Improving Both Domain Robustness and Domain Adaptability in Machine Translation

Wen Lai[1], Jindřich Libovický[2], Alexander Fraser[1]

[1]Center for Information and Language Processing, LMU Munich, Germany
[2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

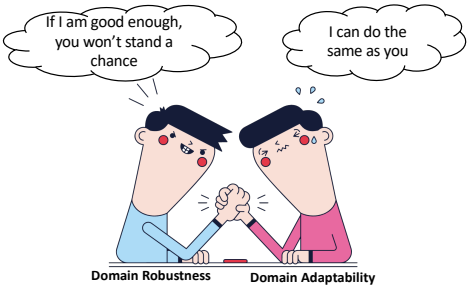
14th October, 2022

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Motivation

- We consider **two problems** of NMT domain adaptation using meta-learning:
  - **Domain Robustness**: we want to reach high quality on both domains seen in the training data and unseen domains.
  - **Domain Adaptability**: we want to make it possible to finetune systems with just hundreds of in-domain parallel sentences.



[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Research Question

- Which of these two properties is more important in NMT domain adaptation?

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Research Question

- Which of these two properties is more important in NMT domain adaptation?
- Can we improve both of the two properties at the same time?

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Related Works

- Domain Adaptability
  - Traditional fine-tuning: an out-of-domain model is continually trained on in-domain data (Freitag and Al-Onaizan, 2016; Dakwale and Monz, 2017).
  - Meta-Learning: train models which can be later rapidly adapted to new scenarios using only a small amount of data (Sharaf et al., 2020; Zhan et al., 2021).
- Domain Robustness
  - Müller et al. (2020) defined the concept of domain robustness and propose to improve the domain robustness by subword regulaization, defensive distillation, reconstruction and neural noisy channel ranking.
  - Jiang et al. (2020) proposed using individual modules for each domain with a word-level domain mixing strategy, which shows domain robustness on seen domains.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Goal

- Design a novel approach to improve both of the properties simultaneously.

Goal

- Design a novel approach to improve both of the properties simultaneously.
  - Domain Robustness

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Goal

- Design a novel approach to improve both of the properties simultaneously.
  - Domain Robustness
    - We first train a word-level domain mixing model to improve the robustness on seen domains, and show that, surprisingly, this improves robustness on unseen domains as well.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Goal

- Design a novel approach to improve both of the properties simultaneously.
  - Domain Robustness
    - We first train a word-level domain mixing model to improve the robustness on seen domains, and show that, surprisingly, this improves robustness on unseen domains as well.
  - Domain Adaptability

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Goal

- Design a novel approach to improve both of the properties simultaneously.
  - Domain Robustness
    - We first train a word-level domain mixing model to improve the robustness on seen domains, and show that, surprisingly, this improves robustness on unseen domains as well.
  - Domain Adaptability
    - we train a domain classifier based on BERT to score training sentences; the score measures similarity between out-of-domain and general-domain sentences.

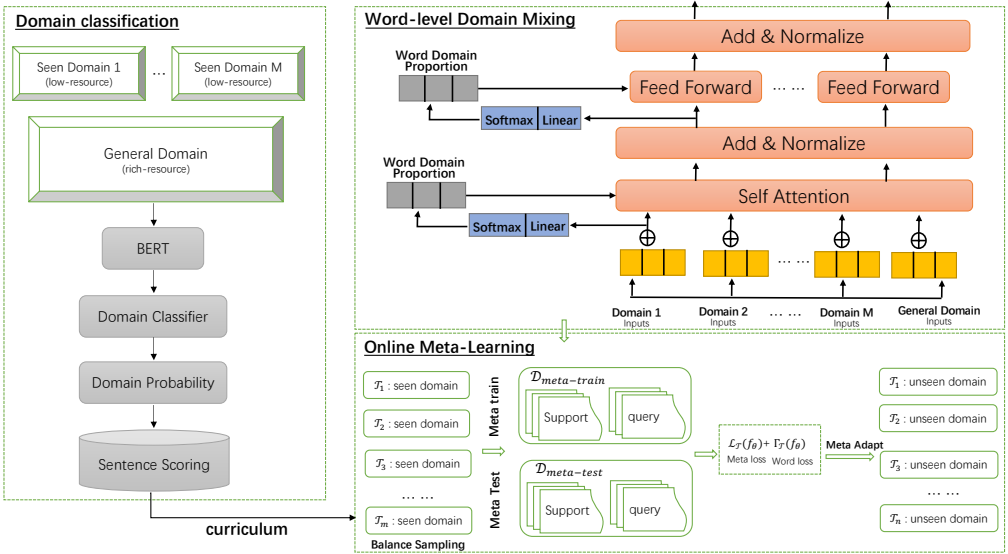
[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Goal

- Design a novel approach to improve both of the properties simultaneously.
  - Domain Robustness
    - We first train a word-level domain mixing model to improve the robustness on seen domains, and show that, surprisingly, this improves robustness on unseen domains as well.
  - Domain Adaptability
    - we train a domain classifier based on BERT to score training sentences; the score measures similarity between out-of-domain and general-domain sentences.
    - we integrate the domain-mixing model into a meta-learning framework with the domain classifier using a balanced sampling strategy.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic



[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Word-level Domain Mixing

- The domain of a word in the sentence is not necessarily consistent with the sentence domain. Therefore, we assume that every word in the vocabulary has a domain proportion, which indicates its domain preference.
- Each domain has its own multi-head attention modules. Therefore, we can integrate the domain proportion of each word into its multi-head attention module.
- The model can be efficiently trained by minimizing the composite loss:

$$L^* = L_{\text{gen}}(\theta) + L_{\text{mix}}(\theta)$$

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Domain Classification

- RieSS et al. (2021) show that using scores from simple domain classifier are more effective than scores from language models for NMT domain adaptation.
- We compute domain similarity using a sentence-level classifier, but in contrast with previous work, we based our classifier on a pre-trained language model (BERT).

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Online Meta-Learning

- We use domain classification scores as the curriculum to split the corpus into small tasks, so that the sentences more similar to the general domain sentences are selected in early tasks.
- Previous meta-learning approaches (Sharaf et al., 2020; Zhan et al., 2021) are based on token-size based sampling, which proved to be not balanced since some tasks did not contain all seen domains, especially in the early tasks. To address these issues, we sample the data uniformly from the domains to compensate for imbalanced domain distributions based on domain classifier scores.
- Following the balanced sampling, the process of meta-training is to update the current model parameter from θ to θ' using a MAML (Finn et al., 2017) objective with the traditional sentence-level meta-learning loss $\mathcal{L}_{\mathcal{T}}(f_\theta)$ and the word-level loss $\Gamma_{\mathcal{T}}(f_\theta)$ (L^* of $\mathcal{T}$).

$$L_{\mathcal{T}}(f_\theta) = \mathcal{L}_{\mathcal{T}}(f_\theta) + \Gamma_{\mathcal{T}}(f_\theta)$$

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Datasets

- For English→German translation task, we evaluate ten domains, which publicly avaliable on OPUS (Tiedemann et al., 2012);
  - *Bible, Books, ECB, EMEA, GlobalVoices, JRC, KDE, TED, WMT-News, COVID-19*
- For English→Chinese translation task, we use UM-Corpus (Tian et al., 2014) containing eight domains.
  - *Education, Microblog, Science, Subtitles, Laws, News, Spoken, Thesis*

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Baselines

- **Vanilla**. A standard Transformer-based NMT system. Note that we also use the $\mathcal{D}_{\text{meta-train}}$ corpus, which is more fair and stronger baseline;
- **Plain finetuning**. Fine-tune the vanilla system for each domain;
- **Plain finetuning + tag**. Using domain tag to fine-tune the system (Kobus et al., 2017);
- **Meta-MT**. Standard meta-learning approach (Sharaf et al., 2020);
- **Meta-Curriculum (LM)**. Meta-Learning approach using LM scores as the curriculum to sample the task (Zhan et al., 2021);
- **Meta-based w/o FT**. This series of experiments uses the meta-learning system prior to adaptation to the specific domain, which can be used to evaluate the domain robustness of meta-based models.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Domain Robustness

- Domain robustness shows the effectiveness of the model both in seen and unseen domains. Hence, we use the model without fine-tuning to evaluate the domain robustness.
- English→German

	Models	Unseen					Seen				
		Covid	Bible	Books	ECB	TED	EMEA	Globalvoices	JRC	KDE	WMT
1	Vanilla	24.34	**12.08**	12.61	29.96	27.89	37.27	24.19	39.84	27.75	27.38
2	Vanilla + tag	24.86	12.04	12.46	30.03	27.93	38.37	24.56	40.75	28.23	27.26
3	Meta-MT w/o FT	23.69	11.07	12.10	29.04	26.86	30.94	23.73	38.82	23.04	26.13
4	Meta-Curriculum (LM) w/o FT	23.70	11.16	12.24	28.22	27.21	33.49	24.27	39.21	27.60	25.83
5	RMLNMT w/o FT	**25.48**	11.48	**13.11**	**31.42**	**28.05**	**47.00**	**26.35**	**51.13**	**32.80**	**28.37**

- RMLNMT shows the best domain robustness compared with other models both in seen and unseen domains.
- The traditional meta-learning approach (Meta-MT, Meta-Curriculum) without fine-tuning is even worse than the standard transformer model in seen domains.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Domain Adaptability

- We evaluate the domain adaptability by testing that the model quickly adapts to new domains using just hundreds of in-domain parallel sentences.
- English→Chinese

	Models	Unseen				Seen			
		Education	Microblog	Science	Subtitles	Laws	News	Spoken	Thesis
1	Plain FT	27.05	26.31	32.09	17.77	47.64	28.28	25.73	28.47
2	Plain FT + tag	27.13	26.48	32.12	17.94	47.91	28.84	26.35	29.58
3	Meta-MT + FT	29.33	27.48	33.12	18.77	45.21	28.43	26.82	29.20
4	Meta-Curriculum (LM) + FT	28.91	27.20	33.19	18.93	45.46	28.17	27.84	29.47
5	RMLNMT + FT	**30.91**	**28.52**	**34.51**	**20.13**	**57.58**	**30.42**	**28.03**	**32.25**

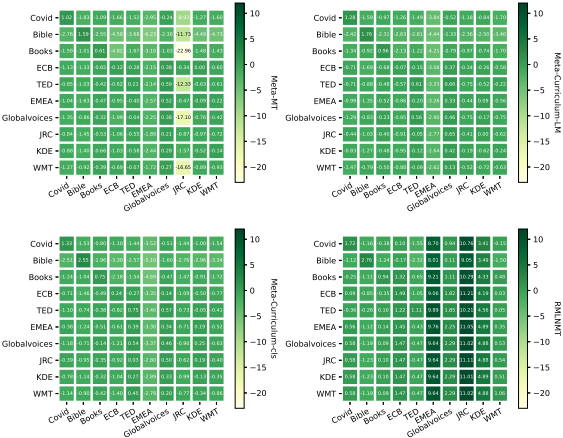
- we observe that the traditional meta-learning approach shows high adaptability to unseen domains but fails on seen domains due to limited domain robustness.
- In contrast, RMLNMT shows its domain adaptability both in seen and unseen domains, and maintains the domain robustness simultaneously.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Cross-Domain Robustness

- we use the fine-tuned model of one specific domain to generate the translation for other domains.
- Given k domains, we use the fine-tuned model M_J with the domain label of J to generate the translation of k domains.

Methods	Avg
Meta-MT	-1.97
Meta-Curriculum (LM)	-0.96
Meta-Curriculum (cls)	-0.98
RMLNMT	**2.64**



[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Different Classifiers

Classifier	Unseen					Seen				
	Covid	Bible	Books	ECB	TED	EMEA	Globalvoices	JRC	KDE	WMT
CNN	24.12	13.57	12.74	30.31	28.14	46.12	25.17	50.52	31.15	26.34
BERT-many-labels	25.89	14.77	13.71	**32.10**	29.28	**47.41**	**26.70**	51.34	32.76	28.17
BERT-2-labels	26.10	14.85	13.58	31.99	29.17	46.80	26.46	**51.56**	32.83	28.37
mBERT-many-labels	26.10	14.73	13.69	31.93	29.11	47.02	26.33	51.13	32.69	27.91
mBERT-2-labels	**26.53**	**15.37**	**13.71**	31.97	**29.47**	47.02	26.55	51.13	**32.88**	**28.37**

- **2-labels**: we distinguish only two classes: out-of-domain and in-domain
- **many-labels**: sentences are labeled with the respective domain labels.

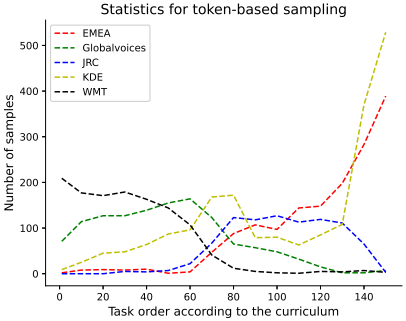
- the performance of RMLNMT is not directly proportional to the accuracy of the classifier. This is because the accuracy of the classifier is close between BERT-based models and the primary role of the classifier is to construct the curriculum for splitting the tasks.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Different Sampling Strategy

Sampling Strategy	Unseen					Seen				
	Covid	Bible	Books	ECB	TED	EMEA	Globalvoices	JRC	KDE	WMT
Token-based sampling	25.30	11.38	12.70	31.61	28.01	47.51	26.50	**51.31**	32.88	28.03
Balance sampling	**25.47**	**11.51**	**12.79**	**32.08**	**28.98**	**47.64**	**26.58**	51.25	**32.91**	**28.07**

- Our methods can result in small improvements in performance.



[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Different Fine-tuning Strategy

Finetune Strategy	Unseen					Seen				
	Covid	Bible	Books	ECB	TED	EMEA	Globalvoices	JRC	KDE	WMT
FT-unseen	25.23	13.18	12.73	**32.45**	28.41	46.35	25.83	50.85	32.30	26.88
FT-seen	24.58	11.73	12.57	30.79	27.29	46.58	25.73	50.91	31.78	26.51
FT-all	15.00	7.77	9.06	21.33	16.98	24.69	14.63	27.59	12.77	15.75
FT-specific	**26.53**	**15.37**	**13.71**	31.97	**29.47**	**47.02**	**26.33**	**51.13**	**32.83**	**28.37**

- **FT-unseen**: fine-tuning using all unseen domain corpora
- **FT-seen**: fine-tuning using all seen domain corpora
- **FT-all**: fine-tuning using all out-of-domain corpora (seen and unseen domains)
- **FT-specific**: using the specific domain corpus to fine-tune the specific models

- Fine-tuning in one specific domain obtains robust results among all the strategies.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Conclusion

- We presented RMLNMT, a robust meta-learning framework for low-resource NMT domain adaptation reaching both high domain adaptability and domain robustness (both in the seen domains and unseen domains).
- We found that domain robustness dominates the results compared to domain adaptability in meta-learning based approaches.
- The results show that RMLNMT works best in setups that require high robustness in low-resource scenarios.

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Thank You!

Wen Lai
OettingenstraSSe 67, 80538 Munich, Germany
lavine@cis.lmu.de
https://lavine-lmu.github.io