ADAPTING PRE-TRAINED LANGUAGE MODELS TO LOW-RESOURCE TEXT SIMPLIFICATION: THE PATH MATTERS

Cristina Gârbacea
Department of EECS
University of Michigan, Ann Arbor
garbacea@umich.edu

Qiaozhu Mei
School of Information, Department of EECS
University of Michigan, Ann Arbor
qmei@umich.edu

ABSTRACT

We frame the problem of text simplification from a task and domain adaptation perspective, where neural language models are pre-trained on large-scale corpora and then adapted to new tasks in different domains through limited training examples. We investigate the performance of two popular vehicles of task and domain adaptation: meta-learning and transfer learning (in particular fine-tuning), in the context of low-resource text simplification that involves a diversity of tasks and domains. We find that when directly adapting a Web-scale pre-trained language model to low-resource text simplification tasks, fine-tuning based methods present a competitive advantage over meta-learning approaches. Surprisingly, adding an intermediate stop in the adaptation path between the source and target, an auxiliary dataset and task that allow for the decomposition of the adaptation process into multiple steps, significantly increases the performance of the target task. The performance is however sensitive to the selection and ordering of the adaptation strategy (task adaptation vs. domain adaptation) in the two steps. When such an intermediate dataset is not available, one can build a "pseudostop" using the target domain/task itself. Our extensive analysis serves as a preliminary step towards bridging these two popular paradigms of few-shot adaptive learning and towards developing more structured solutions to task/domain adaptation in a novel setting.

1 INTRODUCTION

Large-scale language models (such as those adopting the Transformer Vaswani et al. (2017) architectures) have shown outstanding text generation capabilities. This demonstrates that with sufficient data, model capacity, and computational resource, generative models can learn distributions powerful enough to produce high-quality samples from complex domains. These conditions are however unrealistic in many NLP tasks and application scenarios, where abundant training examples either do not exist or are costly to label, and the computational resource required to train large neural language models from the scratch is a luxury. Text simplification, which aims to transform specialized/complex content into simpler text so that it is accessible to readers with low literacy skills, is such an scenario. Despite its difficulty, text simplification is critical for providing fairness and equitable information services to the broad population.

The predominant approach for building NLP solutions for low-resource scenarios relies on a transfer learning paradigm, which works by first training a language model on large general-domain datasets and then adapting or fine-tuning the pre-trained model to the downstream task in a specific domain and/or with a specific objective functions. Nevertheless, such transfer learning methods usually assume the source and target domains consist of the same feature space, which limits their performance in many practical situations where the target domain is qualitatively different from the generic corpora used to train the original language models Day & Khoshgoftaar (2017). Furthermore, the effectiveness of vanilla fine-tuning methods is still heavily dependent upon having adequate amounts of in-domain training data for the target task Chen et al. (2019); when this pre-requisite is not met, the generalization performance of deep models can be considerably limited, leading to model over-fitting, catastrophic forgetting of general-domain knowledge, and negative transfer across tasks Kirkpatrick et al. (2017), Thompson et al. (2019), Xu et al. (2020).

New approaches have been proposed in the literature to address these challenging issues, claiming various degrees of success on a diversity of benchmarks Dumoulin et al. (2021), Delange et al. (2021). Among these, *meta-learning* Thrun (1998), Schmidhuber (1987), Hospedales et al. (2021) has emerged as a promising general learning strategy suitable for few-shot learning and cross-domain generalization Li et al. (2018), Wang et al. (2020b). A typical meta-learning approach frames the learning problem at two levels: *i) base learning*, where an inner/lower/base learning algorithm is focused on the quick acquisition of knowledge within each separate task it encounters, and *ii) meta-learning*, where an outer/upper/meta algorithm is focused on the slower extraction of information learned across all

tasks and updates the inner learning algorithm such that the model it learns improves an outer learning objective. To solve a few-shot learning problem, meta-learning leverages a good number of similar few-shot *tasks* to learn how to adapt the base-learner to a *new* task for which also only a few labeled samples are available. Different approaches to meta-learning include metric learning Vinyals et al. (2016), Snell et al. (2017), Sung et al. (2018), memory networks Santoro et al. (2016), Oreshkin et al. (2018), Mishra et al. (2018), Munkhdalai et al. (2018) and gradient based learning Finn et al. (2017), Zhang et al. (2018), Sun et al. (2019). Among them, gradient / optimization based meta-learning has emerged as an effective approach to addressing the few-shot learning problem Rajeswaran et al. (2019). Such learning settings lend themselves applicable to resource constrained problems where there is a distribution of tasks available.

In this work, we frame the problem of low-resource text simplification from a task and domain adaptation perspective. We consider the everyday use of text simplification in a wide variety of domains, including news and scientific articles (which naturally contain many subject areas), and view parallel complex-simple English language examples in different domains as samples drawn from a distribution over text generation tasks with varying constraints on the level of text complexity and readability. Once such a distribution is learned from large-scale, general purpose corpora (i.e., a pre-trained language model), it is fast adapted to new tasks and domains (in our case different text simplification scenarios) with few training examples. We consider two approaches to this problem: 1) a standard transfer learning practice that fine-tunes the general language model to the new domains of text simplification with limited in-domain data, and 2) simulate many domain adaptation tasks and use gradient based meta-learning to learn model parameters that can generalize to new tasks, again with few examples. We extensively compare these two approaches in our low-resource adaptation settings, and our experiments reveal that when directly adapting a general language model to the target tasks/domains, fine-tuning (i.e., domain-adaptation) remain competitive compared with meta-learning (i.e., task adaptation). Surprisingly, we find that adding an intermediate destination in between the source and target, i.e., first adapting the pre-trained language model to an auxiliary task/domain and then adapt the model to the target tasks/domains, significantly increases the performance of the target tasks. Adding a *stop* in the adaptation path allows each segment to use a different adaptation strategy (akin to a transportation method, or a "vehicle"), and the performance on target tasks is sensitive to which vehicle is taken in each segment. In particular, it is essential to perform domain adaptation through transfer learning (fine-tuning) in the second stage, and performing task adaptation via meta-learning in the first stage further improves the performance. Interestingly, when such an intermediate dataset is not available, one can build a "pseudostop" simply based on the target task/domain itself. Our findings serves as a novel step bridging the two popular paradigms of few-shot adaptive learning and towards developing more structured solutions to task/domain adaptation.

2 Related Work

The task of neural text simplification is similar in nature to neural machine translation, where transfer learning and meta-learning approaches have been widely applied to low-resource settings. Knowledge extracted from multilingual high-resource language pairs is leveraged for adapting machine translation systems to low-resource target languages, demonstrating the benefit of meta-learning over conventional multilingual translation approaches when limited in-domain training data is available Gu et al. (2018). Similarly, a meta-learning strategy is used to simulate many few-shot domain adaptation tasks to learn model parameters for fast adaptation to unseen language pairs in machine translation Sharaf et al. (2020). We consider a similar problem in a different application context.

Relevant to our work, a few-shot evaluation protocol is used to compare recent advances in transfer learning and meta-learning on standard visual classification benchmarks for task adaptation Dumoulin et al. (2021). The authors find that meta-learning approaches struggle to generalize to out-of-distribution test tasks, and that their overall performance is inferior to transfer learning methods. While pre-training then fine-tuning remains a highly competitive baseline for few-shot classification tasks, simply scaling up the size of these pre-trained models does not result in any significant performance gain on out-of-distribution tasks. On the contrary, meta-learning methods are data efficient, but computational bottlenecks and implementation difficulties prohibit their use in combination with large scale backbones. Furthermore, having sufficient heterogeneous training tasks is a critical pre-requisite for meta-model training Kang & Feng (2018); when source tasks present different characteristics from target tasks, the performance of meta-learning algorithms declines and results in poor generalization on unseen tasks. On the particular task of text classification, combining task-adaptive pre-training with domain adaptive pre-training results in performance gains Gururangan et al. (2020). This finding is in line with our work which confirms that multiple stages of adaptation result in improved performance on the end task/ domain. However, unlike our work which is focused on text simplification in a multitude of domains (32 scientific domains and the news domain), Gururangan et al. (2020) are focusing on the different task of text classification in four domains only and report that task-adaptive pre-training yields performance gains after domain adaptive pre-training; instead, our empirical results show the opposite order of adaptation is more effective.

Similarity between source and target (tasks and domains) represents an important predictive factor of successful adaptation Vu et al. (2020), while increasing the size of the source dataset does not necessarily result in the largest transfer gains. When only scarce data is available for the target task, transfer learning remains beneficial.

Many recent approaches combine the strengths of meta-learning with pre-training. Meta-parameterized pre-training Raghu et al. (2021) "meta"-learns the pre-training hyperparameters, demonstrating that optimized meta-parameters improve the learnt representations and the predictive performance of the pre-trained model. Meta-finetuning Wang et al. (2020a) improves the fine-tuning of neural language models by meta-learning class prototypes and domain-invariant representations which are useful in solving groups of similar natural language tasks. Moreover, feature representations meta-learnt are clustered more tightly in the feature space than representations obtained through conventional training of neural networks / pre-train then fine-tune approaches, demonstrating that minimizing within-class feature variation is critical for robust few-shot performance on complex tasks Goldblum et al. (2020).

3 Preliminaries to Few-shot Task/Domain Adaptation

We aim to learn how to adapt a pre-trained neural language model to text simplification contexts that involve new tasks and domains, with only few in-domain training examples available. For this purpose, we use widely popular adaptation strategies, namely gradient-based meta-learning and fine-tuning-based transfer learning. In what follows we formally introduce these two approaches for task and domain adaptation in the context of neural text simplification.

3.1 Gradient-based Meta-learning

In the context of few-shot learning, meta-learning models are designed to find parameters that can be fine-tuned in few optimization steps and with few labeled examples to achieve fast adaptation on a task not seen during training.

In typical machine learning settings, we are given a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, with a training split D_{train} and a testing split D_{test}. The goal is to train a model $\hat{y} = f_\theta(x)$ parameterized by θ such that model parameters θ are optimized on the training subset D_{train}: $\theta^* = \arg\min_\theta \mathcal{L}(\mathcal{D}; \theta, \omega)$, where $\mathcal{L}$ represents the loss function measuring the error between model predictions and ground-truth labels, and ω denotes assumptions such as the choice of the optimizer or function class for f. After training the model, we then evaluate its generalization performance on the testing subset D_{test}. The conventional assumption is that optimization is performed from scratch for every dataset D.

In meta-learning, we assume a distribution over tasks $p(\mathcal{T})$ we want our model to be able to adapt to, and that a set of tasks can be sampled from this distribution, $\{\mathcal{T}_i\}_{i=1}^n \sim p(\mathcal{T})$. Each task $\mathcal{T}_i = (\mathcal{T}_i^s, \mathcal{T}_i^q)$ consists of two small sets of labeled data, the support data $\mathcal{T}_i^s$ which is used for fine-tuning, and the query data $\mathcal{T}_i^q$ which is used for measuring the performance of the resulting fine-tuned model. Note that for different tasks $\mathcal{T}_i$ and $\mathcal{T}_j$, they may share the same data distribution (X) but just deal with different labels (Y). When the data distribution X_i and X_j are different across tasks, we can also describe them as sampled from different "domains". The task $\mathcal{T}_i$ is described as *n-way k-shot* if it consists of n classes, and there are k examples available for each class. In Algorithm 1 Goldblum et al. (2019), we present the general gradient-based meta-learning framework, noting that variations of this approach exist in the literature.

Algorithm 1 The gradient-based meta-learning framework

Require: $p(\mathcal{T})$: distribution over tasks, F_θ: base model, $\mathcal{A}$: fine-tuning algorithm, γ: learning rate
Ensure: Initialize θ, the weights of F
  while not done **do**
    Sample a batch of tasks $\{\mathcal{T}_i\}_{i=1}^n \sim p(\mathcal{T})$, where $\mathcal{T}_i = (\mathcal{T}_i^s, \mathcal{T}_i^q)$
    for $i = 1, \ldots n$ **do**
      Fine-tune model on task $\mathcal{T}_i$ and obtain new network parameters $\theta_i = \mathcal{A}(\theta, \mathcal{T}_i^s)$ ▷ Inner Loop
      Compute gradient $g_i = \nabla_\theta \mathcal{L}(F_{\theta_i}, \mathcal{T}_i^q)$
    end for
    Update base model parameters ▷ Outer Loop
    $\theta \leftarrow \theta - \frac{\gamma}{n} \sum_i g_i$
  end while

In general, meta-learning algorithms employ a bi-level optimization scheme consisting of an "inner" loop and an "outer" loop, where the outer loop searches for the best global parameter initialization and the inner loop optimizes individual models that share a common parameter initialization for a range of tasks. A meta-learning iteration starts with the outer loop, where a batch of tasks are sampled from the distribution over tasks $p(\mathcal{T})$. Then in the inner loop, given as input a base model F_θ parameterized by network parameters θ, F_θ is in turn fine-tuned on the support

data $\mathcal{T}_i^s$ of each task; the resulting fine-tuned model F_{θ_i} is used to make predictions on the query data $\mathcal{T}_i^q$ of each task. After the inner loop completes for all sampled tasks in the batch, the outer loop minimizes the loss on the query data with respect to the pre-finetuned weights; this outer optimization step is achieved by differentiating through the inner loop computation and updating the base model parameters θ such that the inner loop fine-tuning becomes as fast and efficient as possible. Importantly, meta-learning algorithms differentiate through the entire fine-tuning loop, unlike transfer learning approaches that only use simple first-order gradient information to update network parameters. Nevertheless, back-propagating meta-gradients through the inner loop comes with practical constraints such as high-order derivatives, big memory footprints, and the risk of vanishing or exploding gradients Jamal et al. (2021). To alleviate these issues of hierarchical optimization where the outer optimization is constrained on the inner optimization, the number of inner fine-tuning steps k is often set to small value, for eg., $k = 1$; alternatively, approximations for higher order gradients are used when the outer/ meta model and the inner/ task-specific model lie in the same space.

3.2 Transfer learning

Transfer learning Caruana (1994), Pan & Yang (2009) focuses on knowledge transfer across domains. It aims to improve the learning process of a target task with limited or no labeled training data by exploiting knowledge acquired from a different and related source domain/task which has sufficient data available. By enhancing the data in target domain with the additional data from the source domain, model performance on the target task can be considerably improved. Compared to meta-learning which includes an outer optimization loop to evaluate the benefit of prior knowledge when learning a new task, transfer learning extracts prior knowledge by learning on the source task directly (i.e. without the use of a meta-objective). Furthermore, while meta-learning seeks an "algorithmic" solution to the few-shot learning problem and does not necessarily focus on datasets and architectures, transfer learning approaches emphasize learning robust representations from large-scale datasets and models Dumoulin et al. (2021). To this end, one of the most commonly employed approaches to transfer learning is *pretrain-then-finetune*, which first trains a model on massive datasets and then fine-tunes a pre-trained model on new tasks of interest that requires less data.

Numerous successes of large-scale pre-trained models on a wide variety of tasks and domains are reported in the literature Vaswani et al. (2017), Brown et al. (2020), Devlin et al. (2019). Nevertheless, transferred knowledge does not always have a positive impact on new tasks. In the extremely data-scarce regime when only few samples are available in the target domain, transfer learning is less effective and performs subpar Goldblum et al. (2019). Furthermore, when there is little in common between domains or when domain similarities are misleading, the target learner is negatively impacted by the transferred knowledge and negative transfer occurs Zhuang et al. (2020). The brittleness of the fine-tuning process in settings where there is data distribution shift and different label space than seen during pre-training leads to poor out-of-domain generalization. Therefore, the main challenge in transfer learning becomes how to distinguish beneficial source knowledge from inherent cross-domain noise Day & Khoshgoftaar (2017).

Relevant to our work, including a second stage of pre-training with intermediate supervised tasks is reported to improve the robustness and effectiveness of the resulting target task model in few-shot settings Phang et al. (2018). Crucially, a careful selection of source tasks to fine-tune on in the intermediate stage is still required, which is not always clear.

4 Experiment Setup

In our experiments we aim to investigate the robustness and efficacy of meta-learning and transfer learning methods when applied to the scenario of neural text simplification in a wide diversity of real-world data-constrained settings.

4.1 Datasets

We use three datasets which cover different domains and application scenarios of text simplification. In particular, we focus on generating simpler and more readable versions of news articles and scientific papers from a multitude of research fields that are disseminated to the general public. We use a third dataset, Wikipedia, as an auxiliary.

News Simplification. *Newsela* Xu et al. (2015) is a corpus of news articles simplified by professional news editors for children of different age and grade levels for pre-college classroom use. Each article has been rewritten four times, resulting in a parallel sentence-aligned monolingual corpus with different reading levels. In our experiments we use the parallel Newsela dataset made available in Zhang & Lapata (2017), which we further divide into distinct subsets according to the ground-truth labels provided for complex-simple sentence pairs. In other words, we consider the different degrees (or difficulty level) of simplification as different *tasks* of new simplification. Table 6 in Appendix A summarizes our meta-train, meta-dev and meta-test splits according to the complexity level of sentence pairs.

Scientific Press Release. *Biendata* [1] dataset consists of research papers from various scientific disciplines matched with press releases that describe them. Notably, rewriting scientific papers into press releases has mixed objectives that are not solely about text simplification. However, it presents a valuable use-case scenario of text simplification employed in the real world and it also provides a nice out-of-distribution test for our adaptative learning methods. The corpus consists of alignments at the title level, and for each scientific paper we extract domain meta-data via the Microsoft Academic Knowledge API [2]. We define subsets of the Biendata corpus according to scientific domains, so simplifying articles in each scientific domain is also considered as a different *task*; in Table 8 in Appendix A we present statistics regarding the data splits used for the meta-train, meta-dev and meta-test subsets.

Auxiliary. *WikiLarge* Zhang & Lapata (2017) is a Wikipedia-based corpus created by combining existing simplification datasets. The training subset of WikiLarge is obtained by assembling datasets of parallel aligned Wikipedia - Simple Wikipedia sentence pairs available in the literature Kauchak (2013), Woodsend & Lapata (2011), Zhu et al. (2010), while the development and test subsets contain complex sentences with simplifications provided by Amazon Mechanical Turk workers Xu et al. (2016). We split the WikiLarge dataset at random ensuring equal number of complex-simple sentence pairs in each subset; please see Table 7 in Appendix A for details on the meta-train, meta-dev and meta-test subsets.

WikiSmall Zhu et al. (2010) is a smaller text simplification benchmark containing 89,042 training sentence pairs, and 100 testing complex-simple sentence pairs. We only use this corpus sparingly to augment WikiLarge model training.

Each dataset contains a meta-train, meta-dev and meta-test set, and each set includes text simplification tasks that correspond to varying complexity levels (Newsela and WikiLarge) or different scientific domains (Biendata). As these tasks are all based on their own dataset, we can also describe them as different "domains" of text simplification.

4.2 Adaptation Paths

Our main goal is to investigate whether meta-learning or transfer learning is a suitable adaptation strategy when there is a distribution of low-resource text simplification tasks/domains available, how they compare to each other, and whether they can work as a team. In addition, we would like to determine if doing both task adaptation and domain adaptation can improve performance on new target tasks of interest. To answer these research questions, we design the following experiments. In Figure 1, we see multiple possible paths of adaptation process.

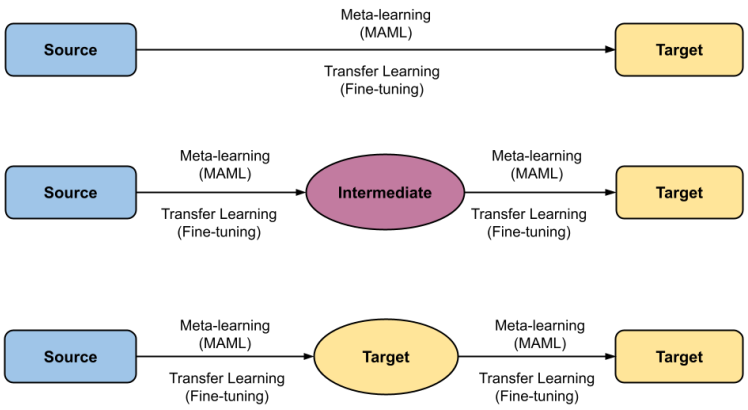

Figure 1: Adapting a pre-trained language model (source) to low-resource text simplification (target).

1. *Direct task adaptation:* we aim to determine if it is possible to adapt a pre-trained model (Source), either trained on general-domain knowledge or specifically designed for text simplification, to text simplification tasks (Target) via meta-learning and achieve good performance on unseen tasks (meta-test);
2. *Direct domain adaptation:* we would like to establish whether a pre-trained model (Source) can be adapted to text simplification domains (Target) via fine-tuning and achieve good performance on unseen domains;

[1] https://www.biendata.xyz/competition/hackathon
[2] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph

3. *Two-stage adaptation via an intermediate:* our goal is to determine if there is any benefit combining task adaptation and domain adaptation through an intermediate task/domain dataset (Intermediate). We would like to find out in which order the two-stage adaptation should be carried out;
4. *Two-stage adaptation via a pseudostop*: if no intermediate dataset is available, we would like to determine if it is possible to use the Target itself as the intermediate.

Note that in each leg of a two-stage process, we have two adaptation methods (meta-learning or transfer learning). Next, we present the specific details of the two methods in our context.

Meta-Learning (Task adaptation) For learning model parameters that facilitate adaptation to a new text simplification task in few steps and with minimal amount of text simplification examples, we use Model Agnostic Meta-Learning (MAML) Finn et al. (2017). In our experiments, we use the publicly available MAML implementation proposed in META-MT Sharaf et al. (2020) [3], originally designed for fast adaptation in the context of neural machine translation with minimal amount of in-domain data. We adapt META-MT for the task of text simplification and use the first-order approximation of MAML (FOMAML) due to computational challenges associated with computing higher order gradients. The meta-learning loss function is optimized using Adam Kingma & Ba (2014) optimizer with a learning rate $\alpha = 1e-5$ and a meta-batch size of 1. We initialize meta-parameters θ in two ways: *i)* by training a Transformer Vaswani et al. (2017) model on the combination of WikiLarge and WikiSmall text simplification datasets, and *ii)* by leveraging the external knowledge encapsulated in the pre-trained Text-to-Text Transformer (T5) Raffel et al. (2020), a general purpose language model not particularly designed for the task of text simplification. We provide details on the hyper-parameter settings for these inner models in the transfer learning section below.

In addition, we also use Reptile Nichol et al. (2018), a first-order meta-learning algorithm designed for fast adaptation to new tasks. We intend to verify whether the same conclusions and insights can be drawn if different meta-learning algorithms are used for task adaptation. Our goal is not to compare the two meta-learning algorithms.

Transfer Learning / Fine-Tuning (Domain Adaptation) For determining the benefit of transfer learning from large-scale general-domain corpora to low-resource text simplification, we use the pre-trained Text-to-Text Transformer (T5) Raffel et al. (2020). T5 is a sequence-to-sequence model with 60 million parameters pre-trained on a multi-task mixture of unsupervised and supervised tasks; each task is converted into a text-to-text format, thus allowing us to use T5 for the purpose of text simplification generation. We fine-tune T5 on the meta-train and meta-valid subsets of each text simplification dataset, then generate simplified outputs for the complex inputs from the meta-test test subset of each dataset by prompting the fine-tuned T5 with the keyphrase "translate English to English"; training and validation batch size are set to 16, and the learning rate $\alpha = 1e-4$.

In addition to T5, we also train a Transformer Vaswani et al. (2017) model for text simplification on WikiLarge and WikiSmall datasets. In our implementation of the Transformer model, we use the Fairseq Ott et al. (2019) library and following Sharaf et al. (2020), we augment the Transformer architecture with adapter modules Houlsby et al. (2019), Bapna & Firat (2019) after each transformer block for more efficient model fine-tuning. The model we train relies on the transformer-base architecture with 6 encoder and 6 decoder layers and multi-head attention with 8 attention heads, the dimensionality of word embeddings is set to 512, feed-forward layers dimension is set to 2,048, and adapter modules have 32 hidden units; we use Adam Kingma & Ba (2014) optimizer with a learning rate $\alpha = 7e-4$.

4.3 Evaluation Metrics

There is no consensus on what is the single best evaluation metric for text simplification. We therefore employ a diverse portfolio of metrics to "meta"-assess the quality of the generated simplifications from different perspectives, including informativeness, relevance, fluency, readability, and adequacy. We use *SARI* Xu et al. (2016) to evaluate the quality of the simplified output by comparing it against the source and reference simplifications, which is one of the most accepted metrics of text simplification in literature. We use *BLEU* Papineni et al. (2002) to measure the similarity between the generated text and gold standard references. We also use *FKGL* Kincaid et al. (1975) to measure the readability of the output. In addition, we also use learnable evaluation metrics that train machine learning models on human annotated datasets to learn a scoring function that reproduces human judgements. *MoverScore* Zhao et al. (2019) measures the semantic distance between system outputs and reference texts using semantically aligned pretrained embeddings; the distance is computed using Word Mover's Distance Kusner et al. (2015) in the embedding space, yielding the amount of flow traveling between the contextualized representations. *MAUVE* Pillutla et al. (2021) rewards model-generated text which resembles human-authored text by comparing the two distributions using Kullback-Leibler information divergence frontiers in a quantized low-dimensional embedding space. *BARTScore*

[3] https://www.dropbox.com/s/jguxb75utg1dmxl/meta-mt.zip?dl=0

Yuan et al. (2021) frames the evaluation of text generation models as a text generation problem, and uses BART Lewis et al. (2020) pre-trained sequence-to-sequence model to assess the probability of the system output (h) being generated from the source (s) and/ or reference text (r). We apply BARTScore in different generation directions to determine *Faithfulness* ($s \rightarrow h$) as the likelihood the hypothesis could be generated based on the source text, *Precision* ($r \rightarrow h$) as the likelihood the hypothesis could be constructed based on the gold reference, *Recall* ($h \rightarrow r$) as the likelihood a gold reference could be generated by the hypothesis, and *F1* ($r \leftrightarrow h$) as the harmonic mean of precision and recall. Because the metric computes the average log-likelihood for the target tokens, the resulting BARTScore values are negative. Each of these evaluation metrics captures different aspects of text simplification generation, therefore in our analysis we account for their overall agreement with regards to the quality of the generated simplified output.

5 Experiment Results

We first intend to find out how well pre-trained language models can be directly adapted to the task of low-resource text simplification, if it is necessary at all (than training a model directly in the low-resource setting). If yes, we would like to understand whether the pre-trained model has to be purposed for text simplification or it can be trained on large-scale general purpose text corpora. If the latter, we would like to establish whether meta-learning or transfer learning works better (via one-stage adaptation) to transfer the general knowledge to text simplification, especially to new simplification domains with scarce data. Furthermore, we would like to determine if task and domain adaptation can complement each other as part of a two-stage adaptation approach; if yes, we would like to determine in which order the two-stage adaptation process should be carried out to optimize performance on out-of-distribution text simplification tasks and domains. Finally, when a natural intermediate dataset is not possible, we would like to determine whether using the target dataset itself for intermediate adaptation is a sensible alternative.

Baseline (No Adaptation). As a baseline, we do not adapt any pre-trained language models but train a Transformer model directly on Newsela or Biendata (based on meta-train and meta-dev) and evaluate its performance on the corresponding meta-test. Because the training data is limited, we anticipate that the performance would be suboptimal. Results are presented in Table 1. While the Transformer model generally scores high in BLEU (indicating similarity to reference) and low in FKGL (indicating high readability), the rest of evaluation metrics all indicate a sub-par performance to adaptation-based methods (see below), including SARI. Note that the FKGL score on Biendata is much lower than that of the ground-truth simplifications, indicating that the model has been oversimplifying the scientific content. For reference, we also include the results on WikiLarge dataset, which are much better than the other two datasets. As WikiLarge is the largest dataset of the three, this suggests that when training data is abundant, neural text simplification could yield good performance without adaptation from a pre-trained language model.

Table 1: Baseline meta-test test set results for the Transformer model trained on each text simplification dataset. Baseline FKGL reference scores - *Newsela: 3.733, Biendata: 9.692, WikiLarge: 5.973*; $\star$ denotes over-simplification.

Dataset	**Method**	**SARI** (↑)	**BLEU** (↑)	**FKGL** (↓)	**MOVER** (↑)	**MAUVE** (↑)	**BARTScore** (↑)			
							Faithfulness	*P*	*R*	*F1*
Newsela	Transformer	38.888	10.294	4.371	0.193	0.471	-3.806	-4.770	-4.686	-4.677
	ACCESS	27.062	13.501	8.097	0.250	0.326	-1.827	-4.276	-3.511	-3.793
	DMLMTL	38.601	8.181	$\star$1.476	0.164	0.229	-4.004	-5.070	-4.767	-4.850
Biendata	Transformer	35.950	0.455	$\star$6.715	0.143	0.019	-6.164	-5.924	-6.517	-6.153
	ACCESS	17.847	2.582	12.552	0.294	0.446	-1.747	-5.821	-5.654	-5.678
	DMLMTL	32.960	0.893	$\star$9.180	0.154	0.053	-4.429	-6.655	-6.280	-6.393
WikiLarge	Transformer	47.361	41.557	$\star$5.520	0.375	0.818	-2.380	-3.408	-3.687	-3.469
	ACCESS	36.075	32.577	8.536	0.356	0.325	-2.016	-3.864	-3.749	-3.734
	DMLMTL	31.233	3.425	$\star$1.650	0.113	0.034	-3.731	-4.871	-5.487	-5.072

Additional Baselines (Pre-trained Text Simplification Models) In addition, we also select the current best text simplification models from the literature which have released pre-trained models. ACCESS Martin et al. (2020) is a controllable sequence-to-sequence simplification model reported highest performance on WikiLarge, while Dynamic Multi-Level Multi-Task Learning for Sentence Simplification (DMLMTL) Guo et al. (2018) reported the highest performance on Newsela. We evaluate these pre-trained text simplification models on our own data splits. The results

on Newsela and WikiLarge are mostly consistent with literature. However, the performance of both pretrained models degrades significantly on Biendata, which demonstrates the critical need for task/domain adaptation. In fact, the Transformer model directly trained on each dataset outperforms both pretrained models on most metrics.

Direct task adaptation. In Table 2 we present results for adapting a pre-trained language model to the target task (Newsela or Biendata) through MAML. We test two source language models: one is the general-purpose T5 model released by Google and the other is the Transformer model we trained on WikiLarge (Wiki), which is purposed for text simplification. We observe that using T5 for MAML meta-parameter initialization yields better performance on new text simplification tasks according to the majority of evaluation metrics. The Transformer model trained on WikiLarge, although already purposed for text simplification, only achieves a higher BLEU and a better FKGL. Overall, adapting from a powerful pre-trained language model outperforms training a model directly from the limited resource (Table 1). To test the robustness of the results, we replace MAML with Reptile and the same pattern is observed (Table 2). These is not a consensus among the metrics whether MAML or Reptile is better in this task - note that our goal is not to compare the two meta learning models but rather the paths of task/domain adaptation.

Table 2: Direct task adaptation results on Newsela and Biendata meta-test test set; we use meta-learning (MAML, Reptile) for adapting an existing language model to new tasks. ⋆ denotes over-simplification according to FKGL.

Dataset	**Method**	**SARI (↑)**	**BLEU (↑)**	**FKGL (↓)**	**MOVER (↑)**	**MAUVE (↑)**	**BARTScore (↑)**			
							Faithfulness	*P*	*R*	*F1*
Newsela	MAML T5	25.343	**16.038**	8.096	**0.276**	**0.785**	**-1.195**	**-3.820**	**-3.208**	**-3.419**
	MAML Wiki	**36.025**	12.310	**5.184**	0.204	0.172	-3.257	-4.731	-5.014	-4.784
	Reptile T5	22.758	16.264	9.133	0.277	0.773	-1.068	-3.800	-3.121	-3.359
	Reptile Wiki	20.391	14.892	9.171	0.267	0.624	-1.114	-3.808	-3.081	-3.339
Biendata	MAML T5	25.804	**1.563**	16.348	**0.178**	**0.149**	**-2.558**	**-5.702**	**-5.755**	**-5.632**
	MAML Wiki	**35.548**	0.240	⋆**7.243**	0.122	0.004	-6.527	-6.584	-6.708	-6.586
	Reptile T5	18.867	1.587	16.916	0.166	0.081	-1.872	-5.429	-5.712	-5.434
	Reptile Wiki	10.004	1.384	18.876	0.137	0.027	-1.524	-4.980	-5.818	-5.262

Direct domain adaptation. In Table 3 we present results for adapting the pre-trained language models (T5 and Wiki) through fine-tuning to new text simplification domains in Newsela and Biendata meta-test test set. In line with our previous findings, using T5 as source yields superior performance to the Transformer trained on WikiLarge, according to most metrics. Comparing one-stage domain adaptation (Table 3) with one-stage task adaptation (Table 2), we observe that by and large domain adaptation (fine-tuning) outperforms task adaptation (either MAML or Reptile); the benefit is more apparent on the out-of distribution scientific press release tasks and domains on Biendata. As using T5 as the source dominates the pre-trained Transformer on WikiLarge, we use T5 as the Source in there after.

Table 3: Direct domain adaptation results on Newsela and Biendata meta-test test set; we use fine-tuning for adapting an existing language model to new domains. ⋆ denotes over-simplification according to FKGL.

Dataset	**Method**	**SARI (↑)**	**BLEU (↑)**	**FKGL (↓)**	**MOVER (↑)**	**MAUVE (↑)**	**BARTScore (↑)**			
							Faithfulness	*P*	*R*	*F1*
Newsela	Fine-tune T5	32.310	**19.547**	7.638	**0.298**	**0.453**	**-1.624**	**-3.874**	**-3.162**	**-3.415**
	Fine-tune Wiki	**35.360**	18.009	**4.397**	0.272	0.158	-2.603	-4.847	-4.677	-4.676
Biendata	Fine-tune T5	35.989	**3.314**	11.279	**0.240**	**0.659**	**-3.324**	**-5.319**	**-5.519**	**-5.352**
	Fine-tune Wiki	**37.314**	1.066	⋆**6.827**	0.177	0.004	-5.652	-6.094	-6.193	-6.068

Two-stage adaptation. Next, we would like to investigate whether it is possible to combine the advantage of adapting to new domains with adapting to new tasks for more robust performance and better generalization on new text simplification tasks and domains. As part of a two-stage adaptation process, we aim to determine the ideal order in which to perform the adaptation, i.e. whether task adaptation should be performed ahead of domain adaptation, or vice

versa. In addition, we explore if consecutive stages of task adaptation and domain adaptation could further improve upon one-stage adaptation results; nevertheless, our expectation is that combining task with domain adaptation adds complementary benefits and can outperform multiple stages of the same type of adaptation. In our analysis, we differentiate two cases: *i)* when there is an intermediate text simplification dataset available, and *ii)* when no other dataset is available, except for the source and target datasets.

Intermediate dataset available. We use WikiLarge as an intermediate dataset for task and domain adaptation. As part of the two-stage adaptation process, we first adapt pre-trained T5 to WikiLarge, then continue to adapt the resulting model to Newsela or Biendata; we explore possible combinations of task and domain adaptation at each stage of the pipeline, and present results for various combinations of the two-stage adaptation process on the Newsela and Biendata target tasks and domains in Table 4. Additionally, in Table 9 we include intermediate adaptation results on WikiLarge.

Table 4: Two-stage adaptation results on Newsela and Biendata meta-test test set when an intermediate dataset (WikiLarge) is used. **BOLD** and Underlined: best and second best within the block of either MAML or Reptile). ⋆ denotes over-simplification according to FKGL. Best single stage adaptation results under each metric included for reference but not highlighted in comparison. (*: results duplicated for comparison purposes.)

Dataset	Method	SARI (↑)	BLEU (↑)	FKGL (↓)	MOVER (↑)	MAUVE (↑)	BARTScore (↑)			
							Faithfulness	*P*	*R*	*F1*
Newsela	T5 domain + domain	**34.722**	19.376	8.319	0.284	0.079	-2.081	-4.182	-3.244	-3.596
	T5 domain + task (MAML)	24.681	16.971	8.415	0.283	0.834	**-1.102**	-3.771	-3.133	-3.354
	T5 task (MAML) + task (MAML)	28.711	15.378	7.462	0.267	0.619	-1.405	-3.821	-3.277	-3.457
	T5 task (MAML) + domain	33.978	**20.876**	**7.398**	**0.312**	**0.858**	-1.226	**-3.592**	**-3.096**	**-3.259**
	T5 domain + domain*	**34.722**	19.376	8.319	0.284	0.079	-2.081	-4.182	-3.244	-3.596
	T5 domain + task (Reptile)	23.467	17.042	9.442	0.285	0.787	**-1.037**	-3.772	**-3.095**	-3.332
	T5 task (Reptile) + task (Reptile)	26.337	17.098	8.727	0.278	0.818	-1.186	-3.829	-3.207	-3.432
	T5 task (Reptile) + domain	34.092	**20.861**	**8.054**	**0.311**	**0.876**	-1.237	**-3.613**	-3.098	**-3.270**
	best single stage	36.025	19.547	4.397	0.298	0.785	-1.195	-3.820	-3.162	-3.415
Biendata	T5 domain + domain	**37.850**	3.342	**10.932**	0.230	0.580	-3.700	**-5.244**	**-5.529**	**-5.315**
	T5 domain + task (MAML)	23.059	2.479	14.939	0.200	0.170	**-2.161**	-5.811	-5.671	-5.671
	T5 task (MAML) + task (MAML)	27.300	1.056	16.545	0.142	0.076	-3.252	-5.658	-5.907	-5.683
	T5 task (MAML) + domain	36.129	**3.419**	11.386	**0.236**	**0.596**	-3.270	-5.307	-5.530	-5.348
	T5 domain + domain*	**37.850**	3.342	**10.932**	0.230	0.580	-3.700	**-5.244**	**-5.529**	**-5.315**
	T5 domain + task (Reptile)	18.407	3.366	14.959	0.225	0.181	**-1.348**	-5.718	-5.532	-5.563
	T5 task (Reptile) + task (Reptile)	22.274	1.982	15.043	0.182	0.118	-2.168	-5.684	-5.771	-5.651
	T5 task (Reptile) + domain	37.175	**3.598**	11.063	**0.238**	**0.626**	-3.418	-5.263	-5.561	-5.341
	best single stage	37.314	3.314	⋆ 6.827	0.240	0.659	-2.558	-5.319	-5.519	-5.352

When an intermediate text simplification dataset is available as part of the two-stage adaptation process, our results indicate that the most promising strategy is to adapt to new tasks (through MAML or Reptile) in the first stage, and continue adapting to new domains (through fine-tuning) in the second stage. The benefit of doing task adaptation first is also supported by the intermediate results on WikiLarge, where adapting the pre-trained T5 model to the new task of text simplification yields better results than adapting to new domains. It is critical to do domain adaptation in the final stage (*domain + domain* is only second to *task + domain*), suggesting that the difference over data distributions is more critical than the difference over tasks in our scenario. This is particularly true on Biendata, where the content in different scientific domains may be very different. Repeating the same type of (task/domain) adaptation in both stages is less effective than *task + domain*, demonstrating the complementary benefit of learning to adapt to both tasks and domains for more robust generalization. Compared to one-stage task and one-stage domain adaptation, a two-stage task and domain adaptation consistently improves the quality of the generated simplifications according to the great majority of evaluation metrics. The findings are consistent when either MAML or Reptile is used for task adaptation.

No intermediate dataset available. While we have established the advantage of a two-stage adaptation procedure to address new text simplification tasks and domains, a potential limitation of this approach is the reliance on a third text simplification dataset as the intermediate. Given the scarcity of labels, we cannot assume the existence of such a related intermediate dataset for adaptation is always guaranteed. In such cases, we investigate whether it is possible to circumvent this additional requirement by using the source or target dataset itself for intermediate adaptation in the two-stage pipeline. In our experiments, we pick the Target dataset for intermediate adaptation, since the simplification model we aim to learn needs to be tailored specifically to target tasks and domains, and also we do not have access to the original dataset that T5 is trained upon. In Table 5 we present results when we adapt from *(Source → Target) → Target*, i.e. from *(pre-trained T5 → Newsela/ Biendata) → Newsela/ Biendata* tasks and domains via a two-

stage adaptation process. In general, task adaptation followed by domain adaptation remains the best performing adaptation strategy, and the benefit is considerably more pronounced on the out-of-distribution tasks and domains from Biendata. We compare these results with Table 4 where an intermediate dataset is used, and observe that using the target dataset directly for intermediate adaptation yields slightly lower but comparable results to relying on Wikipedia for intermediate adaptation, and therefore successfully overcomes the need for extra data. The findings are mostly consistent when either MAML or Reptile is used for task adaptation. For interesting readers, we have included sample outputs of both one-stage and two-stage adaptation paths in Table 10 in Appendix.

Table 5: Two-stage adaptation results on Newsela and Biendata when no intermediate dataset is available and the target dataset is used as a pseudo-intermediate. **BOLD**: best result within block of either MAML and Reptile
.

Dataset	Method	SARI (↑)	BLEU (↑)	FKGL (↓)	MOVER (↑)	MAUVE (↑)	BARTScore (↑)			
							Faithfulness	*P*	*R*	*F1*
Newsela	T5 task (MAML) + domain	**34.690**	**19.799**	**7.601**	0.294	0.251	-1.844	-4.012	-3.183	-3.487
	T5 domain + task (MAML)	30.878	19.141	7.732	**0.299**	**0.882**	**-1.323**	**-3.685**	**-3.136**	**-3.324**
	T5 task (Reptile) + domain	**35.865**	**21.788**	**7.768**	**0.317**	0.794	-1.286	**-3.589**	**-3.112**	**-3.268**
	T5 domain + task (Reptile)	30.501	19.478	8.362	0.301	**0.883**	**-1.172**	-3.689	-3.120	-3.316
Biendata	T5 task (MAML) + domain	**37.945**	**3.354**	**10.681**	**0.238**	**0.634**	-3.735	**-5.312**	-5.511	**-5.343**
	T5 domain + task (MAML)	32.821	2.973	13.582	0.233	0.467	**-3.149**	-5.763	**-5.508**	-5.564
	T5 task (Reptile) + domain	**37.175**	**3.599**	**11.063**	0.238	**0.626**	-3.417	**-5.263**	-5.561	**-5.341**
	T5 domain + task (Reptile)	31.885	3.441	13.853	**0.243**	0.464	**-2.646**	-5.703	**-5.431**	-5.502

6 Conclusion and Future Work

In this work, we frame the problem of low-resource text simplification from a task and domain adaptation perspective and learn how to quickly adapt pre-trained language models to new tasks and domains with few training examples. We examine the performance of state-of-the-art gradient-based meta-learning for task adaptation, and transfer learning from large-scale pre-trained language models for domain adaptation in a variety of tasks and domains. Our analysis reveals that when a direct adaptation approach is used, fine-tuning pre-trained language models outperforms meta-learning models for the task of low-resource text simplification; this trend is in line with previous findings in the literature Dumoulin et al. (2021), Brown et al. (2020). Nevertheless, decomposing the adaptation process into multiple steps can significantly increase target performance, provided that an auxiliary dataset is available for intermediate adaptation and careful attention is paid to performing adaptation in the correct order, i.e. task adaptation ahead of domain adaptation. When such an intermediate auxiliary dataset is not readily available, a "pseudostop" based on the target task/domain itself can be build between the source and the target.

Our findings represent preliminary foundations for proposing adaptation models that simultaneously perform task and domain adaptation in one goal. As we observe, the coupling of task adaptation (difference in Y) and domain adaptation (difference in X) is clearly beneficial comparing to either meta-learning or transfer-learning alone. Therefore creating a model that explicitly and jointly handles these two situations is a promising direction to explore. The utilization of stops (and even pseudostops) between the source and target tasks/domains also suggests that it may be valuable to further investigate a more structured solution of task/domain adaptation. We hope our insights will help inform future directions towards robust adaptation of neural language models to new tasks and domains for few-shot text simplification and other low-resource NLP tasks.

Acknowledgement

The authors thank the anonymous reviewers for their constructive comments. This work was in part supported by the National Science Foundation under grant number 1633370.

References

Ankur Bapna and Orhan Firat. 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)*, pp. 1538–1548, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Rich Caruana. Learning many related tasks at the same time with backpropagation. *Advances in neural information processing systems*, 7, 1994.

Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Oscar Day and Taghi M Khoshgoftaar. A survey on heterogeneous transfer learning. *Journal of Big Data*, 4(1):1–42, 2017.

Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, 2019.

Vincent Dumoulin, Neil Houlsby, Utku Evci, Xiaohua Zhai, Ross Goroshin, Sylvain Gelly, and Hugo Larochelle. Comparing transfer and meta learning approaches on a unified few-shot classification benchmark. *arXiv preprint arXiv:2104.02638*, 2021.

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

Micah Goldblum, Liam Fowl, and Tom Goldstein. Robust few-shot learning with adversarially queried meta-learners. 2019.

Micah Goldblum, Steven Reich, Liam Fowl, Renkun Ni, Valeriia Cherepanova, and Tom Goldstein. Unraveling meta-learning: Understanding feature representations for few-shot tasks. In *International Conference on Machine Learning*, pp. 3607–3616. PMLR, 2020.

Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3622–3631, 2018.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 462–476, 2018.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, 2020.

Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Muhammad Abdullah Jamal, Liqiang Wang, and Boqing Gong. A lazy approach to long-horizon gradient-based meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6577–6586, 2021.

Bingyi Kang and Jiashi Feng. Transferable meta learning across domains. In *UAI*, pp. 177–187, 2018.

David Kauchak. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1537–1546, 2013.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pp. 957–966. PMLR, 2015.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Louis Martin, Éric Villemonte de la Clergerie, Benoît Sagot, and Antoine Bordes. Controllable sentence simplification. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 4689–4698, 2020.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, 2018.

Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler. Rapid adaptation with conditionally shifted neurons. In *International Conference on Machine Learning*, pp. 3664–3673. PMLR, 2018.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31, 2018.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL-HLT (Demonstrations)*, 2019.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

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

Jason Phang, Thibault Févry, and Samuel R Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*, 2018.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Aniruddh Raghu, Jonathan Lorraine, Simon Kornblith, Matthew McDermott, and David K Duvenaud. Meta-learning to improve pre-training. *Advances in Neural Information Processing Systems*, 34, 2021.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pp. 1842–1850. PMLR, 2016.

Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

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

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 403–412, 2019.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1199–1208, 2018.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2062–2068, 2019.

Sebastian Thrun. Lifelong learning algorithms. In *Learning to learn*, pp. 181–209. Springer, 1998.

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

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. Exploring and predicting transferability across nlp tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7882–7926, 2020.

Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. Meta fine-tuning neural language models for multi-domain text mining. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3094–3104, 2020a.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020b.

Kristian Woodsend and Mirella Lapata. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 409–420, 2011.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297, 2015.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016.

Ying Xu, Xu Zhong, Antonio Jose Jimeno Yepes, and Jey Han Lau. Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34, 2021.

Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. *Advances in neural information processing systems*, 31, 2018.

Xingxing Zhang and Mirella Lapata. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 584–594, 2017.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. 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)*, pp. 563–578, 2019.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 1353–1361, 2010.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

A APPENDIX

Table 6: Newsela splits according to complexity level. 0 denotes the most complex level, and 4 represents the simplest.

Complexity level	Sentence pairs	TRAIN (70%)	DEV (15%)	TEST (15%)	
0 - 1	16,611	11,627	2,492	2,492	**META-TRAIN**
0 - 2	20,122	14,086	3,018	3,018	
0 - 3	19,891	13,923	2,984	2,984	
1 - 2	12,888	9,022	1,933	1,933	
1 - 3	13,296	9,308	1,994	1,994	
2 - 3	12,146	8,502	1,822	1,822	
2 - 4	9,780	6,846	1,467	1,467	**META-DEV**
3 - 4	10,185	7,129	1,528	1,528	
0 - 4	16,086	11,260	2,413	2,413	**META-TEST**
1 - 4	10,577	7,403	1,587	1,587	

Table 7: WikiLarge random splits.

Subset	Sentence pairs	TRAIN (50%)	DEV (25%)	TEST (25%)	
Wikipedia 0	20,000	10,000	5,000	5,000	**META-TRAIN**
Wikipedia 1	20,000	10,000	5,000	5,000	
Wikipedia 2	20,000	10,000	5,000	5,000	
Wikipedia 3	20,000	10,000	5,000	5,000	
Wikipedia 4	20,000	10,000	5,000	5,000	
Wikipedia 5	20,000	10,000	5,000	5,000	
Wikipedia 6	20,000	10,000	5,000	5,000	
Wikipedia 7	20,000	10,000	5,000	5,000	
Wikipedia 8	20,000	10,000	5,000	5,000	
Wikipedia 9	20,000	10,000	5,000	5,000	
Wikipedia 10	20,000	10,000	5,000	5,000	**META-DEV**
Wikipedia 11	20,000	10,000	5,000	5,000	
Wikipedia 12	20,000	10,000	5,000	5,000	
Wikipedia 13	20,000	10,000	5,000	5,000	
Wikipedia 14	20,000	10,000	5,000	5,000	
Wikipedia 15	20,000	10,000	5,000	5,000	**META-TEST**
Wikipedia 16	20,000	10,000	5,000	5,000	
Wikipedia 17	20,000	10,000	5,000	5,000	
Wikipedia 18	20,000	10,000	5,000	5,000	

Table 8: Biendata splits according to scientific domain.

Scientific Domain	Sentence pairs	TRAIN (50%)	DEV (25%)	TEST (25%)	
Medicine	7,993	3,997	1,998	1,998	**META-TRAIN**
Biology	10,040	5,020	2,510	2,510	
Internal Medicine	1,095	547	274	274	
Psychology	3,367	1,683	842	842	
Chemistry	1,516	758	379	379	
Cancer Research	1,044	522	261	261	
Neuroscience	1,411	705	353	353	
Virology	1,106	554	276	276	
Pediatrics	812	406	203	203	
Disease	582	292	145	145	
Immunology	2,281	1,141	570	570	**META-DEV**
Genetics	2,151	1,075	538	538	
Social Psychology	1,090	546	272	272	
Surgery	1,261	631	315	315	
Psychiatry	1,045	523	261	261	
Cognition	662	330	166	166	
Demography	992	496	248	248	
Climate Change	847	423	212	212	
Zoology	645	323	161	161	
Endocrinology	1,582	790	396	396	**META-TEST**
Cell Biology	2,154	1,076	539	539	
Molecular Biology	904	452	226	226	
Biochemistry	640	320	160	160	
Physical Therapy	1,189	595	297	297	
Nanotechnology	378	188	95	95	
Gerontology	649	325	162	162	
Computer Science	739	369	185	185	
Physics	1,108	554	277	277	
Materials Science	967	483	242	242	
Ecology	2,869	1,435	717	717	
Geography	658	330	164	164	
Economics	384	192	96	96	

Table 9: Intermediate results on WikiLarge meta-test test set as part of the two-stage adaptation process. Baseline FKGL score for WikiLarge reference sentences: 5.973, $\star$ denotes over-simplification.

Dataset	Method	SARI (↑)	BLEU (↑)	FKGL (↓)	MOVER (↑)	MAUVE (↑)	BARTScore (↑)			
							Faithfulness	*P*	*R*	*F1*
WikiLarge	T5 Task Adaptation	**32.954**	34.722	6.584	**0.344**	**0.324**	**-1.286**	**-3.357**	**-3.411**	**-3.281**
	T5 Domain Adaptation	29.276	**42.608**	$\star$**5.175**	0.319	0.161	-1.816	-3.589	-3.932	-3.638

Table 10: Sample text simplification outputs through different adaptation paths.

Dataset	Model	Output
Newsela	**Complex sentence**	The exploration doubled the estimated gold reserves at El Dorado, within the headwaters of the Lempa River, the nation's most important waterway.
	Transformer	It found that the gold of the nation 's most important event , the nation 's most important event .
	Direct Task Adaptation	El Dorado is within the headwaters of the Lempa River, the nation's most important waterway.
	Direct Domain Adaptation	El Dorado, the nation's most important waterway, is within the headwaters of the Lempa River.
	Two-stage Task + Domain Adaptation	El Dorado is within the headwaters of the Lempa River, the nation's most important waterway.
	Simple sentence	The dig for gold would take place at a mine known as El Dorado.
	Complex sentence	Here , the story of Ebola is one of worldwide public health infrastructure deficiency , delicate trust in post-conflict nations , an exhausted health care workforce and widespread ambivalence to our duty as a global community .
	Transformer	Here , the story of Ebola is one of the public worldwide worldwide worldwide worldwide .
	Direct Task Adaptation	Here, the story of Ebola is one of worldwide public health infrastructure deficiency, delicate trust in post-conflict nations, an exhausted health care workforce and widespread ambivalence to our duty as a global community .
	Direct Domain Adaptation	The story of Ebola is one of the biggest problems in public health.
	Two-stage Task + Domain Adaptation	There is a lack of trust in post-conflict nations.
	Simple sentence	Ebola 's spread is about a worldwide lack of systems that support public health such as health care workers , hospitals and equipment .
Biendata	**Complex sentence**	The One-Two Punch of Alcoholism: Role of Central Amygdala Dynorphins/Kappa-Opioid Receptors
	Transformer	Researchers identify new way to reduce risk of childhood cancer
	Direct Task Adaptation	Alcoholism One-Two Punch of Alcoholism: Role of central Amygdala Dynorphins/Kappa-Opioid Receptors
	Direct Domain Adaptation	Alcoholism: How do we react to alcohol?
	Two-stage Task + Domain Adaptation	Alcoholism: How do we manage alcohol?
	Simple sentence	Alcoholism treatment: Kappa opioid receptors a new target
	Complex sentence	Development of an enhanced human gastrointestinal epithelial culture system to facilitate patient-based assays
	Transformer	Scientists discover new insights into how to fight against cancer
	One-stage Task Adaptation	Development of human gastrointestinal epithelial culture system to facilitate patient-based assayss
	Direct Domain Adaptation	Human gastrointestinal epithelial culture system could help treat patients
	Two-stage Task + Domain Adaptation	New way to test for gastrointestinal disorders
	Simple sentence	Growing human GI cells may lead to personalized treatments