Synthetic Cross-language Information Retrieval Training Data

James Mayfield
HLTCOE, Johns Hopkins University
Baltimore, MD, USA
mayfield@jhu.edu

Eugene Yang
HLTCOE, Johns Hopkins University
Baltimore, MD, USA
eugene.yang@jhu.edu

Dawn Lawrie
HLTCOE, Johns Hopkins University
Baltimore, MD, USA
lawrie@jhu.edu

Samuel Barham
HLTCOE, Johns Hopkins University
Baltimore, MD, USA
samuel.barham@jhuapl.edu

Orion Weller
Johns Hopkins University
Baltimore, MD, USA
oweller@cs.jhu.edu

Marc Mason
HLTCOE, Johns Hopkins University
Baltimore, MD, USA
mmason8@jhu.edu

Suraj Nair
University of Maryland
College Park, MD, USA
srnair@umd.edu

Scott Miller
ISI, University of Southern California
Boston, MA, USA
smiller@isi.edu

ABSTRACT

A key stumbling block for neural cross-language information retrieval (CLIR) systems has been the paucity of training data. The appearance of the MS MARCO monolingual training set led to significant advances in the state of the art in neural monolingual retrieval. By translating the MS MARCO documents into other languages using machine translation, this resource has been made useful to the CLIR community. Yet such translation suffers from a number of problems. While MS MARCO is a large resource, it is of fixed size; its genre and domain of discourse are fixed; and the translated documents are not written in the language of a native speaker of the language, but rather in translationese. To address these problems, we introduce the JH-POLO CLIR training set creation methodology. The approach begins by selecting a pair of non-English passages. A generative large language model is then used to produce an English query for which the first passage is relevant and the second passage is not relevant. By repeating this process, collections of arbitrary size can be created in the style of MS MARCO but using naturally-occurring documents in any desired genre and domain of discourse. This paper describes the methodology in detail, shows its use in creating new CLIR training sets, and describes experiments using the newly created training data.

CCS CONCEPTS

• **Information systems** → **Multilingual and cross-lingual retrieval**; *Document collection models.*

KEYWORDS

cross-language information retrieval, CLIR, synthetic training data. domain shift, GPT-3

1 INTRODUCTION

As with many other human language technologies, neural models have recently achieved state-of-the-art performance in monolingual ad hoc information retrieval (IR). A key enabler of these advances has been the appearance of large IR training sets such as MS MARCO [3]. MS MARCO was developed by mining Bing query logs to identify, for each query, a relevant and a non-relevant document drawn from the Bing index. This makes MS MARCO well-suited to training IR systems for web-style queries where the documents are English webpages. It is less well-suited to other document languages, query styles and document genres as Dai et al. [10] demonstrate. Nonetheless, MS MARCO has been the basis for much of the improvement in IR achieved by neural systems.

In cross-language information retrieval (CLIR) there has been no resource comparable to MS MARCO. A number of CLIR collections are available. HC4[28][1] and TREC NeuCLIR 1 [27][2] are high-quality ad hoc CLIR collections, but are too small to serve as training data for a neural system. Collections such as CLIRMatrix[48],[3] XOR-QA[2],[4] and MIRACL[60][5] cover numerous languages, but like MS MARCO are focused on question answering and are biased towards Wikipedia articles. Their relevant documents are also not paired with non-relevant counterparts.

Given the lack of appropriate training sets for ad hoc CLIR, the research community has used machine translation to translate the MS MARCO documents into other languages. This has resulted in collections such as mMARCO[5], and NeuMARCO[6] of the same size as MS MARCO with queries in English and documents in another language. Using these resources, neural systems have achieved state-of-the-art CLIR performance.

Yet such translated training collections suffer from a number of problems. While MS MARCO is a large resource, it is of fixed size; thus, the amount of available training data is limited. More importantly, the genre and domain of discourse of the collection are fixed; documents are drawn from the Bing index and do not include, for example, informal communications such as email and Tweets. In addition, the translated documents are not written by a native speaker of the language, but rather suffer from a phenomenon known as *translationese* [51]: translation artifacts that have

[1] https://github.com/hltcoe/hc4
[2] https://neuclir.github.io/neuclir1.html
[3] https://github.com/ssun32/CLIRMatrix
[4] https://github.com/AkariAsai/XORQA
[5] https://github.com/project-miracl/miracl
[6] https://ir-datasets.com/neumarco.html

been shown to affect cross-language transfer performance [1]. Furthermore, MS MARCO is available only for research purposes,[7] so commercial systems and other non-research applications cannot make use of it.

To address these problems, we introduce the JH-POLO training set creation methodology. JH-POLO starts with a pair of non-English passages. These passages can be written by native speakers of the language, and can be drawn from any genre or domain. Thus, a collection generated using the JH-POLO methodology can be tailored to any desired retrieval setting.

Once a passage pair has been selected, an English query is automatically generated for which one passage of the pair is relevant and the other passage is not. We use English as the query language to match the available CLIR test collections. This creates an MS MARCO-style training example comprising a query, a relevant passage, and a non-relevant passage. A generative large language model (LLM) such as GPT-3 [8] is used to produce the English query. By repeating this process, a training collection of arbitrary size can be created.

This paper describes the JH-POLO methodology in detail, shows its use in creating new CLIR training sets, and describes experiments that demonstrate the efficacy of the approach.

We make the following contributions:

- We show that it is possible to generate a viable large CLIR training set automatically using only a target document collection and a generative LLM. To our knowledge, this is the first automatically generated CLIR training collection that uses natively-written passages.
- We show that negative training examples can be selected *before* generating the retrieval query to which they are not relevant, thereby allowing some control over the difficulty of negative examples in the generated collection.
- We show that training using the JH-POLO methodology is comparable to using machine-translated MS MARCO data when the documents to be searched are similar to the web documents used by MS MARCO documents, and more effective than training exclusively on MS MARCO when the domain or genre of the evaluation document collection deviates from that of the MS MARCO documents.

2 BACKGROUND

2.1 Cross-Language Information Retrieval

When moving from monolingual IR to CLIR, there is the added complexity of crossing the language barrier between the query expression and the document expression. One popular approach is to use a Machine Translation (MT) system to translate either the queries or the documents, to achieve a monolingual space where a monolingual IR system can be used [12, 37, 61]. Another approach generates dense representations of queries and documents. Matching queries to documents happens in a shared multilingual vector space; this approach is popularly known as *dense retrieval*. Pre-BERT [11] dense retrieval models used non-contextualized cross-language word representations to perform CLIR matching [16, 32, 56]. The adoption of large multilingual pretrained language models (mPLMs) such as mBERT [11] and XLM-R [9] led to dense neural CLIR systems that use contextualized representations for matching [33–35, 44]. Dense retrieval models for CLIR now rely heavily on mPLMs as the backbone of the models. However, Litschko et al. [34] demonstrate that performance using an off-the-shelf mPLM for CLIR is suboptimal. While sufficient training data is available to fine-tune an English system in the form of MS MARCO [3], analogous CLIR data is not natively available. Translated versions of MS MARCO, where the MS MARCO documents are replaced with machine translation output, have been used to fill this gap [5, 35].

This paper focuses on an alternative approach to fine-tuning CLIR systems. We explore the synthetic generation of queries from passages selected from a target document collection. Rather than effectiveness being dependent on the quality of an mPLM or the quality of machine translation, in this approach effectiveness is dependent on the ability of a generative LLM to produce effective training examples.

Dense Passage Retrieval (DPR) [24] and ColBERT [26] are two of the most commonly studied and highest performing dense retrieval models. DPR computes the similarity of the query classification (CLS) token and the CLS token of each document. ColBERT computes similarities between each pair of query and document tokens and scores a document by the sum of the maximum similarity (MaxSim) of each query token [26]. Compared to other neural reranking models such as a cross-encoder [38], dense retrieval models limit ranking latency by separating query and document transformer networks to support offline indexing.

DPR-X [54, 55, 59] and ColBERT-X [35] are the CLIR counterparts of DPR and ColBERT. Both use an mPLM as the underlying language model for crossing the language barrier. Exploiting both multilinguality and improved pre-training from XLM-R [9], DPR-X and ColBERT-X seek to generate similar contextual embeddings for terms with similar meanings, regardless of their language. These are the two retrieval models featured in our experiments.

2.2 LLMs and Retrieval

Language models are now tightly integrated with information retrieval systems. These combined systems are used for a broad range of knowledge-intensive problems, including open-domain question answering [20, 21], conversational assistants [45, 46], fact-checking [49, 50], and even improving language modeling itself [6, 31].

At times these systems are simply combinations of separate processes [15, 20, 40], while other times they are trained end-to-end from retrieval to the downstream task [21, 23, 30]. Due to the size of LLMs, they are typically used as separate components, with retrieval results passed to the LLM [15, 25, 47]. A nascent line of work has even proposed ignoring retrieval entirely and using LLMs to generate a relevant document in lieu of search [57]. In contrast to much of the research cited in this section, our work aims to use LLMs to improve IR models, rather than using retrieval to improve LLMs on NLP tasks.

2.3 Synthetic Query and Document Generation

Using LLMs to improve IR models through synthetic data generation has also been a burgeoning area of interest [19, 41–43, 52]. A

[7] https://microsoft.github.io/msmarco/

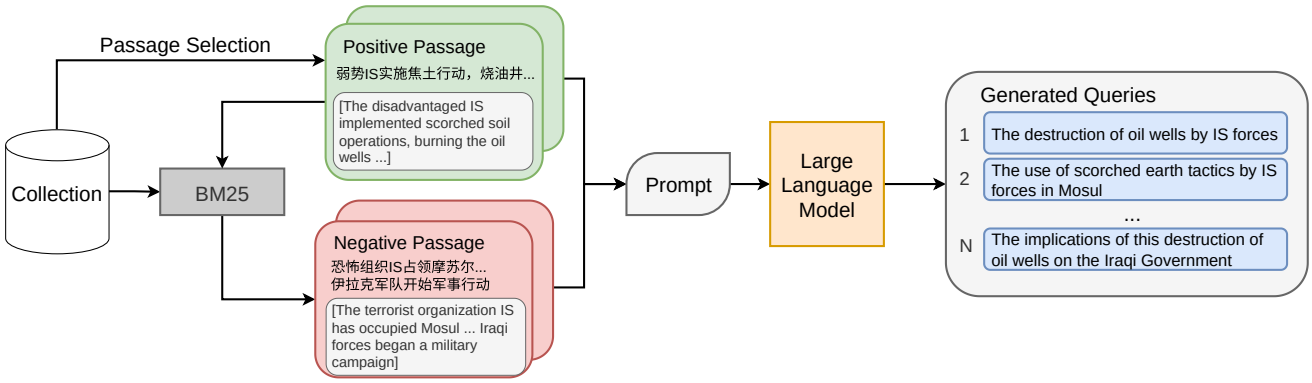

Figure 1: A depiction of the basic JH-POLO methodology. A target language passage (Chinese in this example, translated into English for convenience) is selected randomly from the target passage collection, and BM25 retrieval is used to identify a related passage. The two Passages are presented to a large language model, which is then prompted to generate queries for which one passage is relevant and the other is not.

prominent early example is the doc2query [39] family of algorithms, which supports the generation of a query that is relevant to a given document and which is then appended to it as a form of document expansion. As language models have grown in size and ability [8], there has been a surge of interest in this topic. HyDE [13] uses LLMs to generate a synthetic document that is then used as a query, while the InPars algorithms [4, 7, 22] and PROMPTAGATOR [10] use LLMs to generate queries given document, in the reranking and end-to-end settings respectively. These works differ in how they prompt the LLMs: PROMPTAGATOR uses a different prompt template for each dataset and only shows relevant few-shot examples (i.e., what the LLM should generate) while InPars also uses non-relevant few-shot examples (i.e., what not to generate).

Despite the plethora of recent research in creating synthetic training data for IR, to date, and with a few exceptions (e.g., HyDE [13]), most work has focused on the English language. This leaves it unclear how LLMs can be used to train translingual or multilingual IR systems.

3 JH-POLO METHOD

Generation of a single training example starts with the selection of two passages.[8] A generative LLM is given these passages and prompted to compose an English query for which one passage is relevant and the other is not. This process is repeated to generate as many training examples as desired.

This method has two significant advantages:

(1) It ensures that the passages are naturally-occurring text selected from the language, genre and domain of interest. Use of MS MARCO for CLIR has relied on machine translation of the MS MARCO document collection, which exhibits artifacts of machine translation. Furthermore, there is no way to alter the characteristics of the document collection underlying MS MARCO.

(2) It exploits a generative LLM's strength, which is generating short English[9] texts. LLMs can struggle when trying to generate a long document. Its capabilities in languages other than English are also inconsistent. By generating short English queries these problems with LLMs are ameliorated.

Figure 1 is a pictorial representation of the JH-POLO process. Section 3.1 describes the left side of the figure, while Section 3.2 describes the right side of the figure.

3.1 Passage Selection

Choosing passages at random would be a simple way to select two passages for use in query generation. However, doing so would almost always select two passages with no topic overlap. Any system trained using such pairs would have a difficult time distinguishing passages with a high topic overlap at test time. We would like our training data set to include related passages that exhibit significant overlap with a relevant passage but are not themselves relevant. We hypothesize that the closer the content of the two passages, the more useful the pair will be for training. There are a number of ways to choose two related passages; these include:

- Use an existing document collection and passage pairs. MS MARCO is the obvious target here; it has passages, and topics with an example of a relevant passage and a non-relevant passage for each topic.
- Use an existing ad hoc IR collection. For example the TREC NeuCLIR track[10] provides English topics with documents in Chinese, Persian, or Russian. One way to select a pair is to use the relevance judgments (qrels) to select two passages, one from a randomly chosen judged relevant document and the other from a randomly chosen judged non-relevant document for a given topic. This does not guarantee that the same relevance judgments will apply to the selected passages, but it is likely that those passages will be related but not identical. Alternatively, one could perform retrieval

[8] Full-length documents could exceed length limits imposed by the LLM.

[9] At this writing the major generative LLMs focus on English.

[10] https://neuclir.github.io/

on the original queries, and use the ranked results to select two top-scoring passages.
- Use a collection with relatedness links. One could for example select two linked Wikipedia articles, or two versions of a single Wikipedia article from two different dates.
- Select a passage at random, or one returned by a query, and use that entire passage as a query. Select the top retrieved passage whose BM25 retrieval score is at least a fixed threshold away from the score of the query passage as the negative passage.

The last approach is the one explored in this paper. By requiring at least some separation in the BM25 scores of the two passages, we ensure that the two passages contain some different information. We also require that the passages do not come from the same underlying source document. Different genres may also necessitate additional requirements to ensure the selection of useful training pairs. For instance, we examine the longest common substring between two passages sourced from informal communications; the selected passage must contain both twenty characters and 40% of its total characters outside of that common substring.

3.2 Prompt Specification

Unlike pre-trained language models that are routinely fine-tuned to adapt them to new genres, domains, or tasks, the common and economic way to use a generative LLM such as GPT-3 is to engineer a prompt to guide the desired generation. We experimented with a variety of prompts with the goal of creating suitable CLIR queries. Such a prompt must:

- contain the text of each of the passages.
- indicate what type of output is required. We would like to produce multiple output queries for each prompt to reduce the overall cost of building the collection.
- ensure that the generated queries are written in English regardless of the language of the passages.
- communicate what is meant by relevance.
- require that one of the passages is relevant to the output query and the other is not.

Figure 2 shows the basic prompt we used to create the training collections described in this paper. Here, {first} and {second} are replaced with the complete text of the first and second passages. The prompt requests five outputs for each passage, requires that the output is in English, and stipulates that one passage must be relevant and the other not relevant. Relevance is defined relative to an analyst writing a report; a passage is relevant if it helps the analyst write the report, and not relevant if it does not. The topic of the report is not specified in the prompt; the LLM is free to invent any report topic it likes. Thus the output query can be on any suitable topic.

We experimented with few-shot prompts that included sample outputs. These prompts had two problems. First, they increased the length of the prompt, increasing the cost of the request, which for GPT-3 is dependent on the sum of the lengths of the input and output. Second, there was occasionally bleed-through of the topics of the sample outputs into the queries produced. As a result, we restricted our attention to zero-shot prompts that relied purely on description of the desired output.

3.3 Crossing the Language Barrier

We use GPT-3 Davinci-3[11] as our large language model for two reasons. First, its input buffer is 4000 tokens, allowing us to include passages of up to about 550 words of Chinese, 260 words of Russian, or 370 words of Persian, while allowing an additional 100 or so tokens of English for the prompt. Second, Davinci is far more capable with languages other than English than are the lesser GPT-3 models. We present non-English passages to GPT-3 with no indication of what language they are written in; the prompt indicates only that they are 'documents.' Davinci seems to handle other languages with ease; the lesser models do not.

The ability of GPT-3 Davinci-3 to handle languages other than English varies dramatically by language [8]. If GPT-3 is unable to handle a given foreign language well, an alternative is to use machine translation to produce English versions of the documents. Then an English-only process is applied to these translations. This approach relies on document relevance not changing much when a document is translated. This is plausible, although the claim remains to be proven. It should be noted that while the LLM would process the translated documents in this case, the CLIR fine-tuning would continue to use the original natively written documents.

3.4 Failure Modes

We have identified four categories of error most commonly seen in JH-POLO output. The following describes them in detail.

Underspecification. This occurs when the query could refer to something in the passage, but could just as easily refer to many other things completely unrelated to the passage. For example, "The emergence of images in the media related to the leak" could refer to any of a number of instances of leaked documents. This failure mode can be thought of as inadequate inclusion of context in the query. Despite the apparent problem, we believe underspecified queries are less damaging as training data than other categories of errors because the negative passage is still not relevant to an underspecified query.

Overspecification. This occurs when the selected passage is relevant to the query, but no other passage is likely to be relevant. For example, "The arrest of Moise's bodyguards and 3 security personnel" describes one very specific facet of an event, one that will likely be found in very few of the passages about the event. This failure mode frequently occurs when numbers are part of the generated query, because it limits the query to a particular instance of a topic. While the resulting relevance judgments are still consistent, training with too many queries of this type may not be as useful for system performance since they are unlikely to capture the characteristics that the system needs to learn.

Hallucination. Sometimes the LLM inserts a detail into the query that is completely unrelated to anything in the source passage. For example, the query "I seek news about Hong Kong's women's basketball team" produced from a passage that includes information about a police basketball team and a mother who is a representative for a youth basketball team, but no mention of women's basketball teams, is a hallucination. This is a more problematic failure mode since the passage labeled relevant is not actually relevant to the query.

[11] https://beta.openai.com/docs/models/gpt-3

```
This is document A: <<{first}>>
This is document B: <<{second}>>

I am an analyst writing a report. Only one of the documents will help me write my report.  For each
document, describe in English, one per line, five things my report might be about for which that
document will help me write my report and the other document will not help me write my report.
```

Figure 2: GPT-3 prompt used to create the training examples reported in this paper

Overly broad. Sometimes the LLM fails to detect that the non-relevant passage contains information that means that it is also relevant to the query. As with hallucination, this failure mode leads to inaccurate training data. Unlike hallucination, this type of failure is also found in the MS MARCO training set, where negative examples were not necessarily judged by an assessor.

3.5 Domain and Genre Shift

A key claim of this paper is that building a CLIR training set using the document collection of interest will lead to better retrieval results than just using a generic training collection such as the one underlying MS MARCO. CLIR evaluation collections over genres other than newswire are rare, making this claim challenging to validate empirically. To evaluate JH-POLO performance when the domain or genre does not match that of MS MARCO, we used the HC3 collection [29]. This collection comprises documents, queries, and relevance judgments in Chinese and Persian. The documents are Tweet reply threads of up to 100 Tweets in length. Thus, the documents are short informal conversations – very different from the web documents found in MS MARCO.

When shifting domains or genres, it may be necessary to re-engineer the prompt due to attributes of the data. For instance, we found that when using Tweets as our document collection, underspecificity was particularly egregious. We experimented with prompt variants to ameliorate this problem. We found that adding the sentence "No response should require the recipient to have seen the previous responses" was effective at eliminating many generic noun phrases, which were at the core of most of the underspecificity problems. Figure 3 shows output examples both with and without the additional sentence. Its addition does not eliminate all underspecificity, but it greatly reduces it. The figure illustrates three such queries where "the need for her," "to the country," and "without this" all lead to underspecified queries. This phenomenon is not observed in the queries generated with the sentence. Prompt generation is still a black art; a variety of sentences conveying essentially the same requirement as this sentence did not make an appreciable dent in underspecificity.

4 VALIDATION

We performed two types of validation over the generated data. Section 4.1 describes the manual evaluation undertaken, while Section 4.2 describes an automated validation that improves the quality of the training data.

4.1 Prompt Validation

We manually annotated a small number of system outputs to assess the quality of each prompt. The assessor[12] was provided with the two passages and one of the the resulting queries. Each such example was assigned to one of the following five categories, based on whether the passage labeled relevant was truly relevant, and whether the passage labeled non-relevant was truly not relevant:

- Both assertions were correct.
- The assertion of relevance was incorrect.
- The assertion of non-relevance was incorrect.
- Both assertions were incorrect.
- The generated query was underspecified.

Treating each of these outcomes other than the first as erroneous, the prompt shown in Figure 2 applied to passage pairs selected as described in Section 3.1 had an accuracy of 67% over 61 examples assessed. While underspecified queries are probably not particularly useful training examples, they also are unlikely to damage the training. If the first and last outcomes are treated as correct, accuracy rises to 72%.

4.2 Triple Validation

After generation, we validate the triples with a multilingual cross-encoder reranking model. Validation is a important step to filter out triples that are likely to be hallucinations or that are overly broad. One way to accomplish this validation is to use retrieval to ensure that the positive passage is ranked first, as was done in PROMPTAGATOR [10]. This was necessary in PROMPTAGATOR because negative passages were not included in its prompts. JH-POLO prompts contain both a positive and a negative passage. Therefore, our filtration process relies on the relative rankings of the two passages. In particular, we only include a triple when the positive passage is *more likely* to be relevant to the generated query than is the negative one; this helps to ensure the integrity of the contrastive loss used during training. Furthermore, we use a lower bound threshold on the difference between the two likelihoods to ensure that the two are not too close in meaning with respect to the query.

Specifically, let $F(q, p) : \mathbb{R} \rightarrow \mathbb{R}$ be the cross-encoder model that produces a real-valued score for a given query q and passage p pair. For a given generated query, positive and negative passage triple (q, p_p, p_n), we consider the triple to be valid for training if

$$\frac{e^{F(q,p_p)}}{e^{F(q,p_p)} + e^{F(q,p_n)}} - \frac{e^{F(q,p_n)}}{e^{F(q,p_p)} + e^{F(q,p_n)}} > \tau \qquad (1)$$

[12] Assessors were paper authors using Google passage translations. When there was questionable machine translation, a native speaker reviewed the passage.

Excluding addition	- Taiwanese President Tsai Ing-wen's upcoming visit to Paraguay. - The need for her to pass through the US to demonstrate her presence. - Making fun of Tsai Ing-wen's visit to Paraguay. - Suggestion to give Taiwan several billion in new Taiwan dollars to the country. - Assertion that Taiwanese separatists will not be appeased without this.
Including addition	- Taiwanese President Tsai Ing-wen's planned visit to Paraguay - The possibility of Tsai offering monetary incentives to Paraguay during her visit

Figure 3: Comparison of output when including the prompt addition "No response should require the recipient to have seen the previous responses" (below) or excluding it (above).

where τ must be greater than 0; otherwise, the negative passage is more likely to be relevant to the query than the positive passage. We set τ to 0.15 in our experiments to eliminate noise from the training data.

5 EFFECTIVENESS ANALYSIS

In this section, we explore the effectiveness of JH-POLO-generated triples by training retrieval models on them and comparing the performance of those models over different evaluation datasets. Our purpose here is not to try to match state-of-the-art retrieval effectiveness; doing so is the purview of algorithms, and thus outside of the scope of this paper. We offer no new CLIR algorithms. Rather, we show that JH-POLO-generated training data are as good as machine-translated MS MARCO data for collections that match those data well, and superior to MS MARCO data when the two diverge.

5.1 Evaluation Collections

We analyze the effectiveness of the JH-POLO methodology with two CLIR test suites – TREC NeuCLIR 2022[13] and HC3 [29]. Collection statistics appear in Table 1. These collections form the basis of our effectiveness analysis.

The NeuCLIR 2022 dataset contains three sub-collections in Chinese, Persian, and Russian. Documents in NeuCLIR 1 are news articles extracted from Common Crawl News. The HC3 dataset consists of Chinese and Persian Tweet reply threads each containing a root Tweet and up to 100 replies.

When generating synthetic training data, we draw passages from the target document collection; therefore, passages are in the domain, genre, and language of the test collection. Passage selection for the two collections differed based on the quality of the written language in the passages.

For NeuCLIR 1, a positive passage was chosen randomly from all passages that exceeded a length requirement. Length requirements, which were language-specific, were set to the minimum document lengths imposed by the creators of the NeuCLIR 1 collection: 75 characters for Chinese, 100 characters for Persian, and 200 characters for Russian. To identify a negative passage, the positive passage was used as a query to search the collection, and the resulting passages were ranked using BM25. All BM25 scores were divided by the score of the positive passage. The first passage of sufficient length whose ratio of BM25 scores was less than 0.65 and where no other passages from that document scored higher than 0.65 was selected as the negative passage.

For HC3, the length minimums were reduced to 15 characters for Chinese and 25 for Persian. However, a sample generation revealed that this process was insufficient for selecting Tweets with enough content to generate understandable queries. Consequently, we used 10,200 summaries from the WCEP multi-document summarization dataset [14] as queries to select positive passages (this dataset is time-aligned with HC3, and HC3 topics tended to be event-inspired). Since this summarization dataset is in English, we use Google Translate to translate the summaries into the HC3 languages to use as queries for BM25 retrieval. For Chinese, HC3 contains Tweets written in both Traditional and Simplified characters. To retrieve Tweets in either character set, translations were made into both Traditional and Simplified characters, and the two translations were concatenated to form the queries. Because these events were reported in the English media, not all of them aligned well with topics in non-English Tweets. To provide as much diversity as possible, each relevant passage was uniquely paired with a single non-relevant passage; thus no passage was paired with two different passages. Another observed artifact was that re-Tweets greatly increased the presence of exact duplicate substrings. This made it challenging for an LLM to create queries for which only one of the passages was relevant. We handled this problem by imposing the longest common substring constraint. While the BM25 ratio was still used, we raised the threshold to 0.8 to create more passage pairs. However, because BM25 gives great weight to unusual tokens, URLs present in the Tweets introduced an unusual bias, causing Tweets that were related by advertisements rather than by content to be chosen as pairs. To handle this problem, we stripped URLs from all documents before passages were created. In addition, two passages were paired only if they were both retrieved by the initial retrieval and by the positive passage. This led to the creation of fewer than 10,200 pairs. Finally, the "same document" exclusion criterion used for the NeuCLIR 1 was dropped since Twitter conversations are less coherent than Common Crawl News documents.

5.2 Training Examples Generated

Table 2 summarizes the number of triples generated by GPT-3 Davinci-3. We generated roughly the same number of triples for all three sub-collections in NeuCLIR 1, with Russian pairs having slightly more triples. Despite the prompt asking for five topics for each passage of the pair, GPT-3 would not necessarily respond

[13]The name of document collection is NeuCLIR 1. NeuCLIR 2022 refers to the evaluation suite that contains NeuCLIR 1 and the topics and relevance judgments developed for the TREC NeuCLIR Track 2022.

with the correct number, and not all generated topics would pass the filter, resulting in roughly eight queries per pair. Generation for HC3 is even more challenging, with a fanout of around seven queries per passage pair.

While enforcing unique passage pairs may seem desirable, this repetition of passage pairs is similar to the repetition of query and positive passage that is found in MS MARCO training triples. In fact, our generation process actually has less repetition than MS MARCO. In the small training triple file published by MS MARCO, there are roughly 100 negative passages associated with each query, where the vast majority of the queries have only one positive passage. Repetition of query-positive passage pairs is more than ten times that found in JH-POLO. We argue that JH-POLO provides more diverse information in its triples and thus has the potential to lead to better retrieval models.

5.3 Retrieval Models for Effectiveness Analysis

We used two neural dense retrieval architectures as representatives for analyzing our methodology: DPR-X [54, 59] and ColBERT-X [35]. All models for each retrieval architecture are based on XLM-RoBERTa-base [9] started from the same checkpoint and fine-tuned with a retrieval objective using English MS MARCO for 200,000 update steps. We vary the source of the training data in the second stage fine-tuning, which consists of 1,245 steps. This training scheme is designed to expose differences introduced by a small amount of training data, rather than to train state-of-the-art systems. Note that this training scheme does not include any advanced tricks such as iterative hard-negative mining [17, 53], in-batch negative sampling [24, 58], knowledge distillation [18], etc. The training process here is a for demonstrating the relative effectiveness of JH-POLO as a training resource compared to MS MARCO.

JH-POLO training triples were generated with the passage selection processes for each evaluation collection outlined in Section 5.1. GPT-3 Davinci-3 is prompted for queries using an English description along with a pair of passages from the collection. The generated queries, along with the passages, are passed through a cross-encoder trained on mMARCO [5] [14] for validation, as described in Section 4.2.

To analyze the effectiveness of JH-POLO, we fine-tune the model in the second stage with the following regimens for comparison:

- **English** (*Eng.*). Continues fine-tuning the model with English MS MARCO v1. In this scenario, the model gains knowledge about non-English language during the initial training of the mPLM, but not during fine-tuning.
- **Translate** (*Trans.*). Fine-tuned with MS MARCO v1 documents that have been machine-translated into the language of the target document collection.[15] Queries remain in English, so the model is exposed to the CLIR task during continued fine-tuning, but the documents may contain translationese introduced by the machine translation system. This approach is also known as *translate-train* [35].

[14] https://huggingface.co/cross-encoder/mmarco-mMiniLMv2-L12-H384-v1

[15] We used the NeuMARCO translation provided by the TREC NeuCLIR Track 2022. https://ir-datasets.com/neumarco.html

Table 1: Dataset statistics of NeuCLIR 2022 and HC3.

Collection	Chinese		Persian		Russian	
Set	# Qry	# Docs	# Qry	# Docs	# Qry	# Docs
NeuCLIR	47	3,179,209	45	2,232,016	44	4,627,543
HC3	50	5,584,146	50	7,335,221	-	-

Table 2: Statistics of the generation results.

	Passage Pairs	Generated Triples	Valid Triples	Triples Per Pair
NeuCLIR				
Chinese	19,401	187,908	154,046	7.94
Persian	19,432	180,174	153,933	7.92
Russian	19,348	185,941	159,412	8.24
HC3				
Chinese	9,766	86,532	68,679	7.03
Persian	10,077	88,957	66,535	6.60

Following prior work in CLIR dense retrieval [35, 36], we used the trained models to index the collections by separating the documents into overlapping passages of 180 tokens with a stride of 90. Since both NeuCLIR and HC3 consist of TREC-style topics, we concatenated the titles and descriptions as our search query; these are the same queries used in the official NeuCLIR baseline runs for the reranking subtask. We evaluate the final retrieval effectiveness using nDCG@20 (the primary evaluation metric in TREC NeuCLIR) and Recall at 100 (R@100).

5.4 Effectiveness on News Documents

As presented in Table 3, both ColBERT-X and DPR-X benefit more from further fine-tuning with JH-POLO than with more of the original English MS MARCO for both nDCG@20 and Recall@100. Since JH-POLO is naturally cross-language, a model trained on it could learn the definition of relevance directly from the target language pair. However, English MS MARCO can only provide evidence on the relationship between queries and passages; it cannot inform the CLIR system being trained about the target language. This forces the model to rely solely on the multilinguality of the pretrained language model, resulting in worse retrieval performance than if the training data encapsulated that information.

By translating the MS MARCO passages to the target language, a model being trained can learn the cross-language relationships, although the resulting passages will suffer from translationese. As repeatedly observed by prior work [35, 36, 54], this translate-train approach provides state-of-the-art CLIR effectiveness when training only on MS MARCO but is dependent on the translation quality [35]. When evaluating the models on NeuCLIR 2022, whose documents are similar to MS MARCO passages, models trained with JH-POLO are only slightly worse than their translate-train counterparts. These differences are not statistically significant, indicating that the two approaches are similar and neither consistently outperforms the other on all topics. When evaluating on HC3, which is a very different genre compared to MS MARCO, training with

Table 3: Retrieval Effectiveness. *indicates significance with 95% confidence against fine-tuning with English triples using paired t-tests with Bonferrini correction on three tests (over languages). †indicates significance between JH-POLO and fine-tuning with translated triples using the same statistical test.

	NeuCLIR 2022								HC3					
	nDCG@20				R@100				nDCG@20			R@100		
Triples	Chinese	Persian	Russian	Avg.	Chinese	Persian	Russian	Avg.	Chinese	Persian	Avg.	Chinese	Persian	Avg.
ColBERT-X														
Eng.	0.155	0.131	0.227	0.171	0.236	0.290	0.290	0.272	0.198	0.196	0.197	0.361	0.368	0.364
Trans.	**0.216***	0.220*	**0.267***	**0.234**	**0.320***	**0.389***	**0.325***	**0.345**	0.208	0.254*	0.231	0.385	0.400	0.393
JH-POLO	0.211	**0.223***	0.241	0.225	0.265	0.372	0.322	0.320	**0.236**	**0.270***	**0.253**	**0.442***	**0.419**	**0.430**
DPR-X														
Eng.	0.139	0.088	0.175	0.134	0.224	0.245	0.235	0.235	0.130	0.115	0.123	0.249	0.254	0.251
Trans.	0.191*	**0.155***	**0.192**	**0.179**	0.280	0.317*	**0.278**	0.292	0.177	0.177	0.177	0.322	0.349	0.335
JH-POLO	**0.192***	0.132*	0.181	0.168	**0.294***	**0.343***	0.277	**0.305**	**0.240***	**0.269***†	**0.255**	**0.500***†	**0.483***†	**0.491**

JH-POLO significantly outperforms translate; we will discuss this outcome in detail in the next section.

Comparing the two retrieval models, DPR-X benefits from JH-POLO more than ColBERT-X does, especially in the bottom part of the ranking (measured by recall). Since DPR-X summarizes each query and passage into a single vector, it must rely on general semantics, not on token-level matching. Therefore, training with JH-POLO, which contains queries that are only relevant to part of the positive passage and do not necessarily have overlapping tokens, improves DPR-X's ability to understand subtle differences between the passages. In contrast, ColBERT-X focuses on more token-level cross-language alignments through translated passages, directly enhancing its token-level matching.

However, triple quality is an artifact of the prompt used to generate it. We value the diversity and the rich queries that our prompt can provide by generating topics instead of keywords. This tendency implicitly benefits DPR-X more than ColBERT-X. If one is only considering training a specific type of retrieval model, the prompt can be adjusted to produce the kind of information the model most needs to optimize its effectiveness.

Again, we do not claim to reach the state-of-the-art CLIR effectiveness simply by training on JH-POLO; such performance would require numerous optimizations, such as using XLM-Roberta large rather than XLM-Roberta base, fine-tuning for many steps beyond the reported two-stage regimen, using in-batch negatives, generating perhaps orders of magnitude more training examples, and so on. But what we do see here is that on a collection that is similar to the MS MARCO collection, JH-POLO generates training data that is on par with machine-translated MS MARCO data.

5.5 Effectiveness on Tweets

When building a CLIR engine to search text that differs from the web articles that make up MS MARCO, training on JH-POLO provides dramatic improvements over MS MARCO. When training with JH-POLO-generated triples on HC3, both nDCG@20 and Recall@100 outperform translate-training with MS MARCO. While translating the MS MARCO passages into the target language helps the retrieval model cross the language barrier, the gap between the training genre and the HC3 passages is still large. JH-POLO fills this gap by directly exposing the model to Tweets during retrieval fine-tuning. Such exposure directly translates to effectiveness improvements across all regions of the ranked list.

Interestingly, DPR-X is on par with, and sometimes better than, ColBERT-X when trained with JH-POLO. This is unusual, as ColBERT-X generally outperforms DPR-X [54]. We hypothesize that ColBERT-X requires more training data to learn how to match in a new genre than does DPR-X; while ColBERT-X must adjust all term matches, DPR-X only needs to adjust how its CLS token is created. In this case, DPR-X is more efficient at absorbing the cross-language and cross-genre knowledge provided by JH-POLO. Therefore, we argue that the smaller improvement in ColBERT-X when training on JH-POLO is not necessarily the result of ineffective JH-POLO triples, but of the nature of the retrieval model when searching across genre. Nevertheless, JH-POLO numerically improves ColBERT-X's performance on HC3, although the difference is not statistically significant.

Of particular note is the JH-POLO performance in Persian, where in three of the four collection-retrieval system pairs JH-POLO outperforms translate, one of which is a statistically significant difference. Given that Persian is a lower resources language, machine translation does not perform as well in general [35]. This indicates the using a high performing generative LLM may be able to provide better training data than machine translation.

5.6 Analysis of Examples

Figure 4 presents two passage pairs and the queries that GPT-3 Davinci-3 generated from them. Each passage pair is connected by an identifiable thread (*eruptions* in the top of the figure; *voting* in the bottom). Because of the way they were selected, these passages tend to contain more information than a randomly selected Tweet. Many of the topics for the eruptions are similar, but are specific to the eruption mentioned in the positive passage. We do see that occasionally there is a query for which there is no further information in the passage, such as the location of Popocatépetl.

Because the bottom passages are less formal, the queries are more general. In particular, the bottom Passage A produces only

Passage A: 1月25日,位于菲律宾阿尔拜省的马荣火山喷出火山灰。马荣火山位于菲律宾吕宋岛东南部的阿尔拜省，距菲首都马尼拉约330公里，海拔约2400米,是菲境内最活跃的火山之一。截至24日，已有超过7万人被疏散出马荣火山附近的危险区域。 *Volcanic ash spews from Mayon Volcano in Albay, Philippines, on January 25. Mayon Volcano is located in Albay Province in the southeast of Luzon Island in the Philippines, locating about 330 kilometers away from Manila, the capital of the Philippines, and is about 2,400 meters above sea level. It is one of the most active volcanoes in the Philippines. As of the 24th, more than 70,000 people have been evacuated from the danger zone near Mayon Volcano.*	**Passage B:** 当地时间13日早上5点43分，位于墨西哥市近郊的#波波卡特佩特火山突然喷发，喷出大量火山灰，随即冒出白炽岩。波波卡特佩特火山高达5426米，是墨西哥第二高峰,也是世界上最活跃的火山之一。 *At 5:43 a.m. local time on the 13th, the #Popocatépetl volcano located in the outskirts of Mexico City suddenly erupted, spewing out a large amount of volcanic ash, followed by incandescent rocks. With a height of 5,426 meters, Popocatepetl is the second highest mountain in Mexico and one of the most active volcanoes in the world.*
Passage A being relevant: ✓ The location of Mt. Mayon in the Philippines ✓ The evacuation of more than 70,000 people from the area near Mt. Mayon ✓ The elevation of Mt. Mayon ✓ The activity of Mt. Mayon compared to other volcanoes in the Philippines ✓ The date of the eruption of Mt. Mayon	**Passage B being relevant:** ✓ The location of Popocatépetl volcano near Mexico City ✓ The elevation of Popocatépetl ✓ The activity of Popocatépetl compared to other volcanoes in Mexico ✓ The date of the eruption of Popocatépetl ✓ The composition of the materials emitted by Popocatépetl during the eruption
Passage A: 欢迎和我一起在美国大选中 #成为选民。马上寻找你的投票站，看看你可以为谁投票。我推举我们敬爱的习近平总书记！希望你也能投他一票 *Join me in #becoming a voter in the US election. Find your polling place now and see who you can vote for. I recommend our beloved General Secretary Xi Jinping! I hope you can vote for him too*	**Passage B:** 美国大选日定在礼拜二，是因为当时美国人多为新教徒农民，周日去过教堂，周一动身出发，周二到达投票站投票。由此，大选日历来是亲朋好友难得聚会的好日子，大家会带上面粉、白菜和肉馅，到投票站一起包饺子，包好了一边吃一边投。而饺子不咬开就不知道什么馅，也寓意... *The U.S. election day is set on Tuesday because Americans were mostly Protestant farmers at that time, they went to church on Sunday, leave by Monday, and arrived at polling stations by Tuesday to vote. Therefore, the general election calendar has always been a rare good day for relatives and friends to gather. Everyone will bring flour, cabbage and minced meat to the polling station to make dumplings together, and vote while eating. And you don't know the kind of stuffing of the dumplings until you take a bite, which also means...*
Passage A being relevant: ✓ Endorsement of Xi Jinping ✓ Inviting others to join in voting for a particular candidate ✗ The importance of voting in the US election ✗ The importance of collective voting and participation ✗ The necessity of actively seeking out one's local voting station	**Passage B being relevant:** ✓ A detailed history of the US voting system ✓ Chinese-American cultural customs ✓ How US citizens of different religions view election day ✓ Traditional Chinese foods associated with the US election ✓ The importance of family and friends gathering on election day

Figure 4: Sample queries generated by JH-POLO. Text in italics is the translation of the corresponding Chinese Tweet. ✓ and ✗ indicate whether the generated query passed the cross-encoder filter.

two queries; the other three were filtered out during the validation step and are marked with an ✗. The remaining queries are supported by the passage. In the queries for Passage B, the first one clearly identifies a topic in the passage; however, the third query concerning religions is not well supported, as the passage does not explain the viewpoints of Protestant farmers. While one might question the connection between traditional Chinese food and US elections, the passage does include information on that, and the LLM captures it well.

6 COST

GPT-3 is not free; the cost of producing a CLIR training collection using JH-POLO depends on the size of the collection and the cost of GPT-3 per request. At this writing, GPT-3 Davinci-3 (the most capable model) costs us US$0.02 per 1000 subword tokens (the sum of the number of tokens in the prompt and in the output). Subwords are produced by the GPT-2 tokenizer,[16] which is similar to SentencePiece.[17] Thus, our training corpus built on the Chinese NeuCLIR collection cost us about US$400 to produce, while Persian and Russian cost about 20% more. GPT-3 throughput has been about two prompts per second with ten concurrent processes, allowing us to create the collections in about 16 hours.

The cost per 1000 training examples for the NeuCLIR collections averaged US$3 for the prompt shown in Figure 2. The number of requests is the same as the number of passage pairs shown in Table 2. While the cost does add up, it is orders of magnitudes cheaper than producing a dataset annotated by humans, such as MS MARCO.

7 CONCLUSIONS

This paper introduces the JH-POLO CLIR training set creation methodology, which selects a positive and negative passage from the target document collection and uses a generative large language model to synthetically generate one or more queries per passage

[16] https://beta.openai.com/tokenizer

[17] https://github.com/google/sentencepiece

pair. The methodology suggests that random selection of positive passages works well for high quality texts, and shows how to select passages with meaningful content for noisier texts. It allows negative examples to be selected before the query is generated, thus providing some control over the quality and difficulty of the training collection. It demonstrates effective prompts that describe the desired output without the need for exemplars. We find that GPT-3 Davinci-3 can generate sufficiently good queries so that the resulting training triples can train a CLIR retrieval model as effectively as MS MARCO. In addition, the further the genre of the document collection is from Bing web passages, the more effective the synthetically generated data is. Thus, JH-POLO offers a pathway to automatically creating an effective CLIR training set for any text corpus of interest.

REFERENCES

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation Artifacts in Cross-lingual Transfer Learning. *CoRR* abs/2004.04721 (2020). arXiv:2004.04721 https://arxiv.org/abs/2004.04721

[2] Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual Open-Retrieval Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 547–564. https://doi.org/10.18653/v1/2021.naacl-main.46

[3] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv e-prints*, Article arXiv:1611.09268 (Nov. 2016). https://doi.org/10.48550/arXiv.1611.09268 arXiv:1611.09268

[4] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Data Augmentation for Information Retrieval using Large Language Models. *arXiv e-prints*, Article arXiv:2202.05144 (Feb. 2022). https://doi.org/10.48550/arXiv.2202.05144

[5] Luiz Henrique Bonifacio, Israel Campiotti, Roberto de Alencar Lotufo, and Rodrigo Nogueira. 2021. mMARCO: A Multilingual Version of MS MARCO Passage Ranking Dataset. *CoRR* abs/2108.13897 (2021). arXiv:2108.13897 https://arxiv.org/abs/2108.13897

[6] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, Baltimore, Maryland, 2206–2240. https://proceedings.mlr.press/v162/borgeaud22a.html

[7] Leonid Boytsov, Preksha Patel, Vivek Sourabh, Riddhi Nisar, Sayani Kundu, Ramya Ramanathan, and Eric Nyberg. 2023. InPars-Light: Cost-Effective Unsupervised Training of Efficient Rankers. *arXiv e-prints*, Article arXiv:2301.02998 (Jan. 2023). https://doi.org/10.48550/arXiv.2301.02998

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[9] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR* abs/1911.02116 (2019). arXiv:1911.02116 http://arxiv.org/abs/1911.02116

[10] Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot Dense Retrieval From 8 Examples. *arXiv e-prints*, Article arXiv:2209.11755 (Sept. 2022). https://doi.org/10.48550/arXiv.2209.11755 arXiv:2209.11755

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[12] Petra Galuščáková, Douglas W. Oard, and Suraj Nair. 2021. Cross-language Information Retrieval. *arXiv e-prints*, Article arXiv:2111.05988 (Nov. 2021). https://doi.org/10.48550/arXiv.2111.05988 arXiv:2111.05988

[13] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise Zero-Shot Dense Retrieval without Relevance Labels. *arXiv e-prints*, Article arXiv:2212.10496 (Dec. 2022). https://doi.org/10.48550/arXiv.2212.10496

[14] Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1302–1308. https://doi.org/10.18653/v1/2020.acl-main.120

[15] Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, Rerank, Generate. *arXiv e-prints*, Article arXiv:2207.06300 (July 2022). https://doi.org/10.48550/arXiv.2207.06300 arXiv:2207.06300

[16] Parth Gupta, Rafael E Banchs, and Paolo Rosso. 2017. Continuous space models for CLIR. *Information Processing & Management* 53, 2 (2017), 359–370.

[17] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*. PMLR, 3929–3938.

[18] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation. *arXiv e-prints*, Article arXiv:2010.02666 (Oct. 2020). https://doi.org/10.48550/arXiv.2010.02666 arXiv:2010.02666

[19] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. *arXiv e-prints*, Article arXiv:2212.09689 (Dec. 2022). https://doi.org/10.48550/arXiv.2212.09689 arXiv:2212.09689

[20] Gautier Izacard and Edouard Grave. 2020. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. *arXiv e-prints*, Article arXiv:2007.01282 (July 2020). https://doi.org/10.48550/arXiv.2007.01282 arXiv:2007.01282

[21] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *arXiv e-prints*, Article arXiv:2208.03299 (Aug. 2022). https://doi.org/10.48550/arXiv.2208.03299 arXiv:2208.03299

[22] Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. InPars-v2: Large Language Models as Efficient Dataset Generators for Information Retrieval. *arXiv e-prints*, Article arXiv:2301.01820 (Jan. 2023). https://doi.org/10.48550/arXiv.2301.01820 arXiv:2301.01820

[23] Zhengbao Jiang, Luyu Gao, Jun Araki, Haibo Ding, Zhiruo Wang, Jamie Callan, and Graham Neubig. 2022. Retrieval as Attention: End-to-end Learning of Retrieval and Reading within a Single Transformer. *arXiv e-prints*, Article arXiv:2212.02027 (Dec. 2022). https://doi.org/10.48550/arXiv.2212.02027 arXiv:2212.02027

[24] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[25] Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2022. RealTime QA: What's the Answer Right Now? *arXiv e-prints*, Article arXiv:2207.13332 (July 2022). https://doi.org/10.48550/arXiv.2207.13332 arXiv:2207.13332

[26] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3397271.3401075

[27] Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2023. Overview of the TREC 2022 NeuCLIR Track. In *31st Text REtrieval Conference* (Gaithersburg, Maryland).

[28] Dawn Lawrie, James Mayfield, Douglas W Oard, and Eugene Yang. 2022. HC4: a new suite of test collections for ad hoc CLIR. In *European Conference on Information Retrieval*. Springer, 351–366.

[29] Dawn Lawrie, James Mayfield, Douglas W. Oard, Eugene Yang, Suraj Nair, and Petra Galuščáková. 2023. HC3: A Suite of Test Collections for CLIR Evaluation

over Informal Text. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA.

[30] Haejun Lee, Akhil Kedia, Jongwon Lee, Ashwin Paranjape, Christopher D. Manning, and Kyoung-Gu Woo. 2021. You Only Need One Model for Open-domain Question Answering. *arXiv e-prints*, Article arXiv:2112.07381 (Dec. 2021). https://doi.org/10.48550/arXiv.2112.07381 arXiv:2112.07381

[31] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[32] Bo Li and Ping Cheng. 2018. Learning Neural Representation for CLIR with Adversarial Framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1861–1870. https://doi.org/10.18653/v1/D18-1212

[33] Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, and Avirup Sil. 2021. Learning Cross-Lingual IR from an English Retriever. *arXiv e-prints*, Article arXiv:2112.08185 (Dec. 2021). https://doi.org/10.48550/arXiv.2112.08185 arXiv:2112.08185

[34] Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2021. On Cross-Lingual Retrieval with Multilingual Text Encoders. *arXiv e-prints*, Article arXiv:2112.11031 (Dec. 2021). https://doi.org/10.48550/arXiv.2112.11031 arXiv:2112.11031

[35] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W. Oard. 2022. Transfer Learning Approaches for Building Cross-Language Dense Retrieval Models. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I* (Stavanger, Norway). Springer-Verlag, Berlin, Heidelberg, 382–396. https://doi.org/10.1007/978-3-030-99736-6_26

[36] Suraj Nair, Eugene Yang, Dawn Lawrie, James Mayfield, and Douglas W Oard. 2022. Learning a Sparse Representation Model for Neural CLIR. *Proceedings of Design of Experimental Search & Information REtrieval Systems (DESIRES)* (2022).

[37] Jian-Yun Nie. 2010. Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies* 3, 1 (2010), 1–125.

[38] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv e-prints*, Article arXiv:1901.04085 (Jan. 2019). https://doi.org/10.48550/arXiv.1901.04085 arXiv:1901.04085

[39] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. *arXiv e-prints*, Article arXiv:1904.08375 (April 2019). https://doi.org/10.48550/arXiv.1904.08375 arXiv:1904.08375

[40] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2020. KILT: a Benchmark for Knowledge Intensive Language Tasks. *arXiv e-prints*, Article arXiv:2009.02252 (Sept. 2020). https://doi.org/10.48550/arXiv.2009.02252 arXiv:2009.02252

[41] Andy Rosenbaum, Saleh Soltan, Wael Hamza, Amir Saffari, Marco Damonte, and Isabel Groves. 2022. CLASP: Few-Shot Cross-Lingual Data Augmentation for Semantic Parsing. https://doi.org/10.48550/arXiv.2210.07074 arXiv:2210.07074

[42] Timo Schick and Hinrich Schütze. 2020. Few-Shot Text Generation with Pattern-Exploiting Training. *arXiv e-prints*, Article arXiv:2012.11926 (Dec. 2020). https://doi.org/10.48550/arXiv.2012.11926 arXiv:2012.11926

[43] Timo Schick and Hinrich Schütze. 2021. Generating Datasets with Pretrained Language Models. *arXiv e-prints*, Article arXiv:2104.07540 (April 2021). https://doi.org/10.48550/arXiv.2104.07540 arXiv:2104.07540

[44] Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2021. Cross-Lingual Training of Dense Retrievers for Document Retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 251–253. https://doi.org/10.18653/v1/2021.mrl-1.24

[45] Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2020. Multi-Modal Open-Domain Dialogue. *arXiv e-prints*, Article arXiv:2010.01082 (Oct. 2020). https://doi.org/10.48550/arXiv.2010.01082 arXiv:2010.01082

[46] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv e-prints*, Article arXiv:2208.03188 (Aug. 2022). https://doi.org/10.48550/arXiv.2208.03188 arXiv:2208.03188

[47] Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting GPT-3 To Be Reliable. *arXiv e-prints*, Article arXiv:2210.09150 (Oct. 2022). https://doi.org/10.48550/arXiv.2210.09150 arXiv:2210.09150

[48] Shuo Sun and Kevin Duh. 2020. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for Cross-Lingual Information Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 4160–4170. https://doi.org/10.18653/v1/2020.emnlp-main.340

[49] James Thorne and Andreas Vlachos. 2018. Automated Fact Checking: Task formulations, methods and future directions. *arXiv e-prints*, Article arXiv:1806.07687 (June 2018), arXiv:1806.07687 pages. https://doi.org/10.48550/arXiv.1806.07687 arXiv:1806.07687

[50] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. https://doi.org/10.48550/arXiv.1803.05355 arXiv:1803.05355

[51] Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities* 30, 1 (2015), 98–118.

[52] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2021. GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. *arXiv preprint arXiv:2112.07577* (2021).

[53] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *arXiv e-prints*, Article arXiv:2007.00808 (July 2020). https://doi.org/10.48550/arXiv.2007.00808 arXiv:2007.00808

[54] Eugene Yang, Suraj Nair, Ramraj Chandradevan, Rebecca Iglesias-Flores, and Douglas W. Oard. 2022. C3: Continued Pretraining with Contrastive Weak Supervision for Cross Language Ad-Hoc Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2507–2512. https://doi.org/10.1145/3477495.3531886

[55] Eugene Yang, Suraj Nair, Dawn Lawrie, James Mayfield, and Douglas W. Oard. 2022. Parameter-efficient Zero-shot Transfer for Cross-Language Dense Retrieval with Adapters. *arXiv e-prints*, Article arXiv:2212.10448 (Dec. 2022). https://doi.org/10.48550/arXiv.2212.10448 arXiv:2212.10448

[56] Puxuan Yu and James Allan. 2020. A Study of Neural Matching Models for Cross-Lingual IR. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1637–1640. https://doi.org/10.1145/3397271.3401322

[57] Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than Retrieve: Large Language Models are Strong Context Generators. *arXiv e-prints*, Article arXiv:2209.10063 (Sept. 2022). https://doi.org/10.48550/arXiv.2209.10063 arXiv:2209.10063

[58] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. RepBERT: Contextualized Text Embeddings for First-Stage Retrieval. *arXiv e-prints*, Article arXiv:2006.15498 (June 2020). https://doi.org/10.48550/arXiv.2006.15498 arXiv:2006.15498

[59] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval. *arXiv e-prints*, Article arXiv:2108.08787 (Aug. 2021). https://doi.org/10.48550/arXiv.2108.08787 arXiv:2108.08787

[60] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages. *arXiv e-prints*, Article arXiv:2210.09984 (Oct. 2022). https://doi.org/10.48550/arXiv.2210.09984 arXiv:2210.09984

[61] Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. 2012. Translation techniques in cross-language information retrieval. *ACM Computing Surveys (CSUR)* 45, 1 (2012), 1–44.