CLIRMatrix: A massively large collection of bilingual and multilingual datasets for Cross-Lingual Information Retrieval

Shuo Sun
Johns Hopkins University
ssun32@jhu.edu

Kevin Duh
Johns Hopkins University
kevinduh@cs.jhu.edu

Abstract

We present *CLIRMatrix*, a massively large collection of bilingual and multilingual datasets for Cross-Lingual Information Retrieval extracted automatically from Wikipedia. CLIRMatrix comprises (1) *BI-139*, a bilingual dataset of queries in one language matched with relevant documents in another language for 139×138=19,182 language pairs, and (2) *MULTI-8*, a multilingual dataset of queries and documents jointly aligned in 8 different languages. In total, we mined 49 million unique queries and 34 billion (query, document, label) triplets, making it the largest and most comprehensive CLIR dataset to date. This collection is intended to support research in end-to-end neural information retrieval and is publicly available at `https://github.com/ssun32/CLIRMatrix`. We provide baseline neural model results on BI-139, and evaluate MULTI-8 in both single-language retrieval and mix-language retrieval settings.

1 Introduction

Cross-Lingual Information Retrieval (CLIR) is a retrieval task in which search queries and candidate documents are written in different languages. CLIR can be very useful in some scenarios. For example, a reporter may want to search foreign-language news to obtain different perspectives for her story; an inventor may explore the patents in another country to understand prior art. Traditionally, translation-based approaches are commonly used to tackle the CLIR task (Zhou et al., 2012; Oard, 1998; McCarley, 1999): the *query translation* approach translates the query into the same language of the documents, whereas the *document translation* approach translates the document into the same language as the query. Both approaches rely on a machine translation (MT) system or bilingual dictionary to map queries and documents to the same language, then employ a monolingual information retrieval (IR) engine to find relevant documents.

Recently, the research community has been actively looking at end-to-end solutions that tackle the CLIR task without the need to build MT systems. This line of work builds upon recent advances in Neural Information Retrieval in the monolingual setting, c.f. (Mitra and Craswell, 2018; Craswell et al., 2020). There are proposals to directly train end-to-end neural retrieval models on CLIR datasets (Sasaki et al., 2018; Zhang et al., 2019) or MT bitext (Zbib et al., 2019; Jiang et al., 2020). One can also exploit cross-lingual word embeddings to train a CLIR model on disjoint monolingual corpora (Litschko et al., 2018).

Despite the growing interest in end-to-end CLIR, the lack of a large-scale, easily-accessible CLIR dataset covering many language directions in high-, mid- and low-resource settings has detrimentally affected the CLIR community's capability to replicate and compare with previously published work. For example, among the widely-used datasets, the CLEF collection (Ferro and Silvello, 2015) covers many languages but is not large enough for training neural models. The more recent IARPA MATERIAL/OpenCLIR collection (Zavorin et al., 2020), is not yet publicly accessible. This motivates us to design and build *CLIRMatrix*, a massively large collection of bilingual and multilingual datasets for CLIR.

We construct *CLIRMatrix* from Wikipedia in an automated manner, exploiting its large variety of languages and massive number of documents. The core idea is to **synthesize relevance labels via an existing monolingual IR system, then propagate the labels via Wikidata links** that connect documents in different languages. In total, we were able to mine 49 million unique queries in 139 languages and 34 billion (query, document, label)

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4160–4170,
November 16–20, 2020.

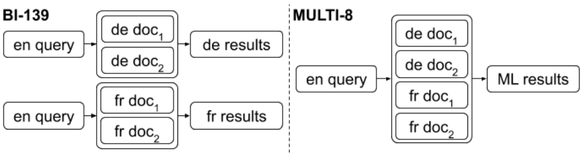

Figure 1: Illustration of our CLIRMatrix collection. The BI-139 portion of CLIRMatrix supports research in bilingual retrieval and covers a matrix of 139 × 138 language pairs. The MULTI-8 portion of CLIRMatrix supports research in multilingual modeling and mixed-language (ML) retrieval, where queries and documents are jointly aligned over 8 languages.

triplets, creating a CLIR collection across a matrix of $139 \times 138 = 19,182$ language pairs. From this raw collection, we introduce two datasets:

- **BI-139** is a massively large bilingual CLIR dataset that covers $139 \times 138 = 19,182$ language pairs. To encourage reproducibility, we present standard train, validation, and test subsets for every language direction.

- **MULTI-8** is a *multilingual* CLIR dataset comprising of queries and documents jointly aligned 8 languages: Arabic (ar), German (de), English (en), Spanish (es), French (fr), Japanese (ja), Russian (ru), Chinese (zh). Each query will have relevant documents in the other 7 languages.

See Figure 1 for a comparison of BI-139 and MULTI-8. The former facilitates the evaluation of bilingual retrieval over a wide variety of languages, while the latter supports research in mixed-language retrieval (a.k.a multilingual retrieval (Savoy and Braschler, 2019)), which is an interesting yet relatively under-explored problem. For both, the train sets are large enough to enable the training of the neural IR models.

We hope *CLIRMatrix* is useful and can empower further developments in this field of research. To summarize, our contributions are:

1. A massive CLIR collection supporting both training and evaluation of bilingual/multilingual models.

2. A set of baseline neural results on BI-139 and MULTI-8. On MULTI-8, we show that a single multilingual model can significantly outperform an ensemble of bilingual models.

CLIRMatrix is publicly available at `https://github.com/ssun32/CLIRMatrix`.

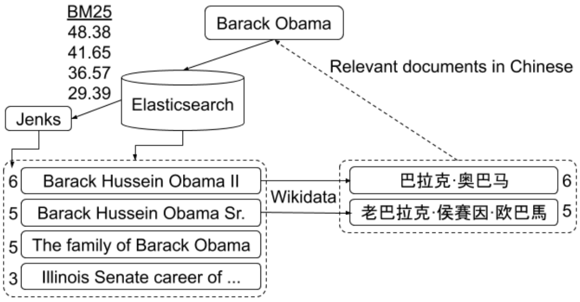

Figure 2: Intuition of CLIR relevance label synthesis. For the English query "Barack Obama", first a monolingual IR engine (Elasticsearch) labels documents in English; then Wikidata links are exploited to propagate the label to the corresponding Chinese documents, which are assumed to be topically similar.

2 Methodology

Let q^X be a query in language X, and d^Y be a document in language Y. A bilingual CLIR dataset consists of I triples

$$\{(q_i^X, d_{ij}^Y, r_{ij})\}_{i=1,2,\ldots,I} \tag{1}$$

where d_{ij}^Y is the j-th document associated with query q_i^X, and r_{ij} is a label saying how relevant is the document d_{ij}^Y to the query q_i^X. Conventionally, r_{ij} is an integer with 0 representing "not relevant" and higher values indicating more relevant.

Suppose there are J documents in total. In the full collection search setup, the index j ranges from $1, \ldots, J$, meaning that each query q_i^X searches over the full set of documents $\{d_{ij}^Y\}_{j=1,\ldots,J}$. In the re-ranking setup, each query q_i^X searches over a subset of documents obtained by an initial full-collection retrieval engine: $\{d_{ij}^Y\}_{j=1,\ldots,K_i}$, where $K_i \ll J$. For practical reasons, machine learning approaches to IR focus on the re-ranking setup with K_i set to 10∼1000 (Liu, 2009; Chapelle and Chang, 2011). We follow the re-ranking setup here.

We now describe the main intuition of our construction method and detail various components and design choices in our pipeline.

2.1 Intuition and Assumptions

To create a CLIR dataset, one needs to decide how to obtain q_i^X and d_{ij}^Y, and r_{ij}. We set q_i^X to be Wikipedia titles, d_{ij}^Y to be Wikipedia articles, and synthesize r_{ij} automatically using a simple yet reliable method. We argue that *Wikipedia* is the best available resource for building CLIR datasets due to two reasons: First, it is freely available and contains articles in more than 300 languages, covering a large variety of topics. Second, Wikipedia articles are mapped to entities in *Wikidata*[1], which is a relatively reliable way to find the same articles written in other languages.

To synthesize relevance labels r_{ij}, we propose first to generate labels using an existing monolingual IR system in language X, then propagate the labels via Wikidata links to language Y. In other words, we assume:

1. the availability of documents d^X in the *same* language as the query, and

2. the feasibility of an existing monolingual IR system in language X to provide labels $\hat{r}_{ij}$ on (q_i^X, d_{ij}^X) pairs

Then for any d_{ij}^Y that links to d_{ij}^X, we assign the relevance label $\hat{r}_{ij}$.

This intuition is illustrated in Figure 2. Suppose we wish to find Chinese documents that are relevant for the English query "Barack Obama". We first run monolingual IR to find English documents that answer the query. In this figure, 4 documents are returned, and we attempt to link to the corresponding Chinese versions using Wikidata information. When the link is available, we set the relevance label r_{ij} for Chinese documents using the English-based IR system's predictions $\hat{r}_{ij}$; all other documents are deemed not relevant. This gives us the triplet $(q_i^X, d_{ij}^Y, r_{ij})$.

2.2 Mining Pipeline

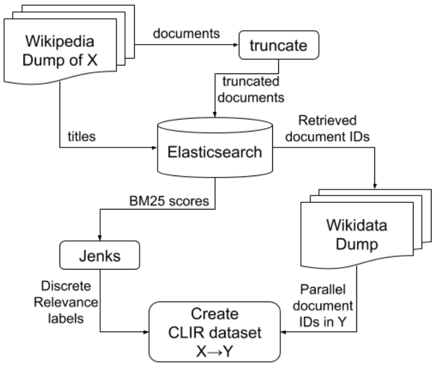

Figure 3: Mining pipeline for constructing a bilingual CLIR dataset with queries in language X and documents in language Y.

Figure 3 is our mining pipeline that implements the intuition in Figure 2. First, we download the Wikipedia dump of language X and then extract the titles and document bodies of every article. We index the documents into an Elasticsearch[2] search engine, which serves as our monolingual IR system. Using the extracted titles as search queries, we retrieve the top 100 relevant documents and their corresponding BM25 scores from Elasticsearch for every query. We then convert the BM25 scores into discrete relevance judgment labels using Jenks natural break optimization. Finally, we propagate these labels to documents in language Y that are linked via Wikidata.

We downloaded Wikidata and Wikipedia dumps released on January 1, 2020. Since Wikipedia dumps contain tremendous amounts of meta-information such as URLs and scripts, it can be expensive to extract actual text directly from those dumps. Inspired by Schwenk et al. (2019), we extracted document ids, titles, and bodies from Wikipedia's search indices[3] instead, which contain raw text data without meta-information.

Wikipedia dumps We discarded dumps with less than ten thousand documents, which are usually the dumps of Wikipedia of certain dialects and less commonly used languages. We are left with Wikipedia dumps in 139 languages, containing a good mix of high-, mid- and low-resource languages. For writing systems that do not use whitespaces such as Chinese, Japanese, and Thai, we truncated documents to approximately the first 600 characters. For other languages, we kept roughly the first 200 tokens of every document. Truncating the documents is necessary for several reasons: First, shorter documents are more friendly to neural models that are bounded by GPU memories. Second, the first few hundred tokens of Wikipedia articles are usually the main points of the full text, thus are more likely to be topically similar across languages. Last but not least, BM25 tends to over-penalize long documents, which can lead to sub-optimal IR performances (Lv and Zhai, 2011). We hypothesize we can get better relevant judgment labels if we use shorter documents.

Wikidata dump We downloaded the JSON dump[4] of Wikidata, a structured knowledge base that links to Wikipedia. We designed a regex rule that efficiently obtains a list of entities IDs from

[1] Wikidata is a knowledge base that contains links to parallel Wikipedia documents in different languages.

[2] https://www.elastic.co/

[3] https://dumps.wikimedia.org/other/cirrussearch/

[4] https://dumps.wikimedia.org/wikidatawiki/entities/

the Wikidata dump. For every entity ID, we also extracted a list of related (language code, document title) pairs. Using our extracted Wikipedia data, we matched the document titles to Wikipedia document IDs[5]. The extracted data allows us to construct two dictionaries: 1) A dictionary that maps the document ID in some language to its Wikidata entity ID. 2) A reverse dictionary that maps a Wikidata entity ID to document IDs in different languages. This enables us to locate a document's counterpart in another language quickly; we use this information to find link relevant documents across languages.[6]

2.3 Design Choices

Document titles as search queries We considered several methods used to generate search queries. One quick way is to acquire human-generated search queries directly from search logs. However, this is not a viable option because search logs are not publicly available for most languages. Alternatively, we can engage human annotators to manually generate search queries, but this can be time-consuming and expensive, and it is not possible to scale the process quickly to 139 languages.

We use document titles as search queries for two reasons: (1) They are readily available in large amounts for each of the 139 languages, which enables us to build large datasets (i.e., I is large). (2) In certain real-world search settings, queries are typically short, spanning only two to three tokens (Belkin et al., 2003) and informational, covering a wide variety of topics (Jansen et al., 2008). We leave the investigation of complex queries to future work. We want to emphasize that our mining pipeline is compatible with all query types; for example, we can use the first sentences of documents as queries (Schamoni et al., 2014; Sasaki et al., 2018) if desired.

BM25 and Elasticsearch The main step of our mining pipeline is to index documents into a monolingual IR system, and then retrieve a list of relevant documents and similarity scores for every query. We assume the similarity score between a query and document accurately reflects the degree of relevance for that document. Since many Wikipedia dumps contain millions of documents, the computations needed to retrieve relevant documents for all 139 languages is non-trivial. We need an efficient retrieval system that can handle the retrieval task efficiently and accurately. For this reason, we chose Elasticsearch[7] as our monolingual IR system.

Elasticsearch is an open-source, highly optimized search engine software based on Apache Lucene[8]. It has built-in analyzers that handle language-specific preprocessing such as tokenization and stemming. By default, Elasticsearch implements the BM25 weighting scheme (Robertson et al., 2009), a bag-of-word retrieval function that calculates similarity scores between queries and documents based on term frequencies and inverse document frequencies. BM25 is a strong baseline that frequently outperforms existing neural IR models on multiple benchmark IR datasets (Chapelle and Chang, 2011; Guo et al., 2016; McDonald et al., 2018).

We used Elasticsearch 6.5.4 and imported the same settings as the official search indices from Wikipedia[9]. For every query, we configured Elasticsearch to search both document titles and document bodies, with twice the weight given to document titles. We limit Elasticsearch to return only the top 100 documents for each query and assume documents not returned by the search engine are irrelevant. We parallelized the retrieval processes by running multiple Elasticsearch instances on numerous servers and dedicated one Elasticsearch instance to every language.

Discrete relevance judgment labels A potential pitfall of using document titles as queries is that some short queries can be ambiguous (Allan and Raghavan, 2002). For example, it is impossible to figure out whether the search query ”Java” refers to the Java programming language or the island in

[5] Note that documents in different languages do not share document IDs. This means that document N in language X does not refer to the same entity as document N in language Y.

[6] We acknowledge that there are potentially missing inter-language links in Wikidata. This implies that our method may miss the labeling of some relevant documents. Wikidata has several policies to improve its data quality, such as requests for editors to link new Wikipedia articles to entities in Wikidata. There are also automated auditing tools that periodically identify articles with missing or inconsistent Wikidata labels and ask human editors for verification. An interesting research problem for future work is to find ways to quantify the coverage of these inter-language links.

[7] Elasticsearch is also used as the backend search engine for Wikipedia.org

[8] https://lucene.apache.org/core/

[9] For example, the settings for English Wikipedia is available at https://en.wikipedia.org/w/api.php?action=cirrus-settings-dump&format=json&formatversion=2. For BM25, $b = 0.3$ and $k_1 = 1.2$.

Indonesia without other context words. Fortunately, Wikipedia disambiguates different document titles by appending category information to the titles, e.g., Java (Programming Language) and Java (Island), etc. Nevertheless, we do not want to rank retrieved documents solely based on their BM25 scores. To prevent potential ambiguity issues, we smooth out the BM25 scores into discrete relevance judgment labels. We achieve this by using the Jenks natural break optimization (McMaster and McMaster, 2002), an algorithm that finds optimal BM25 score intervals for different labels by iteratively reducing the variance within labels and maximizing the variance between labels.

More specifically, for each query q_i^X, we normalized the BM25 scores $\hat{r}_{ij}$ of d_{ij}^X to the unit range and then used Jenks optimization to distribute the normalized scores into 5 different relevance judgment labels $\{1, 2, 3, 4, 5\}$. We want to emphasize that we did not run Jenks optimization globally across all BM25 scores because the scales of BM25 scores are not consistent across different queries. Additionally, documents that are not returned by Elasticsearch or not linked by any Wikidata are deemed irrelevant and given a label 0. We also assigned the label 6 to the document associated with the title query. So final r_{ij} is of **a scale of 0 to 6**, with 0 being irrelevant and 6 being most relevant.

2.4 Bilingual and Multilingual datasets

BI-139 Using the aforementioned pipeline, we build a bilingual dataset $\{(q_i^X, d_{ij}^Y, r_{ij})\}_{i=1,2,...,I}$ for every X→Y language direction. In the "raw" version, there are 49.28 million unique queries and 34.06 billion (query, document, label) triplets across $139 \times 138 = 19,182$ language directions. We also generated a "base" version, which contains standard train, validation, test1, and test2 subsets for each language direction. Train sets contain up to I=10,000 queries, while validation, test1, and test2 sets each contain up to 1,000 queries. We ensured that queries in the train and validation/test sets of one language direction do not overlap with the queries in the test sets from other language directions. For every query, we ensure there are precisely K =100 candidate documents by filling the shortfall with random irrelevant documents.

MULTI-8 This is a multilingual CLIR dataset covering 8 languages from various regions of the world (Arabic, German, English, Spanish, French, Japanese, Russian, and Chinese). First, we restricted queries to those with a relevant document ($r_{ij} = 6$) in all 8 languages. Then, for each query q_i^X, we use the monolingual IR systems to collect 100 documents in the same language d_{ij}^X.[10] Similar to BI-139 base, if ElasticSearch returns less than 100 documents labels ($r_{ij} \geq 1$), then we fill-up the short-fall with random irrelevant documents with label $r_{ij} = 0$. Finally, we merge these document lists such that for any query in language X, we have 7×100 documents in the other 7 languages.

Similar to the base version of BI-139, the train sets contain 10,000 queries, while validation, test1, and test2 sets contain 1,000 queries; but note the query sets are different. This dataset supports two kinds of research: First, one can still evaluate bilingual CLIR (single-language retrieval) like BI-139, but exploit training multilingual models using more than two languages. Second, one can evaluate on multilingual CLIR (mixed-language retrieval), where the document list to be re-ranked contains two or more languages. This research direction is relatively unexplored, with the exception of early work in the 2000s in the CLEF campaign (Savoy and Braschler, 2019).

2.5 File Formats

{"src_id": "6267",
"src_query": "Cultural imperialism",
"tgt_results": [["3383724", 6], ["19028", 5], ["6291141", 4], ["4394682", 2], ["138124", 1], ["1245746", 1], ["1004260", 0], ...}

Figure 4: An example English query "Cultural imperialism" and the document IDs and labels of its relevant Chinese documents.

6499809 ⟨TAB⟩ Structured light is the process of projecting a known pattern (often grids or horizontal bars) on to a scene...

Figure 5: The IDs and texts of documents are stored tab-separated in a text file.

For every language direction, we store queries and their relevant document IDs and labels in the JSON Lines format (Figure 4). For each unique language, we store the IDs and texts of documents in TSV files (Figure 5). Note that we will release both the truncated and the original documents.

3 Experimental Setup

[10] Recall that our Wikidata entities dictionary can map a language-independent entity to query strings (Wikipedia article titles) in any language.

af	**als**	**am**	**an**	**ar**	**arz**	**ast**	**az**	**azb**	**ba**	**bar**	**be**	**bg**	**bn**	**bpy**	**br**
.90	.88	.56	.90	.80	.86	.88	.80	.87	.87	.89	.83	.85	.78	.85	.84
bs	**bug**	**ca**	**cdo**	**ce**	**ceb**	**ckb**	**cs**	**cv**	**cy**	**da**	**de**	**diq**	**el**	**eml**	**eo**
.89	.91	.88	.85	.90	.89	.72	.89	.84	.87	.90	.88	.81	.83	.80	.87
es	**et**	**eu**	**fa**	**fi**	**fo**	**fr**	**fy**	**ga**	**gd**	**gl**	**gu**	**he**	**hi**	**hr**	**hsb**
.87	.83	.86	.85	.86	.87	.84	.90	.78	.79	.87	.78	.82	.79	.88	.86
ht	**hu**	**hy**	**ia**	**id**	**ilo**	**io**	**is**	**it**	**ja**	**jv**	**ka**	**kk**	**kn**	**ko**	**ku**
.88	.86	.82	.90	.00	.88	.86	.83	.84	.84	.89	.81	.85	.67	.86	.76
ky	**la**	**lb**	**li**	**lmo**	**lt**	**lv**	**mai**	**mg**	**mhr**	**min**	**mk**	**ml**	**mn**	**mr**	**mrj**
.82	.88	.88	.85	.83	.86	.85	.80	.88	.84	.92	.86	.87	.86	.74	.82
ms	**my**	**mzn**	**nap**	**nds**	**ne**	**new**	**nl**	**nn**	**no**	**oc**	**or**	**os**	**pa**	**pl**	**pms**
.89	.77	.85	.85	.88	.73	.75	.89	.90	.89	.91	.71	.83	.76	.86	.78
pnb	**ps**	**pt**	**qu**	**ro**	**ru**	**sa**	**sah**	**scn**	**sco**	**sd**	**sh**	**si**	**simple**	**sk**	**sl**
.70	.72	.86	.81	.89	.85	.73	.77	.81	.94	.78	.87	.48	.93	.86	.89
sq	**sr**	**su**	**sv**	**sw**	**szl**	**ta**	**te**	**tg**	**th**	**tl**	**tr**	**tt**	**uk**	**ur**	**uz**
.88	.88	.91	.88	.87	.92	.85	.81	.85	.81	.89	.87	.87	.85	.85	.84
vec	**vi**	**vo**	**wa**	**war**	**wuu**	**xmf**	**yi**	**yo**	**zh**						
0.88	0.89	0.89	0.75	0.86	0.83	0.79	0.65	0.89	0.84						

Table 1: Results of 138 language directions from BI-139 base with English queries. In each cell, the top shows a candidate's language code and the bottom shows the NDCG@10 score for that language direction.

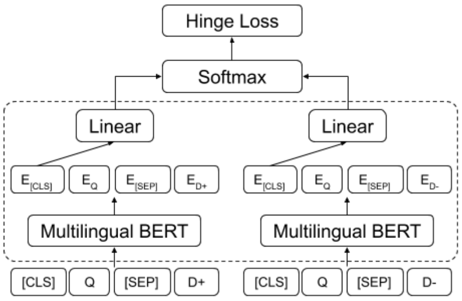

Figure 6: Neural architecture of our baseline CLIR model. Modules in the dotted rectangle share weights.

Baseline neural CLIR model We follow the implementation of the vanilla BERT ranker model (MacAvaney et al., 2019), which obtained strong results in monolingual IR. As shown in Figure 6, the model encodes a query-document pair with BERT (Devlin et al., 2019) and stacks a linear combination layer on top of the [CLS] token. We extended the ranker model to use multilingual BERT[11]. At training time, we sample documents pairs in which the positive documents have higher relevance judgment labels than the negative documents. For each document pair, we obtain scores for both documents using the same BERT ranker model. We then optimize the parameters with pairwise hinge loss and Adam optimizer. We trained all models for 20 epochs and sampled around 1,000 training pairs for each epoch. At inference time, we rerank documents based on the output scores from the BERT ranker model.

Evaluation metric We report all results in NDCG (normalized discounted cumulative gain), an IR metric that measures the usefulness of documents based on their ranks in the search results (Järvelin and Kekäläinen, 2002). Following a common practice from the IR community, we calculate NDCG@10, which only evaluates the top 10 returned documents. For a given query, let ρ_i be the relevance judgment label of the i-th document in the predicted document ranking and ϕ_i be the relevance judgment label of the i-th document in the optimal document ranking. We define DCG@10 and ideal DCG@10 as:

$$\begin{aligned} \text{DCG@10} &= \sum_{i=1}^{10} \frac{2^{\rho_i} - 1}{log_2(i+1)} \\ \text{IDCG@10} &= \sum_{i=1}^{10} \frac{2^{\phi_i} - 1}{log_2(i+1)} \end{aligned} \tag{2}$$

[11] We used BERT-Base, Multilingual Cased

We can calculate NDCG@10 for that query as:

$$\text{NDCG@10} = \frac{\text{DCG@10}}{\text{IDCG@10}} \tag{3}$$

The NDCG@10 of a test set is the arithmetic mean of NDCG@10 values for all queries. The range of the metric is [0, 1] and a higher NDCG@10 score means predicted rankings are closer to the ideal rankings.

3.1 Results on BI-139

We present results on the 138 target languages for English queries. For each language direction, we trained a baseline CLIR model on the base train set and kept the checkpoint with the best NDCG@10 performance on the base validation set. We reranked the documents in the base test1 set and calculated NDCG@10. Table 1 lists the the baseline results.[12] The pleasant surprise here is that the baseline CLIR models also generally did pretty well on languages that are not officially supported by multilingual BERT. For example, the model achieved 0.65 on Yiddish (yi) and 0.75 on Walloon (wa) when multilingual BERT was trained on neither of these languages. There are several explanations for this. For one, we hypothesize that low resource languages such as Yiddish, a high German-derived language, and Walloon, a Romance language, benefit from their similarities to other languages within the same language families. For queries such as named entities, it is also possible that some relevant cross-language Wikipedia document may be multilingual and contain some overlap with the query term untranslated. The details will depend on the query in question.

3.2 Results on MULTI-8

Multilingual IR is a field that has been largely unexplored in recent years. MULTI-8 enables evaluation in two kinds of scenarios (see Table 2):

Single-language retrieval This scenario is similar to BI-139 in terms of evaluation, i.e. during test we only have queries in source language $q^X = S_{test}$ and documents in one target language $d^Y = T_{test}$. We divide MULTI-8 test set into $8 \times 7 = 56$ pairs.

For training, we compare **bilingual model ($\mathbf{BM_{S \rightarrow T}}$)** trained in every language pair, against a **multilingual model (MM)** trained on data concatenated from all 56 language directions. As we can see in Table 3, the MM model performs better than the respective BM models in most language directions. This suggests that multilingual training is a promising research direction even for single-language retrieval.

Mix-language retrieval In this scenario, at test time we have a single source query $q^X = S_{test}$ and wish to retrieve documents $d^Y = A_{test}$ which can be in any of the 8 MULTI-8 languages. The multilingual model (MM) can be applied directly, but the bilingual model (BM) requires some modifications. One can run multiple BM one for each target language, then merge the resulting document lists (Savoy, 2003; Tsai et al., 2008). A common strategy, which we adopt here, is to z-normalize the output scores and rank all the test documents based on z-scores.

As seen in Table 4, the multilingual model performs significantly better than the ensembled/merged bilingual models. The average NDCG@10 of the multilingual model is 0.684, which is 17.1% than bilingual models with z-score merging strategy.

[12] Language codes: af:Afrikaans, als:Alemannic, am:Amharic, an:Aragonese, ar:Arabic, arz:Egyptian Arabic, ast:Asturian, az:Azerbaijani, azb:Southern Azerbaijani, ba:Bashkir, bar:Bavarian, be:Belarusian, bg:Bulgarian, bn:Bengali, bpy:Bishnupriya Manipuri, br:Breton, bs:Bosnian, bug:Buginese, ca:Catalan, cdo:Min Dong, ce:Chechen, ceb:Cebuano, ckb:Kurdish (Sorani), cs:Czech, cv:Chuvash, cy:Welsh, da:Danish, de:German, diq:Zazaki, el:Greek, eml:Emilian-Romagnol, en:English, eo:Esperanto, es:Spanish, et:Estonian, eu:Basque, fa:Persian, fi:Finnish, fo:Faroese, fr:French, fy:West Frisian, ga:Irish, gd:Scottish Gaelic, gl:Galician, gu:Gujarati, he:Hebrew, hi:Hindi, hr:Croatian, hsb:Upper Sorbian, ht:Haitian, hu:Hungarian, hy:Armenian, ia:Interlingua, id:Indonesian, ilo:Ilocano, io:Ido, is:Icelandic, it:Italian, ja:Japanese, jv:Javanese, ka:Georgian, kk:Kazakh, kn:Kannada, ko:Korean, ku:Kurdish (Kurmanji), ky:Kirghiz, la:Latin, lb:Luxembourgish, li:Limburgish, lmo:Lombard, lt:Lithuanian, lv:Latvian, mai:Maithili, mg:Malagasy, mhr:Meadow Mari, min:Minangkabau, mk:Macedonian, ml:Malayalam, mn:Mongolian, mr:Marathi, mrj:Hill Mari, ms:Malay, my:Burmese, mzn:Mazandarani, nap:Neapolitan, nds:Low Saxon, ne:Nepali, new:Newar, nl:Dutch, nn:Norwegian (Nynorsk), no:Norwegian (Bokmål), oc:Occitan, or:Odia, os:Ossetian, pa:Eastern Punjabi, pl:Polish, pms:Piedmontese, pnb:Western Punjabi, ps:Pashto, pt:Portuguese, qu:Quechua, ro:Romanian, ru:Russian, sa:Sanskrit, sah:Sakha, scn:Sicilian, sco:Scots, sd:Sindhi, sh:Serbo-Croatian, si:Sinhalese, simple:Simple English, sk:Slovak, sl:Slovenian, sq:Albanian, sr:Serbian, su:Sundanese, sv:Swedish, sw:Swahili, szl:Silesian, ta:Tamil, te:Telugu, tg:Tajik, th:Thai, tl:Tagalog, tr:Turkish, tt:Tatar, uk:Ukrainian, ur:Urdu, uz:Uzbek, vec:Venetian, vi:Vietnamese, vo:Volapük, wa:Walloon, war:Waray, wuu:Wu, xmf:Mingrelian, yi:Yiddish, yo:Yoruba, zh:Chinese

Scenario	Models	Train	Evaluation
Single-language retrieval	$\{\mathrm{BM}_{S\rightarrow T}\}$	$q^X = S_{train}, d^Y = T_{train}$	$q^X = S_{test}, d^Y = T_{test}$
	MM	$q^X = A_{train}, d^Y = A_{train}$	
Mix-language retrieval	$\{\mathrm{BM}_{S\rightarrow T}\}$	$q^X = S_{train}, d^Y = T_{train}$	$q^X = S_{test}, d^Y = A_{test}$
	MM	$q^X = A_{train}, d^Y = A_{train}$	

Table 2: Different ways of using MULTI-8. A refers to the concatenation of all languages, which is used in mix-language retrieval. S and T refer to the queries/documents in the source and target language under consideration for the bilingual case (i.e., single-language retrieval similar to BI-139 setups). For either, it is possible to train either bilingual models (BM) based on pairwise data or a multilingual model (MM) based on all language data.

q \ d	ar	de	en	es	fr	ja	ru	zh
ar		.65▽	**.60▲**	**.65▲**	.64▽	.65▽	**.60▲**	**.64▲**
de	.75▽		**.75▲**	**.77▲**	**.72▲**	**.72▲**	**.74▲**	**.71▲**
en	**.79▲**	**.82▲**		**.83▲**	**.79▲**	.83▽	.82▽	.82▽
es	**.74▲**	**.72▲**	**.76▲**		**.75▲**	.74▽	**.74▲**	.74▽
fr	**.75▲**	**.75▲**	**.76▲**	**.79▲**		.75▽	**.74▲**	.76▽
ja	.71▽	**.68▲**	**.67▲**	**.68▲**	.67▽		.69▽	.70▽
ru	.73▽	**.71▲**	**.71▲**	**.73▲**	.73▽	.72▽		**.71▲**
zh	**.67▲**	**.67▲**	**.63▲**	**.66▲**	.66▽	**.64▲**	**.66▲**	

Table 3: MULTI-8 single-language retrieval results of bilingual models (BM). The rows are the source query language, and the columns are the target document language. The up arrows next to NDCG@10 scores indicate instances where the multilingual model (MM) outperforms the bilingual models.

	ar	**de**	**en**	**es**	**fr**	**ja**	**ru**	**zh**
BM	.52	.58	.66	.60	.63	.59	.57	.58
MM	.59	.72	.75	.73	.65	.68	.62	.68
△%	13	23	14	22	16	10	20	13

Table 4: MULTI-8 mix-language retrieval results. △% shows percent improvement of MM over BM z-norm.

4 Related Work

Information retrieval (IR) has made a tremendous amount of progress, shifting focus from traditional bag-of-world retrieval functions such as tf-idf (Salton and McGill, 1986) and BM25 (Robertson et al., 2009), to neural IR models (Guo et al., 2016; Hui et al., 2018; McDonald et al., 2018) which have shown promising results on multiple monolingual IR datasets. Recent advances in pre-trained language models such as BERT (Devlin et al., 2019) have also led to significant improvements in IR tasks. For example, MacAvaney et al. (2019) achieves state-of-the-art performances on benchmark datasets by incorporating BERT's context vectors into existing baseline neural IR models (McDonald et al., 2018). Training on synthetic is also a common practice, e.g., Dehghani et al. (2017) show that supervised neural ranking models can greatly benefit from pre-training on BM25 labels.

Cross-lingual Information Retrieval (CLIR) is a sub-field of IR that is becoming increasingly important as new documents in different languages are being generated every day. The field has progressed from translation-based methods (Zhou et al., 2012; Oard, 1998; McCarley, 1999; Yarmohammadi et al., 2019) to recent neural CLIR models (Vulić and Moens, 2015; Litschko et al., 2018; Zhang et al., 2019) that rely on cross-lingual word embeddings. In contrast to the wide availability of monolingual IR datasets (Voorhees, 2005; Craswell et al., 2020), cross-lingual and multilingual IR

Dataset	#Lang	Manual?	Multilingual?	#query	#document	#triplets
(CLEF 2000-2003)	10	yes	yes	2.2K	1.1M	33K
(MATERIAL, 2017)	7	yes	no	11.5K	90K	∼20K
(Schamoni et al., 2014)	2	no	no	245K	1.2M	3.2M
(Sasaki et al., 2018)	25	no	no	10.9M	23.9M	40.1M
CLIRMatrix BI-139 raw	139	no	no	49.3M	50.5M	34.1B
CLIRMatrix BI-139 base	139	no	no	27.5M	50.1M	22.3B
CLIRMatrix MULTI-8	8	no	yes	10.4K	13.4M	72.8M

Table 5: Comparison of CLIR datasets by number of languages (**#Lang**), whether it is manually constructed or supports multilingual retrieval, and data statistics. Large **#query** and **#triplets** are needed for neural training.

datasets are scarce. Examples of the widely used CLIR datasets are the CLEF 2000-2003 collection (Ferro and Silvello, 2015), which focus primarily on European languages, and IARPA MATERIAL/OpenCLIR collection (Zavorin et al., 2020), which focus on a few low-resource language directions. Creating a CLIR dataset for more language directions remains an open challenge.

Extracting CLIR datasets from Wikipedia has been explored in previous work. Schamoni et al. (2014) build a German–English bilingual CLIR dataset from Wikipedia, which contains 245,294 German queries and 1,226,741 English documents. They convert the first sentences from German Wikipedia documents into queries and follow Wikipedia's interlanguage links to find relevant documents in English. Sasaki et al. (2018) apply the same techniques and release a larger CLIR dataset which contains English queries and relevant documents in 25 languages. Both datasets truncate the documents to the first 200 tokens and rely on bidirectional inter-article links to find partially relevant documents. Our contribution differs in three important aspects: (i) BI-139 is a significantly larger dataset, covering more languages and more documents. (ii) MULTI-8 provides a new multilingual retrieval setup, not previously available. (iii) We argue that our method can reliably find more relevant documents by propagating search results from monolingual IR systems to other languages via Wikidata. This is in contrast to directly using bidirectional links extracted from Wikipedia documents to determine relevance, which are much sparser. Further, our method allows for more finer-grained levels of relevance (e.g. as opposed to binary relevance), making the dataset more challenging.

A comparison of various existing CLIR datasets is presented in Table 5.

5 Conclusion and future work

We present *CLIRMatrix*, the largest and the most comprehensive collection of bilingual and multilingual CLIR datasets to date. The *BI-139* dataset supports CLIR in 139×138 language pairs, whereas the *MULTI-8* dataset enables mix-language retrieval in 8 languages. The large number of supported language directions allows the research community to explore and build new models for many more languages, especially the low-resource ones. We document baseline NDCG results using a neural ranker based on multilingual BERT. Our mix-language retrieval experiments on MULTI-8 show that a single multilingual model can significantly outperform the combination of multiple bilingual models.

For future work, we think it will be interesting to look at:

1. zero-shot CLIR models for low-resource languages,
2. comparison of end-to-end neural rankers with traditional translation+IR pipelines in terms of both scalability, cost, and retrieval accuracy,
3. advanced neural architectures and training algorithms that can exploit our large training data,
4. building universal models for multilingual IR.

References

James Allan and Hema Raghavan. 2002. Using part-of-speech patterns to reduce query ambiguity. In *Proceedings of the 25th annual international ACM*

SIGIR conference on Research and development in information retrieval, pages 307–314.

Nicholas J Belkin, Diane Kelly, G Kim, J-Y Kim, H-J Lee, Gheorghe Muresan, M-C Tang, X-J Yuan, and Colleen Cool. 2003. Query length in interactive information retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 205–212.

Olivier Chapelle and Yi Chang. 2011. Yahoo! learning to rank challenge overview. In *Proceedings of the Learning to Rank Challenge*, pages 1–24.

CLEF 2000-2003. *The CLEF Test Suite for the CLEF 2000-2003 Campaigns – Evaluation Package*. https://catalog.elra.info/en-us/repository/browse/ELRA-E0008/.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicola Ferro and Gianmaria Silvello. 2015. CLEF 2000-2014: Lessons learnt from ad hoc retrieval. In *Proceedings of the 6th Italian Information Retrieval Workshop, Cagliari, Italy, May 25-26, 2015*, volume 1404 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 55–64.

Kai Hui, Andrew Yates, Klaus Berberich, and Gerard De Melo. 2018. Co-pacrr: A context-aware neural ir model for ad-hoc retrieval. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 279–287.

Bernard J Jansen, Danielle L Booth, and Amanda Spink. 2008. Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management*, 44(3):1251–1266.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with bert.

Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1253–1256.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331.

Yuanhua Lv and ChengXiang Zhai. 2011. When documents are very long, bm25 fails! In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1103–1104.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. Cedr: Contextualized embeddings for document ranking. In *SIGIR*.

MATERIAL. 2017. *Machine Translation for English Retrieval of Information in Any Language (MATERIAL)*. https://www.iarpa.gov/index.php/research-programs/material.

J Scott McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 208–214. Association for Computational Linguistics.

Ryan McDonald, George Brokos, and Ion Androutsopoulos. 2018. Deep relevance ranking using enhanced document-query interactions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1849–1860.

Robert McMaster and Susanna McMaster. 2002. A history of twentieth-century american academic cartography. *Cartography and Geographic Information Science*, 29(3):305–321.

Bhaskar Mitra and Nick Craswell. 2018. An introduction to neural information retrieval. *Foundations and Trends in Information Retrieval*, 13(1):1–126.

Douglas W Oard. 1998. A comparative study of query and document translation for cross-language information retrieval. In *Conference of the Association for Machine Translation in the Americas*, pages 472–483. Springer.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Gerard Salton and Michael J McGill. 1986. Introduction to modern information retrieval.

Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 458–463.

Jacques Savoy. 2003. Report on clef-2003 multilingual tracks. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 64–73. Springer.

Jacques Savoy and Martin Braschler. 2019. *Lessons Learnt from Experiments on the Ad Hoc Multilingual Test Collections at CLEF*, pages 177–200. Springer International Publishing, Cham.

Shigehiko Schamoni, Felix Hieber, Artem Sokolov, and Stefan Riezler. 2014. Learning translational and knowledge-based similarities from relevance rankings for cross-language retrieval. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–494.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Ming-Feng Tsai, Yu-Ting Wang, and Hsin-Hsi Chen. 2008. A study of learning a merge model for multilingual information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 195–202.

Ellen M Voorhees. 2005. The trec robust retrieval track. In *ACM SIGIR Forum*, volume 39, pages 11–20. ACM New York, NY, USA.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 363–372.

Mahsa Yarmohammadi, Xutai Ma, Sorami Hisamoto, Muhammad Rahman, Yiming Wang, Hainan Xu, Daniel Povey, Philipp Koehn, and Kevin Duh. 2019. Robust document representations for cross-lingual information retrieval in low-resource settings. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 12–20, Dublin, Ireland. European Association for Machine Translation.

Ilya Zavorin, Aric Bills, Cassian Corey, Michelle Morrison, Audrey Tong, and Richard Tong. 2020. Corpora for cross-language information retrieval in six less-resourced languages. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 7–13, Marseille, France. European Language Resources Association.

Rabih Zbib, Lingjun Zhao, Damianos Karakos, William Hartmann, Jay DeYoung, Zhongqiang Huang, Zhuolin Jiang, Noah Rivkin, Le Zhang, Richard Schwartz, and John Makhoul. 2019. Neural-network lexical translation for cross-lingual ir from text and speech. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 645–654, New York, NY, USA. Association for Computing Machinery.

Rui Zhang, Caitlin Westerfield, Sungrok Shim, Garrett Bingham, Alexander Richard Fabbri, William Hu, Neha Verma, and Dragomir Radev. 2019. Improving low-resource cross-lingual document retrieval by reranking with deep bilingual representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3173–3179.

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):1–44.