Issues in Cross-Language Retrieval from Document Image Collections

Douglas W. Oard
College of Library and Information Services
University of Maryland, College Park, MD 20742
oard@glue.umd.edu, http://www.glue.umd.edu/~oard/

Abstract

Over the past decade, broad-coverage cross-language text retrieval has progressed from isolated experiments on small collections to establish credible performance in large-scale evaluations. Extending this capability to document image collections presents some additional challenges that have not yet been well explored. This paper presents a general framework for cross-language retrieval, specializes that framework to retrieval from document image collections, and identifies opportunities for closer integration of the key enabling technologies and resources.

1 Introduction

Information retrieval systems seek to help users obtain information objects from large collections [2]. Early systems typically relied on manually assigned indexing terms, and such "controlled vocabulary" techniques were widely used in libraries to support the retrieval of printed documents. As storage costs declined and processing power improved, "free text" searching became cost effective and was widely deployed. Early applications of free text searching were limited to cases in which character-coded electronic text was available. More recent work on searching document image collections has yielded promising results, however, particularly when high-resolution document images are available [3].

Another trend with important implications for the nature of information retrieval is the rapid expansion in trans-boarder information exchange. Although research libraries and other specialized institutions have always collected documents written in many languages, modern networks now make vast collections of multilingual information available to any user. The past decade has seen substantial progress on the development of techniques for using queries expressed in one natural language to find documents written in another, a task that is typically referred to as Cross-Language Information Retrieval (CLIR) [10]. Present CLIR techniques are limited to electronic text, however. This paper proposes a framework for applying what we know about document image retrieval and cross-language retrieval to search multilingual collections of document images.

2 Framework

Figure 1 depicts a simplified process model for interactive information retrieval.

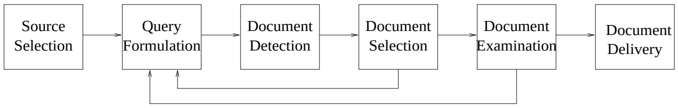

Figure 1: Information Retrieval process Model.

Source Selection. Information retrieval systems seek to provide information objects that contain information relevant to the user's information need. The first challenge is thus to select a system (or set of systems) that might contain information of the type desired. This is often a manual process, and it will not be addressed further in this paper.

Query Formulation. It is usually assumed that the user has a fairly specific information need that can be satisfied by some set of documents within the collection. The goal of the query formulation stage is to help the user develop the best possible formulation of the query. This is often an iterative process, as shown by the feedback loops from subsequent stages in Figure 1.

Document Detection. Detection is a general term that encompasses both searching relatively static collections and filtering dynamic document streams. The typical approach is to compute a figure of merit for each document that reflects the degree to which that document matches the query.

Document Selection. Interactive information retrieval is a synergistic process in which the machine applies relatively simple techniques to quickly cull promising documents from a large collection and then human abilities to rapidly recognize complex patterns are exploited once a manageable number of candidates have been identified. A compact display of important selection cues (title, author, date, etc.) is needed in the selection interface.

Document Examination. When the full text of the document is easily available, users are often able to improve their selection decisions by examining the document itself. Hypertext interfaces that support rapid browsing are often used for this purpose.

Document Delivery. Browsing interfaces provide one form of access, but sometimes additional processing is needed before the document can be used effectively. A printed copy may be desired, for example, or a professional translation of foreign language materials may be needed. Delivery is not discussed further in this paper, but it is identified as a separate stage here in order to emphasize that the purpose of the examination interface is to support choice, rather than use, of the documents being examined.

The remainder of this section explores the design of components to support the four central stages of this process model that are specialized to cross-language document image retrieval.

2.1 Support for Query Formulation

Queries can be posed explicitly, either as some form of selection criteria (using Boolean logic and proximity operators, for example) or as a set of "natural language" search terms. Alternatively, the query might be expressed implicitly by providing one or more examples of desirable (and/or undesirable) document images, and the user might be allowed to specify which aspects of the example(s) are particularly salient. For example, the user might wish to designate the body of a business letter as an example, but the addressee to which the example letter was sent might be of no consequence. The two techniques can be combined, using an explicit query to locate some document images and then enriching the query with selected document images as positive and/or negative examples, a process known as "relevance feedback." The key point here is that the query may contain character-coded electronic text, examples of document images, or a combination of the two. This means that CLIR systems for document image collections must generally search across modalities (between character-coded text and document images) as well as across languages.

2.2 Document Detection

Figure 2 shows the key components of the cross-language document detection stage. Most cross-language retrieval techniques are configured to process a specific language pair. When the document language cannot be reliably inferred from metadata or from the document source, automatic language identification techniques can be used to select appropriate language-specific processing (cf., [7]). If languages for which language-specific processing is not provided might be present in the collection, the language identification component can also be used to reject documents written in those languages.

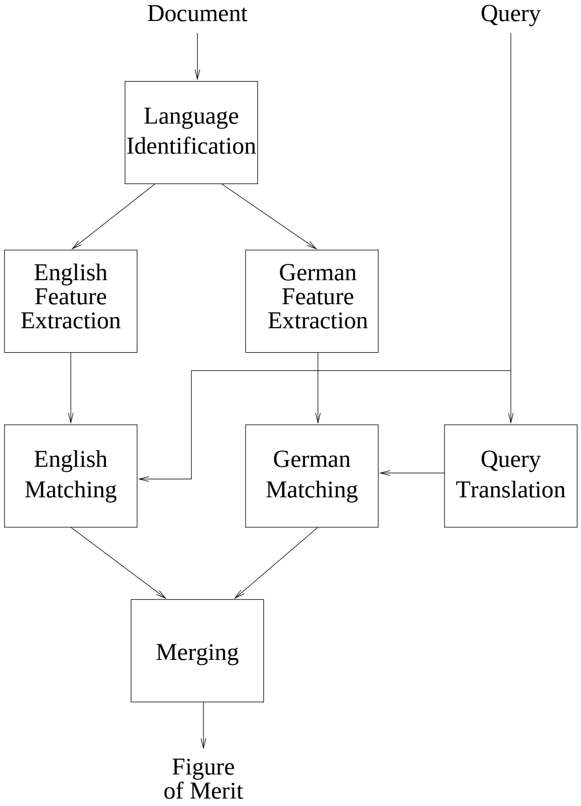

Figure 2: Cross-language document detection using query translation (English queries, English and German document images).

2.2.1 Feature Extraction

Two broad categories of features can be exploited for document image retrieval: document content and document structure. Content is typically characterized by identifying features (known generically as "terms") that are related to meaning and then weighting each term in a way that seeks to characterize that term's contribution to the meaning of a document. Three factors are generally used in the weight computation: the number of instances of that term in the document (more are better), the total number of term instances in the document (fewer are better), and the number of documents in which the term appears (fewer are better) [11].

Terms are extracted from document images by applying Optical Character Recognition (OCR) to identify individual characters and then combining the recognized characters until terms with the desired granularity are formed. The white space (spaces, tabs, etc.) that marks word boundaries can provide a useful cue to the appropriate granularity in some languages, but others (e.g., Chinese) lack reliable orthographic clues. Linguistic constraints and lexical knowledge can be used to identify plausible term boundaries in such cases, but OCR errors could complicate that processing by introducing symbols that appear to violate linguistic constraints. Any choice of terms naturally confounds some meanings and obscures the relationship between others. Often several words can be used interchangeably to convey nearly the same meaning (e.g., happy or glad). Optical Character Recognition (OCR) can exacerbate that problem, sometimes producing different results for separate instances of the same word within a single document.

Two general approaches have been developed for mitigating the effect of OCR errors on feature extraction in information retrieval applications. Both exploit observed regularities in character recognition errors. Character-confusion statistics can be used directly to postulate alternate strings (perhaps with lower weight) that might have resulted in the recognized characters. The same technique can be used with character-recognition algorithms that produce n-best (rather than 1-best) outputs. The other approach is to recognize character classes that exhibit little inter-class confusability rather than to recognize individual characters [13]. Terms formed from resulting "shape codes" exhibit greater ambiguity of meaning than the original words would have. Information retrieval systems perform fairly well in the face of increased ambiguity, particularly if relatively long queries (or examples of desirable documents) are provided [12], and the use of shape codes offers computational advantages over incorporation of character-confusion statistics.

Classification based on physical structure (layout) can be used directly to distinguish different document types such as business letters and newspapers. Physical structure can also provide cues about the logical structure of a document, and the logical structure can help to ascribe context to the terms. In a business letter, for example, it might be useful to know whether a name appears as the originator, as the addressee, or in the body of the letter. This contextual information can be used as an additional source of evidence for term weighting (e.g., giving more weight to terms in the lead paragraph of a news story) or as a basis for supporting queries that are matched against specific document components (e.g., a search for business letters to a specific addressee). Physical structure exhibits both cross-linguistic variations (e.g., vertical vs. horizontal writing) and cross-cultural variations within a single language (e.g., metric vs. U.S. letter paper sizes).

2.2.2 Cross-Language Matching

In cross-language retrieval it is necessary to (1) translate the terms in the query representation into the language(s) in which the documents are written, (2) translate the terms in the document representation into the supported query language(s), or (3) translate the terms in both into some common feature space. Query translation is the most efficient approach, and satisfactory response time is generally easily achieved when the queries are relatively short and are posed as electronic text. Long queries or instances of relevance feedback that might require on-the-fly OCR could shift that balance in favor of advance translation of the terms in every document.

There are four ways of obtaining the knowledge needed to translate the terms in documents and/or queries: (1) looking up term translations in a bilingual (or multilingual) lexicon, (2) algorithmically recognizing terms that are likely to be translation equivalents, (3) extracting useful relationships from a bilingual (or multilingual) corpus, or (4) by asking the user. A bilingual lexicon identifies one or more "target language" translations of each source language term, and it may include additional information such as part of speech or commonly co-occurring words that help to select to the correct translation. Some lexicons list translations in order of predominance in either general usage or in some application domain, and that information can be used as a basis for weighting alternatives when a single translation cannot be identified.

Unfamiliar names and newly introduced terminology pose a problem for systems that depend solely on lexical translation knowledge. When the source and target languages share a common character set, one simple technique is to retain unrecognized terms in the hope that they might be names or some other strings that would have the same representation in the source and target languages. More sophisticated cognate matching techniques can be applied (cf., [8]), and techniques which account for character-recognition errors and character-set differences are also available (cf., [6]).

Corpora (collections of documents that that use terms in representative ways) provide another source of translation knowledge that can be used alone or in conjunction with lexical and/or algorithmic sources. Parallel corpora, bilingual connections of translation equivalent documents, can be aligned to the sentence level fairly easily if sentence boundaries can be accurately detected since sentence length patterns are typically preserved across languages. Term co-occurrence statistics across aligned sentence pairs can then be used to postulate likely translations or as an indication of relative predominance among candidate translations for a term [9]. Cognate matching and/or a bilingual lexicon can be used to identify related regions in "comparable corpora" that contain documents in each language that are topically comparable but that are not translation equivalents (cf., [4]). There are also a number of less direct ways to improve the quality of lexicon-based translation using corpus statistics (cf., [1]).

It is clearly not possible to depend upon the user as the sole source for translation knowledge (since that would not be a cross-language retrieval problem!), but users with no knowledge of the target language might still help improve the accuracy of query translation performed using other techniques. Near-synonyms often group differently in different languages, so retranslation of each target language candidate back into the source language will sometimes provide even a monolingual user with enough cues to select the proper translation. For example, the German word "wagen" translates to either "car" or "risk" in English. The English word "car" retranslates to "wagen" and "auto," which could help a German speaker recognize the correct translation if reference to an automobile had been intended.

Three factors can adversely affect the performance of document or query translation: (1) translation ambiguity, (2) gaps and mismatches in lexical coverage, and (3) incorrect translation of noncompositional phrases. The first two factors deserve particular attention in the case of cross-language document image retrieval. Uncertain character recognition will necessarily magnify the effect of translation ambiguity somewhat no matter what technique is used, but the use of shape codes rather than confusion statistics could result in explosive growth of translation ambiguity. It is thus likely that shape codes will prove useful only for target language recognition. The second point is that lexical-coverage mismatch problems could be exacerbated in cross-language document image retrieval systems that use monolingual lexical resources for OCR error correction. Closely coupling the correction and translation processes could thus prove beneficial.

The remaining components of the document detection stage are essentially the same as those used in any cross-language retrieval application. Once the query and the document are represented by term weights in the same feature space, standard algorithms such as vector space, probabilistic, or Boolean matching can be performed. The result of this matching is a figure of merit that reflects the degree to which each document is estimated to satisfy the query. In monolingual applications, these values are typically used to construct a best-first ranked list of documents. The values are, however, not generally comparable across collections, nor are they generally comparable across different queries for a given collection. When multiple document languages are searched separately with different query translations, the values computed for each collection must be adjusted if a single rank-ordered list is desired. The nature of the adjustment depends on details of the translation and matching algorithms that are difficult to estimate, so the performance of matching in each language on a training collection with known relevance judgments is typically used as a basis for tuning this "merging" component (cf. [**?**]).

2.3 Selection and Examination

Because information retrieval systems typically make little use of factors such as word order and context, undesirable documents are invariably presented, even near the top of a list ranked in "best-first" order. Effective retrieval is thus a synergistic process in which the machine rapidly culls a manageable set of promising documents from the collection so that the user can quickly choose the most interesting documents. Recognition and translation errors make the machine's task more challenging than would be the case for monolingual retrieval of electronic text, so it is particularly important to provide the best possible support for selection and examination in cross-language document retrieval applications.

In the selection interface, documents are typically presented in a single ranked list, with each document represented using a compact set of features that users might find helpful in recognizing interesting documents. Users generally find document titles to be particularly valuable selection cues, so when structural cues are available to help locate a useful title within a document they should be exploited. Titles are often expressed as noun phrases, and simple techniques that produce readable title translations can be built by leveraging the limited range of linguistic phenomena that must be accommodated in such cases (cf. [5]). Users also typically find a few salient terms chosen from the document to be useful. The techniques used to select and weight terms for retrieval could also facilitate term selection for this purpose (unless shape codes are used). Temporal and numerical information that is typically found in a selection interface (e.g., the date the document was acquired and its length) are easily incorporated since no usual processing is needed.

Full text examination has proven to be quite popular in modern information retrieval systems. Monolingual document image retrieval poses no particular challenges in this regard since page images are easily displayed if adequate storage and bandwidth are available, but if translation is needed then two potential problems arise. One potential problem with a serial combination of OCR and translation is cascading errors. This architecture has be implemented in the Army Research Laboratory's Forward Area Language Converter (FALCon) system, and it appears that the resulting translations are of some value for assessing the contents of a document image that the user would otherwise be unable to read.[1] By more closely coupling the OCR and translation components, it may be possible to further improve the readability of the translation and thus improve the utility for document examination in a document image retrieval application.

The other problem is that both optical character recognition and machine translation are far slower than an interactive user might desire. Although information retrieval tasks are frequently modeled as rather narrowly goal-directed, experience suggests that many interactive search processes are marked by dynamic exploration and serendipitous discovery. Exploration and discovery would benefit from the availability of responsive retrieval systems that are able to operate inside the user's decision cycle. At present the only practical way to assure uniformly responsive support for full-text examination would be to perform massive translation in advance, but caching strategies offer a practical alternative that could provide rapid access to translations of frequently retrieved document images. Faster algorithms for each task, coupled with the ever-faster machines promised by Moore's Law, may ultimately obviate this concern completely.

3 Conclusions

Although broad-coverage cross-language document image retrieval systems do not yet exist, all of the enabling technology is now available and modular approaches based on existing components (page decomposition, optical character recognition, query translation, and machine translation) could easily be constructed. Optimal performance will likely require a closer degree of integration, however, both at the level of lexical resources and between the recognition and translation components. What is needed now is a testbed on which alternative integration strategies can be explored. The development of such a tool would be a significant step towards improved access to that portion of the world's storehouse of knowledge that presently exists only in printed form.

[1] http://rpstl.arl.mil/ISB/falcon.htm

Acknowledgments

The author is grateful to Bob Allen for his comments on an earlier version of this paper. This work has been supported in part by DARPA contract N6600197C8540.

References

[1] Lisa Ballesteros and W. Bruce Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 1997.

[2] D. C. Blair. *Language and Representation in Information Retrieval.* Elsevier, Amsterdam, 1990.

[3] David Doermann. The indexing and retrieval of document images: A survey. Technical Report CS-TR-3876, University of Maryland, Computer Science Department, February 1998.

[4] Martin Franz, J. Scott McCarley, and Salim Roukos. Ad hoc and multilingual information retrieval at IBM. In E. M. Voorhees and D. K. Harman, editors, *The Seventh Text REtrieval Conference (TREC-7)*. Department of Commerce, National Institute of Standards and Technology, November 1998. http://trec.nist.gov.

[5] Genichiro Kikui, Yoshihiko Hayashi, and Seiji Suzaki. Cross-lingual information retrieval on the WWW. In *Proceedings of the First Workshop on Multilinguality in Software Engineering: The AI Contribution (MULSAIC)*. European Coordinating Committee for Artificial Intelligence, August 1996.

[6] Kevin Knight and Johnathan Graehl. Machine transliteration. In *Seventeenth International Conference of the Association for Computational Linguistics*, 1997. http://www.isi.edu/natural-language/projects/nlg-publications.html.

[7] Dar-Shyang Lee, Craig R. Nohl, and Henry S. Baird. Language identification in complex, unoriented, and degraded document images.

In *Proceedings of the Second IAPR Workshop on Document Analysis Systems*, pages 17–39, October 1996. http://cm.bell-labs.com/cm/cs/who/hsb/pub.html.

[8] I. Dan Melamed. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Third Workshop on Very Large Corpora*, 1995. http://www.cis.upenn.edu/~melamed/.

[9] I. Dan Melamed. Emperical methods for MT lexicon construction. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas*, 1998. http://www.cis.upenn.edu/~melamed/.

[10] Douglas W. Oard and Anne Diekema. Cross-language information retrieval. In *Annual Review of Information Science and Technology*, volume 33. American Society for Information Science, 1998.

[11] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beauliew, and M. Gatford. Okapi at TREC-3. In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. Department of Commerce, National Institute of Standards and Technology, November 1994. http://trec.nist.gov.

[12] Mark Sanderson. Word sense disambiguation and information retrieval. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 142–151. Springer-Verlag, July 1994. http://www.dcs.gla.ac.uk/ir/papers/Postscript/sanderson94b.ps.gz.

[13] Alan F. Smeaton and A. L. Spitz. Using character shape coding for information retrieval. In *Proceedings of the 4th International Conference on Document Analysis and Recognition*, pages 974–978, 1997.