Additive Interventions Yield Robust Multi-Domain Machine Translation Models

Elijah Rippeth[*]
Department of Computer Science
University of Maryland
`erip@cs.umd.edu`

Matt Post
Microsoft
`mattpost@microsoft.com`

Abstract

Additive interventions are a recently-proposed mechanism for controlling target-side attributes in neural machine translation. In contrast to tag-based approaches which manipulate the raw source sequence, interventions work by directly modulating the encoder representation of all tokens in the sequence. We examine the role of additive interventions in a large-scale multi-domain machine translation setting and compare its performance in various inference scenarios. We find that while the performance difference is small between intervention-based systems and tag-based systems when the domain label matches the test domain, intervention-based systems are robust to label error, making them an attractive choice under label uncertainty. Further, we find that the superiority of single-domain fine-tuning comes under question when training data size is scaled, contradicting previous findings.

1 Introduction

Multi-domain machine translation (MDMT) is the paradigm in which a single model is trained to service many domains by training on multiple corpora covering disparate labeled domains. The goal of MDMT is not only to provide high quality *general* machine translation enabled by knowledge transfer across domains, but also to enable high quality *domain-specific* machine translation when a model is provided cues about the target domain, used to control the generation. Though an intuitive task, the expectations surrounding the task were only recently formalized by Pham et al. (2021) in which the authors provided both a set of functional requirements demanded of successful MDMT models and an experimental framework under which those requirements can be tested.

Pham et al. (2021) explored several mechanisms for controlling domain, ranging from simple tag-based approaches to meta-learning based mechanisms. According to the functional requirements outlined by the authors, no method meets all the expectations demanded of effective multi-domain machine translators, though the experiments were run on a relatively small dataset of only in-domain data. The primary remaining expectations, according to the authors, are the superiority of fine-tuning based methods as compared to these methods which can control the target domain, and the ability to accommodate fuzzy or uncertain domains.

This framework is useful, but the authors leave open several other questions regarding the state of MDMT. The first of these is data size. Previous experiments focused only on relatively small, in-domain data in an otherwise high-resource setting of English-French and found that most models pale in comparison to models fine-tuned on a single domain. We wonder whether this fine-tuning superiority conclusion holds under a more realistic paradigm in which models trained on large, out-of-domain datasets are fine-tuned on in-domain data. While pretraining and fine-tuning on in-domain data can yield strong in-domain performance—as observed by the authors—this is likely to be at the cost of general domain performance, calling into question the transferability under MDMT.

Next, we wonder if new methods might help with the issue of domain control in MDMT. The authors examine reasonable mechanisms for controlling the domain which were known at the time. Since then, new methods have been developed which we hope to investigate under the prescribed framework. We hypothesize that additive interventions (Schioppa et al., 2021), which learn tag embeddings separately from the encoder, may be harder to ignore, and that the learned interventions may be able to absorb target-side properties more easily, while freeing the encoder to learn strong representations purely for translation.

In this work we scale the original experimental

---

[*] Work was done during an internship at Microsoft

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 220–232
Abu Dhabi, December 7–8, 2022. ©2022 Association for Computational Linguistics

framework presented in Pham et al. (2021) by including a significantly larger, more realistic dataset. We also experiment with additive interventions as an alternative to domain tagging. We find that:

- additive interventions perform roughly equivalently with tag-based approaches in the ideal case where provided tags match the target domain.
- additive interventions are much more robust in the face of incorrect and uncertain domain labels.
- when the experiment is scaled, models fine-tuned targeting a single domain are strong translators, but are never unmatched by other models which can service multiple domains suggesting that MDMT models in a high-resource setting are competitive with best-in-class baselines.

2 Method

As a baseline, we inject domain metadata using the tag-based approach. In this scheme, a token representing the target-side attribute, t, is prepended to source segment x and fed to the encoder E whose hidden representation is finally exposed to decoder D in a “normal" fashion:

$$\hat{y} = D(E([t] + x))$$

where $+$ indicates sequence concatenation. In tag-based approaches, the expectation is that the domain tag as a prefix acts as a conditioning variable which encourages target-side attributes to appear as desired in the final translation.

While effective and architecturally non-invasive, this method is not without downsides. Because the target token’s contribution to the encoder representation is learned, there is a chance that the attribute can be ignored. To address this and other weaknesses of tag-based approaches, Schioppa et al. (2021) present the additive interventions method which requires an encoder E, a decoder D, and a separate attribute embedding layer Emb. Given a source segment x and a sentence-level attribute token t, we have

$$V = Emb(t)$$
$$\hat{y} = D(E(x) \oplus V)$$

where $\oplus$ is defined as addition broadcasted along the token dimension. Importantly, this allows prototypically discrete attributes to be represented and controlled in a *continuous* fashion, allowing for interpolation, scaling, and positionally invariant combinations, among other useful features. We note that these are somewhat analogical to an “additive" version of “source factors" approaches (Hoang et al., 2016; Sennrich and Haddow, 2016) with one major difference: additive interventions happen *after* the encoder rather than *before* the encoder.

Source	Parallel sents (k)	Source tokens (m)
ParaCrawl	229,340	4,190.0
BANK	190	6.3
IT	270	3.6
LAW	501	17.1
TALK	160	3.6
RELIG	130	3.2
MED	2,609	133.0
NEWS	254	5.6

Table 1: Effective training set sizes

While the original work only introduces the interventions to the top-most decoder layers in order to allow for partially freezing pretrained networks, we simplify by applying the intervention to the top layer of the encoder, such that it affects all decoder layers. Further, the authors report that improved general performance can be promoted by randomly inducing a zero-vector intervention. As such, we can specify that t is randomly replaced by $\langle \text{PAD} \rangle$ with some probability with the same effect. We report 20% masking in this paper, though we experiment with 0% masking and find no significant differences between the two.

3 Experimental Setup

3.1 Data

We follow the supervised data settings prescribed by Pham et al. (2021) which includes splits from seven domains of varying disparity: BANK, IT, LAW, TALK, RELIG, MED, and NEWS. These domains are drawn from various sources: the European Central Bank corpus (BANK) (Tiedemann, 2012); the documentation for the KDE, Ubuntu, GNOME, and PHP projects from Opus (Tiedemann, 2009) combined to form IT; The JRC-Acquis corpus (LAW) (Steinberger et al., 2006); TED Talks (TALK) (Cettolo et al., 2012); the Tanzil translation of the Koran (RELIG); the UFAL Medi-

Method	BANK		IT		LAW		TALK		RELIG		MED		WMT15	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
`general base`	42.4	0.485	38.3	0.311	56.2	0.832	40.6	0.585	18.9	0.166	43.9	0.548	41.3	0.639
`combined base`	52.1	0.559	45.6	0.528	59.8	0.855	41.5	0.614	27.8	0.284	49.8	0.651	41.7	0.633
`combined ints`	51.9	**0.573**	44.7	0.512	59.9	0.859	41.3	0.610	27.6	0.268	50.1	0.647	**41.6**	**0.638**
`combined tags`	52.0	0.546	**46.5**	0.492	59.8	0.856	**43.7**	**0.647**	**28.8**	**0.307**	50.1	0.647	36.8	0.606
`in-dom ints`	58.5	0.615	51.9	0.615	66.6	0.891	39.2	0.494	88.7	0.872	55.4	0.695	**30.1**	**0.289**
`in-dom tags`	58.7	0.611	51.1	0.599	66.4	0.893	**39.8**	**0.531**	89.5	0.893	55.4	0.685	26.8	0.243
`multi-dom FT ints`	56.1	0.604	50.6	0.605	64.9	**0.896**	41.3	0.580	79.4	0.791	51.6	0.671	**34.3**	0.433
`multi-dom FT tags`	**56.9**	0.614	50.9	0.595	64.8	0.870	41.6	**0.605**	**83.6**	**0.850**	51.9	0.673	33.4	0.439
`single-dom FT`	58.2	0.637	50.8	0.629	67.0	0.917	45.1	0.653	39.0	0.402	52.6	0.679	–	–

Table 2: MT quality scores per test set. Statistically significant differences between `tags` and `ints` at the 95% confidence interval with 1000 bootstrapped samples **bolded**.

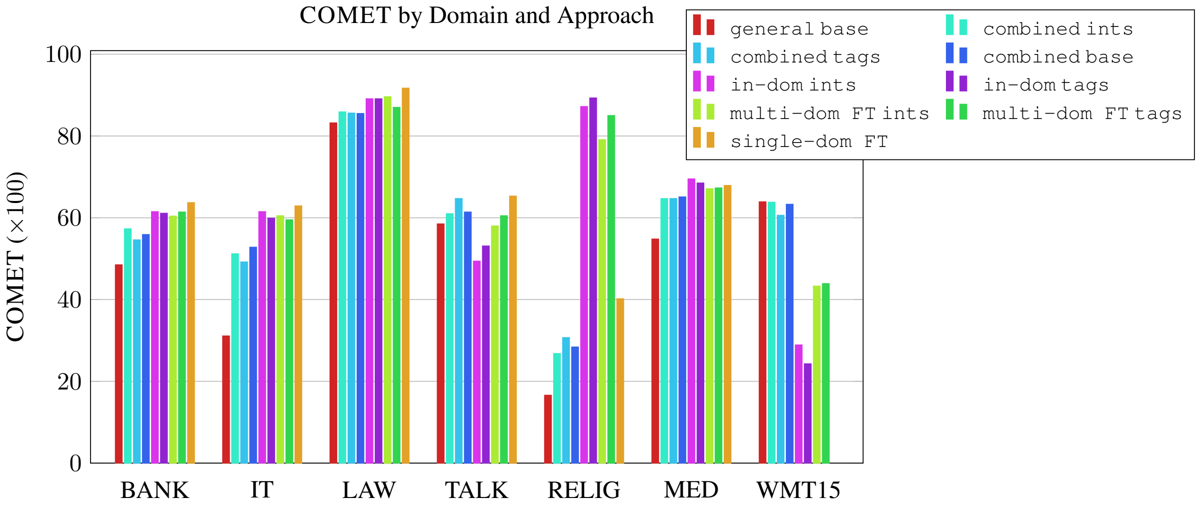

Figure 1: COMET scores (×100) by domain and approach

cal corpus v1.0 (MED)[1]; and News Commentary corpus v12 (NEWS) (Tiedemann, 2012). For sake of consistency, we rely on roughly the same splits as provided by the authors,[2], though we remove duplicates within each domain, which changes the size of each training set slightly. Additionally we include English-French ParaCrawl v9 (Bañón et al., 2020) to serve as a large out-of-domain training set for some experimental settings. The effective training set sizes are summarized in Table 1.

3.2 Models

We consider several models falling into two categories: those trained with (`control`) and without(`no control`) a method for selecting the target domain.

We use approximately the same architecture for all settings, though note that all intervention-based models have an extra embedding layer with the same embedding dimension as the encoder[3]. The basic architecture follows a 12-layer encoder, 6-layer decoder transformer with 8 attention heads each (Vaswani et al., 2017), encoder and decoder feedforward embedding dimensions of 4096, and encoder and decoder embedding dimensions of 1024.

3.2.1 `no control`

We train three models with no training-time information about the domain that the data comes from and, as a consequence, have no ability to explicitly control the target domain:

1. we have an out-of-domain baseline which is trained only on ParaCrawl: `general base`.
2. we have a model which is trained on the in-domain plus out-of-domain training sets:

---

[1] https://ufal.mff.cuni.cz/ufal_medical_corpus

[2] https://github.com/qmpham/experiments

[3] Adding $|D| \times 1024$ parameters, where D is the set of domain labels

`combined` `base`.

3. we have six quasi-oracle fine-tuned models which are produced by fine-tuning the `general base` model on each target domain's training set; we collectively refer to this set of models as single-domain fine-tuned (`single-dom FT`).

3.2.2 `control`

As mechanisms for controlling the target domain we consider:

1. prepending the domain tag to the source sequence, `tags`
2. additive interventions with 20% masking, `ints`

We apply these two methods to three settings:

1. an in-domain plus out-of-domain setting, `combined`
2. an in-domain-only setting, `in-dom`
3. a multi-domain fine-tuning setting, `multi-dom FT`, where `general` `base` is fine-tuned on all in-domain data with domain information available at training time.

This results in six models:

- `combined` `ints`
- `in-dom` `ints`
- `multi-dom FT` `ints`
- `combined` `tags`
- `in-dom` `tags`
- `multi-dom FT` `tags`.

3.3 Training

We train a joint unigram segmentation model (Kudo, 2018) using SentencePiece (Kudo and Richardson, 2018) with a vocabulary of size 32k for each setting in `general` `base`, `combined`, and `in-dom` (reusing `general base`'s model for `multi-dom FT` and `single-dom FT`). We train each model by sampling 10M sentences randomly, splitting on digits and enabling byte-fallback. We add a special token for each domain for which we have splits: ⟨BANK⟩, ⟨IT⟩, ⟨LAW⟩, ⟨TALK⟩, ⟨RELIG⟩, ⟨MED⟩, and ⟨NEWS⟩. We use these models to segment the data as appropriate in each setting.

We use dropout of 0.1 but disable attention dropout and ReLU dropout. We optimize label smoothed cross-entropy loss with a label smoothing factor of 0.1 (Szegedy et al., 2016) using Adam (Kingma and Ba, 2015). All models are built and trained using fairseq (Ott et al., 2019).

For models trained with out-of-domain data, we shard the effective dataset with each shard containing approximately 1b target tokens. For models trained with in-domain data only, we consider the entire combined in-domain dataset to be a single shard. We train for 30 virtual epochs, where a virtual epoch is defined as a single pass over one shard. For models which are fine-tuned, we fine-tune for 10 additional virtual epochs.

Each in-domain training set is assigned a unique special token which is included in the vocabulary and examples drawn from these in-domain training sets are provided the associated special token at training time. Examples from ParaCrawl are assigned no special domain token (i.e., no token is prepended in `tags` models and ⟨PAD⟩ is always provided in `ints` models).

3.4 Evaluation

We evaluate in three settings to probe various aspects of MT quality:

- we evaluate in-domain performance with each model from `control` and `no control` to determine the relative effectiveness of the methods of control against methods without control.
- we evaluate on the WMT15 English-French test set (Bojar et al., 2015) with no domain label provided (i.e., as if the models were in the `no control` setting) to test catastrophic forgetting (Goodfellow et al., 2013) in a general setting. Importantly, while the models trained on in-domain data have been exposed to newswire data, the labels are not provided at test time in this setting.
- we evaluate the effect of providing the incorrect tag to each test set, as computed by SacreBLEU (Post, 2018) and COMET (Rei et al., 2020), to test the resilience of models to label errors

4 Results

No clear winner in ideal case We evaluate the setting in which the provided domain label matches

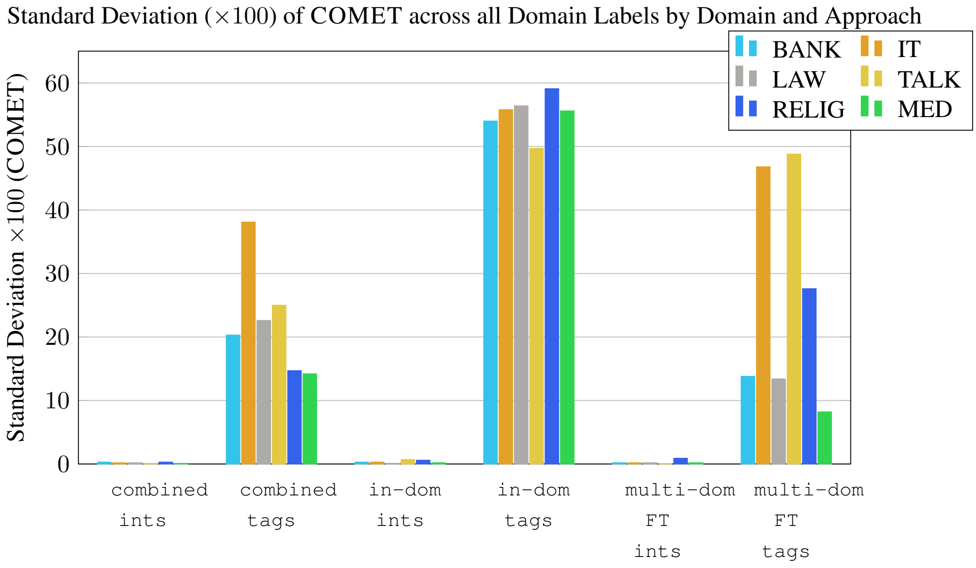

Figure 2: Impact of domain label error on COMET per test set and approach

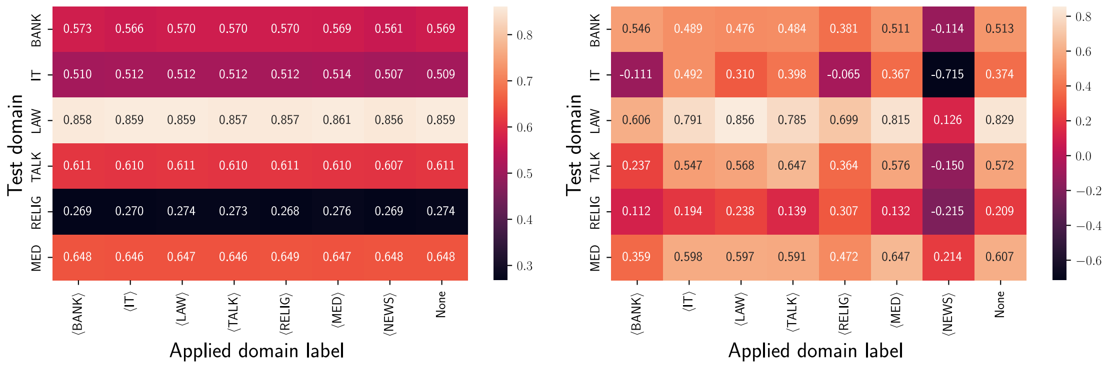

Figure 3: COMET of `combined` models under various domain labels. `ints` left, `tags` right. `ints` maintain high quality translations under mismatching domain labels in all cases, unlike `tags`.

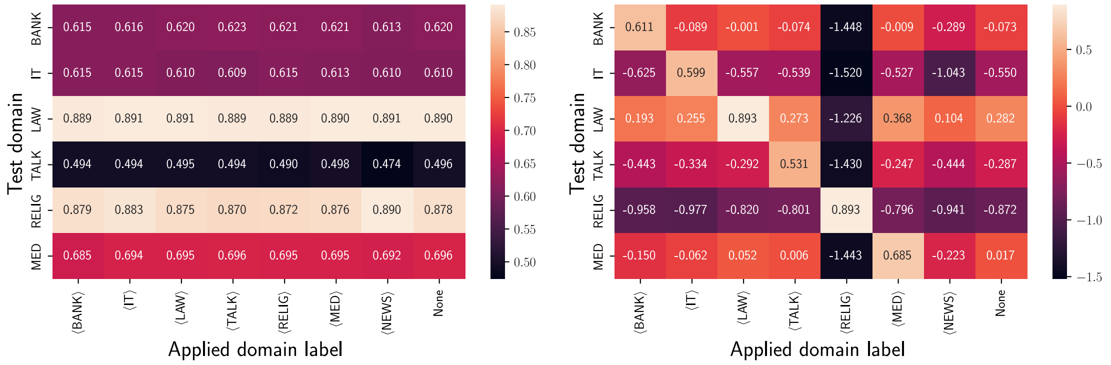

Figure 4: COMET of `in-dom` models under various domain labels. `ints` left, `tags` right. `ints` maintain high quality translations under mismatching domain labels in all cases, unlike `tags`.

the target test domain, and the setting of WMT15 without a provided domain label, for each setting apart from `single-dom FT`. The results can be read in Table 2 and are visualized in Figure 1.

Table 2 shows that when comparing `control` models within a training setting using bootstrap resampling (sample sizes of 1000) (Koehn, 2004), the difference in performance of `tags` and `ints` are insignificant in the majority of cases. While there are a few cases of statistically significant differ-

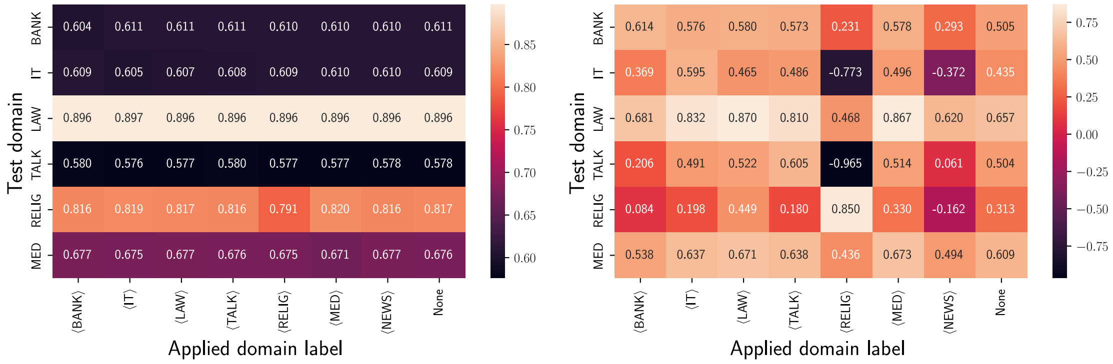

Figure 5: COMET of `multi-dom FT` models under various domain labels. `ints` left, `tags` right. `ints` maintain high quality translations under mismatching domain labels in all cases, unlike `tags`.

ences, neither `tags` nor `ints` are uniformly preferred in these cases. The opposite is observed on the out-of-domain WMT15, where `ints` performs uniformly better than `tags`, often significantly.

We observe that methods with `control` in the `combined` setting perform approximately equally to the `combined base`, showing that naive combination of in-domain and out-of-domain with a mechanism to control the domain does not improve over approaches without control, though `in-dom` and `multi-dom FT` models tend to perform better on average than any model in the `combined` setting.

`ints` are robust under domain label mismatch Next, we perform an ablation study in which we score each test across all domain label assignments (including the correct label and no label), which allows us to observe the effects of test-time labeling error. While we compute both BLEU and COMET, we include only COMET here.[4] We include the full results in Tables 3–8, but summarize the findings in Figures 2-5, which show the robustness of various models and settings to mislabeled domains.

Figures 3–5 show heatmaps resulting from this ablation, but we refer interested readers to Tables 3–8 for the long-form charts. We see that `tags` systems' performances vary dramatically, incurring severe degradation in the face of domain label error but performing strongest along the diagonal. `ints` systems, on the other hand, see only small performance changes when provided with incorrect domain labels and roughly equal performance under all possible labels, as observed in Figure 2. We see that `in-dom tags` have the highest average variation in performance, likely owing to the small amount of data which suggests that `in-dom tags` overfits to the training data. The variation in performance of `ints` systems approaches that of the `general base`, which by definition ignores the domain label and therefore has 0 variance; however, `ints` has demonstrably stronger performance than `general base` in all domains and, indeed, stronger performance than `tags` in a handful of domains and thus seems to learn strong general representations for translation which disentangles the representations of the encoder from the representations of the attribute.

Additionally, through manual analysis we find that `tags` systems are more prone to hallucinating translation artifacts from the corpus associated with the domain label being used, often causing quality degradation. We refer to Table 15 for an example of such artifacts, which includes topical and target language mismatches along with tokens which appear as a result of the HTML-encoded nature of the ⟨IT⟩ dataset.[5]

Single-domain fine-tuning is not as competitive in large-data settings We compare the performance of models trained only with in-domain data and out-of-domain data. From Table 2, we see slightly stronger in-domain performance for `in-dom` models as compared to models fine-tuned with out-of-domain data at the cost of out-of-domain performance on WMT15, suggesting that `multi-dom FT` models generalize better and may surpass `in-dom` models with more training due to the relatively little fine-tuning budget of 10 epochs afforded to them comparatively.

---

[4] Similar results for BLEU are listed in Appendix A.2

[5] Escaping seems to be an artifact of Moses preprocessing leakage of raw data; not germane to all domains in this work.

Finally, we see that while `single-dom FT` is typically among the highest performing systems for a given test set, it is never unmatched by an alternative system in `control`. We observe that `single-dom FT` is uniformly stronger than `general base` and `combined`, `in-dom` and `multi-dom FT` show competitive in-domain performance. We note that because there is one `single-dom FT` model per test set, the effective parameter budget is six times larger than any of the individual models, providing support for both its impracticality and untenability as compared to any other setting. This suggests that single-domain fine-tuning is not as effective as expected in high-resource settings as a strong upper-bound in MDMT.

5 Related Work

Incorporating extra-sentential information has a rich history in NMT. Aside from controlling for the domain, Sennrich et al. (2016) use a politeness tag at training and inference time to accommodate coarse politeness control in machine translation. Additionally, Kuczmarski and Johnson (2018) use tags to afford users the ability to vary binary gender in the translations of gender-neutral inputs, hoping to address gender bias in MT.

At the sub-sequence level, Hoang et al. (2016) and Sennrich and Haddow (2016) included linguistically-informed word-level "source factors", such as part-of-speech tags and dependency relations, as additional feature factors to be concatenated to form a full encoder representation with the goal of reducing ambiguity and sparseness issues.

Perhaps more relatedly, several works have explored the impacts of incorporating domain information into training using various methods. Kobus et al. (2017) explore two methods: a tag-based approach which concatenates a special token to the end of the source sequence, and a "source factors"-style approach which concatenates domain-level embeddings to each token embedding in the source. Sharaf et al. (2020) explore few-shot domain adaptation, rather than domain control, through the lens of meta-learning and show that a meta-learning based approach is generally stronger than other adaptation approaches, though we note that adaptation and control address different needs. Finally, Stojanovski and Fraser (2021) frame machine translation with document-context as an unsupervised domain adaptation problem and incorporate domain embeddings within the encoder, summed with positional and word embeddings, yielding strong improvements over competitive baseline models.

6 Conclusion

In this work we examined the relative impact of additive interventions in a large-scale MDMT setting. We find that typically there are no significant differences between additive interventions and tag-based approaches when the provided domain label matches the test set, but find that additive interventions exhibit *much more desirable degradation properties* when the domain label is unknown or incorrectly provided. In addition, we find that models first trained on a large, general corpus and then fine-tuned on a single-domain—a realistic baseline in machine translation—rarely perform significantly better than approaches which are trained or fine-tuned only on in-domain data, which is in contrast to their generally superior performance in low-resource settings.

In future work we consider developing extensions to additive interventions which can further improve their performance in MDMT settings. Additionally, studying additive interventions in other tasks where tag-based approaches are dominant, such as multi-lingual machine translation, could be an interesting avenue for exploration.

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual* *conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks.

Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. 2016. Improving neural translation models with linguistic factors. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 7–14, Melbourne, Australia.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

James Kuczmarski and Melvin Johnson. 2018. Gender-aware natural language translation. *Technical Disclosure Commons, (October 08, 2018)*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

MinhQuang Pham, Josep Maria Crego, and François Yvon. 2021. Revisiting multi-domain machine translation. *Transactions of the Association for Computational Linguistics*, 9:17–35.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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. Association for Computational Linguistics.

Amr Sharaf, Hany Hassan, and Hal Daumé III. 2020. Meta-learning for few-shot NMT adaptation. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 43–53, Online. Association for Computational Linguistics.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Dario Stojanovski and Alexander Fraser. 2021. Addressing zero-resource domains using document-level context in neural machine translation. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 80–93, Kyiv, Ukraine. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Jörg Tiedemann. 2009. News from opus — a collection of multilingual parallel corpora with tools and interfaces. *Advances in Natural Language Processing*, pages 237–248.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A Raw scores

A.1 Ablation (COMET)

Provided label Test set	⟨BANK⟩	⟨IT⟩	⟨LAW⟩	⟨TALK⟩	⟨RELIG⟩	⟨MED⟩	⟨NEWS⟩	None
BANK	0.573	0.566	0.570	0.570	0.570	0.569	0.561	0.569
IT	0.510	0.512	0.512	0.512	0.512	0.514	0.507	0.509
LAW	0.858	0.859	0.859	0.857	0.857	0.861	0.856	0.859
TALK	0.611	0.610	0.611	0.610	0.611	0.610	0.607	0.611
RELIG	0.269	0.270	0.274	0.273	0.268	0.276	0.269	0.274
MED	0.648	0.646	0.647	0.646	0.649	0.647	0.648	0.648

Table 3: COMET scores of `combined ints` under various domain labels

Provided label Test set	⟨BANK⟩	⟨IT⟩	⟨LAW⟩	⟨TALK⟩	⟨RELIG⟩	⟨MED⟩	⟨NEWS⟩	None
BANK	0.546	0.489	0.476	0.484	0.381	0.511	-0.114	0.513
IT	-0.111	0.492	0.310	0.398	-0.065	0.367	-0.715	0.374
LAW	0.606	0.791	0.856	0.785	0.699	0.815	0.126	0.829
TALK	0.237	0.547	0.568	0.647	0.364	0.576	-0.150	0.572
RELIG	0.112	0.194	0.238	0.139	0.307	0.132	-0.215	0.209
MED	0.359	0.598	0.597	0.591	0.472	0.647	0.214	0.607

Table 4: COMET scores of `combined tags` under various domain labels

Provided label Test set	⟨BANK⟩	⟨IT⟩	⟨LAW⟩	⟨TALK⟩	⟨RELIG⟩	⟨MED⟩	⟨NEWS⟩	None
BANK	0.615	0.616	0.620	0.623	0.621	0.621	0.613	0.620
IT	0.615	0.615	0.610	0.609	0.615	0.613	0.610	0.610
LAW	0.889	0.891	0.891	0.889	0.889	0.890	0.891	0.890
TALK	0.494	0.494	0.495	0.494	0.490	0.498	0.474	0.496
RELIG	0.879	0.883	0.875	0.870	0.872	0.876	0.890	0.878
MED	0.685	0.694	0.695	0.696	0.695	0.695	0.692	0.696

Table 5: COMET scores of `in-dom ints` under various domain labels

Provided label Test set	⟨BANK⟩	⟨IT⟩	⟨LAW⟩	⟨TALK⟩	⟨RELIG⟩	⟨MED⟩	⟨NEWS⟩	None
BANK	0.611	-0.089	-0.001	-0.074	-1.448	-0.009	-0.289	-0.073
IT	-0.625	0.599	-0.557	-0.539	-1.520	-0.527	-1.043	-0.550
LAW	0.193	0.255	0.893	0.273	-1.226	0.368	0.104	0.282
TALK	-0.443	-0.334	-0.292	0.531	-1.430	-0.247	-0.444	-0.287
RELIG	-0.958	-0.977	-0.820	-0.801	0.893	-0.796	-0.941	-0.872
MED	-0.150	-0.062	0.052	0.006	-1.443	0.685	-0.223	0.017

Table 6: COMET scores of `in-dom tags` under various domain labels

Provided label Test set	⟨BANK⟩	⟨IT⟩	⟨LAW⟩	⟨TALK⟩	⟨RELIG⟩	⟨MED⟩	⟨NEWS⟩	None
BANK	0.604	0.611	0.611	0.611	0.610	0.610	0.610	0.611
IT	0.609	0.605	0.607	0.608	0.609	0.610	0.610	0.609
LAW	0.896	0.897	0.896	0.896	0.896	0.896	0.896	0.896
TALK	0.580	0.576	0.577	0.580	0.577	0.577	0.578	0.578
RELIG	0.816	0.819	0.817	0.816	0.791	0.820	0.816	0.817
MED	0.677	0.675	0.677	0.676	0.675	0.671	0.677	0.676

Table 7: COMET scores of `multi-dom FT ints` under various domain labels

Provided label Test set	⟨BANK⟩	⟨IT⟩	⟨LAW⟩	⟨TALK⟩	⟨RELIG⟩	⟨MED⟩	⟨NEWS⟩	None
BANK	0.614	0.576	0.580	0.573	0.231	0.578	0.293	0.505
IT	0.369	0.595	0.465	0.486	-0.773	0.496	-0.372	0.435
LAW	0.681	0.832	0.870	0.810	0.468	0.867	0.620	0.657
TALK	0.206	0.491	0.522	0.605	-0.965	0.514	0.061	0.504
RELIG	0.084	0.198	0.449	0.180	0.850	0.330	-0.162	0.313
MED	0.538	0.637	0.671	0.638	0.436	0.673	0.494	0.609

Table 8: COMET scores of `multi-dom FT tags` under various domain labels

A.2 Ablation (BLEU)

All scores reported are from SacreBLEU[6] (Post, 2018).

Provided label Test set	⟨BANK⟩	⟨IT⟩	⟨LAW⟩	⟨TALK⟩	⟨RELIG⟩	⟨MED⟩	⟨NEWS⟩	None
BANK	51.9	51.7	51.9	51.9	51.9	51.8	51.8	51.9
IT	44.6	44.7	44.8	44.8	44.6	44.7	44.7	44.6
LAW	59.8	59.8	59.9	59.8	59.7	59.8	59.7	59.9
TALK	41.3	41.3	41.4	41.3	41.4	41.3	41.1	41.5
RELIG	27.6	27.8	27.7	27.8	27.6	27.9	27.5	27.7
MED	50.0	50.0	50.0	50.0	49.9	50.1	50.0	50.0

Table 9: BLEU scores of `combined ints` under various domain labels

Provided label Test set	⟨BANK⟩	⟨IT⟩	⟨LAW⟩	⟨TALK⟩	⟨RELIG⟩	⟨MED⟩	⟨NEWS⟩	None
BANK	52.0	43.5	43.0	40.4	39.0	45.2	30.2	44.2
IT	18.5	46.5	36.3	39.9	26.5	37.2	11.0	35.0
LAW	50.2	56.4	59.8	50.7	51.4	55.5	36.9	56.2
TALK	29.5	39.2	38.1	43.7	28.3	39.7	22.7	37.1
RELIG	21.6	24.4	25.5	16.3	28.8	18.9	14.5	22.6
MED	43.5	48.5	48.3	47.3	45.0	50.1	41.6	49.1

Table 10: BLEU scores of `combined tags` under various domain labels

Provided label Test set	⟨BANK⟩	⟨IT⟩	⟨LAW⟩	⟨TALK⟩	⟨RELIG⟩	⟨MED⟩	⟨NEWS⟩	None
BANK	58.5	58.6	58.6	58.6	58.5	58.8	58.2	58.7
IT	52.0	51.9	51.4	51.4	51.8	51.6	51.4	51.8
LAW	66.1	66.2	66.1	66.0	65.9	66.1	66.0	66.1
TALK	39.0	39.1	39.1	39.2	39.1	39.2	38.8	39.0
RELIG	89.2	89.2	89.0	88.7	88.7	89.2	89.3	89.1
MED	55.4	55.5	55.3	55.4	55.4	55.4	55.4	55.5

Table 11: BLEU scores of `in-dom ints` under various domain labels

Provided label Test set	⟨BANK⟩	⟨IT⟩	⟨LAW⟩	⟨TALK⟩	⟨RELIG⟩	⟨MED⟩	⟨NEWS⟩	None
BANK	58.7	31.2	36.0	34.3	3.9	36.1	27.3	34.4
IT	15.5	51.1	16.6	20.0	0.4	18.8	5.9	15.9
LAW	42.2	43.5	66.4	45.3	12.4	48.2	40.2	44.7
TALK	18.6	21.0	20.7	39.8	1.0	23.8	17.2	21.5
RELIG	6.2	6.1	8.2	8.7	89.5	9.0	5.5	7.6
MED	32.2	33.2	32.8	33.3	5.5	55.4	29.5	33.5

Table 12: BLEU scores of `in-dom tags` under various domain labels

---

[6] `BLEU|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.0`

Provided label Test set	⟨BANK⟩	⟨IT⟩	⟨LAW⟩	⟨TALK⟩	⟨RELIG⟩	⟨MED⟩	⟨NEWS⟩	None
BANK	56.1	55.9	56.5	56	56.1	56.4	55.3	56.3
IT	50.6	50.6	50.0	50.4	50.3	50.6	49.8	50.9
LAW	64.8	64.7	64.9	64.9	64.8	65.2	64.5	65.0
TALK	41.2	40.8	41.1	41.3	41.3	41.2	40.4	41.5
RELIG	80.4	81.1	80.5	80.2	79.4	81.8	79.3	82.2
MED	51.7	51.3	51.6	51.7	51.7	51.6	51.3	51.7

Table 13: BLEU scores of `multi-dom FT ints` under various domain labels

Provided label Test set	⟨BANK⟩	⟨IT⟩	⟨LAW⟩	⟨TALK⟩	⟨RELIG⟩	⟨MED⟩	⟨NEWS⟩	None
BANK	56.9	54.5	54.4	52.0	49.6	55.0	43.4	54.9
IT	43.1	50.9	47.4	46.9	28.0	46.9	17.3	40.8
LAW	55.7	63.7	64.8	61.2	59.4	64.2	55.9	60.3
TALK	28.0	37.4	36.1	41.6	8.4	36.1	23.1	36.2
RELIG	32.6	38.6	61.9	22.9	83.6	50.7	19.2	49.1
MED	49.6	51.4	51.8	50.4	49.4	51.9	49.7	51.2

Table 14: BLEU scores of `multi-dom FT tags` under various domain labels

B Figures (BLEU)

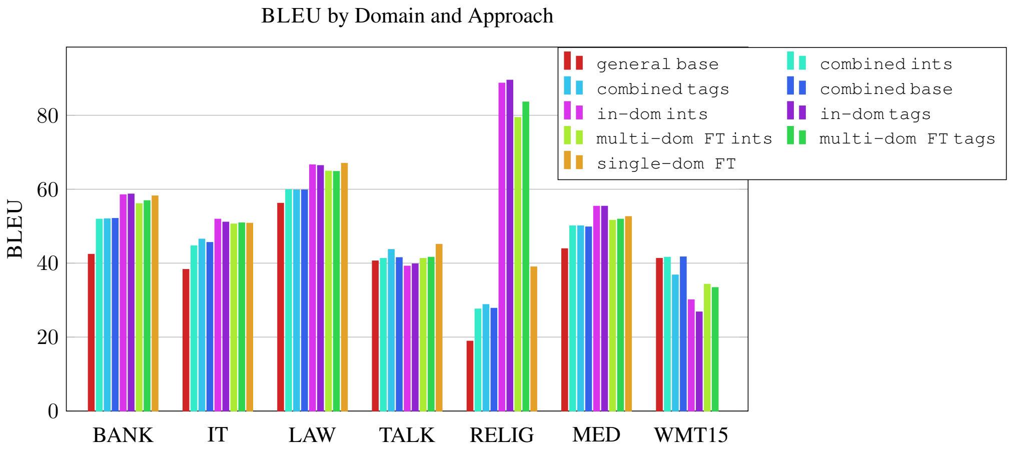

Figure 6: BLEU scores by domain and approach

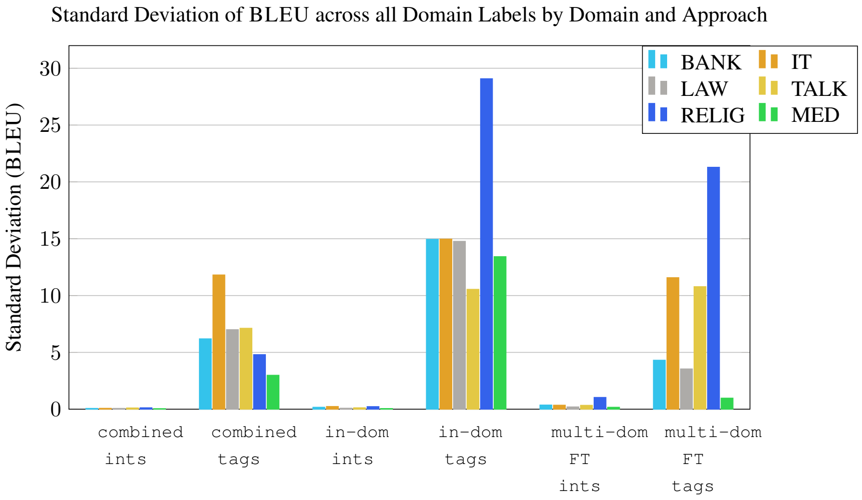

Figure 7: Impact of domain label error on BLEU per test set and approach

C Examples

Src	Never; soon they will deny ever worshipping them, and will turn into their opponents.
Ref	Bien au contraire! [ces divinités] renieront leur adoration et seront pour eux des adversaires.
`multi-dom FT ints`	Bien au contraire! [ces divinités] renieront leur adoration et seront pour eux des adversaires.
`multi-dom FT tags`	You are about to translate the 'None 'COMMAND, there are some rules on how to translate it. Please see http: / / / /www.mysql.com /.

Src	And the evil-doers say: Ye are but following a man bewitched.
Ref	Les injustes disent: «Vous ne suivez qu'un homme ensorcelé».
`in-dom ints`	Les injustes disent: «Vous ne suivez qu'un homme ensorcelé».
`in-dom tags`	Et les « & #160; diaboliques & #160; » disent & #160;: « & #160; fired & #160; » est le suivant d'un homme.

Table 15: Example translation artifacts from incorrect domain label; a translation of ⟨RELIG⟩ sentences with ⟨IT⟩ domain label under different models. We note that the HTML-encoded artifact "& #160;" appears with high frequency in ⟨IT⟩.