Towards Low-resource Language Generation with Limited Supervision

Kaushal Kumar Maurya **Maunendra Sankar Desarkar**
Natural Language and Information Processing Lab (NLIP)
Indian Institute of Technology Hyderabad,
Hyderabad, India
cs18resch11003@iith.ac.in & maunendra@cse.iith.ac.in

Abstract

We present a research narrative aimed at enabling language technology for multiple natural language generation (NLG) tasks in low-resource languages (LRLs). With approximately 7,000 languages spoken globally, many lack the resources required for model training. NLG applications for LRLs present two additional key challenges: (i) The training is more pronounced, and (ii) Zero-shot modeling is a viable research direction for scalability; however, generating zero-shot well-formed text in target LRLs is challenging. Addressing these concerns, this narrative introduces three promising research explorations that serve as a step toward enabling language technology for many LRLs. These approaches make effective use of transfer learning and limited supervision techniques for modeling. Evaluations were conducted mostly in the zero-shot setting, enabling scalability. This research narrative is an ongoing doctoral thesis[1].

1 Introduction

Recently, there has been remarkable progress in natural language processing (NLP) research, primarily due to advancements in large pre-trained language models (PLMs). The global linguistic landscape comprises approximately 7,000 spoken languages worldwide[2]. A notable disparity is evident in NLP research, with the majority of studies conducted on English data (Bender, 2019; Joshi et al., 2020b). This is concerning as the vast majority of the global population — roughly 95% — does not speak English as their primary language, and a staggering 75% do not speak English at all[3]. According to Ruder (2022), out of the 7,000 languages, approximately 400 languages have more than 1 million speakers, and about 1,200 languages have more than 100,000 speakers. Despite this, only around 100 languages are incorporated into large pre-trained models, and limited resources are available for building NLP models for LRLs. Furthermore, a study presented at ACL 2008 (Bender, 2011) revealed that 63% of all papers focused only on English. A more recent study during ACL 2021 (Ruder et al., 2022) concluded that nearly 70% of the papers were evaluated on English. Even a decade later, there has been little change.

The NLP application involving text generation (NLG tasks) in LRLs presents additional challenges in model development: (1) The scarcity of NLG resources for model development in LRLs is more pronounced than other NLP tasks. (2) LRLs often exhibit a long tail, with many lacking annotated data. The preferred solution is zero-shot modeling, though this approach introduces additional challenges for cross-lingual generation tasks. It has been observed that zero-shot generation models frequently encounter issues like catastrophic forgetting (van de Ven et al., 2022) or accidental translation (Xue et al., 2021). Due to these problems, the zero-shot generated text is either code-mixed or not in the intended target language. (3) LRL modeling typically employs a transfer learning setup, where supervision is transferred from HRLs to LRLs. However, performance tends to degrade for LRLs that are different from their HRL and (4) Many LRLs lack monolingual or parallel data, and their representations are absent from PLMs. These LRLs are referred to as Extremely LRLs (ELRLs) or dialects. Despite having millions of speakers, there is a noticeable absence of NLP technology for these ELRLs. This thesis is a step towards addressing these challenges and aims to enable language technology for LRLs, thereby democratizing NLP research for the general population/audience.

Prior to the emergence of transformers-based

[1] From a senior graduate student - the first author of the paper

[2] https://www.ethnologue.com/insights/how-many-languages/

[3] https://www.ethnologue.com/insights/most-spoken-language/

PLMs, most works in cross-lingual generation were primarily reliant on machine translation (MT) systems. Existing models either directly employed the MT system within the modeling (Wan et al., 2010; Shen et al., 2018) or generate training data using MT (Kumar et al., 2019; Chi et al., 2020) to develop models. This dependence on MT not only limits scalability but also propagate error with translation. To address these limitations, multilingual PLMs (mPLMs) have emerged (Zhao et al., 2023), where a large set of languages share a common latent representation space. The cross-lingual models built on top of these mPLMs lead to the remarkable advancement (Hu et al., 2020; Artetxe et al., 2020) in the cross-lingual transfer in zero-shot or few-shot settings. However, most of these advancements are limited to NLU tasks. Furthermore, existing cross-lingual NLG models incorporate one or more challenges mentioned above.

With this thesis, our contributions are as follows:

1. We proposed ZmBART framework (Maurya et al., 2021) to mitigate the catastrophic forgetting and accidental translation issues and enable well-formed zero-shot text generation in LRLs. We evaluated the model's performance across 18 task-setup combinations, including four NLG tasks in three typologically diverse languages.

2. We proposed the first meta-learning approach for cross-lingual generation in LRLs (MetaX$_{\text{NLG}}$; Maurya and Desarkar (2022)). It is based on language clustering to improve the cross-lingual transfer, even for distant LRLs. The model is evaluated across 30 languages, two tasks, and five datasets.

3. We proposed a character span noise augmentation-based model (CHARSPAN; Maurya et al. (2023)) to enable machine translation for closely related HRLs and ELRLs/dialects. It leverages surface-level lexical similarity and uses noise augmentation as a regularization technique to enable zero-shot translation. The model's performance was evaluated across 12 ELRLs from three typologically diverse language groups.

2 The Big Picture

In this section, we provide high-level details of the proposed models. This also includes insights into how we build more recent proposed models based on earlier models and advance the field. Then, we look back and position our research efforts by contextualizing a broader spectrum of multilingual research, specifically for low-resource language generation. Finally, we list our learnings from failed and successful modeling.

2.1 Thesis Overview: Connecting the Dots

Overall, our research contribution includes the development of ZmBART, MetaX$_{\text{NLG}}$, and CHARSPAN models for NLG tasks in LRLs. The primary focus is to extend the English NLG models to LRLs through cross-lingual transfer and generation. These models are developed and evaluated in a zero-shot setting, increasing language coverage. Typical cross-lingual modeling includes fine-tuning multilingual PLMs with the task-specific high-resource English language and learned supervision for transfer to LRLs (referred to as cross-lingual transfer). Then, evaluate the model with a zero-shot setting for target LRLs. In NLG, there are two challenges: mitigation of the CF/AT problem in zero-shot text generation and improvement of cross-lingual transfer. The effort with the ZmBART model mitigates the CF/AT issue and produces well-formed zero-shot generation in LRLs. MetaX$_{\text{NLG}}$builds on top of the ZmBART model and proposes a novel approach to improve cross-lingual transfer, leading to better performance. Finally, with the CHARSPAN model, we design another approach to enhance cross-lingual transfer. This effort scales the coverage to languages with very limited linguistic resources (i.e., ELRLs) and is similar to some HRLs. In summary, with these collective efforts, we advance research in low-resource language generation by mitigating CF/AT, improving cross-lingual transfer, and increasing language coverage to ELRLs.

2.2 Position of the Thesis: Related Work

The research presented in this narrative spans the past few years, during which multilingual Pre-trained Language Models (PLMs) emerged. However, there have been limited concurrent efforts in the field of low-resource language generation. Before the ZmBART model, most research in this area primarily relied on MT (Wan et al., 2010; Shen et al., 2018), parallel (Chi et al., 2020) or task-specific data for LRLs (Kumar et al., 2019), and did not utilize multilingual PLMs. Few attempts were made using Adapter-based models (Houlsby

et al., 2019; Pfeiffer et al., 2021), but they were often limited to MT tasks and may not have zero-shot capabilities. After ZmBART, (1)Vu et al. (2022) presented the alternate method with prompt tuning and compared it to the ZmBART, (2) Li and Murray (2023) proposed a model based on regularization techniques and (3) Pfeiffer et al. (2023) introduced a method for disentangling language-specific information from language-agnostic information. These models mitigate the CF/AT problems and implicitly help improve the cross-lingual transfer. However, their performance gains were limited compared to MetaX$_{\text{NLG}}$which explicitly leverages meta-learning. Furthermore, there are state-of-the-art (SOTA) approaches (Aepli and Sennrich, 2022; Provilkov et al., 2020; Patil et al., 2022) for enhancing cross-lingual transfer for MT for ELRLs. Our recently proposed CHARSPAN model has outperformed existing models and established it as a new SOTA solution. In summary, there has been progress in low-resource language generation, and our models have either pushed this research space or currently represent the SOTA model in the field.

2.3 Learning from Failures and Successes

With many failed and limited successful experiments, here are our key observations and learning: (1) NLG modeling is challenging in LRLs setup, but evaluations are even more challenging. (2) Effective cross-lingual transfer models consider various knowledge, such as semantics, syntax, tokenization, lexical details, typology, and demographics. (3) Better modeling can extend the existing multilingual PLMs capabilities beyond the languages they are trained and (4) Promising research directions to increase language technology coverage are multi-task and adaptive learning among others.

3 Mitigating Catastrophic Forgetting to Enable Zero-shot Language Generation

Our research mission to enable language technology for NLG tasks in LRLs started with ZmBART (Maurya et al., 2021) work. ZmBART is an unsupervised cross-lingual transfer and generation framework that focuses on generative tasks for LRLs in zero-shot and few-shot settings. A typical zero-shot cross-lingual generation modeling involves two main steps: (1) *Training with HRLs:* Train (fine-tune) a model (PLM) using a large annotated dataset from HRLs, typically English. For instance, training with English Abstractive Text Summarization (ATS) dataset. (2) *Zero-shot generation in LRLs:* Utilize the trained model for zero-shot inference. For instance, when given input in an LRL (e.g., Hindi), the model generates a summary in the same LRL (Hindi). Unlike natural language understanding (NLU) tasks, the cross-lingual generation task in zero-shot scenarios is particularly challenging. This is because the zero-shot generated text needs to be in the target LRL, which generally suffers from Catastrophic Forgetting (CF; van de Ven et al. (2022)) or Accidental Translation (AT; Xue et al. (2021)) problems. Due to this, the model fails to generate text in the target LRL or produce code-mixed output with both high-resource and LRLs. *With this work, our objective is to alleviate CF and AT problems with an unsupervised framework, meaning we do not rely on any parallel or pseudo-parallel/back-translated data.* Instead, we harness multilingual pre-trained checkpoints, specifically the mBART model (Liu et al., 2020), to seamlessly enable the generation of well-formed text in LRLs across multiple generative tasks.

Prior to ZmBART, existing cross-lingual generation models were grounded with either machine translation (MT) or parallel/back-translated datasets. Wan et al. (2010) employed the MT pipeline to facilitate cross-language document summarization. This involves the translation of non-English input into English. Subsequently, the English ATS model was employed to procure the summaries, which were finally translated back into non-English languages. Similar approaches are adapted by Shen et al. (2018) and Duan et al. (2019). This direction is not feasible as MT systems are not available for many LRLs and the imperfect translations propagate errors. Considering this, Kumar et al. (2019) and Chi et al. (2020) use back-translated (need MT system) and parallel datasets to develop the few-shot cross-lingual question and answering (Q&A) and zero-shot cross-lingual ATS, respectively. These approaches require an MT system or annotated dataset which limits the model development to a few HRLs. Unlike these, we propose ZmBART, the first unsupervised scalable model based on mBART specialized for zero-shot cross-lingual transfer and generation. Additionally, we have also created *HiDG*[4], a high-quality distractor generation dataset in the Hindi language.

[4]Dataset and code are available here: https://github.com/kaushal0494/ZmBART

3.1 Methodology

In ZmBART, we mitigate Catastrophic Forgetting and Accidental Translation problems by adapting three key modeling modifications, details are presented below:

3.1.1 Unsupervised Auxiliary Task

The mBART model is pre-trained with denoising objectives (masking and sentence permutation) with datasets from 25 languages that encode multilingual latent representation. This can not be used directly for cross-lingual generation because the model is trained with denoising objectives that do not directly follow auto-regressive decoding, thereby causing a mismatch between pretraining and fine-tuning objectives (Chi et al., 2020; Devlin et al., 2019). Considering this, the auxiliary task is formulated with the following objectives: (1) should only utilize monolingual data for selected languages, (2) should enhance the latent representation space for selected languages, (3) maintain close proximity between the auxiliary task objective and NLG tasks and (4) aid in mitigating CF/AT issues. Moreover, the auxiliary task serves as an adaptive pre-training step, facilitating *better warm-start* of the mBART model for downstream natural language generation (NLG) tasks. With these, we have proposed the following auxiliary task: *Given an input passage, generate a few random sentences (called rand-summary) derived from the passage.* Concretely, we take passages with 5-25 sentences as input and 20% of the sentences randomly (1-5 sentences) as the target. We concatenate monolingual datasets for selected languages and fine-tune the mBART model (adaptive training) with this auxiliary task to obtain the ZmBART model.

3.1.2 Freezing Model Components

During supervised training - fine-tuning ZmBART with task-specific HRL data - we freeze all word embeddings and the parameters of the decoder layers. This approach is adapted to ensure that the ZmBART's context and latent space are not overwritten during supervised training.

3.1.3 Adding Language Tag

We have made modifications to the language tag of the mBART model for the cross-lingual generation framework. We concatenate <fxx><2xx> tag in the source side of the training data, where <xx> is the ISO-2 language code. The language tag act as a flag to trigger the zero-shot generation in target <xx> languages.

The ablation study provides evidence that all three components are necessary to effectively mitigate CF/AT problems and enable structured text generation in a zero-shot setting.

3.1.4 Model Training and Generation

We consider four tasks: Question Generation (QG), News Headline Generation (NHG), Abstractive Text Summarization (ATS), and Distractor Generation (DG), in three typologically diverse languages. The HRL is English (en), and the LRLs are Hindi (hi) and Japanese (ja). First, the mBART model undergoes adaptive pre-training with the auxiliary task to obtain the ZmBART model. Then for each NLG task, the ZmBART model is then fine-tuned using the task-specific HRLs data while freezing model components to obtain a task-specific fine-tuned model. This model is used for zero-shot or few-shot (1000 examples) generation in LRLs.

3.2 Experimental Setup and Results

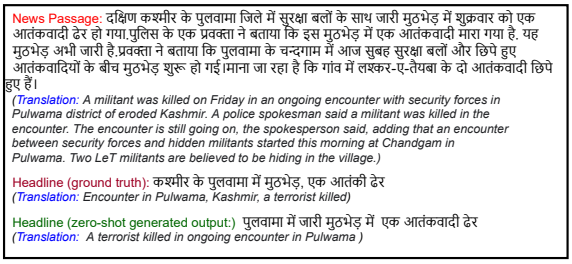

Figure 1: Zero-shot news headline generation from ZmBART in the Hindi language

We have considered three strong baseline models: MT-Pipeline, ZmBART with Masking Auxiliary Task (MAT), and a model inspired by Chi et al. (2020). In total, we conducted experiments across 18 task-setup combinations. The proposed models and baseline models underwent evaluation using three automated evaluation metrics (BLEU, ROUGE-L, and BERTScore) and four manual evaluation metrics (Fluency, Relatedness, Correctness, and Distractibility). The detailed results are presented in (Maurya et al., 2021). Here, we provide a summary of the major results and observations: (1) The ZmBART model consistently outperformed all baseline models across tasks, LRLs, and automated metrics in the zero-shot setting. The few-shot training further boosts the performance. (2) Human evaluation scores exhibited a correlation with automated scores, reinforcing the reliability of the

evaluation process. (3) Among the baselines, the MAT baseline demonstrated superiority, highlighting the importance of an auxiliary task in enriching and mitigating CF/AT problems. However, our proposed auxiliary task exhibited even better results. (4) An ablation study was conducted, indicating that different modeling components (auxiliary task, language tag, and freezing different model components) are necessary to ensure effective zero-shot text generation. A sample generation example is presented in Fig. 1.

3.3 Insights and Limitations

As the auxiliary task is similar to NHG or ATS tasks, it may appear that the auxiliary task is biased towards these tasks, which leads to better performance. However, the model performs equally well for very different tasks like QG and distractor generation (generating incorrect options for MCQ reading comprehension) which nullifies this assumption. We have not modified any single model parameters for different tasks. We also experimented with different objectives for auxiliary tasks; however, the *rand-summary* task performed best. We explored the multiple continual learning techniques (van de Ven et al., 2022) to mitigate CF; however, freezing model components work best. We observed that several generated questions in zero-shot start with English 'wh-words,' and the first word is code-mixed. This is possibly due to English interrogative sentences often introducing 'wh-words' at the beginning, which may not be the case with Hindi and Japanese. However, the high BERTScore indicates semantic correctness. Furthermore, such code-mixing in human evaluation is somewhat acceptable with Hindi evaluators; however, it is not acceptable with Japanese evaluators, resulting in lower human evaluation scores for the QG task. This is concurrent work with the adapter-based models (Houlsby et al., 2019; Pfeiffer et al., 2021). One limitation of this work is the adaption of the new language may require re-training.

4 Meta-Learning Approach to Improve Zero-shot Language Generation

The effort with the ZmBART helps in effectively mitigating CT/AT problems and generating zero-shot outputs in target LRLs seamlessly. In this work, we leverage these findings and extend the study to improve the cross-lingual supervised signals to boost the performance for zero-shot generation.

There are more than 7000 languages across the globe. 95% of the world's population does not speak English as their first language and 75% does not speak English at all[5]. However, the majority of NLP research is focused on the English language (Bender, 2019; Joshi et al., 2020b). To democratize the NLP research for the benefit of the large global community, it is essential to focus on non-English languages. Recently, cross-lingual transfer learning (Hu et al., 2020; Artetxe et al., 2020) has emerged as a promising research direction where a model is trained on HRL(s) and *transfer supervision* to LRL(s). However, the supervision transfer is uneven across languages, which leads to large performance gaps. Such performance gaps are observed because models do not account for cultural and linguistic differences in the modeling (Lai et al., 2019; Blasi et al., 2022). This work was a step towards bridging this performance gap.

Meta-learning or *learning to learn* (Bengio et al., 1990) has emerged as an active research direction to learn *shareable structures* across multiple tasks with limited annotated data. The only constraint is all tasks should share some common structure (or come from a task distribution). Different languages in the world follow this constraint as they come into existence with a common goal of communication and share some structure. So, we consider languages as tasks. The meta-learning approach has been actively applied to multiple NLP tasks (Bansal et al., 2020; Gao et al., 2019) including text classification (van der Heijden et al., 2021), NER (Wu et al., 2020), dialogue systems and Q&A (M'hamdi et al., 2021). There were few efforts made in the multilingual setup (Tarunesh et al., 2021; Nooralahzadeh et al., 2020); however, these are limited to machine translation or NLU tasks only. This work - to the best of our knowledge - was the first attempt to study *meta-learning techniques for cross-lingual natural language generation (X_{NLG})*. Particularly, we focus on zero-shot X_{NLG} for low-resource languages. Unlike NLU tasks, the zero-shot NLG is a more challenging setup due to the typological diversities of languages and CF/AT problems. We refer to this framework as MetaX_{NLG}[6] (Maurya and Desarkar, 2022), a framework for effective cross-lingual transfer and gen-

[5] https://www.ethnologue.com/insights/most-spoken-language/

[6] code & pre-trained models link: https://github.com/kaushal0494/Meta_XNLG

eration based on language clustering and Model-Agnostic Meta-Learning (MAML) algorithm (Finn et al., 2017).

Following are the main contributions: (1) We propose a novel MetaX$_{\text{NLG}}$framework based on language clustering and meta-learning to improve zero-shot generation performance for typologically diverse LRLs. (2) We have conducted an extensive empirical evaluation with 30 languages (29 LRLs), covering two tasks (QG and ATS) and using 5 popular datasets (XL-Sum, Wikilingua, MLQA, TyDiQA, and XQuAD).

4.1 Methodology

The MetaX$_{\text{NLG}}$model has two major components: (a) *Language Clustering*, which clusters 30 selected languages into different clusters and obtains the centroid and non-centroid languages for each cluster. (b) *Meta-learning* algorithms are trained with centroid languages and evaluated with non-centroid (target) LRLs in a zero-shot setting. With this setup, our goal is to achieve *Intra-cluster Generalization* and *Inter-cluster Generalization*. Training with a centroid language leads to improved transfer capability within a cluster, and multiple centroid languages extend the transfer capability to other closely-knit clusters, thereby increasing coverage. The overview of MetaX$_{\text{NLG}}$is presented in Fig. 2.

4.1.1 Language Clustering

In MetaX$_{\text{NLG}}$, we considered 30 languages. To represent each language we have extracted a *multi-view* language representation proposed by Oncevay et al. (2020). It was obtained by fusing typologically learned (Littell et al., 2017) from WALS and URIEL databases and task-learned (e.g., language tag from MT; Malaviya et al. (2017)) language representations using singular vector canonical correlation analysis. We use this representation to obtain centroid and non-centroid based on cosine distance. Formally, given a cluster $C = \{L_1, L_2, \dots L_t\}$, where each L_i is multi-view representation of i^{th} language, the centroid language $L^* \in C$ is defined as:

$$L^* = \arg\min_{L_i \in C} \sum_{L_j \in C} d(L_j, L_i).$$

(1) We use d as the cosine distance.

4.1.2 Meta Training and Generation

The framework comprises five training/generation steps:

1. *Selection of Base PLM:* The proposed approach is model-agnostic; however, due to its large LRLs coverage, we have chosen the multilingual T5 (mT5) (Xue et al., 2021) as the base PLM.
2. *Adaptive Unsupervised Pre-training (ZP_M):* We follow steps outlined in ZmBART to obtain ZmT5 model.
3. *Fine-tuning ZP_M with HRL:* To facilitate the transfer of supervision from HRLs to LRLs, we have fine-tuned ZP_M using a task-specific HRL (e.g., English), which we refer to as $EnZP_M$.
4. *Meta-Training with Low-resource Centroid Languages:* A small, task-specific validation dataset of centroid languages was employed to train the $EnZP_M$ model using the MAML algorithm.
5. *Meta-adaptation for Zero-shot Evaluation with Non-Centroid Languages:* Finally, the meta-learned model is directly evaluated using a task-specific test split of the target languages in the zero-shot scenario.

There is a trade-off between the number of clusters (centroid languages) and generalization. If there is a single cluster (a single meta-training language), then the model tries to over-generalize for different typological structures and fails in the attempt. On the other extreme, if there are too many centroid languages (many typologically diverse structures), then the learning possibly gets distracted. In both cases, the model will be unable to learn a reasonable structure (the required generalization) and perform poorly. The MetaX$_{\text{NLG}}$presents a discussion and empirical evidence on this. Our experiments suggest that *three clusters* across considered languages provide the best performance.

4.2 Experimental Setup and Results

We evaluated the MetaX$_{\text{NLG}}$performance in the following settings: ((1) Two NLG tasks - Question

Cluster-1(14)	Cluster-2(8)	Cluster-3(8)
hi,ur,te,tr,ja,fi,ko,gu, bn,mr,np,ta,pa,sw	es,it,pt,ro, nl,de,en,fr	ru,cs,vi,th, zh,id,el,ar

Table 1: Clustering of considered 30 Languages

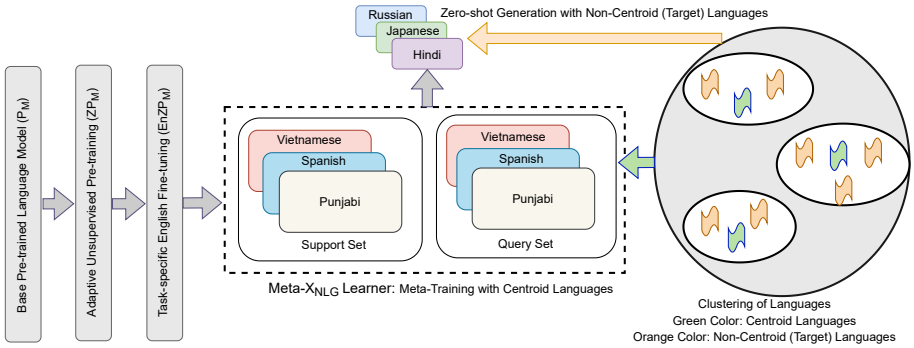

Figure 2: An overview of Meta-X_{XNLG} framework

Generation (QG) and Abstractive Text Summarization (ATS). (2) Five widely-used datasets: XL-Sum, Wikilingua, MLQA, XQuAD, and TyDiQA. (3) 30 languages were selected based on diversity typology, including one HRL (English) and 29 LRLs. Refer to Table 1 for the list of selected languages grouped into three clusters. (4) We employ two automated evaluation metrics (BLEU and ROUGE-L) and three human evaluation metrics (Fluency, Relatedness, and Correctness). (5) LRL evaluation in zero-shot setting on the test split. (6) We compare model performance against two strong baselines: (a) A ZmBART-like model using mT5 as the base checkpoint instead of mBART, and (b) a model fine-tuned directly with centroid languages rather than meta-training, ensuring the performance gain is not due to additional training.

Details of all results and observations are included in the MetaX_{NLG}original paper (Maurya and Desarkar, 2022). In summary, based on automated scores, the proposed MetaX_{NLG}model outperformed baselines in 30 out of 33 LRLs for the ATS task and in 18 out of 19 LRLs for the QG task. Even in cases where it did not perform as well, the difference was marginal. These trends were consistent when considering human evaluation metrics as well, where human scores showed a correlation with automated scores. The MetaX_{NLG}demonstrated above-average fluency and correctness scores, indicating its quick adaptation to various syntactical structures and overall improved performance. The consistent improvement for most of the typologically diverse LRLs provides evidence that supervision transfer is more uniform.

4.3 Insights and Limitations

As discussed in Section 4.1.2, there is a trade-off between the number of clusters and generalization capabilities. To ensure that we have selected the correct number of clusters, we have conducted an extensive adaptation study with 36 experimental setups involving different numbers of clusters and various combinations of languages. We observed that the model with three clusters performs the best. From Table 1, we can observe that most of the clustering results are close to the clustering approach with language family - further validating the correctness of clustering. Furthermore, less improvement is observed for Wikilingual data (ATS). This could be due to the nature of Wikilingual input articles, which consist of instructions for operating software tools/packages. Each instruction is crucial, making it challenging to generate an accurate summary in zero-shot LRLs. One limitation, we need small task-specific annotated data for centroid languages, which will be used in the meta-training.

5 Utilizing Lexical Similarity to Enable Zero-Shot MT for Extremely LRLs

The efforts with ZmBART, MetaX_{NLG}, and the NLP research community on multilingual modeling have extended the coverage of NLP technologies for many LRLs. However, there is a *long-tail* of languages for which there is no parallel/pseudo-parallel data, no/limited monolingual data, and their representations from the multilingual language model are absent. These fall into categories of *extremely low resource languages (ELRLs)* or *dialects*. With this work (Maurya et al., 2023), we made a step towards enabling technology for ELRLs where resources are limited (zero-shot setting). In particular, our focus was on the machine translation (MT) task, driven by the availability of a true evaluation test set from recently released sources such as FLORES-200 (Costa-jussà et al., 2022).

Fortunately, many of these ELRLs are lexically

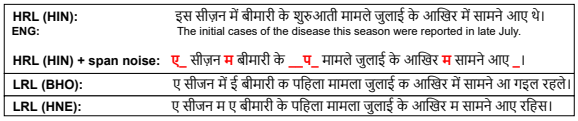

Figure 3: Hindi (HIN; HRL), Bhojpuri (BHO; LRL) and Chhattisgarhi (HNE; LRL/Dialect) parallel sentences. Additionally, the corresponding noisy Hindi example with character-span noise. BHO and HNE are closely related to Hin.

similar to closely related HRLS. *Lexical similarity refers to languages sharing words with similar form (spelling and pronunciation) and meaning*.[7] This includes cognates, lateral borrowings, and loan words. For example, the word lgtA (*lagta*) in Hindi (HRL) is spelled as lAgatA (*laagata*) in Bhojpuri (LRL). Existing cross-lingual transfer methods based on common embedding spaces work best between related languages (Nguyen and Chiang, 2017; Khemchandani et al., 2021). So, if we make the HRL model robust to spelling variations, it will improve cross-lingual transfer to related ELRLs. To achieve this, we introduce unigram character and character-span noise augmentation approaches, CHARSPAN, to improve generalization in zero-shot. The noise injection acts as a regularizer. A sample example is presented in Fig. 3. Formally, *we look at a machine translation task from an ELRL to another language (English) with transfer enabled by a related HRL on the source side.*

The character-level noise augmentation has been employed to improve the robustness and adversarial testing (Sperber et al., 2017; Vaibhav et al., 2019; Karpukhin et al., 2019) for MT systems. There are general noise augmentation techniques (Sennrich et al., 2016a; Wang et al., 2018) that help in cross-lingual transfer. Aepli and Sennrich (2022) introduced unigram character noise augmentation for NLU tasks such as NER, POS tagging, and topic classification. In contrast, we propose CHARSPAN noise augmentation for the more challenging MT task. There is another line of works that leverages lexical similarity based on vocabulary overlap (Patil et al., 2022), non-deterministic segmentations (Provilkov et al., 2020), and soft decoupled encoding (Wang et al., 2019). While these approaches typically require certain amounts of monolingual data, our proposed model operates without such constraints, eliminating the need for monolingual data. With this work, our key contributions are: (a) we show that unigram character and character-span level noise augmentation can improve zero-shot translation from ELRLs to English. CHARSPAN model outperforms the unigram model. (b) The proposed approach is generalized across three typologically diverse language groups which include 6 HRLs and 12 ELRLs.

[7] https://en.wikipedia.org/wiki/Lexical_similarity

5.1 Methodology

5.1.1 Training and Zero-shot Generation

First, we created an augmented parallel corpus from HRL (h) to English (En) as $\hat{\mathcal{D}_{\mathcal{H}}} = \{(\hat{h}, e)|lang(\hat{h}) = \hat{\mathcal{H}}, lang(e) = En\}$, where $\hat{\mathcal{H}} = \eta(\mathcal{H})$ and η is noise function. The input parallel corpus ($\mathcal{D}_{\mathcal{H}}$) was augmented with different kinds of noise (η) in the source HRL side (described later) to create the augmented parallel corpus ($\hat{\mathcal{D}_{\mathcal{H}}}$). We learned the subwords vocabulary $\mathcal{V}$ using ($\hat{\mathcal{D}_{\mathcal{H}}}$). We train the standard encoder-decoder transformer model ($\mathcal{M}$; Vaswani et al. (2017)) from scratch with ($\hat{\mathcal{D}_{\mathcal{H}}}$) and $\mathcal{V}$ to obtain the trained model $\mathcal{M}^{'}$. Finally, zero-shot evaluations are performed with $\mathcal{M}^{'}$ for the source ELR language $\mathcal{L}$ to obtain a target English translation.

5.1.2 Noise Function

We conducted experiments involving two types of noise functions: (1) unigram character noise and (2) character-span noise. For unigram noise, we randomly selected 9-11% of the characters from each source example (excluding punctuation and numbers) and applied insertion, deletion, and replacement operations with equal probabilities[8]. The unigram character noise has the potential to capture limited variations, particularly relevant for very similar languages and dialects. *To address larger lexical divergence, we propose a character-span noising approach, i.e., applying to noise a span of selected characters*. Our particular span noising approach is inspired by SpanBERT (Joshi et al., 2020a).[9] We randomly select 1 to 3-gram character spans with uniform probability and apply span noise until the noise injection budget (ranging from 9-11% of characters) is exhausted. Our approach includes *span deletion* and *span replacement with a single random character*, both with equal probability as the noising operations. In the original paper (Maurya et al., 2023), we conducted various ablation studies involving different combinations of operations, noise budgets, and other parameters.

[8] We explored some linguistically motivated noising schemes as well, but these did not yield any benefits.

[9] SpanBERT applies denoising to subword tokens while we apply it at the character level.

Based on our findings, we concluded that the proposed setup works best.

5.2 Experimental Setup and Results

We have carefully selected three typologically diverse language groups: Indo-Aryan, Romance, and Malay-Polynesian. We consider 6 HRLs and 12 ELRLs (2 HRLs and several ELRLs from each group). All the ELRLs and dialects are lexically similar to corresponding HRLs. Each group has the same writing script for all languages. For training, we use 13.6, 11, and 0.8 million public, parallel examples for Indo-Aryan, Romance, and Malay-Polynesian, respectively. The model's performance was evaluated on the FLORES-200 devtest set. Based on recent literature in low-resource MT, we compare our approach with Vanilla NMT with BPE segmentation (Sennrich et al., 2016b), methods using lexical similarity (Overlap BPE and BPE-Dropout) and their combinations. In alignment with recent studies (Costa-jussà et al., 2022; Siddhant et al., 2022) on MT for ELRLs, the evaluation scores are reported with chrF (Popović, 2015) and BLEU.

We have observed that the unigram noise injection outperformed all the baselines across all three language groups. The CHARSPAN noise model outperformed the unigram model. There were improvements for languages like Konkani which are lexically less similar to corresponding HRLs. We also conducted experiments where the noise was augmented before and after vocabulary preparation. We found that both experiments perform equally well; however, the model where vocabulary created with noisy data performs slightly better. Which scale the proposed model usability to applications where PLMs were involved as they usually have fixed vocab. The CHARSPAN noise model combined with BPE-Dropout emerged as the performing model. However, there is minimal degradation in HRL performance.

5.3 Insights and Limitations

We have conducted several ablation experiments to ensure that the proposed design choices result in the best performance. Furthermore, our analysis indicates that the character-span-based model enhances the performance of languages that are less similar or more distant from HRLs. Additionally, it is important to select lexically similar languages HRLs. Finally, we explore a multilingual setup in which multiple HRLs are trained together, resulting in a performance boost and scale coverage for ELRs. Our model performs equally well with a vocabulary that is learned with clean data. This provides scalability for utilizing PLMs, which typically have a fixed vocabulary.

The current work is only investigated for ELRLs to English MT tasks. We assume that the related languages also use the same script or scripts that can be easily mapped/transliterated to each other. This method might not be effective for transfer between related languages that are written in very different scripts, e.g., Hindi is written in the Devanagari script, while Sindhi is written in the Perso-Arabic script. We will extend this work to English to ELRLs MT and other tasks in the future.

6 Conclusion

With this thesis, we have presented a coherent narrative of our efforts in the field of text generation for multiple LRLs with limited supervision. We began by enabling zero-shot well-formed text generation, then progressed to improving cross-lingual generation, and ultimately enabled zero-shot machine translation for ELRLs and dialects. Our modeling approaches are aligned with adaptive training, meta-learning, language clustering, lexical similarity, and noise augmentation. The evaluations were conducted across a wide range of LRLs across language families, multiple NLG tasks, and datasets. Through these endeavors, we have taken a step towards facilitating language technology for the long tail of languages that possess limited or no linguistic resources. This advancement aims to benefit the general audiences where text needs to be generated in local languages.

In the future, we will explore the following directions: (1) Extend the existing modeling framework to cover 7000+ spoken languages of the world. (2) Design a single unified and scalable framework for many NLG tasks and LRLs. (3) Develop a better modeling approach to adapt the existing Multilingual PLM representations to new/unseen LRLs. (4) Since for many ELRLs there are no evaluation datasets, we will explore a modeling technique where the performance of LRLs is evaluated without reference. (5) Creating a large-scale multilingual NLG benchmark similar to Chen et al. (2022). (6) Investigating active learning, prompting, and other trending methodologies to advance cross-lingual transfer and generation research with limited supervision.

Acknowledgements

I (as the first author of the paper) extend my heartfelt gratitude to my Ph.D. supervisor, Dr. Maunendra Sankar Desarkar, for his unwavering guidance and support throughout my doctoral journey. I also want to acknowledge the invaluable contributions of my collaborators Rahul Kejriwal, Anoop Kunchukuttan, Yoshinobu Kano, and Kumari Deepshikha, whose expertise and collaborative efforts enriched the quality of our research. I am deeply appreciative of the support and resources provided by collaborating organizations Microsoft India, Nvidia AI Center India, and Shizuoka University Japan, which played a pivotal role in facilitating our research endeavors. I thank the dedicated human annotators for evaluation and the anonymous reviewers for their constructive feedback. This research would not have been possible without the collective efforts of these individuals and organizations, and for that, I am profoundly thankful.

References

Noëmi Aepli and Rico Sennrich. 2022. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020. Self-supervised meta-learning for few-shot natural language classification tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 522–534.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Emily M Bender. 2019. The# benderrule: On naming the languages we study and why it matters. *The Gradient*, 14.

Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier. 1990. *Learning a synaptic learning rule*. Citeseer.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. 2022. MTG: A benchmark suite for multilingual text generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2508–2527, Seattle, United States. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7570–7577.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172, Florence, Italy. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. 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 6251–6256, Hong Kong, China. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings*

of Machine Learning Research, pages 2790–2799. PMLR.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020a. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020b. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. 2021. Exploiting language relatedness for low web-resource language model adaptation: An Indic languages study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1312–1323, Online. Association for Computational Linguistics.

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. 2019. Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4863–4872, Florence, Italy. Association for Computational Linguistics.

Guokun Lai, Barlas Oguz, Yiming Yang, and Veselin Stoyanov. 2019. Bridging the domain gap in cross-lingual document classification. *CoRR*, abs/1909.07009.

Tianjian Li and Kenton Murray. 2023. Why does zero-shot cross-lingual generation fail? an explanation and a solution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12461–12476, Toronto, Canada. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. 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.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Kaushal Maurya and Maunendra Desarkar. 2022. Meta-x$_{NLG}$: A meta-learning approach based on language clustering for zero-shot cross-lingual transfer and generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 269–284, Dublin, Ireland. Association for Computational Linguistics.

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. 2021. ZmBART: An unsupervised cross-lingual transfer framework for language generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2804–2818, Online. Association for Computational Linguistics.

Kaushal Kumar Maurya, Rahul Kejriwal, Maunendra Sankar Desarkar, and Anoop Kunchukuttan. 2023. Utilizing lexical similarity to enable zero-shot machine translation for extremely low-resource languages. *arXiv preprint arXiv:2305.05214*.

Meryem M'hamdi, Doo Soon Kim, Franck Dernoncourt, Trung Bui, Xiang Ren, and Jonathan May. 2021. X-METRA-ADA: Cross-lingual meta-transfer learning adaptation to natural language understanding and question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3617–3632, Online. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020,*

Online, November 16-20, 2020, pages 4547–4562. Association for Computational Linguistics.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406, Online. Association for Computational Linguistics.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, Dublin, Ireland. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Francesco Piccinno, Massimo Nicosia, Xinyi Wang, Machel Reid, and Sebastian Ruder. 2023. mmt5: Modular multilingual pre-training solves source language hallucinations. *arXiv preprint arXiv:2305.14224*.

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.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Sebastian Ruder. 2022. The State of Multilingual AI. http://ruder.io/state-of-multilingual-ai/.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2340–2354, Dublin, Ireland. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

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: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, and Mao-song Sun. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 26(12):2319–2327.

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*.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 90–96.

Ishan Tarunesh, Sushil Khyalia, Vishwajeet Kumar, Ganesh Ramakrishnan, and Preethi Jyothi. 2021. Meta-learning for effective multi-task and multilingual modelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3600–3612. Association for Computational Linguistics.

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.

Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. 2022. Three types of incremental learning. *Nat. Mac. Intell.*, 4(12):1185–1197.

Niels van der Heijden, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2021. Multilingual and cross-lingual document classification: A meta-learning approach. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1966–1976, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming

catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926, Uppsala, Sweden. Association for Computational Linguistics.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium. Association for Computational Linguistics.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F Karlsson, Biqing Huang, and Chin-Yew Lin. 2020. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9274–9281.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.