Multilingual Information Access

Douglas W. Oard
College of Information Studies, University of Maryland, College Park

Keywords

Information retrieval, Information seeking behavior, Multilingual, Cross-lingual, Cross-language, CLIR, MLIA, Machine translation, MT.

Abstract

This article describes the process by which systems can be designed to help users find content in a language that may be different from the language of their query. The discussion of the relatively narrowly construed technical issues that are often referred to as Cross-Language Information Retrieval (CLIR) is situated in the context of important related issues such as information seeking behavior, interaction design, and machine translation.

1. Introduction

The central thesis of Tom Friedman's book "The World is Flat" is that we now live in a world in which technological innovation is creating opportunities for more seamless global interaction than has heretofore been possible [1]. It is important to recognize that "technological innovation" encompasses far more than mere technical innovation—equally important is our ability as a society to learn to productively use the technical capabilities that we can create. This chapter examines one such technology: helping users to find information in ways that "flatten" language barriers. In keeping with what is emerging as common usage, we refer to this challenge as "Multilingual Information Access" (MLIA).

This word "multilingual" can be used in many ways, so let us start by saying what we mean. A multilingual collection is a collection of documents that contains more than just a single language (e.g., English and Chinese). These documents may each contain just one language, or some of the documents might contain words from more than one language. Our interest is in helping a searcher to find the documents that they seek, regardless of the language in which they are expressed. For simplicity, we will assume in this chapter that documents are expressed in writing and stored as e-text (i.e., as digital sequences of character codes), but similar approaches have been applied to scanned documents and spoken word collections, and might in the future also be applied to visual languages (e.g., American Sign Language).

Who needs MLIA? We can envision at least two user groups. Perhaps the most obvious is so-called polyglots—people who are able to at least read (and perhaps write) more than one language. For example, more than one billion people who know at least some English are native speakers of some other language. Polyglots can benefit from MLIA in at least three ways: (1) they can find documents in more than one language with a single search, (2) they can formulate queries in the language(s) for which their active vocabulary is largest, and (3) they can move more seamlessly across languages over the course of an information seeking episode than would be possible if documents written in different languages were available only from different information systems. Monoglots (those who know only a single language) form a second important group. For example, many Americans can read only English, while many citizens of China can read only Chinese. Those populations essentially live in different worlds, worlds that MLIA can help to bridge.

MLIA always involves Cross-Language Information Retrieval (CLIR), in which queries in one language are used to find documents in another. When the user cannot read the document language, some form of translation service will usually be needed. This might be as simple as automatic translation of short snippets, or as complex as on-demand translation by a subject matter expert. There are, however, also cases in which adequate results might be presented without translation. For example, someone who knows only Japanese might search a collection of newspaper photographs that are indexed using only English terms and still easily recognize which of the resulting photographs would best meet their needs.

The remainder of this article is organized as follows. The next section places MILA in historical perspective and explains (or at least interprets) how and why modern techniques for MLIA evolved in the way that they did. Section 3 then describes the present state of the art for CLIR, the key technical capability in all MLIA application. Section 4 builds on that, broadening the coverage to address interaction design and information seeking processes. Finally, Section 5 concludes with a brief survey of the present state of practice and elucidation of some important open questions.

2. A Brief History of Multilingual Information Access

Gaining access to information in unfamiliar languages has always been an important problem. The intense technological competition that was emblematic of the Cold War in the second half of the twentieth century created a substantial demand on both sides for translation of scientific and technical papers. After some early, and rather disappointing, experiments with automatic translation, the United States National Research Council recommended in 1966 that basic research continue, but that the work of people, rather than machines, provide the principal means for making foreign-language information accessible for the foreseeable future [2]. This recommendation fostered the development of a part of the information industry that focused on translating scientific and technical literature and indexing those translations. Journal articles and so-called grey literature (e.g., technical reports) were translated either prospectively or on demand by a number of organizations, and the World Translations Index (and its predecessors) evolved to provide the needed indexing service by speakers of English.

The economic growth and linguistic diversity of Europe in the second half of the twentieth century provided the impetus for the second major innovation, the development of multilingual thesauri. Oard and Diekema surveyed the genesis of this work, from the first published report (in 1964 from Germany) through publication of the current (1985) version of ISO Standard 5964, which recommends techniques for construction of multilingual thesauri [3].

Substantial reliance on human translation and thesaurus-based indexing were good choices at the time, but three key events dramatically changed the opportunity space. The most obvious was the end of the Cold War, which resulted in substantial changes in national investment strategies. The International Translations Centre ceased operations in 1997 with the publication of the last volume of the World Translations Index, citing of declining demand for their services that resulted from increasing adoption of English as a lingua franca for scientific communication and from declining funding for information science more generally.

The second key event was the rise of the World-Wide Web, and in particular the widespread adoption of Web search engines such as Lycos, AltaVista, and Google. Furnas et al had remarked on what they referred to as the "vocabulary problem" in human-system interaction, observing that "new or intermittent users often use the wrong words and fail to get the actions or information they want" [4]. Although the fuzzy-match full-text search capabilities of the 1990's-era Web search engines were far from perfect, experience with that technology began the process of incrementally shifting expectations away from intermediated thesaurus-based search and toward end-user "natural language" search.

The third event, which attracted far less attention at the time, was a remarkable payoff from the investments in basic research that the National Research Council had recommended. Earlier approaches, based on hand-coded rules, had proven to be problematic because the rules could interact in ways that were difficult to anticipate. As a result, at some point adding additional rules in an effort to improve things could actually reduce translation quality. In 1990, a group at IBM Research first published a radical new technique based on one simple idea: machines can learn to translate by using statistical analysis to identify regularities in large collections of translations that were generated by people [5]. Importantly, as more examples are provided, translation quality improves. This "data-driven" approach, which came to be called statistical Machine Translation (MT), is thus well matched to a networked world in which assembling ever-larger collections is increasingly tractable.

These three events, unfolding together in the last decade of the twentieth century, came together to transform both the need for, and the opportunities to provide, automated techniques to support multilingual information access by end users. The spark that ignited the process was a 1996 workshop at an information retrieval conference in Zurich [6]. Early techniques were limited by their reliance on online bilingual dictionaries, but techniques based on statistical machine translation were soon introduced. As described in the next section, this ultimately yielded fuzzy-match full-text search capabilities that accommodate language differences between the queries and the documents remarkably well. End-user search requires more than just accurate ways of finding documents that may be useful, however. Equally important, the user must be able to recognize those useful documents, understand their contents, and (sometimes) draw on that understanding to progressively improve their queries; Section 5 addresses those issues.

3. Cross-Language Information Retrieval

The core capabilities that enable MLIA are indexing and query processing for CLIR. Indexing proceeds in three stages: (1) language and character set identification, (2) language-specific processing, and (3) construction of an “inverted index” that allows rapid identification of which documents contain specific terms. Sometimes the language in which a document is written and the character set used to encode it can be inferred from its source (e.g.., New York Times articles are almost always written in English, and typically encoded in ASCII) and sometimes the language and character set might be indicated using metadata (e.g., the HTML standard used for Web pages provides metadata fields for these purposes). In other cases (or to confirm an initial guess), a very simple form of content analysis can be used to identify languages and character sets. The usual approach is to count the frequency of character sequences, and then to guess the language based on similarity to counts computed in the same way for documents written in a known language. For example, the first sentence in this paragraph would yield the following 3-byte sequences: “the”, “he “, “e c”, “ co”, “cor”, “ore”, … The technique is easily extended to accommodate multi-byte character encodings by counting bytes rather than characters. Language and character set classification using this technique is remarkably accurate (typically well over 95%) for any text that is at least as long as a typical paragraph, so language switching within a single document can sometimes also be detected using this technique.

Once the language and character set are known, the character set can be concerted to a standard representation (often Unicode) and two types of language-specific processing are then typically applied: (1) tokenization to identify the terms that could be indexed, (2) stopword removal to identify terms that need not be indexed (for efficiency reasons). For English, tokenization typically involved splitting on white space and then using rule-based techniques to remove common endings (so-called “stemming”). Identifying “word” boundaries is more complex in “freely compounding” languages such as German, Finnish and Tamil, and in “unsegmented” languages such as Chinese. Of course, the spoken form of every language exhibits this same tendency to run words together without pauses, so techniques similar to those used in speech recognition for identifying words can be used to identify. The basic idea is to draw on two sources of evidence: we can know most of the words that exist in the language (using a dictionary), and we can guess which word sequences might make sense (e.g., from statistical analysis of word usage). Using these ideas together would tell us that the word “Washington” found in a German document might (from the dictionary) be segmented as “was”, “hing” and “ton”, but (from usage statistics) that such a segmentation would be unlikely to be correct—in this case we would therefore index the unsegmented word “washington”.

The inverted index used in CLIR is similar in structure to an inverted index used in any information retrieval system, but the information stored in that index may be different. Conceptually, an inverted index includes two parts: (1) a lookup table stored in fast main memory that can be used to rapidly find the “postings” for a specific term (i.e., identifiers for all documents containing that term), and (2) The postings file, which (because of its size) must be stored on the (much slower) hard disk. One of the most important advances in information retrieval system design in the past decade was the widespread introduction of automatic compression techniques for the postings file. Because these techniques are tuned to achieve the greatest compression for the most common terms, stopword removal is no longer essential as an efficiency measure in monolingual applications. In CLIR, however, deficiencies in the translation technique can sometimes yield inappropriate results for translation of very common words. Stopword removal is therefore still common in CLIR applications.

The most obvious distinguishing feature of CLIR is that some form of translation knowledge must be embedded in the system design, either at indexing time or at query time. Essentially, three approaches are possible: (1) translate each term using the context in which that word appears to help select the right translation, (2) count the terms and then translate the aggregate counts without regard to the context of individual occurrences, or (3) compute some more sophisticated aggregate “term weight” for each term, and then translate those weights. Somewhat surprisingly, the first two of these work about equally well in many cases; term weight translation is typically not competitive.

If the user will ultimately require a machine-generated translation, and if that translation is always into the same language, then a strong case can be made for translating every term in context at indexing time. In its simplest form (which is often adequate), this essentially amounts to simply running a full machine translation system as a preprocessing step prior to building the inverted index. Efficiency arguments against this approach would be hard to make: a translation system fast enough for responsive interactive use at query time would also be fast enough to process every document in all but the very largest collections at indexing time.

When full translation is not needed (e.g., for polyglot users), or when translations into many different languages may be needed to serve a linguistically diverse population, indexing documents using the terms in their original language is typically the preferred system architecture. In this case, considerable efficiency improvements can be obtained by translating term counts rather than term occurrences. The basic approach, first discovered by Pirkola [7] (SIGIR, 1998), is to count every possible query-language translation of each term as having been found in the document. Subsequent refinements resulted in further improvements from using translation probability for individual terms to estimate partial counts [8] and from aggregating translation probabilities for synonymous terms [9]. Regardless of the details, the key idea is to compute “term weights” in the query language rather than in the document language. Many of the commonly used term weighting formulae give more weight to rare terms than to common terms, which comports well with the way professional searchers are trained to enhance the precision of their search using terms that they expect will be highly specific. Since specificity is a feature of the query, it makes sense that computing term weights in the query language would work well.

Among all of the advances in CLIR, none has had anywhere near as large an effect as accurate translation probabilities. The best reported results for systems that lack any notion of translation probability (often called "dictionary-based" techniques) are in the range of 70% to 80% of what would have been achieved using queries written in the same language as the documents.[1] The best reported results for systems that use translation probabilities well is closer to 100% of what would have been achieved using same-language queries [9]. It is worth taking a moment to consider what that means—today, we can build systems to search French documents that work (approximately) equally well regardless of whether the query is written in French or in English! Of course, for any specific query the system might do better with French or with English, but on average over repeated use the best systems that can be built today do about equally well in CLIR or monolingual applications.

The key question, therefore, is how to obtain sufficiently accurate translation probabilities. It turns out that this problem was solved for us as one part of statistical MT [5]. The key idea behind statistical MT is that a machine that knows very little about language (e.g., just how to recognize a word) can learn to recognize (and later replicate) patterns of language use by counting what happens in very large collections of examples of language use. Specifically, we give our machine an enormous collection of examples of actual translations (e.g., "man in the moon" and "l'homme dans la lune") and ask it to find the most common alignments of individual terms (e.g., "man" and "l'homme" in this case, but "l'humanite" for "man" in "the evolution of modern man"). If the examples from which the machine learns are representative of the cases to which it will later be applied, the translation probabilities learned by the machine can be quite useful. A full MT system contains additional processing stages, but for CLIR it is often sufficient to simply use the learned translation probabilities directly (with some pruning to suppress the effect of relatively rare random alignments).

4. The Rest of the Story

There is, however, quite a bit more to the search process than simply automatically creating best-first rankings of documents that the user might wish to examine. Three key questions arise: (1) can people learn to formulate effective queries?, (2) can people recognize useful documents in the result set?, and (3) can people adequately understand the contents of those documents to meet their information needs? Research on these topics is still in its infancy, and moreover we can reasonably expect that as translation technology improves the answers to these questions may change. There will be, therefore, substantial scope for important Library and Information Science research on these questions for some time to come.

[1] These results are normally reported as an average across many topics. The most commonly reported search quality statistics in the CLIR literature is "average precision," which is designed to emphasize the density of relevant documents near the top of a ranked list where most searchers are expected to focus their attention.

The most tractable of these questions at present turns out to be the second one: people seem to be remarkably good at recognizing useful documents using even relatively poor translations. In 2001, the Cross-Language Evaluation Forum (CLEF) started an annual interactive track (iCLEF) to foster research on these questions. In that first year, the focus was on interactive assessment of topical relevance using machine translation. Representative users (in this case, university students) were presented with a written topic statement in a language they knew well (e.g., English) and a set of news stories in some other language that they did not know at all (e.g., Spanish) that had been ranked by a CLIR system and then automatically translated back into the language of the topic statement. Topical relevance judgments made by native speakers of the language in which the news stories were written were used as a gold standard. On average (over several users, each working on several topics), the searchers who did not know the document language agreed with the native speakers about as often as two native speakers would be expected to agree with each other [10].

Together, those studies indicate that recognizing relevant documents using automatic translation of short summaries or of entire documents is usually not a particularly difficult task. Considerable scope remains, however, for future research on optimally combining the technology for summary generation and translation, for analysis of specific cases in which present technology is not meeting user needs well, and for determining how best to present those results to the user (e.g., as several lists or as a single integrated list).

The third challenge, translating documents well enough that the user can understand their contents, is exactly the goal for which automatic systems for machine translation are optimized. Translation quality can be measured in two ways: (1) an “intrinsic” evaluation in which we ask how similar an automatic translation is to something that a human translator would actually create, or (2) an “extrinsic” evaluation in which we ask how well the reader can accomplish some task using the translation. Intrinsic evaluations provide an important way of assessing incremental progress in the design of machine translation systems, but extrinsic evaluation sheds more light on the ability of present translation technology to meet user needs in multilingual information access applications.

The iCLEF 2004 user studies provided an initial extrinsic evaluation of translation quality, measuring the user’s ability to answer factual questions when searching a large collection of news stories in an unfamiliar language. The results of those studies indicated that (on average, across users) only about 70% of the questions could be answered at all, and that (on average, across users and answered questions) only about 70% of those answers were correct. Considering both factors together, those factual questions were answered correctly about half the time [11].

Jones et al took this approach further, measuring the utility of an improved machine translation system across four source types (newspaper stories, text-only discussion groups, automatically transcribed news broadcasts, and automatically transcribed talk shows) using a reading comprehension test. They reported about 80% accuracy for answers to factual questions, but only about 50% accuracy for answers to questions that called for some degree of abstract reasoning [12]. From this we can conclude that present machine translation technology can satisfy some user needs, but that further improvements in translation quality will be needed before broadly useful multilingual access applications can be fielded.

The first of the questions posed at the start of this section, whether people can learn to formulate effective queries, is at this point the one we know the least about. The reason for this is simple—to learn very much about this would require long-term studies of real users performing real tasks. But before that can happen we must develop and field real systems capable of supporting those tasks, and those systems don't yet exist. Some insights have begun to accumulate from anecdotal reports of user experiences during structured user studies that have implications for system design. For example, sometimes searchers will recognize a useful term in a translated document and add it to their query, which will only work well if translation of documents and queries are implemented in a consistent manner (which was not the case in early systems). It also seems to be a good idea to inform users when no translation is known for a query term. It is not yet clear, how far to take this idea of informing the user—should we show them the translated query? All possible translations for each query term? Alternate translations for some of the terms in a summary or a full document? Google recently introduced a "translated search" capability that couples automatic query translation, automatic summary translation, and automatic Web page translation. Perhaps soon we will begin to see studies using this system that will begin to shed light on some of these questions.

5. Conclusion

Adoption of MLIA capabilities in deployed systems seems to have been far slower than the progress on the underlying CLIR technology would support. Deficiencies in current MT systems are undoubtedly a limiting factor in many cases, although applications intended for use by polyglot users would naturally be less affected by those deficiencies. Some cases may reflect a chicken-and-egg paradox: MLIA is needed only for large multilingual collections, but collection development policies in many cases predate the availability of these techniques. Web search would seem to be a natural first mover in MLIA (after all "World-Wide" is the Web's first name!), but there too adoption has been slower than the technology base would support. One commonly cited limiting factor for Web search engines has been the challenge of developing a suitable business model for monetizing MLIA. Regardless of the cause, it seems clear that developing a broader experience base with MLIA techniques will be an important next step.

In May, 2007, Google introduced a rudimentary MLIA capability by coupling query translation (to search Web pages in a different language) with document translation (to translate the result list, and individual results). Such an approach can be easily "bolted on" to any Web search engine since Web search engines typically already include automatic language identification and language-specific processing.[2] Similar techniques have been used for CLIR research by modifying freely available information retrieval

[2] Indeed, Yahoo announced a similar service for German users in 2005, although apparently without much success.

systems that had originally been designed for monolingual applications (e.g., Lucene), and at least one freely available system includes provisions for easily incorporating translation probabilities (Indri's "weighted structured queries"). Some degree of adoption is now also becoming evident among commercial providers of search services. For example, Autonomy now offers cross-language search capabilities (e.g., for enterprise search by multinational corporations).

These emerging capabilities are first steps in the direction of developing a richly multilingual information ecology that could support the next generation of research on information seeking behavior in such settings. The few studies that have been conducted in recent years have typically focused on single information systems, relatively narrowly scoped collections, and, of necessity, users who have no prior experience with any MLIA application. As more users gain access to a broader range of increasingly capable systems, richer and more nuanced study designs will be come possible.

Some open issues remain with regard to the technology base as well. For example, development costs for language-specific processing depend on the number of languages that must be accommodated, but the overall value of processing a specific language varies with the importance to the user of the languages that might be written in that document. With hundreds of written languages in use around the world, a point of diminishing returns may be reached beyond which the development costs for language-specific processing can no longer be justified. In such cases, a simple alternative is to count character sequences (in the same way as for language identification) and then simply index those character sequences. While this works reasonably well for monolingual applications in which the query and the document are written in the same language, how best to integrate translation capabilities into such an architecture is presently less clear.

Another open research question in MLIA is how best to present results from different languages. The challenge in this case arises because present systems for ranking documents in decreasing order of probable utility rely on relative "relevance" scores that lack an absolute meaning. The consequence is that we can reasonably hope to determine whether one French document is a better match to the query than another French document, but determining whether an English document is a better match to that query than some French document requires that we create some way of comparing English scores with French scores. Progress on this problem has to date been rather disappointing, with merged result lists often being far less satisfactory than the best single-language result set. Presenting several ranked lists, one per language, is possible, but that approach does not scale well as the number of languages grows.

Result set presentation is a special case of the more general issue of interaction design, for which the research to date has just started to scratch the surface. When first introduced, things that are new are often patterned on things that are already well understood. Newspapers, for example, initially resembled the pamphlets that had preceded them. Later, when newspapers first started providing content on the Web, it resembled a printed newspaper. So it should be no surprise that Google's first try at

MLIA looks like, well, Google. New capabilities tend to create their own dynamics, however, with new users bringing new needs, which drives development of new technologies, sometimes ultimately resulting in something that would have been difficult to imagine at the outset. MLIA has progressed far enough at this point for us to being on that path, but not nearly far enough for us to yet predict where that path will lead us.

Acknowledgments

The author is grateful to Noriko Kando for initial discussions that led to this article. This work has been supported in part by DARPA contract HR0011-06-C-0022 (GALE).

References

[1] Friedman, Thomas L., *The World is Flat: A Brief History of the Twenty-First Century*, Farrar, Straus and Giroux: New York, 2005.

[2] Pierce, J.R., et al., *Languages and Machines—Computers in Translation and Linguistics* (ALPAC Report), National Academy of Sciences, National Research Council: Washington, 1966.

[3] Oard, D.W. and Diekema, A.R., Cross-Language Information Retrieval, *Annual Review of Information Science and Technology* **1998**, Vol 33.

[4] Furnas, G.W., Landauer T.K., Gomez, L.M. and Dumais, S.T., The Vocabulary Problem in Human-System Communication, Communications of the ACM **1987**, 30(11), 964-971.

[5] Brown, P.F. et al., A Statistical Approach to Machine Translation, Computational Linguistics **1990**, 16(2)79-85.

[6] Grefenstette, G., Ed., *Cross-Language Information Retrieval*, Kluwer Academic: Boston, 1998.

[7] Pirkola, A., The Effects of Query Structire and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. In Croft, W.., et al., 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008), Melbourne, Australia, August 24-28, 1998; 55-63.

[8] Xu, J. and Weischedel, R., TREC-9 Cross-Lingual Retrieval at BBN, The Ninth Text Retrieval Conference (TREC-9), Gaithersburg, MD, November 13-16, 2000; 106-115.

[9] Wang, J. and Oard, D.W., Combining Bidirectional Translation and Synonymy for Cross-Language Information Retrieval. In Efthimiadis, E.N. et al., Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006); Seattle, August 6-11; 202-209.

[10] Oard, D.W., et. al., Interactive Cross-Language Document Selection, Information Retrieval **2004**, 7(1-2) ; 205-228.

[11] López-Ostenero, F., Gonzalo, J., Peinado, V. and Verdejo, F., Cross-Language Question Answering: Searching Pasajes vs. Searching Documents, In Peters, C. et al., *Multilingual Information Access for Text, Speech and Images*, 5th Workshop of the Cross-Language Evaluation Forum (CLEF 2004), Bath, UK, September 15-17, 2004; 323-333.

[12] Jones, D., et al., ILR-Basd MT Comprehension Test with Multi-Level Questions, Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2007), Companion Volume, Short Papers; Rochester, NY, April 22-27, 2007; 77-80.

For Further Reading

Excellent sources for the latest work on CLIR include the proceedings of the Cross-Language Evaluation Forum (CLEF), http://www.clef-campaign.org) in Europe, the NACSIS/NII Test Collection Information Retrieval (NTCIR) evaluations (http://research.nii.ac.jp/ntcir) in Japan, and the Forum for Information Retrieval Evaluation (FIRE) (http://www.isical.ac.in/~clia) in India. Contemporaneous reports on earlier CLIR research are also available from the Text Retrieval Conference (TREC) (http://trec.nist.gov) and the Topic Detection and Tracking (TDT) evaluations (http://www.nist.gov/speech/tests/tdt/).

For an historical perspective on the developments in MLIA, see the Annual Review of Information Science and Technology volume 33 (1998). For a broad forward looking treatment of the subject, see the papers and presentations fro the SIGIR 2006 workshop on New Directions in Multilingual Information Access (http://ucdata.berkeley.edu:7101/projects/sigir2006/program.htm).