Effective Domain Mixing for Neural Machine Translation

Reid Pryzant*
Stanford University
rpryzant@stanford.edu

Denny Britz*
Google Brain
dennybritz@google.com

Quoc V. Le
Google Brain
qvl@google.com

Abstract

Neural Machine Translation (NMT) models are often trained on heterogeneous mixtures of domains, from news to parliamentary proceedings, each with unique distributions and language. In this work we show that training NMT systems on naively mixed data can degrade performance versus models fit to each constituent domain. We demonstrate that this problem *can* be circumvented, and propose three models that do so by jointly learning domain discrimination and translation. We demonstrate the efficacy of these techniques by merging pairs of domains in three languages: Chinese, French, and Japanese. After training on composite data, each approach outperforms its domain-specific counterparts, with a model based on a discriminator network doing so most reliably. We obtain consistent performance improvements and an average increase of 1.1 BLEU.

1 Introduction

Neural Machine Translation (NMT) (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014) is an end-to-end approach for automated translation. NMT has shown impressive results (Bahdanau et al., 2015; Luong et al., 2015a; Wu et al., 2016) often surpassing those of phrase-based systems while addressing shortcomings such as the need for hand-engineered features.

In many translation settings (e.g. web translation, assistant translators), input may come from more than one domain. Each domain has unique properties that could confound models not explicitly fitted to it. Thus, an important problem is to effectively mix a diversity of training data in a multi-domain setting.

Our problem space is as follows: how can we train a translation model on multi-domain data to improve test-time performance in each *constituent* domain? This setting differs from the majority of work in domain adaptation, which explores how models trained on some source domain can be effectively applied to *outside* target domains. This setting is important, because previous research has shown that both standard NMT and adaptation methods degrade performance on the original source domain(s) (Farajian et al., 2017; Haddow and Koehn, 2012). We seek to prove that this problem can be overcome, and hypothesize that leveraging the heterogeneity of composite data rather than dampening it will allow us to do so.

To this extent, we propose three new models for multi-domain machine translation. These models are based on discriminator networks, adversarial learning, and target-side domain tokens. We evaluate on pairs of linguistically disparate corpora in three translation tasks (EN-JA, EN-ZH, EN-FR), and observe that unlike naively training on mixed data (as per current best practices), the proposed techniques consistently improve translation quality in each individual setting. The most significant of these tasks is EN-JA, where we obtain state-of-the-art performance in the process of examining the ASPEC corpus (Nakazawa et al., 2016) of scientific papers and SubCrawl, a new corpus based on an anonymous manuscript (Anonymous, 2017). In summary,

*Equal Contribution.

our contributions are as follows:

- We show that mixing data from heterogenous domains leads to suboptimal results compared to the single-domain setting, and that the more distant these domains are, the more their merger degrades downstream translation quality.

- We demonstrate that this problem can be circumvented and propose novel, general-purpose techniques that do so.

2 Neural Machine Translation

Neural machine translation (Sutskever et al., 2014) directly models the conditional log probability $\log p(\boldsymbol{y}|\boldsymbol{x})$ of producing some translation $\boldsymbol{y} = y_1, ..., y_m$ of a source sentence $\boldsymbol{x} = x_1, ..., x_n$. It models this probability through the *encoder-decoder* framework. In this approach, an *encoder* network encodes the source into a series of vector representations $\boldsymbol{H} = \boldsymbol{h}_1, ..., \boldsymbol{h}_n$. The *decoder* network uses this encoding to generate a translation one target token at a time. At each step, the decoder casts an attentional distribution over source encodings (Luong et al., 2015b; Bahdanau et al., 2014). This allows the model to focus on parts of the input before producing each translated token. In this way the decoder is decomposing the conditional log probability into

$$\log p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{t=1}^{m} \log p(y_t|y_{<t}, \boldsymbol{H}) \qquad (1)$$

In practice, stacked networks with recurrent Long Short-Term Memory (LSTM) units are used for both the encoder and decoder. Such units can effectively distill structure from sequential data (Elman, 1990).

The cross-entropy training objective in NMT is formulated as,

$$\mathcal{L}_{gen} = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} -\log p(\mathbf{y}|\mathbf{x}) \qquad (2)$$

Where $\mathcal{D}$ is a set of (source, target) sequence pairs $(\mathbf{x}, \mathbf{y})$.

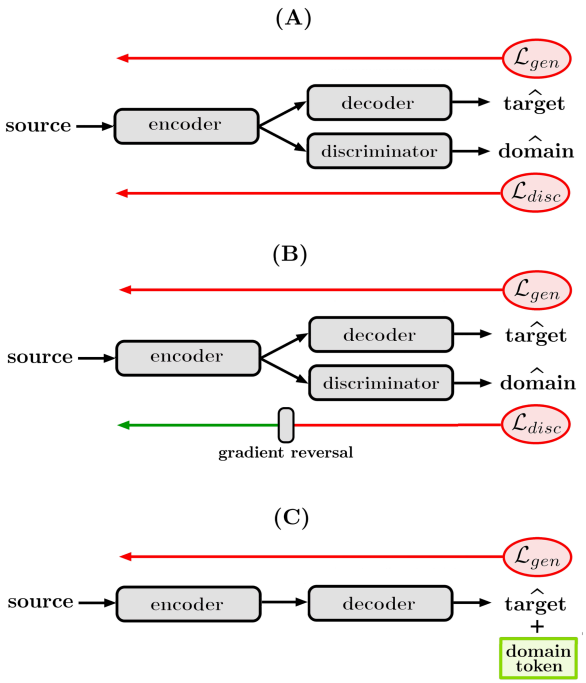


Figure 1: The novel mixing paradigms under consideration. Discriminative mixing (A), adversarial discriminative mixing (B), and target-side token mixing (C) are depicted.

3 Models

We now describe three models we are proposing that leverage the diversity of information in heterogeneous corpora. They are summarized in Figure 1. We assume dataset $\mathcal{D}$ consists of source sequences $\boldsymbol{X}$, target sequences $\boldsymbol{Y}$ and domain class labels $\boldsymbol{D}$ that are only known at training time.

3.1 Discriminative Mixing

In the Discriminative Mixing approach, we add a discriminator network on top of the source encoder that takes a single vector encoding of the source $\boldsymbol{c}$ as input. This network maximizes $P(d|\boldsymbol{H})$, the predicted probability of the correct domain class label d conditioned on the hidden states of the encoder $\boldsymbol{H}$. It does so by minimizing the negative cross-entropy loss $\mathcal{L}_{disc} = -\log p(d|\boldsymbol{H})$. In other words, the discriminator uses the encoded representation of the source sequence to predict the correct domain. Intuitively, this forces the encoder to encode domain-related information into the features it generates. We hypothesize that this information will be useful during the decoding process.

The encoder can employ an arbitrary mechanism to distill the source into a single-vector representation $\boldsymbol{c}$. In this work, we use an attention mechanism over the encoder states $\boldsymbol{H}$, followed by a fully connected layer. We set $\boldsymbol{c}$ to be the attention context, and calculate it according to Bahdanau et al. (2015):

$$\boldsymbol{c} = \sum_j a_j \boldsymbol{h}_j$$
$$\boldsymbol{a} = \text{softmax}(\hat{\boldsymbol{a}})$$
$$\hat{a}_i = \boldsymbol{v}_a^T \tanh(\boldsymbol{W}_a \boldsymbol{h}_i)$$

The discriminator can be an arbitrary neural network. For this work, we fed $\mathbf{c}$ into a fully connected layer with a tanh nonlinearity, then passed the result through a softmax to obtain probabilities for each domain class label.

The discriminator is optimized jointly with the rest of the Sequence-to-Sequence network. If $\mathcal{L}_{gen}$ is the standard sequence generator loss described in Section 2, then the final loss we are optimizing is the sum of the generator and discriminator loss $\mathcal{L} = \mathcal{L}_{gen} + \mathcal{L}_{disc}$.

3.2 Adversarial Discriminative Mixing

We also experiment with an adversarial approach to domain mixing. This approach is similar to that of 3.1, except that when backpropagating from the discriminator network to the encoder, we reverse the gradients by multiplying them by -1. Though the discriminator is still using $\nabla\mathcal{L}_{disc}$ to update its parameters, with the inclusion of the reversal layer, we are implicitly directing the encoder to optimize with $-\nabla\mathcal{L}_{disc}$. This has the opposite effect of what we described above. The discriminator still learns to distinguish between domains, but the encoder is forced to compute domain-invariant representations that are not useful to the discriminator. We hope that such representations lead to better generalization across domains.

Note the connections between this technique and that of the Generative Adversarial Network (GAN) paradigm (Goodfellow et al., 2014). GANs optimize two networks with two objective functions (one being the negation of the other) and periodically freeze the parameters of each network during training. We are training a single network without freezing any of its components. Furthermore, we reverse gradients in lieu of explicitly defining a second, negated loss function. Last, the adversarial parts of this model are trained jointly with translation in a multitask setting.

Note also that the representations computed by this model are likely to be applicable to unseen, outside domains. However, this setting is outside the scope of this paper and we leave its exploration to future work. For our setting, we hypothesize that the domain-agnostic encodings encouraged by the discriminator may yield improvements in mixed-domain settings as well.

3.3 Target Token Mixing

A simpler alternative to adding a discriminator network is to prepend a domain token to the target sequence. Such a technique can be readily incorporated into any existing NMT pipeline and does not require changes to the model. In particular, we add a single special vocabulary word such as "domain=subtitles", per domain and prepend this token to each target sequence therein.

The decoder must learn, similar to the more complex discriminator above, to predict the correct domain token based on the source representation at the first step of decoding. We hypothesize that this technique has a similar regularizing effect as adding a discriminator network. During inference, we remove the first predicted token corresponding to the domain.

The advantage of this approach verses the similar techniques discussed in related work (Section 5) is that in our proposed method, the model must learn to predict the domain based on the source sequence alone. It does not need to know the domain a-priori.

4 Experiments

4.1 Datasets

For the Japanese translation task we evaluate our domain mixing techniques on the standard ASPEC corpus (Nakazawa et al., 2016) consisting of 3M scientific document sentence pairs, and the SubCrawl corpus, consisting of 3.2M colloquial sentence pairs harvested from freely available subtitle repositories on the World Wide Web. We use standard train/dev/test splits (3M, 1.8k, and 1.8k examples, respectively) and preprocess the data

using subword units[1] (Sennrich et al., 2015) to learn a shared English-Japanese vocabulary of size 32,000. To allow for fair comparisons, we use the same vocabulary and sentence segmentation for all experiments, including single-domain models.

To prove its generality, we also evaluate our techniques on a small set of about 200k/1k/1k training/dev/test examples of English Chinese (EN-ZH) and English-French (EN-FR) language pairs. For EN-ZH, we use a news commentary corpus from WMT'17[2] and a 2012 database dump of TED talk subtitles (Tiedemann, 2012). For EN-FR, we use professional translations of European Parliament Proceedings (Koehn, 2005) and a 2016 dump of the OpenSubtitles database (Lison and Tiedemann, 2016).

The premise of evaluating on mixed-domain data is that the domains undergoing mixing are in fact disparate. We need to quantifiably measure the disparity therein to obtain fair, valid, and explainable results. Thus, we measured the distances between the domains of each language pair with A-distance, an important part of the upper generalization bounds for domain adaptation (Ben-David et al., 2007). Due to the intractability of computing A-distances, we instead compute a proxy for A-distance, $\hat{d}_A$, which is given theoretical justification in Ben-David et al. (2007) and used to measure domain distance in Gani et al. (2015); Glorot et al. (2011). The proxy A-distance is obtained by measuring the generalization error ϵ of a linear bag-of-words SVM classifier trained to discriminate between the two domains, and setting $\hat{d}_A = 2(1-2\epsilon)$. Note that by nature of its formulation, $\hat{d}_A$ is only useful in comparative settings, and means little in isolation (Ben-David et al., 2007). However, it has a minimum value of 1, implying exact domain match, and a maximum of 2, implying that domains are polar opposites.

4.2 Experimental Protocol

All models are implemented using the Tensorflow framework and based on the Sequence-to-Sequence implementation of Britz et al. (2017)[3]. We use a 4-layer bidirectional LSTM encoder with 512 units, and a 4-layer LSTM decoder. Recall from Section 3 that we use Bahdanau-style attention Bahdanau et al. (2015). Dropout of 0.2 (0.8 keep probability) is applied to the input of each cell. We optimize using Adam and a learning rate of 0.0001 (Kingma and Ba, 2014; Abadi et al., 2016). Each model is trained on 8 Nvidia K40m GPUs with a batch size of 128. The combined Japanese dataset took approximately a week to reach convergence.

During training, we save model checkpoints every hour and choose the best one using the BLEU score on the validation set. To calculate BLEU scores for the EN-JA task, we follow the instruction from WAT [4] and use the KyTea tokenizer (Neubig et al., 2011). For the EN-FR and EN-ZH tasks, we follow the WMT '16 guidlines and tokenize with the Moses `tokenizer.perl` script (Koehn et al., 2007).

4.3 Results

The results of our proxy-A distance experiment are given in Table 1. $\hat{d}_A$ is a purely comparative metric that has little meaning in isolation (Ben-David et al., 2007), so it is evident that the EN-JA and EN-ZH domains are more disparate, while the EN-FR domains are more similar.

Lanuage	Domain 1	Domain 2	$\hat{d}_A$
Japanese	ASPEC	SubCrawl	1.89
Chinese	News	TED	1.73
French	Europarl	OpenSubs	1.23

Table 1: Proxy A-distances ($\hat{d}_A$) for each domain pair.

To understand the interactions between these models and mixed-domain data, we train and test on ASPEC, SubCrawl, and their concatenation. We do the same for the French and Chinese baselines.

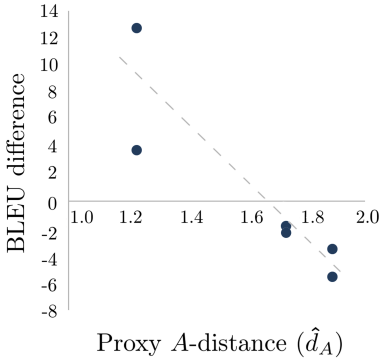
In general, our results support the hypothesis that the naive concatenation of data from disparate domains can degrade in-domain translation quality (Table 2). In both the EN-JA and EN-FR settings, the domains undergoing mixing are disparate enough to *degrade*

[1] Using https://github.com/google/sentencepiece

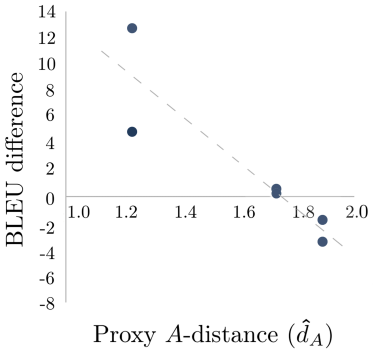
[2] http://www.statmt.org/wmt17/translation-task.html

[3] https://github.com/google/seq2seq

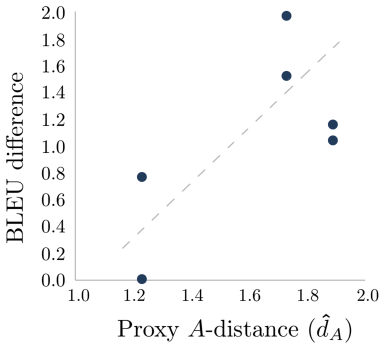
[4] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/



(a) Comparing the mixed-domain and individual-domain baselines ($\text{BLEU}_{mixed} - \text{BLEU}_{individual}$) while varying domain distance. The more different two domains are, the more their mixture degrades performance.



(b) Comparing the proposed discriminator and individual-domain baseline ($\text{BLEU}_{discriminator} - \text{BLEU}_{individual}$) while varying domain distance. Compared to Figure 2a, performance is less degraded when using the discriminator.



(c) Comparing the proposed discriminator approach and mixed-domain baseline ($\text{BLEU}_{discriminator} - \text{BLEU}_{mixed}$) while varying domain distance. The discriminator always improves over the baseline, and this is accentuated when the merged domains are more distant.

Figure 2: Comparative performance and domain distance. Trends corresponding to a least-squares fit are indicated with dashed lines.

performance when mixed, and the proposed techniques recover some of this performance drop. In the EN-ZH setting, we observe that even when similar domains are mixed performance can drop. Notably, in this setting, the proposed techniques successfully *improve* performance over single-domain training.

For a more detailed perspective on this result, Figure 2a depicts the mixed-domain/individual-domain performance differential as a function of domain distance. The two share a negative association, suggesting that the most distant two domains are, the more their merger degrades performance. This degradation is particularly strong in Japanese due the vast structural differences between formal and casual language. The vocabularies, conjugational patterns, and word attachments all follow different rules in this case (Hori, 1986).

We then trained and tested our proposed methods on the same mixed data (Table 2). Our results generally agree with the hypothesis that the diversity of information in heterogeneous data *can* be leveraged to improve in-domain translation. Overall, we find that all of the proposed methods outperform their respective baselines in most settings, but that the discriminator appears the most reliable. It bested its counterparts in 4 of 6 trials, and was

EN-JA Model	**ASPEC**	**SubCrawl**
ASPEC	38.87	3.85
SubCrawl	2.74	16.91
ASPEC + SubCrawl	33.85	14.34
Discriminator	35.01	**15.38**
Adv. Discriminator	29.87	13.31
Target Token	**35.05**	14.92
EN-FR Model	**Europarl**	**OpenSubs**
Europarl	34.51	13.36
OpenSubtitles	13.12	15.2
Europarl + OpenSubs	38.26	27.9
Discriminator	39.03	**27.91**
Adv. Discriminator	38.38	25.67
Target Token	**39.1**	25.32
EN-ZH Model	**News**	**TED**
News	12.75	3.12
TED	2.79	8.41
News + TED	11.36	6.67
Discriminator	**12.88**	**8.64**
Adv. Discriminator	12.15	8.16
Target Token	11.98	7.69

Table 2: BLEU scores for models trained on various domains and languages (both mixed and unmixed). Rows correspond to training domains and columns correspond to test domains. Note that our single-domain ASPEC results are state-of-the-art, indicating the strength of these baselines.

the only approach that outperformed both individually fit and naively mixed baselines in every trial.

Figure 2c depicts the dynamics of the discriminator approach. More specifically, this figure shows the discriminator/naive-mixing performance differential as a function of domain distance. The two share a positive association, suggesting that the more distant two domains are, the more the discriminator helps performance. This may be because it is easier to classify distant domains, so the discriminator can fit the data better and its gradients encourage the upstream encoder to include more useful domain-related structure.

The adversarial discriminator architecture yielded improvements on the small datasets, but underperformed on EN-JA. It is possible that the grammatical differences inherent to casual and polite domains are such that semantic information was lost in the process of forcing their encoded distributions to match. Additionally, adversarial objective functions are notoriously difficult to optimize on, and this model was prone to falling into poor local optimum during training.

The simpler target token approach also yields improvement over the baselines, just barely surpassing that of the Discriminator for ASPEC. This approach has the practical benefit of requiring no architectural changes to an off-the-shelf NMT system.

Our EN-FR results are particularly interesting. Though the data seem like they should come from sufficiently distant domains (parliament proceedings and subtitles), the domains are actually quite close according to $\hat{d}_A$ (Table 1). Since these domains are so close, their merger is able to improve baseline performance. Thus, if the source and target domain are sufficiently close, then their merger does indeed help.

Next, we investigated the optimization dynamics of these models by examining their learning curves. Curves for the baselines and discriminative models trained on EN-JA data are depicted in Figure 3a. Single-domain training clearly outperforms mixed training, and it appears that adding a discriminative strategy provides additional gains. From Figure 3b we can see that the discriminator approach (not reversing gradients), learns to fit the domain distribution quickly, implying that the Japanese domains were in fact quite distant and easily classifiable.

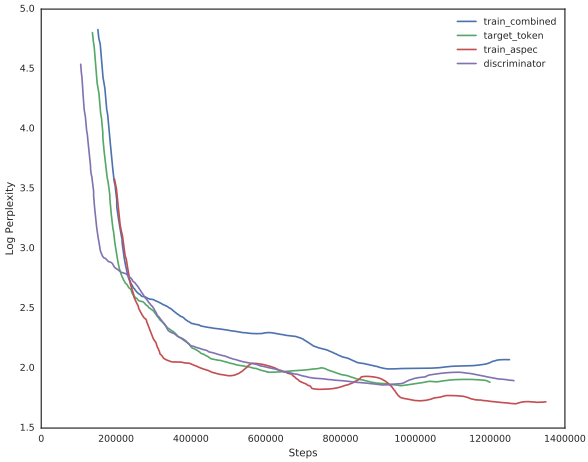
5 Related Work

Our work builds on a recent literature on domain adaptation strategies in Neural Machine Translation. Prior work in this space has proposed two general categories of methods.

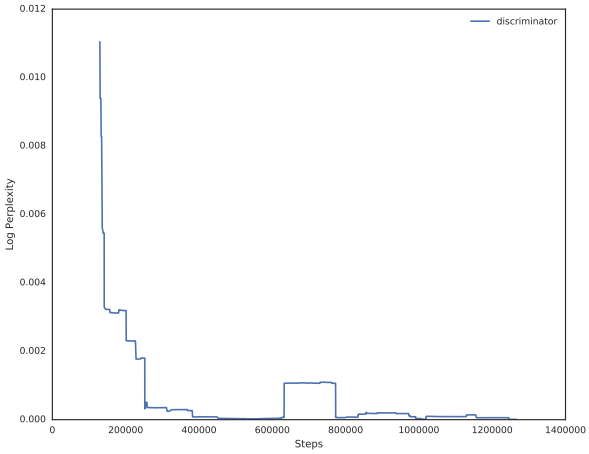
The first proposed method is to take models trained on the source domain and finetune on target-domain data. Luong and Manning (2015); Zoph et al. (2016) explores how to improve transfer learning for a low-resource language pair by finetuning only parts of the network. Chu et al. (2017) empirically evaluate domain adaptation methods and propose mixing source and target domain data during finetuning. Freitag and Al-Onaizan (2016) explored finetuning using only a small subset of target domain data. Note that we did not compare directly against these techniques because they are intended to transfer knowledge to a new domain and perform well on only the target domain. We are concerned with multi-domain settings, where performance on *all* constituent domains is important.

A second strain of "multi-domain" thought in NMT involves appending a domain indicator token to each source sequence (Kobus et al., 2016). Similarly, Johnson et al. (2016) use a token for cross-lingual translation instead of domain identification. This idea was further refined by Chu et al. (2017), who integrated source-tokenization into the domain finetuning paradigm. While it requires no changes to the NMT architecture, these approaches are inherently limited because they stipulate that domain information for unseen test examples be known. For example, if using a trained model to translate user-generated sentences, we do not know the domain a-priori, and this approach cannot be used.

Apart from the recent progress in domain adaptation for NMT, we draw on work that transfers knowledge between domains in semisupervised settings. Our strongest influence is adversarial domain adaptation (Ganin et al., 2015), where feature distributions in the source and target domains are matched



(a) Log perplexity evaluated on the ASPEC validation set. Single-domain training outperforms combined training. The discriminator and target token approaches improve over the naive combined data.



(b) Discriminator training loss over time on the EN-JA data. The discriminator learns to fit the data almost perfectly after a few hundred thousand iterations

Figure 3: Training curves for domain mixing and discriminator loss.

with a Domain-Adversarial Neural Network (DANN). Another approach to this problem is that of Long et al. (2015), which measures and minimizes the distance between domain distribution means before training, thereby negating any unique properties.

There is some overlap between past research in multi-domain statistical machine translation (SMT) and the ideas of this paper. (Farajian et al., 2017) compared the efficacy of phrase-based SMT and NMT on multiple-domain data, observing similar performance degradations as us in mixed-domain settings. However, that study did not seek to understand the issue and offered no explanation, analysis, or solution to the problem. Another line of work merged data by only selecting examples with a propensity for relevance in a multi-domain setting (Mandal et al., 2008; Axelrod et al., 2011). In a strategy that echos NMT fine-tuning, Pecina et al. (2012) used a variety of in-domain development sets to tune hyperparameters to a generalized setting. Similar to our domain discriminator network, Clark et al. (2012) crafted domain-specific features that are used by the decoder. However, some of these systems' features are downstream of binary indicators for domain identity. This approach, then, faces the same inherent limitations as source-tokenization: domain knowledge is required for inference. Furthermore, the domain features of this system are integral to the decoding process, while our discriminator network is an independent module that can be detached during inference.

6 Conclusion

We presented three novel models for applying Neural Machine Translation to multi-domain settings, and demonstrated their efficacy across six domains in three language pairs, and in the process achieved a new state-of-the-art in EN-JA translation. Unlike the naive combining of training data, these models improve their translational ability on each constituent domain. Furthermore, these models are the first of their kind to not require knowledge of each example's domain at inference time. All the proposed approaches outperform the naive combining of training data, so we advise practitioners to implement whichever most easily fits into their pre-existing pipelines, but an approach based on a discriminator network offered the most reliable results.

In future work we hope to explore the dynamics of adversarial discriminative training objectives, which force the model to learn domain-agnostic features, in the related problem of adaptation to unseen test-time domains.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* .

Anonymous. 2017. Subcrawl: A colloquial parallel corpus for english-japanese translation. *Manuscript submitted for publication.* .

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pages 355–362.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. 2007. Analysis of representations for domain adaptation. *Advances in neural information processing systems* 19:137.

D. Britz, A. Goldie, T. Luong, and Q. Le. 2017. Massive Exploration of Neural Machine Translation Architectures. *ArXiv e-prints* .

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of simple domain adaptation methods for neural machine translation. *CoRR* abs/1701.03214. http://arxiv.org/abs/1701.03214.

Jonathan H Clark, Alon Lavie, and Chris Dyer. 2012. One system, many domains: Open-domain statistical machine translation via feature augmentation .

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science* 14(2):179–211.

M Amin Farajian, Marco Turchi, Matteo Negri, Nicola Bertoldi, and Marcello Federico. 2017. Neural vs. phrase-based machine translation in a multi-domain scenario. *EACL 2017* page 280.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *CoRR* abs/1612.06897. http://arxiv.org/abs/1612.06897.

Yaroslav Gani, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2015. Domain-adversarial training of neural networks. arxiv preprint. *arXiv preprint arXiv:1505.07818* .

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. 2015. Domain-Adversarial Training of Neural Networks. *ArXiv e-prints* .

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. pages 513–520.

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.

Barry Haddow and Philipp Koehn. 2012. Analysing the effect of out-of-domain data on smt systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 422–432.

Motoko Hori. 1986. A sociolinguistic analysis of the japanese honorifics. *Journal of pragmatics* 10(3):373–386.

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. http://arxiv.org/abs/1611.04558.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*. volume 5, pages 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open

source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, pages 177–180.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. 2015. Learning transferable features with deep adaptation networks. In *ICML*. pages 97–105.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015a. Effective approaches to attention-based neural machine translation. In *EMNLP*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015b. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* .

Arindam Mandal, Dimitra Vergyri, Wen Wang, Jing Zheng, Andreas Stolcke, Gokhan Tur, D Hakkani-Tur, and Necip Fazil Ayan. 2008. Efficient data selection for machine translation. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*. IEEE, pages 261–264.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*. pages 2204–2208.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pages 529–533.

Pavel Pecina, Antonio Toral, and Josef Van Genabith. 2012. Simple and effective parameter tuning for domain adaptation of statistical machine translation. In *COLING*. pages 2209–2224.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* .

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*. volume 2012, pages 2214–2218.

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.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *CoRR* abs/1604.02201. http://arxiv.org/abs/1604.02201.