HC3: A Suite of Test Collections for CLIR Evaluation over Informal Text

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

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

Douglas W. Oard
University of Maryland
College Park, MD, USA
oard@umd.edu

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

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

Petra Galuščáková
Univ. Grenoble Alpes,
CNRS, Grenoble INP*, LIG
Grenoble, France

ABSTRACT

While there are many test collections for Cross-Language Information Retrieval (CLIR), none of the large public test collections focus on short informal text documents. This paper introduces a new pair of CLIR test collections with millions of Chinese or Persian Tweets or Tweet threads as documents, sixty event-motivated topics written both in English and in each of the two document languages, and three-point graded relevance judgments constructed using interactive search and active learning. The design and construction of these new test collections are described, and baseline results are presented that demonstrate the utility of the collections for system evaluation. Shallow pooling is used to assess the efficacy of active learning to select documents for judgment.

CCS CONCEPTS

• **Information systems → Information retrieval**; **Evaluation of retrieval results**; **Test collections**; **Multilingual and cross-lingual retrieval**; *Relevance assessment*.

KEYWORDS

Test Collection, Cross-Language Information Retrieval, CLIR, Evaluation, Tweet-based documents

ACM Reference Format:
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 (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3539618.3591893

1 INTRODUCTION

Many test collections exist to facilitate the evaluation of cross-language information retrieval (CLIR) algorithms. Starting with the TREC-6 Cross-language track that supported evaluation across English, German and French documents, international evaluations such as CLEF [1], NTCIR [15], and FIRE [23] have produced cross-language evaluation sets covering a diversity of human languages. The documents in most of these collections are newswire articles. Recent collections have added Wikipedia articles [42] or even general Web documents [28]. But almost all prior collections use documents that are reasonably well-edited, third-person texts. Short, informal, conversational texts, such as emails, Tweets, or Reddit comment threads, have not generally been a part of these collections. As a consequence, it is difficult to predict how experimental CLIR systems will compare when faced with such documents.

The HLTCOE CLIR Conversation Collection[1], or HC3, is a new set of test collections for the evaluation of CLIR algorithms. HC3 is designed to fill this gap in the coverage of publicly available CLIR test sets. HC3 documents are Tweets or Tweet reply chains, which we call *conversations*. Tweets are short and often informal texts, sometimes with bad spelling or grammar, and with some notably different writing conventions than more formal sources such as news. Tweets can also be forwarded ("Retweeted") or replied to, thus establishing a conversational thread structure. These features of Tweets present problems to traditional statistical CLIR algorithms, which have often been designed for documents that can be assessed without conversational context, relying on lexical matches between (translated) query and document terms. Recently developed neural approaches to CLIR rely less directly on lexical matching [26, 40], but it is not yet clear how they can be best adapted to handle informal, ill-formed texts. HC3 will facilitate this study, spurring progress in CLIR.

A CLIR test collection to allow the evaluation of CLIR algorithms over informal text should optimally satisfy the following desiderata:

- The document collection should be large, include multiple disparate languages, and focus on short texts.
- The topics must be expressed in a different language than at least some of the documents, and there should be enough topics to measure statistically significant differences between meaningfully different retrieval algorithms.
- The relevance judgments should be sufficiently consistent, accurate, and extensive to permit reliable calculation of insightful evaluation measures.
- To the extent possible, the collection should be "future-proof" in the sense that it can be used to compare present systems

*Institute of Engineering Univ. Grenoble Alpes.

[1] https://github.com/hltcoe/HC3

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591893

against future systems that use technologies not available when the collection was built.

We aimed to meet each of these desiderata in the creation of HC3.

This paper makes the following contributions:

- It presents the HC3 collections, how they were built, and how they can be used.
- It compares the use of active learning for the selection of documents for relevance assessment with a more traditional method based on pooling.
- It provides evaluation results using the test collection for a set of strong baseline systems to which future systems can be compared. Moreover, it provides software for recomputing those baseline results in the event that some portions of the collection become unavailable, as can happen when using short text from Twitter.
- It describes aligned topics across languages and aligned topics across collections (with the HC4 news test collection [18]) that can be used in multilingual or multi-genre experiments.

2 RELATED WORK

We review prior work on the creation of information retrieval test collections, with a focus on test collections for CLIR and short text.

2.1 Test Collections

Information retrieval test collections that follow the Cranfield paradigm represent retrieval tasks in which queries specify a desired topic and content units ("documents") are considered relevant if any part of that content substantially addresses the specified topic [35]. In common usage, *topic* is a broad term, encompassing for example subjects, entities, and events. Content characteristics such as reading level, veracity, and sentiment are typically considered non-topical, as are author characteristics such as identity, mental state, and national origin. Although there are specialized test collections that focus on such non-topical characteristics, our focus in this paper is on representing topical retrieval tasks.

A test collection is typically thought of as containing topic descriptions ("topics" for short), documents, and relevance judgments. Because of scale, relevance judgments are typically defined over a small sample of the cross product of topics and documents. This sample is typically drawn to maximize the coverage of true positives. Prominent among the sampling techniques are pooling [13], interactive search [5], and active learning [6].

A common use of test collections, and the one we are concerned with in this paper, is to compare ranked retrieval systems. In general, systems that rank a greater number of documents that are more highly relevant nearer the top of the list are to be preferred. Some evaluation measures that model this broad goal include Average Precision (AP), normalized Discounted Cumulative Gain (nDCG), Expected Reciprocal Rank (ERR), and Rank Biased Precision (RBP) [25]. Although different assessors might disagree on the relevance of a specific document to a specific topic, experiments have shown that system comparisons using such measures are relatively insensitive to which assessor judgments are used [34].

When using a test collection to simulate an interactive retrieval task in which users formulate queries, a model of query formulation is also needed. A common approach is to encode one or more query variants as fields in the topic description. Systems are then typically compared based on the consistent choice of one of those query variants. It is common to refer to these as "TREC-style" topics, with a short variant (e.g., with a length similar to a Web query) known as the "title" field, and a longer variant (e.g., similar to what a searcher might first say to someone who is helping them to find something) known as the "description" field [13]. Because the description field is sometimes intended by its author to be read in the context of the title field, some information retrieval experiments report results for a concatenation of the two fields. Our focus in this paper is on the use of these kinds of query fields, which is often referred to as *ad hoc* ranked retrieval.

Because the construction of test collections can be expensive, reuse of test collections has long been a topic of interest as a way of amortizing those costs. The emergence of machine learning techniques with a voracious appetite for training data has further exacerbated both the challenge of achieving reusability (because not only are more documents needed, but also more queries) and the value of achieving reusability (because training future systems depends on the ability to perform formative evaluation). Experiments with older (smaller) test collections had indicated that pooling highly ranked results from a diverse set of systems was often sufficient to produce system rankings that were relatively insensitive to ablation of relevance judgments from any one contributor to those pools [43]. However, there is now clear evidence that this is less true for larger (and thus more sparsely sampled) collections [4]. One consequence of that observation has been increased interest in the use of sampling methods based on active learning [6].

2.2 CLIR Test Collections

Test collections for CLIR generally follow the design of monolingual test collections, with the notable difference that query fields are also provided in one or more languages different from the language of the documents [2, 11, 17, 18]. These query fields are generally produced by human translators, and the usual practice is for those translators to be instructed to produce queries that are representative of natural expression in the chosen query language [29]. It is common for the topic descriptions to also include query fields that are written in the language of the documents, for use in monolingual experiments. Relevance judgments are typically performed by assessors who are fluent in the language of the documents, sometimes with reference to a document-language version of the topic description.

2.3 Test Collections for Short Text

Most information retrieval test collections, including those for CLIR, have been built using news or Web documents, although specialized test collections for other genres such as blogs [22], medical documents [30], patents [21], and scholarly papers [15] have also been created. A common characteristic of most of these text collections is that the documents are relatively long. Moreover, such documents are often professionally written (blogs and some Web text being notable exceptions), and they are often intended to be interpreted individually rather than as a part of some larger unit.

There has, however, been considerable interest across multiple research communities in the design and evaluation of computational methods for working with short texts, a collection of genres that have been referred to collectively as microtext [39]. Genres subsumed in this broader term microtext include multi-party text "chatrooms" and short text broadcast services such as Twitter [20] and Sina Weibo [33]. In addition to the usual challenges of test collection design, test collections designed for short text must make genre-specific decisions as to what constitutes the "document" that is to be the unit of retrieval. In asynchronous genres such as Twitter, email and message boards, automated reconstruction of hierarchical thread structures (e.g, based on reply chains) allows parts or all of entire threads to be indexed [19].

There are several English-only IR test collections for short text. The NTCIR Short Text Conversations (STC) track [16] developed a short text test collection for Chinese; however, we understand that the collection was available only to track participants and cannot be redistributed.[2]

3 COLLECTION CREATION METHODOLOGY

HC3 was developed as a test collection for CLIR over document content that is less formal than newswire, thereby allowing the assessment of CLIR algorithms over informal text. Two design decisions framed the construction of the collection. First, a single Tweet may not have sufficient context to support a relevance assessment. On the other hand, the entire tree of responses from a root Tweet could be too long and nonlinear to assess in a reasonable time. We therefore used a single path in the reply tree, which we refer to as a Twitter conversation, as a document.

The second design decision was the alignment of HC3 to HC4 [18], a sister collection of CommonCrawl News documents. The documents in the two collections are broadly time-aligned, having been written or posted between August 2016 and August 2019. To the extent that it was possible, the HC3 topics were selected to be identical or similar to the topics in HC4. This alignment can support comparisons of algorithms across formal and informal text. In addition to aligning the content of the two collections, we also aligned the development of the two collections. Like HC4, the annotation effort was divided into two phases. While the first phase of identifying initial documents through interactive search was different between the two collections, the use of active learning to complete relevance assessment was shared.

3.1 Document Creation

Using Tweets as the basis of a retrieval task is not as straightforward as using newswire, where the articles produced are naturally seen as documents. An individual Tweet can lack context, making relevance assessment difficult. Tweets are both responded to and copied in the form of Retweets, so a rich tree structure exists that can be used to combine Tweets. We aggregated replies to Tweets to form linear conversation threads. Each such thread served as a single document.

We began with Twitter's random 1% feed, known at the Twitter "spritzer" stream available from the Internet Archive.[3] We separated the data by language using Twitter's language ID, selecting the Chinese[4] and Persian subsets of the collection.

We augmented our Tweet collection in two ways. First, to ensure we could generate coherent topics, and to support alignment with HC4, we allowed our assessors to include in the collection Tweets they identified through live search of Twitter. Second, our 1% random sample of the Twitter feed was unlikely to contain complete reply chains from given Tweets back to their roots. Ancestor Tweets that were not part of the 1% were fetched and added to the collection when they were still available from Twitter.

This construction process led to some threads that were subsequences of longer threads. In such cases we chose the longest such thread, although we capped document length at one hundred Tweets. Moreover, no Tweet is included in more than one document, to limit duplication of content. A total of 499,267 (9%) and 1,636,024 (22.3%) documents include more than one Tweet in Chinese and Persian, respectively.

3.2 Topic Development

Standard practice for initial development of TREC-style topics is for an annotator to express a topic, then use interactive search to identify its prevalence in the collection [36]. This practice was not followed for HC3 because of the desire to align the topics with those of HC4. Given that HC4 topics generally focus on news events, we hypothesized that the original document collection drawn from a random 1% sample of Tweets would likely not include sufficient relevant information on the topics. Therefore, annotators were given all HC4 topics available at the time in any language, and searched "live" Twitter using search terms in the language of the Tweets to find Tweets about each topic. The goal of this development phase was to identify between three and ten Tweets relevant to each topic.

Figure 1 shows the main interface for the task, which was fielded on a Mechanical Turk-style platform called Turkle (Turk in the Local Environment).[5] Using the "LAUNCH TWITTER SEARCH" button, a new tab in the browser would open with some search parameters pre-set. These search parameters identified the language of the Tweet and the time interval during which the Tweet was posted. Annotators added their own search terms to these default parameters, and using one or more searches, identified relevant Tweets.

Once the annotator had either identified a sufficient number of Tweets or failed to find anything on the topic, they clicked the "FINALIZE TASK" button, made a recommendation on the topic, and provided other observations about it.

3.3 Relevance Judgments

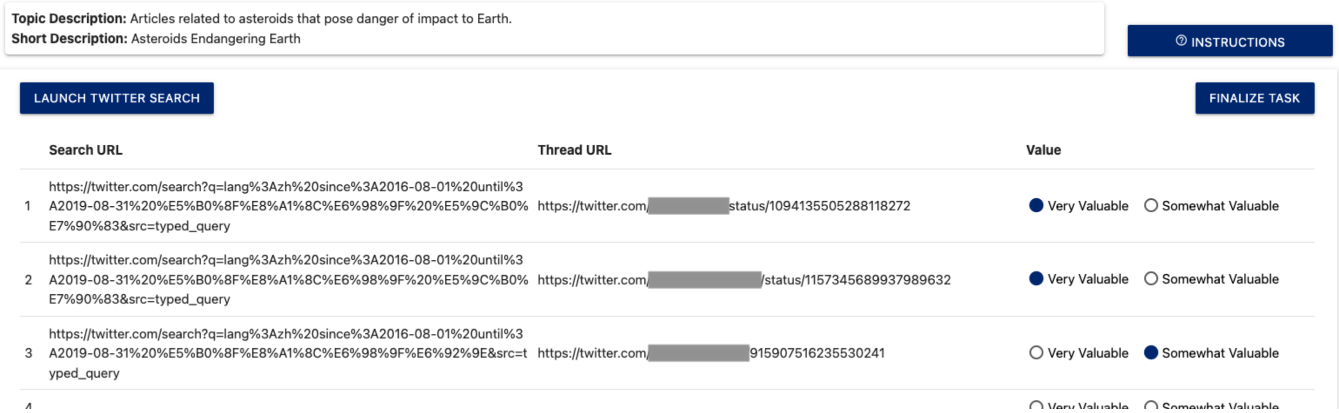
After topic development, all topics with at least one relevant document were selected for more complete assessment. As is standard practice given the impracticality of judging millions of documents, a vast majority of which are not relevant, we adopt the common practice of assessing as many relevant documents as is feasible. As in HC4, we used the active learning system HiCAL [7], to iteratively

[2] Personal communication with Tetsuya Sakai.

[3] https://archive.org/details/twitterstream

[4] At the time the annotation was completed, Twitter did not distinguish between Simplified and Traditional Chinese characters. In addition, some Cantonese was identified as Chinese when it was expressed in Chinese characters. Thus the collection contains a mixture of Chinese representations.

[5] https://github.com/hltcoe/turkle



Topic Description: Articles related to asteroids that pose danger of impact to Earth.
Short Description: Asteroids Endangering Earth

	Search URL	Thread URL	Value
1	https://twitter.com/search?q=lang%3Azh%20since%3A2016-08-01%20until%3A2019-08-31%20%E5%B0%8F%E8%A1%8C%E6%98%9F%20%E5%9C%B0%E7%90%83&src=typed_query	https://twitter.com/[redacted]status/1094135505288118272	● Very Valuable ○ Somewhat Valuable
2	https://twitter.com/search?q=lang%3Azh%20since%3A2016-08-01%20until%3A2019-08-31%20%E5%B0%8F%E8%A1%8C%E6%98%9F%20%E5%9C%B0%E7%90%83&src=typed_query	https://twitter.com/[redacted]/status/1157345689937989632	● Very Valuable ○ Somewhat Valuable
3	https://twitter.com/search?q=lang%3Azh%20since%3A2016-08-01%20until%3A2019-08-31%20%E5%B0%8F%E8%A1%8C%E6%98%9F%E6%92%9E&src=typed_query	https://twitter.com/[redacted]915907516235530241	○ Very Valuable ● Somewhat Valuable
4			○ Very Valuable ○ Somewhat Valuable

Figure 1: Interface used to capture results of interactive search during Topic Development.

Table 1: Topic development time, in minutes.

Language	Topics	Average	Median	Total
Chinese	76	37.64	15.83	2860.77
Persian	131	47.34	21.55	6201.05
All Languages	207	43.78	18.00	9061.82

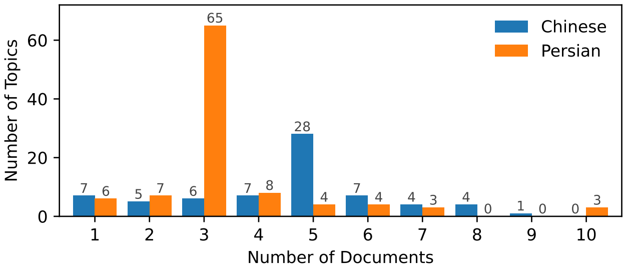


Figure 2: Relevant documents found in topic development.

select documents to be judged instead of using *pooling* [37, 43], because the collection was not built as part of a shared task that has a diverse set of systems contributing to the pools.[6] HiCAL builds a classifier based on the known relevant and non-relevant documents using relevance feedback. As the assessor judges documents, the classifier is retrained using the new assessments. To seed HiCAL's classifier, we used the documents that were identified during topic development. At least one document must be judged relevant to initialize the classifier. As Figure 2 shows, there were between one and nine relevant documents per topic from the initial assessment (mean 4.6 for Chinese, 3.4 for Persian). Because the relevance assessor is likely not the person who identified the Tweets during topic development, and because assessors tend to be more discerning during relevance assessment than during topic development, all documents were rejudged. The leftmost category in Figure 3 shows the number of topics where the second annotator asserted that no identified Tweets were relevant to the topic.

We used the same parameters as HC4 to terminate the judgment of a particular topic. Following the recommendation of the designers of HiCAL,[7] one can reasonably infer that almost all findable relevant documents have been found if an assessor judges twenty documents in a row as not relevant. If an annotator identified a relevant document after the eighty-fifth judged document, we assumed that the topic had too many relevant documents be fully judged in a reasonable amount of time. Topics in this category are identified as "Incomplete" in Figure 3. To reach our target number of topics, some failed topics were iteratively refined by having the annotator propose a new title and description. A new annotation would then begin using documents the annotator indicated that they wanted to see again during the prior round of annotation to initialize the active learning session. Thus more topics were judged during relevance assessment than were created during topic development.

As with HC4, once the assessment was complete, assessors provided a translation of the title and description fields into the language of the documents, and briefly explained (in English) how relevance judgments were made; these explanations were placed in the topic's narrative field. In contrast to the narrative in a typical TREC ad hoc collection, which is written during topic development, these narratives were written after judgments were made; users of the collection should bear in mind that the narrative field is based on known items, and so may not be useful as a model of a long ad hoc query.

We established four relevance levels, defined from the perspective of a user writing a report on the topic:

Very-valuable. a nugget of information or take on the topic that one would quote in a report to support an idea.

Somewhat-valuable. information or commentary not likely to be quoted in a report, but that nonetheless supports an idea in the report.

Not-that-valuable. information that adds no new information beyond what is written in the topic description.

Not-central. not about the topic (*i.e,.* not relevant).

[6] While we evaluate HiCAL's performance using pooling, we believe our nine baseline systems may be insufficient to create an unbiased sample of relevant documents that can fairly evaluate future systems, which is the purpose of a test collection.

[7] Personal communication with Gordon Cormack.

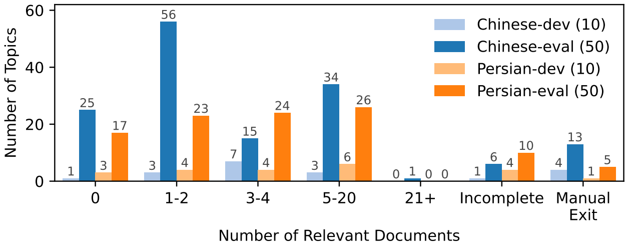


Figure 3: Distribution of the number of the relevant documents found by HiCAL.

Table 2: Collection statistics.

	Chinese		Persian	
	Dev	Eval	Dev	Eval
Documents	5,584,146		7,335,221	
Topics	10	50	10	50
Judged Documents	359	2,192	434	2,021
Partially Relevant Documents	29	212	21	81
Highly Relevant Documents	35	165	43	243

These definitions are less strict than those used in HC4, given that Tweets are less formal and might be expected to contain less useful information on a topic than one might expect of a news document.

To map graded relevance values onto the binary relevance values required by HiCAL, documents judged as ***very-valuable*** or ***somewhat-valuable*** were treated as relevant, while documents judged ***not-that-valuable*** were considered ***not-relevant***. Therefore, the final collection maps the ***not-that-valuable*** category to ***not-relevant***. This means that a document can mention a topic without being considered relevant to that topic if it lacks information beyond that found in the topic description. Because an assessor could judge a topic over multiple days, assessors took copious notes to foster consistency.

To save cost, annotators were given the ability to stop annotating a topic if it became too broad (i.e., if there were so many relevant documents that it seemed unlikely that most could be found in the available time); however, this feature was used sparingly. Such topics are listed under the category ***Manual Exit*** in Figure 3 and were likely to be revised. Most failed topics were too narrow rather than too broad, as is shown in that figure where the categories ***0*** and ***1-2*** are too narrow and the categories ***Incomplete*** and ***Manual Exit*** indicate too broad.

4 COLLECTION CONSTRUCTION DETAILS

This section provides collection details and discusses annotation costs in terms of time. Table 2 describes the size of the collection in documents and topics, and presents counts of the number of annotations used in the final collection. Disjoint subsets of Dev and Eval topics are defined to encourage consistent choices by users of the test collections. As in most information retrieval collections, the vast majority of the unjudged documents are not relevant. However, because we used active learning to suggest documents for assessment, and because of our desire to create topics with relatively few relevant documents, on average there are only about 42 judged documents per topic. This number ranges from 23 (when no additional relevant documents were discovered during the active learning phase) to 103 documents. The Dev topics can be used to tune hyperparameters. As Table 3 shows, four development topics and thirteen evaluation topics have judged documents in both languages. While we sought to maximize the number of multilingual topics, some topics are simply not discussed in both languages. The topic overlap with HC4 is somewhat higher. Close to half of the topics in each language can be found in HC4.

The judgments described in Section 3.3, were mapped to three levels in the qrels used to evaluate systems. The ***not-valuable category*** was treated as not relevant. The ***somewhat-valuable*** category was mapped to a value of 1 in the qrels and ***very-valuable*** was mapped to a value of 3.

Twelve annotators performed topic development and relevance assessment. All were at least bilingual, proficient in both English and either Chinese or Persian, and all use their bilingual skills regularly as part of their work. A majority were native English speakers, although a few were native speakers in the language of the documents that they were assessing. None were proficient in both Chinese and Persian. Topic translations were verified by people highly fluent in the target language of the translation.

4.1 Topic Development Analysis

During topic development, annotators searched live Twitter for Chinese Tweets on 76 topics and for Persian Tweets on 131 topics. After this, the document set in each language was fixed. Relevance assessment was undertaken for a topic if at least one relevant Tweet had been identified during topic development. There were seven topics in Chinese and thirty-one in Persian for which no document was found to be relevant during topic development.

A minimum of three somewhat or very valuable documents was required after relevance assessment was completed for a topic to be included in the collection. We also excluded topics in which relevance assessment did not end with a negative streak of twenty non-relevant documents, because such a negative streak gave us reasonable confidence that HiCAL could not have found many more relevant documents. In Chinese, thirty-four topics met both requirements, while in Persian forty-seven topics did. The remaining topics were iteratively refined.

We also asked assessors to respond to the prompt "I recommend that this topic be part of the collection" using a Likert scale. We treated "Disagree" and "Strongly Disagree" as negative recommendations, while the other three categories were treated as positive. While assessor recommendations for inclusion were not entirely accurate, the annotators were more reliable at identifying topics that would not be included in the collection; 85% of Chinese and 73% of Persian topics with a negative recommendation failed to meet the inclusion criteria. They were less able to identify topics that would be included, with only about 50% of topics in each language that received a positive recommendation meeting the inclusion criteria.

Table 3: Multilingual/Multi-collection topic counts.

	Chinese+Persian HC3	Chinese HC3+HC4	Persian HC3+HC4
Dev	4	8	6
Eval	13	23	21

Table 4: Median relevance judgment time in seconds.

		Relevance Level				
	Topics	0	1	2	3	All
Chinese	101	10.80	30.99	25.72	29.74	12.87
Persian	97	9.69	34.57	29.06	26.00	11.50
All Languages	146	10.32	32.93	26.09	27.42	12.25

4.2 Assessment Time

Assessing a document (most of which were single Tweets) required a median of 12 seconds. As a Tweet was shorter than 140 characters, it was straightforward for the assessors to read and form opinions on relevance. Interestingly, we only found a weak correlation between the length of the document and the assessment time (Pearson's correlation 0.10 with the number of Tweets in a document, and 0.16 with the number of characters). For long documents, the assessors can confidently judge the document by skimming through it if it is clearly not relevant to the topic.

Relevant documents took longer to judge; it can sometimes be difficult to distinguish the boundary around not-that valuable, whether deciding it contains somewhat-valuable information or is not relevant. For documents that are clearly relevant, assessors must determine the degree of relevance, which typically takes more time to judge than does recognizing a non-relevant document.

5 BASELINES

This section demonstrates the utility of the HC3 test collection to distinguish the effectiveness of retrieval systems. We evaluate a variety of CLIR systems on the Eval topics, including sparse retrieval models that use translation and state-of-the-art end-to-end neural retrieval systems (e.g., ColBERT-X [26]).

5.1 Baseline Systems

BM25 is an unsupervised monolingual sparse retrieval model that can perform CLIR with translations. In this work, we include four translation strategies to cover a wide spectrum of applications, including document translation by a Sockeye 2-based MT model (DMT), query translation by the same MT model (QMT), query translation by Google Translate (QGT), and query translation by a human (QHT). The MT models were trained on parallel sentences in general domains, similar to the ones used in Nair et al. [26] and Lawrie et al. [18], along with additional 2000 Tweets collected by McNamee and Duh [24].[8] Since the Tweets used for training the MT models were collected independently from our collection, there may be Tweets used in both training the MT and evaluation in our retrieval collection. The baseline BM25 runs use the default parameter values $b = 0.9$ and $k_1 = 0.7$, which were tuned on newswire collections [31]. Since the parameters were not tuned for short documents such as Tweets, the effectiveness of these BM25 runs may be improvable with other parameter values. Patapsco [8] was used to complete these runs.

Probabilistic Structured Query (PSQ) [9] a sparse CLIR method that uses a statistical translation table to translate document terms into multiple hypotheses in the query language, allocating the term frequency in a document in proportion to the translation probability. The implementation used in this work maps documents into the query language vocabulary at indexing time using the translation table. We use an HMM as the matching function, as it has been shown to work well with PSQ [38].

We also evaluate several dense supervised neural end-to-end models, including mContriever [14], DPR-X [40, 41], SPLADE-X [27] and ColBERT-X [26]. These systems index dense representations of the Tweets and directly search the Chinese or Persian Tweets with English queries. For mContriever, we use the publicly available model with supervised fine-tuning on MS-MARCO.[9] For DPR-X and ColBERT-X, we fine-tuned the pretrained XLM-RoBERTa Large models with translated MS-MARCO training triples (English queries and translated passages) for 200,000 steps with a batch size of 64 and a learning rate of 5×10^{-6}. Note that ColBERT-X indexing and search are implemented with PLAID [32], which is a faster implementation than the original ColBERT that does not impede effectiveness. For SPLADE-X, we fine-tuned the multilingual BERT [10] model with the same training scheme. URLs are removed from documents for the neural models to ensure that pieces of URLs do not mislead the contextual embeddings.

5.2 Results

Table 5 summarizes the effectiveness of the baseline retrieval models using the title queries, which consist of three to five English words.[10] To support retrieval tasks with longer queries, we also provide retrieval effectiveness in Table 6 using both description queries, which are English sentences describing the information need, and the concatenation of title and description queries. Such longer queries are useful for retrieval systems based on large, generative, prompt-based language models such as GPT-3 [3].

BM25 with document machine translation (DMT) provides a strong baseline, but its effectiveness is subject to the quality of the MT model [26]. As the MT models were trained with more formal text (e.g., news articles and subtitles), translating Tweets poses a challenge for the model. We demonstrate the impact of the quality of MT models on retrieval effectiveness in Table 7. The first two systems were only fine-tuned with typical MT parallel corpora (M2 is a larger Sockeye model), such as subtitles, while the model fine-tuned with Tweets (the one reported in Tables 5 and 6) provides better Tweet translations and thus better retrieval effectiveness. Translating the queries leads to lower effectiveness than translating the Tweets, probably because translation systems are better tuned to the way language is used in documents than in queries. Google and

[8] https://pmcnamee.net/research/mmtc/mmtc.html

[9] https://huggingface.co/facebook/mcontriever-msmarco

[10] Scores assigned to each document by each system for each query variant will be released along with the collection to enable future comparisons and for reproducibility.

Table 5: Retrieval effectiveness of baseline CLIR systems using title queries.

		Chinese					Persian				
Model	Translation	nDCG@20	Judged@20	MAP	R@100	R@1k	nDCG@20	Judged@20	AP	R@100	R@1k
Sparse Models with Translation (Unsupervised)											
BM25	DMT	0.212	0.170	0.168	0.452	0.685	0.229	0.131	0.173	0.423	0.632
	QMT	0.191	0.114	0.130	0.418	0.617	0.211	0.141	0.162	0.379	0.610
	QGT	0.245	0.213	0.194	0.556	0.752	0.294	0.193	0.226	0.490	0.667
	QHT	0.237	0.195	0.176	0.535	0.755	0.302	0.211	0.237	0.474	0.630
HMM	PSQ	0.298	0.201	0.238	0.553	0.760	0.383	0.197	0.309	0.603	0.751
End-to-End Neural CLIR Models (Supervised)											
mContriever	–	0.234	0.137	0.158	0.471	0.677	0.121	0.060	0.080	0.287	0.529
SPLADE-X	–	0.260	0.189	0.195	0.568	0.738	0.322	0.166	0.235	0.519	0.710
DPR-X	–	0.323	0.199	0.237	0.578	0.772	0.342	0.164	0.247	0.499	0.745
ColBERT-X	–	0.359	0.206	0.280	0.601	0.779	0.378	0.182	0.291	0.578	0.728

Table 6: Retrieval effectiveness using description(D) and title+description(T+D) queries.

		nDCG@20		R@100	
Model	Translation	D	T+D	D	T+D
Chinese					
BM25	DMT	0.236	0.262	0.432	0.498
BM25	QMT	0.220	0.261	0.450	0.537
BM25	QGT	0.232	0.287	0.455	0.544
BM25	QHT	0.258	0.259	0.513	0.568
HMM	PSQ	0.309	0.361	0.545	0.602
mContriever	–	0.254	0.252	0.497	0.491
SPLADE-X	–	0.322	0.328	0.584	0.595
DPR-X	–	0.367	0.376	0.621	0.649
ColBERT-X	–	0.390	0.397	0.610	0.631
Persian					
BM25	DMT	0.247	0.261	0.394	0.486
BM25	QMT	0.187	0.228	0.303	0.404
BM25	QGT	0.261	0.319	0.368	0.476
BM25	QHT	0.265	0.329	0.414	0.492
HMM	PSQ	0.369	0.410	0.567	0.641
mContriever	–	0.182	0.172	0.323	0.330
SPLADE-X	–	0.348	0.348	0.585	0.580
DPR-X	–	0.399	0.400	0.596	0.598
ColBERT-X	–	0.409	0.400	0.598	0.594

Table 7: BM25 searching translated documents using different MT systems with title queries.

	BLEU	nDCG@20	Judged@20	R@100
M1	24.8	0.146	0.142	0.350
M2	28.0	0.190	0.158	0.391
M2 + 2k Tweets	33.0	0.212	0.170	0.452

human translation of the queries demonstrate strong performance at both the upper (nDCG@20) and lower (R@1000) portions of the ranked results.

Beyond mapping each token to one translation in the target language, PSQ translates each document token into multiple English tokens at indexing time, which helps to alleviate vocabulary mismatches; this leads to stronger performance than any machine translation approach. For longer queries such as descriptions, PSQ expands the documents to more potentially related terms, which leads to a larger performance gain than other systems.

Dense neural retrieval models fine-tuned on translated MS MARCO [26, 27] exhibit stronger effectiveness. The performance of mContriever is lower because it was only fine-tuned on English MS MARCO, not on translations. Interestingly, PSQ outperforms the neural systems in Persian by both nDCG@20 and in recall. However, DPR-X and ColBERT-X are more effective than PSQ in Chinese. This is likely due to better MT models for translating MS MARCO documents into Chinese than into Persian, and to the monolingual embedding quality of high-resource Chinese vs. medium-resource Persian.

SPLADE-X, which uses the masked language modeling (MLM) head to predict English tokens corresponding to foreign language Tweets is worse than PSQ in both Persian and Chinese. As SPLADE-X is essentially a token-level translation model using the MLM head, with a result format similar to that of PSQ, training PSQ with parallel data is the better approach.

While the percentage judged in the top twenty retrieved documents may seem low, the utility of an evaluation collection depends on whether the set of judged documents, especially the relevant ones, is systematically biased toward certain systems. Figure 4 shows that the documents retrieved by each system are diverse; that is, there is low overlap between systems. No system has a substantially higher rate of retrieving judged documents, indicating that the judgments are fair to all of these systems. In the next section, we conduct a deeper analysis of this claim by comparing our relevance assessments acquired through HiCAL to those obtained from the standard pooling approach over our baseline systems.

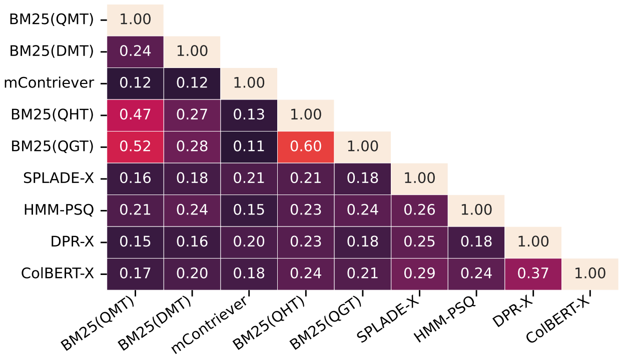


Figure 4: Percentage of overlapping top 20 retrieved documents of systems searching with title queries on the Chinese collection. Systems are ordered by nDCG@20 values.

Overall, HC3 poses challenges for retrieval systems that rely on machine translation. Such challenges reward the modeling of relationships between the query and the Tweet conversations that go beyond token matching. In other words, evaluating systems on HC3 pushes the development of CLIR beyond simply improving machine translation to cross the language barrier.

6 POOLING ASSESSMENTS

To assess the utility of the relevance judgments created by HiCAL, we sampled 25 Chinese topics. For each, we created pools to provide complete judgments down to rank 20 from the baseline systems, searching using title queries. We want to assess whether the set of judgments acquired through HiCAL is systematically biased toward any system. Note that these additional judgments are designed to evaluate the HiCAL assessment process, not for evaluating retrieval systems. These additional judgments may be biased towards the baseline systems, but they enable concrete comparisons of these systems using the nDCG@20 measure.

We employed six native Chinese speakers, each assessing four to five topics.[11] Assessors were tasked to read topics in English and to assess document relevance based on the criteria described in Section 3.3. Since the assessors did not develop the topics themselves, the search intent of the assessors may differ from those of the original topic developers. We provided as much guidance as possible to mitigate such differences. Each pool, on average, has 98.32 Tweets, varying from 48 to 164. Most documents are shorter than 800 characters,[12] with 23 exceptions that are longer conversations.

We evaluate the baseline systems with the new relevance judgments on these 25 topics and summarize the results in Table 8. For comparison, we also report average scores using HiCAL judgments over only the 25 sampled topics; this leads to different scores than shown in Table 5 and a slightly different system ordering.

Although the relative system rankings over the sample are not identical, the correlation between scores calculated based on the two sets of judgments is strong; this is seen both by measuring the Pearson Rank correlation [12] and by Spearman's ρ. The former is head-weighted (giving more weight to swaps between better systems) and gap-sensitive (giving less weight to swaps between similar-scoring systems), while the latter treats all swaps equally. Most ordering differences are swaps between systems with similar nDCG@20 values when evaluated against the HiCAL judgments, such as BM25 with QHT and SPLADE-X ($p = 0.87$ by a paired t-test). PSQ exhibits the largest ordering change, dropping from fourth to seventh. However, since the difference between PSQ and BM25-DMT (ranked seventh with HiCAL judgments) is not statistically significant ($p = 0.34$ by a paired t-test), retrieval results from the two systems are not statistically distinguishable.

Table 8: nDCG@20 and the system ranks using judgments by HiCAL and Pooling. Pearson Rank correlation coefficient [12] is 0.851. Spearman's ρ rank correlation coefficient is 0.800.

Model	Translation	HiCAL		Pooling	
BM25	QMT	(9)	0.137	(8)	0.247
mContriever	–	(8)	0.175	(9)	0.239
BM25	DMT	(7)	0.202	(6)	0.314
BM25	QGT	(6)	0.237	(4)	0.360
BM25	QHT	(5)	0.249	(3)	0.438
HMM	PSQ	(4)	0.253	(7)	0.263
SPLADE-X	–	(3)	0.256	(5)	0.328
DPR-X	–	(2)	0.341	(2)	0.462
ColBERT-X	–	(1)	0.376	(1)	0.463

Table 9: Average number of documents in each relevance level over the sampled 25 Chinese pooling topics over qrels' values.

	# Judged	Other (0)	Somewhat Valuable (1)	Very Valuable (3)
HiCAL	1164	38.00	5.84	2.72
Pooling	2458	78.80	14.84	7.68

Table 9 shows the average number of documents at each relevance level over the 25 sampled topics. The pooling annotators marked a greater number of documents as relevant. As a result, the absolute system scores are higher with the pooling assessors than with the HiCAL assessors in every case. They also marked about half of the documents that had been judged as relevant by the HiCAL assessors as not relevant. However, relevance is an opinion, not a fact, and it is not unusual for assessors to have different opinions. As Voorhees has noted, the key question in information retrieval evaluation is not whether two assessors agree on the relevance of specific documents, but rather whether system comparisons remain stable, whichever set of judgments are used [34].

As Figure 5 shows, the number of documents marked as relevant by rank 20 varied markedly by topic with both HiCAL and pooling, as is expected, but topics with the greatest number of relevant documents differ between the two methods (in the figure, the topics are

[11] These assessors were Taiwanese, and thus may be biased toward Tweets written in Traditional Chinese characters. Twitter is popular in Taiwan but not widely used in mainland China, so this is a good match for the content. Moreover, all six assessors can read Simplified Chinese at a near-native level and are proficient in English.
[12] Counting English, Chinese, and emoji characters.

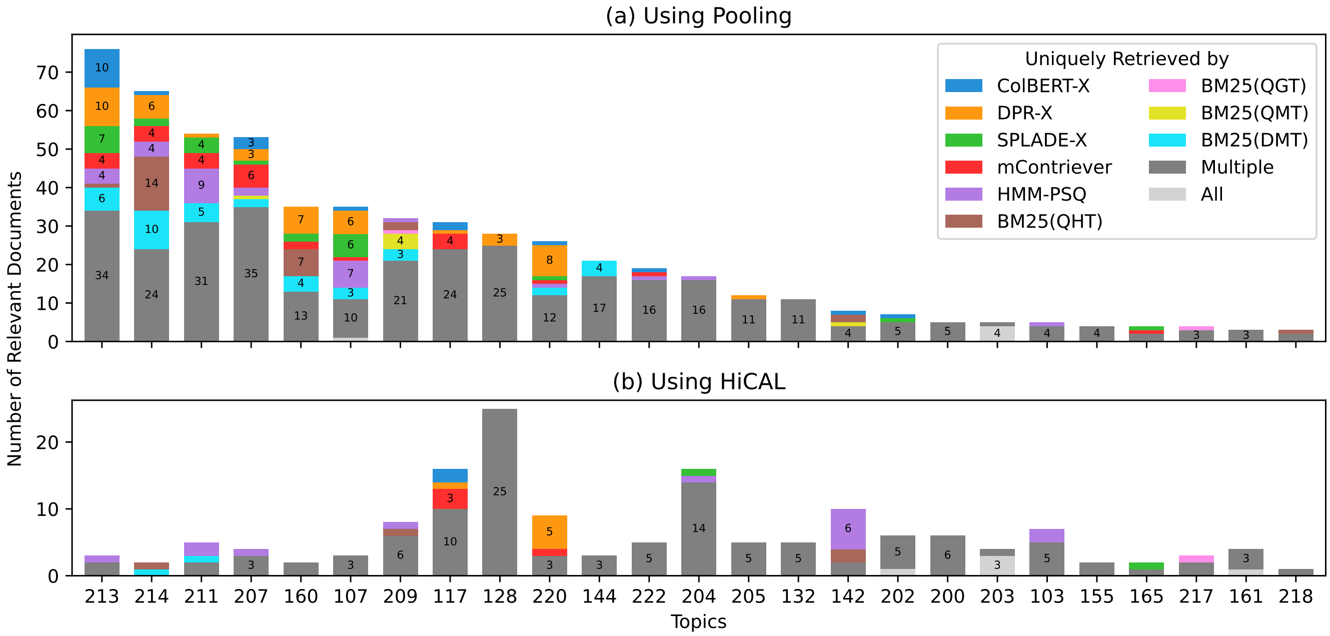


Figure 5: Number of relevant documents uniquely retrieved by each system at rank 20. Documents retrieved by more than one but not all systems are labeled as "Multiple"; documents all systems have retrieved are labeled as "All."

consistently sorted for both methods in largest-first order according to the pooling assessors). Despite these differences, no system was measurably benefited or penalized with one or the other set of judgments. From this we conclude that the relevance judgments acquired through HiCAL are fair to each of the baselines we have implemented.

7 FUTURE USE

A challenge with releasing an IR collection over Tweets is the limitations that Twitter imposes when sharing the collection.[13] At the time of this writing, academic researchers are allowed to share an unlimited number of Tweet IDs for non-commercial research purposes; however, we are not allowed to share the content of the Tweets. Each institution wishing to perform experiments with this collection must acquire permission to use the Twitter version 2 API. The HC3 repository describes each document in the collection as a list of Tweet IDs, with their language as identified by Twitter when the collection was constructed. A document consists of all the Tweets in a conversation thread in the language of the particular document set. We release software that will download available Tweets and assemble the document collection. Because Tweets become unavailable over time, the contents of the documents are checked against a hash code that is also released for each document. Any document whose hash changes will be removed from the collection.

This means that scores across experimenters may become incompatible as documents retrieved by baseline systems become unavailable. To mitigate this problem and allow future researchers to compare against baselines derived from those in this paper, we release the run files of all baseline runs and a post hoc evaluation script. When provided with a run file and either a document file or a list of document ids, the script will produce a new run file with any missing documents removed. It also creates a new qrels file with unavailable documents removed. These script outputs can be used to accurately compare baseline runs to runs produced by new research systems, in spite of changes in the underlying document collection. We encourage future users of the collections to publish their run files as well, so that comparisons can continue to be made over time.

8 CONCLUSION

We have developed a new CLIR test collection for short informal text, using English queries to rank Tweet conversations in Chinese and Persian. This new HC3 test collection is a companion to the previously released HC4 test collection, allowing experiments with some topics that are common to both Twitter and news sources. Using a set of baseline systems, we have shown that system comparisons based on relevance judgments of documents selected by HiCAL, an active learning approach, are comparable to system comparisons based on relevance judgments of documents selected using pooling. Building test collections using Twitter poses challenges that are not present in more traditional CLIR content types such as news, including how best to define a document, and how to define baseline results in a way that can accommodate future content deletion. We have presented solutions to both challenges that we believe may also be useful more generally for evaluation of short informal text retrieval, even in a single language.

[13] https://developer.twitter.com/en/developer-terms/agreement-and-policy

REFERENCES

[1] Martin Braschler. 2001. CLEF 2001—Overview of Results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 9–26.

[2] Martin Braschler. 2003. CLEF 2003–Overview of results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 44–63.

[3] 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[4] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. 2007. Bias and the limits of pooling for large collections. *Information retrieval* 10 (2007), 491–508.

[5] Gordon V Cormack, Christopher R Palmer, and Charles LA Clarke. 1998. Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 282–289.

[6] Gordon V Cormack, Haotian Zhang, Nimesh Ghelani, Mustafa Abualsaud, Mark D Smucker, Maura R Grossman, Shahin Rahbariasl, and Amira Ghenai. 2019. Dynamic sampling meets pooling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1217–1220.

[7] Gordon V. Cormack, Haotian Zhang, Nimesh Ghelani, Mustafa Abualsaud, Mark D. Smucker, Maura R. Grossman, Shahin Rahbariasl, and Amira Ghenai. 2019. Dynamic Sampling Meets Pooling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 1217–1220. https://doi.org/10.1145/3331184.3331354

[8] Cash Costello, Eugene Yang, Dawn Lawrie, and James Mayfield. 2022. Patapasco: A Python Framework for Cross-Language Information Retrieval Experiments. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR)*.

[9] Kareem Darwish and Douglas W. Oard. 2003. Probabilistic Structured Query Methods. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* (Toronto, Canada) *(SIGIR '03)*. Association for Computing Machinery, New York, NY, USA, 338–344. https://doi.org/10.1145/860435.860497

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Nicola Ferro and Carol Peters. 2009. CLEF 2009 Ad Hoc Track Overview: TEL and Persian Tasks. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 6241)*, Carol Peters, Giorgio Maria Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas, and Giovanna Roda (Eds.). Springer, 13–35. https://doi.org/10.1007/978-3-642-15754-7_2

[12] Ning Gao, Mossaab Bagdouri, and Douglas W Oard. 2016. Pearson rank: a head-weighted gap-sensitive score-based correlation coefficient. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 941–944.

[13] Donna K. Harman. 2005. The TREC Test Collections. In *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*, Ellen M Voorhees and Donna K Harman (Eds.). The MIT Press.

[14] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118* (2021).

[15] Noriko Kando, Kazuko Kuriyama, and Masaharu Yoshioka. 2001. Overview of Japanese and English Imformation Retrieval Tasks (JEIR) at the Second NTCIR Workshop.. In *NTCIR*.

[16] Kazuaki Kishida and Makoto P Kato. 2016. Overview of NTCIR-12.. In *NTCIR*.

[17] Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldanini, and Eugene Yang. 2022. Overview of the TREC 2022 NeuCLIR Track. In *The Thirty-first Text REtrieval Conference (TREC 2022) Proceedings*.

[18] Dawn Lawrie, James Mayfield, Douglas W. Oard, and Eugene Yang. 2022. HC4: A New Suite of Test Collections for Ad Hoc CLIR. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR)*.

[19] David D Lewis and Kimberly A Knowles. 1997. Threading electronic mail: A preliminary study. *Information processing & management* 33, 2 (1997), 209–217.

[20] Jimmy Lin, Miles Efron, Yulu Wang, and Garrick Sherman. 2014. Overview of the TREC-2014 microblog track. In *TREC*.

[21] Mihai Lupu, Atsushi Fujii, Douglas W Oard, Makoto Iwayama, and Noriko Kando. 2017. Patent-related tasks at NTCIR. *Current Challenges in Patent Information Retrieval* (2017), 77–111.

[22] Craig Macdonald, Rodrygo LT Santos, Iadh Ounis, and Ian Soboroff. 2010. Blog track research at TREC. In *ACM SIGIR Forum*, Vol. 44. ACM New York, NY, USA, 58–75.

[23] Prasenjit Majumder, Mandar Mitra, Dipasree Pal, Ayan Bandyopadhyay, Samaresh Maiti, Sukomal Pal, Deboshree Modak, and Sucharita Sanyal. 2010. The FIRE 2008 Evaluation Exercise. *ACM Transactions on Asian Language Information Processing* 9, 3 (2010), 10:1–10:24. https://doi.org/10.1145/1838745.1838747

[24] Paul McNamee and Kevin Duh. 2022. The multilingual microblog translation corpus: Improving and evaluating translation of user-generated text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 910–918.

[25] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 1–27.

[26] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas Oard. 2022. Transfer Learning Approaches for Building Cross-Language Dense Retrieval Models. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR)*.

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

[28] Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. ClueWeb22: 10 Billion Web Documents with Rich Information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3360–3362.

[29] Carol Peters, Martin Braschler, and Paul Clough. 2012. *Multilingual information retrieval: From research to practice*. Springer.

[30] Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, and William R Hersh. 2016. Overview of the TREC 2016 clinical decision support track.. In *TREC*.

[31] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[32] Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. 2022. PLAID: an efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1747–1756.

[33] Lifeng Shang, Tetsuya Sakai, Zhengdong Lu, Hang Li, Ryuichiro Higashinaka, and Yusuke Miyao. 2012. Overview of the NTCIR-12 Short Text Conversation Task. (2012).

[34] Ellen M Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management* 36, 5 (2000), 697–716.

[35] Ellen M Voorhees. 2019. The evolution of Cranfield. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF* (2019), 45–69.

[36] Ellen M Voorhees and Donna Harman. 2003. Overview of TREC 2003. *NIST special publication SP* 251 (2003), 1–16.

[37] William Webber, Alistair Moffat, and Justin Zobel. 2010. The Effect of Pooling and Evaluation Depth on Metric Stability.. In *EVIA@ NTCIR*. 7–15.

[38] Jinxi Xu and Ralph Weischedel. 2000. Cross-lingual information retrieval using hidden Markov models. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. 95–103.

[39] Zhenzhen Xue, Dawei Yin, and Brian D Davison. 2011. Normalizing microtext. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[40] 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

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

[42] 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 preprint arXiv:2210.09984* (2022).

[43] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 307–314.