Seeking for Help or Standing on Its Own Feet? Rethinking Semi-Supervised Domain Adaptation for Machine Translation

Di Jin[1] **Zhijing Jin**[2] **Joey Tianyi Zhou**[3] **Peter Szolovits**[1]
[1]CSAIL, MIT [2]Max Planck Institute [3]A*STAR, Singapore
{jindi15,psz}@mit.edu
zhijing.jin@connect.hku.hk
zhouty@ihpc.a-star.edu.sgu

Abstract

Semi-supervised domain adaptation for machine translation aims to learn the translation between two languages in a specific domain, given only (1) monolingual in-domain data, and (2) parallel out-of-domain data. The majority of existing research focuses on innovating domain adaptation (DA) methods to benefit from the out-of-domain data, but uses only simple methods for the monolingual in-domain data. In this paper, we analyze such imbalanced research efforts, and find that by simply using an improved unsupervised learning on the monolingual data only, we can outperform the best DA approach by up to +18.66 BLEU scores. Moreover, we find that combining this improved unsupervised learning with a standard DA technique leads to a even more significant improvement, up to +20.00 higher than the strongest previous model, and up to +47.69 over an unadapted model.[1]

1 Introduction

Semi-supervised domain adaptation for machine translation (semi-DAMT) aims to learn the translation between two languages in a certain domain (i.e., in-domain) with only non-parallel data by leveraging parallel data from other domains (i.e., out-of-domain). This task has wide applications, as it is common to have only monolingual, unsupervised data for domains that need specific expertise, such as law, medicine, and technology (Nakov and Tiedemann, 2012), while in some other domains, a large amount of supervised data could be available, such as European parliament proceedings (e.g., 4.5M translation pairs in WMT-14 dataset).

Existing approaches for semi-DAMT take extensive efforts to propose better domain adaptation (DA) techniques to *learn the supervised out-of-domain data*, while using relatively simple methods to *learn unsupervised in-domain data*. In this paper, we examine this loss of balance – excessive research efforts on making use of out-of-domain supervised data versus much less attention to in-domain unsupervised data. We first discuss the limitations of out-of-domain learning, and then we present two investigations and several ways to **boost** the **d**omain **a**daptation using our model DABooster. Investigation #1 checks whether the current approaches have made the best use of monolingual data, and shows that $\text{DABooster}_{\text{m-only}}$, which simply uses **m**onolingual data, can outperform all previous semi-DAMT methods (Section 3). Investigation #2 explores the performance improvement after incorporating $\text{DABooster}_{\text{m-only}}$ into a common domain adaptation learning scheme using the out-of-domain **p**arallel data. This new model, $\text{DABooster}_{\text{m\&p(bt)}}$, boosts the performance by a *surprisingly* large margin (Section 4).

For each investigation, we set up experiments covering a wide range of settings, including two language pairs, nine domain pairs, and five baseline models. Empirical results show that our $\text{DABooster}_{\text{m-only}}$ in Investigation #1 leads to at most +18.66 BLEU scores over the strongest baseline, and our $\text{DABooster}_{\text{m\&p(bt)}}$ in Investigation #2 achieves an even higher improvement, up to +20.00 BLEU points over the best previous method, and +47.69 over the unadapted model.

2 Background of semi-DAMT

Previous work The setting of semi-DAMT provides two types of data: (1) supervised data that are out-of-domain, and (2) unsupervised data which are in-domain and sometimes also additional general text corpora. Most semi-DAMT work carefully designs domain adaptation (DA) methods to make the best of type (1) data. For example, Duh et al. (2013) use language models to rank the out-of-

[1]Our code is available at https://github.com/jind11/DAMT

domain data and select the top-ranked parallel sentences to train on. Sennrich et al. (2015) use models learned on (1) to back-translate and generate pseudo-parallel corpora. Currey et al. (2017) copy the target-side in-domain sentences to the source side and then mix with the out-of-domain corpus for translation training. Compared with research efforts in (1), there is less work focused on (2). Most uses auto-encoding (Cheng et al., 2016) and language modeling (Ramachandran et al., 2017; Dou et al., 2019) on the in-domain target language corpus.

Note that the setting of our task, semi-DAMT, is different from the settings of the other two tasks, supervised DAMT and unsupervised DAMT. Supervised DAMT assumes that both in-domain and out-of-domain supervised data, while unsupervised DAMT assumes unsupervised data in all domains.

Limitations of DA Although previous work focuses extensively on how to use the out-of-domain data by different DA methods. There are some intrinsic limitations of using the out-of-domain data. For example, there can be a limited amount of out-of-domain parallel data as some low-resource languages have little parallel data in any domain. Another problem is the domain shift between out-of-domain data and in-domain data. With a large domain drift, DA could fail to bring in performance gain (Guzmán et al., 2019).

3 Question 1: Have we made the best of monolingual data yet?

Considering the limitations mentioned above, we propose that the solution to semi-DAMT should be an appropriate combination of (1) out-of-domain DA and (2) unsupervised learning on monolingual data. As we have discussed previously that the monolingual data is largely under-explored, our first investigation is to check how much more improvement we can make of the monolingual data. In this section, we first establish a simple but strong baseline using the techniques in unsupervised machine translation (UMT) to learn monolingual-only data, and compare it with previous methods using both monolingual and out-of-domain parallel data.

3.1 DABooster on monolingual data only

Our baseline, $\text{DABooster}_{\text{m-only}}$, adopts the Transformer (Vaswani et al., 2017) with the encoder-decoder structure as the sequence-to-sequence translation model. We add the language embeddings to the standard token and position embeddings via the element-wise summation operation, so that the model can be shared for any translation direction.

We first pre-train the encoder and decoder by language modeling (LM) over the Wikipedia monolingual corpora for both languages. Then, on in-domain monolingual data, we optimize two learning objectives, LM and iterative back-translation (IBT). LM is implemented via denoising auto-encoding, by minimizing the loss to reconstruct a sentence from noises introduced by random drop, mask, or swapping of the words (Freitag and Roy, 2018). For IBT, we denote two translation directions: $\text{Model}_{\text{s2t}}(\cdot)$ which translates from the source language s to the target language t, and $\text{Model}_{\text{t2s}}(\cdot)$ vice versa. In each iteration, we translate on the fly from each source language sentence $\boldsymbol{x} \in X_{\text{in}}$ to the target language $\text{Model}_{\text{s2t}}(\boldsymbol{x})$, and do so for the other direction. Then the pairs of $(\boldsymbol{x}, \text{Model}_{\text{s2t}}(\boldsymbol{x}))$ and $(\text{Model}_{\text{t2s}}(\boldsymbol{y}), \boldsymbol{y})$ can be used as synthetic parallel data to train the translation model in both directions (Hoang et al., 2018).

3.2 Previous approaches using DA

We compare with five existing approaches for semi-DAMT: (1) UNADAPTED, a supervised machine translation model trained on the out-of-domain parallel data, and directly tested on in-domain data. (2) COPY (Currey et al., 2017) copies the in-domain sentences from the target language corpus to the source language corpus, and trains on both this data and out-of-domain parallel data. (3) BACK (Sennrich et al., 2015) uses back-translation as a one-time data augmentation, which first learns a target-to-source translation model from the out-of-domain parallel data, and then uses it to generate in-domain pseudo parallel data. (4) DALI (Hu et al., 2019) fine-tunes the model trained on out-of-domain data on word-to-word translation of monolingual in-domain data by a learned in-domain lexicon. (5) DAFE (Dou et al., 2019) multitasks both supervised translation on out-of-domain data, and LM on in-domain target language corpus, with additional domain and task embedding learners.

3.3 Experiment design

The goal of the experiments is to compare the previous DA-based approaches with our baseline which simply uses a more effective learning on monolingual data only. We conduct this comparison under

a wide range of settings:

1. Language pairs: German-English (De-En), and Romanian-English (Ro-En)
2. General domain datasets: WMT-14 for De-En, and WMT-16 for Ro-En
3. Specific domain datasets: MED, LAW, IT (Tiedemann, 2012), and TED (Duh, 2018).

In terms of adaptation directions, we experiment on domain adaptations from general domains with large supervised data (i.e., WMT) to specific domains with only monolingual data (incl., MED, LAW, IT, and TED). In addition, we also test domain adaptations among specific domains (incl., MED, LAW, and IT), resulting in six domain pairs (e.g., MED→LAW, MED→IT, LAW→IT, etc.).

Note that all original datasets are parallel. For experiments involving non-parallel data, we randomly shuffle the original parallel corpus, and extract the source language text in the first half, and the target language text in the second half, which follows the practice in (Hu et al., 2019; Dou et al., 2019). More dataset details are in Appendix A.1.

3.4 Results

Adapting from general to specific domains Table 1 lists the results of domain adaptation from the general domain, WMT (W), to specific domains including LAW(L), MED(M), and TED (T). Our method DABooster$_{\text{m-only}}$ is better than the unadapted baseline, UNADAPTED, by a large margin, and on par with/better than the best previous approaches for De-En W→M, Ro-En W→L, and Ro-En W→M. It does not outperform the baseline DAFE in some other cases.

Models	De-En			Ro-En		
	W→L	W→M	W→T	W→L	W→M	W→T
UNADAPTED	23.77	24.42	23.36	33.26	18.39	23.59
COPY	26.17	29.33	29.79	36.49	22.73	26.44
BACK	30.24	33.16	34.46	42.08	30.45	32.70
DAFE	31.46	38.79	34.89	49.63†	46.77†	37.05†
DABooster$_{\text{m-only}}$	27.89	38.67	30.88	49.45	61.55	34.48

Table 1: BLEU scores of semi-DAMT from the general domain, WMT (W), to specific domains, incl. LAW(L), MED(M), and TED (T). DAFE reports the best results from the original paper, except for several re-implementations marked by †.

Adapting between specific domains In Table 2, we have the results of domain adaptation between specific domains. Notably, DABooster$_{\text{m-only}}$ achieves +11.71∼18.66 BLEU score improvement over the strongest baseline in all cases, and +19.91∼27.77 BLEU scores over UNADAPTED.

Models	L→M	L→I	M→L	M→I	I→L	I→M
UNADAPTED	18.76	6.62	7.92	5.94	6.19	10.90
COPY	19.23	7.57	6.01	5.89	5.11	11.15
BACK	22.53	11.34	7.62	7.02	8.06	14.56
DALI	11.32	8.75	26.98	19.49	11.65	10.99
DAFE	26.96	15.41	14.28	13.03	11.67	21.30
DABooster$_{\text{m-only}}$	**38.67**	**31.69**	**27.89**	**31.69**	**27.89**	**38.67**

Table 2: BLEU scores of semi-DAMT on De-En in between specific domains, LAW(L), MED(M), and IT (I).

Summary It is astounding that our model DABooster$_{\text{m-only}}$, which only uses monolingual data, can be on par with or, in many settings, substantially outperform all previous approaches that use DA on parallel out-of-domain data. It indicates that previous methods have not thoroughly explored the potential of the in-domain data. DABooster$_{\text{m-only}}$ can have an even larger advantage when there is less out-of-domain data, or a larger domain shift.

4 Question 2: How much can the improved unsupervised learning boost semi-DAMT?

As our proposed improvement in unsupervised learning on monolingual data is orthogonal to DA on out-of-domain data, we further investigate, after combining our model with DA techniques, how much improvement the model can finally achieve.

4.1 Combined approach

As illustrated in Figure 1, we add in out-of-domain data and use DA techniques in addition to our UMT approach. First, we use the same pretraining on Wikipedia as DABooster$_{\text{m-only}}$. Then we train the model by iteratively optimizing the (1) supervision loss on the out-of-domain data, (2) IBT loss in both translation directions, and (3) two LM losses on both languages in the in-domain data. We denote this method which uses both **m**onolingual and **p**arallel data as **DABooster$_{\text{m\&p}}$**.

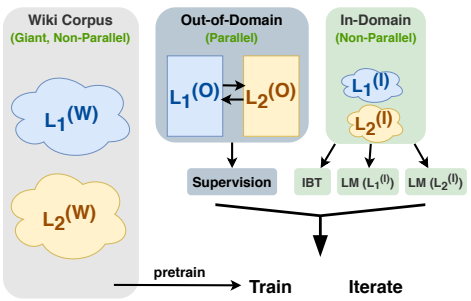

Figure 1: Illustration of our method DABooster$_{\text{m\&p}}$.

Instead of using the supervised out-of-domain

Models	De-En									Ro-En		
	L→M	L→I	M→L	M→I	I→L	I→M	W→T	W→L	W→M	W→T	W→L	W→M
UNADAPTED	18.76	6.62	7.92	5.94	6.19	10.90	23.36	23.77	24.42	23.59	33.26	18.39
Best Previous	26.96	15.41	14.28	13.03	11.67	21.30	**34.89**	31.46	38.79	37.05	49.63	46.77
$\text{DABooster}_{\text{in-only}}$	38.67	31.69	27.89	31.69	27.89	38.67	30.88	27.89	38.67	34.48	49.45	61.55
$\text{DABooster}_{\text{m\&p}}$	**41.22**	34.33	29.54	32.47	30.20	39.77	33.23	32.81	41.40	38.68	53.49	60.98
$\text{DABooster}_{\text{m\&p(bt)}}$	40.40	**35.41**	**30.27**	**35.76**	**30.49**	**40.28**	34.15	**33.35**	**42.08**	**38.90**	**54.39**	**66.08**

Table 3: BLEU scores of semi-DAMT on domains including LAW (L), MED (M), IT (I), WMT (W), and TED (T).

data directly, another alternative setting is to use BACK's method to create in-domain pseudo-parallel data by a translation model learned on out-of-domain data. The remaining training strategies are the same as the approach outlined above. We denote this method as $\mathbf{DABooster_{m\&p(bt)}}$.

4.2 Results

In Table 3, $\text{DABooster}_{\text{m\&p}}$ can bring in $+1\sim4$ BLEU score improvements over $\text{DABooster}_{\text{m-only}}$. Moreover, $\text{DABooster}_{\text{m\&p(bt)}}$, whose additional supervision comes from the back-translated in-domain data, achieves even more significant performance. Results of these two new models demonstrate the benefit from the supervised data. $\text{DABooster}_{\text{m\&p(bt)}}$ is better than $\text{DABooster}_{\text{m\&p}}$ in most cases, indicating that supervised data with a more similar distribution can, in general, be more helpful to the model.

Overall, our best setting can gain up to $+20.00$ BLEU score improvement over the best baseline model and $+47.69$ over the UNADAPTED model.

4.3 Ablation study

To check the effect of the each component in our model, we conduct an ablation study on our best performing model, $\text{DABooster}_{\text{m\&p(bt)}}$. In Table 4, we report the validation set BLEU by adapting from LAW to MED and IT. The most important components are *Pre-training* and *IBT*, without which the model suffers from a substantial performance degradation by around $8\sim22$ BLEU scores. We find that *LM* is also important, but not as crucial as the first two components.

Model	LAW→MED	LAW→IT
$\text{DABooster}_{\text{m\&p(bt)}}$	**42.13**	**47.64**
– Pre-training	31.80 (↓10.33)	27.71 (↓19.93)
– IBT	33.75 (↓8.38)	25.37 (↓22.27)
– LM	40.97 (↓1.16)	40.82 (↓6.82)

Table 4: Ablation study of $\text{DABooster}_{\text{m\&p(bt)}}$.

4.4 Additional in-domain monolingual data

One advantage of our method is that it makes more thorough use of the monolingual data than other previous methods. Therefore, we hypothesize that given additional monolingual in-domain data, it will gain more improvement. We verify it through a case study on domain adaptation from WMT to TED with extra $\sim$200K monolingual TED data we collected from TED talks. Details of this dataset are in Appendix A.3.

Model	+ Data	De-En W→T	Ro-En W→T
$\text{DABooster}_{\text{m-only}}$	–	30.88	34.48
$\text{DABooster}_{\text{m-only}}$	✓	**33.34**	**39.01**
$\text{DABooster}_{\text{m\&p(bt)}}$	–	34.15	38.90
$\text{DABooster}_{\text{m\&p(bt)}}$	✓	**36.45**	**40.92**
Supervised	–	38.97	42.22

Table 5: Test set BLEU scores of WMT-to-TED (W→T) adaptation with extra TED data ("+Data").

We add the additional TED data to the in-domain monolingual corpus, and analyze the performance improvement of $\text{DABooster}_{\text{m-only}}$ and $\text{DABooster}_{\text{m\&p(bt)}}$. As shown in Table 5, our model can improve substantially by the extra non-parallel data by up to $+4.53$, and our best setting achieves performance very close to supervised translation performance. This shows a large potential, as it is often feasible to collect more monolingual data to boost our model performance.

5 Conclusion

In this paper, we made two important contributions. First is identification of the potential of unsupervised learning on monolingual data for semi-DAMT. Second is the proposal of a much stronger baseline for semi-DAMT by integrating our improved unsupervised learning with the DA techniques on parallel out-of-domain data, which brings up to $+20$ BLEU score improvement over the best previous model. Extensive experiments have verified the effectiveness of our model.

References

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. Association for Computational Linguistics.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *WMT*.

Zi-Yi Dou, Junjie Hu, Antonios Anastasopoulos, and Graham Neubig. 2019. Unsupervised domain adaptation for neural machine translation with domain-aware feature embeddings. In *EMNLP/IJCNLP*.

Kevin Duh. 2018. The multitarget ted talks task. http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/.

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. Association for Computational Linguistics.

Markus Freitag and Scott Roy. 2018. Unsupervised natural language generation with denoising autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3922–3929, Brussels, Belgium. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–305, Jeju Island, Korea. Association for Computational Linguistics.

Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

A Appendices

A.1 Dataset Statistics

Statistics of the training set of the original parallel datasets are in Table 6. The size of validation and test sets for WMT-14 are 3K, and all the other domains are 2K.

Lang.	Corpus	Words	Sentences	W/S
De-En	MED	14,533,613	1,104,752	13.2
	LAW	18,461,140	715,372	25.8
	IT	3,212,130	337,817	9.5
	TED	3,110,970	151,627	20.5
	WMT-14	126,735,962	4,468,840	28.4
Ro-En	MED	13,142,512	990,220	13.3
	LAW	10,631,517	450,715	23.6
	TED	3,328,621	161,291	20.6
	WMT-16	10,796,138	399,375	27.0

Table 6: Statistics of the original parallel datasets (target language). W/S is the number of words per sentence.

A.2 Implementation Details

Both encoder and decoder in the transformer model have 6 layers, 8 heads, and a dimension of 1024. For the word corruption function, word dropping and blanking adopt a uniform distribution with a probability of 0.1, and word shuffling is implemented with a window of 3 tokens. The Adam optimizer uses a learning rate of 0.0001. Sentence tokenization was done by the SpaCy tokenizer.[2]

A.3 Newly Collected TED Data

Language Pair	Size of Source	Size of Target
De-En	245,837	229,994
Ro-En	225,154	213,958

Table 7: Sentence numbers of our additional TED data.

We collected an additional dataset of TED talks. After scraping all the TED talk web-pages[3] until the early January 2020, we extracted the transcripts in three languages, English, German and Romanian, and kept the unique TED talk identifier of the transcript.[4] Note that we removed the TED talks that have already appeared in our original TED dataset to avoid duplication by checking the TED talk identifier. Table 7 summarizes the sentence numbers of the added monolingual data. When combining the extra non-parallel data with the original in-domain data, we always up-sample the original data via replication so that it can have the same size as the additional data.

[2] https://spacy.io/

[3] https://www.ted.com/

[4] Our newly collected TED dataset will be available after acceptance.