A Survey of Domain Adaptation for Neural Machine Translation

Chenhui Chu
Institute for Datability Science
Osaka University
chu@ids.osaka-u.ac.jp

Rui Wang
National Institute of Information
and Communications Technology
wangrui@nict.go.jp

Abstract

Neural machine translation (NMT) is a deep learning based approach for machine translation, which yields the state-of-the-art translation performance in scenarios where large-scale parallel corpora are available. Although the 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 NMT.

1 Introduction

Neural machine translation (NMT) (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015) allows for end-to-end training of a translation system without the need to deal with word alignments, translation rules and complicated decoding algorithms, which are characteristics of statistical machine translation (SMT) systems (Koehn et al., 2007). NMT yields the state-of-the-art translation performance in resource rich scenarios (Bojar et al., 2017; Nakazawa et al., 2017). However, currently, high quality parallel corpora of sufficient size are only available for a few language pairs such as languages paired with English and several 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 (Duh et al., 2013; Sennrich et al., 2013; Zoph et al., 2016; Koehn and Knowles, 2017).

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 (Koehn and Knowles, 2017). Leveraging out-of-domain parallel corpora and in-domain monolingual corpora to improve in-domain translation is known as domain adaptation for MT (Wang et al., 2016; Chu et al., 2018). For example, the Chinese-English patent domain parallel corpus has 1M sentence pairs (Goto et al., 2013), but for the spoken language domain parallel corpus there are only 200k sentences available (Cettolo et al., 2015). 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 (Chu et al., 2017). Furthermore, there are monolingual corpora containing millions of sentences for the spoken language domain, which can also be leveraged (Sennrich et al., 2016b).

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 a language model (LM) (Moore and Lewis, 2010; Axelrod et al., 2011; Duh et al., 2013; Hoang and Sima'an, 2014; Durrani et al., 2015; Chen et al., 2016) or generating pseudo parallel data (Utiyama and Isahara, 2003; Wang et al., 2014; Chu, 2015; Wang et al.,

Proceedings of the 27th International Conference on Computational Linguistics, pages 1304–1319
Santa Fe, New Mexico, USA, August 20-26, 2018.

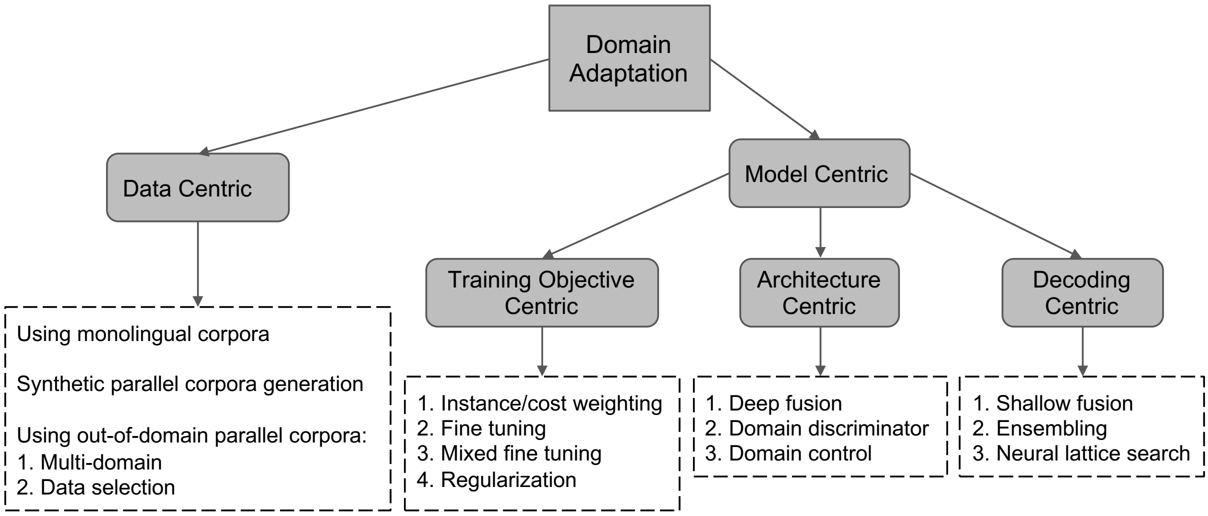

Figure 1: Overview of domain adaptation for NMT.

2016; Marie and Fujita, 2017). Model centric methods interpolate in-domain and out-of-domain models in either a model level (Sennrich et al., 2013; Durrani et al., 2015; Imamura and Sumita, 2016) or an instance level (Matsoukas et al., 2009; Foster et al., 2010; Shah et al., 2010; Rousseau et al., 2011; Zhou et al., 2015). 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 two 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 is no single 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 (Neubig, 2017; Koehn, 2017); however, they focus on general NMT and more diverse topics. Domain adaptation surveys have been done in the perspective of computer vision (Csurka, 2017) and machine learning (Pan and Yang, 2010; Weiss et al., 2016). However, such survey has not been done for NMT. To the best of our knowledge, this is the first comprehensive survey of domain adaptation for NMT.

In this paper, similar to SMT, we categorize domain adaptation 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 (Zhang and Zong, 2016b; Cheng et al., 2016; Currey et al., 2017; Domhan and Hieber, 2017), synthetic corpora (Sennrich et al., 2016b; Zhang and Zong, 2016b; Park et al., 2017), or parallel copora (Chu et al., 2017; Sajjad et al., 2017; Britz et al., 2017; Wang et al., 2017a; van der Wees et al., 2017). 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 (Luong and Manning, 2015; Sennrich et al., 2016b; Servan et al., 2016; Freitag and Al-Onaizan, 2016; Wang et al., 2017b; Chen et al., 2017a; Varga, 2017; Dakwale and Monz, 2017; Chu et al., 2017; Miceli Barone et al., 2017), the NMT architecture (Kobus et al., 2016; Gülçehre et al., 2015; Britz et al., 2017) or the decoding algorithm (Gülçehre et al., 2015; Dakwale and Monz, 2017; Khayrallah et al., 2017). An overview of these two categories is shown in Figure 1. Note that as model centric methods also use either monolingual or parallel corpora, there are overlaps between these two categories.

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 briefly review the historical domain adaptation techniques being developed for SMT (Section 3); Under these background knowledge, we then present and compare the domain adaptation methods for

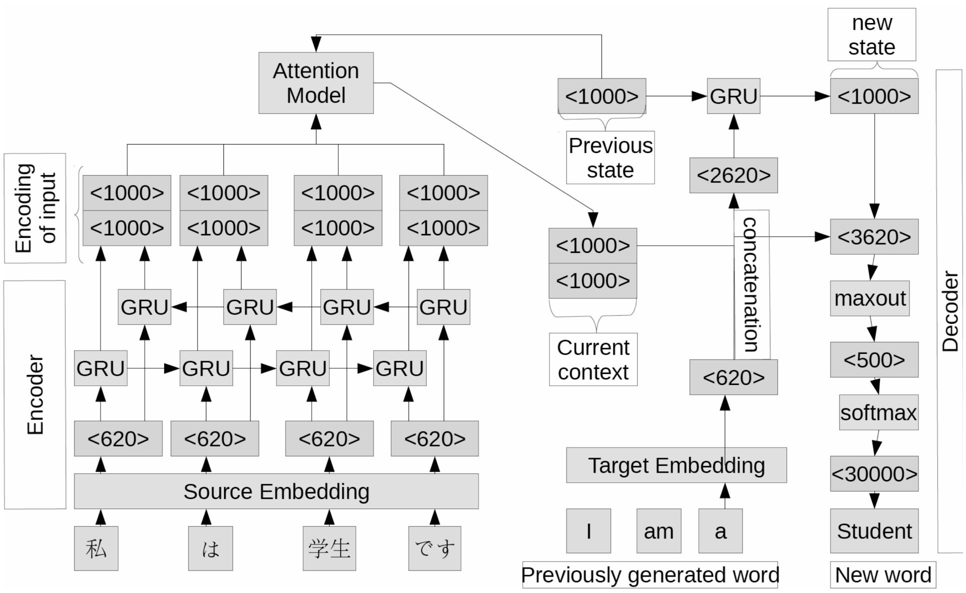

Figure 2: The architecture of the NMT system with attention, as described in (Bahdanau et al., 2015). The notation “<1000>” means a vector of size 1000. The vector sizes shown here are the ones suggested in the original paper.

NMT in detail (Section 4); After that, we introduce domain adaptation for NMT in real word scenarios, which is crucial for the practical use of MT (Section 5); Finally, we give our opinions of future research directions in this field (Section 6) and conclude this paper (Section 7).

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 (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015). The encoder-decoder model with attention (Bahdanau et al., 2015) is the most commonly used NMT architecture. This model is also known as RNNsearch. Figure 2 describes the RNNsearch model (Bahdanau et al., 2015), which 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:

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

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 C consisting of a set of parallel sentence pairs $(\mathbf{x}, \mathbf{y})$. The training object is to minimize the cross-entropy loss L w.r.t θ:

$$L_\theta = \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{C}} -\log p(\mathbf{y}|\mathbf{x};\theta). \tag{2}$$

The model 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, gated recurrent unit (GRU) in this case, 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 all past words that have been seen. There are two GRU layers which encode forward and backward information. Each word x_i is represented by concatenating the forward hidden state $\overrightarrow{h_i}$ and the backward one $\overleftarrow{h_i}$ as $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$. In this way, the source sentence $\mathbf{x} = \{x_1, ..., x_n\}$ can be represented as $\mathbf{h} = \{h_1, ..., h_n\}$. 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, a GRU 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 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 GRU 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 s_j for time j of the decoder is computed by:

$$s_j = f(s_{j-1}, y_{j-1}, c_j), \tag{3}$$

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

$$c_j = \sum_{j=1}^{n} a_{ji} h_i, \quad a_{ji} = \frac{\exp(e_{ji})}{\sum_{k=1}^{m} \exp(e_{ki})}, \quad e_{ji} = a(s_{j-1}, h_i), \tag{4}$$

where a 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 maxout layer which is a feedforward layer with max pooling. The maxout layer takes the recurrent hidden state generated by the decoder GRU, the previous word and the context vector to compute a final representation, which is fed to a simple softmax layer:

$$P(y_j | y_{<j}, \mathbf{x}) = \mathrm{softmax}(\mathrm{maxout}(s_j, y_{j-1}, c_j)). \tag{5}$$

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, and attention model. This is the main bottleneck of NMT for low resource domains and languages.

3 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:

3.1 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 (Moore and Lewis, 2010; Axelrod et al., 2011; Duh et al., 2013), as well as joint models (Hoang and Sima'an, 2014; Durrani et al., 2015), and more recently convolutional neural network (CNN) models (Chen et al., 2016) 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 (Utiyama and Isahara, 2003), self-enhancing (Lambert et al., 2011) or parallel word embeddings (Marie and Fujita, 2017). Besides sentence generation, there are also studies that generate monolingual n-grams (Wang et al., 2014) and parallel phrase pairs (Chu, 2015; Wang et al., 2016).

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 (Wang et al., 2017a).

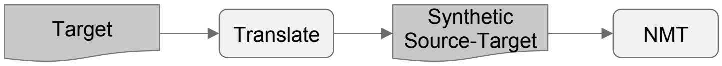

Figure 3: Synthetic data generation for NMT (Sennrich et al., 2016b).

3.2 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 (Foster and Kuhn, 2007; Bisazza et al., 2011; Niehues and Waibel, 2012; Sennrich et al., 2013; Durrani et al., 2015; Imamura and Sumita, 2016).

ii) Instance level interpolation. Instance weighting has been applied to several natural language processing (NLP) domain adaptation tasks (Jiang and Zhai, 2007), especially SMT (Matsoukas et al., 2009; Foster et al., 2010; Shah et al., 2012; Mansour and Ney, 2012; Zhou et al., 2015). They firstly score each instance/domain by using rules or statistical methods as a weight, and then train SMT models by giving each instance/domain the weight. An alternative way is to weight the corpora by data re-sampling (Shah et al., 2010; Rousseau et al., 2011).

For NMT, several methods have been proposed to interpolate model/data like SMT does. For model-level interpolation, the most related NMT technique is model ensemble (Jean et al., 2015). For instance-level interpolation, the most related method is to assign a weight in NMT objective function (Chen et al., 2017a; Wang et al., 2017b). However, the model structures of SMT and NMT are quite different. SMT is a combination of several independent models; in comparison, NMT is an integral 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

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. (2015) train an RNNLM on monolingual data, and fuse the RNNLM and NMT models. Currey et al. (2017) copy the target monolingual data to the source side and use the copied data for training NMT. Domhan and Hieber (2017) propose using target monolingual data for the decoder with LM and NMT multitask learning. Zhang and Zong (2016b) use source side monolingual data to strengthen the NMT encoder via multitask learning for predicting both translation and reordered source sentences. Cheng et al. (2016) 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 (Sennrich et al., 2016b). Figure 3 shows the flowchart of this method. It has also been shown that synthetic data generation is very effective for domain adaptation using either the target side monolingual data (Sennrich et al., 2016c), the source side monolingual data (Zhang and Zong, 2016b), or both (Park et al., 2017).

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. (2017) is originally motivated by Sennrich et al. (2016a), which uses tags to control the politeness of NMT. The overview of this method is shown in

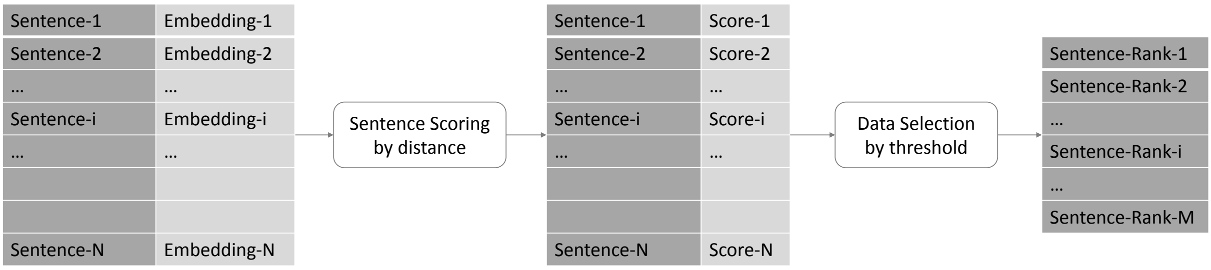

Figure 4: Data selection for NMT (Wang et al., 2017a).

the dotted section in Figure 6. 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.

Sajjad et al. (2017) 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. (2017) 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.

Data Selection As mentioned in the SMT section (Section 3.1), the data selection methods in SMT can improve NMT performance modestly, because their criteria of data selection are not very related to NMT (Wang et al., 2017a). To address this problem, Wang et al. (2017a) 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 (Figure 4). Van der Wees et al. (2017) 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.

Although all the data centric methods for NMT are complementary to each other in principle, there are no studies that try to combine these methods, which is considered to be one future direction.

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. There is only one work concerning instance weighting in NMT (Wang et al., 2017b). They set a 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 (Axelrod et al., 2011) (Figure 5). Instead of instance weighting, Chen et al. (2017a) 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. Recently, Wang et al. (2018) proposed a joint framework of sentence selection and weighting for NMT.

Fine Tuning *Fine tuning* is the conventional way for domain adaptation (Luong and Manning, 2015; Sennrich et al., 2016b; Servan et al., 2016; Freitag and Al-Onaizan, 2016). In this method, an NMT

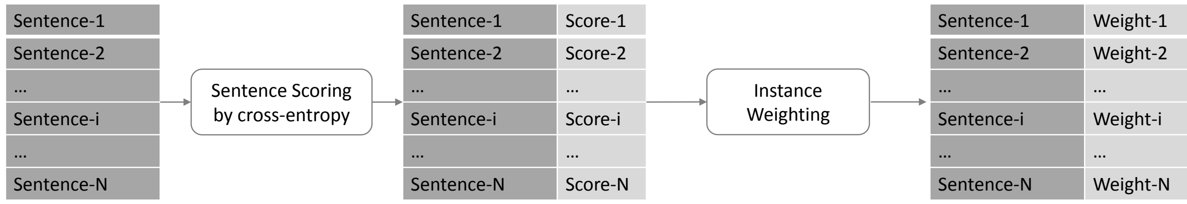

Figure 5: Instance weighting for NMT (Wang et al., 2017b).

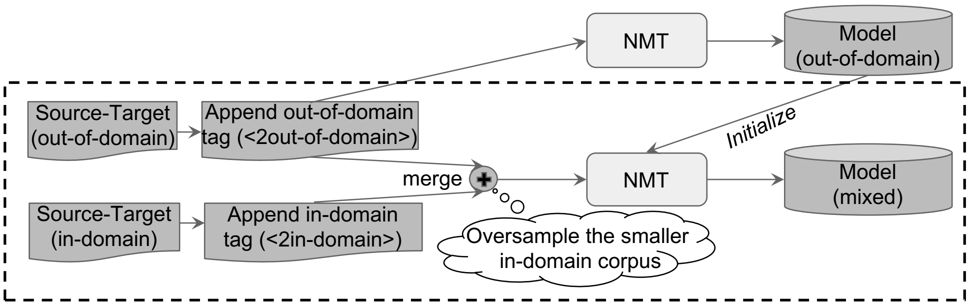

Figure 6: Mixed fine tuning with domain tags for domain adaptation (Chu et al., 2017). The section in the dotted rectangle denotes the *multi-domain* method .

system on a resource rich out-of-domain corpus is trained until convergence, and then its parameters are fine tuned on a resource poor in-domain corpus. Conventionally, fine tuning is applied on in-domain parallel corpora. Varga et al. (2017) apply it on parallel sentences extracted from comparable corpora. Comparable corpora have been widely used for SMT by extracting parallel data from them (Chu, 2015). To prevent degradation of out-of-domain translation after fine tuning on in-domain data, Dakwale and Monz (2017) propose an extension of fine tuning that keeps the distribution of the out-of-domain model based on knowledge distillation (Hinton et al., 2015).

Mixed Fine Tuning This method is a combination of the *multi-domain* and *fine tuning* methods (Figure 6). 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.

Mixed fine tuning addresses the overfitting problem of fine tuning due to the small size of the in-domain data. It is easier to train a good model with out-of-domain data, compared to training a multi-domain model. Once we obtained good model parameters, we can use these parameters for fine tuning on the mixed domain data to obtain better performance for the in-domain model. In addition, mixed fine tuning is faster than multi-domain because training an out-of-domain model convergences faster than training a multi-domain model, which also convergences very fast in fine tuning on the mixed domain data. Chu et al. (2017) show that mixed fine tuning works better than both *multi-domain* and *fine tuning*. In addition, mixed fine tuning has the similar effect as the ensembling method in Dakw and Monz (2017), which does not decrease the out-of-domain translation performance.

Regularization Barone et al. (2017) also realize the overfitting problem during fine tuning. Their strategy to address this problem is to explore regularization techniques such as dropout and L2-regularization. In addition, they also propose *tuneout* that is a variant of dropout for regularization. We think that mixed fine tuning and regularization techniques are complementary to each other.

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 (Gülçehre et

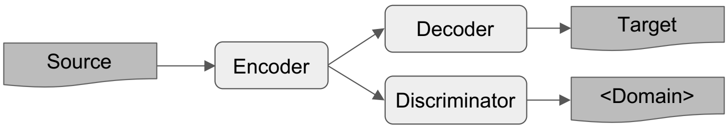

Figure 7: Domain discriminator (Britz et al., 2017).

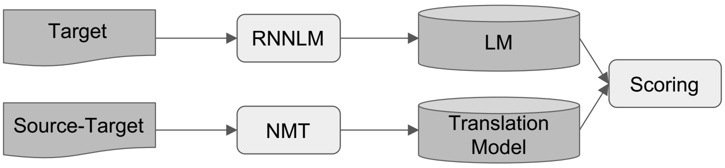

Figure 8: LM shallow fusion (Gülçehre et al., 2015).

al., 2015). 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).

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. Domhan and Hieber (2017) propose a method similar to the deep fusion method (Gülçehre et al., 2015). However, unlike training the RNNLM and NMT model separately (Gülçehre et al., 2015), Domhan and Hieber (2017) train RNNLM and NMT models jointly.

Domain Discriminator To leverage the diversity of information in multi-domain corpora, Britz et al. (2017) propose a discriminative method. In their discriminative method, they add a feed-forward network (FFNN) as a discriminator on top of the encoder that uses the attention to predict the domain of the source sentence. The discriminator is optimized jointly with the NMT network. Figure 7 shows an overview of this method.

Domain Control Besides using domain tokens to control the domains, Kobus et al. (2016) 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.

4.2.3 Decoding Centric

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 (Gülçehre et al., 2015). In the shallow fusion (Gülçehre et al., 2015), the next word hypotheses generated by an NMT model is rescored by the weighted sum of the NMT and RNNLM probabilities (Figure 8).

Ensembling Freitag and Al-Onaizan (2016) propose to ensemble the out-of-domain domain and the fine tuned in-domain models. Their motivation is exactly the same as the work of Dakwale and Monz (2017), which is preventing degradation of out-of-domain translation after fine tuning on in-domain data.

Neural Lattice Search Khayrallah et al. (2017) propose a stack-based decoding algorithm over word lattices, while the lattices are generated by SMT (Dyer et al., 2008). In their domain adaptation experiments, they show that stack-based decoding is better than conventional decoding.

5 Domain Adaptation in Real-World Scenarios

A domain adaptation method should be adopted according to the 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

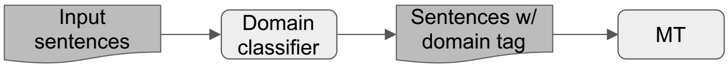

Figure 9: Domain adaptation in an input domain unknown scenario.

of different methods possible. Chu et al. (2018) conduct a study that applys mixed fine tuning (Chu et al., 2017) on synthetic parallel data (Sennrich et al., 2016b), 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.

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 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 (Huck et al., 2015). Xu et al. (2007) 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. (2015) extend the work of Xu et al. (2007) on the sentence level. They use LMs and maximum entropy classifiers to predict the target domain. Banerjee et al. (2010) 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. (2012) rely on averaged perceptron classifiers with various phrase-based features.

For NMT, Kobus et al. (2016) 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. (2016) 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. Farajian et al. (2017) follow the strategy of Li et al. (2016), 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. Figure 9 shows an overview of domain adaptation for MT in the input domain unknown scenario.

6 Future Directions

6.1 Domain Adaptation for State-of-the-art NMT Architectures

Since the success of RNN based NMT (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015), other architectures of NMT have been developed. One representative architecture is CNN based NMT (Gehring et al., 2017). Compared to RNN based models, CNN based models can be computed fully parallel during training and are much easier to optimize. Another representative architecture is the *Transformer*, which is based on attention only (Vaswani et al., 2017). It has been shown that CNN based NMT and the *Transformer* significantly outperform the state-of-the-art RNN based NMT model of Wu et al. (2016) in both the translation quality and speed perspectives. However, currently, most of the domain adaptation studies for NMT are based on the RNN based model (Bahdanau et al., 2015). The research of domain adaptation techniques for these latest state-of-the-art NMT models is obviously an important future direction.

6.2 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 (Arthur et al., 2016; Zhang and Zong, 2016a). Arcan and Buitelaar (2017) use a domain specific dictionary for terminology translation, but they simply apply the unknown word replacement method proposed by Luong et al. (2015), which suffers from noisy attention.

6.3 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 (Johnson et al., 2016). This approach is known as cross-lingual transfer learning, which transfers NMT model parameters among multiple languages. 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 (Zoph et al., 2016). There are studies where either multilingual (Firat et al., 2016; Johnson et al., 2017) or multi-domain models (Sajjad et al., 2017) are trained, but none that attempt to package multiple language pairs and multiple domains into a single translation system. 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. Therefore, we think that multilingual and multi-domain adaptation for NMT can be another future direction. Chu and Dabre (2018) conduct a preliminary study for this topic.

6.4 Adversarial Domain Adaptation and Domain Generation

Generative adversarial networks are a class of artificial intelligence algorithms used in unsupervised machine learning, which are introduced by (Goodfellow et al., 2014). Adversarial methods have become popular in domain adaptation (Ganin et al., 2016), which minimize an approximate domain discrepancy distance through an adversarial objective with respect to a domain discriminator (Tzeng et al., 2017). They have been applied to domain adaptation tasks in computer vision and machine learning (Tzeng et al., 2017; Motiian et al., 2017; Volpi et al., 2017; Zhao et al., 2017; Pei et al., 2018). Recently, some of the adversarial methods began to be introduced into some NLP tasks (Liu et al., 2017; Chen et al., 2017b) and NMT (Britz et al., 2017).

Most of the existing methods focus on adapting from a general domain into a specific domain. In the real scenario, training data and test data have different distributions and the target domains are sometimes unseen. Irvine et al. (2013) analyze the translation errors in such scenarios. Domain generalization aims to apply knowledge gained from labeled source domains to unseen target domains (Li et al., 2018). It provides a way to match the distribution of training data and test data in real-world MT, which may be a future trend of domain adaptation for NMT.

7 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 the first comprehensive review of the techniques mainly being developed in the last two 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 NMT to the techniques in general NLP, computer vision and machine learning is our future work. We hope that this survey paper could significantly promote the research in domain adaptation for NMT.

Acknowledgement

This work was supported by Grant-in-Aid for Research Activity Start-up #17H06822, JSPS. We are very appreciated 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

Mihael Arcan and Paul Buitelaar. 2017. Translating domain-specific expressions in knowledge bases with neural machine translation. *CoRR*, abs/1709.02184.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas, November. Association for Computational Linguistics.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, U.K.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, USA, May. International Conference on Learning Representations.

Pratyush Banerjee, Jinhua Du, Baoli Li, Sudip Naskar, Andy Way, and Josef Genabith. 2010. Combining multi-domain statistical machine translation models using automatic classifiers. In *The Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado.

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *IWSLT*, pages 136–143. ISCA.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark, September. Association for Computational Linguistics.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126, Copenhagen, Denmark, September. Association for Computational Linguistics.

M Cettolo, J Niehues, S Stüker, L Bentivogli, R Cattoni, and M Federico. 2015. The iwslt 2015 evaluation campaign. In *Proceedings of the Twelfth International Workshop on Spoken Language Translation (IWSLT)*.

Boxing Chen, Roland Kuhn, George Foster, Colin Cherry, and Fei Huang. 2016. Bilingual methods for adaptive training data selection for machine translation. In *The Twelfth Conference of The Association for Machine Translation in the Americas*, pages 93–106, Austin, Texas.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017a. Cost weighting for neural machine translation domain adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 40–46, Vancouver.

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. 2017b. Adversarial multi-criteria learning for chinese word segmentation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1193–1203, Vancouver, Canada. Association for Computational Linguistics.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany, August. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Chenhui Chu and Raj Dabre. 2018. Multilingual and multi-domain adaptation for neural machine translation. In *Proceedings of the 24st Annual Meeting of the Association for Natural Language Processing (NLP 2018)*, pages 909–912, Okayama, Japan, Match.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, July. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2018. A comprehensive empirical comparison of domain adaptation methods for neural machine translation. *Journal of Information Processing (JIP)*, 26(1):1–10.

Chenhui Chu. 2015. Integrated parallel data extraction from comparable corpora for statistical machine translation. *Doctoral Thesis, Kyoto University*.

Gabriela Csurka. 2017. Domain adaptation for visual applications: A comprehensive survey. *CoRR*, abs/1702.05374.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark, September. Association for Computational Linguistics.

Praveen Dakwale and Christof Monz. 2017. Fine-tuning for neural machine translation with limited degradation across in- and out-of-domain data. In *Proceedings of the 16th Machine Translation Summit (MT-Summit 2017)*, pages 156–169.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark, September. Association for Computational Linguistics.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683, Sofia, Bulgaria, August.

Nadir Durrani, Hassan Sajjad, Shafiq Joty, Ahmed Abdelali, and Stephan Vogel. 2015. Using joint models for domain adaptation in statistical machine translation. In *Proceedings of MT Summit XV*, pages 117–130, Miami, FL, USA.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*, pages 1012–1020, Columbus, Ohio, June. Association for Computational Linguistics.

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137, Copenhagen, Denmark, September. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 866–875.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for smt. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 128–135, Stroudsburg, PA, USA. Association for Computational Linguistics.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Cambridge, MA.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. *CoRR*, abs/1705.03122.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Isao Goto, Ka-Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. 2013. Overview of the patent machine translation task at the ntcir-10 workshop. In *Proceedings of the 10th NTCIR Conference*, pages 260–286, Tokyo, Japan, June. National Institute of Informatics (NII).

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Cuong Hoang and Khalil Sima'an. 2014. Latent domain translation models in mix-of-domains haystack. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1928–1939, Dublin, Ireland.

Matthias Huck, Alexandra Birch, and Barry Haddow. 2015. Mixed-domain vs. multi-domain statistical machine translation. *Proceedings of MT Summit XV*, 1:240–255.

Kenji Imamura and Eiichiro Sumita. 2016. Multi-domain adaptation for statistical machine translation based on feature augmentation. In *Proceedings of the 12th Conference of the Association for Machine Translation in the Americas*, Austin, Texas, USA.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daume III, and Dragos Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China, July. Association for Computational Linguistics.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a Vigas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Huda Khayrallah, Gaurav Kumar, Kevin Duh, Matt Post, and Philipp Koehn. 2017. Neural lattice search for domain adaptation in machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 20–25, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Catherine Kobus, Josep Crego, and Jean Senellart. 2016. Domain control for neural machine translation. *arXiv preprint arXiv:1612.06140*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2017. Neural machine translation. *CoRR*, abs/1709.07809.

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 284–293, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. One sentence one model for neural machine translation. *CoRR*, abs/1609.06490.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. 2018. Domain generalization via conditional invariant representations. In *The Thirty-Second AAAI Conference on Artificial Intelligence*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada. Association for Computational Linguistics.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, pages 76–79, Da Nang, Vietnam, December.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China, July. Association for Computational Linguistics.

Saab Mansour and Hermann Ney. 2012. A simple and effective weighted phrase extraction for machine translation adaptation. In *The 9th International Workshop on Spoken Language Translation*, Hong Kong.

Benjamin Marie and Atsushi Fujita. 2017. Efficient extraction of pseudo-parallel sentences from raw monolingual data using word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 392–398, Vancouver, Canada, July. Association for Computational Linguistics.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 708–717, Singapore.

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization techniques for fine-tuning in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1489–1494, Copenhagen, Denmark, September. Association for Computational Linguistics.

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 220–224, Uppsala, Sweden.

Saeid Motiian, Quinn Jones, Seyed Mehdi Iranmanesh, and Gianfranco Doretto. 2017. Few-shot adversarial domain adaptation. *CoRR*, abs/1711.02536.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Hideto Kazawa, Yusuke Oda, Graham Neubig, and Sadao Kurohashi. 2017. Overview of the 4th workshop on asian translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 1–54, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Graham Neubig. 2017. Neural machine translation and sequence-to-sequence models: A tutorial. *CoRR*, abs/1703.01619.

Jan Niehues and Alex H. Waibel. 2012. Detailed analysis of different strategies for phrase table adaptation in smt. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, US-CA.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October.

Jaehong Park, Jongyoon Song, and Sungroh Yoon. 2017. Building a neural machine translation system using only synthetic parallel data. *CoRR*, abs/1704.00253.

Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-adversarial domain adaptation.

Anthony Rousseau, Fethi Bougares, Paul Deléglise, Holger Schwenk, and Yannick Estève. 2011. Liums systems for the iwslt 2011 speech translation tasks. In *International Workshop on Spoken Language Translation*, San Francisco, USA.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Yonatan Belinkov, and Stephan Vogel. 2017. Neural machine translation training in a multi-domain scenario. In *Proceedings of the Twelfth International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan.

Rico Sennrich, Holger Schwenk, and Walid Aransa. 2013. A multi-domain translation model framework for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 832–840, Sofia, Bulgaria.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California, June. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Christophe Servan, Josep Crego, and Jean Senellart. 2016. Domain specialization: a post-training domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06141*.

Kashif Shah, Loïc Barrault, and Holger Schwenk. 2010. Translation model adaptation by resampling. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 392–399.

Kashif Shah, Loïc Barrault, and Holger Schwenk. 2012. A general framework to weight heterogeneous parallel data for model adaptation in statistical machine translation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, US-CA.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. *CoRR*, abs/1702.05464.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79, Sapporo, Japan, July. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark, September. Association for Computational Linguistics.

Adam Csaba Varga. 2017. Domain adaptation for multilingual neural machine translation. *Master Thesis, Saarlandes University*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Riccardo Volpi, Pietro Morerio, Silvio Savarese, and Vittorio Murino. 2017. Adversarial feature augmentation for unsupervised domain adaptation. *CoRR*, abs/1711.08561.

Wei Wang, Klaus Macherey, Wolfgang Macherey, Franz Och, and Peng Xu. 2012. Improved domain adaptation for statistical machine translation. In *Proceedings of AMTA*, San Diego, California, USA.

Rui Wang, Hai Zhao, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita. 2014. Neural network based bilingual language model growing for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 189–195, Doha, Qatar, October. Association for Computational Linguistics.

Rui Wang, Hai Zhao, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita. 2016. Connecting phrase based statistical machine translation adaptation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3135–3145, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017a. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada, July. Association for Computational Linguistics.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017b. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark.

Rui Wang, Masao Utiyama, Andrew Finch, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2018. Sentence selection and weighting for neural machine translation domain adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big Data*, 3(1):9, May.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Jia Xu, Yonggang Deng, Yuqing Gao, and Hermann Ney. 2007. Domain dependent statistical machine translation. In *MT Summit*, Copenhagen, Denmark.

Jiajun Zhang and Chengqing Zong. 2016a. Bridging neural machine translation and bilingual dictionaries. *CoRR*, abs/1610.07272.

Jiajun Zhang and Chengqing Zong. 2016b. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas, November. Association for Computational Linguistics.

Han Zhao, Shanghang Zhang, Guanhang Wu, João P. Costeira, José M. F. Moura, and Geoffrey J. Gordon. 2017. Multiple source domain adaptation with adversarial training of neural networks. *CoRR*, abs/1705.09684.

Xinpeng Zhou, Hailong Cao, and Tiejun Zhao. 2015. Domain adaptation for SMT using sentence weight. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 153–163, Guangzhou, China.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1568–1575.