Multi-Domain Neural Machine Translation with Word-Level Domain Context Discrimination

Jiali Zeng Jinsong Su Huating Wen   Yang Liu   Jun Xie
Yongjing Yin   Jianqiang Zhao

Reporter : Xiachong Feng

Author

Jiali Zeng
厦门大学数字媒体计算研究中心
曾嘉莉 2017级硕士研究生

Jinsong Su
厦门大学数字媒体计算研究中心
苏劲松 副教授、硕士生导师

Challenge

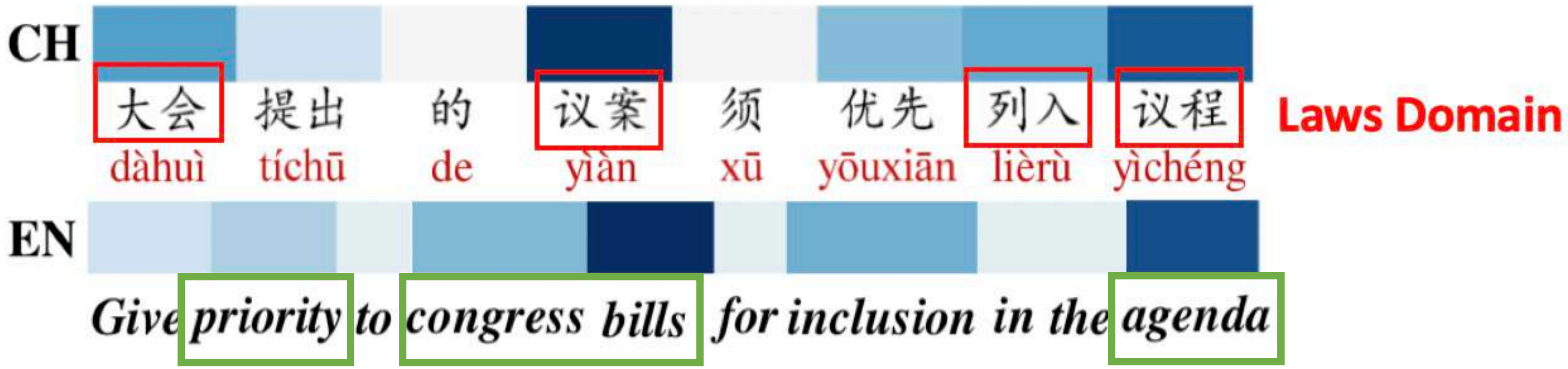
- Training a NMT model for a **specific domain** requires a large quantity of parallel sentences in such domain, which is often **not readily available.**
- The translated sentences often **belong to multiple domains**, thus requiring a NMT model general to different domains.

Previous

- Using mixed-domain parallel sentences to construct a unified model that allows translation to switch between different domains.

Motivation

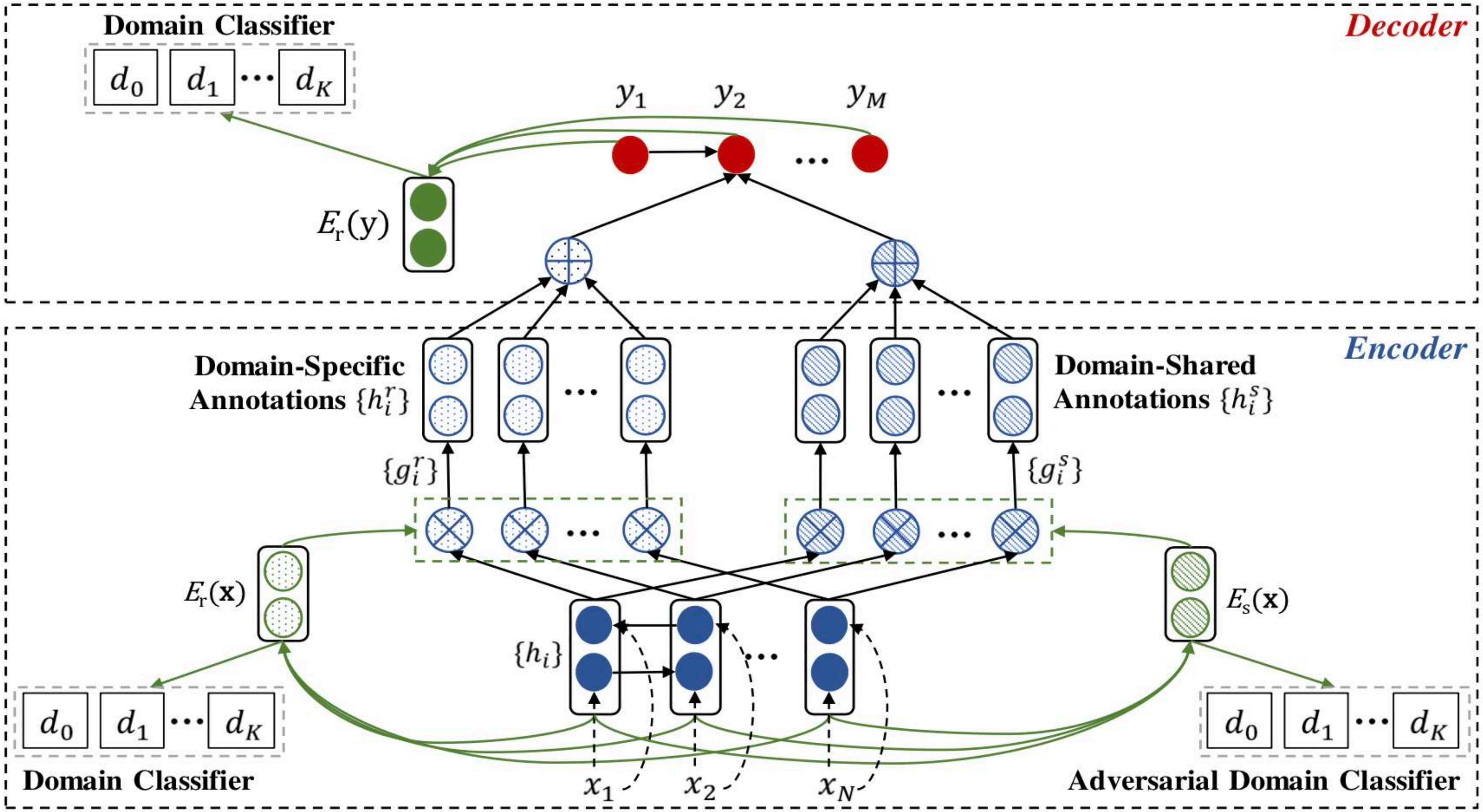
1. Since the textual styles, sentence structures and terminologies in different domains are often remarkably distinctive, whether such domain-specific translation knowledge is effectively preserved could have a direct effect on the performance of the NMT model.
2. Words in a sentence are related to its domain



3. It is also reasonable for our model to pay more attention to these domain-related words than the others during model training.

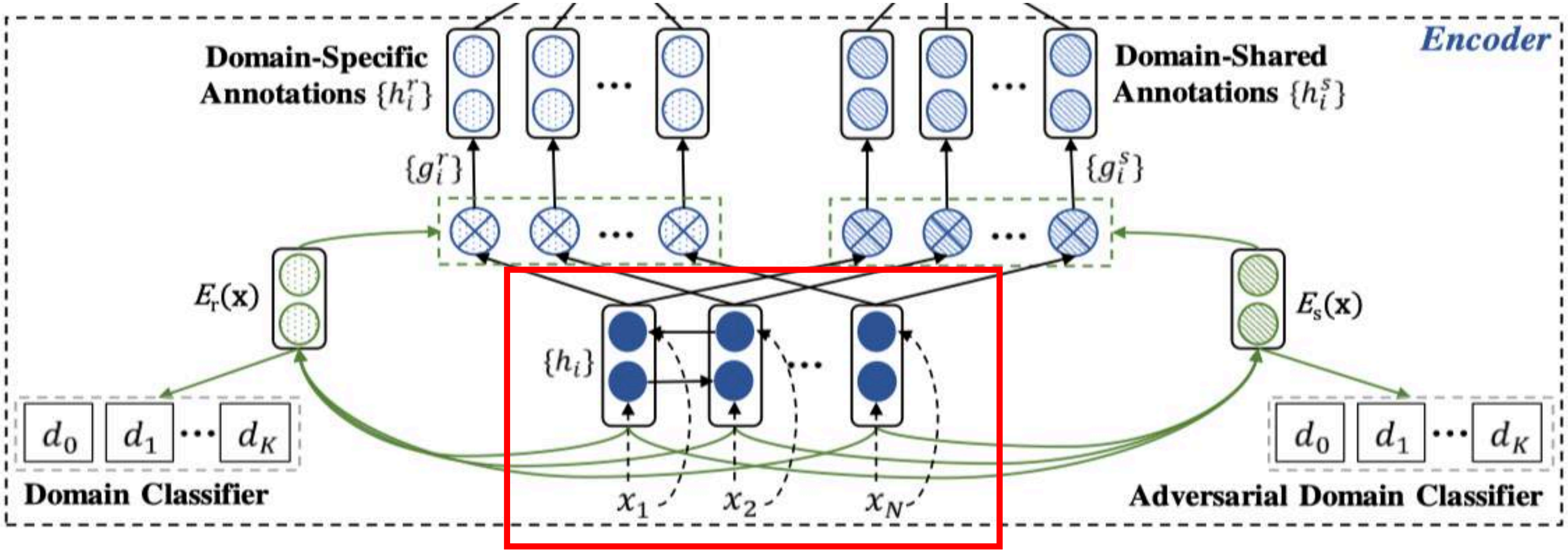
- Context = **domain-specific** + domain-shared

Model



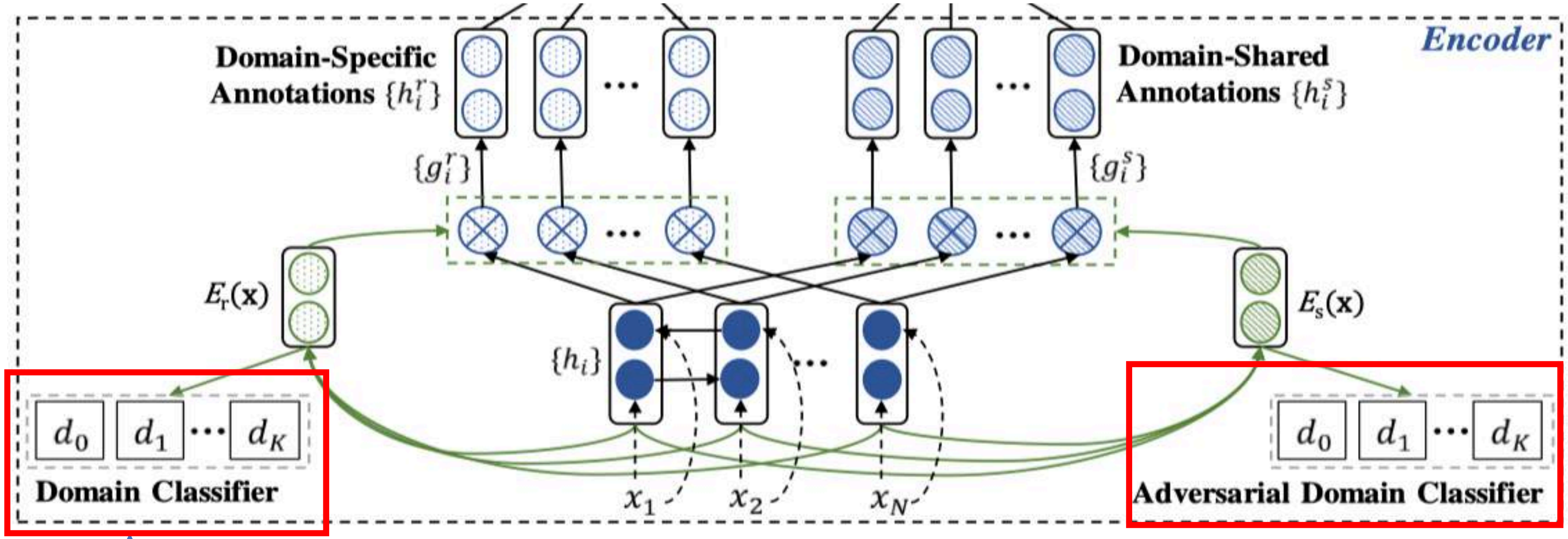
Domain Classifier
d_0 d_1 ... d_K
Decoder
y_1 y_2 y_M
$E_r(y)$
Encoder
Domain-Specific Annotations $\{h_i^r\}$
Domain-Shared Annotations $\{h_i^s\}$
$\{g_i^r\}$
$\{g_i^s\}$
$E_r(\mathbf{x})$
$E_s(\mathbf{x})$
$\{h_i\}$
d_0 d_1 ... d_K
d_0 d_1 ... d_K
Domain Classifier
x_1 x_2 x_N
Adversarial Domain Classifier

Encoder



Bidirectional GRU

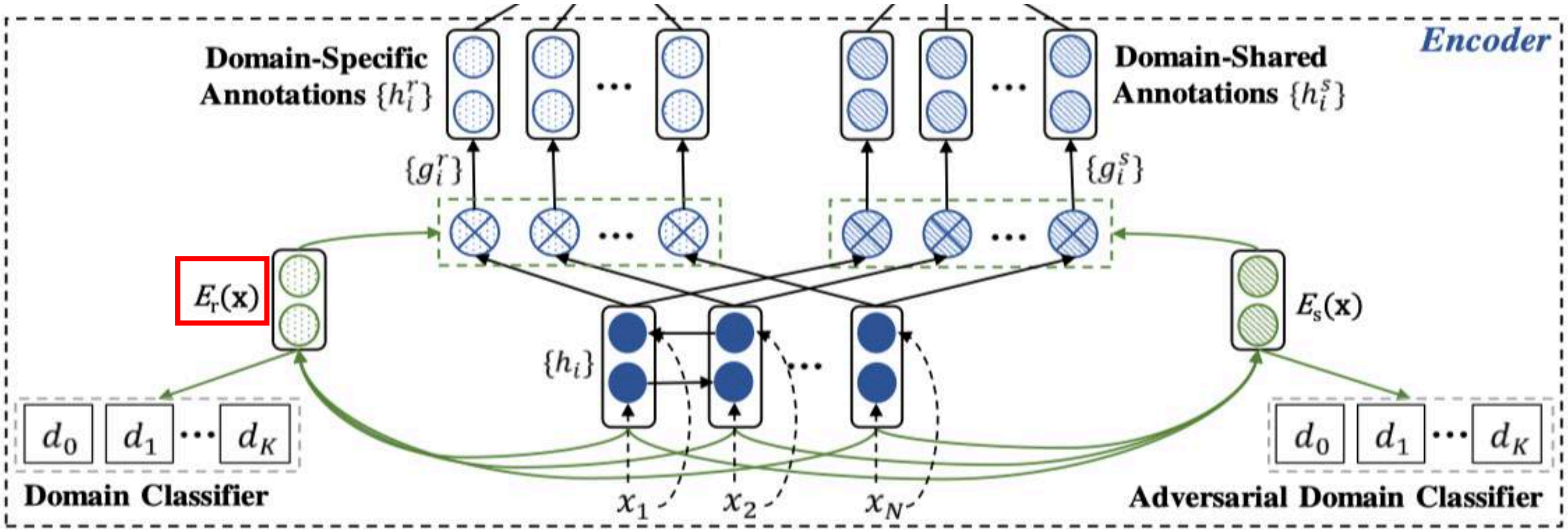
Encoder



Domain classifier that aims to distinguish different domains in order to generate **domain-specific source-side contexts**.

Adversarial domain classifier capturing **source-side domain shared contexts.**

Domain Classifier



Attention

$$E_r(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i h_i,$$

$$\text{where} \quad \alpha_i = \frac{exp(e_i)}{\sum_{i'}^{N} exp(e_{i'})},$$

$$e_i = (v_a)^\top \tanh(W_a h_i),$$

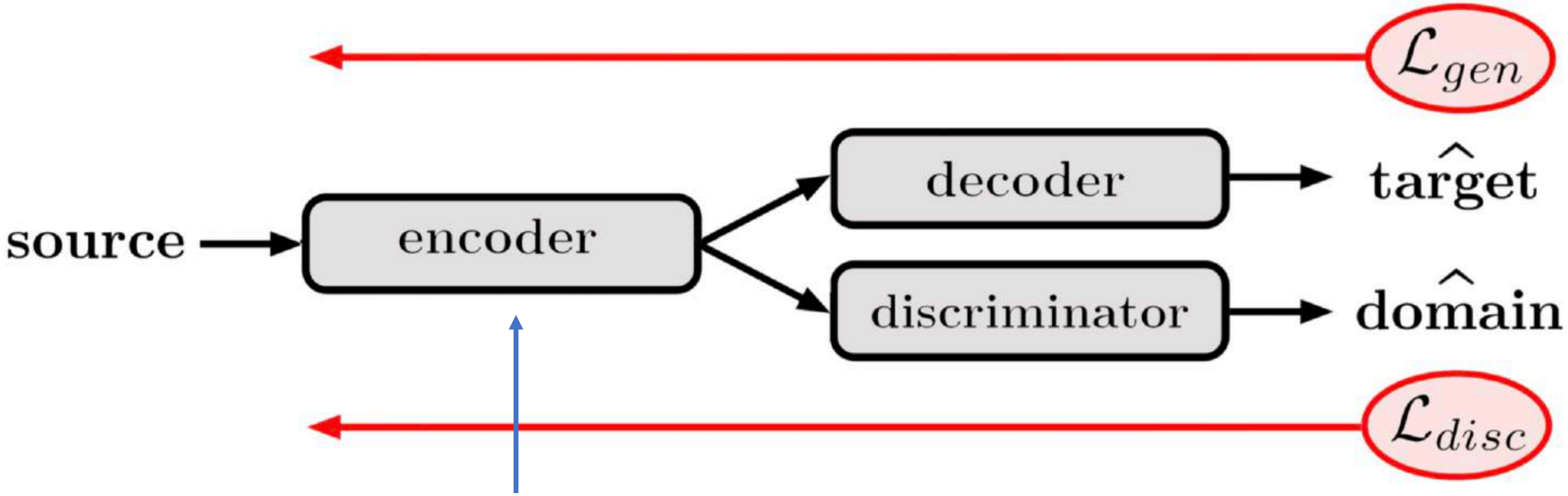
$$\mathcal{J}_{dc}^{s}(\mathbf{x}; \boldsymbol{\theta}_{dc}^{s}) = \log p(d|\mathbf{x}; \boldsymbol{\theta}_{dc}^{s})$$

Object Func

$$p(\cdot|\mathbf{x}; \boldsymbol{\theta}_{dc}^{s})$$

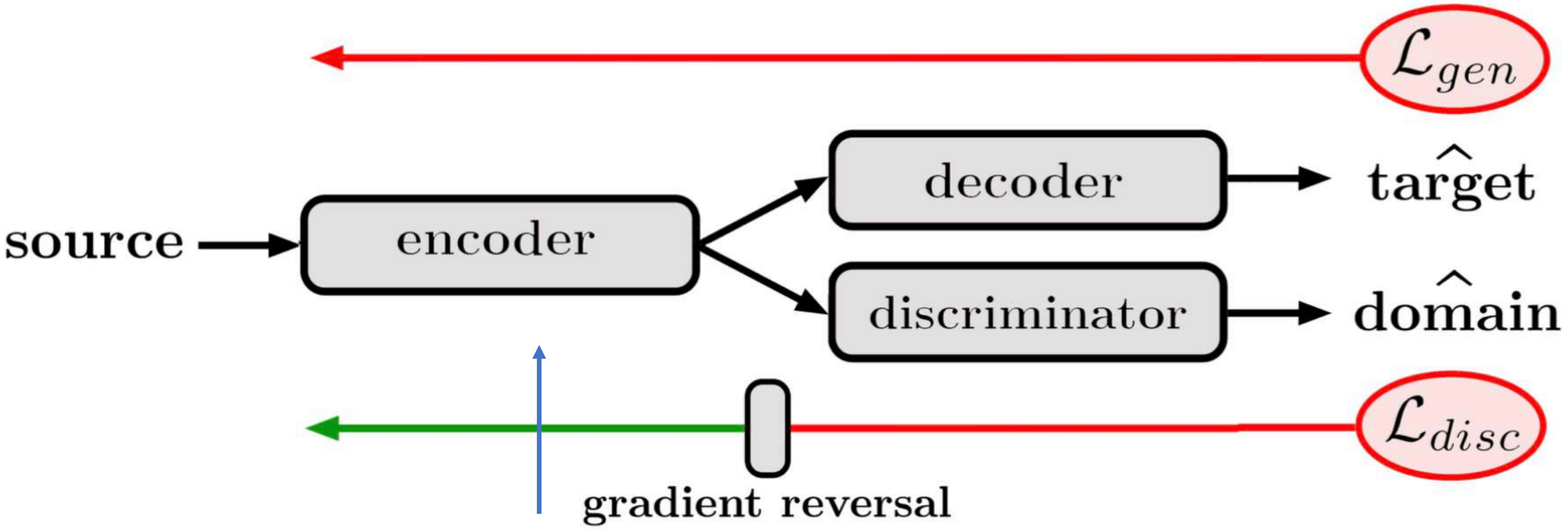
$$= softmax(W_{dc}^{s\top} \, ReLU(E_r(\mathbf{x})) + b_{dc}^{s}),$$

Effective Domain Mixing for Neural Machine Translation *WMT17*



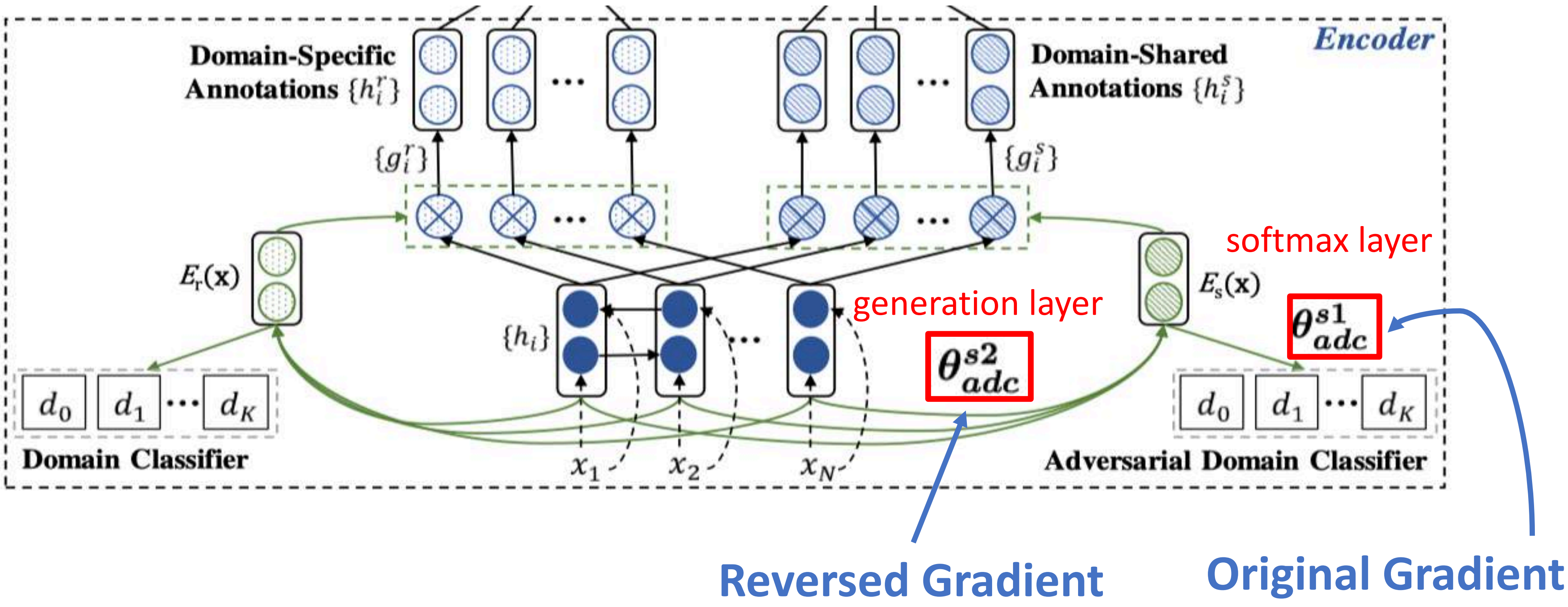
Domain Specific

Effective Domain Mixing for Neural Machine Translation *WMT17*

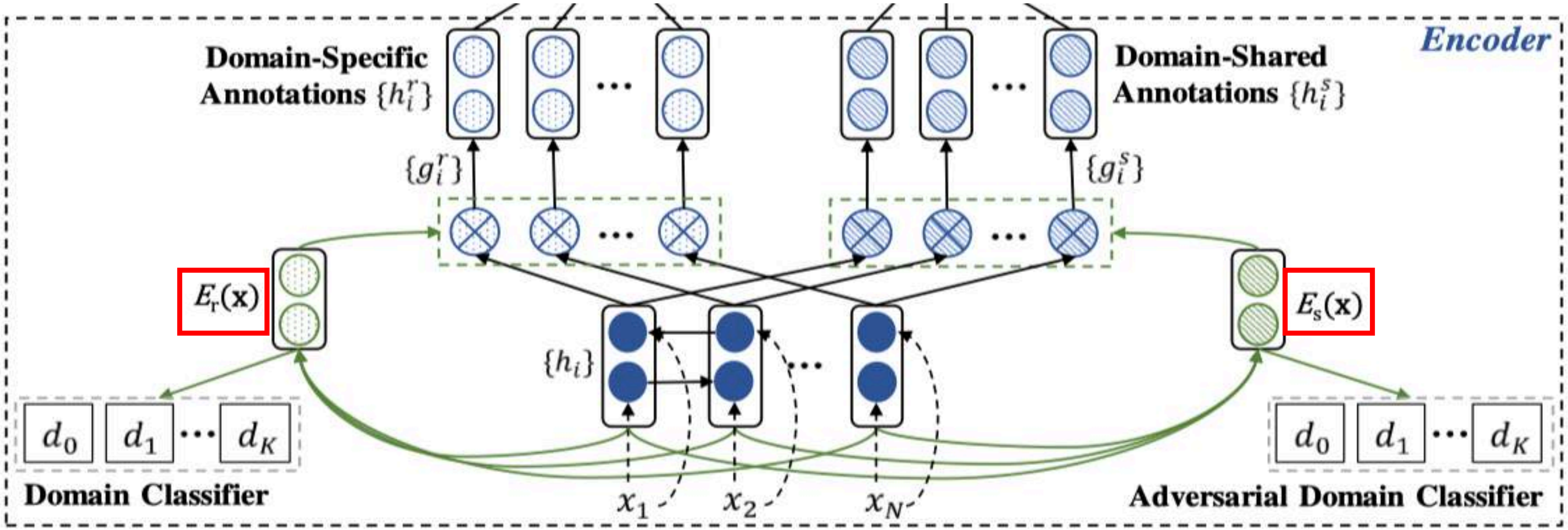


Domain Share

Adversarial Domain Classifier



Encoder



$$g_i^r = sigmoid(W_{gr}^{(1)} \underline{E_r(\mathbf{x})} + W_{gr}^{(2)} \underline{h_i} + b_{gr})$$

$$g_i^s = sigmoid(W_{gs}^{(1)} \underline{E_s(\mathbf{x})} + W_{gs}^{(2)} \underline{h_i} + b_{gs})$$

$$h_i^r = g_i^r \odot h_i,$$
$$h_i^s = g_i^s \odot h_i.$$

Decoder

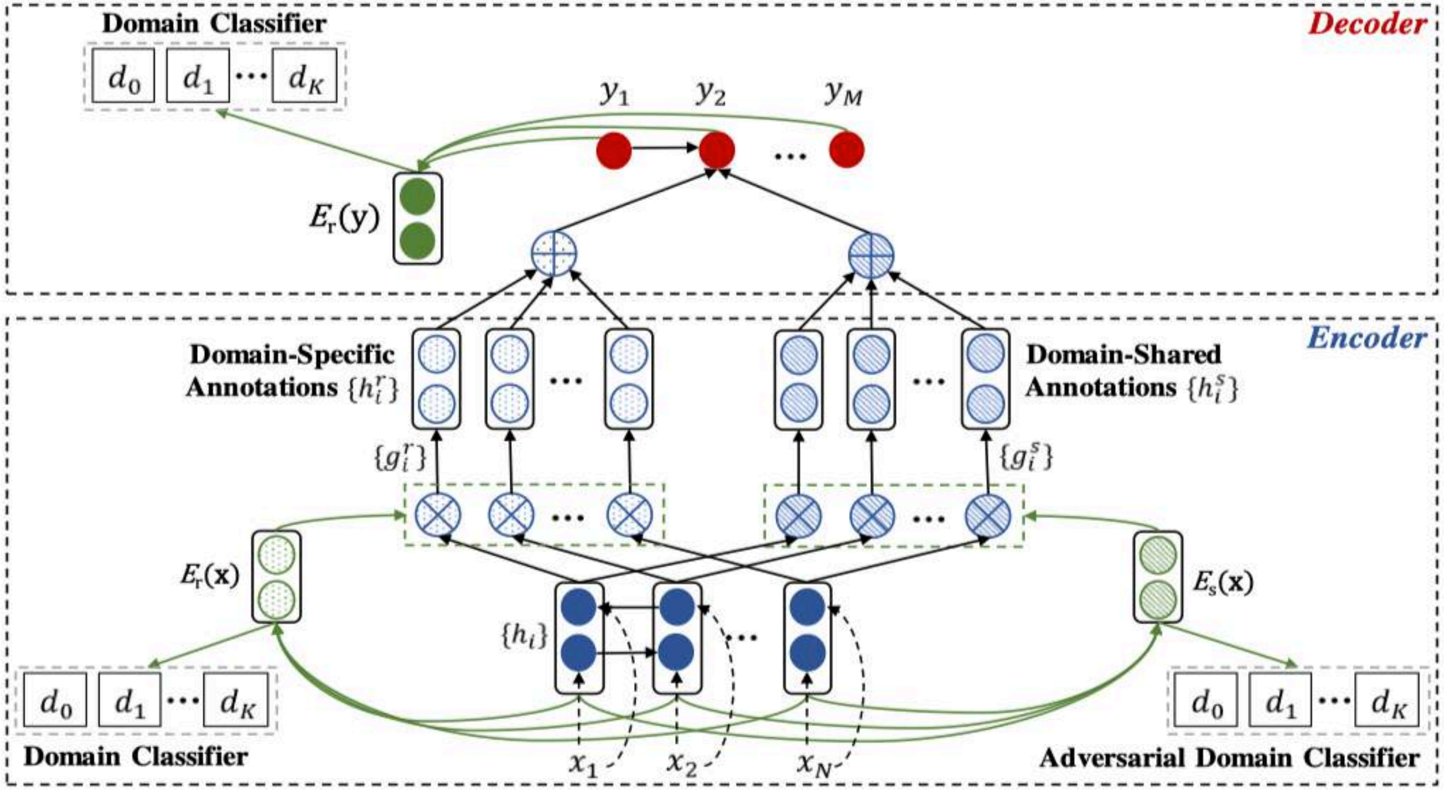
$$s_j = GRU(s_{j-1}, y_{j-1}, c_j^r, c_j^s).$$

GRU Hidden

$$c_j^r = \sum_{i=1}^{N} \frac{\exp(e_{j,i}^r)}{\sum_{i'=1}^{N} \exp(e_{j,i'}^r)} \cdot h_i^r,$$

$$\text{where} \quad e_{j,i}^r = a(s_{j-1}, h_i^r),$$

a *is a feedforward neural network.*



Decoder

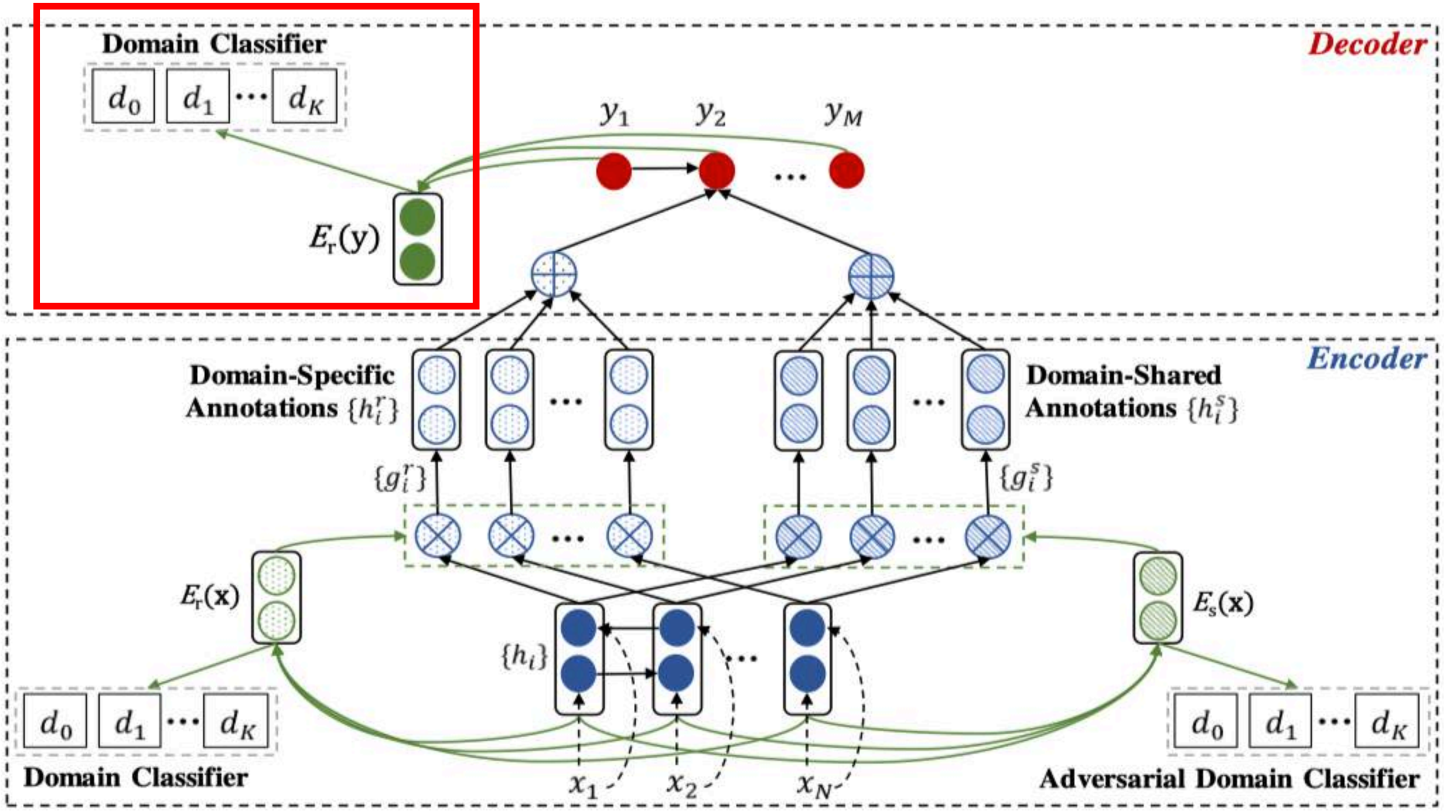
$$E_r(\mathbf{y}) = \sum_{j=1}^{M} \beta_j s_j,$$

$$\text{where} \quad \beta_j = \frac{\exp(e_j)}{\sum_{j'}^{M} \exp(e_{j'})},$$

$$e_j = (v_b)^\top \tanh(W_b s_j),$$

NMT Training Objective with **Word-Level Cost Weighting**.

$$\mathcal{J}_{nmt}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}_{nmt}) = \sum_{j=1}^{M} (1 + \beta_j) \log p(y_j | \mathbf{x}, y_{<j}; \boldsymbol{\theta}_{nmt}),$$



Overall Training Objective

$$
\begin{aligned}
\mathcal{J}(\mathcal{D};\boldsymbol{\theta}) = \sum_{(\mathbf{x},\mathbf{y},d)\in\mathcal{D}} & \{\mathcal{J}_{nmt}(\mathbf{x},\mathbf{y};\boldsymbol{\theta_{nmt}}) \\
& + \mathcal{J}^{s}_{dc}(\mathbf{x};\boldsymbol{\theta^{s}_{dc}}) + \mathcal{J}^{t}_{dc}(\mathbf{y};\boldsymbol{\theta^{t}_{dc}}) \\
& + \mathcal{J}^{s1}_{adc}(\mathbf{x};\boldsymbol{\theta^{s1}_{adc}}) + \lambda\cdot\mathcal{J}^{s2}_{adc}(\mathbf{x};\boldsymbol{\theta^{s2}_{adc}})\}
\end{aligned}
$$

Experiment

- Chinese-English translation
  - Laws, Spoken, Thesis, News
- English-French translation
  - Medical, News, Parliamentary

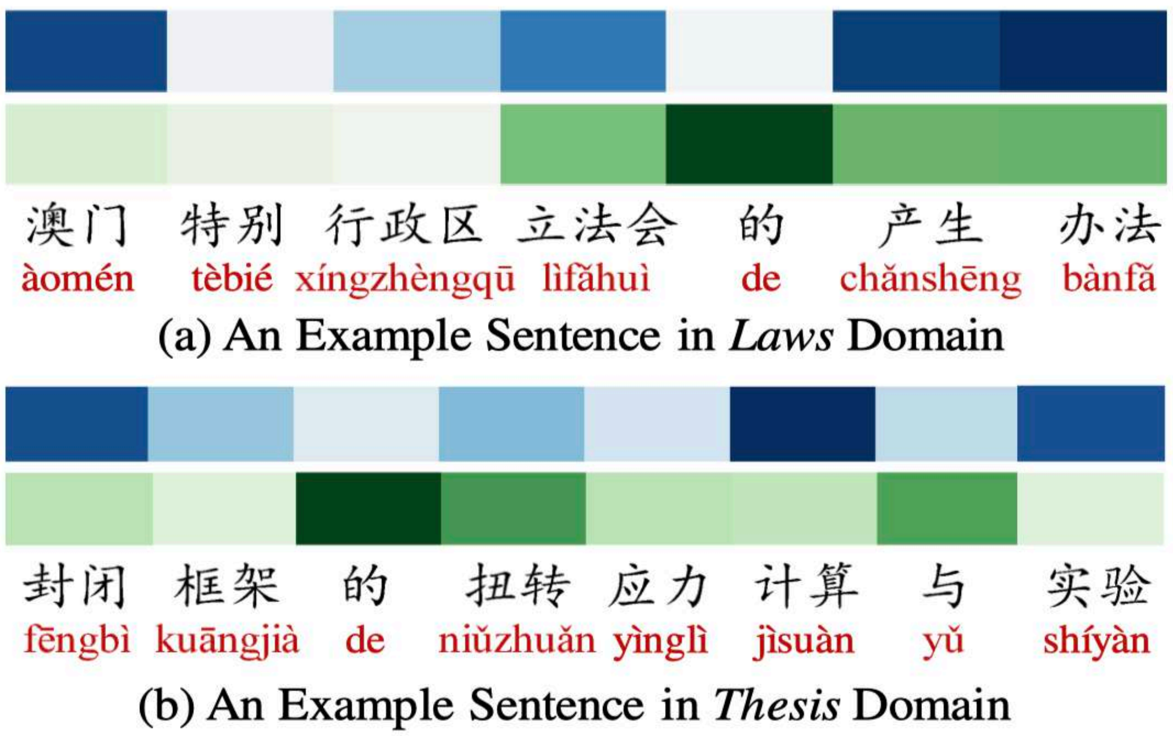
Task	Domain	Train	Dev	Test
CH-EN	Laws	219K	600	456
	Spoken	219K	600	455
	Thesis	299K	800	625
	News	300K	800	650
EN-FR	Medical	1.09M	800	2000
	News	180K	800	2000
	Parliamentary	2.04M	800	2000

Experiment

1. **DL4NMT-single**
   - Attentional NMT trained on a single domain dataset.
2. **DL4NMT-mix**
   - attentional NMT trained on mix-domain training set.
3. **DL4NMT-finetune**
   - first trained using out-of-domain training corpus and then fine-tuned using in-domain dataset.
4. **DC**
   - introduces embeddings of source domain tag
5. **ML1**
   - shares encoder representation and separates the decoder modeling of different domains.
6. **ML2**
   - NMT with domain classification via multitask learning.
7. **ADM**
   - adversarial training to achieve the domain adaptation in NMT.
8. **TTM**
   - adding target-side domain tag

Model	Laws	Spoken	Thesis	News
Contrast Models (1×**hd**)				
OpenNMT	45.82	9.15	13.93	19.73
DL4NMT-single	43.66	5.49	14.54	18.74
DL4NMT-mix	46.82	8.95	15.93	20.33
DL4NMT-finetune	54.19	8.77	16.71	21.55
+DC	49.83	9.18	16.71	20.58
+ML1	46.82	6.66	15.10	20.17
+ML2	48.95	9.45	15.85	20.48
+ADM	48.30	9.41	16.34	20.06
+TTM	49.05	9.36	16.42	20.44
Contrast Models (2×**hd**)				
DL4NMT-single	44.48	6.29	14.66	19.87
DL4NMT-mix	48.74	9.01	16.12	20.14
DL4NMT-finetune	54.69	9.07	17.11	21.85
+DC	50.43	9.38	16.45	20.44
+ML1	49.49	7.67	15.50	20.34
+ML2	50.05	9.35	16.03	20.64
+ADM	48.33	9.06	16.59	19.69
+TTM	49.92	9.01	16.38	21.04
Our Models				
+WDC(S)	54.55	10.12	17.22	22.16
+WDC(T)	51.94	9.76	17.72	21.02
+WDC	**55.03**	**10.20**	**18.04**	**22.29**

Visualizations of Gating Vectors



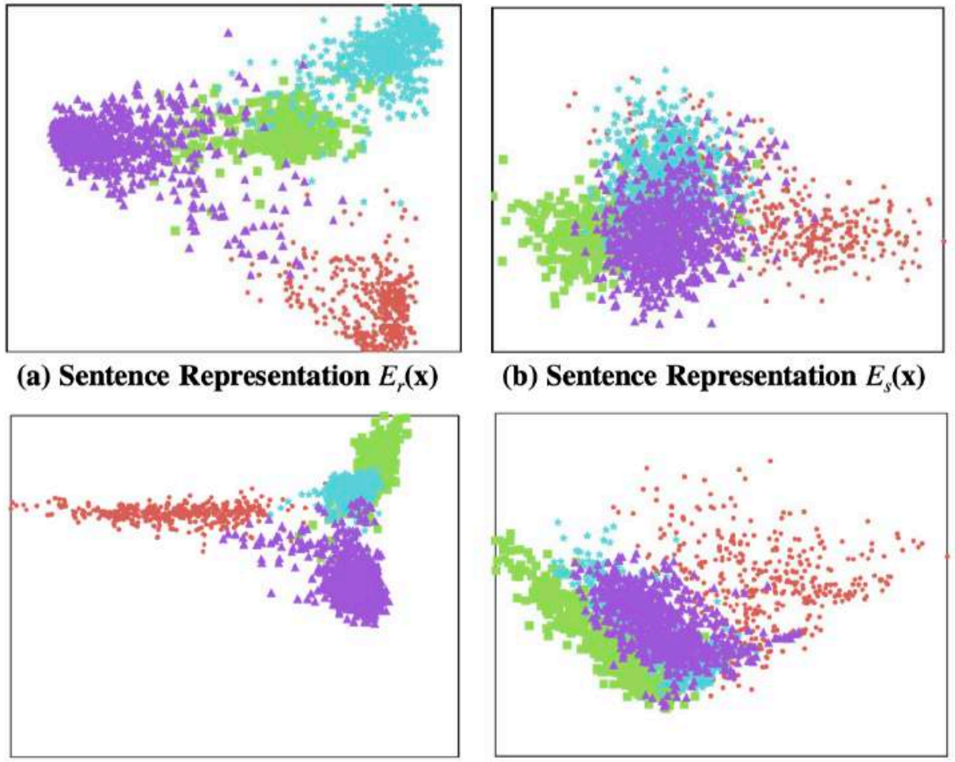
g_i^r Domain-specific

g_i^s Domain-shared

g_i^r Domain-specific

g_i^s Domain-shared

Visualizations of Sentence Representations and Annotations



(a) Sentence Representation $E_r(\mathbf{x})$ (b) Sentence Representation $E_s(\mathbf{x})$

(c) Average of Sentence Annotations $\{h_i^r(\mathbf{x})\}$ (d) Average of Sentence Annotations $\{h_i^s(\mathbf{x})\}$

pentagonal-shaped(blue) points denote *News*, *Laws*, *Spoken* and *Thesis* sentences, respectively.

Illustrations of Domain-Specific Target Words

Domain	Top10 Target Words
Laws	*Article*, *Chapter*, *Principles*, *regulations*, *Provisions*, *Political*, *Servants*, *specify*, *China*, Municipal
Spoken	*meanly*, *Rusty*, *1910s*, *scours*, *mountaintops*, *paralyze*, *Puff*, *perpetrators*, *hitter*, *weightlifting*
Thesis	*aggregation*, *Activities*, *Computation*, *Alzheimer*, *nn*, *Contemporarily*, *EVALUATION*, *ethoxycarbonyl*, *sCRC*, *Announced*
News	*months*, *agency*, *outweighed*, *unconstitutionally*, *Congolese*, *session*, *Asia*, *news*, *hurts*, *francs*

Table 3: Examples of Domain-Specific Target Words.

Thanks!