Self-Supervised Augmentation and Generation for Multi-lingual Text Advertisements at Bing

Xiaoyu Kou
Microsoft Corporation
Beijing, China
xiaoyukou@microsoft.com

Tianqi Zhao
Microsoft Corporation
Beijing, China
tiazhao@microsoft.com

Fan Zhang
Microsoft Corporation
Beijing, China
fanzh@microsoft.com

Song Li
Microsoft Corporation
Beijing, China
sonli@microsoft.com

Qi Zhang
Microsoft Corporation
Beijing, China
zhang.qi@microsoft.com

ABSTRACT

Multi-lingual text advertisement generation is a critical task for international companies, such as Microsoft. Due to the lack of training data, scaling out text advertisements generation to low-resource languages is a grand challenge in the real industry setting. Although some methods transfer knowledge from rich-resource languages to low-resource languages through a pre-trained multi-lingual language model, they fail in balancing the transferability from the source language and the smooth expression in target languages. In this paper, we propose a unified Self-Supervised Augmentation and Generation (SAG) architecture to handle the multi-lingual text advertisements generation task in a real production scenario. To alleviate the problem of data scarcity, we employ multiple data augmentation strategies to synthesize training data in target languages. Moreover, a self-supervised adaptive filtering structure is developed to alleviate the impact of the noise in the augmented data. The new state-of-the-art results on a well-known benchmark verify the effectiveness and generalizability of our proposed framework, and deployment in Microsoft Bing demonstrates the superior performance of our method.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; Information extraction; • **Information systems** → *Computational advertising*.

KEYWORDS

Advertisements Generation, Multi-lingual Language Models, Self-supervised Learning, Low-resource Languages

KDD '22, August 14–18, 2022, Washington, DC, USA
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539091

ACM Reference Format:

Xiaoyu Kou, Tianqi Zhao, Fan Zhang, Song Li, and Qi Zhang. 2022. Self-Supervised Augmentation and Generation for Multi-lingual Text Advertisements at Bing. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3534678.3539091

1 INTRODUCTION

Text advertisement consists of a concise, action-oriented ad copy and a link to the advertiser's website. Advertisements are widely used to make revenue for industrial products such as Microsoft Bing, Facebook, and Instagram, and a high-quality ad can bring a high click-through rate (CTR). Manually designing numerous advertisements is economically inefficient and impractical for large businesses with thousands of products. Especially for international companies, their products are deployed and maintained across hundreds of regions with a large number of languages. For example, the Responsive Search Ads (RSAs)[1] channel of Microsoft Bing has advertisers in 32 different languages, and more than 60% of its non-English pages do not have sufficient text advertisements. Accordingly, high-quality multi-lingual text advertisement generation algorithms can benefit international advertisers by reducing their efforts of manually authorizing ads and increasing the attractiveness of authorized ads across regions and languages.

Recently, deep neural networks (DNN) achieve great performance in text advertisement generation tasks, but such technique will encounter practical challenges when applied to real ads generation products with multi-lingual requirements. DNN requires large amounts of training data to achieve good performance, but it is difficult for companies to collect multi-lingual high-quality ad texts as training data, especially for the low-resource languages with few human annotated ads. On the other hand, ad texts continuously evolve along with the evolution of ad products, e.g., the changes of ad schema by adding, merging, and removing certain ad types

[1] https://docs.microsoft.com/en-us/advertising/guides/responsive-search-ads

and formats. Such schema changes could make the existing training data out-of-date and lead to extra costs in adjusting or even relabeling training data to comply with the new schema. In real industrial scenarios, especially considering multiple languages, the frequent change of data and schema leads to expensive human efforts to judge the quality of generated ads. It is also extremely time-consuming to control the labeling quality in low-resource languages. Therefore, scaling out text advertisement generation to a large number of languages remains a grand challenge to the industry.

To tackle the above issues, multi-lingual text advertisement generation focuses on transferring knowledge from rich-resource languages to low-resource languages. Many pre-trained multi-lingual language models have learned a unified representation and shown good transferability in a wide range of language understanding as well as generation tasks (such as mBART [15] and ProphetNet [19]). Fine-tuning these multi-lingual language models on the English dataset could partially retain the prediction performances for other languages, but these models fail in transferring the format structures of text advertisements from different languages. They often generate sub-spans of the passage, which happens frequently for languages, such as Russian, which have few token overlaps with English. Besides, the code-mixing problem also makes the predicted ad text incoherent, where English words co-occur with words from the target language. An intuitive solution is to pick a proper stop step in the fine-tuning process to balance the transferability from English and the smooth expression in the target language. However, the challenge remains that it is commonly impracticable to determine such a balanced stop step in the real product scenario.

Several studies [13, 17] show that the translated data can substantially enhance the transferability during fine-tuning the pre-trained multi-lingual language models in classification and generation tasks. Other approaches [7, 18] alleviate the data scarcity problem by automatically labeling domain-specific data or generating additional data. Despite that these approaches produce plenty of training data, the ad text generation models still suffer from some critical limitations when using these synthetic data. First, the translation process and generation process may introduce noise and errors, and evaluating the quality of both augmented documentation and ad text are difficult. Secondly, the translated data may not be diverse enough compared to the real target data, and it might introduce some translation artifacts in the data distribution [2]. Finally, generating synthetic training data across languages further adds challenges to the robustness of the generation models. Although some noise filtration methods have been proposed [1, 21], most of them are designed complicatedly by employing an extra filter model or specifically only using a single data augmentation method. It is still challenging to differentiate noisy instances from useful ones in the multi-lingual ad generation task.

In this paper, we propose a unified Self-Supervised Augmentation and Generation (SAG) model to handle the multi-lingual text advertisement generation task, which has been practiced in the Microsoft Bing product team. Our ad generation model is constructed based on the real industrial product scenario, in which it is feasible to collect large amounts of high-quality English training data, while too expensive to label ad text in other languages. We employ three data augmentation strategies to enhance the dataset of the multi-lingual ad generation task: *machine translation, denoising generation, and alternative words substitution*. First, given a large amount of English ad generation data, we utilize a high-quality machine translation model to obtain the dataset in the target language. Then we enhance the diversity by constructing a denoising generation model [15] to generate additional data in the target language. Finally, we apply a masked language model XLM-Roberta [16] to alternatively substitute some words as a supplementary for the data augmentation process.

Furthermore, a self-supervised filtration procedure is developed to alleviate the impact of the noise in the augmentation data and adaptively filter out low-quality examples. To be specific, we construct our ad text generation model based on a pre-trained multi-lingual language model, and fine-tune it using the source language data and augmented target language data. During fine-tuning, we periodically evaluate the augmented data automatically using the trained model to obtain a high-quality subset of the dataset, which is applied as the augmented data to train our model in the next period. In this procedure, the source language data play a central role to teach the model to generate ad text from the document, and the adaptively selected data in the target language make the generated text smooth and grammatically correct. Under such a self-supervised strategy, our proposed SAG gradually reduces the noise and produces high-quality augmented data.

Extensive experimental results show that SAG achieves new state-of-the-art performances on the multi-lingual generation benchmark (XGLUE NTG [14]) and demonstrate the effectiveness and genralizability of our proposed framework. Our model has been deployed in Microsoft Bing, a major search engine with more than 100 million monthly active users [24]. We have evaluate how the method impacts real users by conducting human evaluation and online experiments. The human evaluation shows that the quality of the multi-lingual ads is improved compared with the baselines. Online experiments demonstrate that SAG can significantly increase global coverage and revenue.

The main contributions can be summarized as follows:

- We propose a Self-Supervised Augmentation and Generation (SAG) model to generate high-quality multi-lingual text advertisements. Although we focus on ad generation in this paper, the framework can be generalized to many text generation tasks, e.g., news headline generation and question generation.
- We employ three approaches to augment low-resource language data, and propose an adaptive filtering process to identify high-quality examples as the model evolves. Therefore the quality of text ads generated by our model is insensitive to the noise of the augmented data.
- Our model establishes the new SOTA performance on the NTG benchmark dataset, and has been deployed to the Microsoft Bing product. Automatic offline evaluations, human evaluation, and online experimental metrics demonstrate the superior performance of our method.

2 RELATED WORK

In this section, we briefly review some closely related studies, including text advertisement generation, multi-lingual pre-trained language models, and data augmentation.

2.1 Text Advertisement Generation

Text advertisement generation is a major generation task in online search engines for global companies, by which advertisers are no longer required to fill in a template for each page. Traditional works [6, 22] on text ad generation rely on filling in predefined templates with key phrases or sub-words extracted from documents to generate readable sentences, which can be used in any language. However, predefined templates are often too rigid, lacking diversity, and hard to attract users. Recently, Hughes et al. [10] propose a method using LSTM [8] and attention mechanisms to encode product landing pages [10], which achieve a good success in the automatic generation of text advertisements. However, more than 15% of the text ads it generates fail to meet the high-quality standard in the product. Wang et al. [24] show that deep reinforcement learning can further enhance the ad quality and attractiveness. Notably, our approach is orthogonal to these advertisement generation approaches which are designed for the English domain, and any strategy can be leveraged in our fine-tuning process, such as the reinforcement learning method of Wang et al. [24].

2.2 Multi-lingual Pre-trained Models

Recently, multi-lingual pre-trained language models have achieved great success in a wide range of natural language processing (NLP) tasks [3, 15, 19]. To address the challenge of multi-lingual transfer, the pre-trained language models are fine-tuned by the English training data and then directly apply to target languages. Taking advantage of large-scale unsupervised pre-training, the most relevant generation models like ProphetNet [19], Unicoder [9], mBART [15] and mT5 [25], have achieve prominent results in multi-lingual summarization, question generation, sentiment analysis, etc. However, the performance in target languages is still unsatisfactory due to the lack of corresponding knowledge about target languages. In this work, on top of those relevant pre-trained multi-lingual language models, we propose a generalized framework under the guidance of self-supervised learning to enhance the multi-lingual transferability using augmentation data in low-resource languages.

2.3 Data Augmentation

Data augmentation is a commonly applied technique in the NLP field to generate additional training examples by utilizing all kinds of transformation operations. With the development of machine translation models and the availability of online APIs, machine translation has become a popular augmentation method in many tasks, especially in multi-lingual scenario [11]. Previous studies [12, 23] show that adding translated training data can significantly improve the model performance, especially on languages that are distant from the source language. However, the lack of diversity is a very common phenomenon in translated data, and there may be some translation artifacts in the data distribution. Pre-trained language models have become the mainstream models in recent years due to their excellent performance. Masked language models (MLMs) such as BERT [5] and RoBERTa [16] have obtained the ability to predict the masked words in the text based on the context, which can be used for increasing the diversity of augmented data. In this work, we apply a similar words substitution method by masking multiple words in the translated document and recovering new various sentences. Moreover, there are some sophisticated data generation approaches that can construct additional training data [4, 20]. But these generative methods require a certain amount of labels. We imitate the denoising process of the mBART model, and train translation data to obtain enriched generation corpus. Motivated by LAMBADA [1] method, we evaluate the quality of the augmented data. Due to the difficulty of training another multi-lingual classifier in the text generation scenario, we adopt a self-supervised approach to score the augmented data.

3 METHODOLOGY

In this section, we first formalize the problem and introduce our unified self-supervised learning framework for the multi-lingual text advertisement generation task. Then, we propose our method formed of a data augmentation module and an adaptive filtering module.

3.1 Task Formulation and Overall Framework

The multi-lingual text ad generation task involves generating a qualified relevant description of ads for a product landing page in any given languages. Given a product webpage to be advertised, we concatenate all of its text content into a landing page document $\boldsymbol{Doc} = [d_1, ..., d_n]$, and text ad generation predicts an ad description $\boldsymbol{Ad} = [a_1, ..., a_m]$. Here, $d_i, a_i \in \mathcal{V}$ represents tokens in the document and advertisement, and $\mathcal{V}$ denotes the vocabulary. Denote by $\mathcal{D}^S_{train} = \{(\boldsymbol{Doc}, \boldsymbol{Ad})\}$, the labeled data in the source language, where the superscript S indicates that this is a dataset in the source language. Here, we target at two different multi-lingual settings: *zero-shot* and *few-shot*. For the extreme *zero-shot* setting the source language only contains English (or, in general, a rich-resource language). While for the *few-shot* setting, or called real production setting, there are few labeled data of other low-resource languages in the source language. Besides, all the target languages are available for the test set $\mathcal{D}^T_{test} = \{(\boldsymbol{Doc}, \boldsymbol{Ad})\}$ no matter which setting, and the superscript T indicates those are datasets in the target languages.

We frame our multi-lingual text advertisement generation problem as a three-module sequence-to-sequence learning task. The goal is to learn a function $\boldsymbol{Ad} = f(\boldsymbol{Doc})$, where the f can be represented by using a neural Transformer-based sequence generation model. Here we adopt four pre-trained multi-lingual language models, ProphetNet [19], Unicoder [9], mBART [15], mT5 [25] as f separately, to verify the generalizability of our method. Any pre-trained multi-lingual language models for the sequence-to-sequence generation task could be employed. In this paper, the basic idea of SAG is to ensure that f first uses rich-resource (source) language data to obtain fundamental capabilities (e.g., generate a fluent ad sentence) and then gradually learns abilities to generate high-quality ads in a variety of low-resource (target) languages. This can be achieved conveniently by extending the training schema of pre-trained language model to three modules:

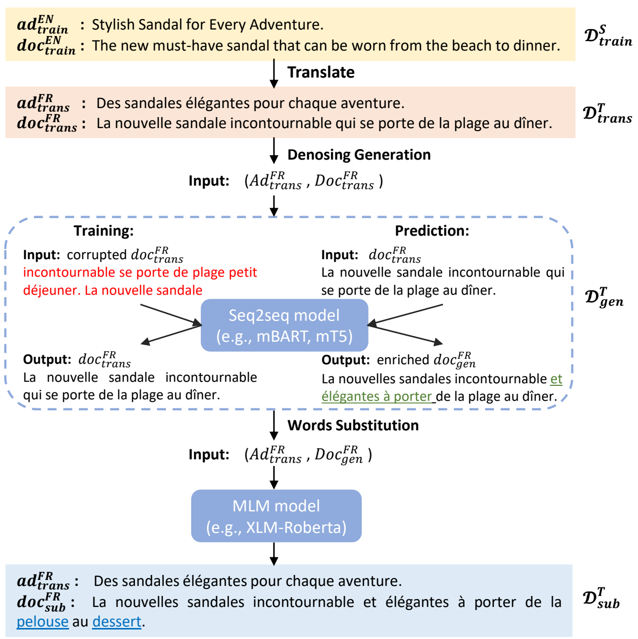

ad^{EN}_{train} : Stylish Sandal for Every Adventure.
doc^{EN}_{train} : The new must-have sandal that can be worn from the beach to dinner.
$\mathcal{D}^S_{train}$
Translate
ad^{FR}_{trans} : Des sandales élégantes pour chaque aventure.
doc^{FR}_{trans} : La nouvelle sandale incontournable qui se porte de la plage au dîner.
$\mathcal{D}^T_{trans}$
Denosing Generation
Input: $(Ad^{FR}_{trans}, Doc^{FR}_{trans})$
Training:
Input: corrupted doc^{FR}_{trans}
incontournable se porte de plage petit déjeuner. La nouvelle sandale
Prediction:
Input: doc^{FR}_{trans}
La nouvelle sandale incontournable qui se porte de la plage au dîner.
Seq2seq model (e.g., mBART, mT5)
Output: doc^{FR}_{trans}
La nouvelle sandale incontournable qui se porte de la plage au dîner.
Output: enriched doc^{FR}_{gen}
La nouvelles sandales incontournable et élégantes à porter de la plage au dîner.
$\mathcal{D}^T_{gen}$
Words Substitution
Input: $(Ad^{FR}_{trans}, Doc^{FR}_{gen})$
MLM model (e.g., XLM-Roberta)
ad^{FR}_{trans} : Des sandales élégantes pour chaque aventure.
doc^{FR}_{sub} : La nouvelles sandales incontournable et élégantes à porter de la pelouse au dessert.
$\mathcal{D}^T_{sub}$

Figure 1: The data augmentation module of SAG.

- **Pre-training.** Pre-trained with multi-lingual unsupervised corpus, models can be equipped with fundamental natural language processing and generation abilities, e.g., understanding the grammar of different languages and distinguishing stop words.

- **Data Augmentation.** SAG is then fine-tuned with the original source language data and augmented target language data by minimizing the loss function of each backbone model. The augmented data $\mathcal{D}^T_{aug} = \{(\boldsymbol{Doc'}, \boldsymbol{Ad'})\}$ mainly contains three parts: $\mathcal{D}^T_{trans}$ is obtained using pre-trained multi-lingual translation models; $\mathcal{D}^T_{gen}$ is enriched by denoising method on translated corpus; and $\mathcal{D}^T_{sub}$ is supplemented by words substitution with pre-trained MLM model on $\mathcal{D}^T_{trans}$. Formally, multi-lingual ad generation aims to learn a model by leveraging both $\mathcal{D}^S_{train}$ and $\mathcal{D}^T_{aug}$ to obtain good performance on $\mathcal{D}^T_{test}$.

- **Adaptive Filtering Module.** To alleviate the problem of augmented errors, before each training process of our model, we filter the augmented data to obtain high-quality examples. Specifically, we design a self-learning method, which use the model trained from the previous step to calculate BLEU-4 value[2] or loss value for the augmented data. Only the data with scores less than the threshold τ can be added to the next training process. With the gradual optimization of model parameters, the threshold is updated adaptively, and the selected data will be more accurate.

3.2 Data Augmentation Module

In this module, we augment source language data to target languages via translation, denoising generation, and words substitution. The Figure 1 shows the details.

3.2.1 *Translation.*

We use multi-lingual BART (mBART) [15], which is pre-trained in 25 languages, to translate the training corpus $\mathcal{D}^S_{train}$ in the source language (English) to the target languages. During the translation process, according to the mBART model limit, documents will be intercepted with a maximum length of 512 and translated at the same time as the text advertisement. The reasons we use pre-trained language models as a translator rather than open translation apps such as Google Translator or Bing Translator are as follows: (1) There is no absolutely perfect and accurate translator unless human translators are employed. (2) We hope that our model can be independent of the effect of the translator, regardless of which translator. Through the filtering process, we can always obtain high-quality augmented data. (3) In future work, it is possible to integrate the translator into our unified architecture, in which we will propose a joint learning model to train the translation model and ad generation model simultaneously.

3.2.2 *Denoising Generation.*

We observe that the corpus $\mathcal{D}^T_{trans}$ obtained from translation procedure lacks diversity. To alleviate this issue, we leverage a sequence-to-sequence (seq2seq) model as a generator to synthesize additional target language corpus $\mathcal{D}^T_{gen}$. However, training a high-quality seq2seq model usually requires a large number of labeled sequence pairs (like in machine translation). To overcome this label requirement, we apply the idea of the denoising method by regarding the translated corpus $\mathcal{D}^T_{trans}$ as high-quality sequence pairs, and produce augmented data by corrupting their input sequences ($\boldsymbol{Doc}$).

As shown in Figure 1, we can corrupt the translated document *"La nouvelle sandale incontournable qui se porte de la plage au dîner."* into *"incontournable se porte de plage petit déjeuner. La nouvelle sandale"* by several operations, such as **token delete**, **sentence permutation**, **text infilling**, etc. Then the seq2seq model is trained to *invert* these corruption operators by restoring them to the original sentence. In this work, we leverage mBART as the seq2seq model, since its pre-training tasks are similar to these corrupt operations.

During prediction, we apply the fine-tuned mBART on the original translated document doc^T_{trans} and expect to generate doc^T_{gen} with natural additional information. This method can generate natural yet diverse augmentations such as adding some adjectives like *"et élégantes à porter"*, or supply few location information like *"De la plage au coucher du soleil au dîner dans un restaurant haut de gamme"*. Then, we perform preliminary data selection by filtering out the generated utterances which have no overlapping words with translated advertisement ad^T_{trans}. At last, we randomly sample n instances from the candidate set to construct the generated corpus $\mathcal{D}^T_{gen}$, where n is the number of English training data.

3.2.3 *Words Substitution.*

To further enhance the training data of target languages and increase the diversity, in addition to translation operation and denoising generation, we introduce a simple yet effective word substitution task for data augmentation. First, we randomly mask b words for each document from n candidates of doc^T_{trans} and doc^T_{gen}, where the masked words can not exist in the translated advertisement ad^T_{trans}. Then, pre-trained multi-lingual language model (MLMs) XLM-RoBERTa [3] is used to predict the masked words in the document based on the whole context. Data

[2] https://github.com/mjpost/sacrebleu

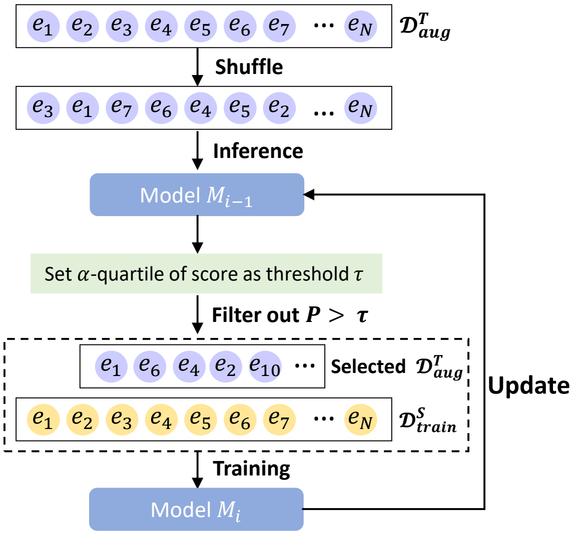

e_1 e_2 e_3 e_4 e_5 e_6 e_7 ⋯ e_N
$\mathcal{D}^T_{aug}$
Shuffle
e_3 e_1 e_7 e_6 e_4 e_5 e_2 ⋯ e_N
Inference
Model M_{i-1}
Set α-quartile of score as threshold τ
Filter out $P > \tau$
e_1 e_6 e_4 e_2 e_{10} ⋯ Selected $\mathcal{D}^T_{aug}$
e_1 e_2 e_3 e_4 e_5 e_6 e_7 ⋯ e_N $\mathcal{D}^S_{train}$
Update
Training
Model M_i

Figure 2: The adaptive filtering module of SAG.

supplement in this way has the following advantages: (1) since the keywords (words in the ad) are not replaced, the general meaning of the document is not affected. (2) MLM task performed by multi-lingual pre-trained language model makes the substituted words conform to context semantics and increase the diversity of selected augmented data.

3.3 Adaptive Filtering Module

To alleviate the noise or errors introduced by the data augmentation module, we design an adaptive filtering module as shown in Figure 2. In order to optimize the model parameters in the right direction and alleviate data noise, we take source language training data as the dominant training data and select high-quality augmented data as the auxiliary to fine-tune the model. At the initialization stage, we leverage the pre-trained multi-lingual generation model to train the labeled corpus in the source language. With the increase of training rounds, the model gradually introduces the selected augmented data into training, and the ratio α of the introduced data also gradually increases. Through the adaptive filtering and training of augmented examples, the generation model trained from the previous step can be improved on the target languages. Therefore, we further iterate this filtering process multiple rounds to drive the final target model. Guided by the source language data and assisted by high-quality augmented data, the model can adapt to the multi-lingual ad generation task well. Now, we need to decide: how to define the quality of the augmented data, and how to adaptively filter out low-quality data.

3.3.1 Define the Data Quality. Inspired by the self-training paradigm, where a model is trained using labels predicted by itself from the previous step, we leverage the model M_{i-1} from the previous step to score the augmented data. Although initial model M_0 is only trained using the labeled data in the source language, it is capable of inferring directly on the cases in the target language, due to benefit from the multi-lingual knowledge in the pre-trained generative language models. In this work, we have tried two scoring strategies: **calculating loss value** and **calculating BLEU value**.

Calculating loss value is a straightforward way to judge the quality of current augmented data. Given an example of augmented data $(doc^T_{aug}, ad^T_{aug}) \in \mathcal{D}^T_{aug}$, the loss value is $P = M_{i-1}(doc^T_{aug}, ad^T_{aug})$. Naturally, the model assumes that the lower the loss value, the higher the quality of the current example. The strategy of calculating loss value can take into account both the document doc^T_{aug} and advertisement ad^T_{aug} quality of augmented data. However, this method tends to select examples which are similar to the current data distribution, thus losing the diversity of the augmented data.

Calculating BLEU value borrows the idea of the model evaluation. Given the document doc^T_{aug} of augmented data, the score P is calculated as the BLEU-4 value between the output of the model $M_{i-1}(doc^T_{aug})$ and the ground truth ad^T_{aug}. Unlike calculating the loss value, the model does not see the ground truth ad^T_{aug} when generating output based on the input document doc^T_{aug}, so this approach can select more diverse augmented data.

3.3.2 Adaptive Data Filtering. Intuitively, we can define a fixed threshold τ to select high-quality augmented data. For the quality score P, a smaller loss value or a larger BLEU-4 value indicates less noise in the augmented sentence and we will add it into the training data in the next round. However, as the augmented data selection ratio α increases, a fixed threshold may not be sufficient. Therefore, we use a simple yet effective method to adjust the threshold adaptively.

Specifically, at the beginning of each round, we first shuffled the augmented data $\mathcal{D}^T_{aug}$ and selected the first m candidates to calculate the quality score P. Then sort the P values of m candidates and set the value in the α-quantile as the threshold τ. Since α-quantile can select exactly α ratio data, we get relatively high-quality candidates after filtering all the augmented data $\mathcal{D}^T_{aug}$ with this threshold. Therefore, the model eventually learns both the structure of ad generation from the source language and the multi-lingual syntactic rules from the augmented data of the target languages.

4 EXPERIMENTS

In this section, we first evaluate SAG on the XGLUE NTG [9] benchmark with the zero-shot setting and then conduct a few-shot setting with automatic offline experiments, human evaluation, and online experiments based on the production of Microsoft Bing.

4.1 Datasets

• **XGLUE NTG.** XGLUE is a benchmark dataset that can be used to train and evaluate large-scale multi-lingual pre-trained language models across a diverse set of tasks. In this paper, we focus on the task closest to generation, news title generation (NTG), which aims to generate a high-quality title for a given news document. NTG task covers 5 languages, including English (EN), German (DE), French (FR), Spanish (ES), and Russian (RU), where the labeled training data (300,000 pairs) is only in English. Examples for the test set of each language are 10,000 pairs. Therefore, we leverage this dataset to test the validity and effectiveness of our proposed

	EN	DE	FR	ES	NL	IT
# T	838,854	109,889	59,108	22,461	69,979	16,201
# V	5,000	5,000	5,000	5,000	5,000	5,000

Table 1: Statistics of BingAd dataset.

framework in the zero-shot scenario. Note that to follow the zero-shot setting, we use the English development set to select the best models and evaluate them directly on the target language test sets.

- **BingAd.** From a large web service (Bing), we acquire a large number of landing pages (LP) for product as documents and corresponding handwritten text ads from advertisers as ground truth. According to the number of advertising pages in various languages, we choose the following languages for the experiment: English (EN), German (DE), French (FR), Spanish (ES), Dutch (NL), Italian (IT). We clean the (***Doc***, ***Ad***) pairs by removing special characters (such as common HTML code elements), and obtain the language identification based on the text ad. To prevent the model from copying the templates which create large numbers of ads, we keep at most 100 samples for each advertiser account. After data pre-processing, each corpus is split into training, development, and test sets. The statistics of those datasets are shown in Table 1.

4.2 Implementation Details

We leverage the PyTorch version of multi-lingual ProphetNet-Multi, Unicoder, mBART-25 and mT5-base in HuggingFace's Transformers[3] as the basic sequence-to-sequence model for all variants. We use Adam as the optimizer and fine-tune the hyper-parameters on the validation set for each dataset. Most hyper-parameters of these models are re-used. We perform a grid search for the hyper-parameters specified as follows: learning rate $lr \in \{1e{-}5, 1e{-}6, 1e{-}7\}$, the proportion of data added per round $\beta \in \{0.02, 0.05\}$. For our method, beam size for evaluation is $B = 10$, beam size for defining the data availability is $B = 1$, and the number of examples for calculating the threshold is $m = 200$. The optimal hyper-parameters on XGLUE NTG dataset are: $lr = 1e - 7$, $\beta = 0.02$, $B = 1$, $m = 200$; and the optimal hyper-parameters on BingAd dataset are: $lr = 1e - 5$, $\beta = 0.05$, $B = 1$, $m = 200$. We leverage the most commonly used BLEU-4 metric, which are computed using the SacreBLEU library, to evaluate these two datasets.

4.3 Baseline Models

As introduced in Section 2, there are many existed works on text advertisement generation and data augmentation, and our proposed algorithm is orthogonal to these techniques. For example, we can leverage various strategies to train multi-lingual data (such as reinforcement learning, LSTM, etc.) or do data augmentation, and then use our denoising method to optimize these data when selecting high-quality corpus in the fine-tuning process. In this paper, the comparison is mainly focused on the pre-trained multi-lingual language models and their variations.

(1) **ProphetNet-Multi**, which is a multi-lingual version encoder-decoder model and can predict n-future tokens for "n-gram" language modeling instead of just the next token.

Models		XGULE NTG					
		FR	DE	ES	RU	EN	Avg
Pro	Paper	11.4	8.7	12.7	8.5	16.7	11.6
	+T	11.83	8.16	13.85	9.29	16.45	11.92
	+T+G	11.84	8.21	13.63	9.37	16.81	11.97
	+T+G+S	11.79	8.25	13.70	9.42	16.76	11.98
	SAG $_{loss}$	12.33	8.68	13.74	**9.60**	16.65	12.20
	SAG $_{bleu}$	**12.34**	**8.71**	**13.80**	9.58	**16.72**	**12.23**
Uni	Paper	9.9	7.5	11.9	8.4	15.8	10.7
	+T	10.20	7.25	11.90	8.48	15.78	10.72
	+T+G	10.13	7.40	11.84	8.53	15.73	10.73
	+T+G+S	10.33	7.40	11.82	8.52	15.82	10.78
	SAG $_{loss}$	10.69	**7.78**	12.48	**8.57**	15.81	11.07
	SAG $_{bleu}$	**10.80**	7.69	**12.74**	8.48	**15.83**	**11.11**
mT5	Repro	9.48	7.60	11.04	8.66	11.51	9.66
	+T+G+S	10.97	7.51	12.83	9.46	12.33	10.62
	SAG $_{loss}$	11.22	7.74	12.97	**9.54**	12.28	10.75
	SAG $_{bleu}$	**11.49**	**7.85**	**13.19**	9.11	**12.40**	**10.81**
mBART	Repro	11.04	7.45	10.96	8.67	16.33	10.89
	+T+G+S	11.64	7.54	12.33	9.23	16.27	11.40
	SAG $_{loss}$	**12.40**	**8.75**	14.12	**9.58**	**17.05**	**12.38**
	SAG $_{bleu}$	12.34	8.69	**14.18**	9.47	16.89	12.31

Table 2: BLEU-4 results on zero-shot XGLUE NTG task. Each bold number is the best result in that column. "Avg" denotes the average score on these 5 languages results. "Pro" and "Uni" represent ProphetNet and Unicoder models, respectively. "+T", "+G", and "+S" denote adding translated data, self-supervised generation data, and word substitution data, respectively. "Repro" represents reproduced results.

(2) **Unicoder**, which includes pre-training tasks including MLM, TLM and introduces a future n-gram prediction mechanism to natural language generation.

(2) **mBART-25**, which is a sequence-to-sequence denoising auto-encoder pre-trained on large-scale corpora in 25 languages using the BART objective. As mentioned in the original paper, language identification tags need to be added before each sentence.

(2) **mT5-base**, which leverages a unified multi-lingual text-to-text format and was pre-trained on a new Common Crawl-based dataset covering 101 languages. Follow the instructions[4], we do not use a task prefix during single-task fine-tuning, since mT5 was pre-trained unsupervised.

4.4 Results on the Zero-shot NTG Dataset

Table 2 shows the BLEU-4 results on the zero-shot multi-lingual NTG task of our method based on four baseline models. For each model, the "Paper" and "Repro" represent results from the original papers or reproduced with the open-sourced framework. "+T", "+T+G", "+T+G+S" represent a random sampling of augmented data at the ratio α of each round. Following the ProphetNet [19], we first fine-tune each model on the training data, then choose the best results for English and all other languages separately. Besides, we

[3] https://github.com/huggingface/transformers

[4] https://huggingface.co/docs/transformers/model_doc/mt5

conduct experiments on different variants of the proposed framework to investigate the contributions of different components. From the above tables, we can see that:

(1) Our proposed SAG model significantly outperforms other baselines and SAG based on mBART achieves state-of-the-art performance almost in all languages. The snapshot of the leaderboard of the XGLUE NTG task is presented in Appendix A. It verifies the effectiveness of our data augmentation methods and adaptive filtering approach in the multi-lingual generation task.

(2) When defining the availability of augmented data, SAG shows consistent improvement using the method of calculating BLEU values on all models except mBART. It indicates that the diversity can be increased by filtering data in different ways from the training process, thus improving the effectiveness of the model.

(3) Our method works much better in Russian than it does in German, since German and English belong to the same language system with high similarity, and the augmented data is of little use. This indicates that it is more useful to supplement language data that is different from the source language.

(4) The overall performance of Unicoder and mT5-base on the NTG benchmark is relatively weak. The reason may be that their pretraining corpus is not rich enough and there is a gap between the pre-training task and the current generation task.

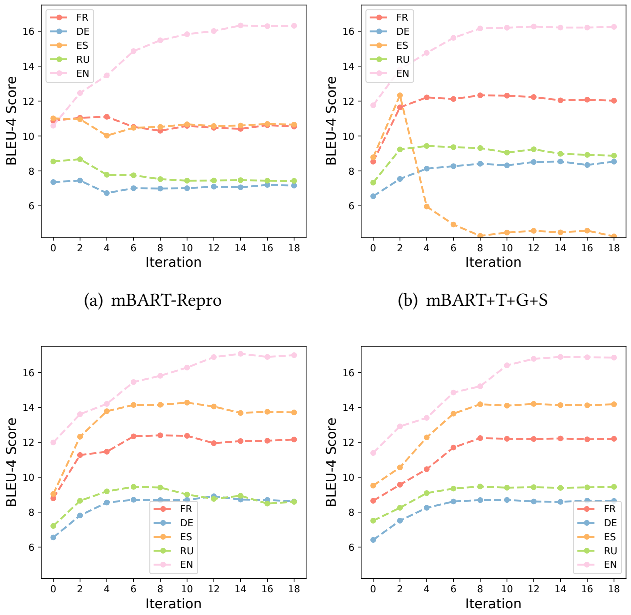
(5) From the ablation study of our augmentation module on the ProphetNet model, we can see that: in general, all of the proposed techniques contribute to the multi-lingual setting. Both the self-supervised generation and words substitution can improve on the basis of translated data, especially on the Russian language dataset, indicating that simple changes and additions to translated data are efficient ways to increase the diversity of augmented data.

(6) One interesting finding is that the performances on German become slightly worse after adding the augmented data. It is because the noise introduced by the augmentation processes may hurt the performance. Our adaptively filtering method is able to handle the noise of synthesized data.

4.5 Effect of Multiple Iterations on NTG

To further investigate how evaluation metrics change while iteration increases, we conduct experiments on the NTG benchmark to study each language dataset in the SAG training framework. Take mBART as an example, the BLEU-4 scores of 20 rounds of iterations are shown in Figure 3. From the figures, we observe that:

(1) With increasing numbers of iterations, the performance of “+T+G+S”, “-loss”, and “-bleu” methods in almost all the languages increases to some level then converges. This indicates that augmented data is the most direct and effective way for the model to learn other language information, and the adaptive filtering process can effectively select high-quality augmented data and further improve the model performance.



(a) mBART-Repro

(b) mBART+T+G+S

(c) SAG-loss

(d) SAG-bleu

Figure 3: Effect of Multiple Iterations on NTG task based on mBART model.

Models	BingAd						
	FR	DE	ES	NL	IT	EN	Avg
Unicoder⋆	1.89	1.75	1.99	1.65	2.21	15.88	4.23
mT5⋆	2.32	1.99	2.21	1.89	2.55	16.78	4.62
mBART⋆	3.01	2.22	3.55	2.81	3.21	18.81	5.60
Pro⋆	3.18	2.65	3.53	3.13	3.41	18.34	5.71
Pro	18.15	15.48	19.86	20.21	17.94	21.01	18.78
Pro+Aug	21.38	14.75	21.75	22.26	20.75	19.25	20.02
SAG $_{loss}$•	3.28	3.40	3.15	4.49	1.04	1.14	2.75
SAG $_{loss}$	**24.31**	17.14	**23.24**	24.05	24.52	20.89	22.53
SAG $_{bleu}$	24.24	**17.29**	23.13	**24.91**	**24.87**	**21.63**	**23.18**

Table 3: BLEU-4 results on the few-shot BingAd dataset. “Pro” represents the ProphetNet model. ⋆ and “+Aug” represent these models only train on English corpus and add augmented data (“T+G+R”), respectively. • denotes that the model is randomly initialized.

(2) In Figure 3(b), the Spanish curve falls off a cliff, and other languages also fluctuate to varying degrees. It indicates that the randomly selected augmented data is unstable and the noise can seriously affect the training effect of the model. After adding the adaptive filtering process, it can be seen from Figure 3(c) and Figure 3(d) that the number of convergent rounds for various languages increases, and the curves are relatively flat and stable.

(3) Although in mBART model, “SAG-loss” can achieve the highest overall result. However, compared with “SAG-loss” 3(c), the curves of each language after convergence of “SAG-bleu” 3(d) are more stable, which demonstrates that calculating BLEU score can obtain high-quality augmented data more stably.

	Landing Page 1 – German https://www.toner-partner.de/panasonic-kx-p-1180-i/	**Landing Page 2 – French** https://www.economybookings.com/fr?idpick=lakselv
Pro+Aug	Ad: Gratis Versand Ab 39,50â,¬ | Schnelle Lieferung | 3 % Skonto. Translate: Free shipping from €39.50 | Fast delivery | 3% discount.	Ad: Comparez Et Ã‰conomisez Avec Kayak. Translate: Compare And Save With Kayak.
SAG $_{bleu}$	Ad: Schnelle Lieferung, GÃ¼nstige Preise. Ãœber 1 Million Zufriedene Kunden! Translate: Fast delivery, cheap prices. Over 1 million satisfied customers!	Ad: Rent-A-Car SpÃ©cialiste. Un Prix Imbattable! Translate: Rent-A-Car Specialist. An Unbeatable Price!

Table 4: Generation output from ProphetNet+Aug and SAG $_{bleu}$ for German and French.

4.6 Results on the Few-shot BingAd Dataset

In BingAd dataset, the task is to generate a text advertisement from a landing page document. We first evaluate the transferring abilities of the baselines in multi-lingual scenarios utilizing BingAd dataset, including Unicoder, mT5-base, mBart and ProphetNet. From Table 3, we observe that if the model is trained only on English corpus, it performs very poorly in other languages, especially Unicoder. There are two main reasons: (1) Sentences on the landing page are generally not coherent, and there may be a large jump between each sentence. Sometimes the sentence is not complete, but rather a combination of descriptive phrases for a particular product. Such documents are different from the coherent sentences most models use in pre-training tasks. For example, a landing page document for selling "Maserati cars" is: *"-SD_DocumentTitle- New Maserati Cars & SUVs in Columbus, OH | Near Delaware & Hilliard -SD_Heading- 2022 Maserati Ghibli Modena Q4 Sedan . Maserati MC20 -SD_Paragraph- Due to the variety of manufacturer incentives, please contact one of our MAG Maserati Brand Specialists at 614-541-2302"*, where *"-SD_DocumentTitle-"*, *"-SD_Heading-"* and *"-SD_Paragraph-"* represent different tags on the web page. (2) Text advertisements are a specific form of language, usually containing interjections to attract users or a concatenation of discount information. In particular, advertising formats vary widely between languages. Therefore, text ad generation learned only in English corpus is difficult to transfer to other languages. Compared with other models, ProphetNet has a better inter-language migration effect, thus we choose the ProphetNet model as our backbone model. Moreover, this is why we chose the few-shot setting in the real production scenario.

After training on real annotation datasets for all languages in the few-shot setting, we observe a significant improvement in the ProphetNet results. On the basis of this result, the model performance is further improved in almost all languages by adding the augmented data, which shows the effectiveness of the three data augmentation methods. Furthermore, the average results of both SAG $_{loss}$ and SAG $_{bleu}$ of our SAG exceed these baselines, which indicates that the adaptive filtering module can effectively reduce the noise of the augmented data. In addition, we try to train the model architecture from scratch without using pre-trained model parameters. It can be seen that SAG $_{loss}\bullet$ almost completely fails, which also illustrates the importance of pre-trained language models for downstream tasks with fewer data.

We further conduct a case analysis of generated ads on the BingAd dataset to examine the capability of our approach. From Table 4, we observe that the generated output is seldom well-formed ad with the ProphetNet model fine-tuned without denoising process. The model either predicts just a plain sub-span of the passage or some irrelevant sentences. Our method generates ads with more fluent sentences and phrases that are relevant to the landing pages. This is consistent with the offline evaluation, which shows that our method achieves a much larger BLEU-4 score.

Models	German (DE)				
	Grammar	Human	Accurate	Relevant	Overall
ProphetNet	**99.0%**	94.5%	88.5%	82.0%	71.4%
Pro+Aug	97.4%	**97.9%**	89.0%	80.0%	70.0%
SAG $_{bleu}$	98.5%	94.4%	**89.1%**	**89.5%**	**73.3%**
Advertiser	95.9%	96.4%	83.6%	**99.0%**	79.5%
LPExtractor	87.8%	87.2%	78.6%	89.8%	77.0%
SAG $_{bleu^{\circ}}$	**97.5%**	**98.5%**	**84.3%**	93.9%	**82.8%**

Table 5: Human evaluation of ad quality for German. Following the LPExtractor model, SAG $_{bleu^{\circ}}$ represents that a context-aware classifier is used after SAG.

4.7 Human Evaluation

Automated metrics like BLEU-4 measures similarity between the predictions and target text ad in terms of word overlap. They do not directly capture whether a prediction is actually an advertisement. In this experiment, we ask human judges in Universal Human Relevance System (UHRS) crowd-sourcing platform to label ads in terms of quality. On average the judges have 12 months of experience in labeling text ads. They have been carefully trained to ensure that they correctly understand the tasks. Detailed guidance is provided before labeling and their initial labeling results have been checked to eliminate misunderstanding. We randomly order the ads so that the judges do not know which method is used to generate each ad.

4.7.1 Criteria. High quality is one of the most important goals for ad generation. To evaluate quality, the following four pivots are evaluated by judges: (1) grammar: language fluency and grammar correctness; (2) human-likeness: whether the ad looks like a human-written one; (3) relevance: whether the ad is relevant to the product landing page; (4) accuracy: whether the information covered in the ad is accurate concerning what is on the advertiser's landing page/website. An ad is considered overall good if and only if all of the four pivots are labeled as good by the judges.

4.7.2 Baselines. In addition to the three generative models (ProphetNet, ProphetNet + Aug, SAG $_{bleu}$) mentioned above, there are two baselines that can be used to evaluate the availability to deploy to the production scenarios. (1) Advertiser: for each website, a text ad written by the advertiser is randomly selected. This baseline is used to evaluate human-written advertising. If the ad generated by the model keeps on-par performance with the ad written by the advertiser, the model can be used in real product scenarios. (2)

	German	French	Spanish	Total
△Coverage	41.0%	5.7%	101.0%	35.5%
△Adoption	-	-	-	33.3%
△Revenue($)	-	-	-	269.8%

Table 6: Online evaluation results for three languages on RSA channel of Bing. △ represents the percentage of SAG increase compared to LPExtractor model.

LPExtractor: text chunks are extracted from the landing page as raw ad candidates, and then these candidate chunks are scored by context-aware classifiers to determine whether they are valid ad.

4.7.3 Results. We randomly sample 200 text ads each for five languages – German, French, Spanish, Dutch, and Italian. Human evaluation results of German are shown in Table 5, and results for other languages are presented in Appendix B. We observe that: (1) Ads generated using SAG $_{bleu}$ have better or comparable quality according to multiple criteria compared with those generated by the baselines. (2) The relevance of SAG $_{bleu}$ is significantly improved compared with other baselines, which indicates that it is very important to filter out low-quality augmented data for parameter optimization of the model. (3) SAG $_{bleu}{}^{\circ}$ is not only better than the direct LPExtractor model but also outperforms the hand-designed ads by advertisers on all metrics except relevance. This means that our model can generate ads that are comparable to humans.

4.8 Online Evaluation

We further deploy SAG to the production scenario, Responsive Search Ads (RSA) of the Microsoft Bing, to illustrate its practical effectiveness. RSA is a major search engine that has the largest market share and has more than 100 million monthly active users and allows for minimum advertiser effort when creating ads. In particular, the advertisers share their landing pages and then RSA infrastructure offline generates ads for landing pages by using different models. We performed an A/B test for 14 days with the LPExtractor model as control and our proposed SAG as treatment in RSA channel across 3 languages: German (DE), French (FR), and Spanish (ES). From Table 6, we observe that our approach has significantly improved coverage in every language. Especially in Spanish, coverage is 101.0% higher than the LPExtractor model. Moreover, the improvements in total are highly statistically significant, especially in revenue metrics, which demonstrates the effectiveness of our framework in the real multi-lingual product scenario.

5 CONCLUSION

In this paper, we propose a self-supervised augmentation and generation framework SAG for the multi-lingual text advertisement generation task. It iteratively filters the augmented data to obtain a high-quality training corpus, which gradually optimizes model parameters to make SAG more transferable and flexible. Extensive experiments verify that SAG outperforms the existed methods and establishes the new state-of-the-art performance on the generation task. For future work, we will apply different weights for the multiple augmented data and incorporate more denoising methods to improve the effectiveness, and conduct an in-depth study on promoting the efficiency based on the current framework.

REFERENCES

[1] Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? Deep learning to the rescue!. In *AAAI*. 7383–7390.

[2] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. *arXiv preprint arXiv:2004.04721* (2020).

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[4] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. 2018. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine* (2018), 53–65.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Atsushi Fujita, Katsuhiro Ikushima, Satoshi Sato, Ryo Kamite, Ko Ishiyama, and Osamu Tamachi. 2010. Automatic generation of listing ads by reusing promotional texts. In *Proceedings of the 12th International Conference on Electronic Commerce: Roadmap for the Future of Electronic Business*. 179–188.

[7] Yingmei Guo, Linjun Shou, Jian Pei, Ming Gong, Mingxing Xu, Zhiyong Wu, and Daxin Jiang. 2021. Learning from multiple noisy augmented data sets for better cross-lingual spoken language understanding. In *EMNLP*.

[8] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[9] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909.00964* (2019).

[10] J Weston Hughes, Keng-hao Chang, and Ruofei Zhang. 2019. Generating better search engine text advertisements with deep reinforcement learning. In *SIGKDD*.

[11] Bohan Li, Yutai Hou, and Wanxiang Che. 2021. Data Augmentation Approaches in Natural Language Processing: A Survey. *arXiv preprint arXiv:2110.01852* (2021).

[12] Xin Li, Lidong Bing, Wenxuan Zhang, Zheng Li, and Wai Lam. 2020. Unsupervised Cross-lingual Adaptation for Sequence Tagging and Beyond. *arXiv preprint arXiv:2010.12405* (2020).

[13] Shining Liang, Ming Gong, Jian Pei, Linjun Shou, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. 2021. Reinforced Iterative Knowledge Distillation for Cross-Lingual Named Entity Recognition. *arXiv preprint arXiv:2106.00241* (2021).

[14] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401* (2020).

[15] 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. *ACL* (2020), 726–742.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[17] Rajarshee Mitra, Rhea Jain, Aditya Srikanth Veerubhotla, and Manish Gupta. 2021. Zero-shot Multi-lingual Interrogative Question Generation for" People Also Ask" at Bing. In *SIGKDD*. 3414–3422.

[18] Baolin Peng, Chenguang Zhu, Michael Zeng, and Jianfeng Gao. 2020. Data augmentation for spoken language understanding via pretrained models. *arXiv e-prints* (2020), arXiv–2004.

[19] Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In *EMNLP: Findings*. 2401–2410.

[20] Giuseppe Russo, Nora Hollenstein, Claudiu Musat, and Ce Zhang. 2020. Control, generate, augment: A scalable framework for multi-attribute text generation. *arXiv preprint arXiv:2004.14983* (2020).

[21] Siamak Shakeri, Noah Constant, Mihir Sanjay Kale, and Linting Xue. 2020. Multilingual Synthetic Question and Answer Generation for Cross-Lingual Reading Comprehension. *arXiv e-prints* (2020), arXiv–2010.

[22] Stamatina Thomaidou, Ismini Lourentzou, Panagiotis Katsivelis-Perakis, and Michalis Vazirgiannis. 2013. Automated snippet generation for online advertising. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 1841–1844.

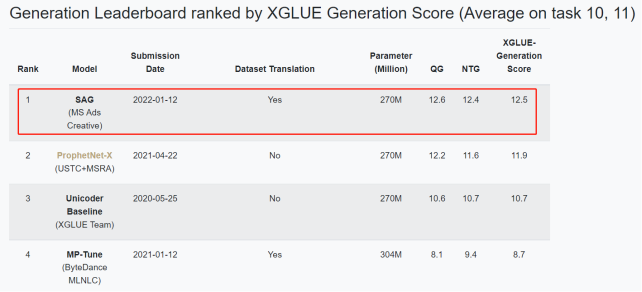
[23] Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (Almost) zero-shot cross-lingual spoken language understanding. In *ICASSP*. IEEE, 6034–6038.

[24] Xiting Wang, Xinwei Gu, Jie Cao, Zihua Zhao, Yulan Yan, Bhuvan Middha, and Xing Xie. 2021. Reinforcing Pretrained Models for Generating Attractive Text Advertisements. In *ACM SIGKDD*. 3697–3707.

[25] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934* (2020).

A SNAPSHOT OF XGLUE NTG TASK

The snapshot of the XGLUE NTG task is shown in Figure 4, where SAG achieved SOTA performance. Besides the NTG task, Question Generation (QG) is another generation task, which aims to generate a question for a given passage. It covers 6 languages, including English, French, German, Spanish, Italian, and Portuguese. BLEU-4 score is used as the metric. We also achieved SOTA performance on the QG task using SAG, demonstrating the generalization of our method and its applicability to a variety of similar generation tasks.



Generation Leaderboard ranked by XGLUE Generation Score (Average on task 10, 11)

Rank	Model	Submission Date	Dataset Translation	Parameter (Million)	QG	NTG	XGLUE-Generation Score
1	**SAG** (MS Ads Creative)	2022-01-12	Yes	270M	12.6	12.4	12.5
2	ProphetNet-X (USTC+MSRA)	2021-04-22	No	270M	12.2	11.6	11.9
3	**Unicoder Baseline** (XGLUE Team)	2020-05-25	No	270M	10.6	10.7	10.7
4	**MP-Tune** (ByteDance MLNLC)	2021-01-12	Yes	304M	8.1	9.4	8.7

Figure 4: Snapshot of XGLUE NTG task.

B RESULTS FOR HUMAN EVALUATION

Human evaluation results of four languages - French, Spanish, Dutch and Italian are shown in Table 7, 8, 9, 10. We can see that our approach is on-par with the advertiser’s manual labeling, thus SAG $_{bleu^{\circ}}$ can be used in real product scenarios in all of these languages.

Models	French (FR)				
	Grammar	Human	Accurate	Relevant	Overall
ProphetNet	91.9%	90.9%	67.2%	94.9%	65.2%
Pro+Aug	87.8%	86.8%	82.2%	**99.5%**	75.1%
SAG $_{bleu}$	**96.3%**	**95.3%**	**93.2%**	93.2%	**90.4%**
Advertiser	96.5%	**100.0%**	94.0%	**99.5%**	90.5%
LPExtractor	**100.0%**	**100.0%**	96.0%	**99.5%**	**96.0%**
SAG $_{bleu^{\circ}}$	99.5%	**100.0%**	**96.5%**	**99.5%**	**96.0%**

Table 7: Human evaluation of ad quality for French.

Models	Spanish (ES)				
	Grammar	Human	Accurate	Relevant	Overall
ProphetNet	**99.6%**	80.9%	80.9%	88.0%	65.7%
Pro+Aug	88.2%	96.7%	81.8%	84.9%	78.4%
SAG $_{bleu}$	98.5%	**99.0%**	**88.5%**	**97.0%**	**88.0%**
Advertiser	99.0%	99.5%	91.9%	99.5%	90.9%
LPExtractor	99.5%	**100.0%**	**97.0%**	**100.0%**	**96.5%**
SAG $_{bleu^{\circ}}$	**100.0%**	**100.0%**	94.8%	**100.0%**	94.8%

Table 8: Human evaluation of ad quality for Spanish.

Models	Dutch (NL)				
	Grammar	Human	Accurate	Relevant	Overall
ProphetNet	97.0%	97.5%	74.1%	**97.0%**	73.6%
Pro+Aug	**98.0%**	**98.5%**	79.2%	95.4%	78.2%
SAG $_{bleu}$	89.4%	89.9%	**88.9%**	90.9%	**84.3%**
Advertiser	**98.0%**	98.0%	71.7%	**99.5%**	70.2%
LPExtractor	96.5%	96.5%	**96.0%**	95.0%	**92.0%**
SAG $_{bleu^{\circ}}$	**98.0%**	**99.0%**	90.0%	96.5%	89.5%

Table 9: Human evaluation of ad quality for Dutch.

Models	Italian (IT)				
	Grammar	Human	Accurate	Relevant	Overall
ProphetNet	98.9%	99.5%	74.0%	**90.5%**	71.8%
Pro+Aug	**100.0%**	**100.0%**	52.5%	**90.5%**	52.5%
SAG $_{bleu}$	98.4%	99.5%	**75.0%**	76.0%	**74.5%**
Advertiser	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**
LPExtractor	99.5%	**100.0%**	97.0%	**100.0%**	96.5%
SAG $_{bleu^{\circ}}$	**100.0%**	97.0%	**100.0%**	**100.0%**	97.0%

Table 10: Human evaluation of ad quality for Italian.