Bridging the Domain Gap: Improve Informal Language Translation via Counterfactual Domain Adaptation

Ke Wang,[1,2] **Guandan Chen,**[3] **Zhongqiang Huang,**[3] **Xiaojun Wan,**[1,2] **Fei Huang**[3]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]The MOE Key Laboratory of Computational Linguistics, Peking University
[3]DAMO Academy, Alibaba Group
{wangke17, wanxiaojun}@pku.edu.cn {guandan.cgd, z.huang, f.huang}@alibaba-inc.com

Abstract

Despite the near-human performances already achieved on formal texts such as news articles, neural machine translation still has difficulty in dealing with ”user-generated” texts that have diverse linguistic phenomena but lack large-scale high-quality parallel corpora. To address this problem, we propose a counterfactual domain adaptation method to better leverage both large-scale source-domain data (formal texts) and small-scale target-domain data (informal texts). Specifically, by considering effective counterfactual conditions (the concatenations of source-domain texts and the target-domain tag), we construct the counterfactual representations to fill the sparse latent space of the target domain caused by a small amount of data, that is, bridging the gap between the source-domain data and the target-domain data. Experiments on English-to-Chinese and Chinese-to-English translation tasks show that our method outperforms the base model that is trained only on the informal corpus by a large margin, and consistently surpasses different baseline methods by +1.12 $\sim$ 4.34 BLEU points on different datasets. Furthermore, we also show that our method achieves competitive performances on cross-domain language translation on four language pairs.

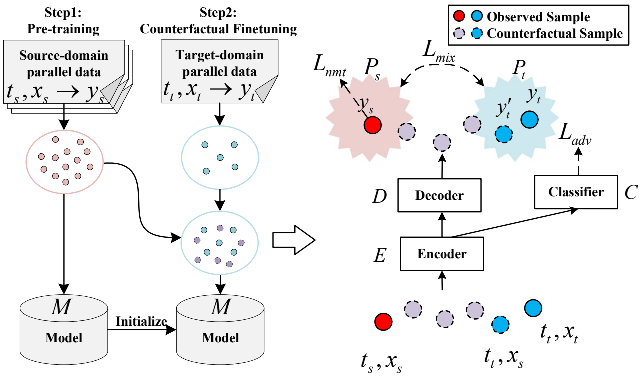
Introduction

With the rapid development of economic globalization and the Internet, machine translation of informal texts can facilitate the communication between people who speak different languages in the world, especially in messaging applications (Messenger, Whatsapp, iMessage, Wechat, DingTalk), content sharing on social media (Facebook, Instagram, Twitter), and discussion forums (Reddit) (Pennell and Liu 2014). However, the lack of large-scale high-quality parallel corpora of informal texts brings challenges to neural machine translation (NMT) (Koehn and Knowles 2017; Chu and Wang 2018; Hu et al. 2019). The existing small-scale parallel corpora of informal texts make NMT models easy to overfit on the small training set and cannot generalize well on the test set. Although NMT has achieved near-human performances on formal texts such as news articles, some researches (Hu et al. 2019) reveal that NMT is sensitive to domain shift and the performance is bounded by the similarity between training data (e.g., formal texts) and test data (e.g., informal texts).

A natural solution to this problem is domain adaptation, which aims to adapt NMT models trained on one or more source domains (formal texts in our case) to the target domain (informal texts in our case). Domain adaptation methods can be roughly categorized into model-based and data-based ones (Chu and Wang 2018). The model-based methods make explicit changes to NMT model such as designing new training objectives (Luong, Pham, and Manning 2015; Chu, Dabre, and Kurohashi 2017), modifying the model architecture (Kobus, Crego, and Senellart 2017; Johnson et al. 2017; Dou et al. 2019), and changing the decoding method (Khayrallah et al. 2017). In contrast, the data-based methods focus on the data being used, such as using monolingual corpus (Gülçehre et al. 2015), combining source-domain and target-domain parallel corpora (Johnson et al. 2017; Caswell, Chelba, and Grangier 2019; Marie, Rubino, and Fujita 2020), dynamic source-domain data selection from easy to complex for target-domain translation (van der Wees, Bisazza, and Monz 2017), etc. Nevertheless, most of these methods train the model based on the observed data, which depends on the quality of corpora used.

To improve target-domain translation (e.g, informal texts) that lacks large-scale high-quality parallel corpora, we explore the use of counterfactual thinking (Swaminathan and Joachims 2015; Pearl and Mackenzie 2018) to augment the training set of models with counterfactual representations that do not exist in the observed data but are useful for improving the target-domain translation. Specifically, we propose a model-agnostic counterfactual domain adaptation (**CDA**) method, which first pre-trains the NMT model on the large-scale source-domain parallel corpus (e.g., formal texts) with artificial domain tags to indicate specific domains, and then fine-tune on a mixture of the source-domain and target-domain parallel corpora (i.e., informal texts) with a counterfactual training strategy. In the counterfactual training strategy, we use the concatenations of source-domain texts and the target-domain tag (i.e, counterfactual conditions) to construct counterfactual representations, and propose three objectives to use both observed and counterfactual data.

Our proposed method is motivated by the observation that not every training sample in the source-domain parallel corpus (e.g., formal texts) is equally useful for target-domain translation (e.g., informal texts), due to the difference be-



(a) Mixed Fine-Tuning Architecture

(b) Illustration

Figure 1: (a) shows the architecture of counterfactual domain adaptation (**CDA**), which first pre-trains the NMT model on a large external source-domain parallel corpus $\mathbb{D}_s$ (e.g., formal texts), and then uses our counterfactual training strategy to fine-tune on a mixture of the source-domain and target-domain parallel corpora. (b) is the illustration of our counterfactual training strategy, where P_s and P_t denote the empirical distributions of $\mathbb{D}_s$ and the small target-domain corpus $\mathbb{D}_t$, respectively. By constructing counterfactual representations based on the concatenations $(t_t, \boldsymbol{x}_s)$ of source-domain text and the target-domain tag (i.e., counterfactual conditions), **CDA** aims to fill the sparse latent space of the target-domain data, thereby improving target-domain translation.

tween the source domain and the target domain. Therefore, in addition to reweighting the observed samples according to their usefulness (i.e, the similarity with the target domain), we construct counterfactual representations to fill the sparse latent space of the target domain caused by the small amount of data, thereby improving target-domain translation by bridging the domain gap between two data distributions.

We compare our method with several advanced baseline methods, covering fine-tuning methods (Luong, Pham, and Manning 2015; Chu, Dabre, and Kurohashi 2017), domain-aware methods (Johnson et al. 2017; Caswell, Chelba, and Grangier 2019; Dou et al. 2019; Marie, Rubino, and Fujita 2020), data augmentation methods (Miyato, Dai, and Goodfellow 2017; Cheng et al. 2020) and domain-adaptation methods (Bapna and Firat 2019; Chu, Dabre, and Kurohashi 2017). For informal language translation, we conduct experiments on two language pairs: English-to-Chinese and Chinese-to-English. We collect the source-domain corpus from multiple sources of formal texts (e.g., news, internet, movies, encyclopedias, government, news dialogue, novels, technical documents and politics) and use the IWSLT-2017 dataset (Cettolo et al. 2017) as the target-domain corpus (i.e., informal texts). Experimental results show that our method outperforms the base model that is trained only on the target-domain corpus by up to +7.93 BLEU points, and consistently surpasses baseline methods by +1.12∼4.34 BLEU points on different test sets, demonstrating the effectiveness and robustness of our proposed method. Encouragingly, we also find that our method achieves competitive performances on the other four cross-domain language translation tasks, including English-to-Chinese, Chinese-to-English, German-to-English and English-to-French.

In summary, our main contributions are:

- Our study demonstrates the promising future of improving informal language translation by using counterfactual samples beyond observable data.
- We propose a counterfactual domain adaptation (**CDA**) method to improve target-domain translation by filling the sparse latent space of the target domain with constructed counterfactual representations.
- Experimental results show that our model consistently outperforms several baseline methods on both informal language translation and cross-domain language translation.

Method

In order to improve target-domain translation (e.g., informal texts) that lacks large-scale high-quality parallel corpus, we propose a counterfactual domain adaptation (**CDA**) method for neural machine translation. The overall architecture of our method is depicted in Figure 1.

Architecture

Formally, for the large-scale source-domain parallel corpus $\mathbb{D}_s$ and the small-scale target-domain parallel corpus $\mathbb{D}_t$, the empirical distributions are P_s and P_t respectively. Each sample is a pair of sentences belonging to different languages (i.e., $(\boldsymbol{x}_s, \boldsymbol{y}_s)$ or $(\boldsymbol{x}_t, \boldsymbol{y}_t)$). We augment all data with the artificial tag (i.e., t_s and t_t) to indicate its specific domain, that is, the input sentence is a tag-text concatenation $(t, \boldsymbol{x})$. The NMT task (Bahdanau, Cho, and Bengio 2015; Gehring et al. 2017; Vaswani et al. 2017) seeks to model the translation probability $P(\boldsymbol{y}|t, \boldsymbol{x})$ based on the encoder-decoder (E and D) paradigm, where the decoder D in the NMT model acts as a conditional language model that operates on a shifted copy of $\boldsymbol{y}$. Here, the tag-text concatenation is the input sequence and $\boldsymbol{y}$ is the output sequence, denoted as,

$$t, \boldsymbol{x} = (t, x_0, \cdots, x_{|\boldsymbol{x}|-1}); \quad \boldsymbol{y} = (y_0, \cdots, y_{|\boldsymbol{y}|-1}). \quad (1)$$

In practice, we add start and end symbols at both ends of the sequence. We apply a mixed fine-tuning architecture (Luong and Manning 2015; Freitag and Al-Onaizan 2016; Johnson et al. 2017; Chu, Dabre, and Kurohashi 2017; Dabre, Fujita, and Chu 2019) to prevent the model over-fitting on the small-scale target-domain corpus, and it works in two steps (shown in Figure 1 (a)):

1) Train an NMT model M on the large source-domain data by minimizing the cross-entropy loss as follows:

$$\mathcal{L}_{pre-train}(\theta_m) = \mathbb{E}_{(\boldsymbol{x}_s, \boldsymbol{y}_s)\sim P_s}[\ell(D(E(t_s, \boldsymbol{x}_s)), \boldsymbol{y}_s)], \quad (2)$$

where ℓ is the cross entropy loss (de Boer et al. 2005) and θ_m is the parameters of the NMT model M.

2) Fine-tune M on a mixture of the source-domain and target-domain parallel corpora based on our counterfactual

training strategy (we will detail it in the next subsection). Due to the imbalance between the source-domain corpus and the target-domain corpus, we over-sample the target-domain data for faster convergence.

Note that we do not need any modifications to the NMT model, and here it is the transformer model (Vaswani et al. 2017) and can be easily extended to other models, such as LSTM (Sutskever, Vinyals, and Le 2014), Fconv-seq2seq (Gehring et al. 2017).

Training

Due to the difference between the source-domain data and the target-domain data, not all instance is equally useful for target-domain translation. Some previous studies have demonstrated the importance of considering the usefulness of training samples in domain adaptation, such as instance weighting (Luong, Pham, and Manning 2015; Chu, Dabre, and Kurohashi 2017), dynamic data selection and curriculum learning (van der Wees, Bisazza, and Monz 2017; Zhang et al. 2019; Caswell, Chelba, and Grangier 2019; Marie, Rubino, and Fujita 2020). However, reweighting the observed samples does not sufficiently take advantages of the use of training data. The core idea behind our method is to construct counterfactual representations that do not exist in the observed data but are more useful for improving the target-domain translation.

Counterfactual samples (Swaminathan and Joachims 2015; Pearl and Mackenzie 2018) allow us to imagine a hypothetical reality that contradicts the observed facts (samples), and have been used to provide explanations for predictions (Feder et al. 2020) and increase the robustness (Kusner et al. 2017; Garg et al. 2019). In this paper, we use the domain tag to explicitly add domain-aware features, and construct counterfactual representations $E(t_t, \boldsymbol{x}_s)$ based on the concatenations of the target-domain tag t_t and source-domain texts $\boldsymbol{x}_s$ that do not exist in the training samples.

Note that we do not have the golden translation output (denoted as $\boldsymbol{y}_t'$) of $(t_t, \boldsymbol{x}_s)$, but it should be different from $\boldsymbol{y}_s$ because it needs to conform to the characteristics of the target domain. Specifically, our **CDA** method constructs counterfactual representation based on the sentence pair $t_t, \boldsymbol{x}_s \rightarrow \boldsymbol{y}_t'$, and consists of the following three objectives:

1) Domain-Aware Adversarial Loss: Due to the lack of the golden translation output $\boldsymbol{y}_t'$ of $t_t, \boldsymbol{x}_s$, we hope to use the adversarial network (Goodfellow et al. 2014) to identify features related to the target domain, and generate counterfactual representations that conform to the target domain under the given target-domain tag.

Specifically, we train a domain classifier with observed samples, and use this classifier as the training objective that the representation of the target-domain tag encoded by E needs to conform to the target domain. Specifically, we use an additional classifier C to predict whether the domain tag t matches the text $\boldsymbol{x}$ based on the mean value of the encoded representation of each word in the sentence. C is a simple multi-layer perceptron (MLP) network whose output is a scalar probability between 0 and 1, and the goal is to minimize the following:

$$\begin{aligned}\mathcal{L}_{dis}(\theta_c) = & \\ & \mathbb{E}_{(\boldsymbol{x}_s,\boldsymbol{y}_s)\sim P_s}[C(E(t_t,\boldsymbol{x}_s)) + (1 - C(E(t_s,\boldsymbol{x}_s)))] \\ & + \mathbb{E}_{(\boldsymbol{x}_t,\boldsymbol{y}_t)\sim P_t}[C(E(t_s,\boldsymbol{x}_t)) + (1 - C(E(t_t,\boldsymbol{x}_t)))],\end{aligned} \quad (3)$$

where θ_c is the parameters of C. Note that our classifier is to judge whether the tag and the text match instead of directly predicting the domain, which is more suitable for the situation where the data sets are not balanced (the size of source-domain data is much bigger than that of target-domain data).

On the contrary, our encoder E aims to generate target-domain representations in the view of the classifier C, as long as the given tag is the target-domain tag t_t. Therefore, the domain-aware adversarial loss $\mathcal{L}_{adv}$ of the NMT model M is:

$$\mathcal{L}_{adv}(\theta_m) = \mathbb{E}_{(\boldsymbol{x}_s,\boldsymbol{y}_s)\sim P_s}[1 - C(E(t_t,\boldsymbol{x}_s))] \quad (4)$$

Unlike the previous adversarial losses (Ganin et al. 2016; Britz, Le, and Pryzant 2017; Zeng et al. 2018; Wang and Wan 2018, 2019; Wang, Hua, and Wan 2019; Gu, Feng, and Liu 2019; Wang and Wan 2020) that want to disentangle domain-invariant and domain-specific information but may risk losing content information, the use of our domain-aware adversarial loss not only makes the encoder E more sensitive to the given domain tag, but also guides the model to explore the latent space of the target-domain distribution, so as to make up for the insufficient coverage (of diverse linguistic phenomena of informal texts) on the small-scale target-domain corpus.

2) Source-Side Mixup Loss: Compared with the domain-aware adversarial loss that indirectly affects the encoder through the adversarial process, the source-side mixup loss directly constructs counterfactual interpolations (representations) between $t_s, \boldsymbol{x}_s \rightarrow \boldsymbol{y}_s$ and $t_t, \boldsymbol{x}_s \rightarrow \boldsymbol{y}_t'$ to train the model.

Mixup is a kind of data augmentation technique that has been proven to improve generalization in the image classification task (Zhang et al. 2018). Given pairs of images (x_1, y_1) and (x_2, y_2), where x_1, x_2 denote the RGB pixels of the input images and y_1, y_2 are their one-hot labels respectively, Mixup chooses a random mixing proportion λ from a Beta distribution $\beta(\alpha, \alpha)$ controlled by the hyper-parameter α, and creates an artificial training example $(\lambda x_1 + (1-\lambda)x_2, \lambda y_1 + (1-\lambda)y_2)$ to train the network by minimizing the loss on mixed-up data points:

$$\begin{aligned}\mathcal{L}(\theta) = & \mathbb{E}_{x_1,y_1\sim p_{\mathcal{D}}}\mathbb{E}_{x_2,y_2\sim p_{\mathcal{D}}}\mathbb{E}_{\lambda\sim\beta(\alpha,\alpha)} \\ & [\ell(\lambda x_1 + (1-\lambda)x_2, \lambda y_1 + (1-\lambda)y_2)] \\ = & \mathbb{E}_{x_1,y_1\sim p_{\mathcal{D}}}\mathbb{E}_{x_2,y_2\sim p_{\mathcal{D}}}\mathbb{E}_{\lambda\sim\beta(\alpha,\alpha)}\mathbb{E}_{z\sim Ber(\lambda)}[z\ell(\lambda x_1 \\ & + (1-\lambda)x_2, y_1) + (1-z)\ell(\lambda x_1 + (1-\lambda)x_2, y_2)] \\ \Rightarrow & \mathbb{E}_{x_1,y_2\sim p_{\mathcal{D}}}\mathbb{E}_{x_2\sim p_{\mathcal{D}}}\mathbb{E}_{\lambda\sim\beta(\alpha+1,\alpha)}\ell(\lambda x_1 + (1-\lambda)x_2, y_1).\end{aligned} \quad (5)$$

Ber represents the Bernoulli distribution. We show the detailed proof of Eq 5 in the Appendix. With the help of Eq 5, we no longer need the blending $(\lambda y_1 + (1-\lambda)y_2)$ of labels y_1 and y_2 under the condition that λ is drawn from

$\beta(\alpha+1,\alpha)$, which is convenient for the situation where we do not have $\boldsymbol{y}_t'$. We mix the following two data points:

$$x_1 = E(t_s, \boldsymbol{x}_s); \quad y_1 = \boldsymbol{y}_s; \qquad (6)$$
$$x_2 = E(t_t, \boldsymbol{x}_s); \quad y_2 = \boldsymbol{y}_t';$$

Therefore, according to Eq 5 and Eq 6, our source-side mix-up loss minimizes the constructed counterfactual interpolations loss from a vicinity distribution (Chapelle et al. 2000) P_v defined in the representation space:

$$\mathcal{L}_{mix}(\theta_m) = \mathbb{E}_{(\boldsymbol{x}_s,\boldsymbol{y}_s)\sim P_s}\mathbb{E}_{\lambda\sim\beta(\alpha+1,\alpha)} \qquad (7)$$
$$[\ell(D(\lambda E(t_s,\boldsymbol{x}_s) + (1-\lambda)E(t_t,\boldsymbol{x}_s)), \boldsymbol{y}_s)].$$

We believe that the advantages of such settings are: 1) We construct a novel vicinal distribution P_v to bridge the generalization gap between P_s and P_t by turning the source-domain distribution into a distribution that is closer to the target-domain distribution, and augment the training data of the model with these constructed counterfactual representations. 2) By linearly interpolating between the source-domain and target-domain representations, we incentivize the network to act smoothly and kind of interpolate nicely between domains - without sharp transitions.

3) Sample-Wise Weighted Loss: Here we directly reweight observed source-domain samples based on the usefulness (i.e, probability of belonging to the target-domain distribution P_t) provided by the classifier C. We assign a sample weight $w_{x_s} = C(E(t_t, \boldsymbol{x}_s))$ for each source-domain sample:

$$\mathcal{L}_{nmt}(\theta_m) = \mathbb{E}_{(\boldsymbol{x}_s,\boldsymbol{y}_s)\sim P_s}\mathbb{E}_{(\boldsymbol{x}_t,\boldsymbol{y}_t)\sim P_t} \qquad (8)$$
$$[w_{x_s} * \ell(D(E(t_s,\boldsymbol{x}_s)),\boldsymbol{y}_s) + \ell(D(E(t_t,\boldsymbol{x}_t)),\boldsymbol{y}_t)]$$

Finally, the overall training objective in our counterfactual training is a combination of the three losses:

$$\theta_m^* = \text{argmin}_{\theta_m}\{\mathcal{L}_{adv}(\theta_m) + \mathcal{L}_{mix}(\theta_m) + \mathcal{L}_{nmt}(\theta_m)\}. \qquad (9)$$

To improve the translation for a data-scarce target domain (informal), we proposed a domain adaptation method based on building counterfactual examples. The application of adversarial training and MIXUP is to indirectly and directly fill the sparse latent space of the target domain, so as to bridge the gap between source and target domains. Our adversarial training is to indirectly guide the model to explore the latent space of the target-domain distribution by making the encoder E more sensitive to the given domain tag, rather than disentangling domain-invariant and domain-specific information like most previous adversarial methods. The use of MIXUP is aiming to directly smooth the space between source and target domains by constructing counterfactual interpolations, which is different from the previous MIXUP methods of interpolating between any two observable samples. Note that although the adversarial samples (Goodfellow, Shlens, and Szegedy 2015) are new samples that do not exist in the data set like ours, the difference is that our counterfactual samples are designed to help the model use counterfactual conditions instead of deceiving it (Wang, Hua, and Wan 2019). We provide the algorithm of the entire training process in the Appendix.

In summary, by constructing counterfactual representations based on the concatenations of source-domain texts and the target-domain tag (i.e, counterfactual conditions), **C-DA** aims to: 1) guide the model to explore the latent space of the target-domain distribution, and 2) turn the source-domain distribution into a one that is closer to the target-domain distribution so as to bridge the gap between two domains. Overall, **CDA** improves the target-domain translation by fill the sparse latent space of the target domain caused by a small amount of data.

Experiments

Setup

Without loss of generality, we evaluate our method on informal language translation tasks for English-to-Chinese (En → Zh) and Chinese-to-English (Zh → En), and apply our method on four cross-domain language translation tasks, including English-to-Chinese (En → Zh), Chinese-to-English (Zh → En), German-to-English (De→En) and English-to-French (En→Fr).

Implementation details: We implement our method on top of the Transformer-base (Vaswani et al. 2017) implemented in Fairseq (Ott et al. 2019). The size of the hidden unit is 512 and the number of the attention heads is 4. Both the encoder and decoder have 6 layers. We apply byte-pair-encoding (BPE) vocabulary (Sennrich, Haddow, and Birch 2016) with 40k merge operations to alleviate the out-of-vocabulary problem. The beam size of the beam search is set to 5. We set α in $\beta(\alpha+1,\alpha)$ to 0.1 and both d-steps and g-steps in the algorithm are set to 1. We tokenize English, German and French sentences using MOSES script (Koehn et al. 2007), and perform word segmentation on Chinese sentences using Jieba segmentation tool. The training takes 2 days on 8 Tesla P100 GPUs.

Datasets: For informal language translation, we collect source-domain corpus containing 25,136,557 sentence pairs from multiple sources that contain formal texts, including:

- **New Commentary v12** (Bojar et al. 2016) contains 227,062 sentence pairs and its genre is news.
- **CWMT** (Chen and Zhang 2019) contains 9,023,454 sentence pairs and its genres are internet, movies, encyclopedias, government, news dialogue, novels and technical documents.
- **UN** (Ziemski, Junczys-Dowmunt, and Pouliquen 2016) contains 15,886,041 sentence pairs and its genre is politics.

We use the IWSLT2017 dataset (Cettolo et al. 2017) as our informal text corpus (i.e., target-domain corpus), which comes from the subtitles of TED talks and can be regarded as a kind of informal spoken-style corpus. It contains a training set of size 231,266 sentence pairs, a validation set of size 879 sentence pairs, and 5 test sets of sizes 1,557 (Test2010),

Train@	Method	Test2010	Test2011	Test2012	Test2013	Test2014	Test2015	Average	△
Target-domain corpus	**Base**	17.67	22.54	21.83	21.41	20.10	24.76	21.39	-
	AdvEmb	17.89	22.98	22.01	22.05	20.41	24.99	21.72	+0.33
	AdvAug	18.05	23.12	22.90	22.03	21.45	25.65	22.20	+0.81
Joint corpus	**Source**	21.68	27.68	25.62	27.08	25.57	30.66	26.38	+4.99
	Joint	22.26	28.87	26.65	28.21	25.71	31.52	27.20	+5.81
	FT	22.21	28.32	26.78	27.33	25.08	30.72	26.74	+5.35
	TAG	22.49	29.24	27.17	28.42	25.98	31.40	27.45	+6.06
	ADAP	22.23	28.04	27.10	27.51	25.12	30.58	26.76	+5.37
	DAFE	22.51	28.94	27.23	28.10	25.86	31.33	27.33	+5.94
	Mixed FT	22.91	29.01	27.41	29.16	26.32	31.25	27.67	+6.28
	AdvAug	22.86	28.91	26.81	27.19	25.16	31.43	27.06	+5.67
	CDA	**24.26**	**30.15**	**29.03**	**30.87**	**28.15**	**33.51**	**29.32**	**+7.93**
	CDA w/o $\mathcal{L}_{mix}$	23.62	29.34	27.41	28.51	26.27	32.19	27.89	+6.50
	CDA w/o $\mathcal{L}_{adv}$	23.97	29.97	28.19	29.15	27.56	32.72	28.59	+7.20
	CDA w/o $\mathcal{L}_{nmt}$	24.16	30.05	28.41	30.16	27.64	33.16	28.93	+7.54

Table 2: Results of informal English-to-Chinese translation. ”w/o” means ”without”.

Task	Domain	#Train	#Valid	#Test
En↔Zh	Thesis	0.29M	1,000	625
	Education	0.21M	1,000	456
	Spoken	0.22M	1,000	455
De→En	Koran	0.54M	1,000	1,000
	IT	0.35M	1,000	1,000
En→Fr	Medical	0.89M	800	2000
	Parliament	2.04M	800	2000

Table 1: Statistics of cross-domain training corpora.

1,426 (Test2011), 1,692 (Test2012), 1,372 (Test2013), 1,297 (Test2014), and 1,205 (Test2015). For cross-domain language translation, we conduct experiments on three different corpora, as listed in Table 1. For Zh→En and En→Zh translation tasks, we use the previous formal text corpus as the large source-domain data, and the three domains in UM-Corpus (Tian et al. 2014) as the target-domain data, as shown in Table 1. Note that the genres of three domains we specifically selected are quite different from the ones of the source-domain corpus. For De → En and En → Fr translation tasks, we use data extracted from OPUS (Tiedemann 2012). They respectively contain two domains, specifically, Koran and information technology domains for De → En, medical and parliament domains for En → Fr. Note that for these two translation tasks, we use one domain as the target domain and the other one as the source domain.

We select the best NMT model according to the validation set of the target-domain corpus in the training process, and report the BLEU (Papineni et al. 2002) scores with SacreBLEU[1] (Post 2018) on test sets.

Baseline Comparisons: Our baselines include the following four categories: 1) fine-tuning methods (i.e, Fine-tuning (Luong, Pham, and Manning 2015), Mixed fine-tuning (Chu, Dabre, and Kurohashi 2017)), 2) domain-aware methods (i.e, TAG (Johnson et al. 2017; Marie, Rubino, and Fujita 2020; Caswell, Chelba, and Grangier 2019), DAFE (Dou et al. 2019)), 3) data augmentation methods (i.e, AdvEmb (Miyato, Dai, and Goodfellow 2017), AdvAug (Cheng et al. 2020)), 4) domain-adaptation methods (i.e, ADAP (Bapna and Firat 2019) (Chu, Dabre, and Kurohashi 2017)). For a fair comparison, we implement all these methods using the Transformer-base (Vaswani et al. 2017) backbone and report results trained on the same corpora.

* **FT**: Luong, Pham, and Manning (2015) first train a NMT model on source-domain corpus, and then fine-tune the model on target-domain corpus.

* **Mixed FT**: Chu, Dabre, and Kurohashi (2017) extend the fine-tuning approach by training on source-domain data, and then fine-tuning on source-domain and target-domain data.

* **AdvEmb**: Miyato, Dai, and Goodfellow (2017) provide a virtual adversarial training method for the label (decoded words in our case) by applying perturbations to the word embeddings, which is less prone to overfitting.

* **TAG**: Johnson et al. (2017); Caswell, Chelba, and Grangier (2019); Marie, Rubino, and Fujita (2020) provide an domain-aware method that introduces domain tag to the source sentence.

* **ADAP**: Bapna and Firat (2019) propose an efficient domain adaptation method that consists of injecting tiny task specific adapter layers into a pre-trained NMT model.

* **DAFE**: Dou et al. (2019) propose a domain-aware approach that adapts models with domain-aware feature embeddings, which are learned via an auxiliary language modeling task.

* **AdvAug**: Cheng et al. (2020) provide an adversarial augmentation method to minimize the vicinal risk over virtual sentences sampled from vicinity distributions for adversarial sentences that describes a smooth interpolated embedding space centered around observed training sentence pairs.

For analysis, we also show training NMT models with different corpora, including:

[1] 4BLEU + case.mixed + lang.LANGUAGE PAIR + numrefs.1 + smooth.exp + test.SET + tok.intl + version.1.2.15

Train@	Method	Test2010	Test2011	Test2012	Test2013	Test2014	Test2015	Average	△
Target-domain corpus	**Base**	21.78	24.37	22.93	23.21	21.62	24.07	23.00	-
	AdvEmb	22.33	25.80	23.95	23.57	21.48	24.11	23.54	+0.54
	AdvAug	22.49	26.02	24.14	24.80	21.53	25.23	24.04	+1.04
Joint corpus	**Source**	19.27	21.71	21.17	24.90	21.76	25.06	22.31	-0.69
	Joint	22.37	26.81	25.11	28.07	24.61	28.13	25.85	+2.85
	FT	25.18	25.71	22.41	25.72	24.40	26.78	25.03	+2.03
	TAG	25.01	28.55	27.78	30.42	25.90	28.67	27.72	+4.72
	ADAP	24.89	25.81	22.54	24.82	25.01	25.91	24.83	+1.83
	DAFE	25.10	27.78	27.18	29.36	25.62	27.27	27.05	+4.05
	Mixed FT	25.48	28.89	27.97	31.01	26.15	28.81	28.05	+5.05
	AdvAug	22.65	27.01	25.63	28.15	24.89	26.71	25.84	+2.84
	CDA	**27.13**	**29.16**	**29.21**	**32.16**	**27.51**	**29.87**	**29.17**	**+6.17**
	CDA w/o $\mathcal{L}_{mix}$	25.81	28.90	27.71	31.47	26.36	28.97	28.20	+5.20
	CDA w/o $\mathcal{L}_{adv}$	26.16	28.98	28.67	31.88	26.90	29.37	28.66	+5.66
	CDA w/o $\mathcal{L}_{nmt}$	26.58	29.06	28.73	31.91	26.58	29.54	28.73	+5.73

Table 3: Results of informal Chinese-to-English translation. ”w/o” means ”without”.

* **Base**: We only use the small target-domain corpus to train the NMT model.

* **Source**: We only use the large source-domain corpus to train the NMT model.

* **Joint**: We combine the target-domain and source-domain corpora to train the NMT model.

Informal Language Translation Results

The translation results of informal language translation for Zh→En and En→Zh are shown in Table 2 and Table 3, respectively. From the results, we can see that:

- Although the data augmentation methods with the target-domain corpus lead to different degrees of improvement, methods using a larger-scale source-domain corpus outperform them by a large margin.
- Even NMT models trained only on the source-domain corpus (**Source**), their performances are close to or better than the ones trained using only the target-domain corpus, which shows that there are a large number of common translation patterns between the two domains of texts.
- All methods that consider domain adaptation (**FT**, **TAG**, **ADAP**, **DAFE**, **Mixed FT** and **CDA**) are significantly better than **Source** method, which shows that it is important to consider the differences between the two domains.
- Experimental results show that our model consistently outperforms several baseline methods on both Zh→En and En→Zh translations. Our method yields more than 1 BLEU point gains over the strongest baseline method Mixed FT (around 2 BLEU points on Test2015 of Zh→En), demonstrating the superiority of CDA.
- The ablation studies of three losses in our method demonstrate that all our losses are useful. The source-side mixup loss has the greatest impact on the performance, which is designed to smooth the transition between domain distributions. This further validates that the superiority of CDA comes from trying to fill the gap between domains.

Cross-domain Language Translation Results

To further demonstrate the effectiveness of CDA in bridging the gap between domains, we also conduct experiments on cross-domain translation tasks. Table 4 shows cross-domain translation results of En→Zh, Zh→En, De→En and En→Fr tasks. The experiment results show that: 1) NMT is sensitive to domain shift and performs poorly if the domains are quite different. For example, the improvement of the formal domain in the spoken domain is much greater than the improvement of the Koran domain in the IT domain. 2) Although the difference between some domains is huge, e.g., Koran and IT, Methods (**FT, TAG, ADAP, DAFT, Mixed FT CDA**) that utilizes corpus from other domains still bring some improvements. 3) Compared with baseline methods, our proposed method still keeps the advantages on different tasks. As shown in Table 4, **CDA** yields more than 1 BLEU point gains over the strongest baseline method in all 10 settings, which is consistent with the findings in informal language translation experiments. It also shows the generality of **CDA**, suggesting its potential in more language pairs and tasks.

Visualization of Counterfactual Representations

To qualitatively analyze how **CDA** bridging the gap between domains, we visualize the formal, informal and two types of counterfactual representations (i.e, Adv and Mix) of part of the train set in the informal En→Zh translation task, with the dimension reduction technique of t-SNE (Maaten and Hinton 2008). As can be seen in Figure 2, the adversarial counterfactual samples (in green color) fill the latent space of target-domain data and the mixup counterfactual samples (in purple color) smooth the space between P_s (source domain samples, in red color) and P_t (target domain samples, in blue color), that is, bridging the generalization gap of two distributions. This encourages **CDA** to produce a more reasonable prediction in the latent space between the source domain and the target domain, and makes it better combine knowledge from two domains.

Train@	Method	En → Zh			Zh → En			De → En		En → Fr	
		Thesis	Edu.	Spoken	Thesis	Edu.	Spoken	Koran	IT	Med.	Par.
Target-domain corpus	**Base**	32.74	28.16	9.28	25.61	23.27	9.73	23.81	35.45	61.25	32.71
	AdvEmb	32.87	28.31	9.67	26.76	23.82	9.90	23.99	35.51	61.21	32.83
	AdvAug	32.98	28.86	10.83	27.43	24.11	10.41	24.20	36.07	62.04	32.96
Joint corpus	**Source**	25.33	20.56	18.01	14.01	20.71	13.41	8.56	9.54	13.65	15.52
	Joint	34.55	32.18	22.51	25.37	24.98	19.46	20.61	32.97	59.35	30.26
	FT	36.24	33.63	22.93	26.31	25.10	20.13	24.05	35.75	62.74	33.10
	TAG	35.29	33.97	22.88	24.45	25.17	19.94	24.43	35.80	62.88	33.41
	ADAP	36.18	33.81	21.73	24.21	25.14	19.98	24.17	35.95	63.01	33.47
	DAFE	34.43	34.10	22.78	25.13	25.08	19.80	24.76	36.11	63.23	33.52
	Mixed FT	37.17	35.86	23.71	26.58	25.21	21.19	24.43	35.91	63.35	34.10
	AdvAug	34.15	33.85	22.81	25.81	25.24	20.03	23.15	33.68	61.36	33.87
	CDA	**39.17**	**37.19**	**24.18**	**27.68**	**26.17**	**22.47**	**25.43**	**37.10**	**64.35**	**35.27**
	CDA w/o $\mathcal{L}_{mix}$	37.91	35.57	23.85	26.81	25.71	21.51	24.89	36.30	63.46	34.31
	CDA w/o $\mathcal{L}_{adv}$	38.54	36.42	24.02	27.15	25.97	22.04	25.03	36.52	63.86	34.69
	CDA w/o $\mathcal{L}_{nmt}$	38.97	36.67	24.10	27.51	26.05	22.31	25.18	36.65	64.10	35.07

Table 4: Results of cross-domain language translation. ”w/o” means ”without”.

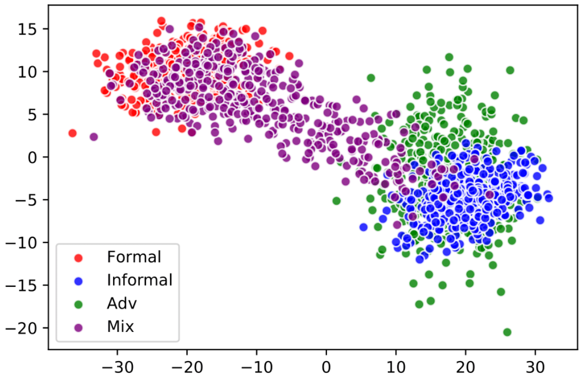


Figure 2: Visualization of encoder representations.

Related Work

Domain Adaptation for Neural Machine Translation are categorized into data-centric and model-centric methods by Chu and Wang (2018). For data-centric methods, **CDA** has connections to dynamic data selection and curriculum learning (van der Wees, Bisazza, and Monz 2017; Zhang et al. 2019; Caswell, Chelba, and Grangier 2019; Marie, Rubino, and Fujita 2020), on which the model is learned by gradually including from easy to complex samples in training so as to increase the entropy of training samples. For data-centric methods, our work is related to fine-tuning and instance weighting (Luong, Pham, and Manning 2015; Chu, Dabre, and Kurohashi 2017), which also up-weight and down-weight certain examples at specific training phases. However, the difference is that we construct new counterfactual samples that are more useful for target-domain translation based on the observed samples.

Robust Neural Machine Translation hopes to improve the translation of noise text, such as natural noise (e.g., spelling/typographical/grammatical errors, code switching and profanity), adversarial samples. The dominant approaches are data cleaning and data augmentation (Cheng, Jiang, and Macherey 2019; Zou et al. 2020; Li and Specia 2019). Different from them, we focus on the diverse linguistic phenomena of informal language translation in this study.

Low-resource Machine Translation has been widely investigated to utilize corpora of other languages with rich resources to improve low-resource language translation (Sennrich and Zhang 2019), such as multilingual machine translation (Johnson et al. 2017; Dabre, Fujita, and Chu 2019) and meta-learning (Gu et al. 2018). However, different from them, we focus on the improvement of translations in specific domains in the same language pair.

Our work is also related to **Counterfactual Augmentation**, which also generates counterfactual examples to create possible alternatives to existing samples. Although it has been used for providing explanations for the predictions (Feder et al. 2020), extending the decision boundary (Swaminathan and Joachims 2015; Besserve et al. 2020; Fu et al. 2019), and increasing the robustness (Kusner et al. 2017; Garg et al. 2019), we believe that this is the first work that applies counterfactual augmentation to domain adaptation for neural machine translation.

Conclusions

In this paper, we propose a counterfactual domain adaptation method to improve the informal language translation by utilizing large-scale parallel corpora of formal texts, which constructs the counterfactual representations that do not exist in the observed samples but can guide the model to explore the latent space of the target-domain distribution and bridge the gap between the source-domain distribution and the target-domain distribution. Experiments on both informal language translation tasks and cross-domain language translation tasks show that our method outperforms the base model and the baseline methods, demonstrating the effectiveness and robustness of the proposed approach.

Acknowledgments

This work was supported by Alibaba Innovative Research (AIR) Program, National Natural Science Foundation of

China (61772036), Beijing Academy of Artificial Intelligence (BAAI) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We would like to appreciate Kaibo Zhang for discussion, and the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR 2015*.

Bapna, A.; and Firat, O. 2019. Simple, Scalable Adaptation for Neural Machine Translation. In *EMNLP-IJCNLP 2019*, 1538–1548. Association for Computational Linguistics.

Besserve, M.; Mehrjou, A.; Sun, R.; and Schölkopf, B. 2020. Counterfactuals uncover the modular structure of deep generative models. In *ICLR 2020*.

Bojar, O.; Graham, Y.; Kamran, A.; and Stanojević, M. 2016. Results of the wmt16 metrics shared task. In *WMT 2016*, 199–231.

Britz, D.; Le, Q. V.; and Pryzant, R. 2017. Effective Domain Mixing for Neural Machine Translation. In *WMT 2017*, 118–126. Association for Computational Linguistics.

Caswell, I.; Chelba, C.; and Grangier, D. 2019. Tagged Back-Translation. In *WMT 2019*, 53–63. Association for Computational Linguistics.

Cettolo, M.; Federico, M.; Bentivogli, L.; Jan, N.; Sebastian, S.; Katsuitho, S.; Koichiro, Y.; and Christian, F. 2017. Overview of the iwslt 2017 evaluation campaign. In *IWSLT 2017*, 2–14.

Chapelle, O.; Weston, J.; Bottou, L.; and Vapnik, V. 2000. Vicinal Risk Minimization. In *NeurIPS 2000*, 416–422. MIT Press.

Chen, J.; and Zhang, J. 2019. *Machine Translation: 14th China Workshop, CWMT 2018, Wuyishan, China, October 25-26, 2018, Proceedings*, volume 954. Springer.

Cheng, Y.; Jiang, L.; and Macherey, W. 2019. Robust Neural Machine Translation with Doubly Adversarial Inputs. In *ACL 2019*, 4324–4333.

Cheng, Y.; Jiang, L.; Macherey, W.; and Eisenstein, J. 2020. AdvAug: Robust Adversarial Augmentation for Neural Machine Translation. In *ACL 2020*, 5961–5970. Association for Computational Linguistics.

Chu, C.; Dabre, R.; and Kurohashi, S. 2017. An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. In *ACL 2017*, 385–391. Association for Computational Linguistics.

Chu, C.; and Wang, R. 2018. A Survey of Domain Adaptation for Neural Machine Translation. In *COLING 2018*, 1304–1319.

Dabre, R.; Fujita, A.; and Chu, C. 2019. Exploiting Multilingualism through Multistage Fine-Tuning for Low-Resource Neural Machine Translation. In *EMNLP-IJCNLP 2019*, 1410–1416. Association for Computational Linguistics.

de Boer, P.; Kroese, D. P.; Mannor, S.; and Rubinstein, R. Y. 2005. A Tutorial on the Cross-Entropy Method. *Ann. Oper. Res.* 134(1): 19–67.

Dou, Z.; Hu, J.; Anastasopoulos, A.; and Neubig, G. 2019. Unsupervised Domain Adaptation for Neural Machine Translation with Domain-Aware Feature Embeddings. In *EMNLP-IJCNLP 2019*, 1417–1422. Association for Computational Linguistics.

Feder, A.; Oved, N.; Shalit, U.; and Reichart, R. 2020. CausaLM: Causal Model Explanation Through Counterfactual Language Models. *CoRR* abs/2005.13407.

Freitag, M.; and Al-Onaizan, Y. 2016. Fast Domain Adaptation for Neural Machine Translation. *CoRR* abs/1612.06897.

Fu, T.; Wang, X.; Peterson, M.; Grafton, S.; Eckstein, M. P.; and Wang, W. Y. 2019. Counterfactual Vision-and-Language Navigation via Adversarial Path Sampling. *CoRR* abs/1911.07308.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. S. 2016. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* 17: 59:1–59:35.

Garg, S.; Perot, V.; Limtiaco, N.; Taly, A.; Chi, E. H.; and Beutel, A. 2019. Counterfactual Fairness in Text Classification through Robustness. In *AIES 2019*, 219–226.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. N. 2017. Convolutional Sequence to Sequence Learning. In *ICML 2017*, volume 70, 1243–1252. PMLR.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS 2014*, 2672–2680.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR 2015*.

Gu, J.; Wang, Y.; Chen, Y.; Li, V. O. K.; and Cho, K. 2018. Meta-Learning for Low-Resource Neural Machine Translation. In *EMNLP 2018*, 3622–3631.

Gu, S.; Feng, Y.; and Liu, Q. 2019. Improving Domain Adaptation Translation with Domain Invariant and Specific Information. In *NAACL-HLT 2019*, 3081–3091. Association for Computational Linguistics.

Gülçehre, Ç.; Firat, O.; Xu, K.; Cho, K.; Barrault, L.; Lin, H.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2015. On Using Monolingual Corpora in Neural Machine Translation. *CoRR* abs/1503.03535.

Hu, J.; Xia, M.; Neubig, G.; and Carbonell, J. G. 2019. Domain Adaptation of Neural Machine Translation by Lexicon Induction. In *ACL 2019*, 2989–3001.

Johnson, M.; Schuster, M.; Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F. B.; Wattenberg, M.; Corrado, G.; Hughes, M.; and Dean, J. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Trans. Assoc. Comput. Linguistics* 5: 339–351.

Khayrallah, H.; Kumar, G.; Duh, K.; Post, M.; and Koehn, P. 2017. Neural Lattice Search for Domain Adaptation in Machine Translation. In *IJCNLP 2017*, 20–25.

Kobus, C.; Crego, J. M.; and Senellart, J. 2017. Domain Control for Neural Machine Translation. In *RANLP 2017*, 372–378. INCOMA Ltd.

Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A.; and Herbst, E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007*.

Koehn, P.; and Knowles, R. 2017. Six Challenges for Neural Machine Translation. In *NMT@ACL 2017*, 28–39.

Kusner, M. J.; Loftus, J. R.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In *NeurIPS 2017*, 4066–4076.

Li, Z.; and Specia, L. 2019. Improving Neural Machine Translation Robustness via Data Augmentation: Beyond Back-Translation. In *W-NUT@EMNLP*, 328–336.

Luong, M.-T.; and Manning, C. D. 2015. Stanford Neural Machine Translation Systems for Spoken Language domain. In *International Workshop on Spoken Language Translation*. Da Nang, Vietnam.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *EMNLP 2015*, 1412–1421.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Marie, B.; Rubino, R.; and Fujita, A. 2020. Tagged Back-translation Revisited: Why Does It Really Work? In *ACL 2020*, 5990–5997. Association for Computational Linguistics.

Miyato, T.; Dai, A. M.; and Goodfellow, I. J. 2017. Adversarial Training Methods for Semi-Supervised Text Classification. In *ICLR 2017*. OpenReview.net.

Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; and Auli, M. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002*, 311–318. ACL.

Pearl, J.; and Mackenzie, D. 2018. *The book of why: the new science of cause and effect*. Basic Books.

Pennell, D.; and Liu, Y. 2014. Normalization of informal text. *Comput. Speech Lang.* 28(1): 256–277.

Post, M. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191. Belgium, Brussels: Association for Computational Linguistics.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural Machine Translation of Rare Words with Subword Units. In *ACL 2016*. The Association for Computer Linguistics.

Sennrich, R.; and Zhang, B. 2019. Revisiting Low-Resource Neural Machine Translation: A Case Study. In *ACL 2019*, 211–221.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In *NeurIPS 2014*, 3104–3112.

Swaminathan, A.; and Joachims, T. 2015. Counterfactual Risk Minimization: Learning from Logged Bandit Feedback. In *ICML 2015*, volume 37, 814–823.

Tian, L.; Wong, D. F.; Chao, L. S.; Quaresma, P.; Oliveira, F.; and Yi, L. 2014. UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation. In *L-REC 2014*, 1837–1842.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. In *LREC*.

van der Wees, M.; Bisazza, A.; and Monz, C. 2017. Dynamic Data Selection for Neural Machine Translation. In *EMNLP 2017*, 1400–1410.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS 2017*, 5998–6008.

Wang, K.; Hua, H.; and Wan, X. 2019. Controllable Unsupervised Text Attribute Transfer via Editing Entangled Latent Representation. In *NeurIPS 2019*, 11034–11044.

Wang, K.; and Wan, X. 2018. SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks. In *IJCAI 2018*, 4446–4452. ijcai.org.

Wang, K.; and Wan, X. 2019. Automatic generation of sentimental texts via mixture adversarial networks. *Artif. Intell.* 275: 540–558.

Wang, K.; and Wan, X. 2020. Adversarial Text Generation via Sequence Contrast Discrimination. In *Findings, EMNLP 2020*, 47–53.

Zeng, J.; Su, J.; Wen, H.; Liu, Y.; Xie, J.; Yin, Y.; and Zhao, J. 2018. Multi-Domain Neural Machine Translation with Word-Level Domain Context Discrimination. In *EMNLP 2018*, 447–457. Association for Computational Linguistics.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *I-CLR 2018*. OpenReview.net.

Zhang, X.; Shapiro, P.; Kumar, G.; McNamee, P.; Carpuat, M.; and Duh, K. 2019. Curriculum Learning for Domain Adaptation in Neural Machine Translation. In *NAACL-HLT 2019*, 1903–1915.

Ziemski, M.; Junczys-Dowmunt, M.; and Pouliquen, B. 2016. The United Nations Parallel Corpus v1.0. In *LREC 2016*. European Language Resources Association (ELRA).

Zou, W.; Huang, S.; Xie, J.; Dai, X.; and Chen, J. 2020. A Reinforced Generation of Adversarial Examples for Neural Machine Translation. In *ACL 2020*, 3486–3497.