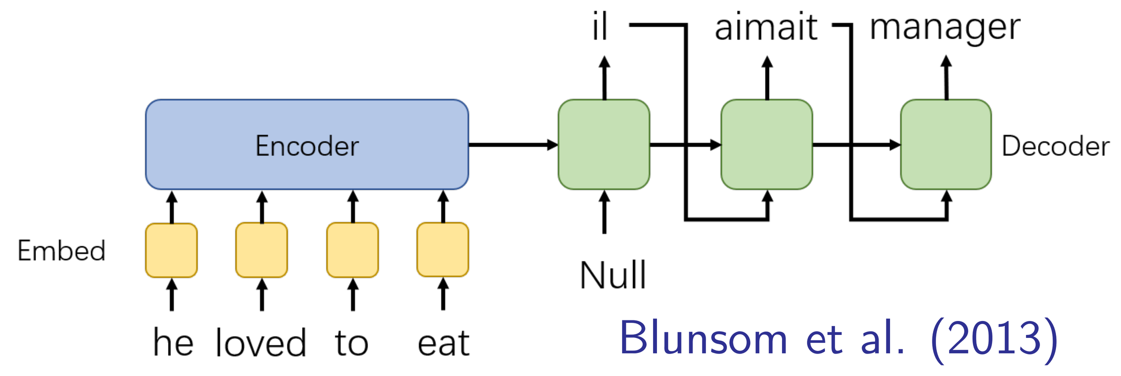
Multi-Domain Neural Machine Translation with Word Level Adaptive Layer-wise Domain Mixing

Haoming Jiang*, Chen Liang*, Chong Wang◇, Tuo Zhao*
*Georgia Tech ◇ByteDance (Seattle)

Objective

Neural Machine Translation (NMT)



Blunsom et al. (2013)

- **Motivation**: Handle data from multiple domains by sharing knowledge (Haddow et al. 2012).
- **Challenges**: Enforcing knowledge sharing lacks adaptivity to each individual domain.
- **Example**: Failure to handle word-level ambiguity across domains: The word "articles" has different meanings in laws and media domains.

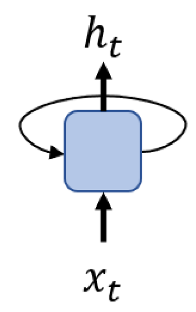
Laws	"**Article** 37 The freedom of marriage ..." "第三十七条:婚姻的自由..."
Media	"... working on an **article** about the poems ..." "... 为了一篇诗的文章 ..."

Question: How to adaptively capture domain-shared and domain-specific knowledge to improve multi-domain NMT?

Background

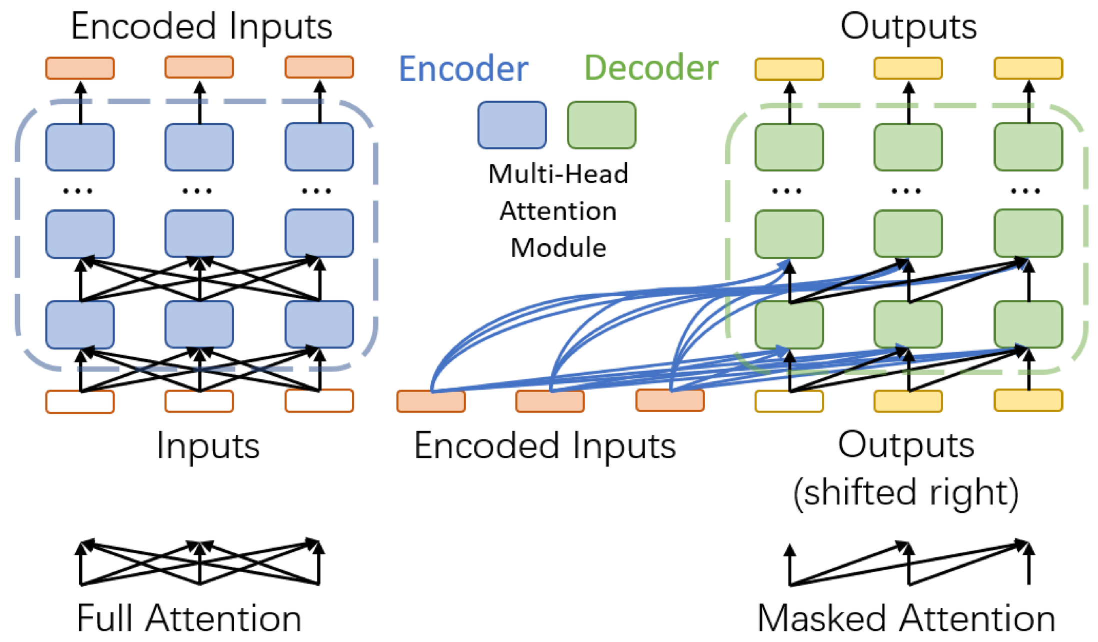
▷ **Recurrent Network based Encoder-Decoder:**

1. Computationally Expensive (Recursive Nature)
2. Fail to Capture Long-term Dependency
3. Various Training Issues (e.g. Gradient Exploding/Vanishing)

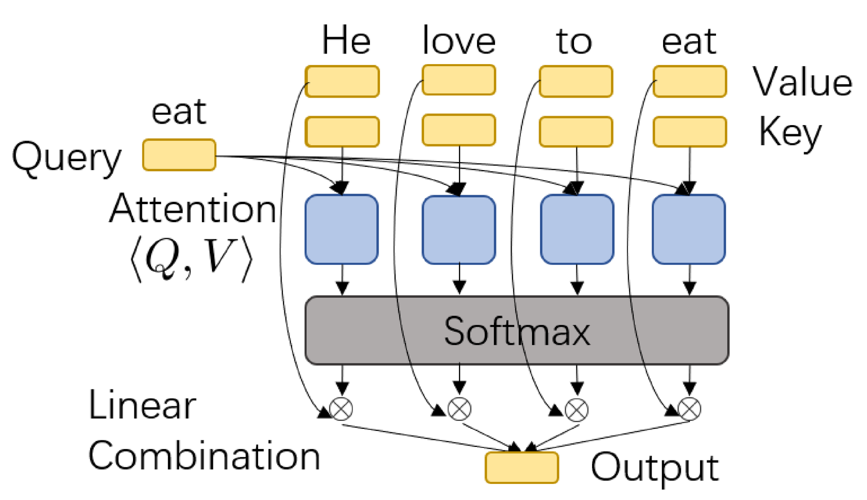


▷ **Transformer Models** (Vaswani et al. 2017)

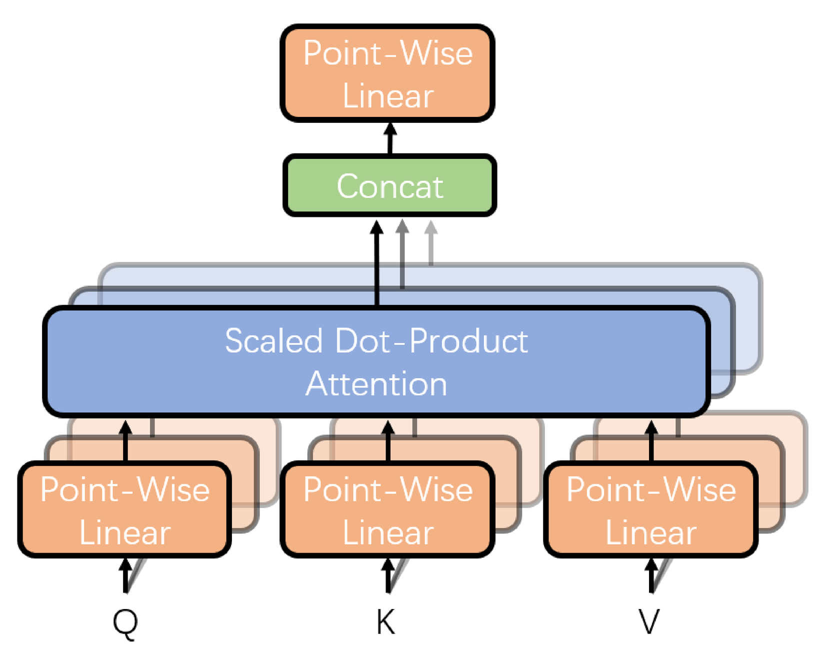
- Feedforward Network based Encoder-Decoder:



- Attention (Bahdanau et al. 2014):



- Multi-Head Attention:



▷ **Benefits of Transformer Models**:

1. High Efficiency (Parallel and Feedforward Structures)
2. Capture Long-term Dependency (Attention Module)
3. Enable Deeper Representation Learning

Proposed Method

▷ **Word-Level Domain Proportion**

$$\mathcal{D}(x) = (1-\epsilon)\cdot \mathrm{softmax}(Rx) + \epsilon/k.$$

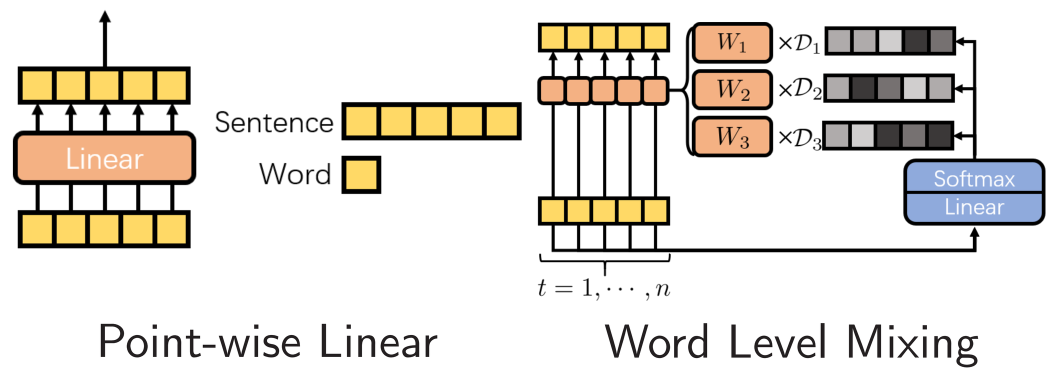
- $\epsilon \in (0,1)$: Smoothing parameter.
- k : Number of domains
- x : Word vector.
- R : Weight matrix of the softmax layer.

▷ **Word Level Adaptive Domain Mixing**

For a given word vector x_{in}, we take the weighted averaging of the point-wise linear transformations from different domains based on the domain proportion.

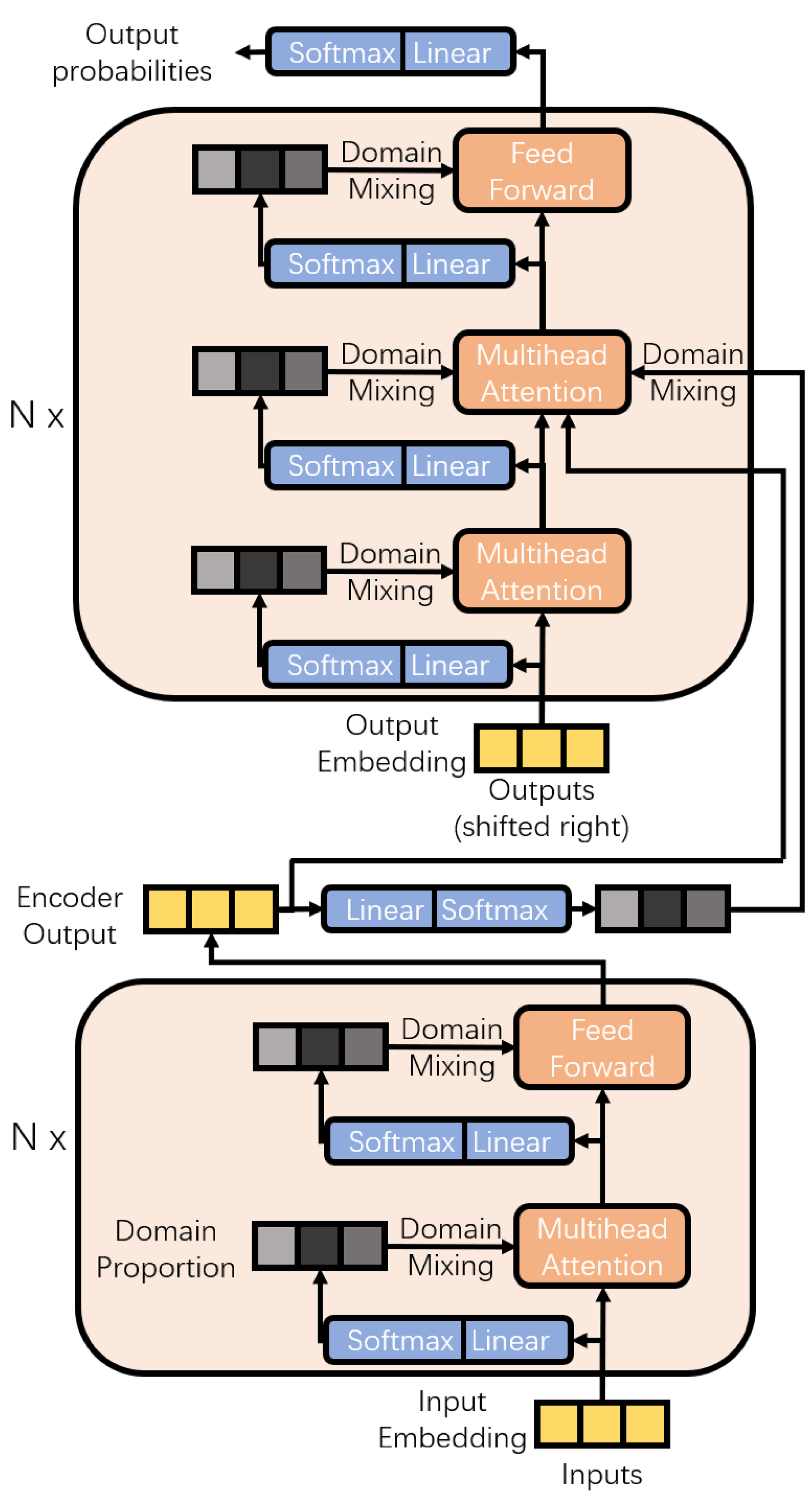
$$x_{\text{out}} = W x_{\text{in}} \quad \Rightarrow \quad x_{\text{out}} = \textstyle\sum_{j=1}^{k} \mathcal{D}_j(x_{\text{in}}) W_j x_{\text{in}}$$

- W_j : Weight matrix for the j-th domain.
- $\mathcal{D}_j(x_{\text{in}})$: The proportion of x_{in} for the j-th domain.



▷ **Layer-wise Domain Mixing**

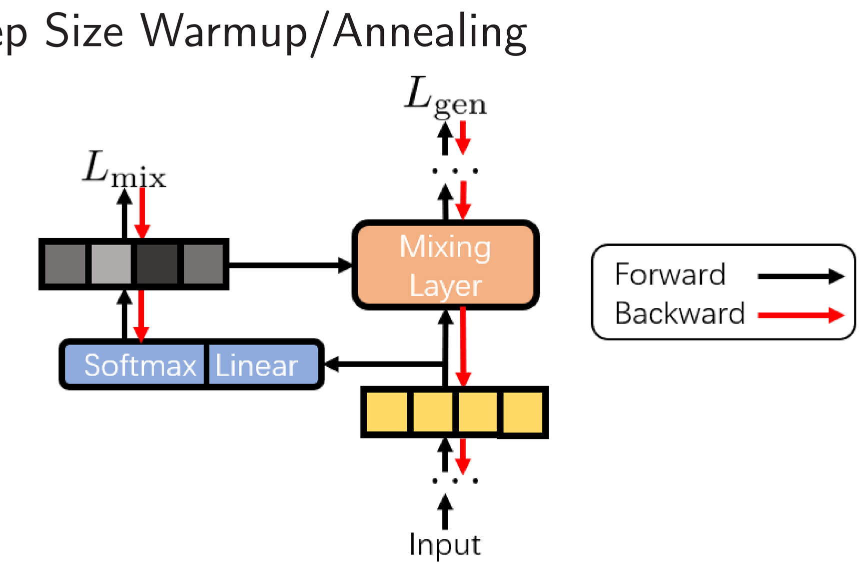
We apply the word-level domain mixing to the transformer at all layers. Note that the domain proportion is **different** at each attention layer.



▷ **Training**

$$\min_{\mathcal{H},\mathcal{D},\mathcal{F}} L^* = L_{\text{gen}}(\mathcal{H},\mathcal{D},\mathcal{F}) + L_{\text{mix}}(\mathcal{H},\mathcal{D},\mathcal{F})$$

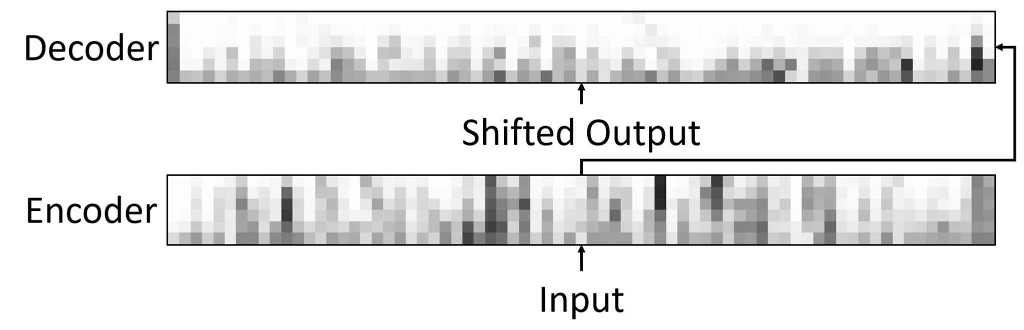
- $\mathcal{H}$: Encoder Module
- $\mathcal{F}$: Decoder Module
- $\mathcal{D}$: Domain Proportion
- L_{gen} : Cross Entropy Loss for Translation
- L_{mix} : Cross Entropy Loss over the Smoothed Domain (Hard) Labels of Words
- Training Algorithm: ADAM + Layer Normalization + Step Size Warmup/Annealing



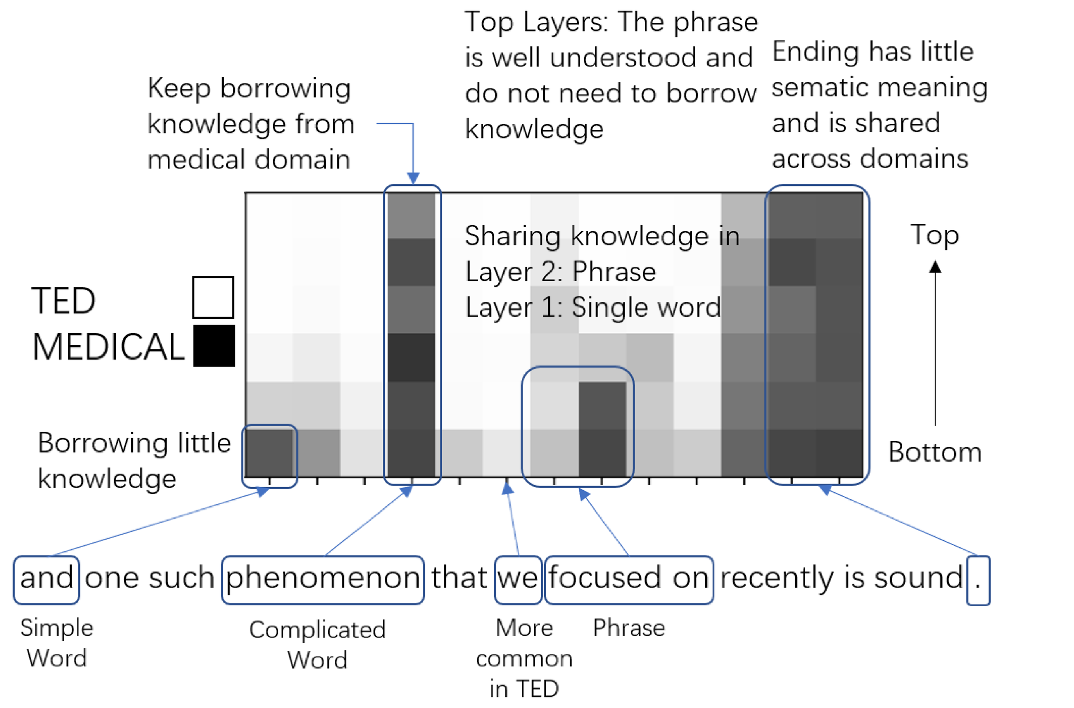
Experiment – Domain Proportion

▷ **Domain Proportion Visualization:**

TED Domain (white) vs. Medical Domain (black)

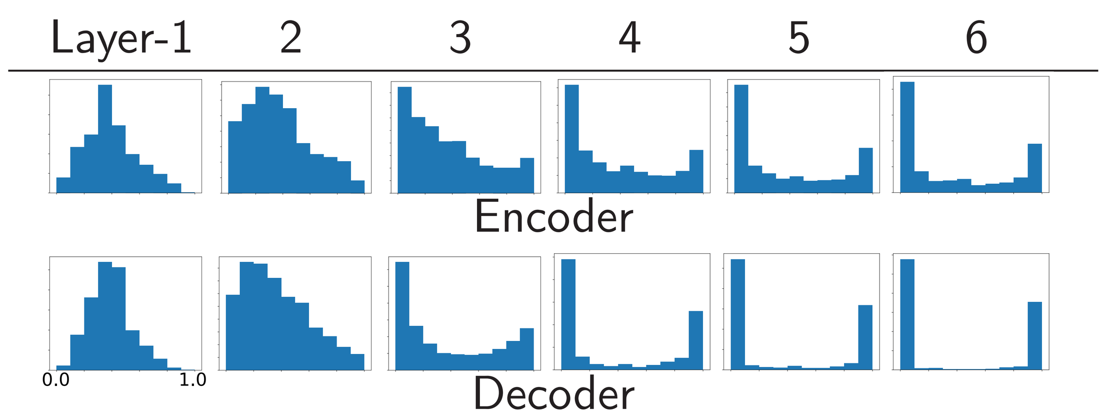


▷ **Word-Level Analysis:**



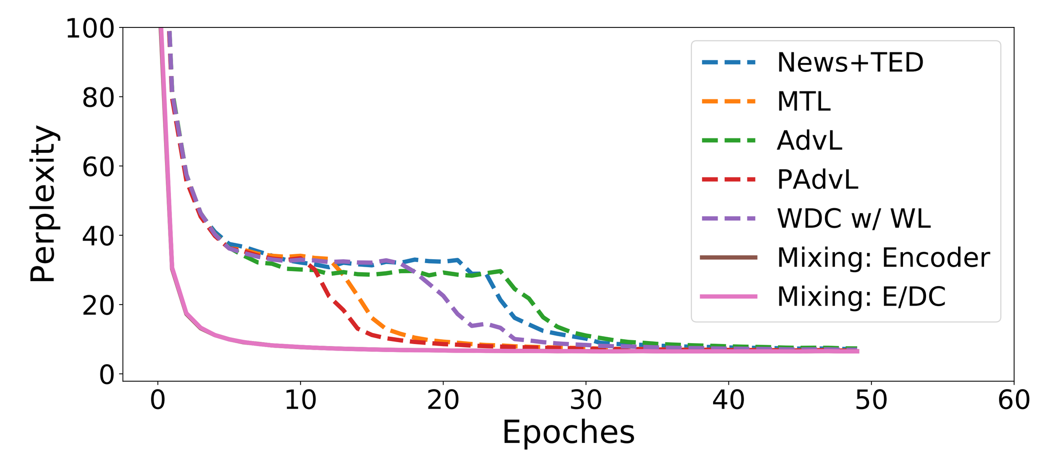
▷ **Histograms of the Domain Proportions**

Within each histogram, 0 means Medical domain, and 1 means TED domain.



Experiment – Translation

▷ **Training Perplexity:**



▷ **Testing BLEU Scores:**

- English to German

Method	News	TED
Direct Training		
News	26.09	6.15
TED	4.90	29.09
News + TED	26.06	28.11
Embedding based Methods		
MTL	26.90	29.27
AdvL	25.68	27.46
PAdvL	27.06	29.49
WDC + WL	27.25	29.43
Our Domain Mixing Methods		
Encoder	**27.78**	30.30
Encoder + WL	27.67	30.11
E/DC	27.58	**30.33**
E/DC + WL	27.55	30.22

- English to French

Method	TED	Medical
Direct Training		
TED	28.22	7.32
Medical	7.03	53.73
Medical + TED	39.21	53.40
Embedding based Methods		
MTL	39.14	53.37
AdvL	39.54	53.46
PAdvL	39.56	53.23
WDC + WL	39.79	53.85
Our Domain Mixing Methods		
Encoder	40.30	54.05
Encoder + WL	40.43	54.14
E/DC	40.52	54.28
E/DC + WL	**40.60**	**54.39**

- Chinese to English

Method	Laws	News	Speech	Thesis
Direct Training				
Laws	51.98	3.80	2.38	2.64
News	6.88	31.99	8.12	4.17
Speech	3.33	4.90	18.63	3.08
Thesis	5.90	5.55	4.77	11.06
Mixed	48.87	26.92	16.38	12.09
Embedding based Methods				
MTL	49.14	27.15	16.34	11.80
AdvL	48.93	26.51	16.18	12.08
PAdvL	48.72	27.07	15.93	**12.23**
WDC + WL	42.16	25.81	15.29	10.14
Our Domain Mixing Methods				
Encoder	50.21	27.94	16.85	12.03
Encoder + WL	50.11	27.48	16.79	11.93
E/DC	**50.64**	**28.48**	17.41	11.71
E/DC + WL	50.04	28.17	**17.60**	11.59