From Priest to Doctor:
Domain Adaptation for Low-Resource Neural Machine Translation

Ali Marashian,[1] Enora Rice,[1] Luke Gessler,[2]
Alexis Palmer,[1] Katharina von der Wense[1,3]
[1]University of Colorado Boulder, [2]Indiana University Bloomington,
[3]Johannes Gutenberg University Mainz
ali.marashian@colorado.edu

Abstract

Many of the world's languages have insufficient data to train high-performing general neural machine translation (NMT) models, let alone domain-specific models, and often the only available parallel data are small amounts of religious texts. Hence, domain adaptation (DA) is a crucial issue faced by contemporary NMT and has, so far, been underexplored for low-resource languages. In this paper, we evaluate a set of methods from both low-resource NMT and DA in a realistic setting, in which we aim to translate between a high-resource and a low-resource language with access to only: a) parallel Bible data, b) a bilingual dictionary, and c) a monolingual target-domain corpus in the high-resource language. Our results show that the effectiveness of the tested methods varies, with the simplest one, DALI, being most effective. We follow up with a small human evaluation of DALI, which shows that there is still a need for more careful investigation of how to accomplish DA for low-resource NMT.

1 Introduction

Neural machine translation (NMT) models have limited ability to deal with languages that lack large-scale monolingual and parallel corpora (Wang et al., 2021). Moreover, NMT systems face challenges when translating text from novel domains characterized by unique style or vocabulary (Koehn and Knowles, 2017; Saunders, 2022). Often, these issues co-occur, a scenario that has been neglected by researchers so far. Most of the world's 7000+ languages are considered low-resource (Joshi et al., 2020), and existing data for them are in limited domains; the languages that could most benefit from domain adaptation (DA) are the ones left behind.

In this paper, we explore a realistic setting in which we aim to translate between a high-resource and a low-resource language and are restricted to the following commonly available resources: a)

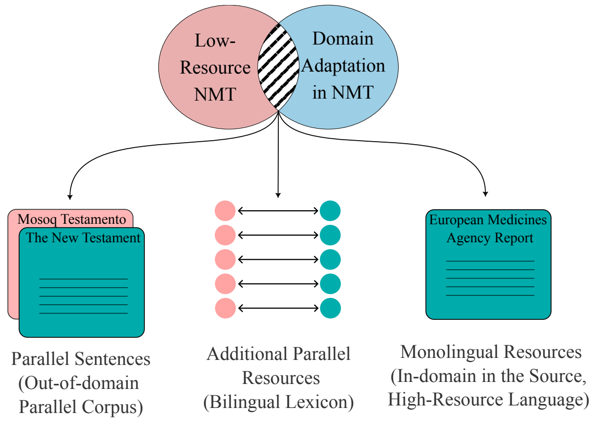

Figure 1: In our work, which looks at the (previously neglected) intersection of low-Resource NMT and domain adaptation in NMT, we consider only these commonly accessible resources.

Bible translations, i.e., a small parallel corpus in the source domain; b) monolingual target-domain texts in the high-resource language; and c) a bilingual dictionary for the two languages. To keep the setting generalizable, we assume neither access to a model pretrained on text in the low-resource language nor access to data in a related high-resource language, as for many truly low-resource languages, those are impossible to find.

We experiment with a set of four DA and low-resource NMT methods and aim to translate from English to a target language, simulating a low-resource setting. We use mBART (Liu et al., 2020) which has been fine-tuned on parallel Bible texts as our base model, and our goal is to adapt it to the target domains of government documents and medicine. The methods we investigate use the bilingual dictionaries in various ways.

Our experiments showcase the varying effectiveness of existing methods: the weakest approach results in models that perform *worse* than the base model, while the best approach – which, surprisingly, is also the simplest – results in a ChrF score

Proceedings of the 31st International Conference on Computational Linguistics, pages 7087–7098
January 19–24, 2025. ©2025 Association for Computational Linguistics

more than twice as high as the base model's. However, as the best model only reaches a ChrF score of 42.47 and a BLEU score of 13.47 (on average), we also perform a small human evaluation, which confirms that there is still a need for the development of better DA methods for low-resource NMT. Our code is available on GitHub.[1]

2 Related Work

Domain Adaptation in NMT As domains are defined by the characteristics of data (Saunders, 2022), many effective DA approaches focus on the data and, thus, can be applied to various underlying architectures. Some works focus on acquiring monolingual in-domain data, which is easier to find than in-domain parallel data. Back-translation uses monolingual target-domain data in the target language and produces artificial source sentences using a target-to-source NMT model (Poncelas et al., 2019; Jin et al., 2020). Chinea-Ríos et al. (2017) use monolingual source-side corpora and a source-to-target NMT model for forward-translation, where it is common to employ self-learning. With access to a small parallel corpus, extra training data can be created by introducing noise (Vaibhav et al., 2019). Synthetic parallel data can be acquired from an external source or generated using a predefined or induced lexicon. Hu et al. (2019) use a lexicon to back-translate target-side sentences. Peng et al. (2020) use a dictionary, injecting dictionary terms into out-of-domain texts to synthesize in-domain training data. Bergmanis and Pinnis (2021) augment the training data by annotating randomly selected source language words with their target language lemmas to integrate terms. Zhang et al. (2022) introduce lexical constraints into iterative back-translation.

Other approaches add parameters to the model, e.g., domain tags (Kobus et al., 2017; Stergiadis et al., 2021). Such a manipulation of the embeddings could extend to more terms in the vocabulary, beyond the tags (Pham et al., 2019; Sato et al., 2020; Man et al., 2023). With adapter-based methods, a domain-specific module is trained (Bapna and Firat, 2019). Chen et al. (2021) use a pointer-generator to copy suggestions from the input, which come from a domain-specific dictionary.

Low-Resource NMT Methods for low-resource MT show some overlap with DA methods. One popular approach is data augmentation, which can be in the form of word or phrase replacement with the help of a bilingual lexicon (Nag et al., 2020). Back-translation, forward-translation, and data selection methods can also be applied (Sennrich et al., 2016; Fadaee and Monz, 2018; Dou et al., 2020). Transfer learning is a useful technique in low-resource NMT (Maimaiti et al., 2019; Kocmi and Bojar, 2020; Cooper Stickland et al., 2021). Liu et al. (2021) continue to pretrain mBART (Liu et al., 2020) on unseen languages, utilizing a bilingual dictionary. Although we do not inspect large language models (LLMs) in our experiments, some recent works explore the potential of LLMs for low-resource NMT. Robinson et al. (2023) observe that ChatGPT's MT capabilities across the 204 languages of the FLORES-200 dataset (Costa-jussà et al., 2022) consistently lag behind traditional NMT models. Ghazvininejad et al. (2023) use dictionaries to suggest words to use in the output translation. Zhang et al. (2024) adopt different strategies for dictionary term lookup and the retrieval of examples for in-context learning. Siddhant et al. (2022); Ranathunga et al. (2023) note that, in the case of many low-resource languages, the problem is more severe since the only available parallel data are religious texts.

3 Data

Parallel Source-Domain Data In all experiments, the only *parallel* data we use for training come from the JHU Bible Corpus (McCarthy et al., 2020).

Target-Domain Data We explore adapting to two different domains, one at a time: government documents and medicine. The domain-specific data mostly come from past WMT translation tasks (Barrault et al., 2020; Akhbardeh et al., 2021; Kocmi et al., 2022, 2023). As we assume only *monolingual* in-domain training data (cf. Section 4), training and pretraining use only source-side sentences from these parallel data sets. Data availability varies across language/domain pairs, and we cap data set sizes to maintain comparability across languages. For **training** we use no more than 200K sentence pairs. If our setting requires **pretraining**, we use the same source-side sentences used for training. For **testing**, we use 1500 sentence pairs.

More details about the data used for each domain and language pair can be found in Appendix A.1.

[1] https://github.com/alimrsn79/da_lr_nmt

Dictionaries The methods we investigate here call for source–target language dictionaries. To build dictionaries, for each language pair we extract the 5000 most frequent lemmas and their inflections from the monolingual training data and use the Google Translate API[2] to translate those words.[3]

We augment this dictionary with word pairs extracted from our small parallel corpora, using standard statistical approaches for lexicon induction. Specifically, we employ Fast Align (Dyer et al., 2013) on the Bible verses.The expansion of the dictionary with statistical methods follows previous work (Hu et al., 2019; Zhang et al., 2024).

Further information about the dictionaries is available in Appendix A.2.

Languages Because it is difficult to source domain-specific evaluation data in truly low-resource languages, we simulate a low-resource setting, selecting languages not seen during mBART's pretraining. For the government domain, we experiment on **Croatian**, **Icelandic**, **Maltese**, **Polish**, and **Ukrainian**. For the medical domain, we use **Croatian**, **Icelandic**, **Maltese**, and **Polish**. In all cases, **English** is our high-resource language.

4 Experimental Setup

Our goal is to translate from English into our low-resource languages, one at a time. In this section, we describe the different approaches we investigate. All of them use mBART as the backbone model and are implemented using `fairseq`.[4]

mBART Baseline Our baseline is the pretrained mBART model, which has been trained on 25 languages and is said to generalize well to unseen languages (Liu et al., 2020).

DALI We adapt the method from Hu et al. (2019), who extract a lexicon by mapping word embeddings from the source to the target language. They then use this lexicon to back-translate from the target monolingual data, by word-for-word replacement. The resulting texts are the pseudo-parallel data that are used for training. We produce pseudo-parallel data using the same method, but use the dictionary described in Section 3. As we have access to monolingual texts in the *source* language, we do forward-translation instead of back-translation.

LeCA Chen et al. (2021) append suggestions to the input to be used in the output. Their model uses a pointer-generator module to potentially copy from the input. Since the model updates just the probability of the next token by also considering copying from the input tokens, it is not a hard constraint. We match their DICTIONARY CONSTRAINT setting, where suggestions are made by looking up source-side terms in a given dictionary. We implement this on top of the base mBART model. Note that LeCA was not originally proposed for low-resource scenarios, and they do not use a pretrained model, instead training the base Transformer model from scratch.

CPT Liu et al. (2021) continue pretraining mBART on mixed-language text, modifying the pretraining scheme of the model. They corrupt the text by replacing some terms with their translation in the new language, and the model is trained to reconstruct the original text. In our setting, we must use source-side monolingual text only, matching their CPT W/ MLT (SRC) method. Note that in our experiments we translate from the high-resource language to the low-resource.

Combined We experiment with merging the above methods: first, we pretrain the model (CPT) and then train it with pseudo-parallel data (DALI) while using pointer-generators (LeCA).

Metrics We evaluate all methods on the test data decribed in Section 3, using BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) as implemented by sacreBLEU (Post, 2018). We consider ChrF our main metric, as it focuses on characters and is more informative when translating into morphologically rich languages.

5 Results and Discussion

The results for all languages and domains appear in Table 1. On average, DALI performs best in the majority of the experiments. It is also the simplest of the methods to implement, as it is model-agnostic and only the training data is manipulated.

LeCA does not help in most cases, supporting Bafna et al. (2024), who observe that pointer-generators are not consistently helpful for low-resource NMT. LeCA was not initially devised for low-resource settings, and also the dictionary here includes just one translation per term, with no guarantee of matching the intended target side meaning. Since mBART was not pretrained on these

[2] https://cloud.google.com/translate/

[3] We expect performance might increase if we had domain-specific bilingual dictionaries for each language pair.

[4] https://github.com/facebookresearch/fairseq

	Metric	Croatian Gov.	Croatian Med.	Icelandic Gov.	Icelandic Med.	Maltese Gov.	Maltese Med.	Polish Gov.	Polish Med.	Ukrainian Gov.	Average Gov.	Average Med.
mBART	BLEU	0.69	1.7	0.76	1.46	1.57	1.68	0.34	0.33	0.9	0.85	1.29
	ChrF	17.34	18.62	18.97	17.72	21.61	19.42	19.11	17.37	17.83	18.97	18.28
DALI	BLEU	4.1	12.74	5.76	13.89	7.92	16.68	4.21	10.57	6.8	5.76	13.47
	ChrF	38.87	**43.32**	36.02	**41.07**	**49.55**	48.77	**36.33**	**36.73**	**37.51**	**39.66**	**42.47**
LeCA	BLEU	0.65	1.68	0.98	0.24	1.41	1.5	0.35	0.41	0.79	0.84	0.96
	ChrF	17.48	18.23	19.24	15.97	20.6	18.56	17.6	17.11	18.74	18.73	17.47
CPT	BLEU	2.62	8.02	3.66	5.26	2.18	5.38	1.57	5.73	4.38	2.88	6.1
	ChrF	20.46	25.19	20.67	20.56	20.42	21.86	19.19	21.03	12.35	18.62	22.16
Combined	BLEU	3.87	12.21	5.63	13.4	7.14	16.75	3.82	10.67	6.69	5.43	13.26
	ChrF	**39.93**	42.11	**36.33**	40.56	48.17	**48.88**	35.72	36.11	36.46	39.32	41.92

Table 1: Performance on the all the test sets for the target domains government (Gov.) and medical (Med.) documents. Best BLEU score per column is underlined, while the best ChrF score is indicated in bold.

languages, its embeddings of words in the low-resource language might not as directly correspond to their source-side, high-resource counterparts; we hypothesize this may be another reason LeCA performs poorly for resource-constrained scenarios.

CPT is helpful in most of the experiments when compared to plain mBART, but not compared to DALI. After pretraining, the model is fine-tuned only on Bible data. In the pretraining, we reconstruct the source side, so the model only learns to output in the target language from the Bible verses. Pretraining helps the model get more familiar with the domain and establish connections between the embeddings of target words and their respective translations in the high-resource language.[5]

Combining the methods together shows some improvements over other individual methods, but generally fails to reach DALI's performance. Note that the same dataset was used to both pretrain the model (the CPT part) and to then make pseudo-parallel data (the DALI part). Since LeCA is not helpful when added to the basic mBART, we also test performance of *Combined* without LeCA, on the medical domain. The results (Table 8) indicate that – when using the same dataset for both – adding pretraining on top of DALI can be detrimental, but removing LeCA increases performance on all languages for the medical domain.

LeCA only uses the dictionary in the final stage, and CPT uses the monolingual data during pretraining, before being fine-tuned on the bible data. DALI and *Combined* are the only methods that have access to source-side target-domain monolingual data during the final stage of training, which could partially explain their superior performance. That a simple method like DALI – that mostly keeps the word order of the sentence language – should be the best performing method hints at the extensive room for growth in future work.

Figure 2 shows averaged sentence-level BLEU and ChrF scores plotted against their respective reference token lengths for DALI models. For length l, we average the scores of the models for different languages if the reference translation is of length l. We can see that generally the scores seem to get higher with longer sentences, especially for ChrF.

Example We see some interesting trends in the outputs. Table 2 showcases an example with the outputs of different methods for one sentence from the Maltese-medical test set, the language–domain pair with the most significant performance boost. *Warning: these outputs could include distressing language against women that may harm some readers.* Both mBART and LeCA translate in a religious tone. The same is true for CPT, which also tends to copy words from the input – as it was a part of the reconstruction procedure during pretraining. It is important to emphasize that Maltese is a morphologically rich language, and the inflections are mostly discarded in the outputs of DALI and *Combined*; for example the words are more likely to be disjoint in their outputs than they are in the target (the first "*il oħra*" vs "*l-oħra*" in the target), or they can be in different forms ("*huwa*" vs "*hija*"). Note that *ointment* was translated to *infusion* by DALI. Given the sensitivity of the domain, a translation like this can potentially be harmful.

Maltese at times has a different word order than that of English ("*il-mediċina tal-għajnejn*" is translated as "*għajn mediċina*", which matches the order

[5]According to Liu et al. (2021), the performance boost is expected to increase if we have monolingual texts in the target language instead and can use them during pretraining.

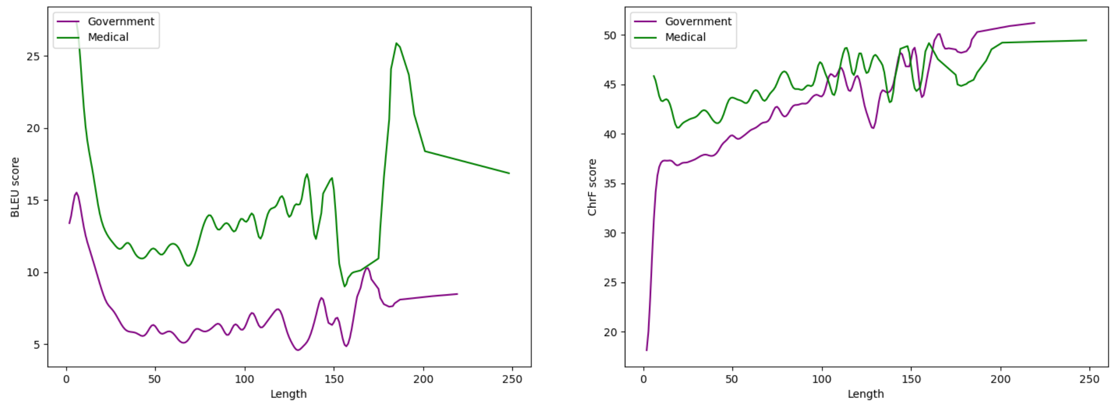

Figure 2: The trend of averaged sentence-level BLEU (left) and ChrF (right) scores against the token length of the reference translation for DALI models. The scores are averaged across all the model outputs of the same length – including averaging across languages, where relevant.

Source: if the other eye medicine is an eye ointment it should be used last

mBART:	jekk il-mara l-ieħor hi çajn oħra , hi çandha tinçatalha l-aħħar fl-aħħar	LeCA:	inkella jekk il-mara l-ieħor hi żejt , tkun maçmula l-aħħar
BT:	if the other woman is someone else, she should be punished in the end	BT:	otherwise if the other woman is a virgin, she will be the worst
DALI:	jekk il oħra għajn mediċina huwa an għajn infużjoni dan għandu tkun użati l-aħħar	CPT:	jekk l-oħrajn ta ' l-ieħor hi çajnejja ointment , it should be used l-aħħar
BT:	If the other eye medicine is an eye infusion, this should be used last	BT:	Even though the other one is a çajnejja ointment, it should be used last
Combined:	jekk il oħra għajn mediċina huwa an għajn ointment dan għandu tkun użati l-aħħar	Target:	jekk il-mediċina tal-għajnejn l-oħra hija ingwent tal-għajnejn , dan għandu jintuża l-aħħar
BT:	If the other eye medicine is an eye ointment, this should be used last	BT:	Although the other eye medicine is an eye ointment, this should be used last

Table 2: **Warning: this table contains harmful language about women that may distress some readers.** An example of different model outputs for a Maltese sentence in the medical domain. For better comparison, the back-translations (BT) of the outputs to English are also included, done via Google Translate.

of its English counterpart "eye medicine"), and it is also more flexible. DALI and *Combined* produce word orders that closely follow the source language.

Human Evaluation Conducting a small-scale human evaluation of the Polish government translations of 25 source sentences, we find that, while DALI improves the communication of the overall semantics of the sentences, there is certainly room for improvement, especially when it comes to fluency and generating grammatical output. Additional model outputs and details of the human evaluation can be found in Appendix C.

6 Conclusion

This paper introduces a realistic setting that has been previously overlooked: DA for NMT into a low-resource from a high-resource language, with available resources restricted to limited parallel text, a dictionary, and monolingual texts in the high-resource language. The simplest approach – DALI – yields the best results, more than doubling baseline performance. A small-scale human evaluation indicates ample room for improvement, and we advocate for increased focus on this setting.

Limitations

It is important that these experiments be conducted for truly low-resource languages. The scope of this work was limited due to the availability of datasets in different domains for such resource-constrained languages; which was the main reason we resorted to experimenting on simulated low-resource languages. Limitations of finding domain-specific corpora for low-resource languages also extend to finding domain-specific dictionaries, and our dictionaries prepared with Google Translate only mimic target-domain dictionaries. In addition, we based our experiments on mBART only, and we leave the study of other multilingual pretrained models and LLMs (or even smaller, non-pretrained models like the base Transformer) in this setting for future work.

Ethics Statement

As our research shows that these methods do not sufficiently enhance performance for the models to be deemed useful, there are some caveats to be mindful of. Specifically, these methods should not be used for real-world MT in critical contexts involving low-resource languages; e.g. providing medical advice based on the translations produced by the model.

All the data used in the study is publicly available (see Appendix A.1).

Acknowledgments

We sincerely thank the anonymous reviewers for their helpful feedback. We also wish to express our gratitude to LECS and NALA labs at the University of Colorado Boulder. This work made use of the Blanca condo computing resource at the University of Colorado Boulder, which is supported through joint funding by computing users and the university.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Niyati Bafna, Philipp Koehn, and David Yarowsky. 2024. Pointer-generator networks for low-resource machine translation: Don't copy that! In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 60–72, Mexico City, Mexico. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Guanhua Chen, Yun Chen, Yong Wang, and Victor OK Li. 2021. Lexical-constraint-aware neural machine translation via data augmentation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3587–3593.

Mara Chinea-Ríos, Álvaro Peris, and Francisco Casacuberta. 2017. Adapting neural machine translation with parallel synthetic data. In *Proceedings of the Second Conference on Machine Translation*, pages 138–147, Copenhagen, Denmark. Association for Computational Linguistics.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2021. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3440–3453, Online. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5894–5904, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

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.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. A simple baseline to semi-supervised domain adaptation for machine translation. *arXiv preprint arXiv:2001.08140*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

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.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2020. Efficiently reusing old models across languages via transfer learning. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 19–28, Lisboa, Portugal. European Association for Machine Translation.

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

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online. Association for Computational Linguistics.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-round transfer learning for low-resource nmt using multiple high-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(4).

Zhibo Man, Zengcheng Huang, Yujie Zhang, Yu Li, Yuanmeng Chen, Yufeng Chen, and Jinan Xu. 2023. Wdsrl: Multi-domain neural machine translation with word-level domain-sensitive representation learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:577–590.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the*

Twelfth Language Resources and Evaluation Conference, pages 2884–2892, Marseille, France. European Language Resources Association.

Sreyashi Nag, Mihir Kale, Varun Lakshminarasimhan, and Swapnil Singhavi. 2020. Incorporating bilingual dictionaries for low resource semi-supervised neural machine translation. *arXiv preprint arXiv:2004.02071*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Wei Peng, Chongxuan Huang, Tianhao Li, Yun Chen, and Qun Liu. 2020. Dictionary-based data augmentation for cross-domain neural machine translation. *arXiv preprint arXiv:2004.02577*.

MinhQuang Pham, Josep Crego, François Yvon, and Jean Senellart. 2019. Generic and specialized word embeddings for multi-domain machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. Adaptation of machine translation models with back-translated data using transductive data selection methods. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 567–579. Springer.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

Shoetsu Sato, Jin Sakuma, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2020. Vocabulary adaptation for domain adaptation in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4269–4279, Online. Association for Computational Linguistics.

Danielle Saunders. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, 75:351–424.

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

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *arXiv preprint arXiv:2201.03110*.

Emmanouil Stergiadis, Satendra Kumar, Fedor Kovalev, and Pavel Levin. 2021. Multi-domain adaptation in neural machine translation through multidimensional tagging. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 396–420, Virtual. Association for Machine Translation in the Americas.

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation. *arXiv preprint arXiv:2107.04239*.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024. Teaching large language models an unseen language on the fly. *arXiv preprint arXiv:2402.19167*.

Hongxiao Zhang, Hui Huang, Jiale Gao, Yufeng Chen, Jinan Xu, and Jian Liu. 2022. Iterative constrained back-translation for unsupervised domain adaptation of machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5054–5065, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Data

A.1 Datasets

Here are the details for the data used in training, testing and potential pretraining and pseudo data generation. All the datasets are lower-cased.

A.1.1 Parallel Data

The parallel data come from the New Testament verses from the Johns Hopkins University Bible Corpus (McCarthy et al., 2020). For all experiments, 8% of the verses are extracted to be used as validation data. The number of verses per language is in the range 7k-8k. The test dataset is of another domain, and it is discussed in A.1.2. The sizes of the train and validation datasets for different languages are shown in Table 3.

Language	Train	Validation
Croatian	7290	634
Icelandic	7167	624
Maltese	7122	620
Polish	7293	635
Ukrainian	6799	592

Table 3: Number of parallel Bible verses used in training and validation across different languages.

A.1.2 Pseudo-Parallel Data

We use the Tilde MODEL corpus (Rozis and Skadiņš, 2017) for the majority of our experiments, as it is listed as an available resource for many of WMT tasks during the last few years (Barrault et al., 2020; Akhbardeh et al., 2021; Kocmi et al., 2022, 2023). In all experiments, we retain 1500 sentence pairs for testing. This is the only portion for which we keep the target side, as we only manipulate the source side from the rest of them. In case there are more than 200K available pairs, we use seed $= 42$ to randomly choose 200K pairs from the dataset.

Government Domain

Croatian: We use EESC from the Tilde MODEL, that comprises document texts from the "European Economic and Social Committee" document portal. The full 200K sentence pairs are used for training and pretraining.

Icelandic: We use the concatenation of the following three datasets: "Government Offices in Iceland - Reports", "Government Offices in Iceland – Legislation and regulations", and "Bilingual English-Icelandic parallel corpus from the official Nordic cooperation website" from the European Language Resource Coordination.[6] This makes for a dataset of size 87233 that is used for both training and pretraining.

Maltese: As was the case with Croatian, we use the English-Maltese subsection of EESC, and we choose 200K sentence pairs from all available pairs.

Polish: We utilize RAPID from the Tilde MODEL, composed of the press releases of "Press Release Database of European Commission" released between 1975 and the end of 2016. 200K sentence pairs are extracted.

Ukrainian: We use "EU acts in Ukrainian" from the European Language Resource Coordination, resulting in 116,568 sentence pairs.

Medical Domain

For the four languages investigated (Croatian, Icelandic, Maltese, Polish), we use EMA from the Tilde MODEL. It is compiled from texts available via the European Medicines Agency document portal. All of these languages had more than 200K sentence pairs, from which 200K were extracted.

A.2 Dictionaries

The method with which the dictionaries are composed is described in 3. Since many of the lemmas might have several inflected forms that appear in the text, the dictionary sizes are larger than 5000, usually varying between 8k-10k. Here are the exact size of the dictionaries. In Table 4, the column 'Bible' denotes the number of terms extracted from the Bible and added to the *in-domain* terms that were drawn out from the monolingual source-side corpus. Note that if the term already exists in the in-domain dictionary, we do not replace it with the one from the Bible. The columns 'Government' and 'Medical' indicate the final size of the dictionary of their respective domains, including the new terms from the Bible.

B Training

The details of training are as follows. Each setting was trained once, and the experiments were done on NVIDIA A100 GPUs.

[6] https://language-data-space.ec.europa.eu/related-initiatives/elrc_en

	Bible	Government	Medical
Croatian	182	9948	8142
Icelandic	359	10004	8383
Maltese	319	10004	8309
Polish	417	10192	8337
Ukrainian	283	9437	-

Table 4: Sizes of different dictionaries used for different languages and domains.

Note that some of the experiments rely on others; for example, *Combined* has three stages of updating the model: 1) continual pretraining on the domain-specific texts, 2) training the model from step 1 on the Bible dataset (CPT), 3) training the model from step 2 on the pseudo parallel data + Bible (which is the DALI part). Some notable libraries we use include:

- fairseq v0.12.2 (which we modified to run our methods)
- torch v1.13.1
- sentencepiece v0.1.99
- transformers v4.30.2.

B.1 Training Hyperparameters

We implemented LeCA and CPT on `fairseq` for mBART, and had to change parts of the main library for compatibility. Since mBART needs a language id, we added new tokens for these new languages. We initialized their embeddings randomly (following the method for parameter initialization in Liu et al. (2020)). The `fairseq` hyperparameters used in pretraining and training are listed in Table 5.

B.2 Batches for experiments with DALI

In experiments containing DALI - DALI, *Combined* (and CPT + DALI which is done for medical domain experiments, as presented in Table 8) - batches are constructed in a particular way.

In each batch, we have the same number of instances from out-of-domain parallel data and in-domain pseudo-parallel data. Training batches do not contain overlapping in-domain pseudo-parallel data, but we do use the same out-of-domain parallel data in every batch, because we are limited to Bible verses for parallel data.

C Additional Outputs and Evaluation

Table 6 shows the model outputs for an example sentence from the Polish test set in the government domain. We can see the same patterns of religious phrasing in mBART, LeCA and CPT. Some words have different translations than those used in the target translation; e.g. the model translates *banking* as “bankowość” while “bankowej” is used in the reference. Polish is also a morphologically rich language and it sometimes does not match English’s word order. Here, for example, *banking union* should be translated as “unii bankowej” while in DALI and and *Combined* the phrase is translated as “bankowość unia”, in the same order as in the English sentence.

Human evaluation We conduct a small-scale human evaluation on a set of 25 randomly-selected sentences from the test set of Polish government data. A Polish native speaker annotator scored the translations for both communication of the **intended meaning** and the correctness of the overall **grammatical structure**, using a scale from 0 to 5. Only the translations of the original mBART (baseline) and DALI are compared. The average scores for **meaning** for baseline and DALI were 0.12 and 0.2, respectively. For the **grammar**, both models were given an average score close to 0. (Perhaps not surprising for a language with the morphological richness of Polish.) Of course, more in-depth study of the results is needed to draw any strong conclusions about usability.

Output statistics The average number of words, number of tokens, and number of characters of the outputs of different methods against the reference translations are presented in Table 7. For number of words, an output is split by white-spaces. For tokens, the mBART tokenizer is used. We average the results across languages. We report the averages because relative length patterns tend to be consistent across languages. The full table containing language specific statistic is available on the GitHub repository: https://github.com/alimrsn79/da_lr_nmt.

Hyperparameter	Pretraining	Training
arch	mbart_large	
lr-scheduler	polynomial_decay	
lr	3e-5	
optimizer	adam	
adam-eps	1e-06	
adam-betas	(0.9, 0.98)	
dropout	0.3	
attention-dropout	0.1	
bpe	sentencepiece	
max-tokens	1024	
save-interval	5	
criterion	label_smoothed_cross_entropy	
no-epoch-checkpoints	True	
layernorm-embedding	True	
encoder-normalize-before	True	
decoder-normalize-before	True	
share-decoder-input-output-embed	True	
encoder-learned-pos	True	
required-batch-size-multiple	1	
label-smoothing	0.2	
update-freq	2	
seed	42	
warmup-updates	2000	1000
min-epoch	20	75
min-epoch	60	150
patience	10	50
total-num-update	(number of steps in one epoch) * max-spoch	
task	denoising	translation_from_pretrained_bart
mask	0.35	-
tokens-per-sample	384	-
poisson-lambda	3.5	-
mask-length	span-poisson	-
replace-length	1	-
rotate	0	-
permute-sentences	0	-

Table 5: Pretraining and training hyperparameters

Source: in the banking union , those funds are pooled together gradually .			
mBART:	przetoż zgromadzi się wszystkie , które są w łodzi .	LeCA:	przetoż zgromadzi one członki w lichwiarze .
DALI:	w the bankowość unia , te fundusze czy poszczepiony razem stopniowo .	CPT:	w banking union wespół to zgromadziło , i nader to zgromadziło .
Combined:	w the bankowość unia , te fundusze czy pooled razem stopniowo .	Target:	fundusze te będą gromadzone stopniowo w ramach unii bankowej .

Table 6: An example of different model outputs for a Polish sentence in the government domain.

	Domain	Average Words	Tokens	Characters
Reference	Gov.	23.51	44.65	154.55
	Med.	19.65	39.43	120.42
mBART	Gov.	25.12	49.25	132.7
	Med.	20.91	44.27	110.94
DALI	Gov.	26.83	43.99	155.28
	Med.	20.3	37.38	115.7
LeCA	Gov.	24.82	49.15	133.92
	Med.	22.94	46.02	117.94
CPT	Gov.	26.44	47.85	139.7
	Med.	20.5	38.23	109.46
Combined	Gov.	26.93	43.97	155.95
	Med.	20.31	37.41	116.09

Table 7: The average number of words, tokens, and characters of the outputs of different methods against the reference translation. The results are averaged over all the experiments.

	Metric	Croatian	Icelandic	Maltese	Polish	Average
DALI	BLEU	12.74	13.89	16.68	10.57	13.47
	ChrF	**43.32**	**41.07**	48.77	**36.73**	**42.27**
Combined	BLEU	12.21	13.4	16.75	10.67	13.26
	ChrF	42.11	40.56	48.88	36.11	41.92
CPT + DALI	BLEU	12.59	13.28	17.03	10.88	13.45
	ChrF	42.6	38.67	**49.1**	36.36	41.68

Table 8: Comparing CPT + DALI with DALI and *Combined* on the medical domain.

	Metric	Icelandic Gov.	Med.
DALI	BLEU	5.76	13.89
	ChrF	36.02	41.07
Combined	BLEU	5.63	13.4
	ChrF	36.33	40.56
Full	BLEU	34.46	55.98
	ChrF	59.1	74.05

Table 9: Comparing the model trained on the full parallel dataset with DALI and *Combined* that only had access to the source side, for Icelandic. The full models were trained with the same hyperparameters as the training column in Table 5, but the training was done on the full in-domain parallel text instead of the Bible and pseudo-parallel sentences.