EMNLP 2022

Specializing Multi-Domain NMT via Penalizing Low Mutual Information

Jiyoung Lee[*†], Hantae Kim[‡], Hyunchang Cho[‡], Edward Choi[†] and Cheonbok Park[‡]

[†]KAIST [‡]Papago, NAVER Corp.

[*]Work done during internship at Papago Team, NAVER Corp.

KAIST AI Kim Jaechul Graduate School

NAVER papago

Motivation

What is Multi-Domain Neural Machine Translation (NMT)?

- Multi-Domain NMT translates multiple domains within one model.
- The model should capture both **general** and **domain-specific** knowledge.

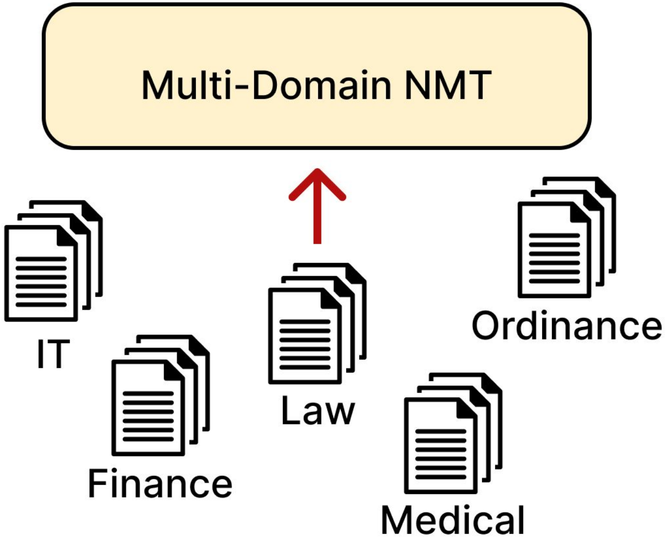


Fig 1. Multi-Domain Neural Machine Translation

Motivation

What is Multi-Domain Neural Machine Translation (NMT)?

- Multi-Domain NMT translates multiple domains within one model.
- The model should capture both **general** and **domain-specific** knowledge.

borrow **Mutual Information (MI)**

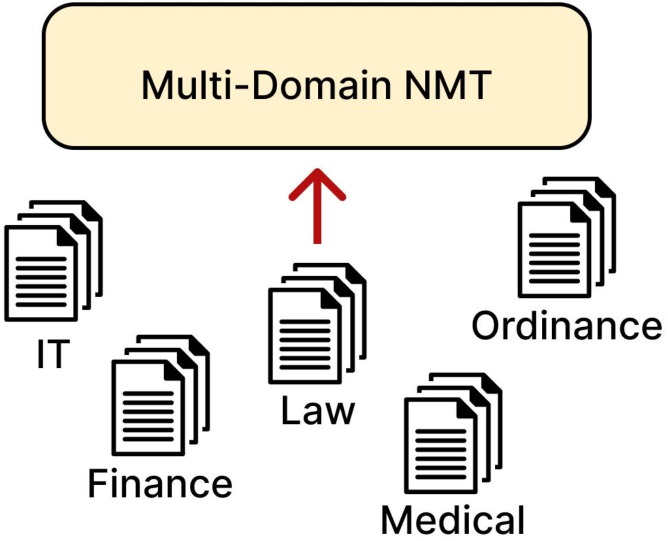


Fig 1. Multi-Domain Neural Machine Translation

Motivation

What is Mutual Information?

Mutual Information indicates the **mutual dependency** between two random variables.

$$MI(A;B) = \mathbb{E}_{A,B}\left[\frac{P(A,B)}{P(A)P(B)}\right]$$

Motivation

What is Mutual Information in Multi-Domain NMT?

In Multi-Domain NMT, we measure mutual dependency between **domain** and **translation**.

Given domain *D*, source sentence *X*, target sentence *Y*, mutual dependency can be written as $MI(D;Y|X)$.

Low MI : the mutual dependency between domain and translation is **low**

$\longrightarrow$ domain information is **less** used

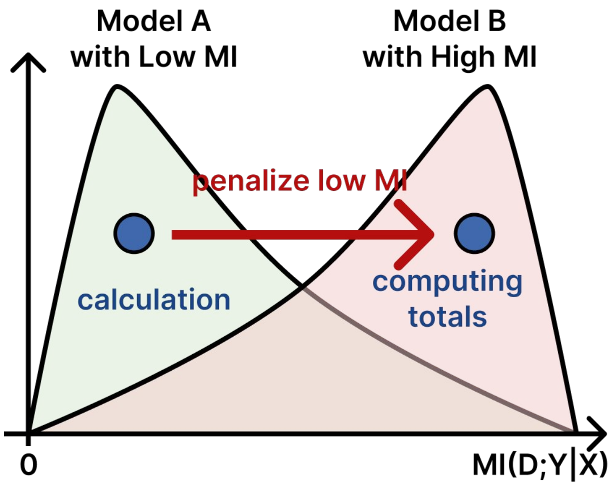
High MI: the mutual dependency between domain and translation is **high**

$\longrightarrow$ domain information is **more** used

Motivation

We compare outputs from two models with different MI distributions:

Model A with low MI, Model B with high MI.



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

Fig 2. Overview of two models with different MI distributions

From the result, high MI value helps in correctly retaining domain-specific terms.

→ In this paper, we aim to penalize low MI to have higher value to encourage model to learn domain knowledge.

Method

How Can We Get Mutual Information?

$$
\begin{aligned}
MI(D;Y|X) &= \mathbb{E}_{D,X,Y}\left[\log \frac{p(D,Y|X)}{p(D|X)\cdot p(Y|X)}\right] \\
&= \mathbb{E}_{D,X,Y}\left[\log \frac{p(D|Y,X)\cdot p(Y|X)}{p(D|X)\cdot p(Y|X)}\right] \\
&= \mathbb{E}_{D,X,Y}\left[\log \frac{p(X,Y,D)\cdot p(X)}{p(X,Y)\cdot p(X,D)}\right] \\
&= \boxed{\mathbb{E}_{D,X,Y}\left[\log \frac{p(Y|X,D)}{p(Y|X)}\right]}
\end{aligned}
$$

log quotient of translation with and without domain information.

Method

How Can We Get Mutual Information?

$$
\begin{aligned}
MI(D;Y|X) &= \mathbb{E}_{D,X,Y}\left[\log \frac{p(D,Y|X)}{p(D|X)\cdot p(Y|X)}\right] \\
&= \mathbb{E}_{D,X,Y}\left[\log \frac{p(D|Y,X)\cdot p(Y|X)}{p(D|X)\cdot p(Y|X)}\right] \\
&= \mathbb{E}_{D,X,Y}\left[\log \frac{p(X,Y,D)\cdot p(X)}{p(X,Y)\cdot p(X,D)}\right] \\
&= \boxed{\mathbb{E}_{D,X,Y}\left[\log \frac{p(Y|X,D)}{p(Y|X)}\right]}
\end{aligned}
$$

← Since we do not know the true distributions, we have to approximate with model output (**XMI**)

Method

How Can We Get Mutual Information?

$$
\begin{aligned}
MI(D;Y|X) &= \mathbb{E}_{D,X,Y}\left[\log \frac{p(D,Y|X)}{p(D|X)\cdot p(Y|X)}\right] \\
&= \mathbb{E}_{D,X,Y}\left[\log \frac{p(D|Y,X)\cdot p(Y|X)}{p(D|X)\cdot p(Y|X)}\right] \\
&= \mathbb{E}_{D,X,Y}\left[\log \frac{p(X,Y,D)\cdot p(X)}{p(X,Y)\cdot p(X,D)}\right] \\
&= \mathbb{E}_{D,X,Y}\left[\log \frac{p(Y|X,D)}{p(Y|X)}\right] \\
&\approx \mathbb{E}_{D,X,Y}[\underbrace{p(Y|X,D)}_{\text{Domain Adapted Model}} - \underbrace{p(Y|X)}_{\text{Generic Domain-Agnostic Model}}]
\end{aligned}
$$

Method

How Can We Get Mutual Information?

We need both Domain-Adapted model and Generic Domain-Agnostic model

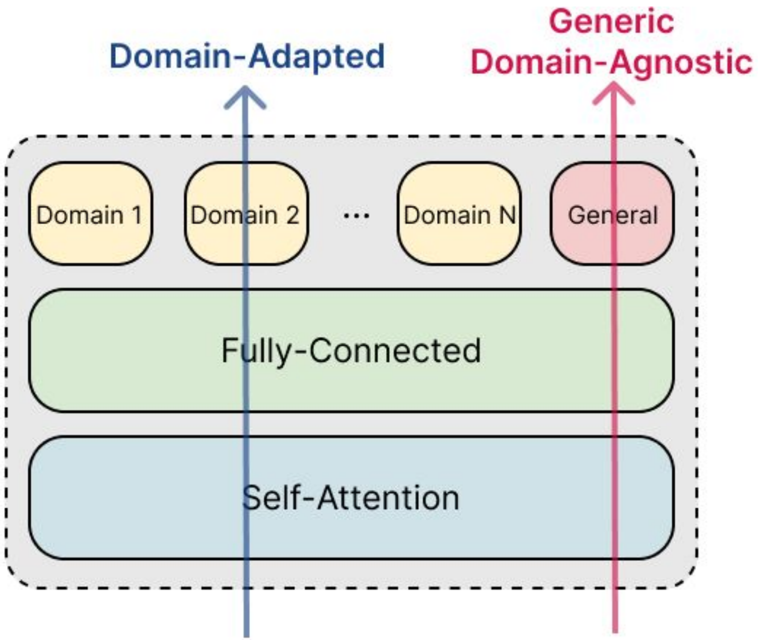


Fig 3. Model Architecture

Method

How Can We Get Mutual Information?

We need both Domain-Adapted model and Generic Domain-Agnostic model

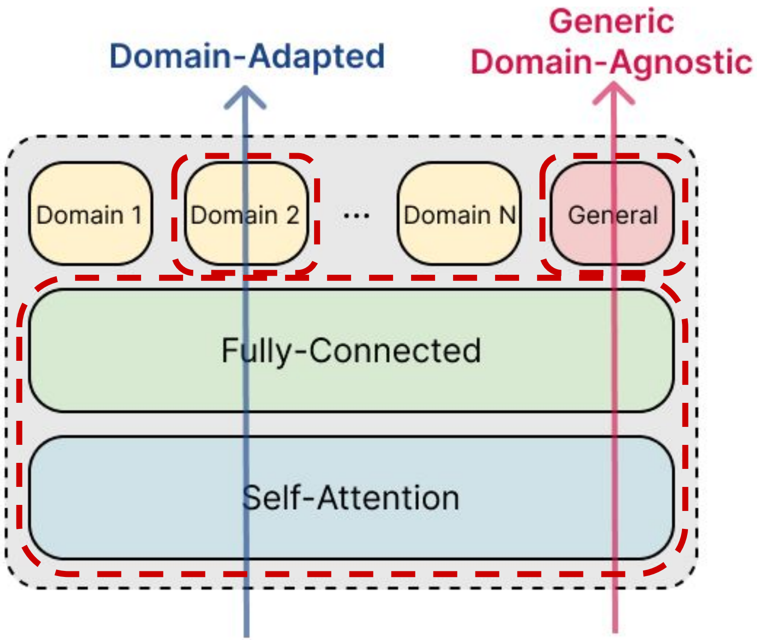


Fig 3. Model Architecture

ϕ_d	domain adapter
ϕ_G	general adapter
θ	shared parameter (Self-Attention + Fully-Connected)

Method

How Can We Get Mutual Information?

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

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

Domain Adapted $p(y_i\|y_{<i}, x, \theta, \phi_d)$		Generic Domain-Agnostic $p(y_i\|y_{<i}, x, \theta, \phi_G)$		$XMI(i)$	
0.8 ↑	−	0.6	=	+0.2	Good ☺️
0.3 ↓	−	0.6	=	-0.3	Bad ☹️

Method

How Can We Get Mutual Information?

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

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

Domain Adapted $p(y_i\|y_{<i}, x, \theta, \phi_d)$		Generic Domain-Agnostic $p(y_i\|y_{<i}, x, \theta, \phi_G)$		$XMI(i)$	
0.8 ↑	−	0.6	=	+0.2	Good ☺️
0.3 ↓		0.6		-0.3	Bad ☹️

high XMI(i) $\rightarrow$ less weight $\rightarrow$ less focus

low XMI(i) $\rightarrow$ more weight $\rightarrow$ more focus

Method

How Can We Penalize Mutual Information?

$$\mathcal{L}_{\text{MI}} = \sum_{i=0}^{n_T} \underbrace{(1 - \text{XMI}(i))}_{\text{XMI weight}} \cdot \underbrace{(1 - p(y_i | y_{<i}, x, \theta, \phi_d))}_{\text{Cross Entropy Loss}}$$

XMI	1-XMI		
Low	High	$\longrightarrow$	More weight on cross entropy loss
High	Low	$\longrightarrow$	Less weight on cross entropy loss

Method

Final Loss

MI Loss : $\mathcal{L}_{\text{MI}} = \sum_{i=0}^{n_T} (1 - \text{XMI}(i)) \cdot (1 - p(y_i|y_{<i}, x, \theta, \phi_d))$

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

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

$$\mathcal{L} = \mathcal{L}_{\text{DA}} + \lambda_1 \mathcal{L}_{\text{G}} + \lambda_2 \mathcal{L}_{\text{MI}}$$

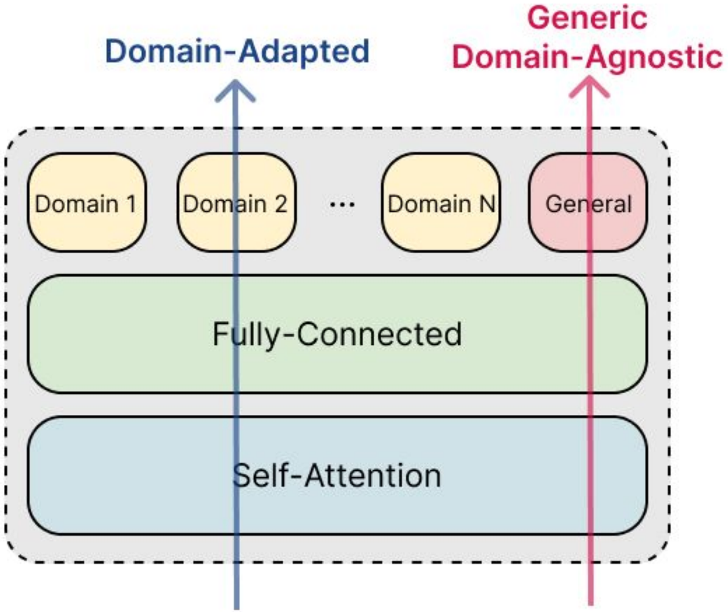


Fig 3. Model Architecture

Experiment

OPUS (De → En)

	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)**

Tab 1. Average BLEU from five random seed experiments on OPUS

AIhub (Ko → En)

	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)**

Tab 2. Average BLEU from five random seed experiments on AIhub

- Baselines performs on par with Mixed (no distinctions among domains)
  - Baseline models are not sufficiently using domain information.
- Our model outperforms all baselines with significant margins.

Experiment

XMI Distributions in OPUS

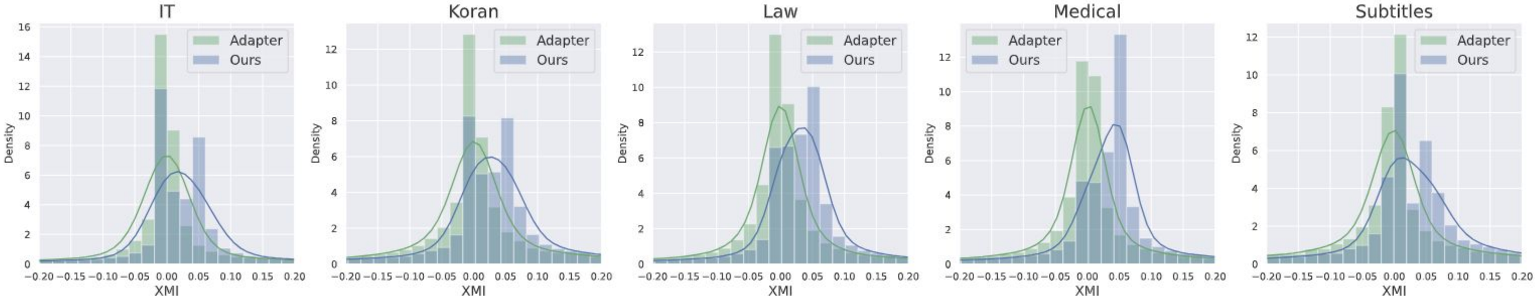


Fig 4. XMI distributions from all domains in OPUS

- XMI values have higher values in all domains in OPUS
- Our proposed loss is effective in maximizing mutual information

XMI Visualization on Generations

MI Value	-1.0	-0.71	-0.43	-0.14	0.14	0.43	0.71	1.0

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

Domain	Medical
Source	Eine Durchstechflasche enthält 150 mg Omalizumab .
Reference	One vial contains 150 mg of omalizumab .
Hypothesis	One vial contains 150 mg of omalizumab .

Fig 5. XMI visualization in generated outputs

Domain-specific terms (*e.g.*, password, omalizumab) are generated with high XMI values.

Conclusion

Take Home Message

- Previous Multi-domain NMTs show similar performance as Mixed
- We encourage model to learn domain-specific knowledge by penalizing low mutual information.

Further Experiments & Analysis

- Increase in Translation Performance on Domain Specialized Sentences
- Comparison on Computation Cost for Training

paper

poster

KAIST AI
Kim Jaechul Graduate School

NAVER papago