Cross-language Information Retrieval

Petra Galuščáková[1]* Douglas W. Oard[2], Suraj Nair[2]

[1]Université Grenoble Alpes, Grenoble, France
[2]University of Maryland, College Park, United States
galuscakova@gmail.com

1 Introduction

Two key assumptions shape the usual view of ranked retrieval: (1) that the searcher can choose words for their query that might appear in the documents that they wish to see, and (2) that ranking retrieved documents will suffice because the searcher will be able to recognize those which they wished to find. When the documents to be searched are in a language not known by the searcher, neither assumption is true. In such cases, Cross-Language Information Retrieval (CLIR) is needed. This chapter reviews the state of the art for CLIR and outlines some open research questions.

The goal of Cross-Language Information Retrieval[1] is to build search engines that use a query expressed in one language (e.g., English) to find content that is expressed in some other language (e.g., French). CLIR and Machine Translation (MT, automatically rendering the content of a document in another language) are closely linked; what we call MT is the use of translation technology to render documents readable, whereas CLIR is the use of translation technology to render documents searchable. So CLIR and MT are both built on the same foundation: *translation technology*. While our focus in this chapter is on retrieval of documents in one language using queries in another language, in practical applications the collection to be searched can potentially contain documents in many different languages, or even documents that themselves contain mixed languages, and the query too may contain terms from different languages.[2]

---

* The author was affiliated with the University of Maryland at the time of writing.

[1] The astute reader might ask why the field is not referred to as "cross-lingual" information retrieval. Participants in the SIGIR 1996 workshop voted on a name for the field, and cross-language won. Language is, however, created through use, and cross-lingual is today also a commonly used form.

[2] Sometimes such cases have been referred to as Multilingual Information Retrieval (MLIR) [153], although that terminology is inconsistently used. For example, many authors have used MLIR to refer to information retrieval systems such as Google that are able to handle many languages, even when no support for cross-language search is provided.

1.1   Some Use Cases

The use cases for CLIR systems can be divided into two groups: (1) cases in which the user lacks proficiency in the language of the documents, and (2) cases in which the user can understand the documents, but prefers to use a different language for their query. These use cases, together with other factors such as relatedness of the languages, frequency of their interaction, collection size and format, or the expected number of queries, then influence the design choices which need to be considered when building a CLIR system. Here we describe several scenarios in which CLIR systems can be useful.

– In Web search, the natural emergence of one or more "lingua franca" languages such as English or Chinese results in an asymmetry in which much of the information is in some small number of languages, but many users prefer to (or must) interact with the Web in some other language. In such cases, CLIR can provide access to information that would otherwise be unavailable.
– Recall-oriented CLIR applications, which include searching in medical documents [76, 82, 84, 150, 161, 170], patents [118] and security applications [34]. The end users of such systems are often professionals for whom finding all documents on a defined topic is a high priority. It can be expected that the end users in such settings might have the ability to pay for high-quality human translation of the relevant documents. Even in such cases, however, automatic translations might provide a fast overview of the search results and they might also be helpful for deciding if individual documents require human translation. In such cases, a summary of a document in the language of the query might be useful.
– CLIR applications can be expected to be particularly important in regions where multiple languages are frequently used, such as in countries that have more than one official language. This might include English/French in Canada, Dutch/French/German in Belgium, Spanish/Catalan/Basque in Spain, the 24 official languages in the European Union, or the 23 official languages in India. It can be expected that substantial amounts of translated text would often be available in such settings, either through commercial practice or to meet legal requirements. Users of CLIR systems in such settings might be able to understand multiple languages, but they may be more fluent in, or more comfortable using, some of those languages.
– Question answering (QA) is a special case of Information Retrieval (IR) in which rather than simply identifying relevant documents to the user, the system seeks to use information from relevant documents to provide a direct answer to a question. Cross-Language Question Answering (CLQA) has received considerable attention among researchers [37, 61, 116, 163] and with the recent attention to conversational search and chatbots it might be expected that this topic would attract continued attention in the coming years.

1.2   The Three Waves of Information Retrieval

We live today in the third wave of research on automated approaches to Information Retrieval (IR). In the first wave, documents were typically described

by controlled vocabulary descriptors (which is one example of what we today call metadata) and queries were expressed using the same descriptors. These descriptors were typically organized hierarchically (e.g., *aircraft* is a narrower term than *vehicle*) in a thesaurus, and cross-language search could thus be supported by adding *entry vocabulary* in other languages (e.g., by linking the French free text term *avion* to the controlled vocabulary term *aircraft*) [144]. The driving technology of that first wave was disk storage, and the driving need was to index large collections of published materials in libraries, and large collections of unpublished technical documents; the American Documentation Institute was an early venue for research on thesaurus-based access to large collections.[3]

In second wave IR (1990-today), documents are typically represented using the terms they contain (i.e., full-text indexing), searchers are no longer limited to thesaurus terms in their queries and instead can use any terms they wish, and retrieved documents are ranked based on the degree to which they are estimated by the system to satisfy the query. A driving force behind the emergence of the second wave was the need to provide search engines for the World-Wide Web, and the ACM Special Interest Group on Information Retrieval (SIGIR) emerged as a focal community for that work.

In second-wave IR research, the usual approach was to represent queries and documents similarly, and then to compare the queries to the documents. Examples of this included vector space and query likelihood models. Recently, however, a third wave has emerged, with different representations for queries and for documents. The defining technology for this new approach is the neural transformer architecture, through which query and document terms learn to attend to each other.

Each of these waves of IR research has shaped a generation of CLIR research. Our goal in this chapter is to present a concise view of CLIR as seen from the third decade of the twenty-first century. It is not possible to review every detail, but neither is that necessary, since several useful overviews already exist. Particularly notable in that regard is Nie's excellent survey, penned a decade ago [138]. Here, therefore, we pay particular attention to influential work published since Nie's 2010 survey. Along the way, our presentation supplements four other surveys that have been published since Nie's survey [44, 80, 216, 220].

2   The Core Technology of CLIR

CLIR research came of age in the second wave, when the key question was how queries and documents that were expressed using different languages could be represented similarly. Because that work still shapes our view of CLIR, we first introduce how that was done. We organize our story around three major questions: (1) 'what to translate?', (2) 'which translations are possible?', and (3) 'how should those translations be used?'

---

[3] Today's successor to the American Documentation Institute is the Association for Information Science and Technology (ASIS&T).

2.1   What to Translate?

For CLIR, the question of what to translate exists at two levels. First we can ask which items should be translated—the queries or the documents? With that question answered, we must then ask how those documents or queries can be broken up into units (which we call terms) that can be translated. First things first, we start with the question of whether it is the query or the documents that should be translated.

The Query or The Documents?

Perhaps it seems obvious that it is the query that should be translated, since queries are short and documents are long. Indeed, query translation is widely used in CLIR experiments, specifically because it is efficient. But efficiency considerations may come out differently when the query workload is very high, as is the case in Web search.[4] On the other hand, query translation has clear advantages when the goal is to cull through streaming content for documents that satisfy a standing query. The point here is that the conventional wisdom that query translation is more efficient may or may not be true—much depends on the details of the task.

However, document translation has another potential advantage beyond the (relatively rare) cases in which it is the more efficient choice. The words in a document occur in sequence, and we have good computational models for leveraging such sequences to make more accurate translations. Of course, some queries include word sequences as well, but word sequences in queries are often very short, and thus possibly less informative. Modern translation models are trained using paired sentences from paired documents, and pairings of documents with their translations are much more widely available than examples of pairings of queries with their translations. If people wrote queries like they wrote documents, this would not be a problem. But they don't. As a result, even when query translation would be more efficient, document translation might be more accurate.

One further factor to consider when choosing between query translation and document translation is which direction yields the simplest index structure. Query translation has advantages for applications in which there are many possible query languages, but only one document language. In such cases, space efficiency argues for indexing in the document language, and thus translating the queries. Symmetrically, document translation has advantages when all the queries are all in one language, but there are many document languages to be searched. In such cases, space efficiency argues for indexing in the one query language.

Translation direction might moreover influence the translation quality. Translating into morphologically rich languages usually performs worse than translating into languages with simpler morphology, such as English. Translating from

---

[4] The indexed Web contains 5.6 billion pages, but Google alone receives 5.6 billion queries per day. At that rate, it doesn't take many days for the number of query words to become larger than the number of indexed document words.

a language with no given word boundaries, such as Chinese or Japanese, might be harder than translating into such language. Translation direction thus might also have an effect on CLIR. Another obvious advantage to document translation is that if the translations produced at indexing time are designed to be read by people (not all translations used in CLIR are) then readers who are unable to read retrieved documents in their original language can be shown cached translations. So the picture is complex, with several factors to be considered when choosing between query translation and document translation. But those are not the only two choices—there are two others.

The third choice is to translate both the queries *and* the documents. For example, if we want to use Navajo queries to search Burmese content we might translate both the Navajo and Burmese into English, simply because there are more language resources for Navajo-English and for Burmese-English than there are for Navajo-Burmese. If we broaden our concept of translation to include any mapping of an item (e.g., a word or a character sequence) from a representation specific to one language to some representation that is not specific to that language, then we can think of bilingual embedding (e.g., [107]) as a form of translation. With bilingual or multilingual embeddings, addressed in detail in Section 3.2, we map from language-specific representations (in this case, sparse vectors that indicate the presence of specific terms in a specific language) to distributed representations (i.e., dense vectors) that can be thought of as language-independent representations of meaning. Here we have in some sense translated both the query and the documents; we have just translated them into an abstract representation rather than a representation associated with any single language.

As long as we're being exhaustive here, we should note that doing no translation at all is the fourth possibility. When character sets are the same, sometimes proper names or loan words will match. Although this is almost never a good idea on its own (with the possible exception of very close dialect differences such as British and American English), the presence of such felicitous "translations" can serve as useful anchor points around which more sophisticated and capable approaches can be constructed. For example, McNamee and Mayfield used blind relevance feedback to enrich queries with related terms, thus producing a surprisingly good baseline for some language pairs to which other techniques could be compared [127].

Finally, we should note that when suggesting translating the query or translating the documents as possibilities, one need not limit themselves to just translating once. Translation activity follows cultural contact patterns, and translation systems are most easily built for the language pairs in which there already is substantial translation activity. For language pairs for which there is little cultural contact (e.g., Navajo and Burmese, as in our earlier example), it can be useful to translate through a "pivot" language (in this case, perhaps English) that has cultural contact with both. For example, to use Navajo queries to search Burmese content with query translation, we might first translate the query from Navajo to English and then from English to Burmese. As Ballesteros

and Sanderson [14] have shown, applying query expansion each time translation is performed can serve to limit the adverse effect of cascading errors.

Which Terms?

When we think of translation, we often think about translating words. But the concept of a word is actually quite slippery. Is *poet* a word? If so, is *poets* a different word, or is it something else—a plural *form* of a word? If we think of words as abstract objects that have forms, then *poet's* and *poets'* and even *poet* are yet more forms of the same word, and *poet* is its *root form*. But information retrieval people have a much simpler view of the world. We don't call what we use in our searches words, but rather *terms*,[5] and a term is simply whatever we choose to index in an IR system. This might be a token (in which case we need to say what choices our tokenizer makes!), it might be a stem (the part of the word which is shared across the inflected variants of this word) [160], or it might simply be a character sequence (which we call character n-grams).

Many other such choices are also made when choosing terms. For example, terms can be in their original case or lowercased, diacritic marks can be retained or removed; substitutable characters can be retained as-is or normalized, and stopwords can be removed or retained. The choices made on each of these points often differ from those used when human-readable translation is the goal, because information retrieval tasks (and thus information retrieval evaluation measures) often place greater emphasis on recall (i.e., on minimizing false negatives) than do machine translation tasks in which human readability is the goal.

None of this is specific to CLIR; these same issues arise in any IR application. What is specific to CLIR, however, is that these choices affect not just the general retrieval approach, but also the translation functions that are indeed specific to CLIR. In the simplest approaches to CLIR, we simply translate terms in one language into terms in the other language. We can pay attention to word order (which often differs between languages) if our retrieval system leverages proximity or word order (or both), but often it may suffice to simply get the right terms in the other language. So, for example, we could even translate from character n-grams in one language to character n-grams in the other [126].

It is worth noting that we have used tokens, stems and character n-grams as examples, but there's more to making terms than just those three options. For example, to our way of thinking in information retrieval, the statistical phrases used in Statistical Machine Translation (SMT) are also terms. So in CLIR, terms are not just what you index; they are also what you translate. Note, however, that this introduces a tension, since longer terms limit translation ambiguity, whereas shorter terms offer more scope for improving recall.

---

[5] The one exception is the so-called "bag of words" representation, which is of course in actuality a bag (in the set theory context) of terms.

2.2   Which Translations are Possible?

Once we choose which terms we need to be able to translate, our next questions are which translations are possible for each term, and (extending that further) how likely is it that each possible translation reflects the intended meaning? There are at least four kinds of places we might look for the answer to that question:

1. We might look translations up in some manually built lexicon. In CLIR, this is referred to as dictionary-based translation (regardless of whether the lexicon was actually created from a human-readable dictionary). Most often, this technique is applied to queries, and thus referred to as Dictionary-Based Query Translation (DBQT). In practice, there are many kinds of manually built lexicons that might be mined; among the most common are bilingual phrase lists, bilingual dictionaries (which contain not just term mappings but also definitions, parts of speech, and examples of usage), multilingual thesauri (thus reusing technology from the first wave), or multilingual ontologies (e.g., titles of pages in different languages on the same topic mined from Wikipedia). Using actual bilingual dictionaries would make it possible to use part of speech tags as constraints, but accurate part of speech tags can be hard to generate automatically for short queries so this constraint has proven to be of less use than might be expected. Some lexicons order alternative translations in decreasing order of general use, and that ordering can be exploited to avoid overweighting uncommon translations.

2. We might estimate translation probabilities from observed language use in one of three ways:
   - By far the most successful way of doing this in CLIR has been to mine parallel text (i.e., translation-equivalent passages; usually sentences) to learn translation mappings that include a probability for each translation. This can be done explicitly using Statistical Machine Translation (SMT) models [203] or Neural Machine Translation (NMT) models [215], or it can be done implicitly using a complete SMT or NMT system that produces one-best or n-best results. Of course, the relative prevalence of translations in the parallel text that is used to learn the probabilities should be similar to that expected in the collection to be searched; learning from translations of tractor manuals might not be wise when the goal is to search poetry.
   - An alternative source would be comparable corpora (i.e., collections of documents on similar topics that are not translations of each other), which can be used to learn plausible translations. When links between pairs of documents on the same language are available, these mappings can either be learned explicitly [179] or implicitly [112, 164]. Indeed, it is even possible to learn implicit mappings with no links between document pairs at all by leveraging regularities in usage that are preserved across languages [29]. These comparable corpus techniques are limited by corpus size, however, since orders of magnitude larger corpora are needed for comparable than for parallel corpora to get similarly good

results for relatively rare words. This issue of modeling the translations of rare words well can be particularly important for CLIR because in many settings relatively rare words are relatively common in queries.

– Finally, it is also possible to mine collections of mixed-language documents for translations. One approach to this is to focus on inline definitions that explain the meaning of a term in one language using terms in another language [101]. Such an approach to writing is used to introduce new technical terms in some settings.

3. We might also generate (or implicitly recognize) "translations" using an orthographic or phonetic transcription algorithm. Buckley, writing somewhat tongue in cheek, famously observed that for purposes of information retrieval, French can be thought of as misspelled English and 'corrected' using a spelling correction algorithm [25]. More generally, transliteration can be performed based on spelling (essentially, character-level translation) or pronunciation, it can be generative (forward transliteration), inferential (backward transliteration), or implicit (cognate matching) [27]. Transliteration can be particularly useful for proper names, but it is also a potentially useful fallback any time translations are not available from lexicons or from corpus-based methods.

4. Finally, the original source of all translation knowledge is human intelligence, so when all else fails we might turn to the user for help. For example, an interactive system might leverage limited evidence of possible translations to engage the searcher in a dialog regarding how best to translate a query term [148], or systems might show searchers snippets that include untranslatable terms in the hope that the searcher would be able to interpret the term in context.

2.3   How to Use those Translations?

Much of the early work on dictionary-based query translation involved simply substituting one document-language term for each query-language term. Since many terms are homonymous,[6] by which is meant that they have more than one quite different meaning, that naturally leads to the question of which translation(s) should be used. A common approach, both in early CLIR work and in MT, has been to look to other terms in the query in an effort to disambiguate the intended meaning of each query term. However, one term might have both distantly and closely related translations (e.g. the Spanish word *banco* can be translated into English as bank, bench, or pew), so a better question (at least when more than one of the translations is plausible) would be which translations should be used. That naturally leads to the question of how several possible translations for a single term should be used, and that turns out to be the right

---

[6] People often say polysemous when they mean to refer to different meanings for terms that are written identically, but when speaking precisely polysemy refers to closely related meanings—often not a problem in IR—whereas homonymy refers to unrelated meanings.

question in general (since a single known translation is simply a special case of that more general question).

Initial attempts at using multiple translations, such as stuffing all of the translations into one very long query, proved to be problematic. Indeed, even careful attention to the balance of weights between query terms did not work particularly well. The fundamental problem was that ranking functions tend to give more weight to query terms that are rare, so it was the rarest translations that were dominating the results—exactly the opposite of what you would want. Still today, it is not uncommon to see unbalanced query translation inappropriately used as a baseline in CLIR experiments. A better approach, based on closer coupling of translation and retrieval, was ultimately introduced by Pirkola [158]. In Pirkola's Structured Query (SQ) method, the key idea was to think of translation as synonymy: every time any synonym (here, any translation) of a query term was seen in a document, the count for the query term in that document was incremented. With this insight in hand, it was a simple matter for Pirkola to use the synonym operator in the Inquery information retrieval system, a generalization of Inquery's query-time stemming, to perform CLIR.

Meanwhile, a parallel line of work on CLIR had developed in which people were simply cascading Machine Translation (MT) and IR to create CLIR systems. This too usually just substituted one word for another, but now with the MT system making the choice of what that single replacement should be. CLIR researchers soon realized that more could be gained by using the internal representations of the MT system, however, to identify synonyms. Thus dictionary-based and MT-based CLIR researchers were on convergent paths.

The unification of CLIR and MT that occurred in the late 1990's led to a further extension to the structured query method, asking not just which translations were possible, but how likely each document term was to have been translated into a query term. Note the direction of the conditional probability here; we are interested in the probability that a query term is a translation of a document term, not the other way around. We can then ask, for every document, what would be the expected counts of the query terms if the document had been written in the query language. If $e_{i,j}$ is the expected count of query term e_i in document j and $f_{k,j}$ is the actual count of document term f_k in document j then we get:

$$e_{i,j} = \sum_{k \in j} p(e_i|f_k) f_{k,j}, \tag{1}$$

where $p(e_i|f_j)$ is the probability that document term f_k (e.g., banco) would translate to query term e_i (e.g., pew). This probability might be estimated in many ways, but by far the most popular approach when parallel (i.e., translation-equivalent) text is available has been to perform term-level alignment and then compute the maximum likelihood estimate for the probabilities (i.e., the approach known in machine translation as IBM model 1) [125].

This approach to using estimated translation probabilities was first tried with document ranking functions based on unigram language models, but the formulation is general and can be applied with any ranking function. In gen-

eral, ranking functions in information retrieval are based on three elements of a term-by-document matrix: (1) an element statistic based on the number of occurrences of a term in a document (for representing the "aboutness" of a document), (2) a row statistic based on the number of documents in which a term occurs (for representing the specificity of a term), and (3) a column statistic based on the number of terms in each document (that can used to compute a density from the element statistic). For brevity, these are typically called term frequency (TF), inverse document frequency (IDF), and length, respectively. It is useful to remember, however, that in practice what is often meant are not raw counts but rather monotonic functions of those counts.

With that in mind, it is easy to see how the same approach can be used to estimate document frequency. Defining k as the number of distinct terms in the document language, $|f_k|$ as the number of documents containing term f_k, and $|e_i|$ as the expected number of documents in which query term e_i would have appeared (if the documents had been written in the query language), we get:

$$|e_i| = \sum_k p(e_i|f_k)|f_k| \tag{2}$$

This formula actually overestimates $|e_i|$ somewhat (because two terms that translate to e_i might be found in the same document), but the error has been shown experimentally to typically be inconsequential [39]. Alternatively, if a suitably large and representative side collection is available in the query language, $|e_i|$ can be estimated directly from that side collection. For length, it typically suffices to directly use the length as computed in the document language. It is well known that different languages can use more or fewer words to express the same content, but for documents of any reasonable length the scaling factor is fairly close to a constant, and most of the widely used ranking functions produce rankings that are insensitive to constant factors. So adjustments to the document length are not typically required. Darwish and Oard called this approach of separately estimating query-language term frequency vectors and a query-language document frequency vector from the observed document-language statistics Probabilistic Structured Queries (PSQ), since it extends Pirkola's SQ approach by allowing partial mappings [39]. The approach can also be seen as an extension of the approach that language modeling researchers had earlier introduced for the case in which only term frequency statistics needed to be mapped across languages [203].

PSQ is essentially vector translation; you start with either the TF vector or the DF vector in the document language, multiply that vector by a translation probability matrix, and get a TF or DF vector in the query language. At that point, ranking proceeds as if the documents had been written in the query language in the first place. The key point is that TF and DF vectors should be translated separately; precombining those two factors (e.g., using TFIDF or

BM25 term weights) before translation does not work nearly as well.[7] The only twist is that these "translated" TF or DF vectors are no longer vectors of integer counts, but rather are now vectors of real-valued (i.e., partial) term counts. This is typically associated with some spreading of the counts across more terms, which has implications for both effectiveness and efficiency. If implemented at query time (i.e., indexing only the document language TF and IDF statistics), then the index will be no larger than would be the case for monolingual IR, but query time will scale linearly with the average translation fanout. If implemented at indexing time (i.e., indexing query-language term weights) then the index size will increase linearly with the average translation fanout, but the only upward pressure on query time would be the increased time to load the larger postings files from secondary storage (e.g., hard disk or solid-state drive) [145]. Empirical results indicate that adverse effects on retrieval effectiveness from that spreading are often balanced by gains in (average) effectiveness from the expansion effect of the translation mapping (which can usefully add synonyms). Adverse effects on efficiency can be limited by pruning the set of possible translations, although overly aggressive pruning can adversely affect effectiveness.

This line of development left one key question unresolved: how translation should be done when the ranking function makes use of term proximity, term order, or both. The problem was that the initial approach to PSQ translated only decontextualized terms. The first to tackle that problem were Federico and Bertoldi, leveraging the observation that translation can change word order, but that long-distance reordering is less common [46]. They therefore experimented with a variant of what has subsequently come to be called the Sequential Dependence (SD) model [129] in which both original and reversed word orders were treated as equally informative.

Work on PSQ still continues, most recently by Rahimi et al. [159], who show that further improvements can be obtained in some cases by balancing discrimination values (e.g., IDF) calculated in the query and the document languages.

3 Updating Nie: CLIR since 2010

If a CLIR counterpart to Rip Van Winkle were to have fallen asleep in 2010, only to awake in 2022, the thing that surely would most impress them is the tremendous energy around neural methods. Neural networks offer ways of learning nonlinear models that have the potential to improve over simpler linear and hand-engineered nonlinear models [73]. Widespread adoption of Rectified Linear Units (ReLU) over the last decade has made it possible to train much deeper models, and this in turn has led to a blossoming of new model architectures. Applications of these techniques to CLIR have faced three challenges: (1) Initial work on Neural MT focused principally on one-best translation for human readability rather than on tuning for retrieval tasks, (2) neural retrieval models

---

[7] The reason for this is that IDF is perhaps better thought of as a measure of the selectivity of a query term than as a measure of selectivity of any particular document-language translation of that query term.

are data intensive, and present training data has largely focused on monolingual applications, and (3) neural methods are computationally intensive, thus placing a renewed premium on efficiency. In this section we adopt key aspects of the structure of the Chapter 2 to review advances in CLIR since 2010.

3.1   What to Translate: Parts of Words

As described above, systems for handling text as tokens must choose how to segment that text into terms. This task has received particular attention among Machine Translation (MT) and Automatic Speech Recognition (ASR) researchers, who have developed new representations that can also be useful for retrieval tasks. Byte-Pair Encoding (BPE) [178] and Wordpiece [177] are now widely used in MT. Both these methods use subword units to deal with rare words, thus mitigating the out of vocabulary problem. In BPE, subword splitting is first trained on a collection so that a subword vocabulary of a pre-specified size is created. Wordpiece works similarly, but the decision about which subwords to generate is made based on the maximizing language model log likelihood rather than word frequency as in BPE. Wordpiece is especially widely used in BERT models (see below). In contrast to these models, which operate on individual tokens, Sentencepiece [91] operates on the sentence level by first joining all the words in the sentence, and then re-segmenting that aggregated character sequence by inserting a special token at the segmentation points.

3.2   Which Translations: Bilingual and Multilingual Embeddings

The third wave of CLIR (characterized by different representations for queries and for documents) is built from three key innovations: (1) some way of creating dense vector representations in which terms with similar meaning are represented by similar vectors, (2) deep neural architectures that can be trained using backpropagation, and (3) the learned self-attention of the neural transformer architecture. The first of these emerged initially in second-wave IR research. Initial work on creating dense vector representations for terms dates to the introduction of Latent Semantic Indexing (LSI) in 1988 [52]. At the time, linear algebra, specifically a truncated Singular Value Decomposition (SVD), was used to compute a dense representation for each term (known as a singular vector). Today, neural autoencoders use non-linear models to do the same thing: compute a dense representation for each term. These dense representations capture an embedding of the original high-dimensional term space in a lower-dimensional vector space. Given a fixed vocabulary (e.g., from BPE), they can be trained in advance, thus reducing the term-level embedding process to a simple table lookup.

It did not take long for researchers to realize that bilingual embeddings could be created that assigned similar dense representations to terms with similar meanings, regardless of their language. Initially these bilingual embeddings were also constructed using linear algebra [112, 113]; today such representations can also be learned using autoencoders [18, 197]. Three broad classes of techniques have been proposed for creating bilingual term embeddings: (1) pseudo-bilingual,

(2) projection-based, and (3) unsupervised. For a detailed treatment of these approaches generally, we refer readers to a survey on this topic by Søgaard et al. [184]. Here we focus on those parts of the story that involve CLIR.

Landauer and Littman [94] were the first to propose the *pseudo-bilingual* approach. The key idea in this approach is to create a set of mixed-language documents from a comparable corpus using explicit links between the documents (as discussed above in Section 2.2), and then to learn embeddings from that collection. Specifically, they concatenated English and French versions of the same document, and then applied a truncated SVD to learn a dense representation for terms in both languages. Because the truncation tended to preserve meaning while suppressing the effect of term choice, the approach learned similar vectors for terms in different languages that had similar meaning. Twenty-five years later, Vulić and Moens [197] did something similar using a neural autoencoder and document-aligned comparable corpora from Wikipedia and Europarl. Instead of an SVD, they trained a skip-gram [130] model on documents created by concatenating an English document and a comparable document in Dutch, and then shuffled the terms. When used with an unusually wide window, this allows the skip-gram model to take into account bilingual context to generate an embedding for each term. However, randomly shuffling the terms might lead to sub-optimal choice of bilingual context for certain terms. Improving over random shuffling, Bonab et al. [18] used a bilingual dictionary to guide the process of interleaving terms from the two languages so that terms that are putative translations of each other would be adjacent in merged sentences. This makes it likely that a valid translation will appear in the neighborhood context of a term, which will be then used to learn better representations for both that term and its translation. In their work, they used sentence-aligned parallel text, and merged on a sentence-by-sentence basis.

Littman et al. [113] was the first to propose the *projection-based* approach for CLIR, where independently learned monolingual embeddings are aligned using a learned linear transformation such that known translations (e.g., from a biligual dictionary, as described in Section 2.2) yield similar representations. This approach is now very widely used, typically being applied to align monolingual embeddings that were created using autoencoders. Because fairly large dictionaries are now available for many language pairs, the projection-based approach is easily scaled to create multilingual embeddings that today can represent terms from more than 50 languages [4, 77, 93]. In recent work with embeddings learned using neural autoencoders, Bhattacharya et al. [16] used linear regression to learn a transformation between a Hindi and English monolingual embeddings using Hindi-English translation dictionary. Similarly, Litschko et al. [107] mapped source and target embeddding spaces into a single shared space by learning two projection matrices, one for each language, to create a new joint space that is different from that used for either language [183]. In that work, they compared three ways of creating projection matrices: Canonical Correlation Analysis (CCA) [45], minimizing the Euclidean distance between the projected vector and the target vector [131], and maximizing a retrieval-oriented

measure they call Cross-Domain Similarity Local Scaling (CSLS) [77, 93]. In multilingual CLIR, Bhattacharya et al. [15] used two different approaches to learn multilingual embeddings, a projection-based method in which Bengali and Hindi were mapped to an English embedding space, and the pseudo-bilingual approach proposed by Vulić and Moens [197], extended to multiple languages.

Recently, there has also been some work on *unsupervised* approaches to learning bilingual embeddings using projection-based techniques [7, 67, 93]. While supervised approaches rely on sources such as bilingual dictionaries to learn alignments, unsupervised approaches instead follow a two-step process: 1) build an initial seed dictionary from monolingual vector spaces using some unsupervised approach, and 2) iteratively refine that dictionary to add more pairs that will eventually be used to learn better embeddings. Although there exist several unsupervised approaches to learn the initial dictionary, the key idea in the second step is to project one embedding space onto another and then to use some of the near neighbors in the aligned spaces as the added dictionary entries. The induced dictionary can be iteratively refined using Procrustes analysis [176], and the refined dictionary can then be used to learn better bilingual embeddings. This does, however, assume that the vector spaces are isomorphic, which needs not be true for languages that are typologically different, and this approach has been shown not to work well in such cases [198]. Litschko et al. [107, 108] used these unsupervised approaches to perform fully unsupervised CLIR in language pairs involving limited training resources. However, the effectiveness of fully unsupervised approaches was found to be rather limited when compared to supervised methods by [198].

One problem with embeddings learned using either linear algebra or autoencoders is that homonomy distorts the representation, since the model has no way of knowing if this *banco* is a financial institution or a wooden bench in a park. The advent of neural methods using self-attention made it possible to produce context-sensitive embeddings, with the embedding for *banco* differing, depending on its surrounding terms [40, 154]. The resulting Bidirectional Encoder Representations from Transformers (BERT) architecture [40] has since been further extended (e.g. RoBERTa [115], XLNet [207], ELECTRA [31], XLM [92]), some of which support more than 100 languages. We should emphasize that this use of BERT and its derivatives to produce contextual embeddings differs from the third-wave use of BERT that we describe in more detail below for direct query-document matching.

All the techniques discussed so far for producing cross-lingual embeddings produce generic representations for queries and documents that need to be tuned for the CLIR task. An alternative is to learn representations that are tailored for CLIR, as had already been done with for representations tailored for monolingual retrieval [70, 213]. The advantage of these "representation-based" learning approaches is that the representations of the documents can be precomputed, and during retrieval the score on which ranking is based can then be more rapidly computed (more on this in Section 3.4). The key challenge then lies in learning specialized representations of queries and documents that are useful for

this task. Gupta et al. [66] learned CLIR-specific representations for queries and documents using a transfer learning approach based on a parallel corpus. First, they learned a representation model similar to DSSM [70] by using a monolingual retrieval collection in English to learn representations such that similarity computed between representations of the queries and representations of the relevant documents are maximized, whereas similarity computed between representations fo the queries and representations of the non-relevant documents is minimized. Using a trained monolingual model, they extended it to the cross-language setting with an English-Hindi parallel corpus. To do so, the monolingual model was applied to the English side of the parallel text, and a cross-lingual model was trained on Hindi side. The goal was maximize the similarity of the English representation computed using the monolingual model and the Hindi representation computed using the cross-lingual model for a given translation pair. In a somewhat different approach, Li and Cheng [99] employed an adversarial learning framework using a Generative Adversarial Network (GAN) [62] to learn a CLIR-specific representation by incorporating both monolingual and cross-lingual matching signals. Their model takes as input three sources: raw queries, documents, and translated queries (in this case, translated using Google Translate). They incorporated several constraints in their model to learn task-specific representations: 1) a monolingual constraint that brings the representations of documents and translated queries closer, 2) a cross-language constraint that brings the representations of documents and raw queries closer, 3) a translation constraint that brings the representations of raw and translated queries closer, and 4) an adversarial constraint that produces source-agnostic representations.

Instead of using an off-the-shelf multilingual contextual language model to generate generic embeddings, Litschko et al. [109, 110] explored the use of multilingual contextual models fine-tuned on sentence-level parallel corpora (e.g., LASER [206], m-USE [6], LaBSE [49]) to produce embeddings specialized for semantic similarity. Naturally, these models work better for matching queries with sentences than with longer documents. While the existing multilingual models are pretrained on generic tasks such as Masked Language Model (MLM), Translation Language Model (TLM) and Next Sentence Prediction (NSP), another line of work explored designing custom pretraining tasks that are tailored for CLIR. Yu et al. [212] introduced Query Language Modeling (QLM), which is a variant of TLM with two differences. While TLM relied on sentence-level parallel-corpora, QLM uses the section titles of Wikipedia in one language as query and the relevant section in another language as the relevant document. Instead of masking tokens in both the source and target side of parallel sentences, as in TLM, QLM only masks tokens in the source side. Yu et al. [212] further introduced a Relevance Ranking (RR) task that produces a higher ranking score using the contextualized embeddings of the concatenated sequence of query and relevant document than a sequence of query and non-relevant document. Yang et al. [204] introduced C3, a continual pretraining recipe which generalized Yu et al.'s RR task to include several non-relevant documents as hard negative samples. This framework, known as contrastive learning [28], has been shown to work

well in many domains. C3 further introduced a token-level similarity task, which is part of the ColBERT [85] framework, to promote cross-lingual alignment.

3.3   Which Translations: MT

MT has advanced quite rapidly over the past decade, building upon the key innovations in distributed word representations and deep learning technologies. For high-resource language pairs, this has resulted in neural systems convincingly surpassing the performance of the statistical MT systems that had dominated MT research activity in the first decade of this century. This started with the advent of attention-based neural methods [13] built on an underlying encoder-decoder architecture [137]. The key idea behind the attention mechanism is to leverage the fact that models can pay attention to different parts of the source sentence while decoding each target token. Later on, the introduction of subword units (described above in Section 3.1) addressed the problem of translating rare words. Ultimately, these two factors, attention and subword units, proved to be the key ingredients.

As mentioned above in Section 2.1, query translation is often a preferred approach owing to its computational efficiency in experimental settings. However, the key challenges lie in 1) overcoming the ambiguity introduced by short word sequences in queries, and 2) mismatch between the style of writing for the queries to be translated and the style of the parallel sentences used for training the MT models (as described in Section 2.1). Prior work has tried to tackle the style mismatch issue by creating an in-domain dataset of queries paired with their translations and using it to train SMT/NMT systems. In the context of CLIR, Nikoulina et al. [140] first explored this idea for building a SMT system using a dataset consisting of human translated queries provided by CLEF. Hu et al. [69] built a small manually-curated in-domain dataset consisting of popular search queries, augmenting it with large amounts of parallel out-of-domain sentences to train a NMT model. As a different approach, Yao et al. [208] utilized click-through data from a multi-lingual e-commerce website[8] to mine query translation pairs. Another way to minimize the style mismatch effect is to force the MT model to to incorporate terms from the retrieval collection in the translation of the query. This has been explored in SMT by using the target retrieval collection to train a language model, which helps in translating the query in a way that better matches the documents [139].

Similarly, Bi et al. [17] trained a neural model by constraining the translation candidates of the query to the set of document terms ranked highly by TFIDF. The set of documents used for this was selected using click-through data for each query. In a somewhat different approach, Sarwar et al. [171] tried to solve the style mismatch problem using a multi-task neural model that optimized for improving the query translation jointly with a relevance-based auxiliary task. The goal of the auxiliary task was to restrict the NMT decoder to produce translations using terms that are equally likely in both the retrieval collection

---

[8] https://www.aliexpress.com/

and the MT training corpus. The auxiliary task involved learning a relevance-based embedding model on the content of the top-ranked document from the retrieval corpus, using the target language sentence from the parallel corpus as the query. Since these word embeddings are shared across the NMT and relevance models, that forces the NMT decoder to choose translations that occur in both the retrieval collection and the parallel corpus.

Document translation, on the other hand, has the advantage of the additional context present in the longer documents. Attention-based encoder-decoder NMT models leverage the context present in both the source and the target side to produce context-dependent translations, in contrast to dictionary-based or PSQ techniques that rely on context-independent term translation. However, owing to the computational overhead involved in translating the entire document collection (as described above in Section 2.1), there has been less work exploring neural models for document translation than there has been for query translation. However, full encoder-decoder models are only one point in a larger design space. Relaxing the contextual constraint to apply only to the document being translated, and not to the sequence of terms that are generated as possible translations, leads to an approach for generating multile translation hypotheses for each term called the Neural Network Lexical Translation Model (NNLTM) [215]. Applying constraints only on the source side implies that the translation(s) of the document terms are assumed to be independent of each other. While this would affect the fluency of translations intended for human readability, the principal goal of a translation model for CLIR is to generate alternative translations, and not necessarily to properly order all those translations. This is very similar to the case of PSQ relying on the word alignments generated from the SMT system rather than using the language model to help select and order those translations. For each document term, NNLTM uses the neighboring terms within a window to produce the contextualized translation alternatives. In order to do so, while training, NNLTM relies on the alignments from a parallel corpus generated using word aligners, and uses an aligned term in the document language along with the context words surrounding it term to predict the translation in the query language.

Ignoring other qualities of query or document translation described in section 2.1, such as speed, a natural question to ask is which approach is most effective? Contrary to our expectation, recent work from Saleh and Pecina [170] found that translating queries actually works better than translating documents while retrieving texts from a medical domain. However, the queries were translated from European languages to English, while the documents were translated from English to other European languages, which is potentially a harder problem. McCarley [125] also tried to answer the question of whether the query or the documents should be translated, finding that neither approach was a consistent winner.

If the quality of MT and CLIR are interconnected, the subsequent question is how a machine translation system can be tuned to achieve better CLIR retrieval effectiveness. One commonly used approach to integrate retrieval signals with

MT is to (re)-rank the translation alternatives produced by an MT system in order to maximize a retrieval objective. This was first attempted by Nikoulina et al. [140], who tuned their SMT system to produce translations that directly optimize Mean Average Precision (MAP). Since then, Sokolov et al. [185] have focused on optimizing SMT decoder weights to maximize a retrieval objective that leads to producing translations geared towards better retrieval. Saleh and Pecina [168] have since built a feature-based reranking model that reranks different alternatives produced by an MT system and then selects the translation that maximizes a retrieval objective.

3.4   Using Translations: Neural IR

Now that we have different ways of generating bilingual embeddings that are either generic or task-specific, as discussed in Section 3.2, the next question to ask is how to use them for CLIR. One possibility, which we have already mentioned, is to use *embedding similarity* between query and document representations. One way to create such representations is to create an aggregated query vector from the embeddings of individual query terms and an aggregated document vector from the embeddings of the individual document terms. Some similarity measure such as the cosine can then be used to compute a retrieval score for each document, with documents sorted in decreasing score order. Vulić and Moens [197] explored different forms of aggregation, including unweighted and weighted summation of word embeddings, using IDF as weights for the weighted summation. They found that IDF-weighted summation worked better than unweighted summation.

A second approach for incorporating embeddings into a retrieval system is to use similarity in a bilingual embedding space of terms from different languages as a basis for term translation. We refer to this approach as *embedding-based translation*. For example, Bhattacharya et al. [16] used the k most similar terms from the document language as translations of a query term. This is similar to query expansion, but in a cross-language setting. In a similar approach, Litschko et al. [107] introduced a model that they called Term-by-Term Query Translation (Tbt-QT) which, as the name suggests, uses embedding-based translation of each term. Specifically, they used a single cross-language nearest neighbor for every query term. As might be expected from the experience with pre-neural CLIR methods, this turns out to be sub-optimal when compared with using more than one nearest neighbor [18].

A third approach, which has now been well studied in monolingual IR, uses embeddings as input to custom neural architectures that are optimized for relevance and that are fast enough to perform full-collection neural ranking on modestly large collections. In one prominent line of work on monolingual information retrieval known as "full-interaction-based" methods, query and document terms are encoded using a single neural network initialized with Word2Vec monolingual representations [130], and the model learns interactions between those representations to maximize a relevance objective (DRMM [64], KNRM [202],

or PACRR [72]). The work of Yu and Allan [211] extends these full-interaction-based approaches to CLIR. These models are initialized with fastText bilingual embeddings aligned using a dictionary [77], which is type of projection-based bilingual word embedding (as described in Section 3.2).

Subsequent research in monolingual retrieval has focused on switching from context-independent Word2Vec embeddings to the context-sensitive BERT-based representations described above in Section 3.2 to initialize the neural ranking models. These third-wave context-sensitive Transformer-based architectures have achieved results superior to the best previously known techniques for ranking documents with respect to a query in monolingual applications [85, 104, 119]. ColBERT [85] introduced a two-stage retrieval approach that relies on learned embeddings to produce an initial set of documents for reranking. These embeddings were trained MaxSim, a late-interaction-based mechanism that finds the document token embedding that is closest to a specific query token embedding. The key idea is that query and document embeddings can be computed separately; that allows us to precompute and store the document embeddings in a nearest neighbor index. During inference, the first-stage retrieval system computes a query embedding on-the-fly and then uses approximate nearest neighbor techniques to find the documents with terms that are closest to the token embeddings of the query. In the next stage, these documents are ranked using the one query token whose contextual embedding is most similar to that of each query term. While initially developed for monolingual applications, ColBERT has been generalized to the CLIR setting [103, 136]. One variant, ColBERT-X [136], used the translations of relevant passages in MS MARCO [19] to build a CLIR model to retrieve documents in non-English languages using English queries. Li et al. [103] showed that a generalization of ColBERT to CLIR also works when retrieving English documents using non-English queries.

Full-interaction-based ranking models that make connections between each query term and every document term can be computationally expensive when ranking all of the documents in a very large collection such as the Web. This leads to the fourth way of employing embeddings in CLIR, *cascade-based approaches*. These involve running an efficient recall-oriented ranker first to get an initial ranked list, which is then re-ranked using a more computationally expensive model [38, 142, 210]. However, the retrieval effectiveness of cascade re-ranking approaches is limited by by the quality of the initial ranked list produced by a fast term-matching recall-oriented ranker. Zhang et al. [219] extended the cascade re-ranking approach to the cross-language setting and found similar improvements as in monolingual retrieval, especially on low-resource languages. Jiang et al. [75] fine-tuned a pretrained multilingual BERT model on cross-language query-sentence pairs constructed from parallel corpora in a weakly supervised fashion and applied that model to perform re-ranking for CLIR. Shi and Lin [181] used a transfer learning approach, applying a retrieval model trained on a large collection in English to retrieve documents from collections in other languages. Litschko et al. [111] aimed on reducing the amount of training data needed to

fine-tune the model on various languages by using Adapters and Sparse Fine-Tuning Masks.

3.5 Fusion

Different approaches to IR have different strengths and weaknesses. This inherent diversity suggests that there is sometimes potential for improving search results by combining different approaches. CLIR systems have even greater potential for benefiting from diversity than do monolingual IR systems because of the additional potential for diversity that translation resources, and ways of using those translation resources, introduces. Combining multiple sources of evidence has a long heritage, and the inclusive name for the general approach is fusion [35, 36, 201]. In IR we distinguish between two classes of fusion, *early* fusion (in which evidence from multiple sources is combined by one or more components of the full system before results are generated) or *late* fusion (in which the results of separate IR systems are combined). Late fusion is often referred to as system combination. In general, late fusion can be employed with systems that search different collections, but here we focus only on cases in which all systems search the same collection.

The earliest work on system combination for CLIR dates to 1997 when, as noted above, McCarley [125] found that neither query translation nor document translation was a clear winner. He went on to test a late fusion combination between the two approaches, finding that the combination of the two approaches yielded the best results. McCarley had endeavored to hold every other aspect of the system design constant, but later experiments by Braschler [24] and by Kishida and Kando [87] demonstrated benefits from combining more diverse systems as well.

Early fusion of translation resources has also proven to be successful in CLIR. Perhaps the best known case is the fusion of parallel sentences with translation pairs from a bilingual lexicon when training machine translation systems. The architecture for this is simple; machine translation systems are trained on pairs of translation-equivalent sentences, and term pairs from a bilingual lexicon are simply treated as very short sentences. Because the term distribution in bilingual lexicons are not as sharply skewed in favor of common terms as is naturally occurring parallel text, this approach can help to overcome a natural weakness of parallel text collections (limited size) and by helping to more reliably learn translations of relatively rare terms.

One challenge that arises in both early and late fusion is that the evidence to be merged may be represented differently. This occurs, for example, when merging translation probabilities (e.g., from statistical alignment) with translation preference order (e.g., from a bilingual dictionary) in early fusion, or when merging document scores from different retrieval systems in late fusion. In such cases, some form of score normalization is needed. This problem has been extensively studied in monolingual fusion applications, and approaches like CombMNZ [96] with sum-to-one normalization that work well in monolingual settings also seem to work well in CLIR [134].

Türe et al. [192] conducted early fusion experiments with three ways of estimating term translation probabilities (one with language model context, one with term context, and one with no context), finding substantial improvements over using the best of those three representations alone. Türe and Boschee [191] later extended this approach to include learned query-specific combination weights for a similar set of three translation probabilities (two with language model context and one with no context), finding further improvements. Shing et al. [182] explored an alternative approach, learning late fusion combination weights without supervised (retrieval effectiveness) training by instead using data programming to estimate system weights. Nair et al. [134] compared early and late fusion using two approaches (one with language model context, one without), finding similar improvements over the best single system from both approaches. Finally, Li et al. [100] experimented with using cross-attention for early fusion between CLIR (for product descriptions) and product attribute matching in an e-commerce application.

One particularly important consideration when using neural methods is that neural methods excel at generalization for cases in which there is adequate training data, but they have more difficulty modeling rare phenomena. Because queries often tend to skew toward more specific, and thus less common, terms, a commonly used approach has been to combine neural and non-neural results. Good results have been reported for that approach in CLIR as well, both with early fusion [18, 71] and with late fusion [175].

3.6 Cross-Language Speech Retrieval

We have so far described retrieval of content (and queries) represented as digital text. But CLIR can be also applied to content that is spoken in a language different from that of the queries. When doing CLIR on recorded speech, a straightforward approach would be to first perform Automatic Speech Recognition (ASR) and then to treat the resulting text as a text CLIR problem. This is the approach that was used by all of the participants in the first four shared-task evaluations on cross-language speech retrieval [47, 48, 147, 151, 200]. This can work well when ASR accuracy is high, but there are still many settings (e.g. conversational speech, or speech recorded under difficult acoustic conditions) in which one-best ASR transcripts exhibit substantial word error rates. As the throughput of computing resources has dramatically increased over the last two decades, it has thus become more common to take advantage of the more complex internal representations of ASR systems.

Internally, ASR systems typically represent what might have been spoken as some form of confusion network (for what might have been spoken between fixed time points) or lattices (for what might have been spoken between arbitrary time points). These representations might be phonetic or word-based. For word-based confusion networks that are aligned to the one-best word boundary, equations (1) and (2) in Section 2.3 are easily extended to incorporate an additional multiplicative factor for transcription probabilities estimated from ASR confidence scores. As Nair et al. [135] found, this can yield better CLIR results

than would be achieved from one-best ASR. Yarmohammadi et al. [209] have also reported a beneficial effect from concatenating multiple recognition hypotheses from utterance-scale (i.e., sentence-like) confusion networks.

Phonetic approaches to cross-language speech retrieval have also been explored. Because translation is typically performed on words, not phonemes, this approach is better matched to a query translation CLIR architecture. This is the approach that was tried in the very first reported work on cross-language speech retrieval [180], in which query terms were first translated using a comparable corpus, then rendered phonetically, and matched with phoneme trigrams created for the spoken content using an ASR system.

Finally, it is worth noting that there has been a spurt of recent activity on cross-language speech retrieval as part of a large multi-site research program in the United States known as MATERIAL [214]. Two things distinguish MATE-RIAL from much of the prior work. First, the program experimented with nine less commonly studied languages (Bulgarian, Farsi, Georgian, Kazakh, Lithuanian, Pashto, Somali, Swahili and Tagalog). Using results from that program, Rubino [162] found that differences in retrieval effectiveness were much better correlated with the amount of available training data for ASR and MT than with the phonetic or syntactic similarity of the query and content languages. Second, the goal of a MATERIAL system was not to rank documents, but rather to select a set of relevant document [79, 141]. Set selection is a challenging task, but it is a task that is well matched to the needs of speech retrieval applications. In text retrieval, the searcher might easily skim a ranked list of results, and in CLIR they might use MT to help them with that task. In speech retrieval, however, the linear nature of audio means that skimming can take longer, and the speech translation systems that would be needed to support that task for cross-language speech retrieval are less accurate than are current text translation systems. In some applications, it may therefore be useful to think of cross-language speech retrieval as a triage application, selecting documents that will then require time from expert speakers of the language. Systems developed for MATERIAL (e.g., [20, 217, 53]) are thus useful as exemplars for complex cross-language text and speech retrieval systems that can draw on the full range of techniques described above.

3.7   Building Real Systems

CLIR research tells us what can be built, and it offers some prescriptions for how best to build it. Compared with the architecture of the IR system, CLIR pipeline contains extra steps which help us crossing the language barrier. They might be done at different spots of the pipeline, depending on the approach. For example embedding-based approaches require using multilingual model instead of the monolingual one and translation-based approaches require running a translation before retrieval. This additional complexity might be also reflected in indexes, which might need to store additional information such as the probability of the translation in the PSQ method. Many details remain to be worked out when

CLIR techniques are employed in real systems. Here we highlight two of those issues: language-specific processing, and accommodating multilinguality.

In research systems, we often strive language-independent methods. In an earlier era, a chapter on CLIR might have devoted considerable attention to the characteristics of specific languages. In recent years, however, CLIR has been evaluated on text and speech collections in a substantial number of language pairs without yet finding pairs for which the same basic approaches would not work. Methods such as Byte-Pair Encoding and Wordpiece, for example, diminished the need for language specific tokenization. Practical systems, by contrast, can still benefit substantially from language-specific processing. As the language of the queries and documents differs, CLIR systems naturally involve two (or more) sets of language-specific issues. For example, in English, stemming is normally applied only to suffixes. In Arabic, by contrast, stemming must pay attention to both prefixes and suffixes, and the agglutination means that Arabic stemming will benefit even more from recursion than would English stemming. Dialect variations in some languages can introduce a whole new level of complexity, and it is not uncommon to find multiple dialects in the same collection. It is not uncommon to find unicode characters used incorrectly, particularly when the visual appearance of different characters is similar. And even when correctly encoded, differences in transliteration conventions (or, more properly, the lack of a single transliteration convention) can result in a substantial variety of representations for the same word when that word has been transliterated from a different character set. Transliteration of queries can also introduce considerable variation, as keyboards on different devices may limit the characters that can easily be generated. The number of language specific issues is so great that one is reminded of the old adage that success is 10% inspiration and 90% perspiration.

Although our focus throughout this chapter has been on cases in which queries are expressed in one language and documents are expressed in one other language, the world is not that simple. Many collections include mixed languages, and many documents are multilingual for the simple reason that they include so-called loan words from other languages. Multilinguality poses a number of challenges. Detecting the predominant language in a document is a fairly easy task, but there are cases in which the language of a query might only be reliably guessed if some information about the person asking the query is known. For example, *Gift* means marries in Swedish, a present in English, and poison in German. The challenges are not limited to the queries or the documents, the system itself has choices to make. For example, if potentially relevant documents in many languages can be found, the system might opt to process and index them in the same way or language specifically. It could also display several ranked lists and allow the user to choose, or it could try to merge the results and display one ranked list.

4   Evaluation

As with IR generally, shared task evaluations and test collections have been driving forces in the development of CLIR, helping researchers to identify and collaborate on emerging problems, providing the training and evaluation data needed to make technical progress, and fostering comparability across research sites.

4.1   Shared-Task Evaluation

Nie provides an overview of the shared-task evaluations organized through 2010. At that time, only the TREC book had been published [196]; since then, edited volumes that review two of the other major CLIR evaluation venues (CLEF and NTCIR) have appeared [153, 167]. The Text REtrieval Conference (TREC) was the venue for the first cross-language retrieval tasks. Early work on CLIR was conducted as a side activity associated with work on non-English collections (in Spanish and Chinese). The main focus of the TREC CLIR tasks between 1997 [194] and 1999 [195] was on CLIR in multiple languages. It was there that a key decision that has influenced CLIR test collection design was first sorted out: the topics (from which the queries are generated) should be translated as natural expressions in the target language, so as to more faithfully represent what a real searcher might pose to a CLIR system. Early in the new century TREC returned to bilingual tasks involving Chinese or Arabic [56, 57, 146] and then again in 2022 as in the NeuCLIR task focused on Chinese, Persian, and Russian. The focus on CLIR for European languages moved to the newly formed Cross-Language Evaluation Forum (CLEF) in Europe. As its name suggests, a range of CLIR-related tasks were organized at CLEF, including ad-hoc CLIR and domain specific CLIR, interactive CLIR, Cross-Language Spoken Document Retrieval and Cross-Language Retrieval in Image Collections.

Starting in 2012, CLEF embraced broader range of text and multimedia processing tasks and was renamed as the Conference and Labs of the Evaluation Forum. CLIR-related tasks at CLEF after 2012 include eHealth, which has included multilingual search in medical data, and Multilingual Cultural Mining and Retrieval (MC2). The CLIR part of the MC2 Lab task [33] was focused on microblog search. Short critiques of movies in French were used as the queries. The main focus was on finding related microblogs in order to provide a diverse summary of a movie based on comments and movie information in French, English, Spanish, Portuguese and Arabic. CLIR has been part of the eHealth tasks since 2014 [81]. The task was in 2014 focused on search by the general public rather than medical professionals. The collection contained one million of medical documents. Queries were originally created for the monolingual task in English and were subsequently translated into Czech, French and German. Though the task in the following year [59] was still oriented towards the general public, queries were created to imitate users searching in the database using observed medical symptoms. Arabic, Farsi and Portuguese were added to the languages of the queries. The ClueWeb 12 B13 collection was used as the collection in the

following years and the queries were mined from Web forums to better imitate realistic conditions [83, 60]. The list of languages also changed, as Hungarian, Polish and Swedish were supported in addition to German, French and Czech. In 2018 [26], webpages from the Common Crawl corpus were used as the collection and queries that had been issued by nonprofessional users of the public HON[9] search service were used.

CLIR tasks involving Japanese and some other East Asian languages were first organized at what is now called the NII Testbeds and Community for Information access Research (NTCIR) evaluations between 2002 and 2007. NTCIR tasks related specifically to CLIR in that period included Cross-Language Q&A, Advanced Cross-Lingual Information Retrieval, Question Answering, Patent Retrieval, and Cross-lingual Link Discovery. Since 2010, CLIR was part of two tasks: GeoTime (between 2010 and 2011), and Crosslink (between 2011 and 2013). In the GeoTime task [54, 55], the search was constrained by geographic and temporal constraints using 'when' and 'where' queries. English and Japanese queries were used to search in English and Japanese news. The main focus of the Crosslink task [189, 190], was automatic linking of related content in different languages.

The Forum for Information Retrieval (FIRE) was created in 2008 to focus on South Asian languages. CLIR tasks at FIRE have included Ad-hoc Cross-language Document Retrieval (between 2010 and 2013), SMS-Based FAQ Retrieval (between 2011 and 2012), Cross-Language Indian News Story Search (CL!NSS) (in 2013), and Mixed Script Information Retrieval (between 2015 and 2016). Similarly to CLEF, the focus of FIRE has gron over time to include a broader range of multilingual access problems such as language identification, event and information extraction, offensive content and hate speech detection and fake news detection.

Recently, CLIR was also the focus of the OpenCLIR challenge,[10] which used a Swahili test collection from the MATERIAL program.

4.2 Test Collections

Shared task evaluations provide test collections and evaluation frameworks to their participants, often with an eye toward making those test collections available to future researchers. We focus in this section on test collections which are presently publicly available; an overview is in Table 1. We list both CLIR test collections and test collections used in closely related problems, such as cross-language question answering. In addition to shared task evaluation venues, which were traditionally the leaders in providing test collections, companies such as Microsoft and Facebook have recently provided access to a range of test collections, especially those oriented towards question answering. Those collections are typically much larger than the collections used in shared task evaluations, which is particularly important for training data-hungry neural techniques. In contrast to

---

[9] https://www.hon.ch/en
[10] https://www.nist.gov/itl/iad/mig/openclir-challenge

this, the collections such as the Large-Scale CLIR Dataset[11] by Sasaki et al. [172], CLIRMatrix[12] by Sun and Duh [188], and the XQA[13] open-domain question answering collection by Liu et al. [114] were created from Wikipedia. Though these collections can also be useful for training neural models, their synthetic nature makes them more suitable for pretraining than for task-specific fine tuning. We therefore omit such collections from our summary. We also omit the early TREC collections, which were well covered in Nie's review. Among question answering test collections, we focus only on those with broad language coverage; more test collections focused on single languages also exist (e.g. MMQA [65]).

One important caveat when using these collections is that older test collections may not include judgments for relevant documents that are found by newer systems [105]. This could produce the paradoxical result that better systems might get worse scores. Moreover, CLIR systems are subject to advances not just in the core IR technology, but also in the translation technology that they use, and (for cross-language speech retrieval) in the ASR technology that they use. A number of evaluation measures that are robust to random omissions of relevance judgments are known (e.g., infAP and bpref), but the problem with

---

[11] `http://www.cs.jhu.edu/~kevinduh/a/wikiclir2018`

[12] `http://www.cs.jhu.edu/~shuosun/clirmatrix`

[13] `https://github.com/thunlp/XQA`

[14] Apart from the data, the CLEF Initiative topics, experiments, pools and performance measures from 2000 to 2013 are also available at `http://direct.dei.unipd.it/`.

[15] `https://catalogue.elra.info/en-us/repository/browse/ELRA-E0008/`

[16] `https://catalog.ldc.upenn.edu/LDC98T25`

[17] `https://catalog.ldc.upenn.edu/LDC2001T57`

[18] `https://catalog.ldc.upenn.edu/LDC2001T58`

[19] `https://catalog.ldc.upenn.edu/LDC2005T16`

[20] `https://catalog.ldc.upenn.edu/LDC2006T19`

[21] `https://www.nii.ac.jp/dsc/idr/en/ntcir/ntcir.html`

[22] `http://catalog.elra.info/en-us/repository/browse/ELRA-E0038/`

[23] `https://www.nii.ac.jp/dsc/idr/en/ntcir/ntcir.html`

[24] `https://catalogue.elra.info/en-us/repository/browse/ELRA-E0036/`

[25] `https://catalogue.elra.info/en-us/repository/browse/ELRA-E0037/`

[26] `https://www.ldc.upenn.edu/collaborations/past-projects/gale/data/gale-pubs`

[27] `https://www.nii.ac.jp/dsc/idr/en/ntcir/ntcir.html`

[28] `http://fire.irsi.res.in/fire/static/data`

[29] `https://www.nii.ac.jp/dsc/idr/en/ntcir/ntcir.html`

[30] `https://sites.google.com/site/clefehealth/datasets`

[31] `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1912`

[32] `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2925`

[33] `https://github.com/facebookresearch/mlqa`

[34] `https://github.com/apple/ml-mkqa`

[35] `https://github.com/deepmind/xquad`

[36] `https://github.com/google-research/xtreme`

[37] `https://nlp.cs.washington.edu/xorqa`

[38] `https://github.com/hltcoe/HC4`

Collection	Published	#Queries	#Documents	Languages	Availability
The CLEF Test Suite for the CLEF 2000-2003 Campaigns – Evaluation Package [21, 22, 23][14]	2001-2003	40/50/50/60 (depends on the year and language)	44-454k (depends on the language)	English, French, German, Italian, Spanish, Dutch, Swedish, Finnish, Russian, Portuguese	ELRA[15]
TDT 1-5 [199, 30, 63]	2001-2006	25/96/80/250 topics	16-98k in total (depends on the year and collection)	Arabic, Chinese, English	LDC[16][17][18][19][20]
NTCIR-3-6 CLIR: IR/CLIR Test Collection [97, 88, 1, 78]	2002-2007	80/60/50/50 topics (depends on the year)	10-900k (depends on the year and language)	Chinese, English, Japanese, Korean	Online application[21]
CLEF Question Answering Test Suites (2003-2008) – Evaluation Package [120, 121, 193, 122, 58, 50]	2003-2008	200 topics per year	55k-454k (depends on language)	Bulgarian, Dutch, English, Finnish, French, German, Italian, Portuguese, Romanian, Spanish	ELRA[22]
NTCIR CLQA (Cross Language Q&A data Test Collection) [173, 174]	2004-2007	200 topics per year	10-900k (depends on the year and language)	Chinese, English, Japanese	Online application[23]
CLEF AdHoc-News Test Suites (2004-2008) – Evaluation Package [152, 143, 41, 42, 2]	2004-2008	50 topics per year	17-454k (depends on the language)	14 languages	ELRA[24]
CLEF Domain Specific Test Suites (2004-2008) – Evaluation Package [89, 90, 186, 156, 155]	2004-2008	25 topics per year	20-303k (depends on the language)	English, German, Russian	ELRA[25]
GALE Phase 1-4 [187, 10, 11]	2005-2009	depends on the collection	depends on the collection	Arabic, Chinese, English	LDC[26]
NTCIR ACLIA (Advanced Cross-Lingual Information Retrieval and Question Answering) [166, 132]	2008-2010	100 topics per language each year	249k-1.6M (depends on the year and language)	Chinese, English, Japanese	Online application[27]
FIRE Information-Retrieval Text Research Collection [133, 124, 123]	2008-2012	50 topics per language each year	95-500k (depends on the year and language)	Assamese, Bengali, English, Gujarati, Hindi, Odia, Marathi, Punjabi, Tamil, Telugu	Online application[28]
NTCIR Crosslink (Cross-lingual Link Discovery) [189, 190]	2010-2013	28 / 25 topics	202k-3.6M (depends on the year and language)	Chinese, English, Japanese, Korean	Online application[29]
CLEF eHealth 2014-2018 [81, 59, 83, 60, 26]	2014-2018	50/50/67/300/50 topics (depends on the year)	23k / 1M / 5.5 M (depends on the year)	Arabic, Czech, English, Farsi, French, German, Hungarian, Polish, Swedish	Differs by collection[30]
Czech Malach Cross-lingual Speech Retrieval Test Collection	2017	118 topics	353 interviews (592 hours)	Czech, English, French, German, Spanish	CC-BY-NC-ND 4.0[31]
Extended CLEF eHealth 2013-2015 IR Test Collection [169]	2019	50/50/166 (depends on the year) queries	1.1M	Czech, English, French, German, Hungarian, Polish, Spanish, Swedish	CC BY-NC 4.0[32]
MLQA [98]	2020	around 5k Q/A pairs per language	2-6k articles per language	English, Arabic, German, Spanish, Hindi, Vietnamese, Chinese	CC-BY-SA 3.0 [33]
MKQA: Multilingual Knowledge Questions & Answers [117]	2020	10k Q/A pair per language	-	26 languages	CC BY-SA[34]
XQuAD [8]	2020	1190 Q/A pairs	240 paragraphs	11 languages	CC-BY-SA 4.0 [35]
XTREME [68]	2020	differs by task	differs by task	40 languages	Apache License v2.0[36]
XOR QA [9]	2020	2-3 Q/A pairs per language	-	Arabic, Bengali, Finnish, Japanese, Korean, Russian, Telugu	CC BY-SA 4.0[37]
HC4: HLTCOE CLIR Common-Crawl Collection[95]	2022	60 topics	0.5M - 5M (depends on the language)	Chinese, English, Persian, Russian	Online[38]

Table 1. Overview of CLIR collections. If the collection contains more than 10 languages, the number of languages is shown.

new methods tested on older collections is that the omissions may well not be random. Users of older test collections would thus be wise to check the number of highly ranked unjudged documents in their result sets, and in the result sets for the baselines to which they compare, and to caveat their results if substantial differences in the prevalence of unjudged documents is observed.

Finally, we note that the full scope of IR evaluation design issues in IR arise in CLIR as well. Most notably, comparison to weak baselines, as noted generally by Armstrong et al. [5], seems particularly problematic in CLIR. Hopefully, this chapter helps to address that problem going forward. Moreover, test collections on which CLIR results are reported are often smaller than those for IR more generally, which also means that greater attention to issues of experiment design and reporting [51, 165] continue to be particularly important. Finally, the development of standard CLIR systems that are widely available lags well behind the situation for IR more generally [43, 205], thus posing additional challenges for reproducability.

5   Future Directions

As the development of CLIR techniques is closely tied with developments in IR and natural language processing, it can be expected that trends in these areas will be reflected in CLIR. Especially as neural approaches have only recently risen to prominence in IR, further growth of those techniques in CLIR can also be expected. With dominance of neural machine translation techniques in recent years, it can be expected that closer integration of machine translation and information retrieval techniques, such as joint training and tuning, will be further explored. Apart from deep neural networks, other machine learning concepts, such as reinforcement learning, imitation learning, bandit algorithms or metalearning may also find their way into CLIR.

The challenge of interaction design for IR also exists in CLIR. There has been far less reported research on interaction design for CLIR than on monolingual retrieval, and the work that has been done (e.g., [106, 149, 157, 218]) points to a number of open research questions. Interactive CLIR systems could, for example, use information about preferred document languages to decide which documents to rank highly when relevant documents exist in several languages. Another challenge is monetization of interactive search. If the search engine chooses to present advertisements to the user, they must be in a language understandable by the user. Moreover, the advertisement should be connected with the topic of the search and search results, and they need to be for products or services that can actually be provided to that searcher.

One natural direction for CLIR research is continuation of its integration with multimedia retrieval. With the rise of voice assistants (e.g., 39.4% of Internet users in US used a voice assistant at least once a month in 2019)[39] and thus the increasing popularity of conversational search, cross-language conversational

---

[39] https://mobidev.biz/blog/future-ai-machine-learning-trends-to-impact-business

search may become another emerging topic. As language support by voice assistants is still somewhat limited,[40] one may expect that CLIR in conversational search might become particularly important for languages with fewer speakers. In such cases, the relevant content might only exist in a language different from the language of the speaker, and the voice assistant will need to decide whether and how to present answers from such content to the speaker. Cross-language conversational search will thus face new challenges, such as determining the probability that the top relevant document really contains the answer to the user's question, determining whether the answer differs in different languages, or exploring how to present the answer to the user if it only exists in a language different from the query language.

One notable gap in the CLIR literature is the application of CLIR to what has been called microtext: short informal text such as that found on Twitter, Short Message Service (SMS) text messages, or text chat. Such sources pose unusual challenges such as colloquial abbreviations (e.g., 'u r fine'), intentional misspellings (e.g., 'pleeeeeeze'), and device-influenced transliteration conventions. Mixed-language context is common, and many types of information occupy the same channel (e.g., Twitter contains professional announcements, thoughts or activity descriptions, and a good deal of spam). While there has been some experimentation with CLIR from informal content [12], and some work on translation of Twitter content that could perhaps be applied to CLIR (including one Twitter MT paper that actually uses standard CLIR techniques for data augmentation [74]), the dearth of work on CLIR from microtext seems somewhat surprising. It thus seems reasonable to anticipate that this topic may attract attention in the coming years.

Language-based search in speech and video is still an under-explored area, given the vast quantities of content that might be found in such sources. Monolingual search in podcasts recently received attention thanks to increasing commmercial interest in that content,[41] and podcasts offer a natural source of content for cross-language speech retrieval. CLIR for spoken content is today seen as a straightforward extension to CLIR, but of course speech is richer in some ways than text (e.g., with evidence for mood, emotions, or speaker accent). It seems plausible that at least some of this potential will one day be explored in a CLIR context.

Image content might be easier to present to speakers of different languages then speech or text, as noted by Clough and Sanderson [32]. Indeed, caption-based cross-language image search has been the focus of tasks at both Image CLEF and in the iCLEF interactive task. Work on this problem has continued in recent years with CLIR for metadata that describes both images [128, 102] and video [86]. The recent emergence of systems that learn to generate captions directly from image content opens new directions for image retrieval that has yet to be explored in CLIR applications.

---

[40] https://www.globalme.net/blog/language-support-voice-assistants-compared
[41] https://podcastsdataset.byspotify.com

Another emerging topic in Information Retrieval is fairness and bias. Similarly to neural networks and conversational search, these were identified as emerging IR topics by the 2018 SWIRL report [3], and they might reasonably be expected to ultimately be as important in CLIR research as in monolingual IR. A CLIR system might, for example, be biased towards returning documents from a certain category (e.g., research papers from universities might be returned more often than the research papers from government organizations). Especially in multilingual CLIR, one might ask if the system is (or should be!) biased towards returning documents from particular languages. So, as with many things, CLIR simply adds an additional dimension of complexity.

As we have sought to illustrate in this chapter, CLIR is yesterday's problem—much of the groundwork has already been laid. It is also today's problem, since the Web provides the potential to access a vast cornucopia of knowledge that, without CLIR, most of its users simply would not be able to find. And it is tomorrow's problem, since wherever IR research goes, CLIR seems sure to follow!

Acknowledgements

Bibliography

[1] S. Abdou and J. Savoy. Report on CLIR task for the NTCIR-5 evaluation campaign. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, Tokyo, Japan, 2005.

[2] E. Agirre, G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. CLEF 2008: Ad Hoc track overview. In *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 15–37, Aarhus, Denmark, 2009.

[3] J. Allan and et al. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in Lorne (SWIRL 2018). *SIGIR Forum*, 52(1):34–90, Aug. 2018.

[4] W. Ammar, G. Mulcaire, Y. Tsvetkov, G. Lample, C. Dyer, and N. A. Smith. Massively multilingual word embeddings. *arXiv:1602.01925*, 2016.

[5] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proceedings of CIKM*, page 601–610, Hong Kong, China, 2009.

[6] M. Artetxe and H. Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 09 2019. ISSN 2307-387X.

[7] M. Artetxe, G. Labaka, and E. Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th ACL*, pages 789–798, Melbourne, Australia, 2018.

[8] M. Artetxe, S. Ruder, and D. Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, July 2020. ACL.

[9] A. Asai, J. Kasai, J. H. Clark, K. Lee, E. Choi, and H. Hajishirzi. XOR QA: Cross-lingual open-retrieval question answering, 2020.

[10] O. Babko-Malaya. Annotation of nuggets and relevance in GALE distillation evaluation. In *Proceedings of the LREC*, Marrakech, Morocco, 2008.

[11] O. Babko-Malaya, D. Hunter, C. Fournelle, and J. White. Evaluation of document citations in phase 2 GALE distillation. In *Proceedings of the Seventh LREC*, Valletta, Malta, 2010.

[12] M. Bagdouri, D. W. Oard, and V. Castelli. CLIR for informal content in Arabic forum posts. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1811–1814, 2014.

[13] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[14] L. Ballesteros and M. Sanderson. Addressing the lack of direct translation resources for cross-language retrieval. In *Proceedings of the 2003 ACM CIKM*, pages 147–152, New Orleans, Louisiana, USA, 2003.

[15] P. Bhattacharya, P. Goyal, and S. Sarkar. Query translation for cross-language information retrieval using multilingual word clusters. In *Proceedings of the Workshop on South Southeast Asian Natural Language Processing (WSSANLP)*, pages 152–162, Osaka, Japan, 2016.

[16] P. Bhattacharya, P. Goyal, and S. Sarkar. Using word embeddings for query translation for Hindi to English cross language information retrieval. *Computación y Sistemas*, 20(3):435–447, 2016.

[17] T. Bi, L. Yao, B. Yang, H. Zhang, W. Luo, and B. Chen. Constraint translation candidates: A bridge between neural query translation and cross-lingual information retrieval. *arXiv preprint arXiv:2010.13658*, 2020.

[18] H. Bonab, S. M. Sarwar, and J. Allan. Training effective neural CLIR by bridging the translation gap. In *Proceedings of SIGIR*, page 9–18, Xi'an, China, 2020.

[19] L. H. Bonifacio, I. Campiotti, V. Jeronymo, R. Lotufo, and R. Nogueira. mmarco: A multilingual version of the MS MARCO passage ranking dataset. *arXiv preprint arXiv:2108.13897*, 2021.

[20] E. Boschee, J. Barry, J. Billa, M. Freedman, T. Gowda, C. Lignos, C. Palen-Michel, M. Pust, B. K. Khonglah, S. Madikeri, J. May, and S. Miller. SARAL: A low-resource cross-lingual domain-focused information retrieval system for effective rapid document triage. In *Proceedings of ACL: System Demonstrations*, pages 19–24, Florence, Italy, July 2019. Association for Computational Linguistics.

[21] M. Braschler. CLEF 2000 — overview of results. In *Cross-Language Information Retrieval and Evaluation*, pages 89–101, Darmstadt, Germany, 2001.

[22] M. Braschler. CLEF 2001 — overview of results. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems*, pages 9–26. Springer Berlin Heidelberg, 2002.

[23] M. Braschler. CLEF 2002 — overview of results. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Advances in Cross-Language Information Retrieval*, pages 9–27, 2003. ISBN 978-3-540-45237-9.

[24] M. Braschler. Combination approaches for multilingual text retrieval. *Information Retrieval*, 7:183–204, 01 2004.

[25] C. Buckley, M. Mitra, J. Walz, and C. Cardie. Using clustering and superconcepts within SMART: TREC 6. *Information Processing & Management*, 36(1):109–131, 2000.

[26] L. Cappellato, N. Ferro, J. Nie, and L. Soulier, editors. *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*, 2018. CEUR-WS.org.

[27] N. Chen, R. E. Banchs, M. Zhang, X. Duan, and H. Li. Report of NEWS 2018 named entity transliteration shared task. In *Proceedings of the Seventh Named Entities Workshop*, pages 55–73, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[28] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[29] X. Chen and C. Cardie. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Oct.-Nov. 2018. ACL.

[30] C. Cieri, D. Graff, M. Liberman, N. Martey, and S. Strassel. The TDT-2 text and speech corpus. *Proceedings of DARPA Broadcast News Workshop*, 08 2000.

[31] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[32] P. Clough and M. Sanderson. User experiments with the eurovision cross-language image retrieval system. *Journal of the American Society of Information Science and Technology*, 57(5):697 –708, March 2006.

[33] J. Cossu, J. Gonzalo, M. Hajjem, O. Hamon, C. Latiri, and E. San-Juan. CLEF MC2 2018 lab technical overview of cross language microblog search and argumentative mining. In L. Cappellato, N. Ferro, J. Nie, and L. Soulier, editors, *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

[34] M. Coury, E. Salesky, and J. Drexler. Finding relevant data in a sea of languages. Technical report, MIT Lincoln Laboratory, 2016.

[35] W. B. Croft. Combining approaches to information retrieval. In W. B. Croft, editor, *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*, pages 1–36. Springer, Boston, MA, 2002.

[36] J. S. Culpepper and O. Kurland. Fusion in information retrieval. SIGIR Tutorial, 2018.

[37] G. Da San Martino, S. Romeo, A. Barroón-Cedeño, S. Joty, L. Maàrquez, A. Moschitti, and P. Nakov. Cross-language question re-ranking. In *Proceedings of SIGIR*, page 1145–1148, New York, NY, USA, 2017. Association for Computing Machinery.

[38] Z. Dai and J. Callan. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 985–988, 2019.

[39] K. Darwish and D. W. Oard. Probabilistic structured query methods. In C. L. A. Clarke, G. V. Cormack, J. Callan, D. Hawking, and A. F. Smeaton, editors, *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada*, pages 338–344. ACM, 2003.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[41] G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. CLEF 2006: Ad hoc track overview. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of*

Multilingual and Multi-modal Information Retrieval, pages 21–34. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-74999-8.

[42] G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. CLEF 2007: Ad hoc track overview. In C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, A. Peñas, V. Petras, and D. Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval*, pages 13–32. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-85760-0.

[43] F. Diaz. Indri, 2018. URL https://github.com/diazf/indri.

[44] S. Dwivedi and G. Chandra. A survey on cross language information retrieval. *International Journal on Cybernetics & Informatics*, 5:127–142, 02 2016.

[45] M. Faruqui and C. Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, 2014.

[46] M. Federico and N. Bertoldi. Statistical cross-language information retrieval using n-best query translations. In K. Järvelin, M. Beaulieu, R. A. Baeza-Yates, and S. Myaeng, editors, *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, pages 167–174. ACM, 2002.

[47] M. Federico and G. Jones. The CLEF 2003 cross-language spoken document retrieval track. In *CLEF*, volume 3237, page 646, 08 2003.

[48] M. Federico, N. Bertoldi, G.-A. Levow, and G. J. F. Jones. CLEF 2004 cross-language spoken document retrieval track. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images*, pages 816–820. Springer Berlin Heidelberg, 2005.

[49] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*, 2020.

[50] P. Forner, A. Peñas, E. Agirre, I. Alegria, C. Forascu, N. Moreau, P. Osenova, P. Prokopidis, P. Rocha, B. Sacaleanu, R. Sutcliffe, and E. Sang. Overview of the CLEF 2008 multilingual question answering track. In *CLEF*, pages 262–295, 01 2008.

[51] N. Fuhr. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51(3):32–41, 2017.

[52] G. W. Furnas, S. C. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In Y. Chiaramella, editor, *SIGIR'88, Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 465–480. ACM, 1988.

[53] P. Galuščáková, D. W. Oard, J. Barrow, S. Nair, H.-C. Shing, E. Zotkina, R. Eskander, and R. Zhang. MATERIALizing Cross-Language Information Retrieval: A Snapshot. In *Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop*, pages 14–21, 2020.

[54] F. Gey, R. Larson, N. Kando, J. Machado, and T. Sakai. NTCIR-GeoTime overview: Evaluating geographic and temporal search. In *NTCIR*, 07 2010.

[55] F. Gey, R. Larson, J. Machado, and M. Yoshio. NTCIR9-GeoTime overview-evaluating geographic and temporal search: Round 2. *Proceedings of NTCIR-9 Workshop*, pages 9–17, 01 2011.

[56] F. C. Gey and A. Chen. TREC-9 cross-language information retrieval (English-Chinese) overview. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA*, volume 500-249 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2000.

[57] F. C. Gey and D. W. Oard. The TREC-2001 cross-language information retrieval track: Searching Arabic using English, French or Arabic queries. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of The Tenth Text REtrieval Conference, TREC 2001, Gaithersburg, Maryland, USA*, volume 500-250 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2001.

[58] D. Giampiccolo, P. Forner, J. Herrera, A. Peñas, C. Ayache, C. Forascu, V. Jijkoun, P. Osenova, P. Rocha, B. Sacaleanu, and R. Sutcliffe. Overview of the CLEF 2007 multilingual question answering track. In C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, A. Peñas, V. Petras, and D. Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval*, pages 200–236. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-85760-0.

[59] L. Goeuriot, L. Kelly, H. Suominen, L. Hanlen, A. Névéol, C. Grouin, J. Palotti, and G. Zuccon. Overview of the CLEF eHealth evaluation lab 2015. In J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. Jones, E. San Juan, L. Capellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 429–443, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24027-5.

[60] L. Goeuriot, L. Kelly, H. Suominen, A. Névéol, A. Robert, E. Kanoulas, R. Spijker, J. Palotti, and G. Zuccon. CLEF 2017 eHealth evaluation lab overview. In *CLEF*, pages 291–303, 08 2017. ISBN 978-3-319-65812-4.

[61] J. Gonzalo and D. W. Oard. iCLEF 2004 track overview: Interactive cross-language question answering. In F. Borri, C. Peters, and N. Ferro, editors, *Working Notes for CLEF 2004 Workshop co-located with the 8th European Conference on Digital Libraries (ECDL 2004), Bath, UK, September 15-17, 2004*, volume 1170 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2004.

[62] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[63] D. Graff, C. Cieri, S. Strassel, and N. Martey. The TDT-3 text and speech corpus. In *in Proceedings of DARPA Broadcast News Workshop*, pages 57–60. Morgan Kaufmann, 1999.

[64] J. Guo, Y. Fan, Q. Ai, and W. B. Croft. 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, 2016.

[65] D. Gupta, S. Kumari, A. Ekbal, and P. Bhattacharyya. MMQA: A multi-domain multi-lingual question-answering framework for English and Hindi. In N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9.

[66] P. Gupta, R. E. Banchs, and P. Rosso. Continuous space models for CLIR. *Information Processing & Management*, 53(2):359–370, 2017.

[67] Y. Hoshen and L. Wolf. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

[68] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization, 2020.

[69] Q. Hu, H.-F. Yu, V. Narayanan, I. Davchev, R. Bhagat, and I. S. Dhillon. Query transformation for multi-lingual product search. In *SIGIR cComm Workshop*, 2020.

[70] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338, 2013.

[71] Z. Huang, H. Bonab, S. M. Sarwar, R. Rahimi, and J. Allan. Mixed attention transformer for leveraging word-level knowledge to neural cross-lingual information retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 760–770, 2021.

[72] K. Hui, A. Yates, K. Berberich, and G. De Melo. 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, 2018.

[73] S. H.-C. A. Intelligence. Artificial intelligence index annual report 2019, 2019. URL `https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf`.

[74] L. Jehl, F. Hieber, and S. Riezler. Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the seventh workshop on statistical machine translation*, pages 410–421, 2012.

[75] Z. Jiang, A. El-Jaroudi, W. Hartmann, D. Karakos, and L. Zhao. Cross-lingual information retrieval with BERT. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 26–31, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-55-9.

[76] Jimmy, G. Zuccon, J. Palotti, L. Goeuriot, and L. Kelly. Overview of the CLEF 2018 consumer health search task. In *CLEF*, 2018.

[77] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium, 2018. ACL.

[78] N. Kando. Overview of the sixth NTCIR workshop. In *NTCIR*, 04 0002.

[79] D. Karakos, R. Zbib, W. Hartmann, R. Schwartz, and J. Makhoul. Reformulating Information Retrieval from Speech and Text as a Detection Problem. In *Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop*, pages 38–43, 2020.

[80] K. Kayode and E. Ayetiran. Survey on cross-lingual information retrieval. *International Journal of Scientific and Engineering Research*, 9, 10 2018.

[81] L. Kelly, L. Goeuriot, H. Suominen, T. Schreck, G. Leroy, D. Mowery, S. Velupillai, W. Chapman, D. Martinez, G. Zuccon, and J. Palotti. Overview of the ShARe/CLEF eHealth evaluation lab 2014. In *CLEF*, volume 8685, 09 2014.

[82] L. Kelly, L. Goeuriot, H. Suominen, A. Névéol, J. Palotti, and G. Zuccon. Overview of the CLEF eHealth evaluation lab 2016. In N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 255–266, Cham, 2016. Springer International Publishing. ISBN 978-3-319-44564-9.

[83] L. Kelly, L. Goeuriot, H. Suominen, A. Névéol, J. Palotti, and G. Zuccon. Overview of the CLEF eHealth evaluation lab 2016. In *CLEF*, volume 9822, pages 255–266, 09 2016. ISBN 978-3-319-44563-2.

[84] L. Kelly, H. Suominen, L. Goeuriot, M. Neves, E. Kanoulas, D. Li, L. Azzopardi, R. Spijker, G. Zuccon, H. Scells, and J. Palotti. Overview of the CLEF eHealth evaluation lab 2019. In F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 322–339, Cham, 2019. Springer International Publishing. ISBN 978-3-030-28577-7.

[85] O. Khattab and M. Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164.

[86] A. Khwileh, D. Ganguly, and G. J. F. Jones. Utilisation of metadata fields and query expansion in cross-lingual search of user-generated internet video. *J. Artif. Int. Res.*, 55(1):249–281, Jan. 2016. ISSN 1076-9757.

[87] K. Kishida and N. Kando. A hybrid approach to query and document translation using a pivot language for cross-language information retrieval. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *Accessing Multilingual Information*

Repositories, pages 93–101. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-45700-8.

[88] K. Kishida, K. Chen, S. Lee, K. Kuriyama, N. Kando, H. Chen, S. Myaeng, and K. Eguchi. Overview of CLIR task at the fourth NTCIR workshop. In N. Kando and H. Ishikawa, editors, *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization, NTCIR-4, National Center of Sciences, Tokyo, Japan, June 2-4, 2004.* National Institute of Informatics (NII), 2004.

[89] M. Kluck. The domain-specific track in CLEF 2004: Overview of the results and remarks on the assessment process. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images*, pages 260–270. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-32051-7.

[90] M. Kluck and M. Stempfhuber. Domain-specific track CLEF 2005: Overview of results and approaches, remarks on the assessment analysis. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *Accessing Multilingual Information Repositories*, pages 212–221. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-45700-8.

[91] T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, 2018. ACL.

[92] G. Lample and A. Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[93] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[94] T. K. Landauer and M. L. Littman. A statistical method for language-independent representation of the topical content of text segments. In *Proceedings of the 11th International Conference: Expert Systems and Their Applications, 1991*, 1991.

[95] D. Lawrie, J. Mayfield, D. W. Oard, and E. Yang. HC4: A new suite of test collections for ad hoc CLIR. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR)*, 2022.

[96] J. H. Lee. Analyses of multiple evidence combination. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, page 267–276, New York, NY, USA, 1997. Association for Computing Machinery. ISBN 0897918363.

[97] S. Lee, S.-H. Myaeng, H. Kim, J. Seo, B. Lee, and S. Cho. Characteristics of the Korean test collection for CLIR in NTCIR-3. In *NTCIR*, 01 2002.

[98] P. Lewis, B. Oğuz, R. Rinott, S. Riedel, and H. Schwenk. MLQA: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*, 2019.

[99] B. Li and P. Cheng. Learning neural representation for CLIR with adversarial framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1861–1870, 2018.

[100] J. Li, C. Liu, J. Wang, L. Bing, H. Li, X. Liu, D. Zhao, and R. Yan. Cross-lingual low-resource set-to-description retrieval for global e-commerce. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8212–8219. AAAI Press, 2020.

[101] Q. Li, S. H. Myaeng, Y. Jin, and B.-Y. Kang. Translation of unknown terms via web mining for information retrieval. In *Asia Information Retrieval Symposium*, pages 258–269. Springer, 2006.

[102] X. Li, C. Xu, X. Wang, W. Lan, Z. Jia, G. Yang, and J. Xu. COCO-CN for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360, 2019.

[103] Y. Li, M. Franz, M. A. Sultan, B. Iyer, Y.-S. Lee, and A. Sil. Learning cross-lingual IR from an English retriever. *arXiv preprint arXiv:2112.08185*, 2021.

[104] J. Lin. The neural hype, justified! a recantation. *SIGIR Forum*, 53(2): 88–93, 2019. ISSN 0163-5840.

[105] J. Lin, R. Nogueira, and A. Yates. Pretrained transformers for text ranking: BERT and beyond, 2020.

[106] C. Ling, B. Steichen, and A. G. Choulos. A comparative user study of interactive multilingual search interfaces. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, page 211–220, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450349253.

[107] R. Litschko, G. Glavaš, S. P. Ponzetto, and I. Vulić. 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, 2018.

[108] R. Litschko, G. Glavaš, I. Vulic, and L. Dietz. Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1109–1112, 2019.

[109] R. Litschko, I. Vulić, S. P. Ponzetto, and G. Glavaš. On cross-lingual retrieval with multilingual text encoders. *arXiv preprint arXiv:2112.11031*, 2021.

[110] R. Litschko, I. Vulić, S. P. Ponzetto, and G. Glavaš. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *European Conference on Information Retrieval*, pages 342–358. Springer, 2021.

[111] R. Litschko, I. Vulić, and G. Glavaš. Parameter-efficient neural reranking for cross-lingual and multilingual retrieval, 2022. URL `https://arxiv.org/abs/2204.02292`.

[112] M. L. Littman, S. T. Dumais, and T. K. Landauer. Automatic cross-language information retrieval using latent semantic indexing. In *Cross-language information retrieval*, pages 51–62. Springer, 1998.

[113] M. L. Littman, F. Jiang, and G. A. Keim. Learning a language-independent representation for terms from a partially aligned corpus. In *ICML*, pages 314–322, 1998.

[114] J. Liu, Y. Lin, Z. Liu, and M. Sun. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of ACL*, pages 2358–2368, Florence, Italy, July 2019. ACL.

[115] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[116] E. Loginova, S. Varanasi, and G. Neumann. Towards multilingual neural question answering. In A. Benczúr, B. Thalheim, T. Horváth, S. Chiusano, T. Cerquitelli, C. Sidló, and P. Z. Revesz, editors, *New Trends in Databases and Information Systems*, pages 274–285, Cham, 2018. Springer International Publishing.

[117] S. Longpre, Y. Lu, and J. Daiber. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *CoRR*, abs/2007.15207, 2020.

[118] M. Lupu, A. Fujii, D. W. Oard, M. Iwayama, and N. Kando. *Patent-Related Tasks at NTCIR*, pages 77–111. Springer Berlin Heidelberg, 2017. ISBN 978-3-662-53817-3.

[119] S. MacAvaney, F. M. Nardini, R. Perego, N. Tonellotto, N. Goharian, and O. Frieder. Efficient document re-ranking for transformers by precomputing term representations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 49–58, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164.

[120] B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, M. Verdejo, and M. Rijke. The multiple language question answering track at CLEF 2003. In *CLEF*, volume 3237, 08 2003.

[121] B. Magnini, A. Vallin, C. Ayache, G. Erbach, A. Peñas, M. de Rijke, P. Rocha, K. Simov, and R. Sutcliffe. Overview of the CLEF 2004 multilingual question answering track. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images*, pages 371–391. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-32051-7.

[122] B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, V. Jijkoun, P. Osenova, A. Peñas, P. Rocha, B. Sacaleanu, and R. Sutcliffe. Overview of the CLEF 2006 multilingual question answering track. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Re-*

trieval, pages 223–256. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-74999-8.

[123] P. Majumder, F. (Workshop), and A. D. Library. *FIRE 2012 & 2013 : Post-proceedings of the 4th and 5th Workshops of the Forum for Information Retrieval Evaluation: Fourth International Workshop, FIRE 2012, Kolkata, India, December 19-21, 2012 and Fifth International Workshop, FIRE 2013 New Delhi, India.* ACM, 2007.

[124] P. Majumder, M. Mitra, P. Bhattacharyya, L. V. Subramaniam, D. Contractor, and P. Rosso, editors. *Multilingual Information Access in South Asian Languages - Second International Workshop, FIRE 2010, Gandhinagar, India, February 19-21, 2010 and Third International Workshop, FIRE 2011, Bombay, India, December 2-4, 2011, Revised Selected Papers*, volume 7536 of *Lecture Notes in Computer Science*, 2013. Springer. ISBN 978-3-642-40086-5.

[125] J. S. McCarley. Should we translate the documents or the queries in cross-language information retrieval? In R. Dale and K. W. Church, editors, *27th Annual Meeting of the Association for Computational Linguistics, University of Maryland, College Park, Maryland, USA, 20-26 June 1999*, pages 208–214. ACL, 1999.

[126] P. McNamee. *Textual representations for corpus-based bilingual retrieval.* PhD thesis, University of Maryland, Baltimore County, 2008.

[127] P. McNamee and J. Mayfield. Comparing cross-language query expansion techniques by degrading translation resources. In K. Järvelin, M. Beaulieu, R. A. Baeza-Yates, and S. Myaeng, editors, *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, pages 159–166. ACM, 2002.

[128] E. Ménard and V. Girouard. Image retrieval with SINCERITY: A search engine designed for our multilingual world! *OCLC Syst. Serv.*, 31(4):204–218, 2015.

[129] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479, 2005.

[130] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[131] T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013.

[132] T. Mitamura, H. Shima, T. Sakai, N. Kando, T. Mori, K. Takeda, C.-Y. Lin, C.-J. Lin, and C.-W. Lee. Overview of the NTCIR-8 ACLIA tasks: Advanced cross-lingual information access. *Proceedings of 8th NTCIR Workshop Meeting*, 01 2010.

[133] M. Mitra and P. Majumdar. FIRE: Forum for information retrieval evaluation. In *Proceedings of the 2nd workshop on Cross Lingual Information Access (CLIA) Addressing the Information Need of Multilingual Societies*, 2008.

[134] S. Nair, P. Galuscakova, and D. W. Oard. Combining contextualized and non-contextualized query translations to improve CLIR. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1581–1584, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164.

[135] S. Nair, A. Ragni, O. Klejch, P. Galuščáková, and D. Oard. Experiments with cross-language speech retrieval for lower-resource languages. In *AIRS*, pages 145–157, 02 2020. ISBN 978-3-030-42834-1.

[136] S. Nair, E. Yang, D. Lawrie, K. Duh, P. McNamee, K. Murray, J. Mayfield, and D. W. Oard. Transfer learning approaches for building cross-language dense retrieval models. In M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, and V. Setty, editors, *Advances in Information Retrieval*, pages 382–396, Cham, 2022. Springer International Publishing. ISBN 978-3-030-99736-6.

[137] R. P. Neco and M. L. Forcada. Asynchronous translations with recurrent neural nets. In *Proceedings of International Conference on Neural Networks (ICNN'97)*, volume 4, pages 2535–2540. IEEE, 1997.

[138] J.-Y. Nie. Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies*, 3(1):1–125, 2010.

[139] V. Nikoulina and S. Clinchant. Domain adaptation of statistical machine translation models with monolingual data for cross lingual information retrieval. In *European Conference on Information Retrieval*, pages 768–771. Springer, 2013.

[140] V. Nikoulina, B. Kovachev, N. Lagos, and C. Monz. Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 109–119, Avignon, France, Apr. 2012. Association for Computational Linguistics.

[141] NIST. *The Official Original Derivation of AQWV*, 2017. URL https://www.nist.gov/system/files/documents/2017/10/26/aqwv_derivation.pdf.

[142] R. Nogueira and K. Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.

[143] G. M. D. Nunzio, N. Ferro, G. J. F. Jones, and C. Peters. CLEF 2005: Ad hoc track overview. In C. Peters and N. Ferro, editors, *Working Notes for CLEF 2005 Workshop co-located with the 9th European Conference on Digital Libraries (ECDL 2005), Wien, Austria, September 21-22, 2005*, volume 1171 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2005.

[144] D. W. Oard and A. R. Diekema. Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 33:223–56, 1998.

[145] D. W. Oard and F. Ertunc. Translation-based indexing for cross-language retrieval. In *European Conference on Information Retrieval*, pages 324–333. Springer, 2002.

[146] D. W. Oard and F. C. Gey. The TREC 2002 Arabic/English CLIR track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of The Eleventh Text REtrieval Conference, TREC 2002, Gaithersburg, Maryland, USA, November 19-22, 2002*, volume 500-251 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2002.

[147] D. W. Oard, J. Wang, G. J. Jones, R. W. White, P. Pecina, D. Soergel, X. Huang, and I. Shafran. Overview of the CLEF-2006 cross-language speech retrieval track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 744–758. Springer, 2006.

[148] W. Ogden, J. Cowie, M. Davis, E. Ludovik, S. Nirenburg, H. Molina-Salgado, and N. Sharples. Keizai: An interactive cross-language text retrieval system. In *Proceeding of the MT SUMMIT VII workshop on machine translation for cross language information retrieval*, volume 416, 1999.

[149] W. C. Ogden and M. W. Davis. Improving cross-language text retrieval with human interactions. In *33rd Annual Hawaii International Conference on System Sciences (HICSS-33), 4-7 January, 2000, Maui, Hawaii, USA*. IEEE Computer Society, 2000.

[150] J. R. M. Palotti, G. Zuccon, Jimmy, P. Pecina, M. Lupu, L. Goeuriot, L. Kelly, and A. Hanbury. CLEF 2017 task overview: The IR task at the eHealth evaluation lab - evaluating retrieval methods for consumer health search. In L. Cappellato, N. Ferro, L. Goeuriot, and T. Mandl, editors, *CLEF (Working Notes)*, volume 1866 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017.

[151] P. Pecina, P. Hoffmannová, G. Jones, Y. Zhang, and D. Oard. Overview of the CLEF-2007 cross-language speech retrieval track. In *CLEF*, pages 674–686, 01 2007.

[152] C. Peters. What happened in CLEF 2004? In *CLEF*, pages 919–919, 08 2005.

[153] C. Peters, M. Braschler, and P. Clough. *Multilingual Information Retrieval: From Research To Practice*. Computer Science. Springer Berlin Heidelberg, 2012.

[154] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. ACL.

[155] V. Petras and S. Baerisch. The domain-specific track at CLEF 2008. In C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. F. Jones, M. Kurimo, T. Mandl, A. Peñas, and V. Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 186–198. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-04447-2.

[156] V. Petras, S. Baerisch, and M. Stempfhuber. The domain-specific track at CLEF 2007. In C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, A. Peñas, V. Petras, and D. Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval*, pages 160–173. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-85760-0.

[157] D. Petrelli and E. Not. User-centred design of flexible hypermedia for a mobile guide: Reflections on the hyperaudio experience. *User Model. User Adapt. Interact.*, 16(1):85–86, 2006.

[158] A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 55–63. ACM, 1998.

[159] R. Rahimi, A. Montazeralghaem, and A. Shakery. An axiomatic approach to corpus-based cross-language information retrieval. *Inf. Retr. J.*, 23(3): 191–215, 2020.

[160] P. Resnik, D. Oard, and G. Levow. Improved cross-language retrieval using backoff translation. In *Proceedings of the First International Conference on Human Language Technology Research*, 2001.

[161] G. Rosemblat, D. Gemoets, A. C. Browne, and T. Tse. Machine translation-supported cross-language information retrieval for a consumer health resource. *AMIA Symposium*, pages 564–568, 2003.

[162] C. Rubino. The effect of linguistic parameters in cross-language information retrieval performance evidence from IARPA's MATERIAL program. In *Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop*, pages 1–6, 2020.

[163] A. Rücklé, K. Swarnkar, and I. Gurevych. Improved cross-lingual question retrieval for community question answering. In *The World Wide Web Conference*, WWW '19, page 3179–3186, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748.

[164] S. Ruder, I. Vulić, and A. Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631, 2019.

[165] T. Sakai. On Fuhr's guideline for IR evaluation. *SIGIR Forum*, 54(1):1–8, 2020.

[166] T. Sakai, N. Kando, C.-J. Lin, T. Mitamura, H. Shima, D. Ji, K.-h. Chen, and E. Nyberg. Overview of the NTCIR-7 ACLIA IR4QA task. In *NTCIR*, 01 0001.

[167] T. Sakai, D. W. Oard, and N. Kando. Evaluating information retrieval and access tasks: NTCIR's legacy of research impact, 2020.

[168] S. Saleh and P. Pecina. Reranking hypotheses of machine-translated queries for cross-lingual information retrieval. In *International Conference of the Cross-Language Evaluation Forum for European (CLEF)*, volume 9822, pages 54–66, 09 2016. ISBN 978-3-319-44563-2.

[169] S. Saleh and P. Pecina. An extended CLEF eHealth test collection for cross-lingual information retrieval in the medical domain. In L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra, editors, *Advances in Information Retrieval*, pages 188–195, Cham, 2019. Springer International Publishing. ISBN 978-3-030-15719-7.

[170] S. Saleh and P. Pecina. Document translation vs. query translation for cross-lingual information retrieval in the medical domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6849–6860, Online, July 2020. ACl.

[171] S. M. Sarwar, H. Bonab, and J. Allan. A multi-task architecture on relevance-based neural query translation. In *Proceedings of ACL*, pages 6339–6344, Florence, Italy, July 2019. Association for Computational Linguistics.

[172] S. Sasaki, S. Sun, S. Schamoni, K. Duh, and K. Inui. 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, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[173] Y. Sasaki, H. hsi Chen, K. hua Chen, and C. jie Lin. Overview of the NTCIR-5 cross-lingual question answering task (CLQA1). In *In Proceedings of NTCIR*, pages 175–185, 2005.

[174] Y. Sasaki, C.-J. Lin, K.-h. Chen, and H.-H. Chen. Overview of the NTCIR-6 cross-Lingual question answering (CLQA) task. In *NTCIR*, 04 2007.

[175] S. Schamoni, F. Hieber, A. Sokolov, and S. Riezler. 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, Baltimore, Maryland, June 2014. ACL.

[176] P. H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[177] M. Schuster and K. Nakajima. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, 2012.

[178] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, Aug. 2016. ACL.

[179] P. Sheridan and J. P. Ballerini. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 58–65, 1996.

[180] P. Sheridan, M. Wechsler, and P. Schäuble. Cross-language speech retrieval: Establishing a baseline performance. *SIGIR Forum*, 31(SI):99–108, July 1997. ISSN 0163-5840.

[181] P. Shi and J. Lin. Cross-lingual relevance transfer for document retrieval, 2019.

[182] H.-C. Shing, J. Barrow, P. Galuščáková, D. W. Oard, and P. Resnik. Unsupervised system combination for set-based retrieval with expectation maximization. In F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 191–197, Cham, 2019. Springer International Publishing. ISBN 978-3-030-28577-7.

[183] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[184] A. Søgaard, I. Vulić, S. Ruder, and M. Faruqui. Cross-lingual word embeddings. *Synthesis Lectures on Human Language Technologies*, 12(2):1–132, 2019.

[185] A. Sokolov, F. Hieber, and S. Riezler. Learning to translate queries for clir. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1179–1182, 2014.

[186] M. Stempfhuber and S. Baerisch. The domain-specific track at CLEF 2006: Overview of approaches, results and assessment. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval*, pages 163–169. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-74999-8.

[187] S. Strassel, C. Cieri, A. Cole, D. Dipersio, M. Liberman, X. Ma, M. Maamouri, and K. Maeda. Integrated linguistic resources for language exploitation technologies. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA).

[188] S. Sun and K. Duh. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online, 2020. ACL.

[189] E. Tang, S. Geva, A. Trotman, Y. Xu, and K. Itakura. Overview of the NTCIR-9 crosslink task: Cross-lingual link discovery. In N. Kando, D. Ishikawa, and M. Sugimoto, editors, *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 437–463. National Institute of Informatics, Japan, 2011.

[190] L.-X. Tang, I.-S. Kang, F. Kimura, Y.-H. Lee, A. Trotman, S. Geva, and Y. Xu. Overview of the NTCIR-10 cross-lingual link discovery task. In *NTCIR*, 07 2013.

[191] F. Türe and E. Boschee. Learning to translate: a query-specific combination approach for cross-lingual information retrieval. In *Proceedings of the*

2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 589–599, 2014.

[192] F. Türe, J. Lin, and D. Oard. Combining statistical translation techniques for cross-language information retrieval. In *Proceedings of COLING 2012*, pages 2685–2702, Mumbai, India, Dec. 2012.

[193] A. Vallin, B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peñas, M. de Rijke, B. Sacaleanu, D. Santos, and R. Sutcliffe. Overview of the CLEF 2005 multilingual question answering track. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *Accessing Multilingual Information Repositories*, pages 307–331. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-45700-8.

[194] E. Voorhees and D. Harman. Overview of the sixth text retrieval conference (TREC-6). *Information Processing & Management*, 36:3–35, 01 2000.

[195] E. M. Voorhees and D. K. Harman, editors. *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*, volume 500-246 of *NIST Special Publication*, 1999. National Institute of Standards and Technology (NIST).

[196] E. M. Voorhees, D. K. Harman, et al. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, 2005.

[197] I. Vulić and M.-F. Moens. 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, 2015.

[198] I. Vulić, G. Glavaš, R. Reichart, and A. Korhonen. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4398–4409, 2019.

[199] C. L. Wayne. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May 2000. European Language Resources Association (ELRA).

[200] R. W. White, D. W. Oard, G. J. F. Jones, D. Soergel, and X. Huang. Overview of the CLEF-2005 cross-language speech retrieval track. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *Accessing Multilingual Information Repositories*, pages 744–759. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-45700-8.

[201] S. Wu. *Data Fusion in Information Retrieval*. Springer, 2012. ISBN 3642288650.

[202] C. Xiong, Z. Dai, J. Callan, Z. Liu, and R. Power. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, pages 55–64, 2017.

[203] J. Xu and R. Weischedel. Cross-lingual information retrieval using hidden Markov models. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 95–103, Hong Kong, China, Oct. 2000. Association for Computational Linguistics.

[204] E. Yang, S. Nair, R. Chandradevan, R. Iglesias-Flores, and D. W. Oard. C3: Continued pretraining with contrastive weak supervision for cross language ad-hoc retrieval. *arXiv preprint arXiv:2204.11989*, 2022.

[205] P. Yang, H. Fang, and J. Lin. Anserini: Reproducible ranking baselines using Lucene. *J. Data and Information Quality*, 10(4), Oct. 2018. ISSN 1936-1955.

[206] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernandez Abrego, S. Yuan, C. Tar, Y.-h. Sung, B. Strope, and R. Kurzweil. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online, July 2020. Association for Computational Linguistics.

[207] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[208] L. Yao, B. Yang, H. Zhang, W. Luo, and B. Chen. Exploiting neural query translation into cross lingual information retrieval. *arXiv preprint arXiv:2010.13659*, 2020.

[209] M. Yarmohammadi, X. Ma, S. Hisamoto, M. Rahman, Y. Wang, H. Xu, D. Povey, P. Koehn, and K. Duh. 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, Aug. 2019. European Association for Machine Translation.

[210] Z. A. Yilmaz, W. Yang, H. Zhang, and J. Lin. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3481–3487, 2019.

[211] P. Yu and J. Allan. A study of neural matching models for cross-lingual ir. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1637–1640, New York, NY, USA, 2020. ACL. ISBN 9781450380164.

[212] P. Yu, H. Fei, and P. Li. Cross-lingual language model pretraining for retrieval. In *Proceedings of the Web Conference 2021*, pages 1029–1039, 2021.

[213] H. Zamani, M. Dehghani, W. B. Croft, E. Learned-Miller, and J. Kamps. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 497–506, 2018.

[214] I. Zavorin, A. Bills, C. Corey, M. Morrison, A. Tong, and R. Tong. Corpora for cross-language information retrieval in six less-resourced languages. In

Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop, pages 7–13, 2020.

[215] R. Zbib, L. Zhao, D. G. Karakos, W. Hartmann, J. DeYoung, Z. Huang, Z. Jiang, N. Rivkin, L. Zhang, R. M. Schwartz, and J. Makhoul. Neural-network lexical translation for cross-lingual IR from text and speech. In B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, and F. Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 645–654. ACM, 2019.

[216] L. Zhang and X. Zhao. An overview of cross-language information retrieval. In X. Sun, J. Wang, and E. Bertino, editors, *Artificial Intelligence and Security*, pages 26–37, Cham, 2020. Springer International Publishing. ISBN 978-3-030-57884-8.

[217] L. Zhang, D. Karakos, W. Hartmann, M. Srivastava, L. Tarlin, D. Akodes, S. K. Gouda, N. Bathool, L. Zhao, R. S. Zhuolin Jiang, and J. Makhoul. The 2019 BBN Cross-lingual Information Retrieval System. In *Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop*, pages 44–51, 2020.

[218] P. Zhang, L. Plettenberg, J. L. Klavans, D. W. Oard, and D. Soergel. Task-based interaction with an integrated multilingual, multimedia information system: a formative evaluation. In E. M. Rasmussen, R. R. Larson, E. G. Toms, and S. Sugimoto, editors, *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2007, Vancouver, BC, Canada, June 18-23, 2007, Proceedings*, pages 117–126. ACM, 2007.

[219] R. Zhang, C. Westerfield, S. Shim, G. Bingham, A. Fabbri, W. Hu, N. Verma, and D. Radev. Improving low-resource cross-lingual document retrieval by reranking with deep bilingual representations. In *Proceedings of ACL*, pages 3173–3179, Florence, Italy, July 2019.

[220] D. Zhou, M. Truran, T. Brailsford, V. Wade, and H. Ashman. Translation techniques in cross-language information retrieval. *ACM Comput. Surv.*, 45(1), Dec. 2012. ISSN 0360-0300.