

# A Dataset and Baselines for Multilingual Reply Suggestion

**Mozhi Zhang\***  
University of Maryland  
mozhi@cs.umd.edu

**Wei Wang†**  
Qualtrics  
wwang@qualtrics.com

**Budhaditya Deb**  
Microsoft AI  
budeb@microsoft.com

**Guoqing Zheng**  
Microsoft Research  
zheng@microsoft.com

**Milad Shokuhi**  
Microsoft AI  
milads@microsoft.com

**Ahmed Hassan Awadallah**  
Microsoft Research  
hassanam@microsoft.com

## Abstract

Reply suggestion models help users process emails and chats faster. Previous work only studies English reply suggestion. Instead, we present MRS, a multilingual reply suggestion dataset with ten languages. MRS can be used to compare two families of models: 1) retrieval models that select the reply from a fixed set and 2) generation models that produce the reply from scratch. Therefore, MRS complements existing cross-lingual generalization benchmarks that focus on classification and sequence labeling tasks. We build a generation model and a retrieval model as baselines for MRS. The two models have different strengths in the monolingual setting, and they require different strategies to generalize across languages. MRS is publicly available at <https://github.com/zhangmozhi/mrs>.

## 1 Multilingual Reply Suggestion

Automated reply suggestion is a useful feature for email and chat applications. Given an input message, the system suggests several replies, and users may click on them to save typing time (Figure 1). This feature is available in many applications including Gmail, Outlook, LinkedIn, Facebook Messenger, Microsoft Teams, and Uber.

Reply suggestion is related to but different from open-domain dialog systems or chatbots (Adiwardana et al., 2020; Huang et al., 2020). While both are conversational AI tasks (Gao et al., 2019), the goals are different: reply suggestion systems help the user quickly reply to a message, while chatbots aim to *continue* the conversation and focus more on multi-turn dialogues.

Ideally, we want our model to generate replies in any language. However, reply suggestion models require large training sets, so previous work mostly

\*Work mostly done as an intern at Microsoft Research.

†Work done at Microsoft Research.

Could you send me a screenshot for reply suggestion?

Sure, here it is.

No, I can't.

What do you mean?

Figure 1: An example of reply suggestion system. User can click on the suggestions for a quick reply.

focuses on English (Kannan et al., 2016; Henderson et al., 2017; Deb et al., 2019). To investigate reply suggestion for other languages with possibly limited data, we build a multilingual dataset, dubbed MRS (Multilingual Reply Suggestion). From publicly available Reddit threads, we extract message-reply pairs, response sets, and machine-translated examples in ten languages (Table 1).

One interesting aspect of the reply suggestion problem is that there are two modeling approaches. Some models follow the retrieval framework and select the reply from a predetermined response set (Henderson et al., 2017). Others follow the generation framework and generate the reply from scratch (Kannan et al., 2016). The two approaches have different advantages. Generation models are more powerful because they are not constrained by the response set. In comparison, retrieval models are easier to train and runs faster, and a curated response set guarantees the coherence and the safety of the model output.

The two frameworks make reply suggestion an interesting task for studying cross-lingual generalization. Most cross-lingual generalization benchmarks use classification and sequence labeling tasks (Tjong Kim Sang, 2002; Nivre et al., 2016; Strassel and Tracey, 2016; Conneau et al., 2018; Schwenk and Li, 2018; Clark et al., 2020; Hu et al., 2020; Lewis et al., 2020b). In contrast, reply suggestion has two formulations that require different cross-lingual generalization strategies. While some recent work explores cross-lingual transfer











Language	Code	Family	Examples	Tokens	Response Set
 English	EN	West Germanic	48,750,948	1,700,066,696	36,997
 Spanish	ES	Romance	2,325,877	195,424,517	45,152
 German	DE	West Germanic	1,864,688	118,711,662	34,747
 Portuguese	PT	Romance	1,822,594	114,642,809	45,225
 French	FR	Romance	1,396,806	133,068,740	32,350
 Japanese	JA	Japonic	727,668	46,124,966	38,817
 Swedish	SV	North Germanic	738,254	47,845,497	32,165
 Italian	IT	Romance	736,296	58,715,043	31,855
 Dutch	NL	West Germanic	638,634	43,847,547	32,293
 Russian	RU	East Slavic	516,739	23,109,295	31,475

Table 1: Dataset statistics for MRS. We collect Reddit message-reply pairs for ten language. For each language, we use 80% examples for training, 10% for validation, and 10% for testing. We then create response sets for retrieval models. We also use MT to translate nineteen million English training examples to other languages.

learning in generation tasks, the tasks are *extrac-tive*; i.e., the output often has significant overlap with the input. These tasks include news title generation, text summarization, and question generation (Chi et al., 2020; Liang et al., 2020; Scialom et al., 2020). Reply suggestion is more challenging because the reply often does not overlap with the message (Figure 1), so the model needs to address different cross-lingual generalization challenges (Section 5.2).

We build two baselines for MRS: a retrieval model and a generation model. We first compare the models in English, where we have abundant training data and human referees. We evaluate the models with both automatic metrics and human judgments. The two models have different strengths. The generation model has higher word overlap scores and is favored by humans on average, but inference is slower, and the output is sometimes contradictory or repetitive (Holtzman et al., 2020). In contrast, the retrieval model is faster and always produces coherent replies, but the replies are sometimes too generic or irrelevant due to the fixed response set.

Next, we test models in other languages. We compare different training settings and investigate two cross-lingual generalization methods: initializing with pre-trained multilingual models (Wu and Dredze, 2019; Conneau et al., 2020; Liang et al., 2020) and training on machine-translated data (Banea et al., 2008). Interestingly, the two models prefer different methods: multilingual pre-training works better for the retrieval model, while the generation model prefers machine translation.

In summary, we present MRS, a multilingual

reply suggestion dataset. We use MRS to provide the first systematic comparison between generation and retrieval models for reply suggestion in both monolingual and multilingual settings. MRS is also a useful benchmark for future research in reply suggestion and cross-lingual generalization.

The rest of the paper is organized as follows. Section 2 describes the data collection process for MRS. Section 3 introduces task formulations, experiment settings, and evaluation metrics. Section 4 describes the baseline generation and retrieval models. Section 5 presents our experiment results. Section 6 discusses how MRS can help future research.

## 2 Dataset Construction

To study reply suggestion in multiple languages, we build MRS, a dataset with message-reply pairs based on Reddit comments. The dataset is available at <https://github.com/zhangmozhi/mrs>.

We download Reddit comments between January 2010 and December 2019 from the Pushshift Reddit dataset (Baumgartner et al., 2020).<sup>1</sup> We extract message-reply pairs from each thread by considering the parent comment as an input message and the response to the comment as the reference reply. We remove comments starting with *[removed]* or *[deleted]*, which are deleted messages. We also skip comments with a rating of less than one, since they are likely to contain inappropriate content.

After extracting examples, we identify their languages with fastText language detector (Joulin et al., 2016). For each example, we run the model

<sup>1</sup><https://files.pushshift.io/reddit/comments>

on the concatenation of the message and the reply. We discard low-confidence examples where none of the languages has a score higher than 0.7. For the remaining examples, we use the highest-scoring label as the language.

We only use English data from 2018 because English data is abundant on Reddit. Non-English examples are much more scarce, so we use data from the last ten years. We select the top ten languages with at least 100K examples. We create three splits for each language: 80% examples for training, 10% for validation, and 10% for testing.

Table 1 shows some dataset statistics. MRS is heavily biased towards English. We have more than 48 million English examples, but fewer than one million examples for half of the languages. This gap reflects a practical challenge for reply suggestion—we do not have enough data for most languages in the world. Nevertheless, we can use MRS to test models in different multilingual settings, including cross-lingual transfer learning, where we build non-English reply suggestion models from English data (Section 3.2).

We also build response sets and filter out toxic examples. We describe these steps next.

## 2.1 Response Set

We build a response set of 30K to 50K most frequent replies for each language, which are used in the retrieval model. We want the response set to cover generic responses, so we select replies that appear at least twenty times in the dataset. This simple criterion works well for English, but the set is too small for other languages. For non-English languages, we augment the response set by translating the English response set to other languages with Microsoft Translator. The non-English response set is sometimes smaller than the English set, because different English responses may have the same translation.

## 2.2 Filtering Toxic Examples

Exchanges on Reddit are sometimes uncivil, inappropriate, or even abusive (Massanari, 2017; Mohan et al., 2017). We try to filter out toxic contents, as they are not desirable for reply suggestion systems.

We use two toxicity detection models. First, we use an in-house multilingual model. The model is initialized with multilingual BERT (Devlin et al., 2019, MBERT) and fine-tuned on a mixture of proprietary and public datasets with toxic and offen-

sive language labels. The model outputs a score from zero to one, with a higher score corresponding to a higher level of toxicity. Second, we use Perspective API<sup>2</sup>, a publicly available model. Perspective API has limited free access (one query per second), so we only use the API on the English validation, test, and response set. For other languages, we rely on our in-house model. We filter message-reply pairs if it has greater than 0.9 score according to the in-house model, or greater than 0.5 score according to Perspective API (Gehman et al., 2020). About one percent of examples are filtered. After filtering the data, we manually validate three hundred random examples and do not find any toxic examples, which confirms that our filter method have a high recall.

While we hope the filtered dataset leads to better reply suggestion models, existing filtering methods are not perfect and can introduce other biases (Dixon et al., 2018; Sap et al., 2019; Hutchinson et al., 2020). Therefore, models trained on all MRS data may still have undesirable behavior. MRS is intended to be used as a benchmark for testing cross-lingual generalization of generation and retrieval models. **The dataset should not be directly used in production systems.** To use the dataset in practice, additional work is required to address other possible biases and toxic or inappropriate content that may exist in the data.

# 3 Experiment Settings

After presenting the dataset, we explain how we use MRS to compare reply suggestion models. We describe the two frameworks for reply suggestion, our experiment settings, and evaluation metrics.

## 3.1 Task Formulation

In reply suggestion, the input is a message  $x$ , and the output is one or more suggested replies  $y$ . In practice, reply suggestion systems can choose to not suggest any replies. This decision is usually made by a separate trigger model (Kannan et al., 2016). In this paper, we focus on reply generation, so we assume that the models always need to suggest a fixed number of replies. Reply suggestion can be formulated as either a *retrieval* problem or a *generation* problem.

**Retrieval Model.** A retrieval model selects the reply  $y$  from a fixed response set  $\mathcal{Y}$  (Section 2.1).

<sup>2</sup><https://www.perspectiveapi.com>

Given an input message  $\mathbf{x}$ , the model computes a relevance score  $\Theta_{\mathbf{x}\mathbf{y}}$  for each candidate reply  $\mathbf{y} \in \mathcal{Y}$ . The model then selects the highest-scoring replies as suggestions; e.g., the top-1 reply is  $\arg \max_{\mathbf{y} \in \mathcal{Y}} \Theta_{\mathbf{x}\mathbf{y}}$ .

**Generation Model.** A generation model generates the reply  $\mathbf{y}$  from scratch. Generation models usually follow the sequence-to-sequence framework (Sutskever et al., 2014, SEQ2SEQ), which generates  $\mathbf{y}$  token by token. Given an input message  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  of  $n$  tokens, a SEQ2SEQ model estimates the probability of a reply  $\mathbf{y} = (y_1, y_2, \dots, y_m)$  of  $m$  tokens as following:

$$p(\mathbf{y} | \mathbf{x}) = \prod_{i=1}^m p(y_i | \mathbf{x}, y_{<i}). \quad (1)$$

The model computes probability for the next token  $p(y_i | \mathbf{x}, y_{<i})$  based on the input  $\mathbf{x}$  and the first  $(i - 1)$  tokens of the output  $\mathbf{y}$ . The model is trained to maximize the probability of reference replies in the training set. At test time, we find the top replies that approximately maximize (1) with beam search.

The two models have different strengths. The generation model is more flexible, but the retrieval model is faster (Henderson et al., 2017), and the output can be controlled by curating the response set (Kannan et al., 2016).

We compare a retrieval model and a generation model as baselines for MRS. To our knowledge, we are the first to systematically compare the two models in both monolingual and multilingual settings. We explain our training settings and metrics next.

### 3.2 Training Settings

For each language in MRS, we train and compare models in four settings. Future work can experiment with other settings (discussed in Section 6).

**Monolingual.** Here, we simply train and test models in a single language. This setting simulates the scenario where we have adequate training data for the target language. Previous reply suggestion models were only studied in the English monolingual setting.

**Zero-Shot.** Next, we train models in a zero-shot cross-lingual setting. We train the model on the English training set and use the model on the test set for another language. This setting simulates the scenario where we want to build models for a low-resource language using our large English set.

To generalize across languages, we initialize the models with pre-trained multilingual models (details in Section 4). These models work well in other tasks (Wu and Dredze, 2019; Liang et al., 2020). We test if they also work for reply suggestion, as different tasks often prefer different multilingual representations (Zhang et al., 2020b).

**Machine Translation (MT).** Another strategy for cross-lingual generalization is to train on machine-translated data (Banea et al., 2008). We train models on nineteen million English training examples machine-translated to the target language with Microsoft Translator. We compare against the zero-shot setting to compare the two cross-lingual generalization strategies.

**Multilingual.** Finally, we build a multilingual model by jointly training on the five languages with the most training data: English, Spanish, German, Portuguese, and French. We oversample non-English training data to have the same number of training examples data across all languages (Johnson et al., 2017). We make two comparisons: 1) for the five training languages, we compare against the *monolingual* setting to test whether fitting multiple languages in a single model hurts performance; and 2) for other languages, we compare against the *zero-shot* setting to check if adding more training languages helps cross-lingual generalization.

### 3.3 Evaluation Metrics

The goal of reply suggestion is to save user typing time, so the ideal metrics are click-through rate (CTR), how often the user chooses a suggested reply, and time reduction, how much time is saved by clicking the suggestion instead of typing. However, these metrics require deploying the model to test on real users, which is not feasible at full-scale while writing this paper. Instead, we focus on automated offline metrics that can guide research and model development before deploying production systems. Specifically, we evaluate models using a test set of message-reply pairs.

To identify a good metric, we compare several metrics in a pilot study by deploying an English system. We collect millions of user interactions and measure Pearson’s correlation between CTR and automated offline metrics. The next paragraph lists the metrics. Based on the study, we recommend weighted ROUGE F1 ensemble (**ROUGE** in tables), which has the highest correlation with CTR.

For the retrieval model, we follow previous work and consider mean reciprocal rank (Kannan et al., 2016, MRR) and precision at one (Henderson et al., 2017). These metrics test if the model can retrieve the reference response from a random set of responses. Alternatively, we compute MRR and precision on a subset of examples where the reference reply is in the response set so that we can directly measure the rank of the reference response in the response set. This set also allows us to compute MRR for individual responses, so we can compute macro-MRR, the average MRR over each response in the set. Higher macro-MRR can indicate diversity but has a worse correlation than computing MRR over the entire test set. For the generation model, we consider model perplexity (Adiwardana et al., 2020). Finally, we consider two word overlap scores, BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), which can be used for both retrieval and generation models.

Our pilot study shows that ROUGE has the best correlation. However, individual ROUGE F1 scores (ROUGE-1/2/3) are sensitive to small changes in sequence lengths (more so because our responses are generally short). Therefore, we use a weighted average of the three scores:

$$\frac{\text{ROUGE-1}}{6} + \frac{\text{ROUGE-2}}{3} + \frac{\text{ROUGE-3}}{2}. \quad (2)$$

This weighted score leads to the highest correlation with CTR. Intuitively, the weights balance the differences in the average magnitude of each metric and thus reduce variance on short responses.

Popular reply suggestion systems (such as Gmail and Outlook) suggest three replies for each message, while the user only selects one. To simulate this setting, we predict three replies for each message. For the retrieval model, we use the three highest-scoring replies from the response set. For the generation model, we use top-three results from beam search. Out of the three replies, we only use the reply with the highest ROUGE compared to the reference reply when computing the final metrics; i.e., the model only has to provide one “correct” reply to have a full score.

We compare models primarily with ROUGE, since the metric has the best correlation in the pilot study. Nevertheless, word overlap scores have known limitations (Liu et al., 2016), as there are different ways to reply to a message. We encourage future research to investigate other metrics to understand different aspects of the model.

As examples, we also report two diversity scores: the proportion of distinct unigrams (**Dist-1**) and bigrams (**Dist-2**) in the generated replies (Li et al., 2016). While ROUGE measures the relevance of the replies, higher diversity can also increase CTR (Deb et al., 2019). We can improve the diversity of the three replies with diversity-promoting decoding (Li et al., 2016; Vijayakumar et al., 2018; Zhang et al., 2018) or latent variable models (Deb et al., 2019), but we leave this direction to future work.

For our English monolingual experiments, we also complement automatic metrics with human judgments (**Human** in Figure 2). For each example, we display the input message and sets of three suggested replies from both generation and retrieval models to three human annotators (crowd workers). We then ask the annotators to select the set with more responses that they prefer to send as a reply. We leave evaluations for other languages to future work due to resource limitations.

## 4 Baseline Models

This section introduces the two baseline models: a retrieval model and a generation model.

### 4.1 Retrieval Model

For the retrieval model, we use the architecture from Henderson et al. (2017), except we replace the feedforward network encoders with Transformers (Vaswani et al., 2017). Given an input message  $\mathbf{x}$  and candidate reply  $\mathbf{y}$ , two Transformer encoders  $\Phi_x$  and  $\Phi_y$  map the message and the reply to two vectors  $\Phi_x(\mathbf{x})$  and  $\Phi_y(\mathbf{y})$ . The relevance score  $\Theta_{xy}$  between the message  $\mathbf{x}$  and the reply  $\mathbf{y}$  is the dot product of the two vectors:

$$\Theta_{xy} = \Phi_x(\mathbf{x})^\top \Phi_y(\mathbf{y}). \quad (3)$$

Henderson et al. (2017) also adds a language model score to encourage more frequent replies. We do not use language model score for simplicity.

We train the model with the symmetric loss from Deb et al. (2019). Suppose the batch size is  $n$ . For a batch of training messages  $\{\mathbf{x}_i\}_{i=1}^n$  and corresponding replies  $\{\mathbf{y}_i\}_{i=1}^n$ , we maximize:

$$\sum_{i=1}^n \frac{e^{\Theta_{x_i y_i}}}{\sum_{j=1}^n (e^{\Theta_{x_i y_j}} + e^{\Theta_{x_j y_i}}) - e^{\Theta_{x_i y_i}}}. \quad (4)$$

In a regular softmax loss, the denominator only sums over one variable. The denominator in the

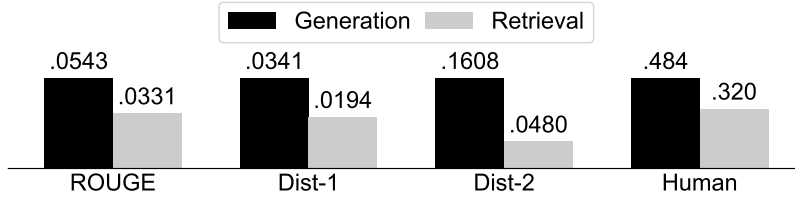


Figure 2: Generation vs. retrieval model on English. Both automatic metrics and human referees prefer the generation model. The **human** score measures how often the human referee prefers the replies from one model over another. The human score does not add up to one, because the referee can choose tie.

<b>Message:</b>	I think I want to play it so it better be available for PS4.
<b>Generation:</b>	I’m sure it’ll be available for PS4 as well.
<b>Retrieval:</b>	I’d play it.
<b>Message:</b>	Very nice! What model is this? Obviously new Softail/Dyna, but which?
<b>Generation:</b>	It’s a Softail, but I’m not sure what model it is.
<b>Retrieval:</b>	Delta.
<b>Message:</b>	Interesting proposal. Especially since they just announced Mixon is going to have a scope.
<b>Generation:</b>	I’m pretty sure it’s going to have a scope, but I don’t think it’s going to have a scope.
<b>Retrieval:</b>	Where did they say that?
<b>Message:</b>	Oh the stickers! I envy you, Lush didn’t sent me any with my body lotion order.
<b>Generation:</b>	I’m so sorry. I’m so sorry. I’m so sorry.
<b>Retrieval:</b>	What did you order?

Table 2: Example replies suggested by English models. In the top two examples, the generation model produces more relevant replies than the retrieval model. In the bottom two examples, the generation model produces contradictory and repetitive replies, while the retrieval model selects appropriate replies from the response set.

symmetric loss sum over both variables to encourage bidirectional compatibility: the message should be predictive of the reply, and the reply should be predictive of the message. This encourages the model to select responses specific to the message, similar to the Maximum Mutual Information objective from Li et al. (2016).

The two encoders  $\Phi_x$  and  $\Phi_y$  are initialized with MBERT (Devlin et al., 2019), a Transformer with 110 million parameters pre-trained on multilingual corpora. Initializing with MBERT allows the model to generalize across languages (Wu and Dredze, 2019). In Appendix A, we experiment with another pre-trained multilingual Transformer, XLM-R (Conneau et al., 2020). We use the “base” version with 270 million parameters.

## 4.2 Generation Model

For the generation model, we follow the SEQ2SEQ architecture (Section 3.1). We use a Transformer encoder to read the input  $\mathbf{x}$ , and another Transformer decoder to estimate  $p(y_i | \mathbf{x}, y_{<i})$  in (1).

We cannot initialize the generation model with MBERT or XLM-R, because the model also has a decoder. Instead, we use Unicoder-XDAE (Liang et al., 2020), a pre-trained multilingual SEQ2SEQ model, which can generalize across languages in extractive generation tasks such as news title generation and question generation. We test if Unicoder-XDAE also generalizes in the more challenging reply suggestion task. There are other generation models we can use, which we discuss as future work in Section 6.

## 4.3 Training Details

We train the retrieval model using Adam optimizer (Kingma and Ba, 2015) with  $1e-6$  learning rate, default  $\beta$ , and 256 batch size. For monolingual and zero-shot settings, we use twenty epochs for English and fifty epochs for other languages. We use ten epochs for MT and multilingual settings. The first 1% training steps are warmup steps. During training, we freeze the embedding layers and the bottom two Transformer layers of both en-

	Monolingual			Zero-Shot			MT			Multilingual		
	ROUGE	Dist-1	Dist-2	ROUGE	Dist-1	Dist-2	ROUGE	Dist-1	Dist-2	ROUGE	Dist-1	Dist-2
EN	.0331	.0194	.0480	.0331	.0194	.0480	-	-	-	.0265	.0158	.0376
ES	.0187	.0157	.0353	<b>.0156</b>	.0113	.0271	.0139	.0164	.0350	.0181	.0151	.0333
DE	.0215	.0134	.0298	<b>.0178</b>	.0098	.0240	.0141	.0152	.0333	.0190	.0140	.0314
PT	.0509	.0158	.0393	<b>.0115</b>	.0121	.0323	.0110	.0184	.0449	.0460	.0161	.0401
FR	.0216	.0191	.0468	<b>.0168</b>	.0133	.0343	.0166	.0196	.0461	.0212	.0169	.0411
JA	.0311	.0220	.0540	<b>.0213</b>	.0236	.0250	.0153	.1031	.0444	.0144	.0677	.0286
IT	.0200	.0357	.0768	<b>.0172</b>	.0246	.0576	.0150	.0378	.0811	.0171	.0278	.0614
SV	.0188	.0287	.0658	.0168	.0203	.0506	.0176	.0302	.0677	<b>.0169</b>	.0224	.0518
NL	.0184	.0316	.0766	.0167	.0199	.0533	.0169	.0297	.0710	<b>.0170</b>	.0221	.0551
RU	.0142	.0486	.0946	.0138	.0298	.0604	.0130	.0431	.0804	<b>.0246</b>	.0405	.0761

Table 3: Results for retrieval model initialized with MBERT (Devlin et al., 2019). The settings are in Section 3.2. Gray cells indicate when the model is trained on the target language training set. White cells indicate cross-lingual settings where the target language training set is not used for training. For each language, we **boldface** the best ROUGE scores in cross-lingual settings (white cells). The zero-shot setting has better ROUGE scores than using MT data for most languages, and the results are sometimes close to monolingual training, confirming the effectiveness of MBERT. Multilingual training hurts training languages (gray cells compared to monolingual) but sometimes improves cross-lingual generalization (white cells compared to zero-shot).

coders, which preserves multilingual knowledge from the pre-trained model and improves cross-lingual transfer learning (Wu and Dredze, 2019). All hyperparameters are manually tuned on the English validation set.

We use almost the same hyperparameters as Liang et al. (2020) to train generation models. Specifically, we use Adam optimizer with  $1e-5$  initial learning rate, default  $\beta$ , and 1024 batch size. For the monolingual and zero-shot setting, we use four epochs for English and 5000 steps for other languages (equivalent to two to nine epochs depending on the language). We use one epoch for the MT setting and 40,000 steps for the multilingual setting. The first 20% training steps are warmup steps. We freeze the embedding layer during training for faster training.

All models are trained with eight Tesla V100 GPU. It takes about an hour to train the generation model for 1000 steps (covering about one million examples). For the retrieval model, an epoch on the English training set (about 48 million examples) takes about seven hours.

## 5 Results and Discussion

We experiment with the two baselines from Section 4 on MRS. We first compare the models in English, where we have enough training data and human referees. We then build models for other

languages and compare training settings listed in Section 3.2.

### 5.1 Results on English

Figure 2 compares the generation and retrieval models in the English monolingual setting. Generation model not only has higher relevance (ROUGE) score but also can generate more diverse replies (higher DIST scores). For English, we also ask three human referees to compare the model outputs on a subset of 500 test examples. Again, the referees prefer the generation model more often than the retrieval model (Figure 2).

We look at some generated responses to understand the models qualitatively. In the top two examples in Table 2, the generation model produces replies highly specific to the input message. In contrast, the retrieval model fails to find a relevant reply, because the response set does not cover these topics. This explains why the generation model has much higher ROUGE and distinct  $n$ -gram scores than the retrieval model.

However, the expressiveness comes at the cost of a lack of control over the generated replies. The generation model sometimes produces incoherent replies that are repetitive and/or contradictory, as shown in the bottom two examples of Table 2. For the retrieval model, we can easily avoid these problems by curating the fixed response set. These degenerative behaviors are observed in other text

	Monolingual			MT			Multilingual		
	ROUGE	DIST1	DIST2	ROUGE	DIST1	DIST2	ROUGE	DIST1	DIST2
EN	.0543	.0341	.161	-	-	-	.0412	.0352	.175
ES	.0397	.0214	.182	<b>.0270</b>	.0261	.190	.0366	.0209	.175
DE	.0469	.0332	.228	<b>.0288</b>	.0244	.142	.0454	.0321	.220
PT	.0566	.0209	.194	<b>.0276</b>	.0221	.161	.0564	.0207	.190
FR	.0446	.0207	.174	<b>.0271</b>	.0165	.109	.0428	.0211	.175
JA	.0139	.1931	.245	.0042	.2812	.216	<b>.0114</b>	.0954	.179
IT	.0493	.0322	.243	<b>.0316</b>	.0393	.240	.0295	.0312	.222
SV	.0387	.0376	.236	<b>.0369</b>	.0359	.203	.0241	.0380	.227
NL	.0377	.0337	.230	<b>.0320</b>	.0284	.162	.0233	.0334	.219
RU	.0286	.0825	.349	<b>.0238</b>	.0310	.094	.0165	.0607	.224

Table 4: Results for generation model. The settings are in Section 3.2. Gray cells indicate when the model is trained on the target language training set. White cells indicate cross-lingual settings where the target language training set is not used for training. For each language, we **boldface** the best ROUGE scores in cross-lingual settings (white cells). Despite initializing with Unicoder-XDAE (Liang et al., 2020), the model fails to generalize across languages in zero-shot settings. The table does not include zero-shot results because the model only produces English replies and thus has near-zero ROUGE. Multilingual training hurts training languages (gray cells compared to monolingual), but the model can now generalize to unseen languages. Training on MT data is the best cross-lingual generalization method for the generation model.

generation tasks and can be mitigated by changing training and decoding objectives (Holtzman et al., 2020; Welleck et al., 2020). We leave these directions for future research.

## 5.2 Results on Other Languages

After comparing English models, we experiment on other languages using the settings from Section 3.2.

**Retrieval Model.** Table 3 shows results for the retrieval model when initialized with MBERT. The retrieval model can generalize fairly well across languages, as the ROUGE in the zero-shot setting is often close to the monolingual setting. This result confirms that initializing with MBERT is an effective strategy for cross-lingual generalization. Training on MT data is usually worse than training in the zero-shot setting. This is possible because the MT system may create artifacts that do not appear in organic data (Artetxe et al., 2020). For the multilingual model, the training language ROUGE scores are lower than monolingual training (gray cells in Table 3). However, multilingual training sometimes leads to better ROUGE on unseen languages compared to transferring from only English (zero-shot). Previous work observes similar results on other tasks, where multilingual training hurts training languages but helps generalization to unseen languages (Johnson et al., 2017; Con-

neau et al., 2020; Wang et al., 2020). Finally, Appendix A shows similar results when initializing with XLM-R (Conneau et al., 2020).

**Generation Model.** Table 4 shows results for the generation model. In the monolingual setting, the generation model has higher scores than the retrieval model on most languages, consistent with the English result (Figure 2). However, unlike the retrieval model, the generation model fails to generalize across languages in the zero-shot setting, despite using Unicoder-XDAE for initialization. We do not show zero-shot results in Table 4, because ROUGE are close to zero for non-English languages. After training on English data, the model always produces English replies, regardless of the input language; i.e., the generation model “forgets” multilingual knowledge acquired during pre-training (Kirkpatrick et al., 2017). This result is surprising because Unicoder-XDAE works in the zero-shot setting for other generation tasks (Liang et al., 2020), which suggests that reply suggestion poses unique challenges for cross-lingual transfer learning. Interestingly, the multilingual model can generalize to unseen languages; perhaps training on multiple languages regularizes the model to produce replies in the input language. Overall, the best method to generalize the generation model across languages is to use machine-translated data.



## 6 Future Work

MRS opens up opportunities for future research. Our experiments use four training settings (Section 3.2), but there are many other settings to explore. For example, we can use other combinations of training languages, which may work better for some target languages (Ammar et al., 2016; Cotterell and Heigold, 2017; Ahmad et al., 2019; Lin et al., 2019; Zhang et al., 2020a). We are also interested in training on both organic data and MT data; i.e., mixing the zero-shot and MT setting.

We can also compare other models on MRS. For the English monolingual setting, we can initialize the generation model with state-of-the-art language models (Radford et al., 2019; Brown et al., 2020; Zhang et al., 2020c). For cross-lingual settings, we can initialize the generation model with several recent pre-trained multilingual SEQ2SEQ models (Chi et al., 2020, 2021; Liu et al., 2020; Tran et al., 2020; Lewis et al., 2020a; Xue et al., 2020). For retrieval models, we can experiment with other multilingual encoders that use different pre-training tasks (Artetxe and Schwenk, 2019; Chidambaram et al., 2019; Reimers and Gurevych, 2020; Feng et al., 2020).

Another idea is to combine the two models. Given an input message, we first use a generation model to create a set of candidate replies. We then use a retrieval model to compute relevance scores and rerank these candidates. Reranking the output of a generation model helps other natural language processing tasks (Shen et al., 2004; Collins and Koo, 2005; Ge and Mooney, 2006), and previous work uses a similar idea for chatbots (Qiu et al., 2017).

Our experiment shows that reply suggestion poses unique challenges for cross-lingual generalization, especially for the generation model. Future work can study methods to improve cross-lingual generalization methods. Some examples include applying adversarial learning (Chen et al., 2018, 2019; Huang et al., 2019), using adapters (Pfeiffer et al., 2020), adaptive transfer (Xia et al., 2021), mixing pre-training and fine-tuning (Phang et al., 2020), and bringing a human in the loop (Yuan et al., 2020).

## 7 Conclusion

We present MRS, a multilingual dataset for reply suggestion. We compare a generation and a retrieval baseline on MRS. The two models have dif-

ferent strengths in the English monolingual setting and require different strategies to transfer across languages. MRS provides a benchmark for future research in both reply suggestion and cross-lingual transfer learning.

## Ethical Considerations

**Data Collection.** No human annotators are involved while creating MRS. The examples and response sets of MRS come from publicly available Reddit dumps from Pushshift, which are used in more than a hundred peer-reviewed publications (Baumgartner et al., 2020).

**Privacy.** Examples in MRS do not have the username and are from publicly available data. Therefore, we do not anticipate any privacy issues. In the pilot study (Section 3.3), we measure the correlation of user CTR with different evaluation metrics. To protect user privacy, we only collect aggregated statistics (CTR) and use no other information.

**Potential Biased and Toxic Content.** Despite our best effort to filter toxic contents (Section 2.2), the dataset may not be perfectly cleansed and may have other biases that are typical in open forums (Massanari, 2017; Mohan et al., 2017). Users should be aware of these issues. We will continue to improve the quality of the dataset.

**Intended Use of MRS.** Because of the possible biases and inappropriateness in the data, MRS should *not* be directly used to build production systems (as mentioned in Section 2.2). The main use of MRS is to test cross-lingual generalization for text retrieval and generation models, and researchers should be aware of possible ethical issues of Reddit data before using MRS.

## Acknowledgement

We appreciate the feedback from anonymous reviewers. MZ is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the BETTER Program contract #2019-19051600005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

## References

- Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.
- Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Conference of the North American Chapter of the Association for Computational Linguistics*.
- Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of Empirical Methods in Natural Language Processing*.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. In *International Conference on Weblogs and Social Media*.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the Association for Computational Linguistics*.
- Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.
- Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *Association for the Advancement of Artificial Intelligence*.
- Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXML: An information-theoretic framework for cross-lingual language model pre-training. In *Conference of the North American Chapter of the Association for Computational Linguistics*.
- Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of ACL Workshop on Representation Learning for NLP*.
- Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.
- Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–70.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Association for Computational Linguistics*.
- Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of Empirical Methods in Natural Language Processing*.
- Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of Empirical Methods in Natural Language Processing*.
- Budhaditya Deb, Peter Bailey, and Milad Shokouhi. 2019. Diversifying reply suggestions using a matching-conditional variational autoencoder. In *Conference of the North American Chapter of the Association for Computational Linguistics (Industry Papers)*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society*.

- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*.
- Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval*, 13(2-3):127–298.
- Ruifang Ge and Raymond J. Mooney. 2006. [Discriminative reranking for semantic parsing](#). In *Proceedings of the Association for Computational Linguistics*.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. [RealToxicityPrompts: Evaluating neural toxic degeneration in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP*. Association for Computational Linguistics.
- Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for Smart Reply. *arXiv preprint arXiv:1705.00652*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the International Conference on Learning Representations*.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the International Conference of Machine Learning*.
- Lifu Huang, Heng Ji, and Jonathan May. 2019. [Cross-lingual multi-level adversarial transfer to enhance low-resource name tagging](#). In *Conference of the North American Chapter of the Association for Computational Linguistics*.
- Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems*, 38(3):1–32.
- Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denny. 2020. [Social biases in NLP models as barriers for persons with disabilities](#). In *Proceedings of the Association for Computational Linguistics*.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Anjali Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, et al. 2016. Smart Reply: Automated response suggestion for email. In *Knowledge Discovery and Data Mining*.
- Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.
- Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. In *Proceedings of Advances in Neural Information Processing Systems*.
- Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020b. [MLQA: Evaluating cross-lingual extractive question answering](#). In *Proceedings of the Association for Computational Linguistics*.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. [A diversity-promoting objective function for neural conversation models](#). In *Conference of the North American Chapter of the Association for Computational Linguistics*.
- Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fengei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. [XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation](#). In *Proceedings of Empirical Methods in Natural Language Processing*.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*.
- Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. [Choosing transfer languages for cross-lingual learning](#). In *Proceedings of the Association for Computational Linguistics*.

- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. [How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation](#). In *Proceedings of Empirical Methods in Natural Language Processing*.
- 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.
- Adrienne Massanari. 2017. #Gamergate and The Fapping: How Reddit’s algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3):329–346.
- Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The impact of toxic language on the health of Reddit communities. In *Canadian Conference on Artificial Intelligence*.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. [Universal Dependencies v1: A multilingual treebank collection](#). In *Proceedings of the Language Resources and Evaluation Conference*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: a method for automatic evaluation of machine translation](#). In *Proceedings of the Association for Computational Linguistics*.
- Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. [MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer](#). In *Proceedings of Empirical Methods in Natural Language Processing*.
- Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. [English intermediate-task training improves zero-shot cross-lingual transfer too](#). In *Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*.
- Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. 2017. [AliMe chat: A sequence to sequence and rerank based chatbot engine](#). In *Proceedings of the Association for Computational Linguistics*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.
- Nils Reimers and Iryna Gurevych. 2020. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of Empirical Methods in Natural Language Processing*.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. [The risk of racial bias in hate speech detection](#). In *Proceedings of the Association for Computational Linguistics*.
- Holger Schwenk and Xian Li. 2018. [A corpus for multilingual document classification in eight languages](#). In *Proceedings of the Language Resources and Evaluation Conference*.
- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. [MLSUM: The multilingual summarization corpus](#). In *Proceedings of Empirical Methods in Natural Language Processing*.
- Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. [Discriminative reranking for machine translation](#). In *Conference of the North American Chapter of the Association for Computational Linguistics*.
- Stephanie Strassel and Jennifer Tracey. 2016. [LORELEI language packs: Data, tools, and resources for technology development in low resource languages](#). In *Proceedings of the Language Resources and Evaluation Conference*.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of Advances in Neural Information Processing Systems*.
- Erik F. Tjong Kim Sang. 2002. [Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition](#). In *Conference on Computational Natural Language Learning*.
- Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *Proceedings of Advances in Neural Information Processing Systems*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*.
- Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search: Decoding diverse solutions from neural sequence models. In *Association for the Advancement of Artificial Intelligence*.
- Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. [On negative interference in multilingual models: Findings and a meta-learning treatment](#). In *Proceedings of Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

- Sean Welleck, Ilya Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *Proceedings of the International Conference on Learning Representations*.
- Shijie Wu and Mark Dredze. 2019. [Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT](#). In *Proceedings of Empirical Methods in Natural Language Processing*.
- Mengzhou Xia, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Newbig, and Ahmed Hassan Awadallah. 2021. [MetaXL: Meta representation transformation for low-resource cross-lingual learning](#). In *Conference of the North American Chapter of the Association for Computational Linguistics*.
- 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*.
- Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2020. [Interactive refinement of cross-lingual word embeddings](#). In *Proceedings of Empirical Methods in Natural Language Processing*.
- Mozhi Zhang, Yoshinari Fujinuma, and Jordan Boyd-Graber. 2020a. Exploiting cross-lingual subword similarities in low-resource document classification. In *Association for the Advancement of Artificial Intelligence*.
- Mozhi Zhang, Yoshinari Fujinuma, Michael J. Paul, and Jordan Boyd-Graber. 2020b. [Why overfitting isn't always bad: Retrofitting cross-lingual word embeddings to dictionaries](#). In *Proceedings of the Association for Computational Linguistics*.
- Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Proceedings of Advances in Neural Information Processing Systems*.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020c. [DialoGPT: Large-scale generative pre-training for conversational response generation](#). In *Proceedings of the Association for Computational Linguistics: System Demonstrations*.

## A Results for XLM-R

	Monolingual			Zero-Shot			MT			Multilingual		
	ROUGE	Dist-1	Dist-2	ROUGE	Dist-1	Dist-2	ROUGE	Dist-1	Dist-2	ROUGE	Dist-1	Dist-2
🇬🇧 EN	.0354	.0177	.0454	.0354	.0177	.0454	-	-	-	.0319	.0152	.0398
🇪🇸 ES	.0158	.0069	.0172	<b>.0140</b>	.0065	.0160	.0122	.0079	.0181	.0155	.0076	.0182
🇩🇪 DE	.0179	.0098	.0261	<b>.0141</b>	.0064	.0162	.0132	.0071	.0170	.0171	.0069	.0170
🇵🇹 PT	.0345	.0088	.0239	<b>.0126</b>	.0076	.0209	.0120	.0071	.0178	.0332	.0086	.0230
🇫🇷 FR	.0161	.0062	.0168	<b>.0143</b>	.0066	.0177	.0135	.0073	.0184	.0161	.0069	.0185
🇯🇵 JA	.0271	.0132	.0364	<b>.0181</b>	.0097	.0277	.0157	.0106	.0293	.0166	.0123	.0328
🇮🇹 IT	.0157	.0123	.0291	<b>.0144</b>	.0123	.0306	.0155	.0156	.0375	.0143	.0136	.0337
🇸🇩 SV	.0172	.0129	.0333	.0165	.0133	.0333	.0153	.0140	.0341	<b>.0168</b>	.0125	.0321
🇳🇱 NL	.0171	.0142	.0390	.0161	.0134	.0371	.0155	.0134	.0353	<b>.0162</b>	.0135	.0370
🇷🇺 RU	.0128	.0259	.0541	.0123	.0223	.0467	.0111	.0248	.0506	<b>.0130</b>	.0244	.0510

Table 5: Results for retrieval model initialized with XLM-R (Conneau et al., 2020). The settings are in Section 3.2. Gray cells indicate when the model is trained on the target language training set. White cells indicate cross-lingual settings where the target language training set is not used for training. For each language, we **boldface** the best ROUGE scores in cross-lingual settings (white cells). We observe similar trends as MBERT (Table 3).