Specializing Multi-domain NMT via Penalizing Low Mutual Information

Jiyoung Lee†, Hantae Kim‡, Hyungchang Cho‡, Edward Choi†, Cheonbok Park‡

†KAIST ‡Papago, NAVER Corp.

EMNLP 2022

Motivation

- Multi-domain NMT, which translates multiple domains within a single model, should capture both **general** and **domain-specific** knowledge.
- Mutual Information (MI) between domain and translation represents the **dependency between the domain and the translated sentence**.
- A model with high MI tends to retain domain-specific terms in its translation.

Source	Beschreib ... **Summenberechnung** fur ein gegebenes Feld oder einen gegebenen Ausdruck.
Reference	Describe a way of **computing totals** for a given field or expression.
Model A with **Low MI**	Describe the kind of **calculation** for a given field or expression.
Model B with **High MI**	Describe the way of **computing totals** for a given field or expression

Table 1: Examples from different MI distributions in IT

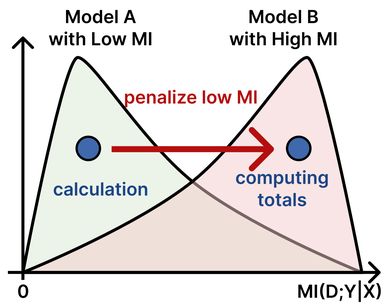

Fig 1. Overview of two models with different MI distributions

- In this study, we specialize multi-domain NMT by penalizing low MI to have higher value.

MI in Multi-domain NMT

Given X (source sentence), Y (target sentence), D (domain), MI is calculated as follows:

$$\mathrm{MI}(D;Y|X) = \mathbb{E}_{D,X,Y}\left[\log \frac{p(Y|X,D)}{p(Y|X)}\right]$$

We approximate with a parameterize model, namely cross-MI (XMI). (DA = domain adapted, G = general)

$$\mathrm{XMI}(D;Y|X) = \mathbb{E}_{D,X,Y}\left[\log \frac{p_{DA}(Y|X,D)}{p_G(Y|X)}\right]$$

Model Architecture

We assign adapters for each domain and an extra adapter for gerenal.

Notations

- θ : shared parameter (*e.g.*, self-attention and feed-forward layer)
- ϕ_d : corresponding domain adapter
- ϕ_G : general adapter

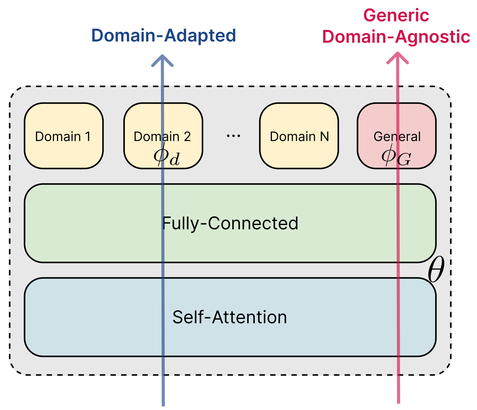

Fig 2. Model Architecture

Proposed Loss

Given the model architecture, XMI for y_i is calculated as below.

$$XMI(i) = p(y_i|y_{<i}, x, \theta, \phi_d) - p(y_i|y_{<i}, x, \theta, \phi_G)$$

We are going to peanlize low MI by weighting more on CE loss.

$XMI(i)$	$1 - XMI(i)$		
High	Low	→	Less weight on CE
Low	High	→	More weight on CE

Our MI-based proposed loss is as follows:

$$\mathcal{L}_{\mathrm{MI}} = \sum_{i}^{n_T} \underbrace{(1 - XMI(i))}_{\text{XMI weight}} \cdot \underbrace{(1 - p(y_i|y_{<i}, x, \theta, \phi_d))}_{\text{Cross Entropy Loss}}$$

Our final loss is as follows:

$$\mathcal{L}_{\mathrm{DA}} = -\sum_{i=0}^{n_T} \log(p(y_i|y_{<i}, x, \theta, \phi_d))$$

$$\mathcal{L}_{\mathrm{G}} = -\sum_{i=0}^{n_T} \log(p(y_i|y_{<i}, x, \theta, \phi_G))$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{DA}} + \lambda_1 \mathcal{L}_G + \lambda_2 \mathcal{L}_{\mathrm{MI}}$$

Multi-domain Translation

OPUS (De → En)

- Domains: IT, Koran, Law, Medical, Subtitles

	IT	Koran	Law	Medical	Subtitles	Average
Mixed	43.87	20.31	58.33	55.19	30.36	41.61
Domain-Tag	44.29	20.44	58.47	55.39	30.61	41.84
WDC	44.44	20.75	58.49	55.43	30.52	41.93
Adapter	44.50	20.37	58.22	56.00	31.02	42.02
Ours	**45.89 (+1.39)**	**20.80 (+0.43)**	**59.22 (+1.00)**	**56.34 (+0.34)**	**31.56 (+0.54)**	**42.76 (+0.74)**

Table 2: Average BLEU from five random seed experiments on OPUS.

Alhub (Ko → En)

- Domains: Finance, Ordinance, Tech

	Finance	Ordinance	Tech	Average
Mixed	52.50	56.65	66.00	58.38
Domain-Tag	52.71	56.60	66.03	58.45
WDC	52.75	56.56	65.93	58.41
Adapter	53.13	56.97	66.25	58.78
Ours	**53.87 (+0.74)**	**57.47 (+0.50)**	**66.66 (+0.41)**	**59.33 (+0.55)**

Table 3: Average BLEU from five random seed experiments on Alhub.

Qualitative Analysis of MI Distirbutions

XMI Distributions in OPUS

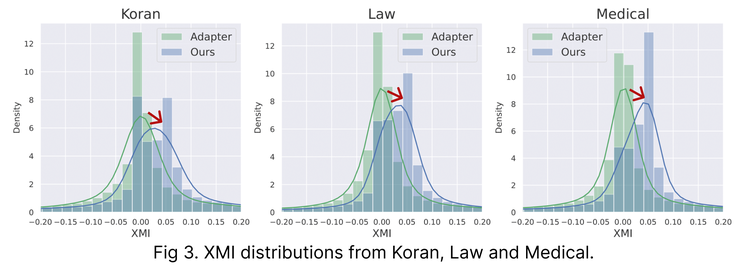

Fig 3. XMI distributions from Koran, Law and Medical.

Generation with XMI

Domain	IT
Source	Microsoft Office ; Importieren passwortgeschützter Dateien
Reference	Microsoft Office ; importing password protected files
Hypothesis	Microsoft Office ; importing password protected files

Fig 4. Example visualizations with XMI values

Contribution

- We reinterpret domain-specific knowledge in multi-domain NMT from MI perspective and present a new objective that penalizes low MI.
- From our experiment results, our proposed method yields domain-specialized model.