Domain Adaptive Inference for Neural Machine Translation

Danielle Saunders[†] and **Felix Stahlberg**[†] and **Adrià de Gispert**[‡] and **Bill Byrne**[‡†]
[†]Department of Engineering, University of Cambridge, UK
[‡]SDL Research, Cambridge, UK
{ds636, fs439, wjb31}@cam.ac.uk, {agispert, bbyrne}@sdl.com

Abstract

We investigate adaptive ensemble weighting for Neural Machine Translation, addressing the case of improving performance on a new and potentially unknown domain without sacrificing performance on the original domain. We adapt sequentially across two Spanish-English and three English-German tasks, comparing unregularized fine-tuning, L2 and Elastic Weight Consolidation. We then report a novel scheme for adaptive NMT ensemble decoding by extending Bayesian Interpolation with source information, and show strong improvements across test domains without access to the domain label.

1 Introduction

Neural Machine Translation (NMT) models are effective when trained on broad domains with large datasets, such as news translation (Bojar et al., 2017). However, test data may be drawn from a different domain, on which general models can perform poorly (Koehn and Knowles, 2017). We address the problem of adapting to one or more domains while maintaining good performance across all domains. Crucially, we assume the realistic scenario where the domain is unknown at inference time.

One solution is ensembling models trained on different domains (Freitag and Al-Onaizan, 2016). This approach has two main drawbacks. Firstly, obtaining models for each domain is challenging. Training from scratch on each new domain is impractical, while continuing training on a new domain can cause catastrophic forgetting of previous tasks (French, 1999), even in an ensemble (Freitag and Al-Onaizan, 2016). Secondly, ensemble weighting requires knowledge of the test domain.

We address the model training problem with regularized fine-tuning, using an L2 regularizer (Barone et al., 2017) and Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017). We fine-tune sequentially to translate up to three domains with the same model.

We then develop an adaptive inference scheme for NMT ensembles by extending Bayesian Interpolation (BI) (Allauzen and Riley, 2011) to sequence-to-sequence models.[1] This lets us calculate ensemble weights adaptively over time without needing the domain label, giving strong improvements over uniform ensembling for baseline and fine-tuned models.

1.1 Adaptive training

In NMT fine-tuning, a model is first trained on a task A, typically translating a large general-domain corpus (Luong and Manning, 2015). The optimized parameters θ^*_A are fine-tuned on task B, a new domain. Without regularization, catastrophic forgetting can occur: performance on task A degrades as parameters adjust to the new objective. A regularized objective is:

$$L(\theta) = L_B(\theta) + \Lambda \sum_j F_j(\theta_j - \theta^*_{A,j})^2 \qquad (1)$$

where $L_A(\theta)$ and $L_B(\theta)$ are the likelihood of tasks A and B. We compare three cases:

- **No-reg**, where $\Lambda = 0$
- **L2**, where $F_j = 1$ for each parameter index j
- **EWC**, where $F_j = \mathbb{E}\left[\nabla^2 L_A(\theta_j)\right]$, a sample estimate of task A Fisher information. This effectively measures the importance of θ_j to task A.

For L2 and EWC we tune Λ on the validation sets for new and old tasks to balance forgetting against new-domain performance.

[1] See `bayesian` combination schemes at https://github.com/ucam-smt/sgnmt

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 222–228
Florence, Italy, July 28 - August 2, 2019.

1.2 Adaptive decoding

We extend the BI formalism to condition on a source sequence, letting us apply it to adaptive NMT ensemble weighting. We consider models $p_k(\mathbf{y}|\mathbf{x})$ trained on K distinct domains, used for tasks $t = 1, \ldots, T$. In our case a task is decoding from one domain, so $T = K$. We assume throughout that $p(t) = \frac{1}{T}$, i.e. that tasks are equally likely absent any other information.

A standard, fixed-weight ensemble would translate with:

$$\underset{\mathbf{y}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{x}) = \underset{\mathbf{y}}{\operatorname{argmax}} \sum_{k=1}^{K} W_k p_k(\mathbf{y}|\mathbf{x}) \quad (2)$$

The BI formalism assumes that we have tuned sets of ensemble weights $\lambda_{k,t}$ for each task. This defines a task-conditional ensemble

$$p(\mathbf{y}|\mathbf{x}, t) = \sum_{k=1}^{K} \lambda_{k,t}\, p_k(\mathbf{y}|\mathbf{x}) \quad (3)$$

which can be used as a fixed weight ensemble if the task is known. However if the task t is not known, we wish to translate with:

$$\underset{\mathbf{y}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{x}) = \underset{\mathbf{y}}{\operatorname{argmax}} \sum_{t=1}^{T} p(t, \mathbf{y}|\mathbf{x}) \quad (4)$$

At step i, where h_i is history $\mathbf{y}_{1:i-1}$:

$$\begin{aligned} p(y_i|h_i, \mathbf{x}) &= \sum_{t=1}^{T} p(t, y_i|h_i, \mathbf{x}) \\ &= \sum_{t=1}^{T} p(t|h_i, \mathbf{x})\, p(y_i|h_i, t, \mathbf{x}) \\ &= \sum_{k=1}^{K} p_k(y_i|h_i, \mathbf{x}) \sum_{t=1}^{T} p(t|h_i, \mathbf{x}) \lambda_{k,t} \\ &= \sum_{k=1}^{K} W_{k,i}\, p_k(y_i|h_i, \mathbf{x}) \end{aligned} \quad (5)$$

This has the form of an adaptively weighted ensemble where, by comparison with Eq. 2:

$$W_{k,i} = \sum_{t=1}^{T} p(t|h_i, \mathbf{x}) \lambda_{k,t} \quad (6)$$

In decoding, at each step i adaptation relies on a recomputed estimate of the *task posterior*:

$$p(t|h_i, \mathbf{x}) = \frac{p(h_i|t, \mathbf{x}) p(t|\mathbf{x})}{\sum_{t'=1}^{T} p(h_i|t', \mathbf{x}) p(t'|\mathbf{x})} \quad (7)$$

1.2.1 Static decoder configurations

In static decoding (Eq. 2), the weights W_k are constant for each source sentence $\mathbf{x}$. BI simplifies to a uniform ensemble when $\lambda_{k,t} = p(t|\mathbf{x}) = \frac{1}{T}$. This leads to $W_{k,i} = \frac{1}{K}$ (see Eq. 6) as a fixed equal-weight interpolation of the component models.

Static decoding can also be performed with task posteriors conditioned only on the source sentence, which reflects the assumption that the history can be disregarded and that $p(t|h_i, \mathbf{x}) = p(t|\mathbf{x})$. In the most straightforward case, we assume that only domain k is useful for task t: $\lambda_{k,t} = \delta_k(t)$ (1 for $k = t$, 0 otherwise). Model weighting simplifies to a fixed ensemble:

$$W_k = p(k|\mathbf{x}) \quad (8)$$

and decoding proceeds according to Eq. 2. We refer to this as decoding with an *informative source* (IS).

We propose using G_t, an collection of n-gram language models trained on source language sentences from tasks t, to estimate $p(t|\mathbf{x})$:

$$p(t|\mathbf{x}) = \frac{p(\mathbf{x}|t)p(t)}{\sum_{t'=1}^{T} p(\mathbf{x}|t')p(t')} = \frac{G_t(\mathbf{x})}{\sum_{t'=1}^{T} G_{t'}(\mathbf{x})} \quad (9)$$

In this way we use source language n-gram language models to estimate $p(t = k|\mathbf{x})$ in Eq. 8 for static decoding with an informative source.

1.2.2 Adaptive decoder configurations

For adaptive decoding with Bayesian Interpolation, as in Eq. 5, the model weights vary during decoding according to Eq. 6 and Eq. 7. We assume here that $p(t|\mathbf{x}) = p(t) = \frac{1}{T}$. This corresponds to the approach in Allauzen and Riley (2011), which considers only language model combination for speech recognition. We refer to this in experiments simply as BI. A refinement is to incorporate Eq. 9 into Eq. 7, which would be Bayesian Interpolation with an informative source (BI+IS).

We now address the choice of $\lambda_{k,t}$. A simple but restrictive approach is to take $\lambda_{k,t} = \delta_k(t)$. We refer to this as *identity-BI*, and it embodies the assumption that only one domain is useful for each task.

Alternatively, if we have validation data V_t for each task t, parameter search can be done to optimize $\lambda_{k,t}$ for BLEU over V_t for each task. This is straightforward but relatively costly.

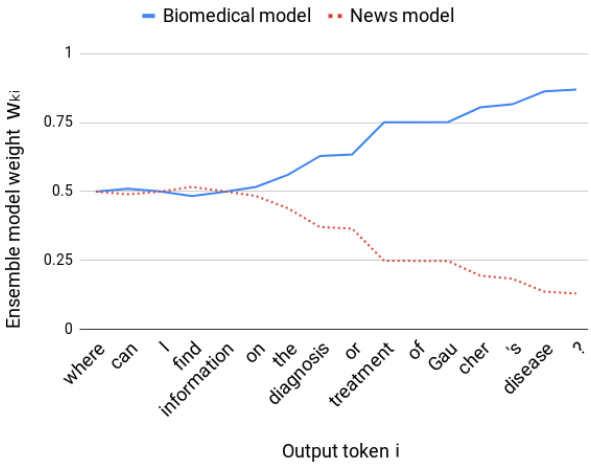

Figure 1: Adaptively adjusting ensemble model weights $W_{k,i}$ (Eq. 6) during decoding with BI

We propose a simpler approach based on the source language n-gram language models from Eq. 9. We assume that each G_t is also a language model for its corresponding domain k. With $\overline{G}_{k,t} = \sum_{\mathbf{x} \in V_t} G_k(\mathbf{x})$, we take:

$$\lambda_{k,t} = \frac{\overline{G}_{k,t}}{\sum_{k'} \overline{G}_{k',t}} \tag{10}$$

$\lambda_{k,t}$ can be interpreted as the probability that task t contains sentences $\mathbf{x}$ drawn from domain k as estimated over the V_t.

Figure 1 demonstrates this adaptive decoding scheme when weighting a biomedical and a general (news) domain model to produce a biomedical sentence under BI. The model weights $W_{k,i}$ are even until biomedical-specific vocabulary is produced, at which point the in-domain model dominates.

1.2.3 Summary

We summarize our approaches to decoding in Table 1.

	Decoder	$\mathbf{p(t\|x)}$	$\lambda_{\mathbf{k,t}}$
Static	Uniform	$\frac{1}{T}$	$\frac{1}{T}$
	IS	Eq. 9	$\delta_k(t)$
Adaptive	Identity-BI	$\frac{1}{T}$	$\delta_k(t)$
	BI	$\frac{1}{T}$	Eq. 10
	BI+IS	Eq. 9	Eq. 10

Table 1: Setting task posterior $p(t|\mathbf{x})$ and domain-task weight $\lambda_{k,t}$ for T tasks under decoding schemes in this work. Note that IS can be combined with either Identity-BI or BI by simply adjusting $p(t|h_i, \mathbf{x})$ according to Eq. 7.

1.3 Related Work

Approaches to NMT domain adaptation include training data selection or generation (Sennrich et al., 2016a; Wang et al., 2017; Sajjad et al., 2017) and fine-tuning output distributions (Dakwale and Monz, 2017; Khayrallah et al., 2018).

Vilar (2018) regularizes parameters with an importance network, while Thompson et al. (2018) freeze subsets of the model parameters before fine-tuning. Both observe forgetting with the adapted model on the general domain data in the realistic scenario where the test data domain is unknown. Barone et al. (2017) fine-tune with L2 regularization to reduce forgetting. Concurrently with our work, Thompson et al. (2019) apply EWC to reduce forgetting during NMT domain adaptation.

During inference, Garmash and Monz (2016) use a gating network to learn weights for a multi-source NMT ensemble. Freitag and Al-Onaizan (2016) use uniform ensembles of general and no-reg fine-tuned models.

2 Experiments

We report on Spanish-English (es-en) and English-German (en-de). For es-en we use the Scielo corpus (Neves et al., 2016), with Health as the general domain, adapting to Biological Sciences ('Bio'). We evaluate on the domain-labeled Health and Bio 2016 test data.

The en-de general domain is the WMT18 News task (Bojar et al., 2017), with all data except ParaCrawl oversampled by 2 (Sennrich et al., 2017). We validate on newstest17 and evaluate on newstest18. We adapt first to the IWSLT 2016 TED task (Cettolo et al., 2016), and then sequentially to the APE 2017 IT task (Turchi et al., 2017).

We filter training sentences for minimum three tokens and maximum 120 tokens, and remove sentence pairs with length ratios higher than 4.5:1 or lower than 1:4.5. Table 2 shows filtered training sentence counts. Each language pair uses a 32K-merge source-target BPE vocabulary trained on the general domain (Sennrich et al., 2016b).

We implement in Tensor2Tensor (Vaswani et al., 2018) and use its base Transformer model (Vaswani et al., 2017) for all NMT models. At inference time we decode with beam size 4 in SGNMT (Stahlberg et al., 2017) and evaluate with case-sensitive detokenized BLEU using SacreBLEU (Post, 2018). For BI, we use 4-gram KENLM models (Heafield, 2011).

Language pair	Domain	Training sentences
es-en	Health	586K
	Bio	125K
en-de	News	22.1M
	TED	146K
	IT	11K

Table 2: Corpora training sentence counts

2.1 Adaptive training results

	Training scheme	Health	Bio
1	Health	**35.9**	33.1
2	Bio	29.6	36.1
3	Health and Bio	35.8	37.2
4	1 then Bio, No-reg	30.3	36.6
5	1 then Bio, L2	35.1	37.3
6	1 then Bio, EWC	35.2	**37.8**

Table 3: Test BLEU for es-en adaptive training. EWC reduces forgetting compared to other fine-tuning methods, while offering the greatest improvement on the new domain.

	Training scheme	News	TED	IT
1	News	37.8	25.3	35.3
2	TED	23.7	24.1	14.4
3	IT	1.6	1.8	39.6
4	News and TED	38.2	25.5	35.4
5	1 then TED, No-reg	30.6	**27.0**	22.1
6	1 then TED, L2	37.9	26.7	31.8
7	1 then TED, EWC	**38.3**	**27.0**	33.1
8	5 then IT, No-reg	8.0	6.9	56.3
9	6 then IT, L2	32.3	22.6	56.9
10	7 then IT, EWC	35.8	24.6	**57.0**

Table 4: Test BLEU for en-de adaptive training, with sequential adaptation to a third task. EWC-tuned models give the best performance on each domain.

We wish to improve performance on new domains without reduced performance on the general domain, to give strong models for adaptive decoding. For es-en, the Health and Bio tasks overlap, but catastrophic forgetting still occurs under no-reg (Table 3). Regularization reduces forgetting and allows further improvements on Bio over no-reg fine-tuning. We find EWC outperforms the L2 approach of Barone et al. (2017) in learning the new task and in reduced forgetting.

In the en-de News/TED task (Table 4), all fine-tuning schemes give similar improvements on TED. However, EWC outperforms no-reg and L2 on News, not only reducing forgetting but giving 0.5 BLEU improvement over the baseline News model.

The IT task is very small: training on IT data alone results in over-fitting, with a 17 BLEU improvement under fine-tuning. However, no-reg fine-tuning rapidly forgets previous tasks. EWC reduces forgetting on two previous tasks while further improving on the target domain.

2.2 Adaptive decoding results

At inference time we may not know the test data domain to match with the best adapted model, let alone optimal weights for an ensemble on that domain. Table 5 shows improvements on data without domain labelling using our adaptive decoding schemes with unadapted models trained only on one domain (models 1+2 from Table 3 and 1+2+3 from Table 4). We compare with the 'oracle' model trained on each domain, which we can only use if we know the test domain.

Uniform ensembling under-performs all oracle models except es-en Bio, especially on general domains. Identity-BI strongly improves over uniform ensembling, and BI with λ as in Eq. 10 improves further for all but es-en Bio. BI and IS both individually outperform the oracle for all but IS-News, indicating these schemes do not simply learn to select a single model.

The combined scheme of BI+IS outperforms either BI or IS individually, except in en-de IT. We speculate IT is a distinct enough domain that $p(t|x)$ has little effect on adapted BI weights.

In Table 6 we apply the best adaptive decoding scheme, BI+IS, to models fine-tuned with EWC. The es-en ensemble consists of models 1+6 from Table 3 and the en-de ensemble models 1+7+10 from Table 4. As described in Section 2.1 EWC models perform well over multiple domains, so the improvement over uniform ensembling is less striking than for unadapted models. Nevertheless adaptive decoding improves over both uniform ensembling and the oracle model in most cases.

With adaptive decoding, we do not need to assume whether a uniform ensemble or a single model might perform better for some potentially unknown domain. We highlight this in Table 7 by reporting results with the ensembles of Tables 5 and 6 over concatenated test sets, to mimic the realistic scenario of unlabelled test data. We additionally include the uniform no-reg ensembling approach given in Freitag and Al-Onaizan (2016) using models 1+4 from Table 3 and 1+5+8 from Table 4.

Uniform no-reg ensembling outperforms unadapted uniform ensembling, since fine-tuning gives better in-domain performance. EWC

Decoder configuration	es-en		en-de		
	Health	Bio	News	TED	IT
Oracle model	35.9	36.1	37.8	24.1	39.6
Uniform	33.1	36.4	21.9	18.4	38.9
Identity-BI	35.0	36.6	32.7	25.3	42.6
BI	35.9	36.5	38.0	26.1	**44.7**
IS	**36.0**	36.8	37.5	25.6	43.3
BI + IS	**36.0**	**36.9**	**38.4**	**26.4**	**44.7**

Table 5: Test BLEU for 2-model es-en and 3-model en-de unadapted model ensembling, compared to oracle unadapted model chosen if test domain is known. Uniform ensembling generally underperforms the oracle, while BI+IS outperforms the oracle.

Decoder configuration	es-en		en-de		
	Health	Bio	News	TED	IT
Oracle model	35.9	37.8	37.8	27.0	57.0
Uniform	36.0	36.4	**38.9**	26.0	43.5
BI + IS	**36.2**	**38.0**	38.7	**26.1**	**56.4**

Table 6: Test BLEU for 2-model es-en and 3-model en-de model ensembling for models adapted with EWC, compared to oracle model last trained on each domain, chosen if test domain is known. BI+IS outperforms uniform ensembling and in some cases outperforms the oracle.

Language pair	Model type	Oracle model	Decoder configuration	
			Uniform	BI + IS
es-en	Unadapted	36.4	34.7	36.6
	No-reg	36.6	34.8	-
	EWC	37.0	36.3	**37.2**
en-de	Unadapted	36.4	26.8	38.8
	No-reg	41.7	31.8	-
	EWC	42.1	38.6	**42.0**

Table 7: Total BLEU for test data concatenated across domains. Results from 2-model es-en and 3-model en-de ensembles, compared to oracle model chosen if test domain is known. No-reg uniform corresponds to the approach of Freitag and Al-Onaizan (2016). BI+IS performs similarly to strong oracles with no test domain labeling.

achieves similar or better in-domain results to no-reg while reducing forgetting, resulting in better uniform ensemble performance than no-reg.

BI+IS decoding with single-domain trained models achieves gains over both the naive uniform approach and over oracle single-domain models. BI+IS with EWC-adapted models gives a 0.9 / 3.4 BLEU gain over the strong uniform EWC ensemble, and a 2.4 / 10.2 overall BLEU gain over the approach described in Freitag and Al-Onaizan (2016).

3 Conclusions

We report on training and decoding techniques that adapt NMT to new domains while preserving performance on the original domain. We demonstrate that EWC effectively regularizes NMT fine-tuning, outperforming other schemes reported for NMT. We extend Bayesian Interpolation with source information and apply it to NMT decoding with unadapted and fine-tuned models, adaptively weighting ensembles to out-perform the oracle case, without relying on test domain labels. We suggest our approach, reported for domain adaptation, is broadly useful for NMT ensembling.

Acknowledgments

This work was supported by EPSRC grant EP/L027623/1 and has been performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service[2] funded by EPSRC Tier-2 capital grant EP/P020259/1. Initial work by Danielle Saunders took place during an internship at SDL Research.

References

Cyril Allauzen and Michael Riley. 2011. Bayesian Language Model Interpolation for Mobile Speech Input. In *Proceedings of the Twelfth Annual Conference of the International Speech Communication Association*.

[2] http://www.hpc.cam.ac.uk

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.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2016. The IWSLT 2016 evaluation campaign. In *IWSLT 2016, International Workshop on Spoken Language Translation*.

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

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for Neural Machine Translation. *CoRR*, abs/1612.06897.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source Neural Machine Translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Huda Khayrallah, Brian Thompson, Kevin Duh, and Philipp Koehn. 2018. Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 36–44.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521–3526.

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.

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*, pages 76–79.

Mariana L Neves, Antonio Jimeno-Yepes, and Aurélie Névéol. 2016. The ScieLO Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. In *LREC*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. *CoRR*, abs/1804.08771.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Yonatan Belinkov, and Stephan Vogel. 2017. Neural Machine Translation training in a multi-domain scenario. In *IWSLT 2017, International Workshop on Spoken Language Translation*.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1715–1725.

Felix Stahlberg, Eva Hasler, Danielle Saunders, and Bill Byrne. 2017. SGNMT–A Flexible NMT Decoding Platform for Quick Prototyping of New Models and Search Strategies. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 25–30.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson, and Philipp Koehn. 2018. Freezing subnetworks to analyze domain adaptation in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation*, pages 124–132.

Marco Turchi, Rajen Chatterjee, and Matteo Negri. 2017. WMT17 en-de APE shared task data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for Neural Machine Translation. *CoRR*, abs/1803.07416.

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 6000–6010.

David Vilar. 2018. Learning hidden unit contribution for adapting neural machine translation models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2, pages 500–505.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. 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)*, volume 2, pages 560–566.