Pre-training via Paraphrasing

Mike Lewis **Marjan Ghazvininejad** **Gargi Ghosh**

Armen Aghajanyan **Sida Wang** **Luke Zettlemoyer**
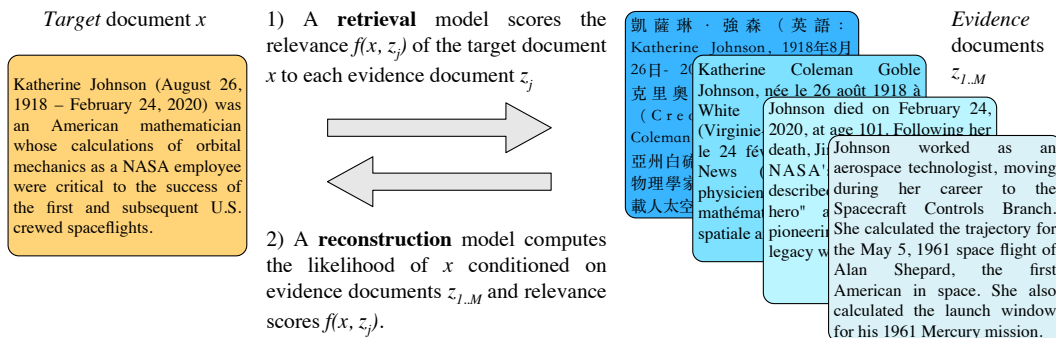
Facebook AI
mikelewis@fb.com

Figure 1: Pre-training via Paraphrasing: a retrieval model maps a document to a set of related documents, which a reconstruction model paraphrases to maximize the likelihood of the original. Example text adapted from `https://{en,es,de,it,fr,zh}.wikipedia.org/wiki/Katherine_Johnson`

Abstract

We introduce MARGE, a pre-trained sequence-to-sequence model learned with an unsupervised multi-lingual multi-document paraphrasing objective. MARGE provides an alternative to the dominant masked language modeling paradigm, where we self-supervise the *reconstruction* of target text by *retrieving* a set of related texts (in many languages) and conditioning on them to maximize the likelihood of generating the original. We show it is possible to jointly learn to do retrieval and reconstruction, given only a random initialization. The objective noisily captures aspects of paraphrase, translation, multi-document summarization, and information retrieval, allowing for strong zero-shot performance on several tasks. For example, with no additional task-specific training we achieve BLEU scores of up to 35.8 for document translation. We further show that fine-tuning gives strong performance on a range of discriminative and generative tasks in many languages, making MARGE the most generally applicable pre-training method to date.

1 Introduction

Variations on masked language models (MLMs) [Devlin et al., 2019, Liu et al., 2019, Yang et al., 2019b, Conneau et al., 2019, Lewis et al., 2019a, Raffel et al., 2019, Clark et al., 2020] provide highly effective self supervision for pre-training by removing and then reconstructing parts of an input text. In this paper, we present the first viable pretraining alternative to MLMs; self supervision is instead provided by learning to paraphrase collections of related documents in many languages.

More specifically, we introduce MARGE, a **M**ultilingual **A**utoencoder that **R**etrieves and **Ge**nerates. We train MARGE by self-supervising the *reconstruction* of target text by first *retrieving* a set of related texts (in many languages) and then conditioning on them to maximize the likelihood of generating the original. We pre-train a multi-source sequence to sequence model that separately encodes each retrieved document and decodes the target, piecing together and translating content from the appropriate inputs as needed to provide the best reconstruction possible. The retrieval model scores are used to bias the cross attention to the most relevant retrieved documents, allowing the retrieval model to be trained jointly from the reconstruction loss.

Our approach can be viewed as a new type of denoising auto-encoder where the noise comes from the retrieval step and is much more diverse than masking; retrieved documents may have little lexical overlap with the target, and may not even be in the same language, but should communicate the same underlying information. The pre-training task emphasizes paraphrasing and reduces the amount of encyclopedic knowledge the model must memorize. The set of retrieved documents and relevance scores are an autoencoder bottleneck from which the input must be reconstructed. MARGE is related to recent work that learns to do retrieval as part of the end task model, for example to find evidence documents in open domain question answering [Guu et al., 2020, Lewis et al., 2020]. This leads to a more challenging retrieval problem that, unlike ours, requires a separate pre-training phase.

Overall, our pre-trained models capture elements of traditional paraphrasing, translation, multi-document summarization, and information retrieval tasks.[1] This enables effective zero-shot learning; with no fine-tuning we achieve BLEU scores of up to 35.8 for document translation, and outperform strong baselines for cross-lingual transfer in summarization — providing a step towards pre-trained models that can perform any task with little or no fine-tuning. With fine-tuning, we achieve competitive performance with masked language models on a range of discriminative and generative tasks in many languages, making MARGE the most generally applicable pre-training method to date.

2 Model

2.1 Overview

During pre-training, the input to the model is a batch of evidence documents[2] $z_{1..M}$ and target documents $x_{1..N}$. The model is trained to maximize the likelihood of the targets, conditioned on the evidence documents, and the relevance of each evidence document to each target:

- The model first computes a relevance score $f(x_i, z_j)$ between every pair of documents x_i and z_j, by embedding each document and computing their cosine similarities (§2.2).

- The model then computes the likelihood of reconstructing each x_i conditioned on $z_{1..M}$ and each $f(x_i, \cdot)$, using a modified seq2seq model. The similarity score encourages the model to attend more to relevant evidence documents. Backpropagating the reconstruction loss therefore improves both the sequence-to-sequence model and the relevance model (§2.3).

- We construct batches so that evidence documents are relevant to the targets, using the relevance model for retrieval (§2.4).

Training this model is a chicken-and-egg problem. The reconstruction and relevance models cannot be effectively updated if the batches do not contain relevant evidence documents, but batch construction relies on a relevance model. However, we found that, in practice, the model is able to learn from a random initialization, which effectively provides a type of hashing of random features for each word.

---

[1]Masked language models, in contrast, are less directly related to target fine tuning tasks and significant ongoing research focuses on understanding why they work so well, see Rogers et al. [2020] for a survey.

[2]We use *document* to refer to contiguous chunks of text up to maximum length (here, 512 tokens).

2.2 Relevance Scores

To learn the relevance scores $f(x_i, z_j)$ for a pair of documents, we train a document encoder g that maps a list of tokens to a fixed size representation. We apply the same encoder to both the target and evidence document, and take the cosine similarity between their representations:

$$f(x, z) = \begin{cases} \frac{g(x) \cdot g(z)}{\|g(x)\| \|g(z)\|} & \text{if } x \neq z \\ -\infty & \text{otherwise} \end{cases} \tag{1}$$

This function is used in the reconstruction model (§2.3), and trained by the reconstruction loss. It is also used to construct batches of relevant documents (§2.4).

Using the same encoder for both the target and evidence documents allows even random models to compute meaningful similarity functions, as documents with high lexical overlap are more likely to be projected to more similar representations [Wieting and Kiela, 2019]. This is crucial at initialization.

We encode documents by taking the representation of the first token from the top of a 4-layer Transformer [Vaswani et al., 2017]. We share parameters with the first four layers of the reconstruction-model encoder, which saves computation and allows multitask learning.

2.3 Reconstruction Model

Given a set of evidence documents $z_{1..M}$ and similarity scores $f(x_i, z_j)$, the reconstruction model computes the likelihood of target document x_i.

$$L_\theta = -\sum_i \log p_\theta(x_i | z_{1..M}, f(x_i, z_1), \dots, f(x_i, z_M)) \tag{2}$$

This provides an auto-encoder loss where the reconstruction of document x_i is indirectly conditioned on x_i, but with an intermediate bottleneck provided by the retrieved documents and relevance scores.

First, the input documents are encoded individually with a bidirectional Transformer, and then the resulting embeddings are concatenated. The similarity score is used to bias the cross-attention from the decoder to the encoder, so that the decoder will pay more attention to more relevant evidence documents. Using more relevant evidence documents will improve the likelihood of reconstructing x_i, so gradient descent on (2) will improve the quality of the similarity scores.

Standard Transformer sequence-to-sequence models [Vaswani et al., 2017] compute a matrix of cross-attention probabilities between all elements of target document x_i and evidence document z_j:

$$\alpha = softmax_{z_j}(Q^{lh}(x_i)K^{lh}(z_j)) \in \mathbb{R}^{|x_i| \times |z_j|} \tag{3}$$

where Q^{lh} and K^{lh} compute query and key representations for layer l and head h, and $softmax_{z_j}$ denotes a softmax normalised over elements of z_j.

We instead compute cross attention over a set of evidence documents $z_{1..M}$, biasing the attention scores with the document relevant score from (1):

$$\alpha = softmax_{z_{1..M}}(Q^{lh}(x_i)K^{lh}(z_{1..M}) + \beta f(x_i, z_j)) \in \mathbb{R}^{|x_i| \times \sum_j |z_j|} \tag{4}$$

where β is a trainable scalar parameter that weights the importance of the document similarity score.

Guu et al. [2020] propose a related approach in which the likelihood of a target x is calculated by marginalizing out latent documents z: $p(x) = \sum_j p(x|z_j)p(z_j)$. Our attention-like mechanism is (1) more expressive, because it can pay complete attention to a token from one document at one timestep and a token from another document at another timestep, and (2) more efficient because $p(x|z)$ is not computed separately for each z_j. However, our method does not allow attention from z to x.

2.4 Batch Construction

Batches are constructed to create evidence document sets $z_{1..M}$ that give useful information for reconstructing target documents $x_{1..N}$, as detailed in this section. Overall, we divide the data into

shards of related documents. Periodically, we compute the similarities between pairs of documents within each shard, using the relevance model, and apply a threshold to keep the strongest connections. The final batches are constructed to maximize connectivity between evidence and target documents.

Document similarity   We compute document similarity in the same way as §2.2. All documents x are encoded as a vector $g(x) \in \mathbb{R}^d$, and then all pair-wise similarities between documents are computed with a single matrix multiplication.

Data Sharding   We use simple heuristic constraints to divide documents into related shards, to improve both the accuracy and efficiency of retrieval. Specifically, for news text, documents are in the same shard iff they were published on the same date. For Wikipedia, we split articles into chunks of length 512. We create 1000 shards, where all chunks from the same article, or the equivalent article in another language, are in the same shard (otherwise dividing chunks randomly). Shards typically contain 50-250k entries.

Indexing   While we backpropagate through the relevance model in (4), the construction of the batch itself is inherently non-differentiable. For convenience we perform the nearest neighbour search offline. Every 10k model updates, we sample a set of shards of documents. For each shard, we compute $f(x, z)$ for every pair of target and evidence documents, using the current relevance model.

Thresholding   We select which documents are sufficiently related by taking the top k most similar document pairs across all pairs in the shard. Some targets may have no sufficiently relevant evidence documents, and are unused until the shard is re-indexed with an updated relevance model.

Batching   We aim to construct batches containing clusters of related target and evidence documents, to maximize available information for reconstructing each target. The output from the thresholding step is a bipartite graph of evidence and target documents with edges between them. A batch is a subgraph, and we perform a small local search to find subgraphs maximizing the sum of the weights of all edges in the subgraph. To encourage the model to build multilingual batches, edges where the evidence and target are in different languages are given weight 100, and other edges have weight 1. To create batches, we iterate over seed evidence documents x_i with an edge to at least one evidence document. We then greedily add evidence and target documents to the batch to maximize the sum of the weights of edges, until the maximum number of tokens that can fit in GPU memory is reached.

3   Training

Architecture   We use a Transformer model [Vaswani et al., 2017]. The encoder consists of 12 Transformer layers of dimension 1024, with feedforward layers of size 4096. Recent work showed that large models train more efficiently [Li et al., 2020, Kaplan et al., 2020]. The decoder is similar to the encoder, but we increase the size of the feed-forward layers in the Transformer decoder to 16536. We also add 4 additional Transformer layers to the base of the decoder with only self-attention and feedforward layers of size 4096, which allows words in the target to contextualize locally before the more expensive cross-attention and feed-forward layers. We focus on scaling up the decoder, because it has access to more information than the encoder (which sees only evidence documents). In total, the model contains roughly 960M parameters. For the relevance model, we use the first 4 layers of the encoder, and take the documents representation from the beginning-of-sentence token.

Pre-training   During pre-training, workers process sub-batches containing an average of 2 evidence documents and 2 target documents, and accumulate gradients across workers. Using a multilingual version of the CC-NEWS corpus [Liu et al., 2019], we train initially using the with 64 workers for 450k steps (linearly annealing the learning rate from 1e-04 to 0 with 10k warmup steps), and then continue training with 2048 workers with 550k steps (annealing the learning rate from 2e-04 to 0).[3] We refer to this model as MARGE-NEWS. To explore domain effects, we further pre-train for 100k steps on Wikipedia data, annealing the learning rate from 1e-04 to 0, and refer to the resulting model as MARGE. We rebuild the index every 10k updates. We set retrieval thresholds such that we take on average 4 monolingual and 4 crosslingual links per target document.

---

[3]Initially training with a smaller learning rate reduced instability with an untrained retrieval model.

	#Parameters	#Languages	Pretraining task	Pretraining GPU Days (estimated)	Pretraining Data (GB; estimated)
mBERT	172M	104	MLM	Unknown	60
XLM	570M	100	MLM	640	60
XLM-R	550M	100	MLM	27000	2394
MMTE	192M	100	Translation	Unknown	Unknown
mBART	680M	25	seq2seq MLM	4500	1370
MARGE	963M	26	Retrieval+Reconstruction	4700	206

Table 1: Comparison models: MARGE is pre-trained on a scale between XLM and XLM-R.

	ar	de	fr	ja	zh	WMT19 de
			IWSLT2017			
Into English	26.8	28.5	34.3	12.6	19.9	35.8
From English	12.9	14.4	25.5	10.7	12.9	13.4

				Target		
		de	en	it	nl	ro
Source	de	-	30.6	14.0	14.8	11.6
	en	18.8	-	14.3	15.0	14.0
	it	14.0	31.7	-	11.3	12.7
	nl	14.3	27.5	12.6	-	9.3
	ro	14.3	32.8	14.4	9.8	-

Table 2: **Zero-shot unsupervised document level machine translation** BLEU scores using the pre-trained model, with no fine-tuning or special constraints on generation. Performance varies considerably across languages, but is non-trivial with even distantly related languages.

Data Pre-processing We de-duplicate the data, and identify languages using FastText [Joulin et al., 2016]. We select documents published in 26 different languages (based on their prevalence in downstream tasks), summarized in the Appendix. We divide documents into chunks of length 512. We allow all chunks to be evidence documents. For the news domain, we only allow the first chunk in each document to be used as a target, which we found improved performance during development. We prepend a language identifier token as the first decoder input, to control the output language.

Fine-tuning For fine-tuning, we use a similar procedure to Lewis et al. [2019a]. For generation problems, such as translation and summarization, the task input is fed into the encoder, and the output is generated by the decoder. For classification problems the task input is fed into both the encoder and decoder, and a representation is used from the decoder's final layer hidden state. For zero-shot transfer experiments, we freeze word embeddings and the first 4 decoder layers.

4 Experiments

As a multi-lingual sequence-to-sequence model, MARGE is applicable to a very broad range of tasks. We focus on multi-lingual tasks with elements of retrieval, document comprehension, and document generation, because they are the most directly related to our pre-training.

Table 1 lists the strongest available multilingual pre-trained models, along with relevant model statistics. We compare performance to published numbers for these models.

4.1 Cross-lingual Sentence Retrieval

Our pre-training task requires the model to retrieve similar texts, which may be in different languages. As an extrinsic evaluation of this functionality, we study cross-lingual sentence retrieval, in which a model must identify the correct translation of a sentence from a set of distractors. We report performance on BUCC2018 [Zweigenbaum et al., 2018] and Tatoeba [Artetxe and Schwenk, 2019].

We follow the setup of Hu et al. [2020], using no fine-tuning. As a document representation, we use the average embedding of the fifth encoder layer (tuned on BUCC development data).

On BUCC (Table 3), MARGE outperforms other unsupervised models by almost 10 points. On Tatoeba (see Appendix), there is significant variation across languages, but overall MARGE performs comparably to XLM-R and significantly better than other pre-trained models. Better results have been achieved on both tasks using labeled bitext for training [Artetxe and Schwenk, 2019], but our results suggest that our pre-training objective learns an effective cross-lingual retrieval function.

	de	fr	ru	zh	avg
mBERT	62.5	62.6	51.8	50.0	56.7
MMTE	67.9	63.9	54.3	53.3	59.8
XLM	56.3	63.9	60.6	46.6	56.8
XLM-R	67.5	66.5	73.5	56.7	66.0
MARGE	**78.8**	**75.9**	**77.3**	**71.6**	**75.9**

Table 3: Unsupervised Sentence Retrieval results on BUCC. MARGE outperforms other unsupervised models.

	en-de	zh-en
Random Initialization	7.7	3.2
HAN [Miculicich et al., 2018]	-	24.0
mBART (sentence)	38.0	28.4
mBART (document)	38.5	**29.6**
MARGE	**39.2**	28.4

Table 4: Supervised document-level machine translation. Comparison results are from Liu et al. [2020]. MARGE performs similarly to mBART.

4.2 Document-Level Machine Translation

During pre-training, the model can retrieve evidence documents in different languages to the target—in contrast to mBERT, XLM and mBART where instances are monolingual. We explore how well this pre-training approach learns to translate. We focus on document level translation tasks, and report document-level BLEU scores.[4] Following Liu et al. [2020], we segment documents into chunks of 512 tokens for training and generation, and then concatenate chunks of the same document.

Zero-Shot Unsupervised Document Translation   Translation offers a direct measure of how well the pre-trained model encoder and decoder work for different languages, and the extent to which the interface between them is language independent. Therefore, in contrast to prior work on unsupervised translation, we do not further fine-tune the model with iterative back-translation [Lample et al., 2017, Artetxe et al., 2017], or bitext in other language pairs [Johnson et al., 2017, Liu et al., 2020].

We measure both translation into English, which compares encoder performance for other languages, and translation out of English, which measures the decoder performance. Generation hyperparameters were minimally tuned on German/English development, and are shared across all translation pairs. We use a beam of size 6 and block repeated n-grams of length 8 [Fan et al., 2017].

Results are shown in Table 2. Performance varies considerably by language, but reaches 35.8 for German to English, which is the highest score we are aware of for system trained with no bitext. Performance is also strong for some languages using different scripts, such as Arabic to English. However, some languages work less well, notably Japanese. Generating non-English languages proves harder in all cases, particularly those with non-Latin alphabets, but English to French works well. Future work should explore up-sampling rarer languages during pre-training.

Qualitatively, we note that the translations are often good but less literal translations than the reference. This may cause BLEU scores to underestimate performance. It is likely that unsupervised performance could be further improved using iterative back-translation using MARGE as an initialization, but we focus here on examining the pre-trained model directly.

Supervised Document Translation   We also evaluate how well our models can be fine-tuned for translation using labeled bitext. To compare with mBART, we use the same English-German and Chinese-English document translation tasks from WMT19 and IWSLT2015. Table 4 show that MARGE and mBART perform similarly, with MARGE performing better on English-German and mBART on Chinese-English. Both outperform baselines by a wide margin.

4.3 Summarization

We evaluate monolingual sequence-to-sequence generation performance on text summarization tasks. We use the MLSum dataset [Scialom et al., 2020] to compare performance in several languages.

Results are shown in Table 5. MARGE outperforms an extractive mBERT model—the extractive oracle performance suggests that extractive models are very competitive on this dataset—and a seq2seq model without pre-training. In some cases, training one model on all languages (train all) improves results. Finally, we explore zero-shot summarization, where the model is trained on all

---

[4]All sentences in a document are concatenated prior to calculating BLEU, using SacreBLEU [Post, 2018].

		MLSum					
Model	Setting	de	es	fr	ru	tr	avg
Extractive Oracle	Oracle	52.30	35.78	37.69	29.80	45.78	29.81
Lead 3	Deterministic	33.09	13.70	19.69	5.94	28.90	13.65
Pointer-Generator	Train One	35.08	17.67	23.58	5.71	32.59	15.91
M-BERT	Train One	42.01	20.44	25.09	9.48	32.94	17.59
MARGE-NEWS	Zero-shot Transfer	30.01	17.81	19.39	8.67	29.39	15.05
MARGE-NEWS	Train One	42.60	22.31	**25.91**	10.85	**36.09**	19.03
MARGE	Train All	42.70	22.27	25.78	10.85	35.47	18.87
MARGE-NEWS	Train All	**42.77**	**22.72**	25.79	**11.03**	35.90	**19.09**

Table 5: ROUGE-L scores on MLSum. MARGE generates abstractive summaries that outperform an extractive mBERT model. We also demonstrate zero-shot transfer learning, where the model is trained only on languages it is not trained on, and results from training on all languages.

	en	ar	de	es	hi	vi	zh	avg	en	de	es	fr	ja	ko	zh	avg
mBERT	80.2	52.3	59.0	67.4	50.2	61.2	59.6	61.4	94.0	85.7	87.4	87.0	73.0	69.6	77.0	81.9
MMTE	78.5	56.1	58.4	64.9	46.2	59.4	58.3	60.3	93.1	85.1	87.2	86.9	72.0	69.2	75.9	81.3
XLM	68.6	42.5	50.8	54.7	34.4	48.3	40.5	48.5	94.0	85.9	88.3	87.4	69.3	64.8	76.5	80.9
XLM-R	83.5	**66.6**	**70.1**	**74.1**	**70.6**	**74.0**	62.1	**71.6**	**94.7**	**89.7**	90.1	90.4	78.7	**79.0**	82.3	86.4
MARGE	**83.7**	64.5	68.7	73.4	67.2	71.5	**67.8**	71.0	**94.7**	89.4	**91.6**	**90.9**	**78.9**	77.7	**82.5**	**86.5**

| (a) F1 scores on the MLQA question answering task. | (b) Paraphrasing accuracy on PAWS-X. |

Table 6: Cross-lingual transfer: models are trained on English (en) and tested on other languages. MARGE performs competitively with XLM-R, with 20% of the pre-training compute.

languages except the test language—this model outperforms a strong lead-3 baseline, and even a supervised pointer-generator model on Spanish and Russian. On this domain, we achieve better results with MARGE-NEWS, a version of the model trained only on news.

4.4 Paraphrasing

We measure how well our pre-training task learns paraphrasing on the PAWS-X paraphrase detection dataset [Yang et al., 2019a]. Models must determine whether two sentences are paraphrases; examples were constructed adversarially to have high lexical overlap. Models are trained on English, and we test zero-shot transfer to other languages. MARGE edges out a new state of the art (Table 6b).

4.5 Question Answering

Question answering offers another document level reasoning task that is easily posed in many languages. We use the MLQA dataset [Lewis et al., 2019b], in which models are trained on the English SQuAD dataset [Rajpurkar et al., 2016] and then tested in other languages. Results in Table 6a show that MARGE achieves competitive performance with XLM-R, setting the state of the art for Chinese, and outperforms other models by a wide margin.

5 Analysis

What does the reconstruction model learn? To build intuitions about what the reconstruction model learns, we examine model outputs for inputs in different languages on the same topic (Table 9). Even for a fixed topic, the model output varies significantly with the input, showing that it is not simply memorizing text. Almost all facts in the outputs are supported by the input, with few hallucinations—suggesting pre-training has taught the model to translate and paraphrase information from its source, rather than memorize facts in its parameters. However, the outputs are not literal translations—in particular, some important facts from the source are not expressed in the output. The model was not trained on literal translations, so it is perhaps surprising that the output is so closely aligned to the input. Translations may represent a mode of a diverse distribution over paraphrases.

7

Target \ Evidence	de	en	nl	es	fr	it	pt	ro	el	bg	pl	ru	bn	hi	ur	te	fi	tr	ar	id	th	vi	ja	ko	zh	sw
de	42.1	14.8	6.2	6.5	7.5	6.5	3.3	1.2	0.2	0.3	4.3	1.5	0	0	0	0	2.1	1	0.1	1	0	0.8	0.4	0.1	0.1	0
en	6.8	53.1	3.4	8.6	6.8	6.1	4.2	1.1	0.3	0.2	2.1	1.4	0	0	0	0	0.9	0.8	0.4	1.6	0	1.4	0.3	0.1	0.2	0
nl	1.2	1.4	91.7	1	0.8	0.7	0.7	0.2	0	0	0.6	0.1	0	0	0	0	0.3	0.2	0	0.5	0	0.3	0	0	0	0
es	5.6	15.9	4.3	40.8	7.3	7.8	7.3	1.3	0.3	0.3	2.6	1.4	0	0	0	0	1.2	0.9	0.2	1.2	0	1.1	0.3	0.1	0.2	0
fr	7.7	15	4.3	8.8	41.3	8.5	4.4	1.6	0.2	0.2	2.8	0.9	0	0	0	0	1.1	0.8	0.1	1.1	0	0.9	0.2	0.1	0.1	0
it	7.1	14.6	4.1	10	9.1	36.4	5.5	2	0.3	0.2	3.5	1.3	0	0	0	0	1.6	1	0.2	1.2	0	1.1	0.4	0.1	0.2	0
pt	5.2	14.2	5.5	13.3	6.8	7.9	32.7	1.8	0.4	0.4	3.2	1.4	0	0	0	0	1.5	1.3	0.3	1.8	0	1.6	0.3	0.1	0.2	0
ro	3.7	7.4	3.5	4.8	4.8	5.7	3.6	55.3	0.4	0.4	2.8	0.9	0	0	0	0	1.5	1.3	0.2	1.7	0	1.6	0.2	0.1	0.1	0
el	2.8	8	2.5	4.5	2.3	3.8	3.1	1.7	53.2	1.4	2.8	3.2	0	0.1	0	0	1.6	1.5	1.2	1.6	0	1.7	1.3	0.9	0.7	0
bg	3.7	5.1	2.7	4.3	2.3	2.9	3.1	1.8	1.4	36.4	3.3	25.1	0	0	0	0	1.3	0.9	0.8	1.5	0	1.1	1	0.5	0.7	0
pl	5.2	5.5	3.5	3.6	3.3	3.8	2.5	1.1	0.3	0.3	64.1	1.5	0	0	0	0	1.9	0.9	0.1	1	0	0.8	0.3	0.1	0.1	0
ru	2.5	5.2	1.2	2.7	1.5	2	1.4	0.5	0.4	3.2	2.1	73.7	0	0	0	0	0.5	0.4	0.3	0.5	0	0.4	0.8	0.3	0.4	0
bn	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
hi	0.2	2.2	0.1	0.5	0.2	0.3	0.3	0.2	0.3	0.1	0.1	0.5	0	92.2	0.2	0	0	0.2	1	0.3	0	0.3	0.2	0.2	0.2	0
ur	0.1	0.3	0.1	0.1	0.1	0.1	0.1	0	0.1	0	0.1	0.1	0	0.2	97.2	0	0	0.1	1	0.1	0	0.1	0.1	0.1	0.1	0
te	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	99.6	0	0	0.1	0	0	0	0	0	0	0
fi	9	8.3	6.4	6	4.8	6.2	4.3	2.1	0.6	0.4	7	1.4	0	0	0	0	34.8	2.4	0.2	2.5	0	2.2	0.7	0.4	0.3	0.1
tr	5.3	9.6	4	5.6	3.8	4.8	4.5	2.1	0.6	0.4	4	1.3	0	0	0	0	2.9	44.1	0.4	2.8	0	2.4	0.5	0.4	0.4	0.1
ar	0.5	3.8	0.3	1.2	0.5	0.7	0.8	0.3	0.4	0.3	0.4	0.7	0	0.1	0.2	0.1	0.2	0.3	87	0.5	0	0.5	0.4	0.3	0.3	0
id	1.4	4.8	4	2	1.5	1.6	1.7	0.8	0.2	0.2	1.1	0.4	0	0	0	0	0.8	0.8	0.2	76.5	0	1.5	0.2	0.1	0.2	0
th	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99.9	0	0	0	0	0
vi	2.3	8.7	4.7	3.5	2.4	2.8	2.9	1.4	0.4	0.2	1.9	0.7	0	0	0	0	1.4	1.3	0.3	3	0	60.5	0.6	0.3	0.5	0
ja	0.5	1	0.3	0.5	0.3	0.5	0.3	0.1	0.1	0.1	0.4	0.7	0	0	0	0	0.2	0.1	0.1	0.2	0	0.3	85.9	1.6	6.7	0
ko	0.5	1.2	0.4	0.7	0.3	0.6	0.5	0.2	0.4	0.2	0.6	1	0	0	0	0	0.6	0.5	0.4	0.4	0	0.5	6.7	82.6	1.7	0
zh	0.2	0.6	0.2	0.3	0.1	0.2	0.2	0.1	0.1	0.1	0.2	0.3	0	0	0	0	0.1	0.1	0.1	0.2	0	0.2	6.2	0.4	90.3	0
sw	0.7	1.7	0.8	1	0.6	0.7	0.7	0.7	0.1	0.1	1	0.1	0	0	0	0	1.7	1	0.3	1.4	0	1	0	0.1	0	86.2

Language family groupings: germanic, romance, slavic, indo-iranian, sino-tibetan
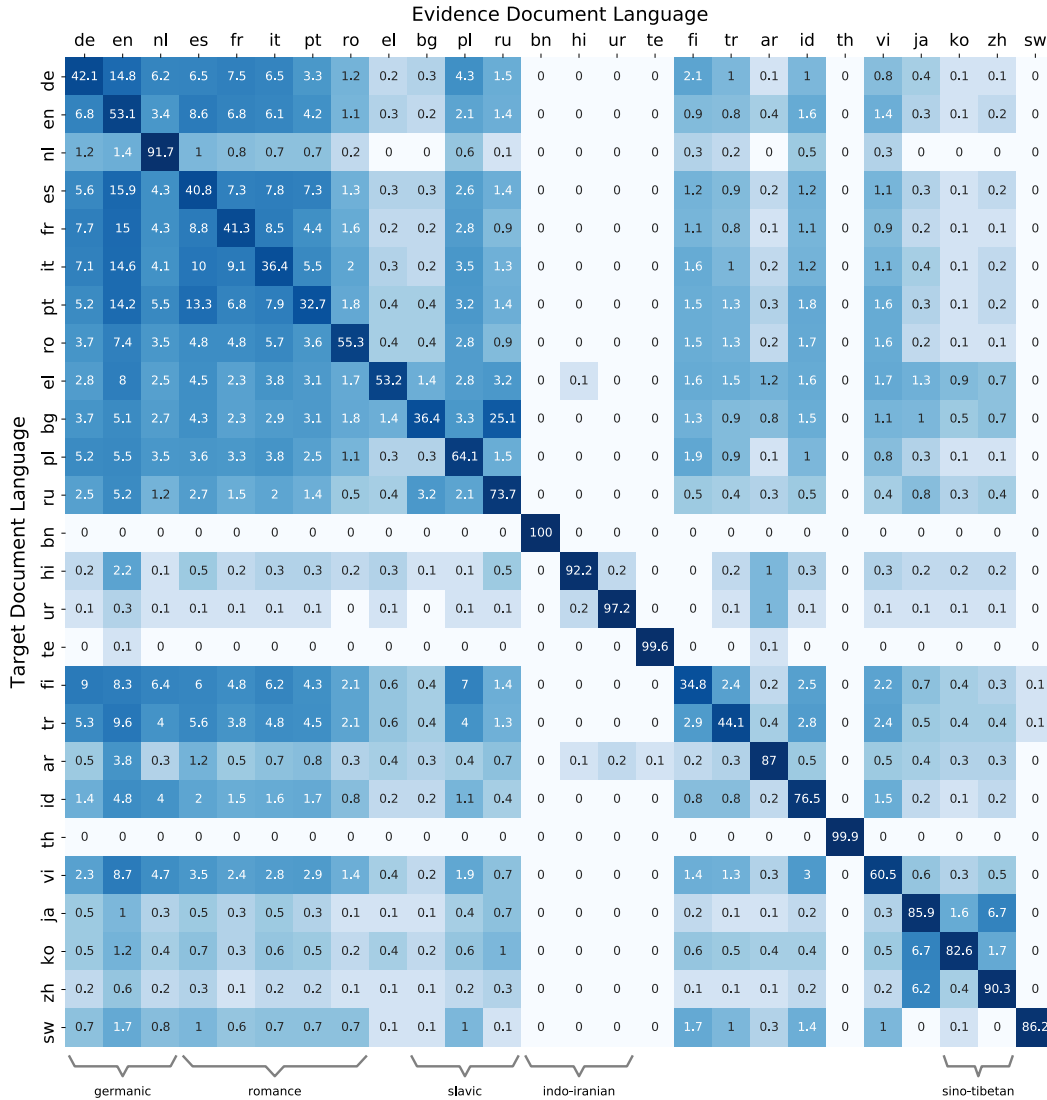
Figure 2: Percentage of retrieved links to documents in target languages (y-axis) from evidence documents in different source languages (x-axis) on Wikipedia.

What does the retrieval model learn? Figure 2 shows statistics of the retrieval model. Differences across languages are due to many factors, including the frequency of languages in the corpus, and how related languages are to each other. Our pre-training also introduces feedback loops, because if the reconstruction model is unable to translate between two languages, it may train the retrieval model that documents in these languages are less relevant to each other.

All languages retrieve the highest proportion of documents within their own language (represented by the diagonal), but otherwise the retrieved documents tend to be distributed over a number of other languages. There tend to be closer affinities between geographically or linguistically related languages, such as Bulgarian and Russian, or Chinese and Japanese. For some languages, the model fails to retrieve many documents in other languages—particularly Indo-Iranian languages, and those which are the only example of their language family we include (such as Telugu and Thai). For these cases, the pre-training reduces to independent updates for each language, as in mBART and mBERT.

Discussion Overall, MARGE shows strong performance on a wider range of tasks than any previous pre-trained models, and is effective at discriminative and generative tasks in many languages. Results are competitive with less general models, even XLM-R, which was trained with significantly

higher pre-training resources. The pre-training task is more closely related to downstream tasks than masked language modeling, allowing pre-trained models to achieve BLEU scores as high as 35.8 for translation. MARGE also broadens the range of known effective pre-training tasks beyond MLMs, which we hope will lead to further exploration and understanding of pre-training objectives.

However, there are several limitations that future work should address. We pre-trained on news and Wikipedia, where simple metadata can be used to constrain the similarity search, improving efficiency and accuracy. Broadening the domains may require approximate nearest neighbor search [Johnson et al., 2019]. Learning the retrieval model requires batch sizes greater than one, so model-parallel training would be required to train significantly larger models.

6 Related Work

NLP pre-training   Since BERT [Devlin et al., 2019], pre-training for NLP has been dominated by variants of masked language models. For example, Yang et al. [2019b] predicts the masked tokens auto-regressively, Dong et al. [2019] multitasks MLM and language modeling objectives, Clark et al. [2020] trains a discriminator to classify the correctness of MLM samples, and Lewis et al. [2019a] and Raffel et al. [2019] use seq2seq models with masked inputs. MARGE departs significantly from these objectives in that the inputs during pre-training are complete, uncorrupted text.

Bitext Mining   Recent work has shown impressive results on machine translation through bitext mining [Schwenk et al., 2019], in which a retrieval model is used to search for parallel sentences in a large multilingual corpus, which are then used as training data for a machine translation model. A key conceptual difference is that literal bitext is not optimal for our approach, as we hope to learn linguistic information by training on noisy document-level paraphrases. We also learn to retrieve and translate with no manually translated sentences, unlike existing bitext mining methods.

Cross-lingual Learning   Several works attempt to pre-train language-independent representations. McCann et al. [2017] and Siddhant et al. [2019] pre-train on translation tasks. Better resuts are achieved using MLMs on the concatenation of monolingual corpora, relying on parameter sharing to learn cross-lingual representations [Lample and Conneau, 2019, Conneau et al., 2019, Liu et al., 2020]. We instead pre-train on loose cross-lingual paraphrases.

Language Models with Retrieval   Several recent papers have shown that word prediction can be improved by retrieving relevant evidence documents. Guu et al. [2020] and Lewis et al. [2020] improve MLMs and text generation by learning to retrieve relevant evidence documents. Guu et al. [2018] perform language modeling by retrieving and editing sentences. kNN-LM [Khandelwal et al., 2019] shows that language models can be improved with retrieving from the training set, by interpolating a language model with a nearest neighbor classifier. In contrast, we learn retrieval during training but do not require it for inference. Perhaps most relevantly, Liu et al. [2018] generate Wikipedia articles conditioned on a set of evidence documents.

7 Conclusion

We introduced a new approach to pre-training models for natural language understanding and generation, by using retrieved documents to reconstruct the original. MARGE exhibits strong performance on a range of discriminative and generative tasks in many languages, both with and without fine-tuning. These results establish MARGE as a viable alternative to masked language modeling and provide a step towards pre-trained models that can perform any task with little or no fine-tuning. Future work should scale MARGE to more domains and languages.

Broader Impact

This work has a broad scope, covering discriminative and generative tasks in many languages. As such, the broader impact is similar to that of the field of NLP; there exist many potential good and bad applications. The pre-training method is likely to capture and potentially amplify any biases found in the pre-training corpus. The model should not be used in applications with real world consequences without carefully considering any potential for harm.

References

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2017.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*, 2019.

Angela Fan, David Grangier, and Michael Auli. Controllable abstractive summarization. *arXiv preprint arXiv:1711.05217*, 2017.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450, 2018.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*, 2020.

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

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 2019.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*, 2019.

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

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019a.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*, 2019b.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020.

Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joseph E Gonzalez. Train large, then compress: Rethinking model size for efficient training and inference of transformers. *arXiv preprint arXiv:2002.11794*, 2020.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

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

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*, 2020.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305, 2017.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. Document-level neural machine translation with hierarchical attention networks. *arXiv preprint arXiv:1809.01576*, 2018.

Matt Post. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*, 2018.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*, 2020.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*, 2019.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. Mlsum: The multilingual summarization corpus. *arXiv preprint arXiv:2004.14900*, 2020.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Arivazhagan, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. *arXiv preprint arXiv:1909.00437*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

John Wieting and Douwe Kiela. No training required: Exploring random encoders for sentence classification. *arXiv preprint arXiv:1901.10444*, 2019.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*, 2019a.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019b.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th Workshop on Building and Using Comparable Corpora*, pages 39–42, 2018.

A Additional Results

	ar	bg	bn	de	el	es	fi	fr	hi	id	it	ja
XLM-R	47.5	**71.6**	**43.0**	88.8	**61.8**	75.7	**71.6**	73.7	**72.2**	**77.0**	68.3	**60.6**
MARGE	**49.9**	70.5	16.9	**88.9**	57.2	**82.9**	55.8	**77.0**	67.1	73.8	**76.5**	60.1

	ko	nl	pt	ru	sw	te	th	tr	ur	vi	zh
XLM-R	**61.4**	80.8	82.2	74.1	20.3	**35.9**	29.4	**65.7**	24.3	74.7	68.3
MARGE	50.6	**84.3**	**84.8**	**78.7**	**22.8**	16.2	**38.0**	63.2	**41.9**	**77.3**	**77.2**

Table 7: **Tatoeba** zero-shot sentence retrieval results. MARGE performs comparably to XLM-R, but with significant variation across languages. We only show results for languages in all model's pre-training data.

B Pre-training Data

Language	Code	Language Family	CCNews	Wikipedia
Arabic	ar	Afro-Asiatic	2416996	747891
Bulgarian	bg	Slavic	496023	297989
Bengali	bn	Indo-Iranian	741	134560
German	de	Germanic	13320055	2735591
Greek	el	Hellenic	1793198	317780
English	en	Germanic	57061325	6372976
Spanish	es	Romance	16990991	2111406
Finnish	fi	Uralic	471029	496988
French	fr	Romance	7281926	2749382
Hindi	hi	Indo-Iranian	1907850	124816
Indonesian	id	Austronesian	1295060	435599
Italian	it	Romance	6865752	1776998
Japanese	ja	Japonic	458675	1311915
Korean	ko	Sino-Tibetan	1241560	442675
Dutch	nl	Germanic	2091796	1359535
Polish	pl	Slavic	1153817	1219494
Portuguese	pt	Romance	2971009	1107798
Romanian	ro	Romance	1960236	348036
Russian	ru	Slavic	6579113	1939546
Swahili	sw	Niger-Congo	11878	34107
Telugu	te	Dravidian	7155	80131
Thai	th	Kra-Dai	5412	156505
Turkish	tr	Turkic	3524089	353028
Urdu	ur	Indo-Iranian	154912	96773
Vietnamese	vi	Austro-Asiatic	1019445	566375
Chinese	zh	Sino-Tibetan	434378	1027950

Table 8: Number of documents per language used for pre-training. Languages represent a range of families and geographical regions. The Germanic, Hellenic, Romance, Slavic, and Indo-Iranian families are part of a broader Indo-European family.

Language	Source	Zero-shot Unsupervised Output
French	Katherine Coleman Goble Johnson, née le 26 août 1918 à White Sulphur Springs (Virginie-Occidentale) et morte le 24 février 2020 à Newport News (Virginie), est une physicienne, mathématicienne et ingénieure spatiale américaine.Elle contribue aux programmes aéronautiques et spatiaux du National Advisory Committee for Aeronautics (NACA) puis de la National Aeronautics and Space Administration (NASA).	Katherine Coleman Goble Johnson (August 26, 1918 – February 24, 2020) was an American physicist, mathematician, and space engineer. She contributed to the aeronautics and space programs of the National Advisory Committee for Aeronautics (NACA) and later the National Aeronautics and Space Administration (NASA).
Spanish	Katherine Coleman Goble Johnson (White Sulphur Springs, Virginia Occidental; 26 de agosto de 1918 - Newport News, Virginia; 24 de febrero de 2020) fue una física, científica espacial y matemática estadounidense que contribuyó a la aeronáutica de los Estados Unidos y sus programas espaciales con la aplicación temprana de las computadoras electrónicas digitales en la NASA. Conocida por su precisión en la navegación astronómica, calculó la trayectoria para el Proyecto Mercury y el vuelo del Apolo 11 a la Luna en 1969.	Katherine Coleman Goble Johnson (August 26, 1918 – February 24, 2020) was an American physicist, space scientist, and mathematician who contributed to the United States aeronautics and its space programs with the early application of digital electronic computers at NASA. Known for her accuracy in astronomical navigation, she calculated the trajectory for Project Mercury and the Apollo 11 flight to the Moon in 1969.
German	Katherine G. Johnson (gebürtig Coleman, zwischenzeitlich verheiratet Goble; * 26. August 1918 in White Sulphur Springs, West Virginia; † 24. Februar 2020 in Newport News, Virginia) war eine US-amerikanische Mathematikerin afroamerikanischer Abstammung. Für ihre Beiträge zur Berechnung der Flugbahnen für das Mercury-Programm und den ersten bemannten Flug zum Mond im Rahmen der Apollo-11-Mission wurde sie Ende 2015 mit der Presidential Medal of Freedom ausgezeichnet.	Katherine G. Johnson (née Coleman; August 26, 1918 – February 24, 2020) was an American mathematician. She was awarded the Presidential Medal of Freedom in 2015 for her contributions to the calculation of the flight paths for the Mercury program and the first manned flight to the Moon in the Apollo 11 mission.
Italian	Katherine Coleman Goble Johnson (White Sulphur Springs, 26 agosto 1918 – Hampton, 24 febbraio 2020) è stata una matematica, informatica e fisica statunitense. Contribuì alla scienza dell'aeronautica statunitense e ai programmi spaziali, già dal primo utilizzo dei computer elettronici digitali da parte della NASA. Venne molto apprezzata per l'accuratezza che poneva nel calcolo della navigazione spaziale computerizzata e per il lavoro tecnico dirigenziale pluridecennale svolto alla NASA: da quando calcolava le traiettorie delle orbite, paraboliche e iperboliche, le finestre di lancio e i percorsi di ritorno di emergenza per molti voli, al Project Mercury, incluse le prime missioni NASA di John Glenn, Alan Shepard, le traiettorie di inserzione lunare nei voli lunari del programma Apollo, continuando con il lavoro sul programma dello Space Shuttle, infine con la progettazione dei primi piani per la missione su Marte.	Katherine Coleman Goble Johnson (White Sulphur Springs, August 26, 1918 – Hampton, February 24, 2020) was an American mathematician, computer scientist, and physicist. She contributed to the science of the U.S. Air Force and space programs, as early as the first use of digital electronic computers by NASA. She was highly regarded for the accuracy she put into computerized space navigation calculations and for the decades-long technical leadership work she performed at NASA: from calculating orbital trajectories, parabolic and hyperbolic, launch windows, and emergency return paths for many flights, to Project Mercury, including the first NASA missions of John Glenn, Alan Shepard, lunar insertion trajectories in the Apollo lunar flights, continuing work on the Space Shuttle program, and finally designing the initial plans for the Mars mission.

Table 9: Example zero-shot unsupervised inputs and outputs (truncated for clarity).