liv ⇐	Lv-Liv ⇒	Lv-Liv ⇐
Liv4ever-MT Rikters et al.	14.3	19.3	20.5	24.4	22.3	29.3
M2M04 (T=5)	21.1	27.7	25.3	29.2	26.8	36.6
+ CMEA	**23.0**	**28.4**	**27.2**	**30.7**	**28.5**	**37.6**

- **Pre-trained model** can significantly improve translation performance.
- **CMEA** further improves the performance through transferring the pre-trained embeddings.

SYNTHETIC DATA

	Valid (multi-way) En⇒Liv	Valid (multi-way) Liv⇒En	Round-Trip (En-original)
M2M04 (T=5) +CMEA	23.0	28.4	23.4
Add synthetic data and retrain			
En-original	17.2	17.5	30.7
Liv-original	21.5	27.4	25.8
Both	17.0	19.3	32.7

- Different types (**En/Liv-original**) of synthetic data have different effects.
- Synthetic data improves round-trip BLEU but **degrades** the performance on parallel validation sets.

FINE-TUNE & POST-PROCESS

	Test Set En-Liv ⇒	Test Set En-Liv ⇐	Round-Trip BLEU
Before fine-tuning	15.8	29.4	32.7
+Fine-tuning	16.3	30.1	37.1
+Post-proc.	17.0	30.4	37.1

- The **best** generalization performance is obtained after fine-tuning and post-processing.