Journal of Information Processing Vol.28 413–426 (Aug. 2020)

[DOI: 10.2197/ipsjjip.28.413]

Regular Paper

A Survey of Domain Adaptation for Machine Translation

Chenhui Chu[1,a)] Rui Wang[2,b)]

Received: January 31, 2020, Accepted: April 13, 2020

Abstract: Neural machine translation (NMT) is a deep learning based approach for machine translation, which outperforms traditional statistical machine translation (SMT) and yields the state-of-the-art translation performance in scenarios where large-scale parallel corpora are available. Although a high-quality and domain-specific translation is crucial in the real world, domain-specific corpora are usually scarce or nonexistent, and thus vanilla NMT performs poorly in such scenarios. Domain adaptation that leverages both out-of-domain parallel corpora as well as monolingual corpora for in-domain translation, is very important for domain-specific translation. In this paper, we give a comprehensive survey of the state-of-the-art domain adaptation techniques for MT. Because of the current dominance of NMT in MT research, we give a brief review of domain adaptation for SMT, but put most of our effort into the survey of domain adaptation for NMT. We hope that this paper will be both a starting point and a source of new ideas for researchers and engineers who are interested in domain adaptation for MT.

Keywords: neural machine translation, domain adaptation, survey

1. Introduction

Neural machine translation (NMT) [5], [16], [117] allows for end-to-end training of a translation system without the need to deal with explicit word alignments, translation rules, and complicated decoding algorithms, which are characteristics of traditional statistical machine translation (SMT) systems [72]. NMT yields a state-of-the-art translation performance in resource rich scenarios [41], [91]. However, currently, high quality parallel corpora of sufficient size are only available for a few language pairs such as the languages paired with English and other widely-used European language pairs. Furthermore, for each language pair the sizes of the domain specific corpora and the number of domains available are limited. As such, for the majority of language pairs and domains, only few or no parallel corpora are available. It has been known that both vanilla SMT and NMT perform poorly for domain specific translation in low resource scenarios [33], [73], [108], [149].

High quality domain specific machine translation (MT) systems are in high demand whereas general purpose MT has limited applications. In addition, general purpose translation systems usually perform poorly and hence it is important to develop translation systems for specific domains [73]. Leveraging out-of-domain parallel corpora and in-domain monolingual corpora to improve in-domain translation is known as domain adaptation for MT [22]. For example, the Chinese-English patent domain parallel corpus has 1M sentence pairs [45], but for the spoken language domain parallel corpus there are only 200 k sentences available [12]. MT typically performs poorly in a resource poor or domain mismatching scenario and thus it is important to leverage the spoken language domain data with the patent domain data [20]. Furthermore, there are monolingual corpora containing millions of sentences for the spoken language domain, which can also be leveraged [106].

There are many studies of domain adaptation for SMT, which can be mainly divided into two categories: data centric and model centric. Data centric methods focus on either selecting training data from out-of-domain parallel corpora based on a language model (LM) [4], [14], [33], [34], [52], [90] or generating pseudo parallel data [17], [84], [123], [131], [132]. Model centric methods interpolate in-domain and out-of-domain models in either a model level [34], [55], [108] or an instance level [42], [85], [103], [110], [148]. However, due to the different characteristics of SMT and NMT, many methods developed for SMT cannot be applied to NMT directly.

Domain adaptation for NMT is rather new and has attracted plenty of attention in the research community. In the past four years, NMT has become the most popular MT approach and many domain adaptation techniques have been proposed and evaluated for NMT. These studies either borrow ideas from previous SMT studies and apply these ideas for NMT, or develop unique methods for NMT. Despite the rapid development in domain adaptation for NMT, there lacks a single up-to-date compilation that summarizes and categorizes all approaches. As such a study will greatly benefit the community, we present in this paper a survey of all prominent domain adaptation techniques for NMT. There are survey papers for NMT [70], [92]; however, they focus on general NMT and more diverse topics. Domain adaptation surveys have been done in the perspective of computer vision [23] and machine learning [96], [136].

In this paper, similar to SMT, we categorize domain adapta-

[1] Institute for Datability Science, Osaka University, Suita, Osaka 565–0871, Japan
[2] National Institute of Information and Communications Technology, Soraku District, Kyoto 619–0237, Japan
[a)] chu@ids.osaka-u.ac.jp
[b)] wangrui@nict.go.jp

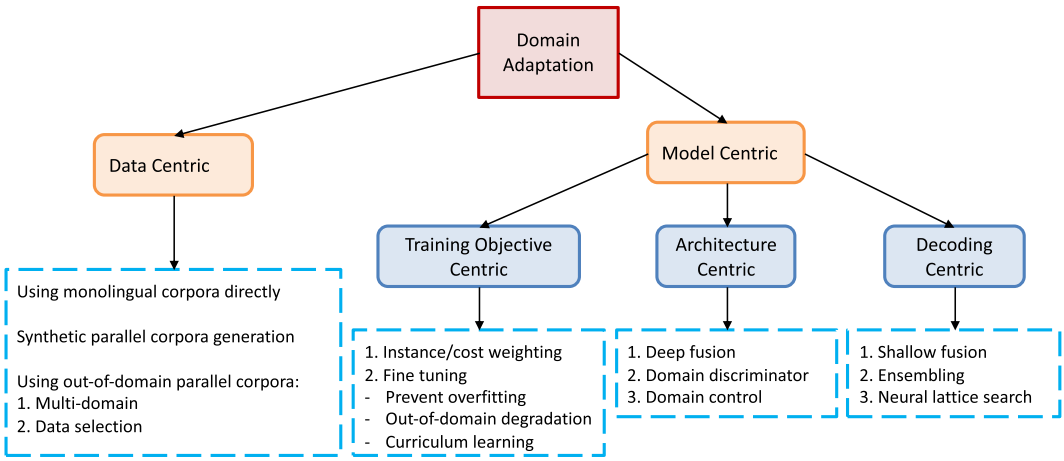


Fig. 1 Overview of domain adaptation for NMT.

tion for NMT into two main categories: data centric and model centric. The data centric category focuses on the data being used rather than specialized models for domain adaptation. The data used can be either in-domain monolingual corpora [15], [24], [30], [47], [144], synthetic corpora [36], [53], [94], [97], [106], [107], [144], [147], or parallel corpora [10], [20], [29], [48], [69], [80], [89], [101], [102], [104], [118], [124], [128]. On the other hand, the model centric category focuses on NMT models that are specialized for domain adaptation, which can be either the training objective [13], [20], [27], [28], [44], [64], [65], [67], [71], [77], [81], [86], [106], [109], [113], [119], [120], [125], [130], [134], [140], [141], [145], [146], the NMT architecture [10], [19], [31], [32], [46], [47], [59], [68], [87], [100], [115], [135], [137], [142] or the decoding algorithm [28], [32], [47], [66], [98]. An overview of these two categories is shown in **Fig. 1**. Note that as model centric methods also use either monolingual or parallel corpora, there are overlaps between these two categories.

We previously conduct a survey of domain adaptation for NMT [22] in about two years ago. However, due to the rapid development of this field, the state-of-the-art NMT model and research trends have changed, and many newly appeared studies are not covered in our previous survey. This paper extends our previous survey paper [22] as follows:

- We update the formulation of NMT by adding the state-of-the-art Transformer based model [126]. In addition, we formulate and optimize the categorization of training objective centric approaches.
- We add and categorize dozens of new studies appearing after our previous survey. We also newly discuss incremental domain adaptation, multilingual, and multi-domain adaptation as specific scenarios.
- We introduce datasets and resources that are useful for NMT domain adaptation studies. In addition, we update future directions of domain adaptation for NMT considering the new trend after our previous survey.

These extension and updates make our paper more comprehensive, and up-to-date.

The remainder of this paper is structured as follows: We first give a brief introduction of NMT, and describe the reason for the difficulty of low resource domains and languages in NMT (Section 2); Next, we introduce the background of domain adaptation and briefly review the historical domain adaptation techniques being developed for SMT (Section 3); Under this background knowledge, we then present and compare the domain adaptation methods for NMT in detail (Section 4); After that, we introduce domain adaptation for NMT in specific scenarios that are crucial for the practical use of MT (Section 5) and the commonly used datasets and resources in research (Section 6); Finally, we give our opinions of future research directions in this field (Section 7) and conclude this paper (Section 8).

2. Neural Machine Translation

NMT is an end-to-end approach for translating from one language to another, which relies on deep learning to train a translation model [5], [16], [117]. NMT takes in an input sentence $\mathbf{x} = \{x_1, ..., x_n\}$ and its translation $\mathbf{y} = \{y_1, ..., y_m\}$. The translation is generated as:

$$p(\mathbf{y}|\mathbf{x};\theta) = \prod_{j=1}^{m} p(y_j|y_{<j}, \mathbf{x};\theta),$$

where θ is a set of parameters, m is the entire number of words in $\mathbf{y}$, y_j is the current predicted word, and $y_{<j}$ are the previously predicted words. Suppose we have a parallel corpus $\mathbf{C}$ consisting of a set of parallel sentence pairs $(\mathbf{x}, \mathbf{y})$. The training object is to maximize the log-likelihood $\mathcal{L}$ w.r.t θ:

$$\mathcal{L}_\theta = \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{C}} \log p(\mathbf{y}|\mathbf{x};\theta). \quad (1)$$

RNN based Model The encoder-decoder model with attention [5] is the most commonly used NMT architecture. **Figure 2** shows an overview of this model. It consists of three main parts, namely, the encoder, decoder and attention model. The encoder uses an embedding mechanism to convert words into their continuous space representations. These embeddings by themselves do not contain information about relationships between words and their positions in the sentence. Using a recurrent neural network (RNN) layer such as gated recurrent unit and long short-term memory, this can be accomplished. An RNN maintains a hidden state (also called a memory or history), which allows it to generate a continuous space representation for a word given

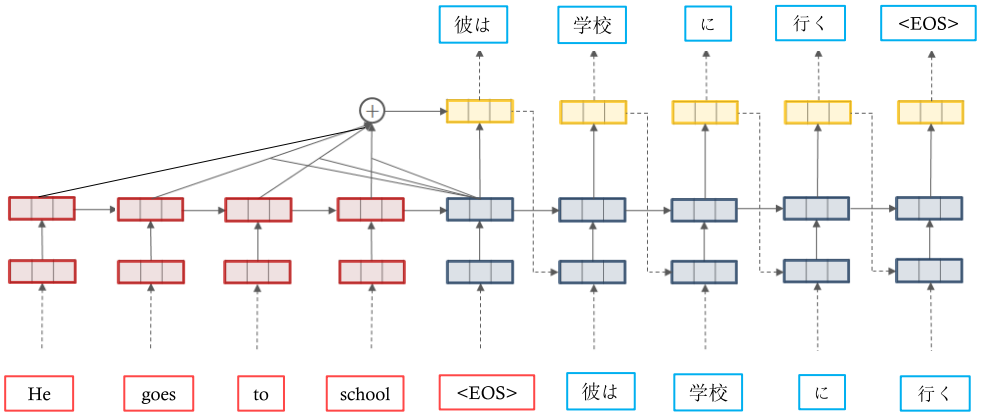


Fig. 2 The architecture of the RNN based NMT system with attention.

all past words that have been seen. There are two RNN layers which encode forward and backward information. Each word x_i is represented by concatenating the forward hidden state $\overrightarrow{\mathbf{h_i}}$ and the backward one $\overleftarrow{\mathbf{h_i}}$ as $\mathbf{h_i} = [\overrightarrow{\mathbf{h_i}}; \overleftarrow{\mathbf{h_i}}]$. In this way, the source sentence $\mathbf{x} = \{x_1, ..., x_n\}$ can be represented as $\mathbf{h} = \{\mathbf{h_1}, ..., \mathbf{h_n}\}$. By using both forward and backward recurrent information, one obtains a continuous space representation for a word given all words before as well as after it.

The decoder is conceptually an RNN language model (RNNLM) with its own embedding mechanism, an RNN layer to remember previously generated words and a softmax layer to predict a target word. The encoder and decoder are coupled by using an attention mechanism, which computes a weighted average vector of the recurrent representations generated by the encoder thereby acting as a soft alignment mechanism. This weighted averaged vector, also known as the context or attention vector, is fed to the decoder RNN along with the previously predicted word to produce a representation that is passed to the softmax layer to predict the next word. In equation, an RNN hidden state $\mathbf{s_j}$ for time j of the decoder is computed by:

$$\mathbf{s_j} = g(\mathbf{s_{j-1}}, \mathbf{y_{j-1}}, \mathbf{c_j}),$$

where g is an activation function of RNN, $\mathbf{s_{j-1}}$ is the previous RNN hidden state, $\mathbf{y_{j-1}}$ is the embedding of the previous word, $\mathbf{c_j}$ is the context vector. $\mathbf{c_j}$ is computed as a weighted sum of the encoder hidden states $\mathbf{h} = \{\mathbf{h_1}, ..., \mathbf{h_n}\}$, by using alignment weight a_{ji}:

$$\mathbf{c_j} = \sum_{i=1}^{n} a_{ji}\mathbf{h_i}, \tag{2}$$

where

$$a_{ji} = \frac{\exp(e_{ji})}{\sum_{k=1}^{n}\exp(e_{jk})},$$
$$e_{ji} = align(\mathbf{s_{j-1}}, \mathbf{h_i}),$$

where *align* is an alignment model that scores the match level of the inputs around position i and the output at position j. The softmax layer contains a feedforward layer f. The feedforward layer takes the recurrent hidden state generated by the decoder RNN, the previous word and the context vector to compute a final representation, which is fed to the softmax layer:

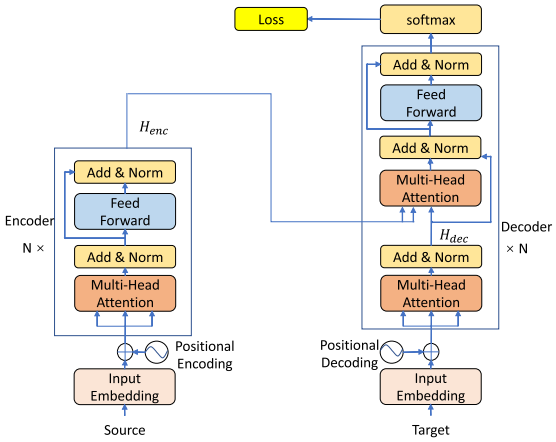


Fig. 3 The architecture of the Transformer based NMT system with self-attention.

$$P(y_j|y_{<j}, \mathbf{x}) = \text{softmax}(f(\mathbf{s_j}, \mathbf{y_{j-1}}, \mathbf{c_j})). \tag{3}$$

Transformer based Model relies on self-attention networks (**SANs**), and has become the dominant model in the MT community. SANs are neural networks with no recurrent or convolution operations, and fully reliant on a self-attention mechanism [126] to learn the source representation for the NMT as shown in **Fig. 3**. Specifically, the $\boldsymbol{H}^0$ for source input is first packed into a query matrix $\boldsymbol{Q}^0$, a key matrix $\boldsymbol{K}^0$, and a value matrix $\boldsymbol{V}^0$. The multi-head self-attention is performed over the $\boldsymbol{Q}^0$, $\boldsymbol{K}^0$, and $\boldsymbol{V}^0$:

$$\text{MultiHead}(\boldsymbol{Q}_0, \boldsymbol{K}_0, \boldsymbol{V}_0) = \text{Concat}(\boldsymbol{Q}_1^1 : \cdots : \boldsymbol{Q}_H^1)\boldsymbol{W}^O,$$
$$\boldsymbol{Q}_h^1 = \text{softmax}\left(\frac{\boldsymbol{Q}_h^0 \boldsymbol{K}_h^{0T}}{\sqrt{d_{model}}}\right)\boldsymbol{V}_h^0,$$
$$\boldsymbol{Q}_h^0, \boldsymbol{K}_h^0, \boldsymbol{V}_h^0 = \boldsymbol{Q}^0\boldsymbol{W}_h^Q, \boldsymbol{K}^0\boldsymbol{W}_h^K, \boldsymbol{V}^0\boldsymbol{W}_h^V,$$

where $\boldsymbol{Q}_h^0$, $\boldsymbol{K}_h^0$, and $\boldsymbol{K}_h^0$ are respective the query, key, and value matrices of the h-th head. $\{\boldsymbol{W}_h^Q, \boldsymbol{W}_h^K, \boldsymbol{W}_h^V\} \in \mathbb{R}^{d_{model} \times d_k}$ denote parameter matrices, d_{model} and d_k represent the dimensions of the model and the head. For example, if there are H=8 heads and d_{model} is 512, d_k=512/8=64. A position-wise feedforward neural network (FFNN), which is a fully connected network with ReLU activation function, is then applied to each position separately and identically:

$$\boldsymbol{Q}^1 = \text{FFNN}(\text{MultiHead}(\boldsymbol{Q}^0, \boldsymbol{K}^0, \boldsymbol{V}^0)) + \boldsymbol{Q}^0,$$

where Q^1 is the source representation with global feature information.

Similarly, this processing sequence is formally denoted as the function f_{SANs} to learn the source representation Q^1:

$$\boldsymbol{Q}^1 = f_{\text{SANs}}(\boldsymbol{Q}^0, \boldsymbol{K}^0, \boldsymbol{V}^0).$$

According to Vaswani [126]'s work, SANs use a stack of N-identical layers to learn the source representation:

$$\left[\boldsymbol{Q}^n = f^n_{\text{SANs}}(\boldsymbol{Q}^{n-1}, \boldsymbol{K}^{n-1}, \boldsymbol{V}^{n-1})\right]_N,$$

where $[\cdots]_N$ denotes a stack of N-identical layers for the encoder; and n takes each of the values $\{1, 2, \cdots, N\}$ in turn. As a result, the output $\boldsymbol{Q}_N$ of the N-th SANs layer is the final source representation to be fed into the decoder to learn a translation context vector for predicting the target word.

An abundance of parallel corpora are required to train an NMT system to avoid overfitting, due to the large amounts of parameters in the encoder, decoder, attention model, and SANs. This is the main bottleneck of NMT for low resource domains and languages.

3. Domain Adaptation

Transfer learning is a research problem in machine learning that focuses on solving one problem and applying it to a different but related problem. Specifically, transfer learning is using source domain D_s and source task T_s to improve the effect of target domain D_t and target task T_t [95], [121]. The information of D_s and T_s is transferred to D_t and T_t. Domain adaptation can be considered as a type of isomorphic transfer learning where $T_s = T_t$.

3.1 Domain Adaptation for NLP and MT

Domain adaptation is an important problem in natural language processing (NLP) due to the lack of labeled data in novel domains [60]. As Jiang et al. [60] mentioned, the domain adaptation problem is commonly encountered in NLP. For example, in part-of-speech tagging, the source domain may be tagged WSJ articles, and the target domain may be scientific literature that contains scientific terminology. In named entity recognition, the source domain may be annotated news articles, and the target domain may be personal blogs.

Most of the NLP tasks only have a vocabulary with limited size, such as most tagging tasks and classification tasks. In comparison, MT is a classic language generation task and its vocabulary is quite large. Typically, the vocabulary usually contains 30 k∼50 k words or sub-words. Therefore, the domain adaptation problem in MT is also more complicated than other NLP tasks.

Formally, given a small in-domain parallel corpus $\mathbf{C_{in}}$, the problem of domain adaptation for MT is how to improve in-domain translation using either large out-of-domain parallel corpora $\mathbf{C_{out}}$ or monolingual in-domain corpora $\mathbf{M_{in}}$. Note that $\mathbf{M_{in}}$ can be either in the source or target language.

3.2 Domain Adaptation for SMT

In SMT, many domain adaptation methods have been proposed to overcome the problem of the lack of substantial data in specific domains and languages. Most SMT domain adaptation methods can be broken down broadly into two main categories.

Data Centric This category focuses on selecting or generating the domain-related data using existing in-domain data.

i) When there are sufficient parallel corpora from other domains, the main idea is to score the out-domain data using models trained from the in-domain and out-of-domain data and select training data from the out-of-domain data using a cut-off threshold on the resulting scores. LMs [4], [33], [90], as well as joint models [34], [52], and more recently convolutional neural network (CNN) models [14] can be used to score sentences.

ii) When there are not enough parallel corpora, there are also studies that generate pseudo-parallel sentences using information retrieval [123], self-enhancing [75] or parallel word embeddings [84]. Besides sentence generation, there are also studies that generate monolingual n-grams [131] and parallel phrase pairs [17], [132].

Most of the data centric-based methods in SMT can be directly applied to NMT. However, most of these methods adopt the criteria of data selection or generation that are not related to NMT. Therefore, these methods can only achieve modest improvements in NMT [128].

Model Centric This category focuses on interpolating the models from different domains.

i) Model level interpolation. Several SMT models, such as LMs, translation models, and reordering models, individually corresponding to each corpus, are trained. These models are then combined to achieve the best performance [9], [34], [43], [55], [93], [108].

ii) Instance level interpolation. Instance weighting has been applied to several NLP domain adaptation tasks [60], especially SMT [42], [83], [85], [111], [148]. They firstly score each sentence pair/domain by using rules or statistical methods as a weight, and then train SMT models by giving each sentence pair/domain the weight. An alternative way is to weight the corpora by data re-sampling [103], [110].

For NMT, several methods have been proposed to interpolate model/data as SMT does. For model-level interpolation, the most related NMT technique is model ensemble [58]. For instance-level interpolation, the most related method is to assign a weight in NMT objective function [13], [130]. However, the model structures of SMT and NMT are quite different. SMT is a combination of several independent models; in comparison, NMT is an end-to-end model itself. Therefore, most of these methods cannot be directly applied to NMT.

4. Domain Adaptation for NMT

4.1 Data Centric

4.1.1 Using Monolingual Corpora Directly

Unlike SMT, in-domain monolingual data cannot be used as an LM for conventional NMT directly, and many studies have been conducted for this. Gülçehre et al. [47] train an RNNLM on monolingual data, and fuse the RNNLM and NMT models. Currey et al. [24] copy the target monolingual data to the source side and use the copied data for training NMT. Domhan and Hieber [30] propose using target monolingual data for the decoder

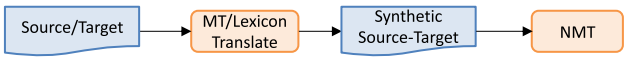


Fig. 4 Synthetic data generation for NMT.

with LM and NMT multitask learning. Zhang and Zong [144] use source side monolingual data to strengthen the NMT encoder via multitask learning for predicting both translation and reordered source sentences. Cheng et al. [15] use both source and target monolingual data for NMT through reconstructing the monolingual data by using NMT as an autoencoder.

4.1.2 Synthetic Parallel Corpora Generation

As NMT itself has the ability of learning LMs, target monolingual data also can be used for the NMT system to strengthen the decoder after back translating target sentences to generate a synthetic parallel corpus [106]. The method of back translation is crucial, and it has been investigated that back translation via sampling or noised beam outputs performs better than pure beam search [36]. It has also been shown that synthetic data generation is very effective for domain adaptation using either the target side monolingual data [107], the source side monolingual data [144], or both [94], [97]. Different from studies that generate a synthetic parallel corpus by MT systems, Hu et al. [53] perform lexicon induction to first obtain an in-domain lexicon and then use the lexicon to generate a synthetic parallel corpus via word-to-word translation. They show that their method is effective for in-domain unseen word translation. Zheng et al. [147] use clear and noisy tags appending to the target sentences to generate social-media-style synthetic sentences for the robust translation task [88]. **Figure 4** summarizes the synthetic data generation approach for NMT.

4.1.3 Using Out-of-Domain Parallel Corpora

With both in-domain and out-of-domain parallel corpora, it is ideal to train a mixed domain MT system that can improve in-domain translation while do not decrease the quality of out-of-domain translation. We categorize these efforts as *multi-domain* methods, which have been successfully developed for NMT. In addition, the idea of data selection from SMT also have been developed for NMT.

Multi-Domain The *multi-domain* method in Chu et al. [20] is originally motivated by Sennrich et al. [105], which uses tags to control the politeness of NMT. The overview of this method is shown in the dotted section in Fig. 7. In this method, the corpora of multiple domains are concatenated with two small modifications:

- Appending the domain tag "<2domain>" to the source sentences of the respective corpora. This primes the NMT decoder to generate sentences for the specific domain.
- Oversampling the smaller corpus so that the training procedure pays equal attention to each domain.

Kocmi et al. [69] show that simple concatenation of out-of-domain and in-domain data can be harmful for in-domain translation. Luo et al. [80] apply the multi-domain method to translate financial listing documents. Mino et al. [89] use the multi-domain method for the newswire Japanese-English task at WAT 2019. Sajjad et al. [104] further compare different methods for training a multi-domain system. In particular, they compare *concatenation* that simply concatenates the multi-domain corpora, *staking* that iteratively trains the NMT system on each domain corpus, *selection* that selects a set of out-of-domain data which is close to the in-domain data, and *ensemble* that ensembles the multiple NMT models trained independently. They find that fine tuning the concatenation system on in-domain data shows the best performance. Britz et al. [10] compare the *multi-domain* method with a discriminative method (see Section 4.2.2 for details). They show that the discriminative method performs better than the *multi-domain* method. Tars and Fishel [118] further study the *multi-domain* method on the setting that the domains are unknown, which automatically clusters parallel sentences into different domains during training and testing. He et al. [48] go to the same direction that treat domains as latent variables, which are learned via optimizing the marginal log-likelihood.

Data Selection As mentioned in the SMT section (Section 3.2), the data selection methods in SMT can improve NMT performance modestly, because their criteria of data selection are not very related to NMT [128]. To address this problem, Wang et al. [128] exploit the internal embedding of the source sentence in NMT, and use the sentence embedding similarity to select the sentences that are close to in-domain data from out-of-domain data (**Fig. 5**). Van der Wees et al. [124] propose a dynamic data selection method, in which they change the selected subset of training data among different training epochs for NMT. They show that gradually decreasing the training data based on the in-domain similarity gives the best performance. Ding et al. [29] label sentences with domain information, and then select sentences based on the labels. Poncelas et al. [101], [102] use a feature decay algorithm that selects sentences most relevant to the source sentences in the test set. They show that the translation performance on the selected subset outperforms that on the full training corpus.

Although all the data centric methods for NMT are complementary to each other in principle, there are few studies that try to combine these methods, which could be further studied.

4.2 Model Centric

4.2.1 Training Objective Centric

The methods in this section change the training functions or procedures for obtaining an optimal in-domain training objective.

Instance/Cost Weighting The main challenge for instance weighting in NMT is that NMT is not a linear model or a combination of linear models, which means the instance weight cannot be integrated into NMT directly. Instead, we have to change the training object in Eq. (1) as:

$$\mathcal{L}_\theta = \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{C_{in}}} \log \lambda_{in} p(\mathbf{y}|\mathbf{x};\theta) + \sum_{(\mathbf{x}',\mathbf{y}')\in\mathbf{C_{out}}} \log p(\mathbf{y}'|\mathbf{x}';\theta),$$

where $\mathbf{C_{in}}$ and $\mathbf{C_{out}}$ denote the in-domain and out-of-domain corpora, respectively, λ_{in} is the in-domain instance weight. Wang et al. [130] set an in-domain weight for the objective function, and this weight is learned from the cross-entropy by an in-domain LM and an out-of-domain LM [4] (**Fig. 6**). The in-domain weight can be set in either instance, corpus or batch level. Zhang and Xiong [145] learn the similarity of a sentence to the in-domain

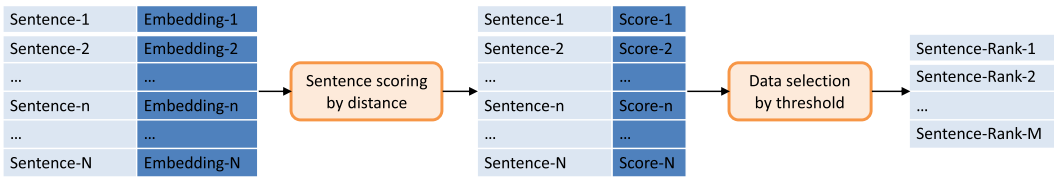


Fig. 5 Data selection for NMT [128].

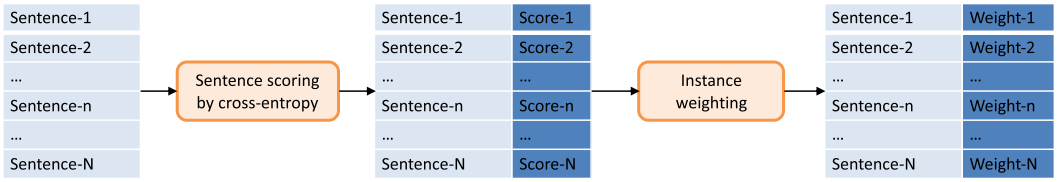


Fig. 6 Instance weighting for NMT [130].

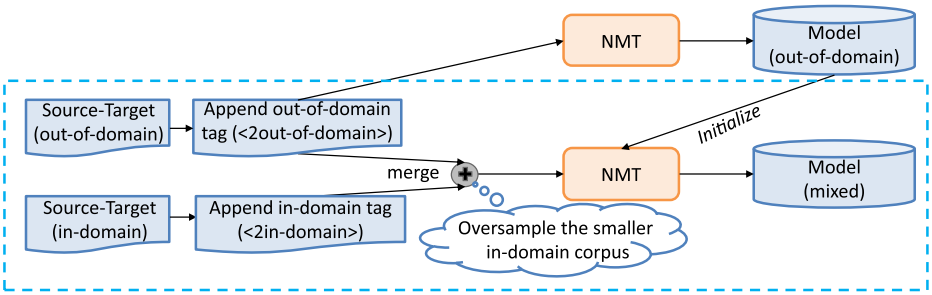


Fig. 7 Mixed fine tuning with domain tags for domain adaptation [20]. The section in the dotted rectangle denotes the *multi-domain* method.

corpus and use the similarity for the objective function. Yan et al. [140] further compute the word weights in out-of-domain datasets based on an in-domain LM and an out-of-domain LM, which gives higher weights to in-domain words. Instead of instance weighting, Chen et al. [13] modify the NMT cost function with a domain classifier. The output probability of the domain classifier is transferred into the domain weight. This classifier is trained using development data. Wang et al. [129] propose a joint framework of sentence selection and weighting for NMT.

Fine Tuning *Fine tuning* is the conventional way for domain adaptation [44], [81], [106], [109]. In this method, an NMT system on a resource rich out-of-domain corpus is trained until convergence with the training objective:

$$\mathcal{L}_{\theta_{\text{out}}} = \sum_{(\mathbf{x}',\mathbf{y}')\in\mathbf{C}_{\text{out}}} \log p(\mathbf{y}'|\mathbf{x}';\theta_{\text{out}}),$$

and then its parameters θ_{out} are used for initializing θ_{in} during in-domain training and fine tuned on a resource poor in-domain corpus as:

$$\mathcal{L}_{\theta_{\text{in}}} = \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{C}_{\text{in}}} \log p(\mathbf{y}|\mathbf{x};\theta_{\text{in}}),$$

Conventionally, fine tuning is applied on in-domain parallel corpora. Varga et al. [125] apply it on parallel sentences extracted from comparable corpora. Comparable corpora have been widely used for SMT by extracting parallel data from them [17]. Khan et al. [65] study a setting that there are multiple in-domain data, and fine tune the model on the multiple in-domain data in order. Thompson et al. [120] study the effect of the encoder, decoder, word embedding, and softmax components by freezing them individually during fine tuning, and find that any single component has little impact on the performance. Vilar [127] applies learning hidden unit contribution to amplify the contribution of the hidden states during fine tuning. Li et al. [77] split the parameters into model and meta ones, and only update the meta parameters during fine tuning for fast domain adaptation.

Prevent Overfitting. Due to the small-scale of in-domain data, fine tuning tends to overfit very quickly. Chu et al. [20] propose mixed fine tuning to address this problem, which is a combination of the *multi-domain* and *fine tuning* methods (**Fig. 7**). The training procedure is as follows:

(1) Train an NMT model on out-of-domain data until convergence.
(2) Resume training the NMT model from step 1 on a mix of in-domain and out-of-domain data (by oversampling the in-domain data) until convergence.

Chu et al. [20] show that mixed fine tuning works better than both *multi-domain* and *pure fine tuning*. In addition, mixed fine tuning has a similar effect as the ensembling method in Dakw and Monz [28], which does not decrease the out-of-domain translation performance. Kawara et al. [64] and Dabre et al. [27] apply the mixed fine tuning method for the low resource Myanmar-English translation task at WAT 2018. Sostaric et al. [113] apply mixed fine tuning on a low resource language pair of English-Croatian. Barone et al. [86] address this problem by exploring regularization techniques such as dropout and L2-regularization. In addition, they also propose *tuneout* that is a variant of dropout for regularization. Unlike dropout that drops columns of the

weight matrices and sets them to zero, tuneout sets them to the corresponding out-of-domain parameter columns. We think that mixed fine tuning and regularization techniques are complementary to each other.

Prevent Out-of-domain Degradation. As fine tuning is applied on the in-domain data, it tends to decrease the out-of-domain performance. To prevent degradation of out-of-domain translation after fine tuning on in-domain data, several studies have been conducted. Dakwale and Monz [28] propose an extension of fine tuning that keeps the distribution of the out-of-domain model based on knowledge distillation [50]. Zeng et al. [141] further use knowledge distillation in a bi-directional (both out-of-domain to in-domain, and in-domain to out-of-domain) way iteratively. Khayrallah et al. [67] add an additional term that minimizes the cross-entropy between the in-domain and out-of-domain's output word distribution to the NMT training objective. Thompson et al. [119] adapt elastic weight consolidation aiming to learn a new task without forgetting previous tasks.

Curriculum Learning. Fine tuning can also be applied on the sentences selected from out-of-domain data by data selection, which are relevant to in-domain data and thus can boost the performance [71]. Wang et al. [134] use small trusted data to measure noise in selected sentences and sort sentences by their noise level, making fine tuning perform on gradually noise-reduced data batches. Zhang et al. [146] apply curriculum learning in this direction in that they use the similarity scores given by data selection to rearrange the order of the selected sentences, making sentences more similar to in-domain data being seen earlier and used more frequently during fine tuning.

4.2.2 Architecture Centric

The methods in this section change the NMT architecture for domain adaptation.

Deep Fusion One technique of adaptation with in-domain monolingual data is to train an in-domain RNNLM for the NMT decoder and combine it (also known as fusion) with an NMT model [47]. Fusion can either be shallow or deep. Formally, deep fusion indicates that the LM and NMT are integrated as a single decoder (i.e., integrating the RNNLM into the NMT architecture). Shallow fusion indicates that the scores of the LM and NMT are considered together (i.e., rescoring the NMT model with the RNNLM model, and more details are presented in Section 4.2.3).

In deep fusion, the RNNLM and the decoder of the NMT are integrated by concatenating their hidden states. When computing the output probability of the next word, the model is fine tuned to use the hidden states of both the RNNLM and NMT models. With the RNN based NMT output probability Eq. (3), deep fusion can be formulated as follows:

$$P(y_j|y_{<j}, \mathbf{x}) = \mathrm{softmax}(f(\mathbf{s}_\mathbf{j}^{\mathbf{MT}}, \mathbf{s}_\mathbf{j}^{\mathbf{LM}}, \mathbf{y_{j-1}}, \mathbf{c_j})),$$

where $\mathbf{s}_\mathbf{j}^{\mathbf{MT}}$ and $\mathbf{s}_\mathbf{j}^{\mathbf{LM}}$ are the hidden states for NMT and RNNLM, respectively. Domhan and Hieber [30] propose a method similar to the deep fusion method [47]. However, unlike training the RNNLM and NMT model separately [47], Domhan and Hieber [30] train RNNLM and NMT models jointly. Dou et al. [32] improves the deep fusion method by using both an in-domain and out-of-domain LMs to simulate the difference between in-domain and out-of-domain NMT.

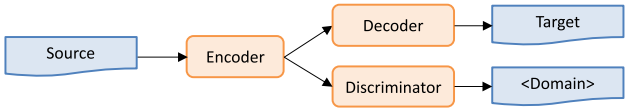


Fig. 8 Domain discriminator [10].

Domain Discriminator To leverage the diversity of information in multi-domain corpora, Britz et al. [10] propose a discriminative method. In their discriminative method, they add an FFNN as a discriminator on top of the encoder that uses the sum of context vector at each position in Eq. (2) to predict the domain of the source sentence, namely:

$$\mathbf{c} = \sum_{j=1}^{m} \mathbf{c_j},$$

$$p(d|\mathbf{c}) = g(\mathbf{c}),$$

where d is domain class label, and g is an FFNN. The discriminator is optimized jointly with the NMT network. **Figure 8** shows an overview of this method.

Zeng et al. [142] propose to discriminate word-level domain-specific and domain-shared context for improving multi-domain NMT. Word-level domain-specific and domain-shared context are learned from source sentences in the encoder with domain classifiers and then used for decoding. In addition, a domain classifier on the decoder side is learned to weight words during decoding. Su et al. [115] extend the work of Ref. [142] by using a source-side context gate to better incorporate domain-specific and domain-shared context, and they also experiment on the Transformer. Pham et al. [100] assign domain-shared and domain-specific dimensions in the word embeddings to discriminate word-level domain context. Chu and Dabre [19] learn domain-shared and domain-specific decoder hidden state representations for multi-domain NMT. Wu et al. [137] mimic the attention mechanism in RNN based NMT for soft domain adaptation, which learns domain context by calculating attention scores on multiple domain representations. Gu et al. [46] further use multiple encoders and decoders for domain-specific and domain-shared translations. Dou et al. [31] design a network to learn domain- and task-specific embeddings via training language modeling and NMT on in- and out-of-domain monolingual data and out-of-domain parallel data sequentially. Wang et al. [135] propose a domain transformation network to transform general embeddings to domain-specific embeddings. The network is supervised by domain distillation that is guided by domain teachers, and domain discrimination that distinguishes general and domain-specific embeddings.

Domain Control Besides using domain tokens to control the domains, Kobus et al. [68] propose to append word-level features to the embedding layer of NMT to control the domains. In particular, they append a domain tag to each word. They also propose a term frequency - inverse document frequency (tf-idf) based method to predict the domain tag for input sentences. Jehl and Riezler [59] extend the work of Kobus et al. [68] in that instead of domain tokens, they propose to use document categories for

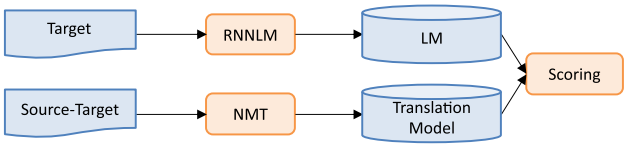


Fig. 9 LM shallow fusion [47].

patent translation. Michel and Neubig [87] propose to add a bias vector to the softmax layer in Eq. (3) to control the domain as:

$$P(y_j|y_{<j}, \mathbf{x}) = \text{softmax}(f(\mathbf{s_j}, \mathbf{y_{j-1}}, \mathbf{c_j}) + \mathbf{b}),$$

where **b** is a bias vector. They experiment their method on a personalized domain adaptation setting that adapts for speakers in the TED talks.

4.2.3 Decoding Centric

Different from architecture centric methods that changes the entire NMT architecture, decoding centric methods focus on the decoding algorithm for domain adaptation, which are essentially complementary to the other model centric methods.

Shallow Fusion Shallow fusion is an approach where LMs are trained on large monolingual corpora, following which they are combined with a previously trained NMT model [47]. In the shallow fusion [47], the next word hypotheses generated by an NMT model is rescored by the weighted sum of the NMT and RNNLM probabilities (see **Fig. 9**) as:

$$p(y_j = k) = p_{MT}(y_j = k) + \lambda p_{LM}(y_j = k),$$

where k is a candidate target word being output at y_j, and λ is the weight to be tuned. Dou et al. [32] also improves this shallow fusion method by using both in-domain and out-of-domain LMs.

Ensembling Freitag and Al-Onaizan [44] propose to ensemble the out-of-domain and the fine tuned in-domain models as:

$$p(y_j = k) = ens(p_{in}(y_j = k), p_{out}(y_j = k)),$$

where *ens* is a ensemble function that usually uses either a majority voting or consensus building scheme, p_{in} and p_{out} are the output probabilities for the in-domain and out-of-domain models, respectively. Their motivation is exactly the same as the work of Dakwale and Monz [28], which is preventing degradation of out-of-domain translation after fine tuning on in-domain data. Peng et al. [98] apply ensembling for the biomedical translation task at WMT 2019.

Neural Lattice Search Khayrallah et al. [66] propose a stack-based decoding algorithm over word lattices, while the lattices are generated by SMT [35]. In their domain adaptation experiments, they show that stack-based decoding is better than conventional decoding.

5. Domain Adaptation in Specific Scenarios

A domain adaptation method should be adopted according to certain scenarios. For example, when there are some pseudo parallel in-domain data in the out-of-domain data, sentence selection is preferred; when only additional monolingual data is available, LM and NMT fusion can be adopted. In many cases, both out-of-domain parallel data and monolingual in-domain data are available, making the combination of different methods possible.

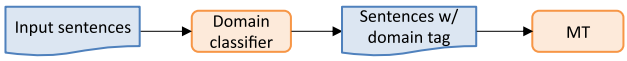


Fig. 10 Domain adaptation in an input domain unknown scenario.

Chu et al. [21] conduct a study that applies mixed fine tuning [20] on synthetic parallel data [106], which shows better performance than either method. Therefore, we do not recommend any particular techniques in this paper but recommend readers to choose the best method for their own scenarios. In addition, in this section we discuss three specific scenarios, which have not been covered in Section 4.

5.1 Input Domain Unknown

Most of the above domain adaptation studies assume that the domain of the data is given. However, in a practical view such as an online translation engine, the domain of the sentences input by the users are not given. For such a scenario, predicting the domains of the input sentences is crucial for good translation. To address this problem, a common method in SMT is to firstly classify the domains and then translate input sentences in classified domains using corresponding models [54]. Xu et al. [139] perform domain classification for a Chinese-English translation task. The classifiers operate on whole documents rather than on individual sentences, using LM interpolation and vocabulary similarities. Huck et al. [54] extend the work of Xu et al. [139] on the sentence level. They use LMs and maximum entropy classifiers to predict the target domain. Banerjee et al. [6] build a support vector machine classifier using tf-idf features over bigrams of stemmed content words. Classification is carried out on the level of individual sentences. Wang et al. [133] rely on averaged perceptron classifiers with various phrase-based features.

For NMT, Kobus et al. [68] propose an NMT domain control method, by appending either domain tags or features to the word embedding layer of NMT. They adopt an in-house classifier to distinguish the domain information. Li et al. [78] propose to search similar sentences in the training data using the test sentence as a query, and then fine tune the NMT model using the retrieved training sentences for translating the test sentence. This method also has been used when the domain of input sentences is known [51]. Farajian et al. [39] follow the strategy of Li et al. [78], but propose to dynamically set the hyperparameters (i.e., learning rate and number of epochs) of the learning algorithm based on the similarity of the input sentence and the retrieved sentences for updating the NMT model. Farajian et al. [37] further report that the method of Ref. [39] can boost in-domain terminology translation. Tars and Fishel [118] cluster parallel sentences into different domains during training and testing. **Figure 10** shows an overview of domain adaptation for MT in the input domain unknown scenario.

5.2 Incremental Domain Adaptation

Another scenario is applying domain adaptation in an interactive translation environment where the model is adapted to the translation post-edited by human translators incrementally. Incremental domain adaptation has been shown to be useful for computer aided translation in NMT. Kothur et al. [74] apply

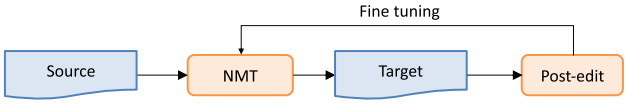


Fig. 11 Incremental domain adaptation.

fine tuning in an incremental manner, where the sentence already post-edited by a human translator is used to fine tune the model to translate the next sentence in a document. They show that incremental fine tuning can correctly translate novel vocabulary items and adapt to document-specific terminology usage and style. Turchi et al. [122] study different ways of incremental domain adaptation, including fine tuning on the current post-edited sentences, or similar sentences retrieved from the training and previously post-edited sentences based on the current source sentence, and both. Peris and Casacuberta [99] find the importance of the learning rate during fine tuning based incremental domain adaptation. Wuebker et al. [138] apply group lasso regularization in incremental domain adaptation, which freezes most of the parameters and thus makes the adapted model compact. Karimova et al. [63] conduct a user study for incremental domain adaptation, and find that it can significantly reduce the human post-editing cost. Simianer et al. [112] address the evaluation problem in incremental domain adaptation. They propose three content word based metrics to measure zero- and one-shot vocabulary acquisition in incremental adaptation and compare several fine tuning based domain adaptation methods. **Figure 11** shows an overview of incremental domain adaptation.

5.3 Multilingual and Multi-Domain Adaptation

It may not always be possible to use an out-of-domain parallel corpus in the same language pair and thus it is important to use data from other languages [62]. This approach is known as cross-lingual transfer learning, which transfers NMT model parameters among multiple languages [25]. It is known that a multilingual model, which relies on parameter sharing, helps in improving the translation quality for low resource languages especially when the target language is the same [149]. Even if out-of-domain data in the same language pair exists, it is possible that using both multilingual and multi-domain data can boost the translation performance. There are studies where either multilingual [40], [61] or multi-domain models [104] are trained. Recently, many studies also have been conducted on using multilingual and multi-domain data for in-domain data adaptation.

Chu and Dabre [18] conduct a preliminary study for this topic, where they apply mixed fine tuning using out-of-domain data from language pairs different from the in-domain data. Chu and Dabre [19] focus on training a single translation model for multiple domains by either learning domain-specific hidden state representations or predictor biases for each domain, and incorporate multilingualism into the domain adaptation framework. Imankulova et al. [56] use out-of-domain data from other languages to train a multilingual NMT model, and then fine tune it on in-domain parallel and back-translated pseudo-parallel data. Dabre et al. [26] apply multi-stage fine tuning using out-of-domain data from other languages to improve one-to-many in-domain translation in a N-way corpus setting. They find that multi-stage fine tuning performs better than single-stage fine tuning in that setting. Bapna and Firat [7] improve the scalability of fine-tuning for both domain adaptation and multilingual NMT. Instead of fine-tuning the entire NMT system, they propose using light-weight adapter layers that are suitable for the target task.

6. Datasets and Resources

Supervised MT is usually evaluated on several typical datasets, such as the WMT English to French and English to German shared tasks, the NIST Chinese to English shared task, and the WAT Japanese to English shared task. In this section, we summarize the mostly used datasets and resources for domain adaptation in MT. As shown in **Table 1**, IWSLT is used as a typical in-domain dataset and WMT is used as a typical out-of-domain dataset. The reason is that IWSLT contains TED talks and speeches, which contain approximately 100–200 k domain-specific sentences and are quite different from other corpora. WMT contains more than 10 million sentences and the domains are quite general; therefore, it is usually used as the out-of-domain or general-domain corpus.

There are also many domain specific parallel corpora such as restaurant reviews [8], clinical [114], civil engineering [49] and crisis-related texts [11], which could be good resources for studying NMT domain adaptation.

7. Future Directions

Through this survey, we see various studies trying to address domain adaptation for NMT. However, there are still unaddressed problems that need further investigation.

Universal Domain Adaptation Model. Based on different data scenarios, we have summarized different approaches for domain adaptation. However, most of these approaches are independent from others. Can we develop a model that is universal and robust for all data scenarios? Achieving this will significantly reduce the deployment footprint for domain-specific translation.

Error Tracking in Domain Adaptation. When apply a domain adaptation approach but get unsatisfied performance, how to track the errors remains an unclear problem. Are they coming from the out-of-domain model or the domain adaptation approach? There is no easy way to analyze this. For instance, although fine tuning is a very promising in domain adaptation, it is unclear how the errors in the out-of-domain model will affect the performance in the in-domain model. Therefore, studying an error tracking mechanism in domain adaptation is important.

Domain Specific Dictionary Incorporation. How to use external knowledge such as dictionaries and knowledge bases for NMT remains a big research question. In domain adaptation, the use of domain specific dictionaries is a very crucial problem. In the practical perspective, many translation companies have created domain specific dictionaries but not domain specific corpora. If we can study a good way to use domain specific dictionaries, it will significantly promote the practical use of MT. There are some studies that try to use dictionaries for NMT, but the usage is limited to help low frequent or rare word translation [3], [143]. Arcan and Buitelaar [1] use a domain specific dictionary for terminology translation, but they simply apply the unknown word

Table 1 The corpora used for domain adaptation in NMT.

Studies	Language pairs	In-domain corpora	Out-of-domain corpora
[128], [129], [130]	En-De & En-Fr	IWSLT	WMT
[20]	Zh-En	IWSLT	NTCIR
	Zh-Ja	WIKI-CJ	ASPEC
[13]	Zh-En	Dev data of NIST	Training data of NIST
	En-Fr	Dev data of WMT	Training data of WMT
[87]	En-Fr & En-De & En-Es	Certain speaker of IWSLT	IWSLT
[124]	En-De	Dev data of TED & WMT & Movie dialogues & EMEA medical	Training data of TED & WMT & Movie dialogues & EMEA medical
[146]	En-De & Ru-En	IWSLT & Patents	Paracrawl
[38]	En-Fr	Multi-domain	Multi-domain
[146]	Ru-En & En-De	IWSLT & Patent	Web-crawled
[46]	En-Zh	Laws	LDC
	En-De	News Commentary	WMT (NEWS)

replacement method proposed by Luong et al. [82], which suffers from noisy attention.

Domain Generation. Most of the existing methods focus on adapting from a general domain into a specific domain. In many scenarios, training data and test data have different distributions and the target domains are sometimes unseen. Irvine et al. [57] analyze the translation errors in such scenarios. Domain generalization aims to apply knowledge gained from labeled source domains to unseen target domains [79]. It provides a way to match the distribution of training data and test data in MT, which may be a future trend of domain adaptation for NMT.

Unsupervised NMT. Unsupervised NMT [2], [76] has achieved remarkable success. Because unsupervised NMT only uses monolingual data for training, the data scenarios differ from supervised NMT in domain adaptation. Sun et al. [116] summarize the different domain data scenarios for unsupervised NMT. Under such scenarios, how to conduct domain adaptation could be an interesting future direction.

8. Conclusion

Domain adaptation for NMT is a rather new but very important research topic to promote MT for practical use. In this paper, we gave a comprehensive survey of the techniques mainly being developed in the last four years. We compared domain adaptation techniques for NMT with the techniques being studied in SMT, which has been the main research area in the last two decades. In addition, we outlooked the future research directions. Connecting domain adaptation techniques in MT to the techniques in general NLP, computer vision and machine learning in detail is our future work. We hope that this survey paper could significantly promote research in domain adaptation for MT.

Acknowledgments This work was supported by Grant-in-Aid for Young Scientists #19K20343, JSPS and Microsoft Research Asia Collaborative Research Grant. We are very indebted to Dr. Raj Dabre for the deep discussion of the structure for this paper. We also thank the anonymous reviewers for their insightful comments.

References

[1] Arcan, M. and Buitelaar, P.: Translating Domain-Specific Expressions in Knowledge Bases with Neural Machine Translation, *CoRR*, Vol.abs/1709.02184 (2017).

[2] Artetxe, M., Labaka, G., Agirre, E. and Cho, K.: Unsupervised Neural Machine Translation, *International Conference on Learning Representations* (2018).

[3] Arthur, P., Neubig, G. and Nakamura, S.: Incorporating Discrete Translation Lexicons into Neural Machine Translation, *Proc. 2016 Conference on Empirical Methods in Natural Language Processing*, pp.1557–1567, Association for Computational Linguistics (2016).

[4] Axelrod, A., He, X. and Gao, J.: Domain Adaptation via Pseudo In-Domain Data Selection, *Proc. 2011 Conference on Empirical Methods in Natural Language Processing*, pp.355–362 (2011).

[5] Bahdanau, D., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *Proc. 3rd International Conference on Learning Representations* (*ICLR 2015*), International Conference on Learning Representations (2015).

[6] Banerjee, P., Du, J., Li, B., Naskar, S., Way, A. and Genabith, J.: Combining multi-domain statistical machine translation models using automatic classifiers, *The 9th Conference of the Association for Machine Translation in the Americas*, Denver, Colorado (2010).

[7] Bapna, A. and Firat, O.: Simple, Scalable Adaptation for Neural Machine Translation, *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), pp.1538–1548, Association for Computational Linguistics (online), DOI: 10.18653/v1/D19-1165 (2019).

[8] Berard, A., Calapodescu, I., Dymetman, M., Roux, C., Meunier, J.-L. and Nikoulina, V.: Machine Translation of Restaurant Reviews: New Corpus for Domain Adaptation and Robustness, *Proc. 3rd Workshop on Neural Generation and Translation*, pp.168–176, Association for Computational Linguistics (2019).

[9] Bisazza, A., Ruiz, N. and Federico, M.: Fill-up versus interpolation methods for phrase-based SMT adaptation, *IWSLT*, pp.136–143, ISCA (2011).

[10] Britz, D., Le, Q. and Pryzant, R.: Effective Domain Mixing for Neural Machine Translation, *Proc. 2nd Conference on Machine Translation*, pp.118–126, Association for Computational Linguistics (2017).

[11] Cadwell, P., O'Brien, S. and DeLuca, E.: More than tweets: A critical reflection on developing and testing crisis machine translation technology, *Translation Spaces*, Vol.8, pp.300–333 (online), DOI: 10.1075/ts.19018.cad (2019).

[12] Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Cattoni, R. and Federico, M.: The IWSLT 2015 Evaluation Campaign, *Proc. 12th International Workshop on Spoken Language Translation* (*IWSLT*) (2015).

[13] Chen, B., Cherry, C., Foster, G. and Larkin, S.: Cost Weighting for Neural Machine Translation Domain Adaptation, *Proc. 1st Workshop on Neural Machine Translation*, pp.40–46 (2017).

[14] Chen, B., Kuhn, R., Foster, G., Cherry, C. and Huang, F.: Bilingual Methods for Adaptive Training Data Selection for Machine Translation, *The 12th Conference of The Association for Machine Translation in the Americas*, pp.93–106 (2016).

[15] Cheng, Y., Xu, W., He, Z., He, W., Wu, H., Sun, M. and Liu, Y.: Semi-Supervised Learning for Neural Machine Translation, *Proc. 54th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp.1965–1974, Association for Computational Linguistics (2016).

[16] Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, *Proc. 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp.1724–1734, Association for Computational Linguistics (2014).

[17] Chu, C.: *Integrated Parallel Data Extraction from Comparable Corpora for Statistical Machine Translation*, Doctoral Thesis, Kyoto University (2015).

[18] Chu, C. and Dabre, R.: Multilingual and Multi-Domain Adaptation for Neural Machine Translation, *Proc. 24th Annual Meeting of the Association for Natural Language Processing* (*NLP 2018*), pp.909–912 (2018).

[19] Chu, C. and Dabre, R.: Multilingual Multi-Domain Adaptation Approaches for Neural Machine Translation, *CoRR*, Vol.abs/1906.07978 (2019).

[20] Chu, C., Dabre, R. and Kurohashi, S.: An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation, *Proc. 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, Association for Computational Linguistics (2017).

[21] Chu, C., Dabre, R. and Kurohashi, S.: A Comprehensive Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation, *Journal of Information Processing* (*JIP*), Vol.26, No.1, pp.1–10 (2018).

[22] Chu, C. and Wang, R.: A Survey of Domain Adaptation for Neural Machine Translation, *Proc. 27th International Conference on Computational Linguistics*, pp.1304–1319, Association for Computational Linguistics (2018).

[23] Csurka, G.: Domain Adaptation for Visual Applications: A Comprehensive Survey, *CoRR*, Vol.abs/1702.05374 (2017).

[24] Currey, A., Miceli Barone, A.V. and Heafield, K.: Copied Monolingual Data Improves Low-Resource Neural Machine Translation, *Proc. 2nd Conference on Machine Translation*, pp.148–156, Association for Computational Linguistics (2017).

[25] Dabre, R., Chu, C. and Kunchukuttan, A.: A Survey of Multilingual Neural Machine Translation, *CoRR*, Vol.abs/1905.05395 (2019).

[26] Dabre, R., Fujita, A. and Chu, C.: Exploiting Multilingualism through Multistage Fine-Tuning for Low-Resource Neural Machine Translation, *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), pp.1410–1416, Association for Computational Linguistics (online), DOI: 10.18653/v1/D19-1146 (2019).

[27] Dabre, R., Kunchukuttan, A., Bhattacharyya, P., Fujita, A. and Sumita, E.: NICT's Participation in WAT 2018: Approaches Using Multilingualism and Recurrently Stacked Layers, *Proc. 5th Workshop on Asian Language Translation*, Hong Kong, China (2018).

[28] Dakwale, P. and Monz, C.: Fine-Tuning for Neural Machine Translation with Limited Degradation across In- and Out-of-Domain Data, *Proc. 16th Machine Translation Summit* (*MT-Summit 2017*), pp.156–169 (2017).

[29] Ding, L., He, Y., Zhou, L. and Qingmin, L.: Combining Domain Knowledge and Deep Learning Makes NMT More Adaptive, *Proc. 13th China Workshop on Machine Translation* (*CWMT 2017*), Dalian, China (2017).

[30] Domhan, T. and Hieber, F.: Using Target-side Monolingual Data for Neural Machine Translation through Multi-task Learning, *Proc. 2017 Conference on Empirical Methods in Natural Language Processing*, pp.1500–1505, Association for Computational Linguistics (2017).

[31] Dou, Z.-Y., Hu, J., Anastasopoulos, A. and Neubig, G.: Unsupervised Domain Adaptation for Neural Machine Translation with Domain-Aware Feature Embeddings, *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), pp.1417–1422, Association for Computational Linguistics (2019).

[32] Dou, Z.-Y., Wang, X., Hu, J. and Neubig, G.: Domain Differential Adaptation for Neural Machine Translation, *Proc. 3rd Workshop on Neural Generation and Translation*, pp.59–69, Association for Computational Linguistics (online), DOI: 10.18653/v1/D19-5606 (2019).

[33] Duh, K., Neubig, G., Sudoh, K. and Tsukada, H.: Adaptation Data Selection using Neural Language Models: Experiments in Machine Translation, *Proc. 51st Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), pp.678–683 (2013).

[34] Durrani, N., Sajjad, H., Joty, S., Abdelali, A. and Vogel, S.: Using Joint Models for Domain Adaptation in Statistical Machine Translation, *Proc. MT Summit XV*, pp.117–130 (2015).

[35] Dyer, C., Muresan, S. and Resnik, P.: Generalizing Word Lattice Translation, *Proc. ACL-08: HLT*, pp.1012–1020, Association for Computational Linguistics (2008).

[36] Edunov, S., Ott, M., Auli, M. and Grangier, D.: Understanding Back-Translation at Scale, *Proc. 2018 Conference on Empirical Methods in Natural Language Processing*, pp.489–500, Association for Computational Linguistics (online), DOI: 10.18653/v1/D18-1045 (2018).

[37] Farajian, M.A., Bertoldi, N., Negri, M., Turchi, M. and Federico, M.: Evaluation of Terminology Translation in Instance-Based Neural MT Adaptation, *Proc. 21st Annual Conference of the European Association for Machine Translation*, pp.149–158 (2018).

[38] Farajian, M.A., Turchi, M., Negri, M., Bertoldi, N. and Federico, M.: Neural vs. Phrase-Based Machine Translation in a Multi-Domain Scenario, *Proc. 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp.280–284, Association for Computational Linguistics (2017).

[39] Farajian, M.A., Turchi, M., Negri, M. and Federico, M.: Multi-Domain Neural Machine Translation through Unsupervised Adaptation, *Proc. 2nd Conference on Machine Translation*, pp.127–137, Association for Computational Linguistics (2017).

[40] Firat, O., Cho, K. and Bengio, Y.: Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.866–875 (2016).

[41] Fonseca, E., Yankovskaya, L., Martins, A.F.T., Fishel, M. and Federmann, C.: Findings of the WMT 2019 Shared Tasks on Quality Estimation, *Proc. 4th Conference on Machine Translation* (*Volume 3: Shared Task Papers, Day 2*), pp.1–10, Association for Computational Linguistics (online), DOI: 10.18653/v1/W19-5401 (2019).

[42] Foster, G., Goutte, C. and Kuhn, R.: Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation, *Proc. 2010 Conference on Empirical Methods in Natural Language Processing*, pp.451–459 (2010).

[43] Foster, G. and Kuhn, R.: Mixture-model Adaptation for SMT, *Proc. 2nd Workshop on Statistical Machine Translation, StatMT '07*, Stroudsburg, pp.128–135, Association for Computational Linguistics (2007).

[44] Freitag, M. and Al-Onaizan, Y.: Fast Domain Adaptation for Neural Machine Translation, arXiv preprint arXiv:1612.06897 (2016).

[45] Goto, I., Chow, K.-P., Lu, B., Sumita, E. and Tsou, B.K.: Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop, *Proc. 10th NTCIR Conference*, pp.260–286, National Institute of Informatics (NII) (2013).

[46] Gu, S., Feng, Y. and Liu, Q.: Improving Domain Adaptation Translation with Domain Invariant and Specific Information, *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (*Long and Short Papers*), pp.3081–3091, Association for Computational Linguistics (2019).

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

[48] He, X., Haffari, G. and Norouzi, M.: Sequence to Sequence Mixture Model for Diverse Machine Translation, *Proc. 22nd Conference on Computational Natural Language Learning*, pp.583–592, Association for Computational Linguistics (2018).

[49] Hedberg, L.E., Labaka, G. and Gojenola, K.: Spanish-Swedish Neural Machine Translation for the Civil Engineering Domain (2019).

[50] Hinton, G., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, *NIPS Deep Learning and Representation Learning Workshop* (2015).

[51] Hira, N.-E., Abdul Rauf, S., Kiani, K., Zafar, A. and Nawaz, R.: Exploring Transfer Learning and Domain Data Selection for the Biomedical Translation, *Proc. 4th Conference on Machine Translation* (*Volume 3: Shared Task Papers, Day 2*), pp.156–163, Association for Computational Linguistics (online), DOI: 10.18653/v1/W19-5419 (2019).

[52] Hoang, C. and Sima'an, K.: Latent Domain Translation Models in Mix-of-Domains Haystack, *Proc. 25th International Conference on Computational Linguistics: Technical Papers*, pp.1928–1939 (2014).

[53] Hu, J., Xia, M., Neubig, G. and Carbonell, J.: Domain Adaptation of Neural Machine Translation by Lexicon Induction, *The 57th Annual Meeting of the Association for Computational Linguistics* (*ACL*), Florence, Italy (2019).

[54] Huck, M., Birch, A. and Haddow, B.: Mixed-Domain vs. Multi-Domain Statistical Machine Translation, *Proc. MT Summit XV*, Vol.1, pp.240–255 (2015).

[55] Imamura, K. and Sumita, E.: Multi-domain Adaptation for Statistical Machine Translation Based on Feature Augmentation, *Proc. 12th Conference of the Association for Machine Translation in the Americas*, Austin, Texas, USA (2016).

[56] Imankulova, A., Dabre, R., Fujita, A. and Imamura, K.: Exploiting Out-of-Domain Parallel Data through Multilingual Transfer Learning for Low-Resource Neural Machine Translation, *Proc. Machine Translation Summit XVII Volume 1: Research Track*, pp.128–139, European Association for Machine Translation (2019).

[57] Irvine, A., Morgan, J., Carpuat, M., III, H.D. and Munteanu, D.: Measuring Machine Translation Errors in New Domains, *Trans. Association for Computational Linguistics*, Vol.1, pp.429–440 (2013).

[58] Jean, S., Cho, K., Memisevic, R. and Bengio, Y.: On Using Very Large Target Vocabulary for Neural Machine Translation, *Proc. 53rd*

Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (*Volume 1: Long Papers*), pp.1–10, Association for Computational Linguistics (2015).

[59] Jehl, L. and Riezler, S.: Document-Level Information as Side Constraints for Improved Neural Patent Translation, *Proc. 13th Conference of the Association for Machine Translation in the Americas* (*Volume 1: Research Papers*), pp.1–12, Association for Machine Translation in the Americas (2018).

[60] Jiang, J. and Zhai, C.: Instance Weighting for Domain Adaptation in NLP, *Proc. 45th Annual Meeting of the Association of Computational Linguistics*, pp.264–271, Czech Republic (2007).

[61] Johnson, M., Schuster, M., Le, Q., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M. and Dean, J.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, *Trans. Association for Computational Linguistics*, Vol.5, pp.339–351 (2017).

[62] Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M. and Dean, J.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, *Tran. Association for Computational Linguistics*, Vol.5, pp.339–351 (2017) (online), available from ⟨https://www.aclweb.org/anthology/Q17-1024⟩.

[63] Karimova, S., Simianer, P. and Riezler, S.: A User-study on Online Adaptation of Neural Machine Translation to Human Post-edits, *Machine Translation*, Vol.32, No.4, pp.309–324 (online), DOI: 10.1007/s10590-018-9224-8 (2018).

[64] Kawara, Y., Takebayashi, Y., Chu, C. and Arase, Y.: Osaka University MT Systems forWAT 2018: Rewarding, Preordering, and Domain Adaptation, *Proc. 5th Workshop on Asian Language Translation*, Hong Kong, China (2018).

[65] Khan, A., Panda, S., Xu, J. and Flokas, L.: Hunter NMT System for WMT18 Biomedical Translation Task: Transfer Learning in Neural Machine Translation, *Proc. 3rd Conference on Machine Translation: Shared Task Papers*, pp.655–661, Association for Computational Linguistics (2018).

[66] Khayrallah, H., Kumar, G., Duh, K., Post, M. and Koehn, P.: Neural Lattice Search for Domain Adaptation in Machine Translation, *Proc. 8th International Joint Conference on Natural Language Processing* (*Volume 2: Short Papers*), pp.20–25, Asian Federation of Natural Language Processing (2017).

[67] Khayrallah, H., Thompson, B., Duh, K. and Koehn, P.: Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation, *Proc. 2nd Workshop on Neural Machine Translation and Generation*, pp.36–44, Association for Computational Linguistics (2018).

[68] Kobus, C., Crego, J. and Senellart, J.: Domain Control for Neural Machine Translation, arXiv preprint arXiv:1612.06140 (2016).

[69] Kocmi, T., Variš, D. and Bojar, O.: CUNI NMT System for WAT 2017 Translation Tasks, *Proc. 4th Workshop on Asian Translation* (*WAT2017*), pp.154–159, Asian Federation of Natural Language Processing (2017).

[70] Koehn, P.: Neural Machine Translation, *CoRR*, Vol.abs/1709.07809 (2017).

[71] Koehn, P., Duh, K. and Thompson, B.: The JHU Machine Translation Systems for WMT 2018, *Proc. 3rd Conference on Machine Translation: Shared Task Papers*, pp.438–444, Association for Computational Linguistics (2018).

[72] 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.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proc. 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proc. Demo and Poster Sessions*, pp.177–180, Association for Computational Linguistics (2007).

[73] Koehn, P. and Knowles, R.: Six Challenges for Neural Machine Translation, *Proc. 1st Workshop on Neural Machine Translation*, pp.28–39, Association for Computational Linguistics (2017).

[74] Kothur, S.S.R., Knowles, R. and Koehn, P.: Document-Level Adaptation for Neural Machine Translation, *Proc. 2nd Workshop on Neural Machine Translation and Generation*, pp.64–73, Association for Computational Linguistics (2018).

[75] Lambert, P., Schwenk, H., Servan, C. and Abdul-Rauf, S.: Investigations on Translation Model Adaptation Using Monolingual Data, *Proc. 6th Workshop on Statistical Machine Translation, WMT '11*, Stroudsburg, PA, USA, pp.284–293, Association for Computational Linguistics (2011).

[76] Lample, G., Conneau, A., Denoyer, L. and Ranzato, M.: Unsupervised Machine Translation Using Monolingual Corpora Only, *International Conference on Learning Representations* (2018).

[77] Li, R., Wang, X. and Yu, H.: MetaMT, a Meta Learning Method Leveraging Multiple Domain Data for Low Resource Machine Translation (2019).

[78] Li, X., Zhang, J. and Zong, C.: One Sentence One Model for Neural Machine Translation, *CoRR*, Vol.abs/1609.06490 (2016).

[79] Li, Y., Gong, M., Tian, X., Liu, T. and Tao, D.: Domain Generalization via Conditional Invariant Representations, *The 32nd AAAI Conference on Artificial Intelligence* (2018).

[80] Luo, L., Yang, H., Siu, S.C. and Chin, F. Y.L.: Neural Machine Translation for Financial Listing Documents, *Neural Information Processing*, Cheng, L., Leung, A.C.S. and Ozawa, S. (Eds.), pp.232–243, Springer International Publishing (2018).

[81] Luong, M.-T. and Manning, C.D.: Stanford Neural Machine Translation Systems for Spoken Language Domains, *Proc. 12th International Workshop on Spoken Language Translation*, pp.76–79 (2015).

[82] Luong, T., Sutskever, I., Le, Q., Vinyals, O. and Zaremba, W.: Addressing the Rare Word Problem in Neural Machine Translation, *Proc. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*Volume 1: Long Papers*), pp.11–19, Association for Computational Linguistics (2015).

[83] Mansour, S. and Ney, H.: A Simple and Effective Weighted Phrase Extraction for Machine Translation Adaptation, *The 9th International Workshop on Spoken Language Translation*, Hong Kong (2012).

[84] Marie, B. and Fujita, A.: Efficient Extraction of Pseudo-Parallel Sentences from Raw Monolingual Data Using Word Embeddings, *Proc. 55th Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), pp.392–398, Association for Computational Linguistics (2017).

[85] Matsoukas, S., Rosti, A.-V.I. and Zhang, B.: Discriminative Corpus Weight Estimation for Machine Translation, *Proc. 2009 Conference on Empirical Methods in Natural Language Processing*, pp.708–717 (2009).

[86] Miceli Barone, A.V., Haddow, B., Germann, U. and Sennrich, R.: Regularization techniques for fine-tuning in neural machine translation, *Proc. 2017 Conference on Empirical Methods in Natural Language Processing*, pp.1489–1494, Association for Computational Linguistics (2017).

[87] Michel, P. and Neubig, G.: Extreme Adaptation for Personalized Neural Machine Translation, *Proc. 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), pp.312–318, Association for Computational Linguistics (2018).

[88] Michel, P. and Neubig, G.: MTNT: A Testbed for Machine Translation of Noisy Text, *Proc. 2018 Conference on Empirical Methods in Natural Language Processing*, pp.543–553, Association for Computational Linguistics (2018).

[89] Mino, H., Ito, H., Goto, I., Yamada, I., Tanaka, H. and Tokunaga, T.: Neural Machine Translation System using a Content-equivalently Translated Parallel Corpus for the Newswire Translation Tasks at WAT 2019, *Proc. 6th Workshop on Asian Translation*, pp.106–111, Association for Computational Linguistics (online), DOI: 10.18653/v1/D19-5212 (2019).

[90] Moore, R.C. and Lewis, W.: Intelligent selection of language model training data, *Proc. 48th Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), pp.220–224 (2010).

[91] Nakazawa, T., Doi, N., Higashiyama, S., Ding, C., Dabre, R., Mino, H., Goto, I., Pa, W.P., Kunchukuttan, A., Parida, S., Bojar, O. and Kurohashi, S.: Overview of the 6th Workshop on Asian Translation, *Proc. 6th Workshop on Asian Translation*, pp.1–35, Association for Computational Linguistics (online), DOI: 10.18653/v1/D19-5201 (2019).

[92] Neubig, G.: Neural Machine Translation and Sequence-to-sequence Models: A Tutorial, *CoRR*, Vol.abs/1703.01619 (2017).

[93] Niehues, J. and Waibel, A.H.: Detailed Analysis of different Strategies for Phrase Table Adaptation in SMT, *Proc. Conference of the Association for Machine Translation in the Americas* (*AMTA*), San Diego, US-CA (2012).

[94] Niu, X., Denkowski, M. and Carpuat, M.: Bi-Directional Neural Machine Translation with Synthetic Parallel Data, *Proc. 2nd Workshop on Neural Machine Translation and Generation*, pp.84–91, Association for Computational Linguistics (2018).

[95] Pan, S.J. and Yang, Q.: A survey on transfer learning, *IEEE Trans. Knowledge and Data Engineering*, Vol.22, No.10, pp.1345–1359 (2009).

[96] Pan, S.J. and Yang, Q.: A Survey on Transfer Learning, *IEEE Trans. Knowledge and Data Engineering*, Vol.22, No.10, pp.1345–1359 (online), DOI: 10.1109/TKDE.2009.191 (2010).

[97] Park, J., Song, J. and Yoon, S.: Building a Neural Machine Translation System Using Only Synthetic Parallel Data, *CoRR*, Vol.abs/1704.00253 (2017).

[98] Peng, W., Liu, J., Li, L. and Liu, Q.: Huawei's NMT Systems

for the WMT 2019 Biomedical Translation Task, *Proc. 4th Conference on Machine Translation* (*Volume 3: Shared Task Papers, Day 2*), pp.164–168, Association for Computational Linguistics (online), DOI: 10.18653/v1/W19-5420 (2019).

[99] Peris, Á. and Casacuberta, F.: Online Learning for Effort Reduction in Interactive Neural Machine Translation, *CoRR*, Vol.abs/1802.03594 (2018).

[100] Pham, M.Q., Crego, J.-M., Yvon, F. and Senellart, J.: Generic and Specialized Word Embeddings for Multi-Domain Machine Translation, *International Workshop on Spoken Language Translation*, *Proc. 16th International Workshop on Spoken Language Translation* (*IWSLT*), Hong-Kong, China (online), DOI: 10.5281/zenodo.3524978 (2019).

[101] Poncelas, A., de Buy Wenniger, G.M. and Way, A.: Feature Decay Algorithms for Neural Machine Translation, *Proc. 21st Annual Conference of the European Association for Machine Translation*, pp.239–248 (2018).

[102] Poncelas, A., Way, A. and Sarasola, K.: The ADAPT System Description for the IWSLT 2018 Basque to English Translation Task, *Proc. 15th International Workshop on Spoken Language Translation* (*IWSLT*), pp.76–82 (2018).

[103] Rousseau, A., Bougares, F., Deléglise, P., Schwenk, H. and Estève, Y.: LIUMâs systems for the IWSLT 2011 Speech Translation Tasks, *International Workshop on Spoken Language Translation*, San Francisco, USA (2011).

[104] Sajjad, H., Durrani, N., Dalvi, F., Belinkov, Y. and Vogel, S.: Neural Machine Translation Training in a Multi-Domain Scenario, *Proc. 12th International Workshop on Spoken Language Translation* (*IWSLT*), Tokyo, Japan (2017).

[105] Sennrich, R., Haddow, B. and Birch, A.: Controlling Politeness in Neural Machine Translation via Side Constraints, *Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.35–40, Association for Computational Linguistics (2016).

[106] Sennrich, R., Haddow, B. and Birch, A.: Improving Neural Machine Translation Models with Monolingual Data, *Proc. 54th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp.86–96, Association for Computational Linguistics (2016).

[107] Sennrich, R., Haddow, B. and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, *Proc. 54th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp.1715–1725, Association for Computational Linguistics (2016).

[108] Sennrich, R., Schwenk, H. and Aransa, W.: A Multi-Domain Translation Model Framework for Statistical Machine Translation, *Proc. 51st Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp.832–840 (2013).

[109] Servan, C., Crego, J. and Senellart, J.: Domain specialization: A post-training domain adaptation for Neural Machine Translation, arXiv preprint arXiv:1612.06141 (2016).

[110] Shah, K., Barrault, L. and Schwenk, H.: Translation model adaptation by resampling, *Proc. Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pp.392–399 (2010).

[111] Shah, K., Barrault, L. and Schwenk, H.: A General Framework to Weight Heterogeneous Parallel Data for Model Adaptation in Statistical Machine Translation, *Proc. Conference of the Association for Machine Translation in the Americas* (*AMTA*), San Diego, US-CA (2012).

[112] Simianer, P., Wuebker, J. and DeNero, J.: Measuring Immediate Adaptation Performance for Neural Machine Translation, *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (*Long and Short Papers*), pp.2038–2046, Association for Computational Linguistics (2019).

[113] Sostaric, M., Pavlović, N. and Boltužić, F.: Domain adaptation for machine translation involving a low-resource language: Google AutoML vs. from-scratch NMT systems, *Proc. Translating and the Computer 41* (online), DOI: 10.13140/RG.2.2.17293.90087 (2019).

[114] Soto, X., Perez-de Viñaspre, O., Labaka, G. and Oronoz, M.: Neural machine translation of clinical texts between long distance languages, *Journal of the American Medical Informatics Association: JAMIA*, Vol.26 (online), DOI: 10.1093/jamia/ocz110 (2019).

[115] Su, J., Zeng, J., Xie, J., Wen, H., Yin, Y. and Liu, Y.: Exploring Discriminative Word-Level Domain Contexts for Multi-domain Neural Machine Translation, *IEEE Trans. Pattern Analysis and Machine Intelligence* (2019).

[116] Sun, H., Wang, R., Chen, K., Utiyama, M., Sumita, E. and Zhao, T.: An Empirical Study of Domain Adaptation for Unsupervised Neural Machine Translation (2019).

[117] Sutskever, I., Vinyals, O. and Le, Q.V.: Sequence to Sequence Learning with Neural Networks, *Proc. 27th International Conference on Neural Information Processing Systems*, pp.3104–3112, MIT Press (2014).

[118] Tars, S. and Fishel, M.: Multi-Domain Neural Machine Translation, *Proc. 21st Annual Conference of the European Association for Machine Translation*, Alacant, Spain, pp.259–268 (2018).

[119] Thompson, B., Gwinnup, J., Khayrallah, H., Duh, K. and Koehn, P.: Overcoming Forgetting During Domain Adaptation of Neural Machine Translation, *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (*Long and Short Papers*), pp.2062–2068, Association for Computational Linguistics (2019).

[120] Thompson, B., Khayrallah, H., Anastasopoulos, A., McCarthy, A.D., Duh, K., Marvin, R., McNamee, P., Gwinnup, J., Anderson, T. and Koehn, P.: Freezing Subnetworks to Analyze Domain Adaptation in Neural Machine Translation, *Proc. 3rd Conference on Machine Translation: Research Papers*, pp.124–132, Association for Computational Linguistics (2018).

[121] Torrey, L. and Shavlik, J.: Transfer learning, *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques*, pp.242–264, IGI Global (2010).

[122] Turchi, M., Negri, M., Farajian, M.A. and Federico, M.: Continuous learning from human post-edits for neural machine translation, *The Prague Bulletin of Mathematical Linguistics*, Vol.108, No.1, pp.233–244 (2017).

[123] Utiyama, M. and Isahara, H.: Reliable Measures for Aligning Japanese-English News Articles and Sentences, *Proc. 41st Annual Meeting of the Association for Computational Linguistics*, pp.72–79, Association for Computational Linguistics (online), DOI: 10.3115/1075096.1075106 (2003).

[124] van der Wees, M., Bisazza, A. and Monz, C.: Dynamic Data Selection for Neural Machine Translation, *Proc. 2017 Conference on Empirical Methods in Natural Language Processing*, pp.1400–1410, Association for Computational Linguistics (2017).

[125] Varga, A.C.: *Domain Adaptation for Multilingual Neural Machine Translation*, Master Thesis, Saarlandes University (2017).

[126] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.U. and Polosukhin, I.: Attention is All you Need, *Advances in Neural Information Processing Systems 30*, Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R. (Eds.), pp.5998–6008, Curran Associates, Inc. (2017).

[127] Vilar, D.: Learning Hidden Unit Contribution for Adapting Neural Machine Translation Models, *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2* (*Short Papers*), pp.500–505, Association for Computational Linguistics (online), DOI: 10.18653/v1/N18-2080 (2018).

[128] Wang, R., Finch, A., Utiyama, M. and Sumita, E.: Sentence Embedding for Neural Machine Translation Domain Adaptation, *Proc. 55th Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), pp.560–566, Association for Computational Linguistics (2017).

[129] Wang, R., Utiyama, M., Finch, A., Liu, L., Chen, K. and Sumita, E.: Sentence Selection and Weighting for Neural Machine Translation Domain Adaptation, *IEEE/ACM Trans. Audio, Speech, and Language Processing*, Vol.26, No.10, pp.1727–1741 (online), DOI: 10.1109/TASLP.2018.2837223 (2018).

[130] Wang, R., Utiyama, M., Liu, L., Chen, K. and Sumita, E.: Instance Weighting for Neural Machine Translation Domain Adaptation, *Proc. 2017 Conference on Empirical Methods in Natural Language Processing*, pp.1482–1488 (2017).

[131] Wang, R., Zhao, H., Lu, B.-L., Utiyama, M. and Sumita, E.: Neural Network Based Bilingual Language Model Growing for Statistical Machine Translation, *Proc. 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp.189–195, Association for Computational Linguistics (2014).

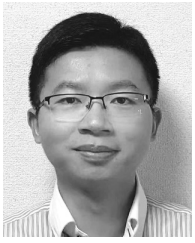
[132] Wang, R., Zhao, H., Lu, B.-L., Utiyama, M. and Sumita, E.: Connecting Phrase based Statistical Machine Translation Adaptation, *Proc. COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp.3135–3145, The COLING 2016 Organizing Committee (2016).

[133] Wang, W., Macherey, K., Macherey, W., Och, F. and Xu, P.: Improved domain adaptation for statistical machine translation, *Proc. AMTA*, San Diego, California, USA (2012).

[134] Wang, W., Watanabe, T., Hughes, M., Nakagawa, T. and Chelba, C.: Denoising Neural Machine Translation Training with Trusted Data and Online Data Selection, *Proc. 3rd Conference on Machine Translation: Research Papers*, pp.133–143, Association for Computational Linguistics (2018).

[135] Wang, Y., Wang, L., Shi, S., Li, V.O.K. and Tu, Z.: Go From the General to the Particular: Multi-Domain Translation with Domain

Transformation Networks, *The 34th AAAI Conference on Artificial Intelligence* (2020).
[136] Weiss, K., Khoshgoftaar, T.M. and Wang, D.: A survey of transfer learning, *Journal of Big Data*, Vol.3, No.1, p.9 (online), DOI: 10.1186/s40537-016-0043-6 (2016).
[137] Wu, S., Zhang, D. and Zhou, M.: *Effective Soft-Adaptation for Neural Machine Translation*, pp.254–264 (2019).
[138] Wuebker, J., Simianer, P. and DeNero, J.: Compact Personalized Models for Neural Machine Translation, *Proc. 2018 Conference on Empirical Methods in Natural Language Processing*, pp.881–886, Association for Computational Linguistics (2018).
[139] Xu, J., Deng, Y., Gao, Y. and Ney, H.: Domain dependent statistical machine translation, *MT Summit*, Copenhagen, Denmark (2007).
[140] Yan, S., Dahlmann, L., Petrushkov, P., Hewavitharana, S. and Khadivi, S.: Word-based Domain Adaptation for Neural Machine Translation, *Proc. 15th International Workshop on Spoken Language Translation* (*IWSLT*), pp.31–38 (2018).
[141] Zeng, J., Liu, Y., Su, J., Ge, Y., Lu, Y., Yin, Y. and Luo, J.: Iterative Dual Domain Adaptation for Neural Machine Translation, *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), pp.845–854, Association for Computational Linguistics (2019).
[142] Zeng, J., Su, J., Wen, H., Liu, Y., Xie, J., Yin, Y. and Zhao, J.: Multi-Domain Neural Machine Translation with Word-Level Domain Context Discrimination, *Proc. 2018 Conference on Empirical Methods in Natural Language Processing*, pp.447–457, Association for Computational Linguistics (2018).
[143] Zhang, J. and Zong, C.: Bridging Neural Machine Translation and Bilingual Dictionaries, *CoRR*, Vol.abs/1610.07272 (2016).
[144] Zhang, J. and Zong, C.: Exploiting Source-side Monolingual Data in Neural Machine Translation, *Proc. 2016 Conference on Empirical Methods in Natural Language Processing*, pp.1535–1545, Association for Computational Linguistics (2016).
[145] Zhang, S. and Xiong, D.: Sentence Weighting for Neural Machine Translation Domain Adaptation, *Proc. 27th International Conference on Computational Linguistics*, pp.3181–3190, Association for Computational Linguistics (2018).
[146] Zhang, X., Shapiro, P., Kumar, G., McNamee, P., Carpuat, M. and Duh, K.: Curriculum Learning for Domain Adaptation in Neural Machine Translation, *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (*Long and Short Papers*), pp.1903–1915, Association for Computational Linguistics (2019).
[147] Zheng, R., Liu, H., Ma, M., Zheng, B. and Huang, L.: Robust Machine Translation with Domain Sensitive Pseudo-Sources: Baidu-OSU WMT19 MT Robustness Shared Task System Report, *Proc. 4th Conference on Machine Translation*, Association for Computational Linguistics (2019).
[148] Zhou, X., Cao, H. and Zhao, T.: Domain adaptation for SMT using sentence weight, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pp.153–163 (2015).
[149] Zoph, B., Yuret, D., May, J. and Knight, K.: Transfer Learning for Low-Resource Neural Machine Translation, *Proc. 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1–4, 2016*, pp.1568–1575 (2016).



Chenhui Chu received his B.S. degree in software engineering from Chongqing University in 2008, and his M.S. and Ph.D. degrees in Informatics from Kyoto University in 2012 and 2015, respectively. He is currently a research assistant professor at Osaka University. His research interests include natural language processing, particularly machine translation and multimodal machine learning.

Rui Wang is a tenure-track researcher at the National Institute of Information and Communications Technology, Japan. He received his B.S. degree from Harbin Institute of Technology in 2009, his M.S. degree from the Chinese Academy of Sciences in 2012, and his Ph.D. degree from Shanghai Jiao Tong University in 2016, all of which are in computer science. He received a joint Ph.D. at Centre Nationnal de la Recherche Scientifique, France in 2014. His research interests are machine translation and natural language processing.