LAFT: Cross-lingual Transfer for Text Generation by Language-Agnostic Finetuning

Xianze Wu[1] Zaixiang Zheng[2] Hao Zhou[3] and Yong Yu[1]

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

[3]Insititute for AI Industry Research, Tsinghua University

Outline

- **Background & Motivation**
- Method
  - Language-agnostic Task Acquisition
  - Language Specialization for Generation
  - Learning
- Experiments
- Takeaways

Task: Zero-resource Cross-lingual Transfer

- [**Cross-lingual Transfer (CLT)**] Transfer knowledge learned from source language(s) to target language(s)
  - source language: rich-resource language in most case

- [**Zero-resource**] No human-annotated task data in target language is available for training

Existing Method: Fine-tuning MLPMs

- Fine-tuning MPLM on task-annotated data in source language
  - MPLM: **M**ulti-lingual **P**re-trained **L**anguage **M**odels (e.g. mBART)

- **How to further improve the fine-tuning method?**

Neural NLG Pipeline

- Understanding input text
  - e.g. convert a news article to hidden representations
- Manipulating semantics representation
  - e.g. filter out redundant content while keep the main idea
- Generating text result
  - e.g. generate abstractive summarization

Neural NLG Pipeline

- Understanding input text
  - e.g. convert a news article to hidden representations

- Manipulating semantic representation **(key step)**
  - e.g. filter out redundant content while keep the main idea
  - **Learn how to manipulate input representations according to downstream tasks.**

- Generating text result
  - e.g. generate abstractive summarization

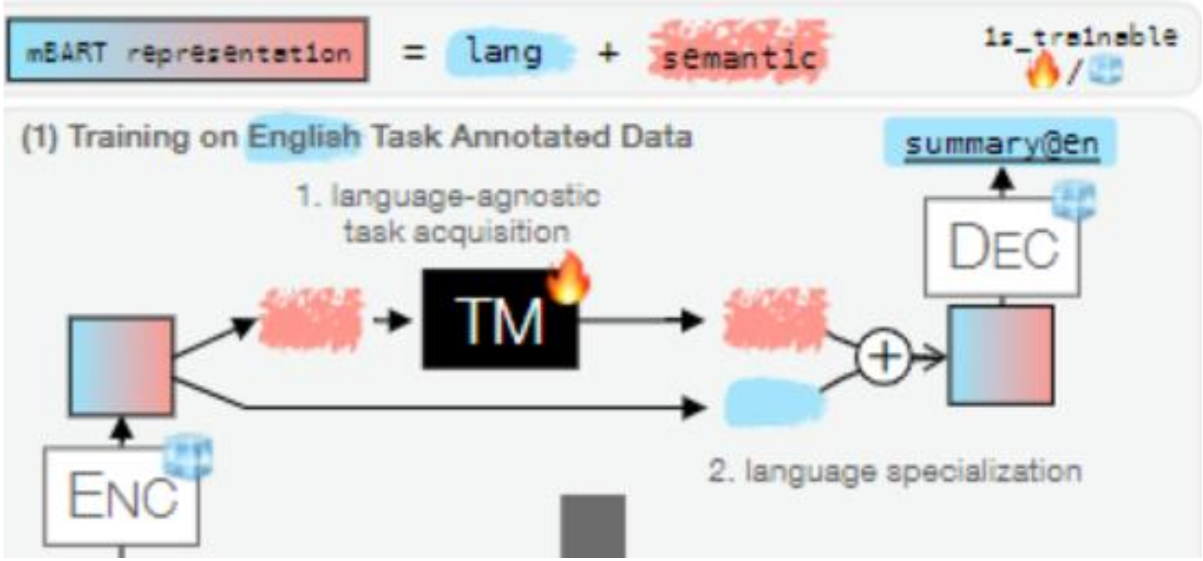
Problem of Fine-tuning MPLM

- Semantic and language component are highly entangled on MPLM's representation
- Knowledge of downstream tasks would be correlated to the source language
  - **Harm transfer ability!**
- Our approach: language-agnostic fine-tuning

Outline

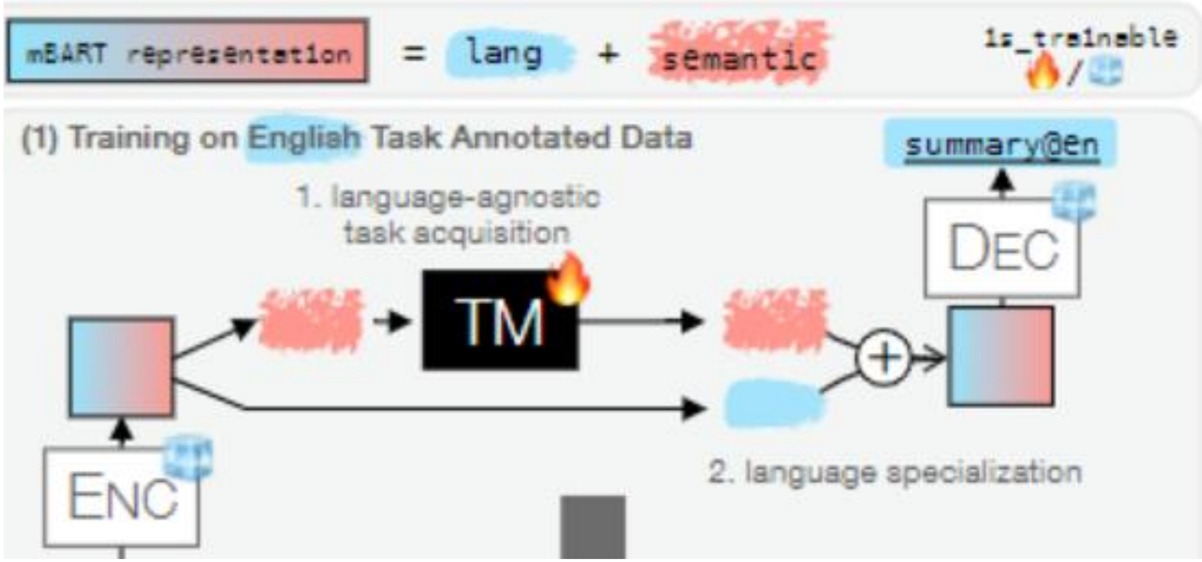
- Background & Motivation
- **Method**
  - **Language-agnostic Task Acquisition**
  - **Language Specialization for Generation**
  - **Learning**
- Experiments
- Takeaways

LAFT Model



- Task Module between encoder and decoder

LAFT Model



- Task Module between encoder and decoder
- **Language-agnostic task acquisition**
  - Only semantic representation is feed into TM
- **Language-specialization**
  - Add language information to the language-agnostic representation obtained by TM

Language-agnostic task acquisition

- How to remove language information?
  - For an MPLM, the representations from the same language L share vector space components, which corresponds to the language identity of language L. [Yang et al. 21]

[Yang et al. 21] Ziyi Yang, Yinfei Yang, Daniel Cer, Eric Darve:
A Simple and Effective Method To Eliminate the Self Language Bias in Multilingual Representations. EMNLP (1) 2021: 5825-5832

Language-agnostic task acquisition

- How to remove language information?
  - Estimation of language component [Yang et al. 21]
    - Construct a language matrix $M_L \in \mathbb{R}^{n \times d}$ by encoding monolingual texts from language L.
    - Perform SVD and extract the first k right singular vectors $c_L \in \mathbb{R}^{d \times k}$

[Yang et al. 21] Ziyi Yang, Yinfei Yang, Daniel Cer, Eric Darve:
A Simple and Effective Method To Eliminate the Self Language Bias in Multilingual Representations. EMNLP (1) 2021: 5825-5832

Language-agnostic task acquisition

- How to remove language information?
  - Estimation of language component [Yang et al. 21]
    - Construct a language matrix $M_L \in \mathbb{R}^{n \times d}$ by encoding monolingual texts from language L.
    - Perform SVD and extract the first k right singular vectors $c_L \in \mathbb{R}^{d \times k}$
  - Removal of language component
    - Subtract the projection of sentence representation onto language components from token representation.

$$r_L^i = e_L^i - c_L \frac{c_L^T e_L}{\|e_L\|_2}.$$

[Yang et al. 21] Ziyi Yang, Yinfei Yang, Daniel Cer, Eric Darve:
A Simple and Effective Method To Eliminate the Self Language Bias in Multilingual Representations. EMNLP (1) 2021: 5825-5832

Language Specialization for Generation

- Representations obtained by the task module is language-agnostic
  - Not enough for generating text
- Two solutions
  - Add language component back with a fusion mechanism

$$\mathbf{B}(h_L^i, c_L) = \mathbf{U}\left(\mathrm{ReLU}\left(\mathbf{D}([h_L^i, c_L])\right)\right) + h_L^i$$

  - Incorporate language adapter to each decoder layer (Pfeiffer et al., 2020)

[Pfeiffer et al., 2020] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulic, Sebastian Ruder, Kyunghyun Cho, Iryna Gurevych: AdapterHub: A Framework for Adapting Transformers. EMNLP (Demos) 2020: 46-54

Learning

- Unsupervised generation pre-training
  - Only the task module and fusion mechanism are trainable
  - training on unsupervised data from the source and target language

- Task fine-tuning
  - Only the task module are trainable
  - training on source language annotated task data

[Pfeiffer et al., 2020] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulic, Sebastian Ruder, Kyunghyun Cho, Iryna Gurevych: AdapterHub: A Framework for Adapting Transformers. EMNLP (Demos) 2020: 46-54

Outline

- Background & Motivation
- Method
  - Language-agnostic Task Acquisition
  - Language Specialization for Generation
  - Learning
- **Experiments**
- Takeaways

Experiment Setting

- Task and Dataset
  - Abstractive text summarization: XGIGA dataset
  - Question Generation: XQG dataset
- Language:
  - source language: En
  - target language: Zh, Fr
- Scenario: Zero-shot and Trans-train

Experiment Setting

- Backbone model: mBART
- Baselines
  - mBART (full): directly finetuning the full parameters of mBART on English annotated data;
  - mBART (enc): only finetuning the encoder parameters of mBART;
  - TM + adv: using adversarial training to force the output of TM to be language-agnostic

Main Result

Setting	Zero-shot		Trans-train	
Language	zh→zh	fr→fr	zh→zh	fr→fr
Baselines				
mBART (full)	43.82	33.40	47.33	42.8
mBART (enc)	45.85	36.55	47.09	42.11
TM + adv	31.41	36.71	**48.04**	43.04
LAFT	**46.37**	**40.78**	47.66	**43.10**

Table 1: Results of abstractive summarization. “full“: finetuning full model. “enc”: finetuning only encoder

Setting	Zero-shot	Trans-train
Language	zh→zh	zh→zh
Baselines		
mBART (full)	21.62	36.58
mBART (enc)	32.08	33.57
TM + adv	21.98	37.02
LAFT	**34.53**	**37.02**

- Zero-shot: LAFT outperform all baselines

Main Result

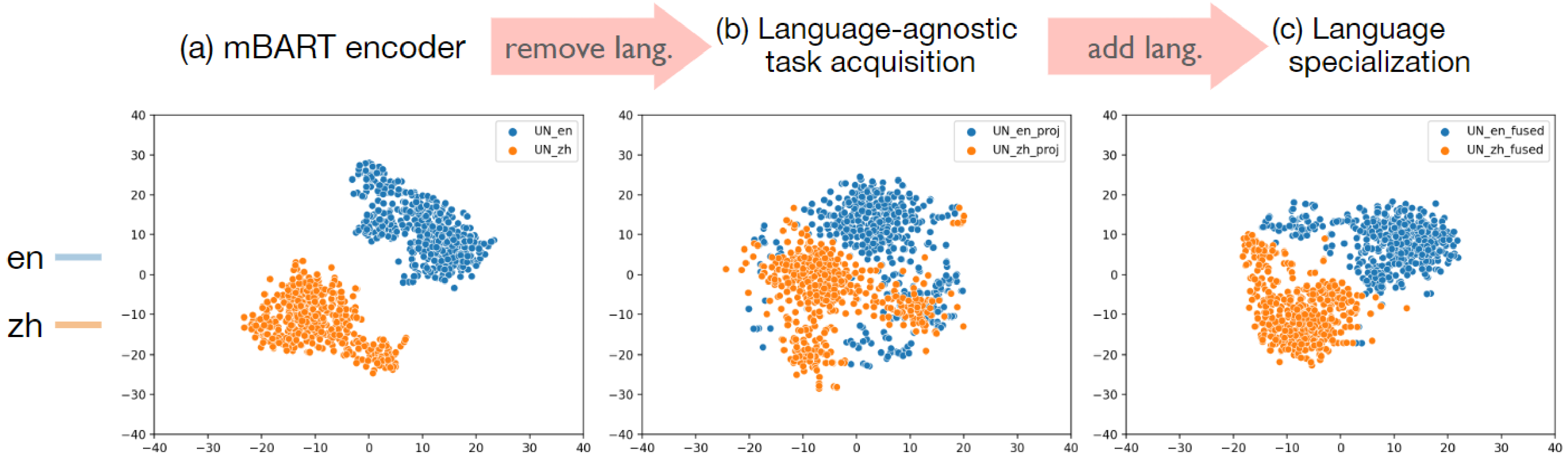
Setting	Zero-shot		Trans-train	
Language	zh→zh	fr→fr	zh→zh	fr→fr
Baselines				
mBART (full)	43.82	33.40	47.33	42.8
mBART (enc)	45.85	36.55	47.09	42.11
TM + adv	31.41	36.71	**48.04**	43.04
LAFT	**46.37**	**40.78**	47.66	**43.10**

Table 1: Results of abstractive summarization. "full": finetuning full model. "enc": finetuning only encoder

Setting	Zero-shot	Trans-train
Language	zh→zh	zh→zh
Baselines		
mBART (full)	21.62	36.58
mBART (enc)	32.08	33.57
TM + adv	21.98	37.02
LAFT	**34.53**	**37.02**

- Trans-train:
  - Baselines function better because of pseudo task data on target languages is accessible.
  - LAFT still perform good

Visualization



- After removing language identity, representations become closer.
- Representations become separable again after language specialization.

Outline

- Background & Motivation
- Method
  - Language-agnostic Task Acquisition
  - Language Specialization for Generation
  - Learning
- Experiments
- **Takeaways**

Takeaways

- Motivation:
  - Improving zero-resource cross-lingual transfer by language-agnostic fine-tuning
- Method
  - Language-agnostic task acquisition with an inserted task module
  - Language specialization for generation
- Experiments
  - Scenario: zero-shot and translate-train.
  - Tasks: Abstractive summarization and question generation

Thanks for your listening!

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

ByteDance AI Lab
字节跳动人工智能实验室