GARETH J. F. JONES

SCHOOL OF COMPUTING, DUBLIN CITY UNIVERSITY
DUBLIN 9, IRELAND

BEYOND ENGLISH TEXT: MULTILINGUAL AND MULTIMEDIA INFORMATION RETRIEVAL

1. INTRODUCTION

In common with many areas of language processing, the origins of information retrieval (IR) research are to be found in the exploration of techniques for electronic English language text archives. The adoption of this research strategy arose, I suspect, from the general competence in the English language of scientific researchers internationally, and more particularly due to the availability of standard English text collections for comparative experimental research. A number of successful models for information retrieval were, and continue to be, developed using these test collections as their primary research focus.

English language document collections, and electronic text documents in any language, represent only a minority of the information sources that a user may wish to search to satisfy their information need. The need to expand the scope of IR research beyond English text has been recognised in the last 10 years. Increasing amounts of work are now being reported which explore non-English IR, cross-language information retrieval (CLIR), multilingual information retrieval (MLIR) and multimedia information retrieval (MIR).

When these efforts to expand the horizons of IR began, it was not at all clear what approaches should be adopted for these new tasks in order to achieve the greatest IR effectiveness. However, as we shall see in this chapter, the techniques of probabilistic information retrieval and the approaches to automatic indexing, developed by Karen Spärck Jones and her various collaborators over the last 30 years, have stood up remarkably well to the new challenges. Indeed at the time of writing, the comment from many researchers seeking to develop novel more effective IR methods for these and other tasks, continues to be *"... it's good, but it still isn't really any better than Okapi ..."*. The reason for this result should perhaps not be too surprising given the rigor and care taken over the years to ground these models in sound theoretical analysis, and the extensive experimental evaluations that have characterized this work (Sparck Jones, Walker, & Robertson, 2000a) (Sparck Jones, Walker, & Robertson, 2000b).

This chapter continues in Section 2 with a brief review of the relevant details and indexing assumptions of the probabilistic model of IR. Section 3 describes experimental work with non-English test collections, this is extended in Section 4 which gives results for cross-language and multilingual IR. Section 5 introduces multimedia IR and highlights some relevant experimental work. Finally, Section 6 draws conclusions from existing work and looks toward future applications and challenges.

John I. Tait (ed.), Charting a New Course.

2. PROBABILISTIC MODELS AND FEATURE INDEXING

IR systems seek to satisfy a user's *information need*. Current IR systems attempt to do this by locating *relevant* documents from within which the user themself extracts the required information. Potentially relevant documents are selected and returned to the user based on a retrieval model. This model can make use of whatever information is made available about the documents from among which it is seeking to locate the relevant ones. Document retrieval models fall into two broad classes of Boolean and best-match, the latter is the most dominant in current IR research and is the only approach considered here. Over the years many best-match or ranked retrieval models have been proposed and evaluated. The most popular models being: the vector-space approach (Salton & Buckey, 1988), the probabilistic model (Robertson & Sparck Jones, 1976), and more recent methods based on statistical language modelling (Ponte & Croft, 1998). For reasons of its demonstrated effectiveness, and Karen Spärck Jones's strong association with its development, this discussion focuses only on the probabilistic approach.

The probabilistic model seeks to evaluate a simple quantity *P(relevance|document)*, the probability of relevance given this document for a *specific* search request. Using this model documents can be returned to the user in decreasing order of probability of relevance. This is more formally stated in the Probability Ranking Principle (Robertson, 1977) (Sparck Jones et al., 2000a):

> P1: If retrieved documents are ordered by decreasing probability of relevance on the data available, then the system's effectiveness is the best to be gotten for the data.

If we had a complete model of each document, describing all potentially important features, with a corresponding model of the information need expressed by the search request, we might expect perfect retrieval with all relevant documents having higher probabilities than non-relevant documents. Alas such document models do not currently exist, and retrieved ranked document lists interleave relevant and non-relevant documents. Even if it were possible to compute *P(relevance|document)* perfectly, the under specification of information need often found in expressed search requests may cause an unavoidable ambiguity in document relevance. In any case, the objective of research in probabilistic IR is to improve the reliability of these imperfect relevance probability estimates.

Every document can be assumed to be a unique event, and in general, we take it that the description of each document used for retrieval is similarly unique. A problem arises with this modelling assumption, since it is difficult to assign probabilities to unique events. A solution comes in the form of decomposing document descriptions into their non-unique components or attributes, whose association with relevance can be estimated. These attributes can be used in combination to synthesise a relevance probability estimate for each unique document. The derivation of the early form of this practical probabilistic model (the "binary independence model") is described in van Rijsbergen (1979), and the more recent extended form of the model (well known as the "Okapi BM25"' model) in Sparck Jones et al. (2000a). In the BM25 model the likelihood of relevance for a document *j* is computed based on the sum of the *combined weights cw(i,j)* of the

independent attributes *i* which occur in both the document and the current search request. *cw(i,j)* values are computed based on the classic IR attribute weighting features of across document collection frequency (the *collection frequency weight cfw(i)*) of attributes *i*, the *within document frequency* of an attribute *i* in the document *j*, and an adjustment of the weight to compensate for document length (Robertson & Walker, 1994).

In general for current IR systems, each document is modelled as a simple "bag-of-words" which lists the attributes occurring within the document and their frequency of occurrence. The degree of match between a document *j* and the search request is then simply computed as a matching score *ms(j)* of the number of attributes in common between the request and the document. A list of documents ranked by matching score is then returned to the users. Documents are thus represented within the IR system as (assumed) independent attributes. The theory of the probabilistic model tells us nothing about the language of these attributes or even the media of the documents. Of course, much of the experimental work that established the effectiveness of this model was carried out using English text collections, but in theory there should be no reason why it cannot be used for other languages or media.

Several well established techniques are typically applied for automatic indexing of English language text documents. These include removal of frequent *stop words*, such as those in van Rijsbergen's list (van Rijsbergen 1979), *suffix stripping*, using a method such as the Porter algorithm (Porter 1980), standardisation of spelling, and conflation of synonyms. Whatever preprocessing is applied, the features used for retrieval are still independent attributes derived from the document. Combined with enhancements such as relevance feedback and pilot searching using large additional document collections, BM25 has shown consistently good effectiveness in comparative retrieval evaluation exercises such as TREC (Robertson, Walker, & Beaulieu, 1998) (Sparck Jones et al., 2000b).

The following sections look at the adaptations required for the application of probabilistic retrieval to non-English documents, cross-language and multilingual information retrieval, and its effectiveness for multimedia information retrieval.

3. NON-ENGLISH INFORMATION RETRIEVAL

A key consideration when developing an IR system for a new language is the selection of the most suitable set of attributes to be used to index the documents. The lexical and structural differences between languages mean that the distributions of attributes within individual documents and across collections will vary between different languages. However, since the probabilistic model makes no explicit language dependent assumptions about these distributions, there is no reason to suppose that, with appropriately selected indexing units, it should not work effectively for any language.

From a linguistic perspective English actually provides a good starting point for the investigation of indexing methods and retrieval models. The basic word units of the language are easily identified, and the types and degrees of inflection of

individual words are relatively simple compared to those of many other languages. There are of course many exceptions to these apparently simple rules of inflexion, and ongoing debate over the basic units of meaning, but generally these concerns can be safely ignored or handled by explicit exception lists for the purposes of IR indexing. Some other languages have similar properties to English while others introduce new issues which must be addressed for effective retrieval. This discussion outlines some of the features relating to indexing and retrieval of a range of representative languages.

From an IR perspective, languages such as French, Italian and Spanish can be addressed using adaptations of the techniques used for English. Thus for each language, we need to develop a suitable set of high frequency stop words that can be removed safely without affecting retrieval effectiveness, suffix stripping algorithms to conflate words to common stems, and appropriate synonym dictionaries (Wechsler, Sheridan, & Schäuble, 1997). Probabilistic IR methods using this approach have been shown to be effective in comparative evaluations of non-English IR tasks, for example within the Cross-Language Evaluation Forum (CLEF) workshop series (Savoy, 2004).

More complex issues are introduced by languages such as German and Dutch which are highly declensional with a rich system of inflections and cases (Braschler & Ripplinger, 2004). In addition, in common with other Germanic languages, such as Swedish, and other languages such as Finnish, there is free compounding of words to express concepts developed from the component words. In these cases, although words are still the building blocks of the language, they are frequently combined into noun compounds without spaces. If one of these noun compounds appears in a search request and a document, there is a very good chance that this is a relevant document. However, the generative nature of the compounds means that often no match will be found for a search compound within the document set. This can lead to many potentially relevant documents being missed, since they don't contain the compound in exactly the form used in the request. The general approach to this problem is to develop methods for compound splitting; these techniques may rely on the use of a compound dictionary or language specific rules for identifying word units within compounds, or a combination of both methods (Braschler & Ripplinger, 2004). Of course, in addition to the decompounding of these concatenated words, indexing of these languages also benefits from the application of effective stemmers and removal of stop words.

Different issues arise in the case of east Asian languages such as Chinese and Japanese. The written form of these languages uses ideograms of Chinese origin. There are many thousands of these characters which usually have some meaning associated with them. Most words are formed by bringing two characters together. The meaning of the word is usually related to those of its constituent characters. Shorter words consisting of one character can express simple concepts and occasional longer words more complex ones. While Chinese is restricted to a single character set, in the case of Japanese three additional character sets are in common usage: *hiragana* whose role is similar to function words and verb suffices in English, *katakana* which are used to transliterate Western concepts, e.g. *computer* appears phonetically in Japanese katakana as *ko n pu ta*, and *romaji*, for Western characters

sometimes used for numbers and proper nouns. The major concern when indexing languages of this type is the observation that there are no spaces between the words of each sentence. The text must thus be segmented into suitable representative units prior to indexing. Further since the ideogram character set is itself so rich, there is a question of what the best units for retrieval actually are.

A number of approaches have been explored for indexing these languages. The most basic method is simply to take each character as an indexing unit, a slightly more elaborate one is to use overlapping n-grams of characters of varying lengths, while the most complex strategy is to apply morphological analysis to identify the most likely word break points. A number of experiments using various Chinese and Japanese test collections exploring different approaches to segmentation have been carried out with inconclusive results, for example Huang & Robertson (1997) and Jones, Sakai, Kajiura, & Sumita (1998). Regardless of the indexing units selected, the probabilistic IR model has achieved consistently good retrieval performance with these languages. This was demonstrated recently for Japanese by the very good performance of the Toshiba BRIDJE system, which is based on BM25, at the NTCIR-4 Asian language evaluation workshop (Sakai, Koyama, Kumano, & Manabe, 2004).

4. CROSS-LANGUAGE AND MULTILINGUAL INFORMATION RETRIEVAL

Another topic moving IR beyond English language text collections, which has received considerable attention in recent years, is retrieval applications working with more than one language. This subject is broadly classified into two areas: cross-language information retrieval (CLIR), and multilingual information retrieval (MLIR). CLIR is concerned with the retrieval of documents in one language using search requests in another language, e.g. French requests used to retrieve Chinese documents. MLIR extends this to retrieval from a collection where documents are uniquely present in one language, but the collection overall covers documents in multiple languages, e.g. using a Japanese request to retrieve from a collection with documents in English, French, German, Spanish, Finnish and Russian. In practice, more complex situations are clearly possible. A single document may contain material in more than one language, and individual documents may be repeated in different languages within a collection. From these definitions it can be argued that CLIR is really a subset of MLIR. This section introduces research questions posed by CLIR amd MLIR, and outlines the main solutions that have been proposed and explored to date.

4.1. Cross-Language Information Retrieval

The principal question that arises in the context of CLIR is: how should the language barrier between the search requests and documents be crossed? Should search requests be translated into the language of the documents, should the documents be translated into the language of the request, or both? Further, what is the best approach to carrying out this translation?

4.1.1 Request Translation vs Document Translation

There are well rehearsed arguments for and against request or document translation, with the main issues relating to translation cost, at what stage it is carried out, its effectiveness for retrieval, the available translation and computational resources, and the storage implications.

Generally it is held that translating requests when they are entered will be fast enough, since they are likely to short, not to interfere with interactive searching. Unfortunately, short requests often have minimal formal linguistic structure, and further because they are short, there is little information of the context in which the request words have been selected by the user. These factors mean that it will often be difficult to perform reliable deep linguistic analysis when attempting to perform translation of the request. One consequence of this is that it can be difficult to select the contextually appropriate translation of polysemous words. A further implication of attempting to translate short requests is that the mistranslation of individual words can have a significant impact on retrieval effectiveness. However, since the document collection to be searched will not have been translated, and is therefore accurate, redundancy effects are often found to help to ameliorate translation errors even for short requests. It is further frequently argued that, since deep linguistic analysis of request may not be possible (or if possible may not be desirable, if it is likely to be unreliable), and since we are only seeking to transfer the words into another language, shallower translation methods may be better for request translation CLIR.

Consider now the alternative approach of document translation. Documents are generally much longer than search requests, and the content will generally be linguistically well structured with large amounts of contextual information available. Thus translation of documents using formal linguistic analysis is potentially more accurate than it is for requests. While they may generally be translated more accurately then short requests, translated documents will nevertheless contain a number of errors arising from incorrect analysis of the source text and limitations of the translation dictionaries. These errors will inevitably impact adversely on retrieval accuracy for CLIR. However, adopting document translation does mean that no translation has to take place when the search request is entered, so the retrieval stage itself is computationally faster and cheaper. Also, the search request is now accurate, with no possibility of translation error. A major disadvantage of document translation is the very high cost of translating all the documents. Although, since translation is done in advance of retrieval and only has to be done once, it can really be regarded as part of a very expensive indexing process. However, there are storage implications which arise from the need to maintain a separate search collection in each request language into which the documents are translated.

Experimentally both request and document translation have been shown to be effective, with at least one study showing that combining the retrieval output of both methods used independently can produce the best overall retrieval effectiveness (McCarley, 1999).

One way to address the problem of storage is to translate all documents into a single "pivot" language, most probably English, and then to translate the requests into this same language when they are entered. This has the disadvantage that since both the requests and documents are being translated, translation errors will be compounded with a consequential impact on retrieval effectiveness. Pivot languages can also be used when resources are not available to translate directly between the request and document languages (Gollins & Sanderson, 2001). In this case they can be used for translation of both requests and the documents into the pivot language, or for sequential translation of either the requests or documents into the language of the other.

4.1.2 Translation Methods for CLIR

Another widely debated issue in CLIR is how the translation should be carried out. The issues here relate both to the actual best means of translation for CLIR, were a perfect translation resource to be available, and the most appropriate method, where technical and resourcing limitations mean that real translation systems are currently far from perfect. Broadly speaking the three translation strategies that have been explored for CLIR can be categorised as: dictionary-based, comparable corpora, and machine translation.

Most early work in CLIR advocated the use of bilingual dictionaries for topic translation, with a variety of elaborations to improve their effectiveness for this task (Hull & Grefenstette, 1996). In its simplest form, this approach replaces each word in the search request with all possible translations of the word in the document language appearing in a bilingual dictionary. As well as including the appropriate translation, if it is available in the dictionary, this simple method often introduces many contextually inappropriate translations of this word. These incorrect translations have been shown to significantly degrade CLIR retrieval effectiveness relative to monolingual IR for the same set of requests and documents. It has been demonstrated that dictionary-based CLIR performance can be improved by using careful phrase translation, and relevance feedback both prior to and after translation of the request (Ballesteros & Croft, 1998).

Given the problems with ambiguity arising from the use of bilingual dictionaries, and the gaps which occur with regard to their coverage of domain specific vocabulary items, alternative methods have been explored which align comparable corpora in the different languages (Sheridan & Ballerini, 1996). Related terms appearing in this aligned content are used to translate requests in a context specific way. One of the problems with this strategy is that suitable related corpora are often not available for alignment. A widely explored way to overcome this problem is to use content from the internet (Nie, Simard, Isabelle, & Durand, 1999). In this approach large numbers of web pages are collected and aligned, and then used for request translation. Nie et al. demonstrated that an improvement in retrieval effectiveness can be obtained by using the aligned web documents in combination with a bilingual dictionary.

Perhaps the most obvious solution to crossing the language barrier between requests and documents is to use a standard commercial machine translation system. Indeed for CLIR using document translation, machine translation would appear to be the only realistic option given the huge amount of ambiguity that the other translation methods would introduce. Certainly I'm not aware of work which attempts to translate whole document collections using a different method. The arguments in favour of machine translation for CLIR centre on the potential for accurate translation of the words, appearing in the request or the document, which can be achieved by bringing sophisticated translation resources to bear on the task. Current machine translation systems often produce rather unnatural prose output. However this is not a problem for CLIR where we are only interested in the reliable translation of words with good relevance selectivity. The arguments against machine translation for CLIR are based on the previously stated issues of poor linguistic structure in search requests, which can render them difficult for formal linguistic analysis using machine translation, with consequential translation failures and inappropriate translation of words. Dictionary limitations can also result in translation problems with domain specific words for both requests and documents.

My former colleagues and I at Toshiba performed a comparative evaluation of progressively more sophisticated request translation strategies ranging from simple bilingual dictionary lookup, to part-of-speech tagging, sense disambiguation, and full machine translation for an English - Japanese CLIR task (Jones, Sakai, Collier, Kumano, & Sumita, 1999). Perhaps surprisingly given the arguments against machine translation for CLIR, the best retrieval effectiveness was found using full machine translation. This result was observed for both natural language request statements, and requests modified to disrupt the linguistic structure by removing the function words prior to translation. More recent experiments have shown that a combination of machine translation and the Okapi BM25 probabilistic model combined with relevance feedback produces among the best reported effectiveness for the CLEF CLIR tasks (Jones & Lam-Adesina, 2001) (Lam-Adesina & Jones, 2003). Analysis of the retrieval behaviour of individual requests showed that there is sensitivity to the failure to translate important words, usually previously unseen proper nouns. For example, failure to translate phonetic loan word proper nouns rendered in katakana in Japanese if they are not present in the translation dictionary, significantly degrades retrieval effectiveness. This will often be a problem for bilingual dictionaries as well; although, its impact on retrieval performance may be masked by translation ambiguity issues. However comparable corpora should be able to capture these domain specific translations, as long as they include documents covering the appropriate related topics in their training set.

Many papers have been published describing CLIR results in recent years. The references included here are generally those which first introduced or advocated a particular translation approach for CLIR, in each case subsequent work has often extended these methods. While machine translation shows good results when available, bilingual dictionaries and aligned corpora remain an important translation resource for CLIR with language pairs for which well developed machine translation tools are not available. There are direct bilingual dictionaries available between most major languages pairs, and even for minority languages there are bilingual

dictionaries to major languages such as English, while the expanding amounts of electronic text available from many sources mean that corpus-based methods will become an increasingly important resource.

4.2. Multilingual Information Retrieval

In MLIR the IR system is expected to respond to a search request in one language by generating a ranked list of potentially relevant documents in multiple languages. Similar to CLIR, MLIR can be approached using either a request or document translation strategy. The challenges of MLIR include similar translation issues to CLIR; however it also introduces a significant new problem which arises because the documents in each language will often be in separate collections. In a practical system document collections may be geographically distributed with no option to merge them into a single collection. However, even if the documents can be combined into a single physical collection, the fact that they are in different languages means that semantically related search terms cannot be conflated, and effectively it will still behave as separate, language specific, sub-collections. The major difficulty that arises for the MLIR is how to take the separate outputs from searching individual collections and merge them into a single output list for delivery to the user, which reliably ranks relevant documents higher than non-relevant ones. For this reason, MLIR is often seen as being akin to monolingual distributed IR, where separate search collections are stored and searched independently for practical or commercial reasons (Callan, 2000).

The merging problem arises since ranked lists from the separate collections will be generated using different indexing strategies, and, as discussed earlier, the features will have varied distributions for the individual languages. This means that the document matching scores from the retrieved ranked document lists will generally be incompatible. For example, documents retrieved from a collection with higher average matching scores will tend to be favoured in the merged list. Thus the list may be biased towards certain collections regardless of the actual relative likelihood of documents retrieved from these collections being relevant. If this problem is overcome, a further concern is that the matching score profiles of the lists may be different. Hence the lists cannot be merged in a simple reliable way. In general for distributed IR, difficulties of list merging vary depending on the number of differences between the IR systems used to compute the separate lists, and potentially the cooperation between the maintainers of the separate search engines (Callan, 2000). If the separate retrieval systems use different retrieval ranking algorithms then the scores will clearly be incompatible, but even if an identical retrieval strategy is used for all the collections, the matching scores will be incompatible due to the different values used to estimate the term weights or other ranking parameters. In MLIR, these issues are compounded by problems arising from the variations in the properties of the languages. For document translation MLIR, if the document index data are located physically together, the index files can be combined to form a single search collection. This removes the need for merging

of separate lists. However, if the collections are distributed or request translation is being used, some method of merging must be adopted.

A variety of list merging algorithms of varying complexity have been proposed for distributed IR. A number of these have been applied for MLIR with varying degrees of success. The simplest approach involves ignoring the score incompatibility problem, and simply merging the ranked lists using their raw scores. More complex methods involve ranking the separate collections in terms of their estimated likelihood of containing relevant documents, combining these collection matching scores with the matching scores of individual documents to form a composite score, and using this combined score to generate the final merged document list. These methods have been shown to effective for monolingual distributed IR (Callan, 2000). Unfortunately, they have not proved so successful for MLIR, where it has been difficult to improve performance beyond that achieved using the simplest methods (Lam-Adesina & Jones, 2003) (Savoy, 2004).

In our experiments for the CLEF workshop MLIR task in 2003, we translated all the documents from their original languages of French, German and Spanish into English using machine translation. We then compared retrieval effectiveness of various list merging strategies with that for a single collection formed from the translated documents. Overall we found that the single collection method worked best indicating that all the merging strategies fell short of the performance that could potentially be achieved using these document sets (Lam-Adesina and Jones, 2003). Once again our results showed that the BM25 Okapi probabilistic model produced among the best retrieval effectiveness for this task. Of course it will not always be possible to translate the entire retrieval collections and then combine them, and thus merging is an important ongoing concern for MLIR requiring further investigation.

5. MULTIMEDIA INFORMATION RETRIEVAL

The current expansion in archives of digital multimedia content is creating the need for tools to automatically search and retrieve material from these collections. Similar to the work on multilingual text documents, recent years have seen a rapid increase in research exploring Multimedia Information Retrieval (MIR). Multimedia archives comprise material in one or more of audio or visual media, often accompanied by some form of electronic text annotation. Retrieval from these collections raises a number of issues with respect to both the indexing and retrieval processes. Multimedia content can be either static, in case of individual digitised images such photographs or paintings, or temporal, comprising audio and/or video content. The static or temporal nature introduces various concerns with respect to the presentation to the user and browsing of retrieved content.

Indexing and retrieval methods for MIR depend on the media under consideration. Let us consider these in order of increasing complexity. Electronic text material available for MIR can either take the form of metadata or direct transcription of content. Metadata may describe the content in some way, e.g. the names or roles of the characters appearing in an image, or the events taking place in a video. Transcriptions of linguistic content may be generated manually or

automatically. For example, the close captioning often broadcast with TV sources can be captured and used as a high quality transcription of the content for the purpose of retrieval and browsing.

Existing IR research has focussed very much on linguistic content, and so can in general be applied directly to manually annotated material associated with multimedia content. The usefulness of manually entered descriptive metadata will depend on the quality of the data, and its relevance to an individual request. Thus, while the visual content of an image may make it relevant to a particular request, if the descriptive metadata is not pertinent to the aspect of this item which makes it relevant, then the MIR system will fail to locate it. Thus the effectiveness of MIR will clearly be affected by the accuracy and richness of the annotation. Additionally, the complexity of the retrieval methods used for textual annotations may be influenced by their complexity; if the annotations are highly structured, this may be taken into account in the retrieval algorithms adopted.

Of more interest within recent and current research, is MIR based on automated annotation of the content. The following sections consider indexing and retrieval for first spoken documents, and then image and video data.

5.1. Spoken Document Retrieval

In many situations it is uneconomic or impractical to manually transcribe the spoken contents of multimedia documents, and thus transcriptions must be generated automatically using speech recognition technologies. Forming transcriptions in this way using current speech recognition tools has a number of limitations. The most significant issue is that, like machine translation systems used for CLIR, these tools make mistakes; incorrect words can be inserted into the transcription, correct words deleted, or one word incorrectly substituted for another one. These errors arise for a number of reasons relating to both the natural language data and the tools themselves. Speech recognition is inherently challenging for a number of reasons including the following: the speech may be poorly articulated, it may not follow expected linguistic patterns, it may be captured using poor quality equipment, there may be high levels of background or environmental noise, or there may be crosstalk where more than one speaker is talking at the same time. The accuracy of a speech recognition system is limited by the effectiveness of its acoustic models to accurately recognise the sound patterns of the current speaker, and of its language models to predict their use of word patterns. Current speech recognition transcription systems are also correctly described as “large vocabulary”, where only the words within a predefined vocabulary can be recognised correctly; other so called “out-of-vocabulary” words will be transcribed incorrectly by definition. In general, the overall accuracy of an automatically generated document transcript will depend on the extent to which the speech deviates from the trained parameters of the speech recognition system and the quality of the input speech signal.

The effect of recognition errors is to produce a “noisy” transcription which will have some similarities to the output of a machine translation system. The characteristics of the errors however are likely to be somewhat different. A machine

translation system can determine its output, although it may experience problems with the naturalness of the word patterns generated, or be subject to limitations in the richness of the available vocabulary or linguistic structures. By contrast, a speech recognition system must do its best to transcribe the data presented to it. Automatic transcriptions often include apparently random insertion and deletion errors. A potential problem for both machine translation and speech recognition though is how to appropriately handle input words outside their vocabulary.

Research into spoken document retrieval (SDR) began with a number of projects in the early 1990's. These examined various approaches to automatically indexing the spoken contents and were evaluated using locally developed test collections (Glavitsch & Schäuble, 1992) (Jones, Foote, Sparck Jones, & Young, 1996). When these projects started, the potential of IR techniques derived from experience with electronic text documents to transfer successfully to errorful spoken document index files was very much an open question.

Video Mail Retrieval using Voice (VMR) at Cambridge University was one of these early SDR projects. Karen Spärck Jones and myself worked with others to investigate the impact on retrieval effectiveness of several approaches to spoken document recognition. The VMR project used a small test collection of 300 voice mail messages to explore SDR effectiveness. We used the BM25 model to compare retrieval behaviour for manually created message transcriptions with those generated using a 20,000 word large vocabulary system and an alternative technique known as *phone lattice spotting (PLS)* (Jones et al., 1996). In neither case was the recognition system specifically adapted for the indexing of these messages. The transcription system was trained for a broadcast news recognition task, and achieved an average word error rate of 47%. PLS uses subword level speech recognition to form a phone lattice structure. The lattice is scanned for phone strings corresponding to possible occurrences of words appearing in a search request, as such it is an open-vocabulary indexing method able to recognise any word appearing in a message. Experiments using the VMR test collection demonstrated retrieval effectiveness of around 70-75% of that for manual transcriptions for both these recognition techniques, rising to around 85% when they were used in combination.

It is a feature of speech recognition that the hardest words to recognise accurately are often short function words. Of course, these are generally not useful for retrieval, and hence SDR systems can still operate with good reliability in the presence of relatively high word recognition error rates. A further issue is that since important words within a document are often repeated, even if the word is recognised incorrectly when it occurs in one place, it may be correctly recognised elsewhere in the document. Whilst errors of this type will degrade the overall quality of term weights, the documents will still be retrieved. This distortion of term weights can result in some distortion of the ranked retrieval list, relative that to that which would be achieved with a perfect document transcription, but overall high levels of retrieval effectiveness can still be achieved.

Interest in SDR increased significantly in the mid-1990's and a track was introduced at the annual TREC series in 1997. For the first time researchers were able to work with a common SDR test collection. The SDR track ran for 4 years, each conference increased the document collection size or the complexity of the

retrieval task. During this time speech recognition technologies continued to advance. Using the best available transcription systems, achieving recognition average word errors rates of around 20% with a vocabulary of around 65,000 words, together with the BM25 model and retrieval enhancement techniques, such as relevance feedback and merging with in-domain large contemporaneous text collections, TREC SDR participants demonstrated similar overall retrieval effectiveness for manual and automatic document transcriptions (Johnson, Jourlin, Sparck Jones, & Woodland, 2001) (Garafolo, Auzanne, & Voorhees, 2000). The success of the TREC SDR track indicated, at least for a task where the transcription system can be well trained for the domain of the document collection, in this case broadcast news, that SDR is effective using current speech recognition technologies. Most MIR research interest has now moved to the new challenges of image and video retrieval.

5.2. Image and Video Retrieval

Whereas it is natural to use the same indexing units for spoken content and written linguistic content, the appropriate mechanism for indexing and retrieving from visual media is much less clear. Visual content can include natural scenes either in static images or moving video, as well as other image content, for example scanned or overlaid textual material.

Considering first the more straightforward case of textual content in images. The first stage in automatically indexing this material is to identify zones or regions in the image containing text. The text in these zones is then recognised using an optical character recognition (OCR) process. After this, it can be indexed using a standard retrieval approach derived from experience with electronic text documents. Unfortunately, similar to speech recognition systems, OCR systems make mistakes; although the errors in this case are often of a different form. Instead of making whole word recognition errors, as is the case for speech recognition, OCR systems typically make errors in the recognition of individual characters. Each of these errors will usually introduce a new word into the indexing vocabulary of the collection. These words will not be useful indexing terms, since they will not match correctly with terms appearing in typed search requests, and they will also have disproportionately high collection frequency weights, since they are very rare within the document collection. A simple way to resolve this problem might be to attempt to correct automatically the spelling of these words using a dictionary. However, it is not always clear what the correct word should be. Indeed sometimes a word not present in the dictionary will actually have been correctly recognised by the OCR system, and attempting to correct OCR errors in this way may replace these accurately recognised words with incorrect words taken from the dictionary. As a consequence of this problem, "correcting" the OCR output with a dictionary may lead to a degrading of retrieval effectiveness. Another issue, similar to spoken document recognition, is that the accuracy of the output of an OCR system will be related to the difficulty of the recognition task. OCR accuracy will depend on the quality of the printing, the fonts used, and the contrast between the print and the

paper. For example, modern laser printed output with a simple font is easier to recognise than older mechanically printed documents for which the paper may be yellowing with age. Significantly more difficult to recognise accurately is handwritten text, for which accuracy will obviously depend on how clearly it has been written, as well as the other factors affecting printed text.

Experimental exploration of scanned text image retrieval has demonstrated that the BM25 model once again performs well for this task with printed data (Jones & Lam-Adesina, 2002). To the best of my knowledge its effectiveness for more difficult hand written documents has at present not been examined, although work using a statistical relevance model for retrieval of handwritten historical documents is reported by Rath, Manmatha, & Lavrenk (2004). Interestingly, while relevance feedback has been shown to be very effective for SDR (Johnson et al., 2001), the differences in error types encountered between OCR and speech generated transcripts, mean that it does not transfer to scanned text documents in a simple way (Jones & Lam-Adesina, 2002).

A much less well defined task is the retrieval of multimedia documents based on non-linguistic visual content. When examining a visual scene, we might want to identify any number of different features. For example, we may wish to recognise the individuals appearing in the image, the place where the scene is taking place, the objects in the picture, or perhaps the events being depicted. Identifying these features is very difficult. Indeed doing this in a robust way outside a very narrow pre-defined domain is currently not possible. Much visual media can be interpreted in a seemingly unlimited, often subjective, number of ways. This type of intelligent analysis will be beyond analysis of visual features alone, often requiring knowledge outside that available in the visual content itself. Of course, texts can frequently be interpreted in many ways as well, but for retrieval purposes, word level indexing has generally been shown to be effective without needing to determine any particular interpretation of the text. In the case of images, not only are attempts at recognising features unreliable, there is no obvious parallel means of selecting indexing units for open domain retrieval. Current video media retrieval systems either focus on very narrow domains, for example identifying pictures of predefined named individuals, or seek to index images using low-level features, such as colour or texture. Indexing images using such low-level features is perhaps comparable to identifying the letters in a text document without determining what the words are.

The difficulty in indexing images and of specifying search queries for them means that retrieval of visual media inherently requires more user interaction than text retrieval. A user will typically initiate a search either using a text request which will locate some potentially relevant images or video based on their textual annotation, or they will select a sample image and request the retrieval system to "find me more like this", in response to which the system returns images with similar colour and texture profiles to those of the example. The user is then able to provide feedback on the images retrieved using this initial query, after which further searches are carried out, with feedback after each one, until the user's information need has been satisfied.

Since 2001 the TRECVID workshop has provided standard document collections for researchers to explore indexing and retrieval tasks for video data (Smeaton,

Kraaji, & Over, 2004). Tasks undertaken in TRECVID include: automated shot boundary detection, visual feature recognition, locating named individuals or events in video, and interactive searching of a video archive. TRECVID is proving instructive in the development and evaluation of MIR technologies, but perhaps the clearest message so far is the large amount of work that remains to be done to achieve mature MIR systems.

6. CONCLUDING THOUGHTS AND FUTURE CHALLENGES

This chapter has demonstrated how fundamental work on English language text information retrieval has been successfully applied for multilingual and multimedia documents. In each case the underlying probabilistic model has contributed to an effective IR system. For text retrieval in a new language it has been illustrated that the need is for the selection of appropriate indexing units and development of automatic indexing methods, including morphological processing, stop word lists, and suffix stripping algorithms. Research issues for CLIR relate primarily to translation methods to cross the language barrier between search requests and documents. For MLIR issues of translation are compounded with the need for effective merging of the document lists retrieved from different language collections. Speech and scanned text document retrieval have been shown to be remarkably robust to indexing errors in automatic recognition of their content. It is only in the area of visual media where Karen Spärck Jones's work in IR has not been fully explored. It is perhaps interesting to speculate as to whether the probabilistic model might be successfully adapted for indexing and retrieval of visual media. The ongoing issues of defining and recognising visual indexing features continue to be the focus of much research in visual media retrieval. However, the lessons from spoken and scanned text document retrieval suggest that a probabilistic IR model applied for visual retrieval would be robust to considerable degrees of indexing errors. However, there is already research underway exploring the use of the alternative language modelling approach to IR in visual retrieval (Westerveld & de Vries, 2004).

Solution of the problems of multilingual and multimedia information retrieval explored in this chapter does not represent the end of the story for research into information access technologies for this data. Research interest continues to evolve to embrace more challenging tasks. For example, work is currently being established in the areas of retrieval from multilingual collections of image and video archives, retrieval from multilingual web collections, and question-answering methods for multilingual and multimedia data.

7. AFFILIATION

Gareth J. F. Jones, School of Computing, Dublin City University, Ireland

8. REFERENCES

Ballesteros, L., & Croft, W. B. (1998). Resolving Ambiguity for Cross-Language Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 64-71, Melbourne, ACM.

Braschler, M., & Ripplinger, B. (2004). How Effective is Stemming and Decompounding for German Text Retrieval? *Information Retrieval*, 7(3-4), 291-316, Kluwer.

Callan, J. (2000). Distributed Information Retrieval. In W. B. Croft, editor, *Advances in Information Retrieval*, pp. 127-150. Kluwer.

Garafolo, J. S., Auzanne, C. G. P., & Voorhees, E. M. (2000). The TREC Spoken Document Retrieval Track: A Success Story. In *Proceedings of the RIAO 2000 Conference: Content-Based Multimedia Information Access*, pp. 1-20, Paris.

Glavitsch, U., & Schäuble. P. (1992). A System for Retrieving Speech Documents. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 168-176. ACM.

Gollins, T., & Sanderson, M. (2001). Improving Cross Language Retrieval with Triangulated Translation, In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Rretrieval*, pp 90-95, New Orleans, ACM.

Huang, X., & Robertson, S. E. (1997). Application of Probabilistic Methods to Chinese Text Retrieval. *Journal of Documentation*, 53(1), 74-79.

Hull, D. A., & Grefenstette. G. (1996). Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49-57, Zürich, ACM.

Johnson, S. E., Jourlin, P., Sparck Jones, K., & Woodland, P. C. (2001). Spoken Document Retrieval for TREC-9 at Cambridge University. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, pp. 117-126. NIST.

Jones, G. J. F., Foote, J. T., Sparck Jones, K., & Young, S. J. (1996). Retrieving Spoken Documents by Combining Multiple Index Sources. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 30-38, Zürich,. ACM.

Jones, G. J. F., Sakai, T., Kajiura, M., & Sumita, K. (1998). Experiments in Japanese Text Retrieval and Routing using the NEAT System. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 197-205, Melbourne, ACM.

Jones, G. J. F., Sakai, T., Collier, N. H., Kumano, A., & Sumita, K. (1999). A Comparison of Query Translation Methods for English-Japanese Cross-Language Information Retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 269-270, San Francisco, ACM.

Jones, G. J. F., & Lam-Adesina, A. M. (2001). Exeter at CLEF 2001: Experiments with Machine Translation for bilingual retrieval. In *Proceedings of the CLEF 2001: Workshop on Cross-Language Information Retrieval and Evaluation*, pp. 59-77, Darmstadt, Springer Verlag.

Jones, G. J. F., & Lam-Adesina, A. M. (2002). An Investigation of Mixed-Media Information Retrieval. In *Proceedings of the 6th European Conference on Digital Libraries*, pp. 463-478, Rome, Springer Verlag.

Lam-Adesina, A. M., & Jones, G. J. F. (2003). Exeter at CLEF 2003: Experiments with Machine Translation for Monolingual and Bilingual and Multilingual Retrieval. In *Proceedings of the CLEF 2003: Workshop on Cross-Language Information Retrieval and Evaluation*, Trondheim, Springer.

McCarley, J. S. (1999). Should we Translate the Documents or the Queries in Cross-language Information Retrieval. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 99)*, pp. 208-214, University of Maryland, MD, ACL.

Nie, J.-Y., Simard, M., Isabelle, P., & Durand, R. (1999). Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In *Proceedings of the 22nd AAnnual International ACM SIGI Conference on Research and Development in Information Retrieval*, pp. 74-81, San Francisco, ACM.

Ponte, J. M., & Croft, W. B. (1998). A Language Modelling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp275-281, Melbourne, ACM.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14, 130-137.

Rath, T., Manmatha, R., & Lavrenko, V. (2004). A Search Engine for Historical Manuscript Images. In *Proceedings of the 27th Annual International ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp369-376, Sheffield, ACM.

Robertson, S. E. (1977). The Probability Ranking Principle in IR. *Journal of Documentation*, 33, 294-304.

Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27, 129-146.

Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232-241, Dublin, ACM.

Robertson, S. E., Walker, S. & Beaulieu, M. M. (1999). Okapi at TREC-7: automatic ad hoc, filtering, vls and interactive track. In E. Voorhees and D. K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pp. 253-264. NIST.

Sakai, T., Koyama, M., Kumano, A., & Manabe, T. (2004). Toshiba BRIDJE at NTCIR-4 CLIR: Monolingual/Bilingual IR and Flexible Feedback. In *Proceedings of NTCIR-4*.

Salton, G, & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, 24, 513-523, Elsiver.

Savoy, J. (2004). Combining Multiple Strategies for Effective Monolingual and Cross-Language Retrieval. *Information Retrieval*, 7(1-2), 121-148, Kluwer.

Sheridan, P. & Ballerini, J. P. (1996). 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*, pp. 58-65, Zürich, ACM.

Smeaton, A. F., Kraaji, W., & Over, P. (2004). The TREC Video Retrieval Evaluation (TRECVID);' A Case Study and Status Report. In *Proceedings of RIAO 2004 – Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, pp. 25-37, Avignon.

Sparck Jones, K., Walker, S., & Robertson, S. E. (2000a). A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing and Management*, 36(6), 779-808, Elisver.

Sparck Jones, K., Walker, S., & Robertson, S. E. (2000b). A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information Processing and Management*, 36(6), 809-840, Elisver.

van Rijsbergen, C. J. (1979). *Information Retrieval*. (2nd edition) Butterworths.

Wechsler, M., Sheridan, P., & Schäuble, P. (1997). Experiments in Multilingual Information Retrieval using the SPIDER System. In *Proceedings of the 5th RIAO Conference, Computer-Assisted Information Searching on the Internet*, Montreal.

Westerveld, T. & de Vries, A. P. (2004). Multimedia Retrieval Using Multiple Examples. In *Proceedings of the Third International Conference on Image and Video Retrieval*, pp. 344-352, Dublin, Springer.