MTG: A Benchmark Suite for Multilingual Text Generation

Yiran Chen[1], **Zhenqiao Song**[1,*], **Xianze Wu**[1], **Danqing Wang**[1], **Jingjing Xu**[1], **Jiaze Chen**[1], **Hao Zhou**[2,†], **Lei Li**[3,†]
[1] ByteDance AI Lab [2] Insititute for AI Industry Research, Tsinghua University
[3] University of California, Santa Barbara
{chenyiran.robin,songzhenqiao,wuxianze.0}@bytedance.com
{wangdanqing.122,chenjiaze}@bytedance.com
jingjingxu@pku.edu.cn zhouhao@air.tsinghua.edu.cn
lilei@cs.ucsb.edu

Abstract

We introduce MTG, a new benchmark suite for training and evaluating multilingual text generation. It is the first-proposed multilingual multiway text generation dataset with the largest human-annotated data (400k). It includes four generation tasks (story generation, question generation, title generation and text summarization) across five languages (English, German, French, Spanish and Chinese). The multiway setup enables testing knowledge transfer capabilities for a model across languages and tasks. Using MTG, we train and analyze several popular multilingual generation models from different aspects. Our benchmark suite fosters model performance enhancement with more human-annotated parallel data. It provides comprehensive evaluations with diverse generation scenarios. Code and data are available at https://github.com/zide05/MTG.

1 Introduction

Natural language generation (NLG) aims to automatically generate meaningful texts with the input in different formats, such as images (Anderson et al., 2018), tables (Ye et al., 2020) or texts (Guan et al., 2019). The generated texts generally target at realizing an underlying communicative goal while remaining coherent with the input information and keeping grammatically correct. Multilingual text generation extends the natural language generation task to produce texts in multiple languages, which is important to overcome language barriers and enable universal information access for the world's citizens (Artetxe et al., 2020; Arivazhagan et al., 2019; Pan et al., 2021).

To achieve this goal, various multilingual text generation datasets have been proposed. Some of them do not incorporate cross-lingual pairs (Liang et al., 2020; Ladhak et al., 2020). This limits the knowledge transfer from one language to another. Others involve cross-lingual pairs while English is included on either source or target side in most cases (Zhu et al., 2019; Ladhak et al., 2020), leading to difficult transfer between low-resource or distant language pairs. Constructing a multilingual text generation dataset that can directly transfer knowledge between any two languages is still under-explored.

To this end, we propose MTG, a human-annotated multilingual multiway dataset. Multiway means that the same sample is expressed in multiple languages. It covers four generation tasks (story generation, question generation, title generation and text summarization) across five languages (English, German, French, Spanish and Chinese). We do not include multilingual machine translation because MT itself is a standard task. The multiway parallel feature enables cross-lingual data construction between arbitrary language pairs. Such direct parallel signal promotes knowledge transfer and cross-lingual generation between any language pairs (even distant pairs such as Spanish-Chinese) without involving an intermediate language such as English (Leng et al., 2019).

The multilingual multiway feature also enables various training and test scenarios. In this paper, we design four scenarios to verify the advantages of our MTG from different aspects. Several representative pretrained multilingual models are employed to test these scenarios, including multilingual BERT (M-BERT) (Devlin et al., 2019), XLM (Lample and Conneau, 2019), mBART (Liu et al., 2020) and mT5 (Xue et al., 2020). We leverage various metrics to assess the coherence and diversity of the outputs generated by these models. Besides, we also propose an ensemble metric, which mainly focuses on relevance, measuring to what degree is the generated text close to human-level. Human evaluation is also conducted to validate models' performances.

*Corresponding author.
†Work is done while at ByteDance.

In summary, the contributions of this paper are listed as follows:
(i) We propose a new human-annotated multilingual multiway text generation benchmark suite MTG.
(ii) We design a new evaluation metric measuring how a text resembles human writing and prove that it has higher correlation scores with human scores compared with other automatic relevance metrics.
(iii) We evaluate several representative pretrained multilingual models on our proposed MTG and make a rigorous analysis to verify its advantages.

2 Related Work

A significant body of works have been committed to the construction of multilingual datasets covering diverse tasks (Hu et al., 2020; Jiang et al., 2020; Longpre et al., 2020). XTREME (Hu et al., 2020) is a multilingual understanding benchmark across 40 languages and 9 tasks, but it does not cover any generation task. Jiang et al. (2020) propose X-FACTR, which is a cross-lingual factual retrieval benchmark. Longpre et al. (2020) propose MKQA, an open-domain question answering evaluation dataset covering 26 diverse languages. Ladhak et al. (2020) present WikiLingua, which is a large-scale, multilingual dataset for cross-lingual abstractive summarization systems. MLSUM (Wang et al., 2021) is a dataset for text summarization in 12 languages. Wiki-40B (Guo et al., 2020) is a multilingual language model dataset across 40+ languages. Although these datasets cover multiple languages, they either belong to natural language understanding tasks or a single, specific generation task, which limits researchers to obtain general findings incorporating a set of generation tasks.

XGLUE (Liang et al., 2020) is a cross-lingual benchmark dataset for nine understanding tasks and two generation tasks. GEM (Gehrmann et al., 2021) is a newly-presented vision-language dataset covering 11 image-language and video-language tasks and 32 languages. These two datasets encompass multiple tasks and languages. However, a remarkable difference of our MTG from XGLUE and GEM is that MTG focuses on text-to-text generation tasks and is parallel across all languages, which facilitates easier knowledge transfer.

3 Dataset Collection and Methodology

This section will introduce how to create the benchmark suite for multilingual text generation (MTG). In order to construct multiway parallel dataset, the initial dataset is translated into other languages by an off-the-shelf translation model. Part of the translated data is randomly selected for further human annotation to increase data quality. The selection of tasks, initial datasets and languages are based on several principles as described below.

3.1 Task and Dataset Selection

It is important to select suitable tasks for our MTG benchmark to make it diverse and challenging. Thus, we define several criteria during the task selection procedure:
Task Definition Tasks should be well-defined, which means that humans can easily determine whether the generated results meet the task requirements.
Task Difficulty Tasks should be solvable by most college-educated speakers. In the meantime, they should be challenging to current models, the performance of which in various test scenarios falls short of human performance.
Task Diversity Tasks should cover a wide range of generation challenges that allow for findings to be as general as possible.
Input Format The input format of the tasks needs to be as simple as possible to reduce the difficulty of data processing. Besides, it should not contain anything but text (e.g., without any images or videos).

In order to meet the above criteria, 8 domain experts are asked to vote from 10 typical generation tasks[1]. Finally, four generation tasks are selected for MTG, which are story generation, question generation, title generation and text summarization. **Story generation** (SG) aims to generate the end of a given story context, which requires the model to understand the story context and generate a reasonable and fluent ending (Guan et al., 2019). **Question generation** (QG) targets at generating a correct question for a given passage and its answer (Duan et al., 2017). For the same passage with different answers, the system should be able to generate different questions. **Title generation** (TG) converts a given article into a condensed sentence while preserving its main idea (Jin and Hauptmann, 2002). The title should be faithful to the original document and encourage users to read the news

[1]These generation tasks are story generation, commonsense generation, style transfer, question generation, question answering, dialogue generation, title generation, text summarization, image caption, and data-to-text generation.

Task	Corpus	Domain	Format	Goal
Story Generation	ROCStories	Daily life	<story>	Generate the end of the story
Question Generation	SQUAD 1.0	Wikipedia	<passage,answer, question>	Generate the question of the answer
Title Generation	ByteCup	News	<article, title>	Generate the title of the document
Text Summarization	CNN/DailyMail	News	<article, summary>	Generate the summary of the document

Table 1: The description of tasks and English datasets included in MTG. For story generation, we use the last sentence as story end to be generated and the rest as input.

at the same time. **Text summarization** (Summ) aims to condense the source document into a coherent, concise, and fluent summary (Mani, 2001). It is similar to title generation but the output of text summarization is relatively longer. These four tasks focus on different generative abilities and realize different goals.

After confirming the tasks, the next step is to choose the dataset for each task. The two selection principles are listed as follows:(1) **License:** Task data must be available under licenses that allow using and redistributing for research purposes. The dataset should be free and available for download. (2) **Quality:** The dataset size should be as large as possible and the quality should be checked.

English datasets are chosen as the initial datasets because they are more accessible in all four tasks and have relatively larger size compared with datasets in other languages. We choose ROCStories (Mostafazadeh et al., 2016) for story generation, SQUAD 1.0 (Rajpurkar et al., 2016) for question generation, ByteCup [2] for title generation and CNN/DailyMail (Nallapati et al., 2016) for text summarization. These datasets are popular in the corresponding fields and have been verified to be high-quality by many works. Moreover, they are all under a permissive license. An overview of all task datasets is shown in Table 1.

3.2 Language Selection

The original datasets are in **English** (en) only and we want to extend them into a multiway parallel form. This means that all English texts should be translated into other languages, which will lead to high annotation costs. Thus, a state-of-the-art translator is leveraged to do the translation and then annotators are asked to correct the translated text. Considering this construction method, MTG should contain languages that (1) have good English-to-X translators and (2) are diverse in language family. Finally, **German** (de), **French** (fr), **Spanish** (es) and **Chinese** (zh) are chosen. German is from the same language branch as English while French and Spanish are from different ones. Chinese is more distant from the rest of languages in the language family tree.

Task	SG, QG, TG, Summ
For each language	
Rough training size	76k/61k/270k/164k
Annotated training size	15k/15k/15k/15k
Annotated development size	2k/2k/2k/2k
Annotated test size	3k/3k/3k/3k
For five languages (en, de, fr, es, zh)	
Total Annotated size	400k
Total dataset size	6.9m

Table 2: The number of samples in MTG. MTG consists of four subsets: *rough training*, *annotated training*, *development* and *test* set. The rough training set is filtered by back translating across five languages. The annotated training, development and test sets are corrected by human experts.

3.3 Data Collection

After determining the tasks and languages, we introduce the data collection process to get the MTG. The Google Translate[3] is used to translate the English datasets to the selected languages. To control the quality of translated texts, we back translate the text to English and filter the samples whose n-gram overlap ratios with the original English texts are lower than a certain threshold. Different threshold values (from 0.3 to 0.6 with 0.1 as step length) are tested and if it is set to 0.6, the training data size of QG will drop more than 60%. Thus we decide to use 0.5 as the threshold number to improve the quality of the filtered data while still maintaining more than 70% of the original training data.[4] Samples in four languages are aligned to ensure that the dataset is multiway parallel.

$20,000$ samples of each task and language are randomly selected for annotation under the premise

[2] https://www.biendata.xyz/competition/bytecup2018/

[3] https://translate.google.com/

[4] The detailed sizes of the filtered datasets with respect to different thresholds are included in appendix A.

Correlation	AdaBoost	DecisionTree	ExtraTree	GradientBoosting	Kneighbors	Linear	RandomForest	SVR	Bagging
Pearson	0.100	0.133	0.190	0.215	0.192	0.173	0.208	0.113	**0.240**
Correlation	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore-P	BERTScore-R	BERTScore-F1	Bagging
Pearson	0.180	0.142	0.163	0.144	0.122	0.142	0.176	0.162	**0.344**

Table 3: The correlation scores between automatic metric scores and human-annotated scores (the average scores of grammar, fluency and relevance). Upper part of the table shows the correlation scores of different regression algorithms in test set of all languages. The lower part demonstrates correlation scores of our ensemble score (the bagging regressor) and other classic automatic scores in test set without Chinese results because Meteor does not support Chinese.

of ensuring inter-language alignment. The annotators are required to further check the translated results based on the following rules: (1) **Semantic aligned** Whether the target text is meaningful and is fully semantic aligned with the source text. (2) **Fluency** Whether the translated text is grammatically correct. (3) **Style** Whether the translation follows the norms of local culture, language conventions, and gender-related words. If the translated text contradicts any of the above rules, annotators will correct it accordingly. The annotated data is then split to 15k/2k/3k as training/development/test subsets.

A team of 10 full-time experts[5] are hired to do the annotation, who are paid daily. Some part-time workers[6] are also employed to increase the annotation throughput, who are paid by the number of annotations. Each annotator is an expert in at least two languages (English and another target language). They are trained to correct translation errors according to the above rules, first a small number of samples for trial, these annotation results are re-checked by us and feedback is given to the annotators to help them understand the tasks better. After this annotation training process, the annotators start to annotate the dataset. For quality control, we sample 2% from the annotations and arrange for 9 experts to double-check them. Each example is assigned to two other experts and the data is qualified only if both of them agree on the annotation[7]. If more than 5% of the annotations fail, then all the data of that annotator for that day will be re-checked.

Then the multiway parallel generation benchmark MTG is finally completed. It contains four different generation tasks in five languages and its quality is improved by the incorporation of human annotation. However, the number of human-annotated data is still small due to cost concerns. Introducing more human-annotated data or carrying out extra filtering for machine-translated data can be future directions to further improve the quality of MTG. The statistics of MTG is shown in Table 2.

4 Experiments

In this section, we conduct extensive experiments to benchmark the difficulty of our proposed MTG via several state-of-the-art multilingual models under different scenarios.

4.1 Baseline Models

The performance of the following four popular multilingual pretrained models is explored[8]:

M-BERT Multilingual BERT (M-BERT) (Devlin et al., 2019) is a language model pretrained from monolingual corpora in 104 languages using Masked Language Modeling (MLM) task.

XLM The Cross-Lingual Language Model (XLM) (Lample and Conneau, 2019) is pretrained with Masked Language Modeling (MLM) task using monolingual data and Translation Language Modeling (TLM) task using parallel data.

mBART Multilingual BART (mBART) (Liu et al., 2020) is a pretrained encoder-decoder model using denoising auto-encoding objective on monolingual data over 25 languages.

mT5 Multilingual T5 (mT5) (Xue et al., 2020) is a multilingual variant of T5 (Raffel et al., 2020) formatting all tasks as text-to-text generation problems. mT5 is pretrained on a span-corruption version of Masked Language Modeling objective over 101 languages.

[5]There are 3 language experts for German, 3 for French, 4 for Spanish and 4 for Chinese

[6]There are 16 part-time workers who are participated in the German annotation, 39 for French, 4 for Spanish and 15 for Chinese.

[7]The grammar, expressions, and punctuation of the annotated text are completely correct and the expressions are in accordance with the foreign language.

[8]Detailed descriptions for models are included in Appendix B.

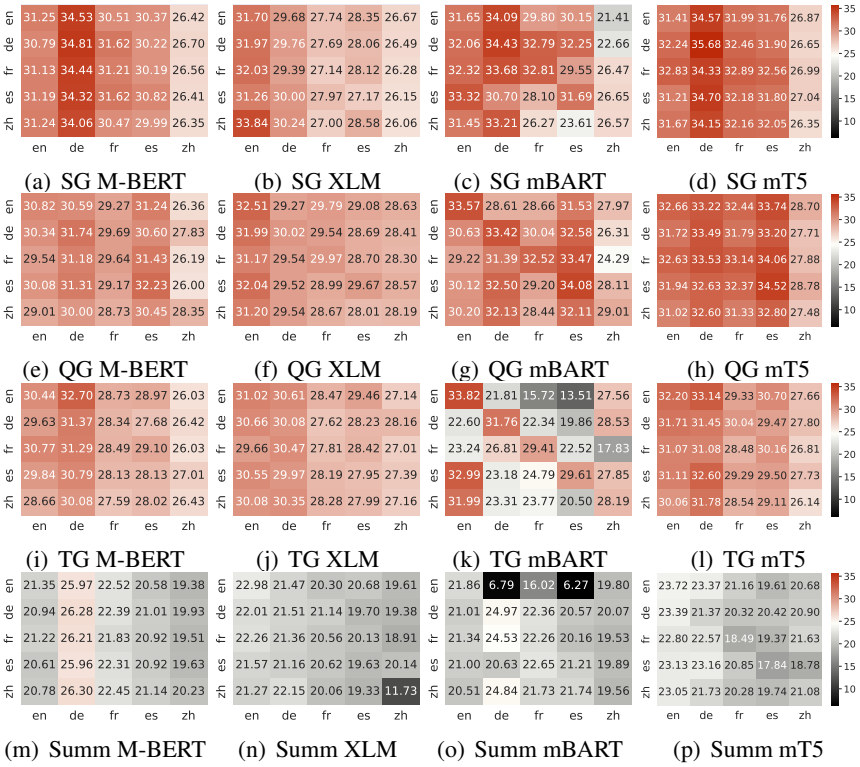

Figure 1: The cross-lingual ensemble metric results for four models in four tasks. Every cell of row lang1 and column lang2 means the result when the languages of input and output are lang1 and lang2 respectively. Deeper red represents better cross-lingual performance while deeper gray indicates worse performance.

4.2 Evaluation Metrics

In order to fully understand the model performance, the quality of generated texts is evaluated from different aspects, including metrics measuring the relevance between outputs and references (e.g., BLEU, ROUGE, and BERTScore) and metrics measuring the diversity of the generated texts (e.g., Distinct). Moreover, we propose a new ensemble metric leveraging relevance metrics to measure how close the generated text is to human writing. It not only has higher correlation scores with human judgments but also is capable of measuring model performances fairly between languages.

N-gram based Metrics N-gram based metrics evaluate the text-overlapping scores between the outputs and references. The following three metrics are used: (1) **BLEU** (Papineni et al., 2002) is a popular metric that calculates the word-overlap scores between the generated texts and gold-standard ones. We use the BLEU-4, which is the average score for unigram, bigram, trigram, and 4-gram. (2) **ROUGE** (Lin, 2004) is a recall-oriented metric that counts the number of overlapping units such as n-gram and word sequences between the produced texts and gold-standard ones. (3) **METEOR** (Banerjee and Lavie, 2005) relies on semantic features to predict the similarity scores between system hypotheses and human references.

Embedding based Metrics The embedding-based metrics can, to a large extent, capture the semantic-level similarity between the generated text and the ground truth. **BERTScore** (Zhang et al., 2019) computes the similarity of candidate and reference as a sum of cosine similarities of tokens using BERT contextual embeddings.

Diversity Metrics We also employ the distinct metric (Li et al., 2016), which calculates the proportion of the distinct n-grams in all the system hypotheses and can be used to evaluate the diversity of the generated texts.

Human Evaluation Human evaluation is also leveraged to better estimate the quality of model outputs. Specifically, 30 cases are randomly sampled from the test set for each task and language while ensuring all 30 cases are aligned among five languages, and then they are presented to human annotators with the model outputs. The generated texts are evaluated under task-agnostic and task-specific aspects. Task-agnostic aspects include **Grammar**, **Fluency**, **Relevance** and **Language**

Task	Model	BLEU		ROUGE-L		METEOR		BERTScore		Distinct-1		Ensemble	
		mono	multi	mono	multi	mono	multi	mono	multi	mono	multi	mono	multi
SG	M-BERT	2.486	**2.836**	16.680	**17.240**	0.139	**0.140**	0.741	**0.743**	0.952	**0.959**	30.891	**30.987**
	XLM	**4.026**	2.992	**24.520**	22.820	**0.145**	0.144	**0.754**	0.744	**0.967**	0.967	28.364	**28.449**
	mBART	4.514	**4.880**	19.320	**19.920**	0.149	**0.156**	0.759	**0.762**	**0.985**	0.983	31.430	**31.907**
	mT5	2.668	**3.832**	16.280	**18.620**	0.126	**0.145**	0.751	**0.759**	**0.976**	0.974	**31.623**	31.482
QG	M-BERT	8.266	**9.980**	27.340	**29.520**	0.240	**0.262**	0.778	**0.785**	0.938	**0.944**	30.553	30.526
	XLM	**16.472**	15.264	**41.100**	40.600	**0.305**	0.298	**0.810**	0.809	0.966	**0.967**	**30.072**	29.979
	mBART	16.256	**17.624**	36.640	**38.140**	0.298	**0.315**	0.811	**0.817**	0.981	**0.983**	32.522	**32.961**
	mT5	15.792	**17.700**	34.100	**37.680**	0.294	**0.313**	0.806	**0.818**	0.977	**0.979**	32.257	**32.944**
TG	M-BERT	9.524	**10.550**	25.440	**26.360**	0.214	**0.228**	0.749	**0.754**	0.930	**0.957**	28.971	**29.422**
	XLM	11.144	**11.926**	26.960	**28.660**	0.236	**0.248**	0.752	**0.759**	**0.946**	0.941	28.808	**29.063**
	mBART	14.726	**14.786**	31.680	**32.120**	0.257	**0.260**	0.773	**0.775**	0.966	**0.968**	**30.556**	30.322
	mT5	11.336	**13.546**	26.460	**29.400**	0.223	**0.257**	0.753	**0.767**	**0.959**	0.956	29.556	**30.010**
Summ	M-BERT	9.766	**10.956**	31.280	**32.220**	0.221	**0.232**	0.748	**0.751**	0.787	**0.815**	**22.122**	22.018
	XLM	9.486	**11.830**	30.160	**34.740**	**0.235**	0.235	0.729	**0.755**	**0.814**	0.772	19.281	**20.770**
	mBART	**12.858**	12.792	**32.940**	32.920	0.256	**0.257**	0.750	**0.750**	0.796	**0.803**	21.972	**22.292**
	mT5	5.022	**6.090**	25.060	**27.980**	0.145	**0.162**	0.724	**0.741**	0.826	**0.870**	20.499	**21.826**

Table 4: Automatic scores averaged across five languages for four models on four tasks. Mono and multi mean models are trained in monolingual and multilingual setting respectively. Higher scores between monolingual and multilingual results are bolded.

Fusion. The former three aspects are scored from 1 to 5 while the language fusion score is set to 1 if all tokens of a model-generated text are in the target language and 0 otherwise.

Besides task-agnostic aspects, the generated text is also evaluated under task-specific aspects. For title generation and summarization, coverage measures the degree to which the generated text covers the main content of the document. Correspondence for question generation measures the extent to which the generated question is matched with both document and answer. For story generation, we further evaluate whether the generated story is logically feasible. All task-specific aspects are scored from 1 to 5.

Ensemble Metric Some N-gram based metrics such as BLEU and ROUGE largely depend on the tokenizer for specific languages. For example, BLEU scores for Chinese outputs are relatively high because it simply uses a character-level tokenizer. This causes unfair comparison between different languages. To this end, we propose an ensemble metric that evaluates the degree to which a piece of text resembles manual writing. It not only enables fair comparison between languages but is also proved to have a better correlation with human-annotated scores at the end of this subsection. We first average the grammar, fluency and relevance scores as targets, then normalize the automatic metrics and human scores among every language to eliminate the score discrepancy between languages. Three relevance metrics (BLEU, ROUGE-L, and BERTScore-F1) are gathered as features. The samples are split into training, development and test sets.

After comparing different regression models' performance as shown in the upper part of Table 3 , we finally choose bagging regression model (Breiman, 1996) as the ensemble metric. Moreover, the bagging ensemble metric shows a higher correlation with human-annotated scores compared with other relevance automatic metrics as shown in the lower part of Table 3.

4.3 Evaluation Scenarios

To validate the effect of different experimental settings on model performance, several state-of-the-art multilingual models are studied under four evaluation scenarios.

Monolingual fine-tuning The pretrained model is tuned for a downstream task using the training data for a specific language and evaluated on the test set for the same language.

Multilingual fine-tuning The pretrained model is jointly fine-tuned with data in all languages for a specific task. Different from the monolingual fine-tuning setting, there is only one model for each downstream task, which can serve all languages.

Cross-lingual generation Since MTG is multi-way parallel, it can be reorganized to create input-output pairs that belong to different languages. In this paper, we make use of the multiway parallel data to do the supervised cross-lingual training, e.g., for English centric cross-lingual training, we take the English source as the input and the parallel German, French, Spanish, Chinese target as the out-

put. Then we evaluate the model on same setting (en->de, en->es, en->fr, en->zh). The cross-lingual generation performances on all $5 * 4$ directions are evaluated.

Zero-shot transfer We also try to explore the zero-shot ability of multilingual pretrained models on the four tasks. The model is fine-tuned on a specific task with English input and output. Then it is used to generate output in other languages with a given language tag.

5 Results

5.1 Monolingual and Cross-lingual

This section displays the monolingual and cross-lingual model comparison to explore their performances in different tasks and languages. Figure 1 contains the five language-centric cross-lingual and monolingual results. Several conclusions can be drawn from the results:

The performance of Cross-lingual is better than monolingual in some cases. As shown in Figure 1, model performances on ensemble scores in cross-lingual setting exceed those in monolingual setting frequently (e.g., the monolingual result of French underperforms the English to French cross-lingual result in Figure 1(b)). This is because the cross-lingual models are trained with more data (e.g., the English centric cross-lingual model is trained with en->de, en->fr, en->es, en->zh data), and the data from different cross-lingual directions can sometimes benefit from each other thus improving the model performance.

Chinese text generation is challenging in cross-lingual setting. As illustrated in Figure 1, nearly all models obtain inferior scores when generating Chinese text. Also, model results on Chinese inputs are usually worse than results on inputs in other languages. The wide discrepancies in grammar and vocabulary between Chinese and other languages lead to the poor performance of cross-lingual generation when either the target language or source language is Chinese.

Multilingual pretrained models obtain lower scores on the Summarization task. Compared with other tasks, summarization task requires longer output, which increases the difficulty of text generation, thus causing poor performance both in cross-lingual and monolingual settings.

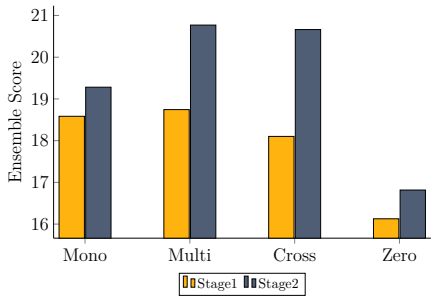

Figure 2: The different stage performances averaged across five languages of XLM in summarization under various settings. Here stage1 represents models trained only on rough training data while stage2 represents models further trained on human-annotated training data based on models in stage1.

5.2 Monolingual and Multilingual

In addition to cross-lingual analysis, we also explore the performance difference between models trained in monolingual and multilingual settings. Table 4 displays the monolingual and multilingual training results for four models in four tasks.

In most cases, multilingual training can improve model performance on relevance. As shown in Table 4, 75 out of 96 multilingual results outperform the monolingual counterparts on various relevance metrics in different tasks. The reason is that the multilingual data in MTG is fully parallel across all five languages and every sample has semantically aligned counterparts in other languages. It makes better semantic fusion among different languages, thus boosting the multilingual training performance.

The advantages of multilingual training are not obvious on diversity measured by distinct-1. Especially in the story generation task, 3 out of 4 models obtain better distinct-1 scores in monolingual setting than in multilingual one. Diversity can not be improved by semantic sharing across languages especially when the samples of them are multiway parallel. This is because the multiway parallel dataset with the semantic aligned samples repeating in different languages encourages models to generate similar texts to some extent.

5.3 Zero-shot results

To test the cross-lingual generation ability of multilingual pretrained models when no direct cross-lingual training data are provided, we evaluate the zero-shot cross-lingual generation performance.

Table 5 presents the zero-shot results for XLM in four tasks. It demonstrates that the multilin-

Task	Language	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore	Distinct-1	Distinct-2	Ensemble
SG	en->de	0.02/3.20	7.20/27.20	0.20/4.00	7.20/25.80	0.05/0.14	0.63/0.73	0.47/0.96	0.50/1.00	18.90/29.70
	en->fr	0.02/4.23	5.90/28.10	0.20/6.30	5.90/26.40	0.04/0.20	0.63/0.74	0.38/0.95	0.41/0.99	14.30/27.70
	en->es	0.09/3.38	8.70/26.30	0.40/4.60	8.50/24.80	0.04/0.14	0.65/0.74	0.52/0.96	0.55/0.99	16.90/28.40
	en->zh	0.00/5.79	0.00/28.80	0.00/8.80	0.00/26.80	-	0.45/0.67	0.61/0.99	0.57/0.34	16.60/26.70
QG	en->de	1.96/10.41	18.10/38.70	2.40/14.70	17.60/37.20	0.10/0.25	0.73/0.78	0.94/0.97	0.98/1.00	29.80/29.30
	en->fr	2.16/14.70	16.80/42.80	2.90/19.00	16.20/39.60	0.08/0.35	0.74/0.80	0.94/0.95	0.99/0.99	28.60/29.80
	en->es	7.46/16.93	25.50/49.50	8.70/22.40	23.90/46.80	0.18/0.37	0.76/0.83	0.94/0.95	0.99/1.00	28.50/29.10
	en->zh	0.00/16.07	0.00/43.10	0.00/22.90	0.00/37.90	-	0.44/0.73	0.10/1.00	0.08/1.00	16.40/28.60
TG	en->de	2.58/9.15	13.40/26.90	4.40/11.10	12.50/24.30	0.12/0.22	0.67/0.73	0.83/0.95	0.88/0.99	26.30/30.60
	en->fr	3.26/11.54	13.90/33.80	4.50/14.70	12.70/29.00	0.12/0.30	0.69/0.75	0.89/0.91	0.93/0.99	25.20/28.50
	en->es	4.90/12.45	21.20/36.30	7.40/15.70	18.50/31.10	0.17/0.31	0.71/0.76	0.88/0.91	0.94/0.99	24.50/29.50
	en->zh	0.01/15.44	0.00/34.50	0.00/19.40	0.00/29.90	-	0.45/0.69	0.37/0.98	0.22/0.58	16.70/27.10
Summ	en->de	1.85/8.36	15.40/34.70	2.90/11.70	14.50/31.10	0.08/0.20	0.65/0.72	0.61/0.81	0.78/0.97	18.50/21.50
	en->fr	1.29/11.79	13.70/39.90	2.60/15.80	13.00/35.50	0.07/0.29	0.68/0.75	0.64/0.75	0.82/0.94	18.60/20.30
	en->es	4.18/11.93	22.50/41.00	5.80/15.60	20.30/36.60	0.14/0.29	0.69/0.75	0.64/0.74	0.82/0.95	17.30/20.70
	en->zh	0.00/14.58	0.00/42.20	0.00/20.40	0.00/38.70	-	0.42/0.71	0.68/0.84	0.27/0.94	12.80/19.60

Table 5: English centric zero-shot and cross-lingual results for XLM on four tasks. Scores on the left and right side of each cell represent the zero-shot and cross-lingual results respectively.

gual pretrained model XLM still lacks the ability to generate high-quality cross-lingual output in zero-shot scenario. **Moreover, English to Chinese and French zero-shot generation shows inferior performance.**[9] The performance decline is rather salient when generating Chinese text. This is because Chinese and French (especially Chinese) are distant from English in the language family tree. **On the other hand, zero-shot results underperform cross-lingual results** which further emphasizes the importance of direct cross-lingual training data for cross-lingual text generation.

5.4 Pseudo and Annotated Data

To answer the question "Does the 400k annotated training data help the model generate better? ", we use the rough training data filtered by back translation for the first stage fine-tuning and the annotated training data for the second stage. The ablation study results on the two-step fine-tuning in summarization under all evaluation scenarios with XLM are illustrated in Figure 2.

The extra human-annotated data boost model performance by at least 3.8% on the ensemble metric. We also make a T-test and prove that the improvement of annotated training data is significant in all settings.[10] **It demonstrates that although the number of annotated data is small, it can significantly improve the performance.** It also highlights the necessity of human-annotated multilingual data compared with pseudo-parallel data via machine translation.

[9]Zero-shot results show the same trend as shown in Table 17 in Appendix.

[10]The t-test details are shown in Appendix C.

Setting	Model	Gram.	Flu.	Rel.	lang fuse	task spec.
SG	mono	4.69	4.81	3.75	1.00	3.79
	multi	4.71	4.80	3.67	1.00	4.02
	cross	4.18	4.23	3.49	0.95	2.53
	zero	4.15	4.18	3.27	0.18	3.00
QG	mono	4.66	4.69	3.03	0.99	3.95
	multi	4.69	4.67	3.06	0.97	4.11
	cross	4.30	4.30	2.70	0.95	2.64
	zero	3.35	4.26	3.18	0.19	3.09
TG	mono	4.53	4.51	3.09	0.96	3.71
	multi	4.66	4.65	3.18	0.93	3.17
	cross	3.73	3.64	2.63	0.90	1.85
	zero	3.52	4.15	3.51	0.18	1.43
Summ	mono	4.19	3.99	3.71	0.68	3.71
	multi	4.19	4.02	3.78	0.64	3.60
	cross	2.14	2.22	2.23	0.68	2.05
	zero	1.57	1.54	1.58	0.03	1.59

Table 6: Human evaluation scores averaged on five languages for mBART on four tasks. 'Gram.', 'Flu.', 'Rel.', 'Lang Fu.', 'Task Spec.' indicates **Grammar**, **Fluency**, **Relevance**, **Language Fusion** and **Task Specific** scores respectively.

5.5 Human evaluation

Table 6 presents the human evaluation scores for mBART in four tasks. Multilingual training results can surpass the monolingual results in QG, TG and Summ on relevance. In terms of task-specific score, multilingual results are also superior in SG and QG. This is consistent with the conclusion in Sec. 5.2. On the other hand, language fusion scores in zero-shot setting are extremely low, indicating the pretrained models still lack the ability to generate texts in correct language in zero-shot setting.

6 Discussions

Considering the annotation cost, it is not realistic to construct a multiway text generation dataset with all data annotated by human. As a consequence, most of the non-English data in MTG are automatically translated from their English counterparts. Although the n-gram consistency check

when round-trip translating the data can guarantee the quality of them to some extent, some translation errors are inevitable. MTG with more annotated data and with data filtered by more reliable methods will be explored in the future.

On the other hand, human often gives an overall evaluation of a generated text rather than measuring it in fine-grained aspects of grammar, fluency and relevance. Thus we try to propose a metric measuring how a text resembles human writing and consider grammar, fluency and relevance as a whole. This metric may not be perfect, but it is a promising direction as there does not exist a really reliable text generation metric nowadays.

7 Conclusion

In this paper, we propose a multilingual multiway benchmark MTG for text generation. It contains four typical generation tasks: story, question, title generation and text summarization. The key feature of MTG is that it has multiway parallel data across five diverse languages: English, German, French, Spanish and Chinese. It provides the benchmark with the ability to create cross-lingual data between any two languages and makes the semantic fusion between languages easier. On the other hand, it provides more evaluation scenarios, such as multilingual training, cross-lingual generation and zero-shot transfer. We also benchmark state-of-the-art multilingual pretrained models on our MTG from different metrics (including a newly proposed ensemble metric) to explore its features and promote research in multilingual text generation.

8 Ethics Consideration

Since we propose a new multilingual text generation benchmark MTG, we solve some possible ethic considerations in this section.

English dataset We choose ROCStories, SQUAD 1.0, ByteCup and CNN/DailyMail as the English datasets for story, question, title generation and text summarization tasks. All of them are available for research use under their licenses. They can be downloaded free from their websites[11]. We ensure that these datasets are only used for academic research and the dataset construction process complies with the intellectual property and privacy rights of the original authors. Also, our proposed benchmark suite MTG should only be used for academic research purposes.

Annotation process As described in Sec. 3.3, we hire some full-time and part-time language experts to do the annotation. Full-time experts are paid \$40 per day and part-time annotators are paid \$0.2 per example[12]. Their salary is higher than the local average hourly minimum wage. All annotators are aware of any risk of harm associated with their participation. The annotation process is in compliance with the intellectual property and privacy rights of the recruited annotators. The annotation protocol is proved by the legal department inside the company.

Risk Concern In this paper we propose a new ensemble metric measuring to what degree is the generated text close to human-level. The further pursue for more human-like multilingual generation will possibly raise safety concerns.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24(2):123–140.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of

[11] ROCStories requires for some necessary contact information

[12] Full-time employees work at most 8 hours per day, and the local minimum hourly wage is \$3.7. The part-time annotators can produce at least 20 examples per hour.

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40b: Multilingual language model dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2440–2452.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-factr: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959.

Rong Jin and Alexander G Hauptmann. 2002. A new probabilistic model for title generation. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization. *arXiv preprint arXiv:2010.03093*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yichong Leng, Xu Tan, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. 2019. Unsupervised pivot translation for distant languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 175–183.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

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 pretraining, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *arXiv preprint arXiv:2007.15207*.

Inderjeet Mani. 2001. *Automatic summarization*, volume 3. John Benjamins Publishing.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Ça glar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL 2016*, page 280.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Danqing Wang, Jiaze Chen, Hao Zhou, Xipeng Qiu, and Lei Li. 2021. Contrastive aligned joint learning for multilingual summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2739–2750.

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*.

Rong Ye, Wenxian Shi, Hao Zhou, Zhongyu Wei, and Lei Li. 2020. Variational template machine for data-to-text generation. *arXiv e-prints*, pages arXiv–2002.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3054–3064.

A Back Translation Threshold Testing

The detailed data sizes of back translation filtered datasets for different tasks are presented in Table 7.

B Experimental settings

The overall statistics for multilingual pretrained models are presented in Table 8 and the detailed descriptions for them are as follows:

1**M-BERT** Multilingual BERT (M-BERT) (Devlin et al., 2019) is a single language model pre-trained from monolingual corpora in 104 languages using Masked Language Modeling (MLM) task. M-BERT leverages a shared vocabulary of 110k WordPiece tokens and has 12 layers with 172M parameters totally.

XLM The Cross-Lingual Language Model (XLM) (Lample and Conneau, 2019) is pre-trained simultaneously with Masked Language Model (MLM) task using monolingual data and Translation Language Model (TLM) task using parallel data. XLM has a shared vocabulary of 200k byte-pair encoded (BPE) subwords (Sennrich et al., 2016) and 16 layers totaling 570M parameters.

1**mBART** Multilingual BART (mBART) (Liu et al., 2020) is a pre-trained encoder-decoder model using denoising auto-encoding objective on monolingual data over 25 languages. mBART has a shared vocabulary of 250k tokens leveraging Sentence Piece tokenization scheme. mBART consists of 12-layer encoder and 12-layer decoder with a total of 680M parameters.

mT5 Multilingual T5 (mT5) (Xue et al., 2020) is a multilingual variant of T5 (Raffel et al., 2020) leveraging a text-to-text format. mT5 is pre-trained on a span-corruption version of Masked Language Modeling objective over 101 languages. It is composed of 24-encoder layers and 24 decoder layers with 13B parameters.

We use the encoder-decoder architecture for our generation tasks. Among the models described above, mBART and mT5 have been pretrained for generation tasks, but M-BERT and XLM are only pretrained for encoder representations. Therefore, we initialize the decoder with the encoder parameters for M-BERT and XLM. During the pretraining phase, there are no language tags in M-BERT and mT5. Thus we manually add the language tag at the beginning of the source and target for M-BERT and add the target language tag to the beginning of source for mT5.

We adjust the input format for each task. For QG, we append the answer to the passage and insert a special token to separate them. For SG, we take the beginning four sentences as the source and make the last sentence as the target.

We take a two-step finetuning to make full use of our MTG benchmark. We first use the large rough parallel training data to train our models on the downstream tasks for 20 epochs, and then finetune the models on the small annotated training data to further improve the generation performance for 10 epochs. We evaluate the model for every 2000 steps and use the loss on development to choose the best model. The batch size is 32. The learning rate and optimizer parameters are set to the default parameters for each model. All models are trained with 32GB Tesla-V100.

Threshold	QG	TG	SG	Summ
0	82306	393792	88161	287083
0.3	80836	355034	88158	243698
0.4	79390	333461	88077	217777
0.5	71819	280376	87003	164355
0.6	32261	144109	75892	58060

Table 7: The data sizes of datasets filtered by back translation with respect to different thresholds.

Models	Arch	# langs	# vocab	# layers	# params
M-BERT	enc	104	110k	12	172M
XLM	enc	17	200k	16	570M
mBART	enc-dec	25	250k	12	680M
mT5	enc-dec	101	250k	24	13B

Table 8: The overall statistics for multilingual pre-trained models. Arch means the architectures of models. # vocab means the vocabulary sizes of models. # langs, # layers and # params mean the number of languages, layers and parameters respectively.

Tasks		Mono	Multi	Cross	Zero
SG	stage1	0.268	0.270	0.258	0.125
	stage2	**0.284**	**0.284**	**0.289**	**0.167**
	p-value	0.000	0.000	0.000	0.000
QG	stage1	0.286	0.287	0.279	0.235
	stage2	**0.301**	**0.300**	**0.295**	**0.258**
	p-value	0.000	0.000	0.000	0.000
TG	stage1	0.257	0.270	0.268	0.223
	stage2	**0.288**	**0.291**	**0.289**	**0.232**
	p-value	0.000	0.000	0.000	0.003
Summ	stage1	0.186	0.187	0.181	0.161
	stage2	**0.193**	**0.208**	**0.207**	**0.168**
	p-value	0.001	0.000	0.000	0.004

Table 9: The average ensemble metric scores for XLM for stage1 and stage2 in four tasks in four settings and the corresponding t test p-values. Here stage1 represents models trained only on rouge training data while stage2 represents models further trained on human-annotated training data based on models in stage1. The bold cell means the significantly higher score between stage1 and stage2 scores.

C Significant Test Results

The average ensemble metric scores for stage1 and stage2 in four tasks and the corresponding significant test p-values are displayed in Table 9. As it shows, adding human-annotated training data can always improve the model performance under different settings. The improvements are significant in all settings.

D Experimental Results

We present detailed experimental results of our four baseline models under four different evaluation settings here.

Task	Model	Language	N-gram-based			Embedding-based	Diversity		Ours
			BLEU	ROUGE-L	METEOR	BERTScore	Distinct-1	Distinct-2	Ensemble
SG	M-BERT	en->en	2.56	18.8	0.103	0.894	0.917	0.99	31.254
		de->de	2.27	13.4	0.131	0.714	0.944	0.994	34.812
		fr->fr	1.38	12.7	0.201	0.715	0.945	0.997	31.209
		es->es	1.81	13.5	0.121	0.72	0.955	0.996	30.825
		zh->zh	4.41	25	-	0.661	1	0.124	26.354
	XLM	en->en	3.71	20.5	0.107	0.895	0.968	0.995	31.701
		de->de	3.02	25.3	0.14	0.729	0.966	0.995	29.758
		fr->fr	4.28	25.6	0.196	0.741	0.948	0.987	27.136
		es->es	3.41	24.9	0.135	0.736	0.959	0.989	27.168
		zh->zh	5.71	26.3	-	0.667	0.996	0.262	26.057
	mBART	en->en	4.2	21.7	0.114	0.902	0.98	1	31.65
		de->de	3.64	15.5	0.142	0.733	0.982	1	34.429
		fr->fr	4	16.5	0.199	0.748	0.985	1	32.808
		es->es	4.21	15.6	0.14	0.741	0.982	1	31.688
		zh->zh	6.52	27.3	-	0.673	0.997	0.31	26.575
	mT5	en->en	2.25	17.5	0.097	0.896	0.981	0.998	31.405
		de->de	1.67	11.8	0.12	0.721	0.971	0.992	35.676
		fr->fr	2.44	14	0.168	0.738	0.96	0.98	32.888
		es->es	2.36	13.1	0.118	0.735	0.966	0.992	31.801
		zh->zh	4.62	25	-	0.664	1	0.171	26.346
SG	M-BERT	en->en	12.2	39.2	0.19	0.896	0.907	0.988	30.817
		de->de	6.06	22.2	0.195	0.744	0.931	0.992	31.737
		fr->fr	5.35	20.8	0.258	0.749	0.923	0.994	29.639
		es->es	7.43	23.3	0.315	0.801	0.929	0.992	32.226
		zh->zh	10.29	31.2	-	0.698	0.999	0.902	28.348
	XLM	en->en	19.69	44.8	0.231	0.915	0.958	0.997	32.513
		de->de	10.59	36.4	0.247	0.776	0.974	0.998	30.017
		fr->fr	16.43	40.2	0.358	0.799	0.947	0.99	29.97
		es->es	19.66	47.4	0.382	0.832	0.953	0.995	29.667
		zh->zh	15.99	36.7	-	0.726	0.996	0.999	28.194
	mBART	en->en	20.9	47.4	0.235	0.92	0.976	0.999	33.566
		de->de	11.69	28.7	0.241	0.78	0.986	1	33.424
		fr->fr	14.97	32.9	0.336	0.793	0.977	0.999	32.524
		es->es	17.61	36.1	0.38	0.835	0.969	0.999	34.085
		zh->zh	16.11	38.1	-	0.728	0.999	0.997	29.01
	mT5	en->en	18.72	42.9	0.216	0.914	0.968	0.999	32.658
		de->de	11.07	24.5	0.231	0.774	0.977	0.999	33.492
		fr->fr	16.52	32.3	0.345	0.798	0.972	0.998	33.143
		es->es	18.39	36.2	0.384	0.838	0.971	0.998	34.515
		zh->zh	14.26	34.6	-	0.707	0.996	0.999	27.478
SG	M-BERT	en->en	14.46	36.2	0.196	0.887	0.931	0.988	30.435
		de->de	6.88	16.7	0.175	0.713	0.943	0.995	31.367
		fr->fr	6.59	21.1	0.22	0.727	0.89	0.984	28.491
		es->es	8.69	25.4	0.264	0.748	0.892	0.988	28.126
		zh->zh	11	27.8	-	0.67	0.993	0.42	26.435
	XLM	en->en	15.52	34.3	0.199	0.889	0.97	0.996	31.022
		de->de	7.1	20.3	0.182	0.714	0.959	0.996	30.084
		fr->fr	10.25	26.6	0.279	0.742	0.915	0.988	27.814
		es->es	11.4	29.3	0.284	0.756	0.912	0.994	27.954
		zh->zh	11.45	24.3	-	0.659	0.976	0.614	27.165
	mBART	en->en	21.78	41.9	0.231	0.905	0.984	1	33.816
		de->de	9.29	22.5	0.2	0.733	0.977	0.999	31.759
		fr->fr	11.86	29	0.278	0.757	0.953	0.999	29.413
		es->es	13.97	32.1	0.319	0.769	0.931	0.998	29.605
		zh->zh	16.73	32.9	-	0.699	0.985	0.547	28.187
	mT5	en->en	15.27	35.4	0.194	0.893	0.979	0.996	32.198
		de->de	7.71	18.1	0.174	0.714	0.962	0.994	31.454
		fr->fr	9.8	24.7	0.245	0.74	0.941	0.992	28.479
		es->es	11.73	27.5	0.277	0.753	0.925	0.993	29.504
		zh->zh	12.17	26.6	-	0.667	0.986	0.52	26.145
SG	M-BERT	en->en	14.7	38.2	0.177	0.87	0.817	0.985	21.351
		de->de	7.42	24.5	0.194	0.713	0.803	0.98	26.276
		fr->fr	7.16	27.4	0.252	0.726	0.77	0.979	21.831
		es->es	8.89	31.3	0.26	0.739	0.762	0.979	20.922
		zh->zh	10.66	35	-	0.691	0.783	0.961	20.23
	XLM	en->en	16.23	38.1	0.194	0.878	0.777	0.968	22.984
		de->de	7.98	29.2	0.189	0.712	0.803	0.956	21.505
		fr->fr	11.48	34.3	0.281	0.743	0.746	0.934	20.562
		es->es	11.33	34.8	0.275	0.743	0.743	0.953	19.629
		zh->zh	0.41	14.4	-	0.569	1	1	11.726
	mBART	en->en	17.33	39.9	0.193	0.875	0.832	0.993	21.863
		de->de	9.7	26.3	0.226	0.714	0.809	0.988	24.97
		fr->fr	12.09	31.9	0.308	0.739	0.778	0.987	22.264
		es->es	13.23	33.3	0.298	0.743	0.742	0.985	21.207
		zh->zh	11.94	33.3	-	0.677	0.818	0.984	19.556
	mT5	en->en	7.46	31	0.119	0.869	0.846	0.941	23.722
		de->de	3.45	18.6	0.117	0.682	0.824	0.921	21.368
		fr->fr	3.45	23.2	0.16	0.708	0.826	0.936	18.487
		es->es	4.02	24.7	0.185	0.704	0.769	0.912	17.837
		zh->zh	6.73	27.8	-	0.656	0.864	0.907	21.083

Table 10: The whole results under the monolingual evaluation scenarios.

Task	Model	Language	N-gram-based			Embedding-based	Diversity		Ours
			BLEU	ROUGE-L	METEOR	BERTScore	Distinct-1	Distinct-2	Ensemble
SG	M-BERT	en->en	2.93	19.8	0.106	0.895	0.937	0.993	31.793
		de->de	2.77	14.6	0.139	0.72	0.944	0.996	34.933
		fr->fr	1.65	13.7	0.196	0.72	0.949	0.998	30.998
		es->es	1.85	13.4	0.119	0.718	0.965	0.997	30.234
		zh->zh	4.98	24.7	-	0.661	0.998	0.374	26.977
	XLM	en->en	3.56	20.2	0.106	0.896	0.966	0.997	31.594
		de->de	3.07	25.4	0.145	0.728	0.957	0.995	29.464
		fr->fr	3.89	26.1	0.192	0.745	0.944	0.982	28.127
		es->es	3.52	24.7	0.132	0.738	0.969	0.994	27.465
		zh->zh	0.92	17.7	-	0.611	1	0	25.593
	mBART	en->en	4.6	22.4	0.117	0.903	0.974	1	32.396
		de->de	4	16.1	0.145	0.735	0.982	1	34.724
		fr->fr	4.79	17.3	0.215	0.751	0.981	1	32.955
		es->es	4.17	15.9	0.145	0.745	0.982	1	32.43
		zh->zh	6.84	27.9	-	0.677	0.998	0.275	27.03
	mT5	en->en	3.49	20.4	0.109	0.9	0.97	0.994	32.014
		de->de	3.01	14.6	0.139	0.731	0.977	0.993	34.758
		fr->fr	3.38	16.1	0.196	0.746	0.96	0.983	31.854
		es->es	3.49	14.9	0.134	0.742	0.966	0.989	32.188
		zh->zh	5.79	27.1	-	0.674	0.999	0.198	26.596
SG	M-BERT	en->en	14.47	41.4	0.204	0.9	0.923	0.992	31.192
		de->de	7.75	24.8	0.226	0.759	0.95	0.995	31.709
		fr->fr	6.55	22.2	0.279	0.753	0.917	0.994	29.918
		es->es	9.06	25.3	0.34	0.803	0.933	0.992	32.032
		zh->zh	12.07	33.9	-	0.708	0.998	0.916	27.779
	XLM	en->en	18.73	44.4	0.223	0.914	0.957	0.996	32.732
		de->de	9.86	36.2	0.245	0.778	0.976	0.997	29.317
		fr->fr	14.82	39.3	0.347	0.797	0.95	0.991	29.82
		es->es	17.38	46.5	0.375	0.829	0.953	0.996	29.435
		zh->zh	15.53	36.6	-	0.727	0.998	1	28.593
	mBART	en->en	21.73	47.9	0.242	0.921	0.976	0.999	34.481
		de->de	13.46	31.2	0.262	0.791	0.988	1	33.59
		fr->fr	16.13	34.4	0.35	0.8	0.978	0.999	33.181
		es->es	19.17	38.4	0.407	0.842	0.974	0.999	34.857
		zh->zh	17.63	38.8	-	0.733	0.997	0.993	28.698
	mT5	en->en	20.9	46.8	0.232	0.92	0.971	0.999	33.449
		de->de	13.22	30.2	0.264	0.789	0.984	1	34.218
		fr->fr	17.08	33.6	0.356	0.801	0.971	0.998	33.392
		es->es	19.45	37.8	0.398	0.842	0.97	0.998	34.95
		zh->zh	17.85	40	-	0.737	0.997	0.999	28.712
SG	M-BERT	en->en	15.87	37.1	0.209	0.891	0.967	0.998	31.467
		de->de	7.59	17.8	0.189	0.719	0.952	0.997	31.835
		fr->fr	8.46	23.1	0.234	0.738	0.942	0.997	28.474
		es->es	10.08	26.3	0.279	0.755	0.93	0.997	29.1
		zh->zh	10.75	27.5	-	0.669	0.993	0.407	26.236
	XLM	en->en	14.91	34.8	0.202	0.89	0.97	0.995	31.291
		de->de	7.86	22.3	0.198	0.719	0.944	0.986	30.049
		fr->fr	11.03	28.2	0.296	0.749	0.903	0.982	28.037
		es->es	11.59	30.2	0.295	0.758	0.907	0.993	28.062
		zh->zh	14.24	27.8	-	0.681	0.982	0.614	27.875
	mBART	en->en	21.91	42.9	0.233	0.907	0.984	0.999	33.625
		de->de	9.58	22.9	0.208	0.735	0.975	1	31.922
		fr->fr	11.75	29.1	0.274	0.759	0.958	0.999	29.147
		es->es	14.11	32.7	0.325	0.775	0.939	0.999	29.714
		zh->zh	16.58	33	-	0.699	0.983	0.641	27.202
	mT5	en->en	17.54	38.3	0.217	0.899	0.978	0.997	32.212
		de->de	9.19	21	0.209	0.728	0.961	0.994	31.694
		fr->fr	12.03	27.5	0.289	0.754	0.937	0.994	29.376
		es->es	14	30.2	0.313	0.766	0.918	0.995	29.539
		zh->zh	14.97	30	-	0.687	0.984	0.576	27.228
SG	M-BERT	en->en	15.88	39	0.182	0.873	0.827	0.989	21.642
		de->de	8.21	25.2	0.204	0.716	0.822	0.987	25.818
		fr->fr	8.89	28.9	0.267	0.73	0.786	0.985	22.296
		es->es	10.42	32.3	0.276	0.742	0.765	0.983	20.451
		zh->zh	11.38	35.7	-	0.692	0.877	0.839	19.884
	XLM	en->en	16.05	38	0.193	0.877	0.763	0.952	22.182
		de->de	7.82	29	0.187	0.712	0.785	0.934	21.8
		fr->fr	11.24	34.9	0.286	0.747	0.746	0.937	20.087
		es->es	11.06	35	0.273	0.742	0.726	0.939	20.473
		zh->zh	12.98	36.8	-	0.698	0.842	0.902	19.306
	mBART	en->en	17.64	39.8	0.196	0.875	0.827	0.993	22.912
		de->de	9.46	26.3	0.222	0.714	0.816	0.99	24.926
		fr->fr	12.2	32	0.309	0.741	0.783	0.987	22.549
		es->es	13.5	33.8	0.301	0.745	0.752	0.986	20.927
		zh->zh	11.16	32.7	-	0.675	0.836	0.982	20.146
	mT5	en->en	8.59	33.9	0.13	0.876	0.882	0.982	24.324
		de->de	4.05	21.4	0.132	0.698	0.874	0.974	22.8
		fr->fr	4.09	25.1	0.176	0.723	0.862	0.977	21.068
		es->es	5.39	27.7	0.211	0.727	0.83	0.978	19.884
		zh->zh	8.33	31.8	-	0.679	0.901	0.96	21.054

Table 11: The whole results under the multilingual evaluation scenarios.

Task	Model	Language	N-gram-based			Embedding-based	Diversity		Ours
			BLEU	ROUGE-L	METEOR	BERTScore	Distinct-1	Distinct-2	Ensemble
SG	M-BERT	en->de	2.28	13.6	0.127	0.721	0.953	0.995	34.529
		en->fr	1.43	14.1	0.182	0.72	0.938	0.995	30.507
		en->es	1.57	13	0.111	0.72	0.943	0.993	30.368
		en->zh	4.72	25.4	-	0.664	0.997	0.302	26.423
	XLM	en->de	3.2	25.8	0.142	0.73	0.964	0.995	29.683
		en->fr	4.23	26.4	0.198	0.744	0.951	0.989	27.735
		en->es	3.38	24.8	0.135	0.737	0.959	0.991	28.353
		en->zh	5.79	26.8	-	0.67	0.994	0.338	26.674
	mBART	en->de	1.81	11.9	0.117	0.723	0.983	0.999	34.089
		en->fr	1.35	12.7	0.133	0.728	0.969	0.989	29.805
		en->es	1.22	11.2	0.098	0.722	0.928	0.978	30.153
		en->zh	2.59	19.1	-	0.599	0.998	0.663	21.407
	mT5	en->de	3.33	15	0.141	0.731	0.973	0.992	34.569
		en->fr	3.2	15.9	0.203	0.746	0.954	0.981	31.988
		en->es	3.2	14.9	0.131	0.743	0.965	0.988	31.759
		en->zh	5.91	27.3	-	0.675	0.997	0.278	26.873
QG	M-BERT	en->de	5.53	22	0.198	0.738	0.892	0.983	30.585
		en->fr	4.15	19.7	0.255	0.741	0.901	0.991	29.265
		en->es	5.43	21.4	0.292	0.79	0.903	0.987	31.241
		en->zh	8.57	30.3	-	0.689	0.994	0.934	26.358
	XLM	en->de	10.41	37.2	0.254	0.78	0.973	0.996	29.273
		en->fr	14.7	39.6	0.353	0.799	0.949	0.992	29.786
		en->es	16.93	46.8	0.373	0.831	0.952	0.996	29.082
		en->zh	16.07	37.9	-	0.733	0.997	1	28.626
	mBART	en->de	2.25	15.9	0.14	0.701	0.929	0.996	28.605
		en->fr	3.1	18.4	0.178	0.739	0.975	0.997	28.656
		en->es	3.8	17.1	0.246	0.797	0.983	0.997	31.535
		en->zh	13.1	35.4	-	0.722	0.998	0.998	27.967
	mT5	en->de	9.84	28.5	0.246	0.782	0.984	0.998	33.221
		en->fr	13.62	32.8	0.346	0.804	0.967	0.997	32.444
		en->es	15.17	35.2	0.381	0.839	0.962	0.997	33.736
		en->zh	15.69	38.4	-	0.736	0.997	0.999	28.698
TG	M-BERT	en->de	7.21	18	0.181	0.719	0.942	0.995	32.703
		en->fr	6.78	21.9	0.221	0.734	0.928	0.994	28.731
		en->es	9.32	26.2	0.278	0.756	0.92	0.996	28.972
		en->zh	10.94	28.5	-	0.676	0.993	0.452	26.029
	XLM	en->de	9.15	24.3	0.216	0.727	0.947	0.987	30.613
		en->fr	11.54	29	0.301	0.753	0.914	0.987	28.465
		en->es	12.45	31.1	0.31	0.763	0.91	0.993	29.464
		en->zh	15.44	29.9	-	0.692	0.981	0.576	27.139
	mBART	en->de	1.01	6.8	0.069	0.612	0.55	0.73	21.811
		en->fr	2.35	14	0.095	0.63	0.526	0.741	15.719
		en->es	3.84	16.2	0.13	0.626	0.513	0.747	13.509
		en->zh	16.39	34	-	0.705	0.989	0.6	27.556
	mT5	en->de	9.34	21.8	0.221	0.733	0.96	0.994	33.14
		en->fr	12.26	28.2	0.302	0.76	0.929	0.989	29.326
		en->es	14.14	31.1	0.324	0.772	0.917	0.993	30.702
		en->zh	15.31	31.1	-	0.694	0.98	0.609	27.661
Summ	M-BERT	en->de	7.69	25.2	0.199	0.718	0.823	0.986	25.97
		en->fr	8.94	29.2	0.284	0.733	0.769	0.982	22.519
		en->es	10.16	32.5	0.276	0.745	0.757	0.979	20.582
		en->zh	13.08	37.7	-	0.702	0.859	0.92	19.381
	XLM	en->de	8.36	31.1	0.199	0.721	0.807	0.965	21.47
		en->fr	11.79	35.5	0.289	0.75	0.748	0.94	20.301
		en->es	11.93	36.6	0.285	0.75	0.737	0.953	20.678
		en->zh	14.58	38.7	-	0.706	0.841	0.937	19.607
	mBART	en->de	0.61	6.7	0.084	0.533	0.24	0.409	6.789
		en->fr	6.23	23.7	0.204	0.679	0.608	0.857	16.018
		en->es	1.45	16.5	0.093	0.581	0.36	0.615	6.274
		en->zh	15.95	39.7	-	0.709	0.841	0.972	19.802
	mT5	en->de	3.92	22.5	0.135	0.702	0.847	0.955	23.368
		en->fr	3.99	25.7	0.179	0.727	0.837	0.959	21.158
		en->es	5.55	28.9	0.217	0.733	0.799	0.956	19.613
		en->zh	8.43	32.5	-	0.685	0.897	0.965	20.679

Table 12: The whole results under the English centric cross-lingual evaluation scenarios.

Task	Model	Language	N-gram-based			Embedding-based	Diversity		Ours
			BLEU	ROUGE-L	METEOR	BERTScore	Distinct-1	Distinct-2	Ensemble
SG	M-BERT	de->en	2.59	18.5	0.1	0.893	0.918	0.988	30.794
		de->fr	1.29	13.5	0.19	0.717	0.927	0.995	31.616
		de->es	1.58	12.6	0.11	0.716	0.929	0.991	30.216
		de->zh	4.57	24.1	-	0.656	0.989	0.518	26.703
	XLM	de->en	3.18	19.7	0.104	0.894	0.961	0.991	31.967
		de->fr	3.85	26	0.189	0.744	0.949	0.986	27.691
		de->es	3.42	25	0.131	0.737	0.961	0.989	28.057
		de->zh	5.66	26.5	-	0.669	0.991	0.437	26.488
	mBART	de->en	1.99	18.4	0.093	0.887	0.94	0.988	32.055
		de->fr	1.39	12.8	0.163	0.733	0.982	0.997	32.795
		de->es	2.99	14.1	0.122	0.741	0.99	1	32.247
		de->zh	3.29	22.4	-	0.62	0.994	0.317	22.664
	mT5	de->en	3.37	20.1	0.105	0.901	0.968	0.995	32.24
		de->fr	2.89	15.5	0.196	0.744	0.95	0.98	32.463
		de->es	2.93	14.1	0.124	0.74	0.965	0.988	31.902
		de->zh	5.39	26.9	-	0.673	0.996	0.275	26.654
QG	M-BERT	de->en	8.46	33.8	0.16	0.889	0.888	0.985	30.344
		de->fr	3.65	18.3	0.237	0.739	0.901	0.993	29.692
		de->es	4.56	20.6	0.276	0.786	0.914	0.988	30.602
		de->zh	8.17	29.8	-	0.688	0.993	0.927	27.83
	XLM	de->en	13.39	39.3	0.192	0.907	0.952	0.994	31.989
		de->fr	12.11	36.7	0.319	0.79	0.949	0.991	29.54
		de->es	14.3	44	0.34	0.822	0.95	0.996	28.685
		de->zh	14.32	35.6	-	0.722	0.997	1	28.406
	mBART	de->en	8.39	35.1	0.151	0.899	0.978	0.999	30.625
		de->fr	3.36	18.8	0.2	0.75	0.988	0.998	30.037
		de->es	4.03	18	0.252	0.799	0.985	0.997	32.577
		de->zh	8.65	31	-	0.691	0.996	0.989	26.313
	mT5	de->en	12.34	39.5	0.187	0.909	0.967	0.998	31.723
		de->fr	10.51	28.4	0.306	0.787	0.966	0.996	31.792
		de->es	11.61	31.2	0.332	0.827	0.962	0.997	33.198
		de->zh	12.39	34.9	-	0.72	0.998	0.999	27.706
TG	M-BERT	de->en	9.2	30.6	0.158	0.879	0.94	0.99	29.629
		de->fr	5.78	20.1	0.201	0.727	0.923	0.991	28.339
		de->es	6.49	22.5	0.233	0.741	0.917	0.99	27.685
		de->zh	9.85	26.7	-	0.669	0.99	0.498	26.422
	XLM	de->en	11.84	30.6	0.183	0.883	0.963	0.994	30.658
		de->fr	10.2	27	0.276	0.744	0.906	0.982	27.617
		de->es	10.82	28.7	0.278	0.755	0.906	0.991	28.232
		de->zh	13.54	27.8	-	0.68	0.98	0.559	28.158
	mBART	de->en	2.72	19.1	0.083	0.833	0.739	0.881	22.604
		de->fr	5.3	19.5	0.148	0.689	0.773	0.883	22.339
		de->es	1.76	14.9	0.132	0.662	0.662	0.838	19.864
		de->zh	14.6	31.1	-	0.694	0.988	0.591	28.533
	mT5	de->en	12.53	34	0.19	0.892	0.976	0.996	31.707
		de->fr	10.32	25.9	0.276	0.75	0.924	0.985	30.037
		de->es	11.64	28.4	0.291	0.761	0.912	0.991	29.469
		de->zh	13.22	28.5	-	0.682	0.979	0.582	27.805
Summ	M-BERT	de->en	10.84	36.2	0.154	0.869	0.822	0.985	20.938
		de->fr	8.08	28.3	0.258	0.729	0.775	0.98	22.388
		de->es	8.78	31.6	0.26	0.742	0.757	0.977	21.011
		de->zh	11.9	36.9	-	0.7	0.865	0.89	19.931
	XLM	de->en	11.69	35.5	0.17	0.873	0.753	0.943	22.014
		de->fr	10.44	34.4	0.275	0.745	0.738	0.93	21.136
		de->es	10.32	35.1	0.268	0.745	0.734	0.945	19.705
		de->zh	13.25	37.3	-	0.7	0.837	0.921	19.377
	mBART	de->en	12.86	37.2	0.17	0.871	0.833	0.991	21.008
		de->fr	10.53	31.9	0.284	0.743	0.788	0.987	22.358
		de->es	11.52	33.6	0.283	0.745	0.752	0.985	20.568
		de->zh	9.8	33	-	0.679	0.86	0.972	20.075
	mT5	de->en	6.04	31.4	0.118	0.872	0.844	0.961	23.391
		de->fr	3.51	24.7	0.176	0.723	0.819	0.944	20.324
		de->es	5.01	28.1	0.207	0.728	0.779	0.94	20.423
		de->zh	7.63	31.1	-	0.679	0.887	0.964	20.901

Table 13: The whole results under the German centric cross-lingual evaluation scenarios.

Task	Model	Language	N-gram-based			Embedding-based	Diversity		Ours
			BLEU	ROUGE-L	METEOR	BERTScore	Distinct-1	Distinct-2	Ensemble
SG	M-BERT	fr->en	2.62	18.8	0.102	0.893	0.907	0.986	31.131
		fr->de	2.47	14.1	0.134	0.721	0.948	0.995	34.44
		fr->es	1.66	12.6	0.11	0.72	0.944	0.994	30.191
		fr->zh	4.83	24.6	-	0.659	0.996	0.42	26.564
	XLM	fr->en	3.19	19.6	0.102	0.895	0.961	0.993	32.028
		fr->de	3.21	25.7	0.141	0.729	0.958	0.99	29.395
		fr->es	3.17	24.9	0.129	0.737	0.964	0.991	28.116
		fr->zh	5.49	25.9	-	0.666	0.986	0.471	26.284
	mBART	fr->en	1.83	18	0.085	0.892	0.953	0.983	32.32
		fr->de	1.65	12.6	0.117	0.724	0.97	0.997	33.679
		fr->es	0.97	9.3	0.085	0.713	0.898	0.962	29.546
		fr->zh	6.07	26.9	-	0.672	0.996	0.346	26.474
	mT5	fr->en	3.4	20.1	0.104	0.901	0.968	0.994	32.827
		fr->de	3.22	14.8	0.14	0.731	0.97	0.99	34.333
		fr->es	2.9	14.7	0.128	0.741	0.966	0.989	32.558
		fr->zh	5.64	27.2	-	0.674	0.997	0.266	26.986
QG	M-BERT	fr->en	8.47	34.3	0.163	0.888	0.867	0.978	29.537
		fr->de	5.29	21	0.185	0.735	0.903	0.984	31.182
		fr->es	5.41	20.9	0.286	0.789	0.911	0.987	31.429
		fr->zh	7.67	28.9	-	0.683	0.997	0.926	26.186
	XLM	fr->en	13.53	39.5	0.197	0.907	0.955	0.994	31.174
		fr->de	8.5	35.1	0.234	0.771	0.973	0.998	29.541
		fr->es	15.76	45.6	0.359	0.827	0.953	0.996	28.703
		fr->zh	15.01	36.5	-	0.726	0.996	0.998	28.297
	mBART	fr->en	5.4	31.6	0.128	0.892	0.981	0.999	29.219
		fr->de	4	21	0.164	0.744	0.981	0.998	31.385
		fr->es	11.74	31.1	0.328	0.825	0.973	0.999	33.473
		fr->zh	7.32	30	-	0.673	0.993	0.989	24.29
	mT5	fr->en	13.07	40.7	0.196	0.911	0.97	0.998	32.631
		fr->de	8.53	26.2	0.224	0.774	0.983	0.999	33.533
		fr->es	14.74	34.3	0.363	0.834	0.965	0.997	34.057
		fr->zh	13.11	35.5	-	0.723	0.997	0.997	27.876
TG	M-BERT	fr->en	10.46	32.4	0.176	0.885	0.965	0.998	30.775
		fr->de	6.11	16.7	0.166	0.715	0.949	0.997	31.288
		fr->es	8.3	25.6	0.265	0.754	0.921	0.996	29.096
		fr->zh	9.98	27.2	-	0.672	0.989	0.503	26.035
	XLM	fr->en	11.64	31.2	0.188	0.884	0.964	0.995	29.661
		fr->de	7.51	21.8	0.196	0.72	0.956	0.996	30.472
		fr->es	10.99	29.3	0.288	0.757	0.911	0.993	28.422
		fr->zh	14.25	28.1	-	0.684	0.978	0.613	27.015
	mBART	fr->en	1.95	21.4	0.079	0.847	0.878	0.946	23.241
		fr->de	1.49	11.5	0.088	0.663	0.838	0.925	26.812
		fr->es	1.35	15.7	0.131	0.689	0.844	0.932	22.517
		fr->zh	1.7	12.1	-	0.542	0.836	0.432	17.833
	mT5	fr->en	12.13	32.9	0.188	0.891	0.973	0.996	31.065
		fr->de	8.03	19.7	0.196	0.723	0.957	0.994	31.076
		fr->es	12.04	28.6	0.298	0.762	0.911	0.993	30.165
		fr->zh	12.1	26.8	-	0.675	0.969	0.67	26.812
Summ	M-BERT	fr->en	11.19	36.4	0.16	0.869	0.815	0.986	21.222
		fr->de	7.16	24.8	0.195	0.716	0.813	0.986	26.214
		fr->es	9.07	31.7	0.264	0.742	0.762	0.98	20.92
		fr->zh	12.45	37.1	-	0.698	0.855	0.916	19.512
	XLM	fr->en	11.82	36.3	0.176	0.876	0.766	0.956	22.262
		fr->de	7.52	29.9	0.19	0.716	0.796	0.951	21.358
		fr->es	10.89	35.5	0.276	0.747	0.732	0.95	20.133
		fr->zh	13.3	37.7	-	0.702	0.842	0.927	18.906
	mBART	fr->en	12.97	38	0.171	0.873	0.844	0.993	21.336
		fr->de	6.75	25.1	0.186	0.704	0.801	0.98	24.534
		fr->es	10.61	32.5	0.271	0.737	0.749	0.982	20.158
		fr->zh	10.46	34.1	-	0.684	0.874	0.967	19.532
	mT5	fr->en	6	31.4	0.118	0.872	0.847	0.964	22.804
		fr->de	3.43	21.4	0.127	0.696	0.824	0.937	22.566
		fr->es	5.1	28.2	0.209	0.729	0.792	0.951	19.37
		fr->zh	7.74	31.3	-	0.68	0.889	0.958	21.627

Table 14: The whole results under the French centric cross-lingual evaluation scenarios.

Task	Model	Language	N-gram-based			Embedding-based	Diversity		Ours
			BLEU	ROUGE-L	METEOR	BERTScore	Distinct-1	Distinct-2	Ensemble
SG	M-BERT	es->en	2.12	18.4	0.1	0.892	0.909	0.985	31.191
		es->de	2.21	13.7	0.134	0.718	0.93	0.993	34.32
		es->fr	1.42	13.4	0.194	0.72	0.931	0.993	31.622
		es->zh	4.51	24.2	-	0.658	0.995	0.337	26.406
	XLM	es->en	3.12	19.6	0.103	0.895	0.966	0.996	31.257
		es->de	2.83	25	0.139	0.727	0.962	0.994	30.004
		es->fr	3.94	26.3	0.192	0.743	0.948	0.986	27.971
		es->zh	5.35	25.9	-	0.668	0.99	0.483	26.153
	mBART	es->en	0.9	15.2	0.065	0.881	0.925	0.995	33.321
		es->de	0.39	5.9	0.063	0.67	0.963	0.997	30.696
		es->fr	0.81	11.2	0.11	0.699	0.97	0.998	28.099
		es->zh	5.08	26.1	-	0.668	0.999	0.305	26.645
	mT5	es->en	3.4	20.2	0.106	0.901	0.966	0.994	31.215
		es->de	3.17	15	0.141	0.732	0.971	0.992	34.703
		es->fr	3.17	16.3	0.201	0.746	0.955	0.982	32.183
		es->zh	5.61	26.8	-	0.673	0.996	0.286	27.04
QG	M-BERT	es->en	8.89	34.5	0.164	0.888	0.87	0.982	30.078
		es->de	5.18	20.9	0.185	0.735	0.919	0.989	31.312
		es->fr	3.99	18.2	0.254	0.735	0.906	0.993	29.173
		es->zh	7.82	29.1	-	0.683	0.994	0.929	26.005
	XLM	es->en	13.48	39.6	0.198	0.907	0.958	0.997	32.042
		es->de	9.61	36.3	0.239	0.774	0.972	0.996	29.52
		es->fr	13.9	38.6	0.343	0.796	0.948	0.991	28.99
		es->zh	14.78	36.7	-	0.727	0.997	1	28.566
	mBART	es->en	4.96	30.8	0.124	0.89	0.98	0.998	30.121
		es->de	2.35	17.6	0.139	0.726	0.966	0.997	32.497
		es->fr	3.29	18.6	0.193	0.745	0.983	0.998	29.198
		es->zh	12.07	34.5	-	0.719	0.999	0.998	28.107
	mT5	es->en	13.34	40.9	0.198	0.911	0.971	0.998	31.936
		es->de	8.53	26.3	0.227	0.774	0.984	0.998	32.628
		es->fr	12.89	31.1	0.333	0.797	0.967	0.996	32.371
		es->zh	13.37	36	-	0.726	0.997	0.998	28.779
TG	M-BERT	es->en	9.94	31.8	0.167	0.884	0.955	0.995	29.842
		es->de	5.81	16.5	0.165	0.714	0.941	0.993	30.79
		es->fr	6.13	21.2	0.237	0.734	0.932	0.996	28.129
		es->zh	9.73	26.9	-	0.669	0.992	0.418	27.009
	XLM	es->en	12.25	31.7	0.188	0.886	0.966	0.994	30.547
		es->de	8.01	22.5	0.206	0.721	0.947	0.991	29.972
		es->fr	10.99	28.2	0.288	0.748	0.906	0.981	28.193
		es->zh	14.36	28.3	-	0.685	0.978	0.612	27.391
	mBART	es->en	16.65	39.1	0.21	0.901	0.984	0.999	32.993
		es->de	0.95	6.4	0.085	0.605	0.492	0.672	23.184
		es->fr	7.21	22.9	0.187	0.712	0.818	0.901	24.795
		es->zh	15.27	32.3	-	0.697	0.988	0.577	27.849
	mT5	es->en	13	34.5	0.196	0.893	0.977	0.996	31.113
		es->de	8.27	20	0.201	0.726	0.959	0.994	32.597
		es->fr	10.83	26.6	0.285	0.753	0.928	0.988	29.288
		es->zh	13.67	29.2	-	0.685	0.979	0.602	27.73
Summ	M-BERT	es->en	11.41	36.6	0.16	0.869	0.821	0.985	20.613
		es->de	7.22	24.9	0.197	0.716	0.808	0.983	25.962
		es->fr	8.61	28.7	0.28	0.732	0.772	0.983	22.311
		es->zh	12.34	37.3	-	0.7	0.854	0.904	19.625
	XLM	es->en	11.18	35.2	0.173	0.872	0.749	0.946	21.567
		es->de	7.01	29.6	0.184	0.715	0.8	0.96	21.158
		es->fr	10.48	34	0.276	0.742	0.728	0.924	20.617
		es->zh	11.01	35.1	-	0.69	0.834	0.929	20.141
	mBART	es->en	13.28	37.8	0.172	0.872	0.833	0.992	20.997
		es->de	1.21	7.2	0.089	0.663	0.675	0.964	20.631
		es->fr	10.82	31.8	0.29	0.742	0.789	0.987	22.652
		es->zh	10.38	33.5	-	0.683	0.87	0.966	19.889
	mT5	es->en	6.66	32.6	0.122	0.875	0.861	0.975	23.134
		es->de	3.88	22.4	0.134	0.703	0.849	0.96	23.164
		es->fr	4	25.7	0.18	0.728	0.847	0.968	20.854
		es->zh	2.97	19.8	-	0.612	0.758	0.814	18.781

Table 15: The whole results under the Spanish centric cross-lingual evaluation scenarios.

Task	Model	Language	N-gram-based			Embedding-based	Diversity		Ours
			BLEU	ROUGE-L	METEOR	BERTScore	Distinct-1	Distinct-2	Ensemble
SG	M-BERT	zh->en	2.33	18.1	0.096	0.895	0.945	0.993	31.242
		zh->de	2.08	13	0.125	0.716	0.962	0.995	34.057
		zh->fr	1.37	13.1	0.189	0.716	0.94	0.997	30.47
		zh->es	1.5	12.5	0.111	0.716	0.957	0.997	29.987
	XLM	zh->en	0.77	12.3	0.074	0.882	1	1	33.839
		zh->de	1.31	17.6	0.106	0.704	0.917	1	30.245
		zh->fr	0.19	14.2	0.056	0.693	1	1	27.004
		zh->es	0.34	17	0.049	0.701	1	1	28.576
	mBART	zh->en	1.42	17.7	0.068	0.88	0.936	0.978	31.453
		zh->de	0.85	9.6	0.098	0.705	0.943	0.991	33.207
		zh->fr	0.76	11	0.08	0.687	0.867	0.938	26.274
		zh->es	0.64	7.5	0.063	0.678	0.815	0.909	23.607
	mT5	zh->en	3.04	18.9	0.098	0.899	0.967	0.995	31.665
		zh->de	2.65	13.9	0.131	0.728	0.971	0.991	34.153
		zh->fr	2.76	15.1	0.194	0.742	0.951	0.979	32.158
		zh->es	2.72	13.6	0.119	0.738	0.964	0.988	32.053
QG	M-BERT	zh->en	6.24	29.5	0.138	0.882	0.852	0.977	29.007
		zh->de	3.71	17.8	0.155	0.716	0.903	0.984	29.999
		zh->fr	2.69	16.1	0.221	0.718	0.887	0.991	28.73
		zh->es	3.36	18.1	0.248	0.771	0.893	0.985	30.45
	XLM	zh->en	9.96	34.9	0.171	0.9	0.963	0.998	31.198
		zh->de	6.81	31.9	0.206	0.759	0.974	0.998	29.542
		zh->fr	9.62	33.9	0.291	0.78	0.948	0.992	28.668
		zh->es	12.59	41.9	0.319	0.815	0.943	0.994	28.012
	mBART	zh->en	6.04	31.3	0.129	0.895	0.981	1	30.199
		zh->de	4.29	19.4	0.163	0.743	0.989	0.999	32.126
		zh->fr	3.31	19.5	0.191	0.748	0.99	0.999	28.441
		zh->es	3.21	14.5	0.224	0.792	0.991	1	32.106
	mT5	zh->en	11.21	37.9	0.181	0.907	0.969	0.998	31.016
		zh->de	7.65	24.5	0.212	0.765	0.982	0.998	32.596
		zh->fr	9.67	27.4	0.292	0.785	0.966	0.994	31.326
		zh->es	10.45	28.5	0.322	0.822	0.955	0.996	32.803
TG	M-BERT	zh->en	7.27	27.4	0.145	0.871	0.915	0.985	28.663
		zh->de	4.86	14.2	0.145	0.701	0.923	0.991	30.084
		zh->fr	4.88	18.4	0.19	0.717	0.91	0.988	27.593
		zh->es	6.09	22.1	0.224	0.736	0.902	0.99	28.024
	XLM	zh->en	9.6	28.4	0.174	0.879	0.958	0.992	30.079
		zh->de	6.25	19.9	0.181	0.712	0.953	0.994	30.351
		zh->fr	8.2	24.7	0.263	0.736	0.909	0.984	28.278
		zh->es	9.12	26.8	0.261	0.748	0.903	0.991	27.991
	mBART	zh->en	15.03	37.4	0.201	0.899	0.985	0.999	31.993
		zh->de	1.47	9.2	0.084	0.632	0.649	0.787	23.311
		zh->fr	5.41	20.1	0.167	0.701	0.822	0.901	23.772
		zh->es	3.76	18.6	0.156	0.683	0.748	0.879	20.496
	mT5	zh->en	10.19	30.9	0.177	0.887	0.972	0.994	30.058
		zh->de	6.74	18.1	0.181	0.719	0.954	0.99	31.781
		zh->fr	8.31	23.7	0.258	0.742	0.907	0.972	28.542
		zh->es	9.92	26.8	0.271	0.756	0.901	0.986	29.113
Summ	M-BERT	zh->en	8.12	34	0.141	0.866	0.826	0.985	20.782
		zh->de	5.84	23.7	0.176	0.712	0.815	0.984	26.296
		zh->fr	6.86	27.3	0.256	0.727	0.774	0.981	22.454
		zh->es	7.54	30.8	0.247	0.739	0.753	0.976	21.144
	XLM	zh->en	9.13	33.8	0.164	0.869	0.742	0.942	21.267
		zh->de	6	27.8	0.174	0.709	0.777	0.939	22.151
		zh->fr	9.2	32.9	0.264	0.74	0.728	0.917	20.063
		zh->es	9.36	34.2	0.259	0.74	0.708	0.928	19.334
	mBART	zh->en	10.05	35.9	0.156	0.87	0.85	0.993	20.509
		zh->de	6.71	25.5	0.192	0.712	0.813	0.988	24.843
		zh->fr	8.9	31	0.267	0.741	0.798	0.988	21.728
		zh->es	9.5	32.2	0.264	0.745	0.776	0.987	21.736
	mT5	zh->en	4.78	29.4	0.111	0.869	0.824	0.945	23.05
		zh->de	2.9	20.1	0.117	0.69	0.789	0.907	21.725
		zh->fr	3.01	23.6	0.172	0.718	0.796	0.923	20.28
		zh->es	4.54	27.4	0.199	0.725	0.756	0.922	19.743

Table 16: The whole results under the Chinese centric cross-lingual evaluation scenarios.

Task	Model	Language	N-gram-based			Embedding-based	Diversity		Ours
			BLEU	ROUGE-L	METEOR	BERTScore	Distinct-1	Distinct-2	Ensemble
SG	M-BERT	en->de	0.06	3.2	0.053	0.694	0.92	0.991	34.856
		en->fr	0.04	3.9	0.041	0.705	0.921	0.991	32.033
		en->es	0.06	3.5	0.04	0.707	0.918	0.991	31.123
		en->zh	0	0.2	-	0.542	0.919	0.991	20.574
	XLM	en->de	0.02	7.2	0.05	0.634	0.469	0.5	18.939
		en->fr	0.02	5.9	0.037	0.626	0.38	0.412	14.273
		en->es	0.09	8.5	0.035	0.646	0.516	0.547	16.903
		en->zh	0	0	-	0.45	0.609	0.574	16.581
	mBART	en->de	0.1	3.8	0.056	0.705	0.976	0.999	32.962
		en->fr	0.07	4.6	0.048	0.715	0.977	0.999	31.386
		en->es	0.08	4.1	0.046	0.719	0.976	0.999	32.222
		en->zh	0	0.2	-	0.55	0.976	0.999	21.89
	mT5	en->de	0.06	2.3	0.046	0.699	0.983	0.997	33.87
		en->fr	0.03	3.1	0.039	0.709	0.983	0.997	31.869
		en->es	0.05	2.7	0.033	0.712	0.983	0.997	32.738
		en->zh	0	0.1	-	0.542	0.983	0.998	20.781
QG	M-BERT	en->de	0.62	4.3	0.071	0.724	0.908	0.987	32.122
		en->fr	0.58	3.3	0.06	0.731	0.911	0.989	29.54
		en->es	0.49	2.4	0.057	0.747	0.907	0.988	29.76
		en->zh	0.04	0.5	-	0.551	0.906	0.987	22.383
	XLM	en->de	1.96	17.6	0.097	0.726	0.938	0.984	29.795
		en->fr	2.16	16.2	0.081	0.742	0.943	0.99	28.605
		en->es	7.46	23.9	0.177	0.757	0.941	0.985	28.485
		en->zh	0	0	-	0.438	0.099	0.075	16.413
	mBART	en->de	1.37	6.2	0.08	0.738	0.982	1	31.099
		en->fr	1.15	6.7	0.061	0.745	0.982	1	28.971
		en->es	0.93	3.4	0.072	0.761	0.982	1	29.98
		en->zh	0.19	8.5	-	0.557	0.982	1	17.838
	mT5	en->de	1.43	5.4	0.082	0.737	0.97	0.998	31.781
		en->fr	1.17	4.9	0.064	0.744	0.971	0.998	29.666
		en->es	1.07	3.8	0.072	0.76	0.971	0.998	29.345
		en->zh	0.22	0.8	-	0.556	0.971	0.998	22.886
TG	M-BERT	en->de	5.14	11.8	0.136	0.697	0.928	0.988	31.819
		en->fr	3.6	10.8	0.117	0.702	0.936	0.989	28.926
		en->es	3.76	10.4	0.142	0.711	0.928	0.988	28.513
		en->zh	0.56	2.1	-	0.557	0.928	0.988	21.746
	XLM	en->de	2.58	12.5	0.12	0.665	0.83	0.877	26.312
		en->fr	3.26	12.7	0.12	0.685	0.887	0.933	25.239
		en->es	4.9	18.5	0.169	0.705	0.875	0.936	24.466
		en->zh	0.01	0	-	0.446	0.372	0.224	16.684
	mBART	en->de	5.49	14.4	0.142	0.706	0.983	0.999	32.232
		en->fr	3.79	12.6	0.118	0.713	0.988	1	28.936
		en->es	3.89	12.3	0.149	0.721	0.983	0.999	29.65
		en->zh	0.71	2.9	-	0.56	0.983	0.999	22.608
	mT5	en->de	5.41	12.9	0.133	0.696	0.974	0.996	31.885
		en->fr	4.04	11.5	0.112	0.703	0.979	0.997	29.098
		en->es	4.09	11.3	0.138	0.712	0.975	0.996	29.73
		en->zh	0.79	2.4	-	0.553	0.973	0.995	21.886
Summ	M-BERT	en->de	2.24	9.8	0.078	0.675	0.816	0.983	23.995
		en->fr	1.49	9.4	0.075	0.698	0.821	0.986	23.347
		en->es	1.38	8.9	0.074	0.701	0.815	0.983	22.207
		en->zh	0.04	0.5	-	0.521	0.815	0.983	13.572
	XLM	en->de	1.85	14.5	0.08	0.652	0.61	0.777	18.486
		en->fr	1.29	13	0.066	0.678	0.639	0.816	18.639
		en->es	4.18	20.3	0.141	0.686	0.643	0.818	17.303
		en->zh	0	0	-	0.424	0.682	0.274	12.837
	mBART	en->de	2.2	10.3	0.07	0.675	0.852	0.992	23.317
		en->fr	1.46	9.9	0.07	0.695	0.853	0.992	22.576
		en->es	1.35	9.3	0.074	0.699	0.852	0.992	21.736
		en->zh	0.06	0.4	-	0.521	0.853	0.992	13.065
	mT5	en->de	1.11	8	0.05	0.654	0.842	0.936	19.314
		en->fr	0.63	7.4	0.047	0.672	0.845	0.939	17.417
		en->es	0.59	7.1	0.055	0.675	0.842	0.937	17.085
		en->zh	0.01	0.3	-	0.514	0.842	0.937	13.39

Table 17: The whole results under the zero-shot evaluation scenarios.