

DHS Research Challenges in Cross-Language Information Retrieval

James Mayfield

The Johns Hopkins University Applied Physics Laboratory

1. The Cross-Language Problem

An important challenge facing DHS is to successfully handle documents that are written or spoken in languages other than English. There are three ways one might search foreign language collections. First, one might translate the documents into English, and use English retrieval techniques. This is an appealing solution in theory, as it allows English-speaking analysts to explore the document space without foreign language expertise. However, human translation is too expensive to apply to any but the most important documents, begging the question of how those documents are found in the first place. Machine translation, in contrast, is perhaps inexpensive enough to apply on a broad scale, but the quality of its output varies considerably with language, genre, transcription accuracy, *etc.*

The second approach is to perform document retrieval in the target language. There are both technical and organizational impediments to this approach. Information retrieval technology is well-developed for English, but less so for other languages. To the extent that accurate retrieval depends upon language-specific processes, such processing must be reworked for each new language to be handled. For example, a system that uses a rule-based stemmer to reduce morphological variation will need new language-specific rules to be properly applied to a new language. If a retrieval system relies on assumptions about the structure of English, entirely new capabilities may need to be developed to handle other languages. For example, most English retrieval systems assume that documents and queries can be broken into words by splitting text at spaces. However, languages such as Chinese do not place spaces between words. Foreign language retrieval also presents organizational challenges, because the need for expertise in a particular language will ebb and flow in response to world events.

The third alternative is to allow English queries to retrieve foreign language documents. This capability falls under the rubric of 'cross-language information retrieval' (CLIR). CLIR has been the focus of much recent work. TREC was the first large-scale evaluation of CLIR. More recently, the CLEF and NTCIR evaluations have focused on CLIR for European and Asian languages respectively.

CLIR is useful for several reasons. First, it allows an English-speaking analyst to identify foreign language documents that are likely to be useful before committing translation resources to them. Second, it assists the analyst who has weak knowledge of the target language. For students of a foreign language, it is easier to read or listen to the language than it is to generate new sentences in it. Thus, allowing the analyst to express queries in English lowers the language barrier for the analyst, and thereby increases the number of analysts that can effectively manipulate the data. Third, when documents written in more than one language must be searched, CLIR eliminates the need to manually rewrite the analyst's query in each of the target languages. Fourth, the technologies developed for effective CLIR will be applicable to other HLT tasks as those tasks are moved into a multilingual setting. For example, question answering typically relies on a phrase-level analysis of the question and of the documents that might contain answers to the question. Thus, phrase-level translation approaches developed for CLIR might benefit cross-language question answering as well.

2. The State of the Art

The big ideas in information retrieval, both monolingual and cross-language (and excluding the Web setting), that lead to today's level of performance, are the following:

Inverted indexes are the basic data structures that allow efficient full-text access to large text collections. An inverted index maps each indexing term to the set of documents that contain it. Compression and ordering in inverted indexes produce small indexes that provide rapid response time.

Pooled evaluation, the mechanism underlying TREC and the other IR evaluations, provides a way to measure performance over collections that are too large to be exhaustively judged. Many systems each provide a list of the top retrieved documents for a query. These documents are pooled and judged for relevance by a person. When averaged over many queries, these judgments provide an acceptable way to conclude that one system is superior to another.

Blind relevance feedback (BRF) is a way to augment a query with new terms that will likely improve search result quality. Retrieval is performed on the original query, and the top few retrieved documents are automatically examined for closely related terms. These terms are added to the original query to produce a new query, the results of which are returned to the user. The technique is 'blind' because there is no human intervention in the process. When applied before translation in a cross-language retrieval task, this technique is usually referred to as *pre-translation expansion*. Both blind relevance feedback and pre-translation expansion typically produce large gains in retrieval quality when averaged over many queries.

Language modeling is a general approach to assigning probabilities to events that involve natural language. A language model is simply a mechanism that produces strings of a language. We can build a simple language model from a document by generating sequences of words in proportion to their presence in the document. We can then ask the question 'what is the probability that a given language model will generate a particular query?' If we ask this question about the language models generated by a set of documents, we obtain a similarity metric that can be used to rank the documents--the higher the probability that a document's language model would generate the query, the better that document's rank. Language modeling is used across the spectrum of human language technologies, thereby tying those technologies together in a unified framework.

Parallel collections are sets of documents written in one language, each of which has a translation into a second language. Parallel collections are plentiful, albeit not in every desirable language pair. Parallel collections are useful for finding mappings from one language to another. The simplest such mapping is a dictionary of words in one language with their translations in the second. However, much more interesting relationships can be extracted from parallel collections; in the limit, alignments can be extracted to support full machine translation. The range of translations between these two extremes is ripe for exploitation, with applications not only in information retrieval, but also in question answering, summarization, and other cross-language tasks.

Language-neutral processing is the manipulation of text in ways that do not make assumptions about the language being processed. For example, word-based processing (which underlies many human language technologies) relies on the ability to identify words in language. However, in languages such as Chinese, breaks between words are not explicitly encoded in text. Thus, to do word-based processing in Chinese, one must first divide the target text into words using a (presumably language-specific) segmentation process. In contrast, processing that uses *character n-grams*, overlapping sequences of n characters, is language-neutral. While characterizing a text such as "Four score and seven" using the terms `four_`, `our_s`, `ur_sc`, *etc.*, seems peculiar (what is the semantic content of `ur_sc`?), n -gram

tokenization nonetheless performs well in retrieval tasks; it handles Chinese without the need for word segmentation, German without the need for decompounding, and OCR data without the need for spelling correction. Retrieval built on such language-neutral techniques is quickly retargetable to new languages as the need arises.

3. Research Directions

Performance on cross-language retrieval tasks roughly tracks performance on monolingual tasks for well-studied languages. This parity is sometimes used to argue that advances in CLIR demand advances in monolingual retrieval, and that it does not therefore make sense to focus research efforts on CLIR at this time. However, this analysis presumes that monolingual retrieval and CLIR gain their power from the same principles. While this certainly explains a part of the common performance levels, I believe it is not a comprehensive explanation. That is, the use of translation resources provides a form of term expansion and disambiguation that is simply not available in monolingual retrieval, and that provides a performance boost for CLIR. It seems likely then that aspects of monolingual retrieval are not currently being exploited properly in CLIR, and that further research could therefore improve CLIR performance without significant new advances in monolingual retrieval.

Information retrieval research topics that would benefit DHS include the following:

Data preparation. This most unglamorous task is eschewed by most researchers as being mindless and uninteresting. Nonetheless, data preparation is a significant barrier both to the research community and to the intelligence community. Tasks such as character set identification, file type identification, file format induction, and duplicate detection, could be presented so as to make the intellectual challenges apparent. Evaluation in a TREC-style setting would also help to attract researchers to these problems.

Partial translation. There are many tasks that would benefit from the ability to translate at the phrase or constituent level. For example, cross-language information retrieval might see performance improvements if the level of chunking for translation were allowed to vary. Similarly, cross-language question answering might be possible without full machine translation if the most important constituents could be accurately translated. Partial translation also ties well with information extraction and knowledge representation. There is room for significant advances in translation between the extremes of word-by-word mapping and full-blown machine translation.

Cross-language resource development. Resources for mapping from one language to another include translation dictionaries, parallel collections and comparable collections. These resources should be compiled *before* they are needed by the intelligence community. DHS might consider pursuing a variety of approaches, including on-line resource discovery; resource extraction; exploitation of government channels not ordinarily available to the research community; CLIR using comparable collections; and ingesting print resources. Making the *resources* open source (as opposed to the code to create or manipulate them) would be a good way to gain leverage from the broader community interested in the world's languages.

Query-specific retrieval. Most IR systems apply the same techniques to every query they process. Each technique works well for some queries and not for others. Selecting the best-performing technique on a query-by-query basis would significantly increase retrieval performance, yet no strong techniques are known for making such a selection. Research questions that would lead to query-specific retrieval, roughly in increasing order of difficulty, include: What makes systems fail? What makes a given query difficult? How many relevant documents are there for a query in a given collection? How difficult is a

given query? How well will system X perform on a given query? And finally, what techniques are best for a given query? The TREC Robust track is a venue for examining these issues. The opportunity for producing fundamental advances in information retrieval technology by performing query-specific retrieval are wide open.

4. Final Thoughts

DHS can provide the research community with more than just money. First, realistic data sets are tremendously important for promoting research that will ultimately be applicable to DHS problems. Large heterogeneous and degraded data sets are difficult to come by, so most current IR work is done over clean collections of newswire text. The rapid and wide distribution of the Enron email collection is testament to the demand in the community for realistic data. Second, most researchers have no access to intelligence analysts, and so target their work toward what they imagine analysts would like rather than at what they would actually like. Providing mechanisms for researchers to interact with analysts would ensure a closer match between the research being conducted and the needs of the intelligence community. Finally, participation in the development of TREC tracks would be a way to gain leverage from a broad range of researchers not directly funded to work on intelligence problems.