APPROACHES FOR CROSS-LANGUAGE INFORMATION RETRIEVAL

[1]Nasir Naveed, [2]Muhammad Tariq Pervez
[1]Department of Computer Science, Virtual University of Pakistan
nasir.naveed@outlook.com
[2]Department of Computer Science, Virtual University of Pakistan
Tariq_cp@hotmail.com

ABSTRACT: *Classical IR system works well for retrieving information from a corpus or internet having mono language documents. But it is difficult to build a multilingual IR system that has the capacity to retrieve information from a mix of documents, books, or content on internet written in variety of languages. The solution is to build a Cross-Language Information Retrieval (CLIR) system. In this paper, we have discussed basics of IR systems and how existing CLIR methods can be built on top of it. Different methods of CLIR are analyzed with the prospects of building a system which can make use of positives of all other systems to deal with issues prevalent in existing CLIR methods.*

Keywords: cross-language information retrieval, CLIR, across languages

INTRODUCTION

The internet contains information and documents in at least 51 languages [1]. Cross-language information retrieval deals with information and knowledge to make it accessible across languages. The first part of this work describes fundamental concepts of information retrieval and the second part expands the concepts to cross-language information retrieval.

1. INFORMATION RETRIEVAL

Information Retrieval (IR) is a sub-domain of computer science, which deals with the computer-based natural language processing. The English term "information retrieval" literally means sourcing information. This requires appropriate representations and possibilities for structuring of complex data and efficient content-based [2] search. An IR system tries to satisfy the information needs of a user adding it to a query matching search results.

The next section provides an insight into the fundamentals of information retrieval. Some essential processes are also explained using detailed examples.

1.1 The naive approach

A user tends to use certain key words to find information that satisfies her information need. To find a book; as an example, which contains words "Conspiracy AND Science AND NOT Illumination". The task for retrieval system is to find all the books that contain Conspiracy and Science words but do not contain the word Illumination. Most IR systems will handle this simple task with Boolean retrieval algorithm. However, in other areas of applications more efficient procedures and algorithms are required.

1.2 Fundamental strategy of Information Retrieval

Systems are to focus on the actual search request, which is often referred to as query. In the previous example from 1.1, the search query is "Conspiracy AND Science AND NOT Illumination". This query will not be processed in one phase. The query processing can be divided in three phases that are identification, data preprocessing and indexing [3]. Below these individual phases are explained with examples.

1.2.1 Identification

Depending on the context and the form in which the data to be searched, there may require different procedures for the preprocessing of the data. For example if a document is in HTML format then data must be extracted from html tags with preprocessing. In order to detect which preprocessing for a specific tag is to be applied, this must therefore first be identified. The identification is carried out via the IR system specific heuristics [4], such as the analysis of file extensions or from file headers.

1.2.2 Specific data preprocessing

File identification and preprocessing can make work convenient. After preprocessing, a file is converted into a uniform data structure (e.g. simple text). It follows two further preprocessing steps:

(a) The subdivision of this text in tokens
(b) Locate and remove the stop words.

a. By token we mean splitting text at its spaces. The next step would be to phrases (such as legal person) and proper names such as (Hong Kong, Mercedes Benz, and Lion King) identification. Dependencies and relationships of words found by a data-dependence and co-occurrence analysis [5]. For example a co-occurrence analysis shows that in English the tokens Hong Kong almost always occur together. Therefore, the token Hong Kong should be joined together. In the next step, the token words are reduced on the basis of its shape and meaning. Several morphological variants of a word in the IR are reduced to their base forms. For this activity another technique which is used is called stemming. There exist various stemming algorithms for English language that has experimentally shown to produce good results.

b. Th tokenization process does not remove the stop words. Stops words are those words which occur very frequently and usually is of no relevance to the contents of the document. Often a stop word in the English referred to as an article, conjunctions, prepositions or punctuation. To remove stop words from text, there exist lists which contain appropriate words. These stop words in a document are matched with words in the list. It is important to note that stop words are only removed after identifying the proper words; otherwise this can distort proper words (King Lion).

1.2.3 Indexing:

To avoid effort of linear search for each new search request through the entire corpus of documents, an index is created. Index helps in finding a posting list of a word, whereas posting list contains document ids in which that specific word is appearing. The simplest form of indexing is the Boolean IR model [6]. For our example book search, it is assumed that the words appearing in each book are indexed (see Table1). This simple index shows if a specific word is present (1) or

not (0). From Table1, the vector of document ids against each word can be identified.

Table 1. A token document Matrix

Document Token	Illumination	Limit	Sacrilege	The Swarm	Pattern Recognition
conspiracy	1	1	1	1	0
Science	1	1	1	1	1
Art	1	0	1	0	0
CERN	1	0	0	0	0
Illumination	1	0	1	0	0
Space	0	1	0	1	0
tidal wave	0	0	0	1	0

For search query "Conspiracy AND Science AND NOT evaluate Illumination, logical operations are applied on the Boolean values:

11110 AND 11111 AND NOT 10100
= 11110 AND 11111 AND 01011
=01010

At first glance it seems simple and effective and it is impractical in reality (at least in this scenario). The limitations of the Boolean IR becomes evident when you look at the example with realistic searches, such as the current books offering at Amazon.com, which is a database with more than 250,000 books and a total of more than 100000 vocabulary words (here, it can be assumed that a word is a token), with a byte of memory per entry, it is expected to be a table of approximately 2.91GB size will be searched. First of all, this size (for an Amazon Database) is still small but using Boolean model with phrase query will not retrieve the information. It will only provide information on books where the searched words appear. This was a simple example where document collection is small, and it is assumed each book is only of two pages long. The result was that the size of the matrix to a three-digit factor. For books with an average of 400 pages the size of indexing table could reach to a size of greater than 580GB [5]. using inverted indexing, one of the most important concepts of information retrieval, the size of the indexing table can be drastically reduced. To understand how inverted index saves space, assume a scenario where we have a collection of one hundred documents with each document is two pages long having 1000 words in each document. Our index table will be of 100*1000 matrix where 99,000 elements are 0. It is therefore possible with inverted index to save 99% of the space, if only the fields with the value 1 will be stored. This is exactly what happens in the inverted indexing: for each word, which occurs in the corpus a posting list is created, where posting list for a word lists the system generated document ids in which that word appears. The words are listed alphabetically in the inverted index and is saved so that you do not have to recreate it each time.

As previously mentioned, the three phases of data processing in information retrieval are depicted with example as given in Table 2.

Table 2. Three phases of data processing

Document1	Document2	Document3
<html> <body> Helium Chemical element, discovered by American explorer Bob Moon </body> </html>	A researcher in his Swiss laboratory was found murdered. He was very extremely promoted to soil.	May 2025: The Supply earth seems assured, since America on the moon promotes the element helium-3.
1- Identification of documents and selection of preprocessing		
Document1: HTML	Document2: TEXT	Document3: TEXT
2- Preprocessing: Documents in uniform format		
Helium chemical element, discovered by the American explorer Bob Moon	A researcher in his Swiss laboratory was found murdered. He was very roughly to floor	May 2025: The Supply earth seems assured, since America on the moon promotes the element helium-3

a) preprocessing: token unification and stemming (example)

Helium chemical element discover by	The American researchers Bob Moon	On researcher becomes in his laboratory murder find	He was very roughly to the floor	May 2025 the supply of the earth seem secure since	America on the moon promote the element helium 3

(b) remove stop words preprocessing:

helium chemistry element discover	America researchers Bob Moon	Researchers find murder laboratory	Very roughly soil inquire	May 2025 seem safe supply earth	America promote moon element helium 3

3- Inverter Indexing: Document 1

Token	Documents in which the Token	Token	Documents in which the Token
America	1	Researchers	1
Bob	1	Discover	1
Chemical	1	Helium	1

Inverter Indexing: Document 2 and Document 3

Token	Documents in which the Token	Token	Documents in which the Token
2025	1	Promote	2,3
America	1,3	Helium	1
Extremely	2	Helium3	3
Bob	1	Laboratory	2
Soil	2	May	3
Chemistry	1	Moon	1,3
Element	1,3	Murder	2
Discover	1	Seem	3
Earth	3	Safe	3
Find	2	Rough	2
Researchers	1,2	Supply	3

So far, we have explained three preparatory stages; identification, data-specific preprocessing and indexing are presented with examples. The importance of inverted index is explained with example. This part of the paper provides a short introduction to the IR. On the basis of these principles, the reader has got an introduction of the processes and procedures of the Information Retrieval which lead to the concept of the cross-language information retrieval, we explain in the next part.

2 Cross Language Information Retrieval

When classical IR is compared with cross-language information retrieval (CLIR), the difference is that the language of the search query is different from the document collection. to handle this difficulty, there exists three approaches [7]:

(1) The search request in real time is translated into all languages. Then for each language a separate search request is started.
(2) All documents should be indexed in all possible languages. Then search queries can be written, translated in the language of classic IR to operate on.
(3) All documents and search queries are translated into a main language. This can be a natural language (such as English, Chinese), or an abstract (language-independent) concept of space.

One difficulty that all these approaches have in common is the recognition of the language. Before we delve in various problems of the three approaches, lets first look at in section 2.1 on language identification.

2.1 Identify language(s)

In cross-language information retrieval, multilingual documents are searched, it is essential to identify the language of each document, in order not to risk issuing Polish queries to English texts (or generally speaking queries in language A to text written in language B) This step is necessary not necessarily to be taken at indexing of the documents, but at time of processing the search queries or when you compile the results.

The different algorithms for language identification in electronic documents are based on one and the same principle: strings in the text to be identified are compared with strings from a previously trained system. This contains information about the frequency distribution of certain words of all discerning languages. It is obvious that the system trained in language with the largest similarity to this text is also the language of the text. The differences between the various voice recognition algorithms are mainly used in the training of the system and the evaluation criteria for similarity of words.

Within the European Language area for language identification word-based approaches are used. The trained system knows common words and word forms of all languages and their frequency (average frequency of occurrence within a regular text). The training of such a system is relatively complex, since it is half done automatically. Especially for languages with more flexion, the texts with which the system is trained to be very long and are automatically classified as words strings manually checked for accuracy. Nevertheless, this word-based approach is less cumbersome when compared with matching of byte sequences-based approach because (1) for the majority of languages trained systems are already available and (2) the detection of very similar languages is improved.

In order to identify the languages for which no trained system is available or where the word-based algorithms are not applicable (e.g. Asian languages), the system is trained with byte sequences instead of words, which is often referred to as N-gram technology. In most cases, this approach is sufficient. Only languages with very similar N-grams, which is indicated by the same word strains within similar languages can lead to errors.

They are no difficulty for voice identification Standard documents of a length of more than 20 words, the regular text is included - that is, they contain at least some common function words or other high-frequency word forms. Here the detection rates of all known algorithms are over 99%.

2. PROBLEMS AND METHODS IN CLIR

In cross language information retrieval, as already mentioned, there are fundamentally different approaches. In the following sections all the individual approaches have been discussed. In addition, problems which they bring with them are also discussed [8].

2.2.1 Translate the requirements

The search request is translated in most CLIR systems, to the language of the documents to be searched. Then classical information retrieval actions are performed. With the help of dictionaries, words can be directly translated from one language to another. In the literature in connection with the translation, CLIR often uses multilingual thesauri of speech [8]. A thesaurus is a word network, whose terms by tables are connected to each other. Multilingual thesauri contain tables of equivalence between terms in different languages. With the help of this information, words that are semantically equivalent can be summarized.

The most words have multiple, partly widely varying meanings. Most of the meanings can in turn by different words are expressed in the target language. This characteristic is called the ambiguity (ambiguity). It is because without context information, the translation has to be chosen from variety of translation options.

There are several different methods to word sense disambiguation. Already in part 1 the co-occurrence analysis has been mentioned, the goal of which is to determine how often terms within a contextual framework are mentioned together. With the co-occurrence analysis many techniques can be used with the search query and its context to exclude translations. What is excluded while translating, depends on the probability that a system looks for the joint appearance of the searched words. This procedure is problematic if the user precisely needs excluded translations. This can easily happen if the information needs of the user are very special.

2.2.2 Compiling the documents

It is sometimes necessary that all documents should be in all the languages in which the search query may be available. In this case, there must nevertheless be some limitations. Either the number of the searchable documents and of the languages used as far as restricted, that a manual translation of the documents may be made and or it will fall back on machine

translation which results in partial decline in the quality of the translated texts. For both possibilities there are real scenarios. These translations are done with the intention of improving the information retrieval in the background.

Careful consideration in CLIR approach is required, since the actual IR performs very little cross language functionality, which is required in the previous steps, such as the indexing. For the major part of the language pairs and thus for the practical use the benefits gained by the translation is not sufficiently sophisticated and can be ignored. The machine translation of documents is a very important disciplines and developments in this area has a crucial impact on the CLIR.

2.2.3 Search Queries and documents in a uniform language translation

(a) The advantages of this technique are very promising. For example, if we select English as the common language of a system, then it will be possible to use all IR tools of the English language. In addition, the indexing table will be much smaller. Even in difficult-to-process languages such as Finnish, English or Asian languages, it would be easier against search query to find and assign the correct result in English language. It remains however the same problems which we have discussed with automatic machine translation in sections 2.2.1 and 2.2.2. Machine translation in certain languages is hard to achieve with good accuracy than translating it vice versa to English with high accuracy. For example, it is simple because of grammatical structures to translate from an Asian into English than vice versa. These properties will be used as to find a common language which can be used as reference language in future as well.

(b) A completely different approach it to move both documents and queries in a language-independent space of concepts without involving costly translation processes. One such method which is capable of moving documents and queries in the concept space is called Latent Semantic Indexing (LSI) [9]. In LSI the document is searched for the concepts without paying attention to the meaning of individual terms. In such space of concepts, a connection between the words Auto, car, truck, car etc. is found and exploited. The procedure is based on the theory that by singular value decomposition of the data, the frequency value of the term is approximated and used to reduce the number of dimensions. The benefit of this method is that the we are able to use monolingual IR without giving the benefits of multilingual techniques. The limitation of this method that low dimensional space is sometime difficult to interpret. It can partially handle the polysemy and suffers from same shortcoming which comes with bag of words model. Sometime the probabilistic method computed using LSI does not tally with real world observation. For the calculation of LSI model, a multilingual language corpus is required. In addition, the mathematical complexity of the singular value decomposition is O (n^2*k^3) [10].

3. CONCLUSION

The difficulty of identification of languages can be viewed as a largely solved problem. For all common languages there exists sufficiently well functioning heuristics, such as the use of the stop word lists or the n-gram procedure. Only one of the three presented CLIR approaches, namely the translation approach of translating the search request in all other system languages, at present is viable logistically and technically for a large area of applications. The results provided by this technique are reusable. The issue of word sense disambiguation remains unsolved in this approach because there is no sophisticated technology of machine translation exists. It depends on the quality of the results of language pair.

The other two methods have their advantages and disadvantages. The translation of all data in all possible languages can be exploited manually but for a very limited area of use with its drawback. As, more sophisticated methods of automatic translation are becoming available, this technique may be helpful in future to larger collection of datasets. An advantage of this technology is the known best quality of the search results. The last method presented has a lot of potential. It remains to be seen whether the quality of the results with the quality of the other procedures will be comparable and whether more efficient algorithms for transformation into a language-independent concept of space can be found.

REFERENCES

1. Cho, J. and Garcia-Molina, H., "The Evolution of the Web and Implications for an Incremental Crawler", *In Proc. of VLDB '00 Proceedings of the 26th International Conference on Very Large Databases*, 200-209(1999).
2. Bhogal, J., Macfarlane, A. and Smith, P., "A review of Ontology based Query Expansion," *Information Processing and Management Journal*, **43**(4): 866-886(2007).
3. Harshit, S., and Singh, A. K., "A More Discerning and Adaptable Multilingual Transliteration Mechanism for Indian Languages," *In Proc. of 3rd International Joint Conference on Natural Language Processing* (2008).
4. Egozi, O., Evgeniy, G., and Shaul, M., "Concept-based Feature Generation and Selection for Information Retrieval," *In Proceedings of the Twenty-Third Conference on Artificial Intelligence (AAAI)*, **2**: 1132-1137(2008).
5. Schoenhofen, P., Benzcur, A., Biro, I. and Csalogany, K. "Performing cross-language retrieval with Wikipedia," *In Proc. of CLEF*(2007).
6. Argaw, A. A., Asker, L., Coster, R., Karlgren, J. and Sahlgren, M., "Dictionary-based Amharic-French Information Retrieval" *In Proc. of CLEF*(2005).
7. Scannell, K., "The Crúbadán Project: corpus-building for under-resourced languages," *In Proc. WAC- 3: Building and Exploring Web Corpora, Louvain-laNeuve, Belgium,* (2007)
8. Ren, F. and Bracewell, D. B., "Advanced Information Retrieval," *Electronic Notes in Theoretical Computer Science*, **225:** 303-317(2009).
9. Peters, C., Braschler, M. and Clough, P., "Multilingual Information Retrieval - From Research to Practice," *Springer Science and Business Media*(2012).
10. Baker, E. and Rob, U., "Accessing the Compatibility of Document," *Department of Computer Science, University of Sheffield* (2012).