Multilingual Information Access: Information Retrieval and Translation in a Digital Library

Vamshi Ambati[1], Rohini U[1], Pramod P[1], N.Balakrishnan[3] and Raj Reddy[2]

[1]International Institute of Information Technology , Hyderabad, India. (vamshi@iiit.ac.in, pramodp@students.iiit.ac.in, rohini@research.iiit.ac.in),

[2]Carnegie Mellon University, USA (rr@cmu.edu)

[3]Indian Institute of Science, Bangalore, India. (balki@serc.iisc.ernet.in)

***Abstract* — Digital libraries have expanded in the recent years in scope and content to include content in a vast variety of languages. The development of technologies that enable access to this varied language information regardless of geographic or language barriers are a key factor for true global sharing of knowledge. Two such technologies that play a major role in success of multilingual digital libraries are Multilingual Information Retrieval and Translation. We describe our approach and implementation of a Multilingual Information Retrieval system that helps users identify multilingual content and also a customized Reading Assistant service that assists users of a digital library in reading multilingual content via translation. We discuss how both these approaches make extensive use of a Universal Dictionary, a collection of dictionaries of various languages across the world.**

***Index Terms* —Digital Library, Multilingual Access, Language Technologies, Machine Translation**

I. INTRODUCTION

Digital libraries have expanded in the recent years in scope and content to include resources in a vast variety of languages. The development of technologies that enable access to this varied language information regardless of geographic or language barriers are a key factor for truly global sharing of knowledge. Users of internationally distributed information networks need tools that allow them to find, retrieve and understand relevant information, in whatever language and form it may have been stored. This can be made possible by technologies such as machine translation, cross lingual information retrieval etc. Cross lingual information retrieval helps to search and retrieve content which is stored in one language using a language understood by the user. Machine translation aids to understand content stored in any language by translating it to a language understood by the user.

Cross lingual Information Retrieval has gained wide focus with the increasing multilingual content made available in the recent years. The content in one language possibly not familiar to the user is made available by querying in a language known to the user. The retrieved documents can then be translated and made available to the user for reading. Cross lingual information retrieval is typically done by first translating the query to the language of the document followed by retrieval of the documents using the translated query. Translation is typically done by using a bi lingual dictionary. Multilingual information retrieval aims at retrieval of content in various languages in response to a query in one language.

Machine translation has been a well explored area in the area of natural language processing. Translation is typically done by using manually fed in rules, a dictionary consisting of word to word (or phrase to phrase) translation from one language to the other. A lot of work has been done translating between two pairs of languages. However, much of the work focused on translation between English and world's major languages like Chinese, French, German, Italian, Japanese, Portuguese etc. Translation between English and other minority languages like Indian languages is much lesser explored. To our knowledge, there is very few work done on translation between languages other than English especially Indian.

The Digital Library of India (DLI) project [1] is a digitization initiative with motivations from the Universal Library Project and aims to digitally preserve all the significant literary, artistic and scientific works of people and make it freely available to anyone, anytime, from any corner of the world, for education, research and also for appreciation by our future generations. Ever since its inception in November, 2002 operating at three centers, the project has been successfully digitizing books, which are a dominant store of knowledge and culture. We now host close to two tenths of a million books online. These books come from various languages and the current collection being hosted consists of books from 18 different languages across the world. Such a truly multilingual digital library requires intelligent ways

of accessing the multilingual content to be useful for a wider range of users.

In this paper, we discuss the deployment of a multilingual information access solution to the Digital Library of India project. In this regard, we make use of Universal Dictionary, consisting of multiple language dictionaries. Universal Dictionary is a rare resource that can play a pivotal role in language research. It is an initiative to collect in digital form dictionaries of all the languages spoken in the world. We discuss how we have exploited such a resource to build our Multi lingual Information Retrieval System and a customized Multi lingual Reading Assistant service which assists a user in reading multilingual content by providing translations of words, phrases or sentences.

The rest of the paper is organized as follows. In section 2 we discuss in detail existing multilingual technologies and resources that are vital for providing multilingual access in a digital library. In section 3 we discuss the multilingual information retrieval feature in a digital library that helps in identifying and retrieving books in a language using a query in another language. In section 4 we discuss a reading assistant service that provides translation of a requested word or phrase or a sentence. This can be used as an indispensable tool by people to read a book from a different language. We conclude in section 5.

II. Multilingual Technology and Resources

Processing multilingual content requires a number of resources. Among these dictionaries is a prominent resource. For example, development of multilingual access solutions for a multilingual digital library consisting of various languages we need dictionaries. However, instead of using multiple bilingual dictionaries for every single pair of languages, the existence of a comprehensive dictionary, consisting of multiple language entries at the same place is an indispensable resource.

Also the representation of the language content and dictionaries for machine readability and processing is a major concern. There is a need that textual representation of content belonging to multiple languages has to be standardized. In this section we first describe the Universal Dictionary Project which acts as a major resource for language engineering and also a transliteration scheme that helps provide a uniform representation to the content. The multilingual information access solutions that we propose in the later sections are dependent on these technologies.

A. Universal Dictionary:

Dictionaries are one of the most important language resources in all Language Engineering activities. A bilingual dictionary or simply a dictionary for a language is a collection of all possible entries of that language mapped to the corresponding meaning of the word in another language.

A Universal Dictionary is a collection of dictionaries of various languages. A simple collection of such dictionaries may not be as useful unless we can traverse from one language to the other to retrieve meanings of a given word in a particular language. The Universal Dictionary Project conceived at Carnegie Mellon University, USA aims at achieving the same – a universal collection of dictionaries for all languages spoken across the world. The representation of the Universal Dictionary helps to traverse across languages and perform lookups from any language to any other language making it very useful for building language technology tools for various languages. In Table 1 we provide a sample list of the languages currently present in the Universal Dictionary.

TABLE 1
SAMPLE LANGUAGES IN THE UNIVERSAL DICTIONARY

Ayapathu	Dutch	Khowar	Russian
Bosnian	French	Kiribati	Votic
Bulgarian	German	Norwegian	Serbian
Canadian	Greek	Polish	Slovak
Cebuano	Hiligaynon	Portuguese	Swedish
Chamorro	Hindi	Roviana	Tagalog
Ukrainian	Kapampangan	Russian	Thai

The project has so far collected a number of dictionaries for different languages available freely over the web as well as by having people across the world to manually enter them. For instance, a lot of Indian language dictionaries are being entered from India. A similar effort is going on at other places of the world. ‘English’ language has been chosen as the pivot and all the entries of any particular language are being mapped to their corresponding English words. Hence in order to translate or retrieve a meaning of a word in one non-English language into another non-English language, we essentially have to perform two look-ups in the Universal Dictionary. A first look-up retrieves the English meaning of the word and the second look-up retrieves the meaning corresponding to the English word in the other non-English language. The Universal

Dictionary can be used as a simple translation tool for all the languages in the collection.

B. Transliteration Scheme and Editors:

In order to operate across various languages there is a compelling need for standardization of the transcription and the transliteration scheme used to represent the language, especially non-English content which is difficult to represent or display on a computer. There is a need for the development of such a digital representation that lays a common foundation for many languages and for seamless adaptation of algorithms in language technologies, this representation must also be parsable by many language processing tools and algorithms, such as for machine translation, information retrieval, text summarization and statistical language modeling.

A transliteration scheme to suit this purpose has been developed by IISc, Bangalore and Carnegie Mellon University, USA to represent the Indian as well as some non-English language scripts. It is called IT3 notation and is an adaptation of the widely known ITRANS developed primarily for Indian languages. IT3 is mapped to the corresponding Unicode font of the language and displayed in the language. The following are the salient points of this transliteration scheme.

1. It is case-insensitive.
2. This scheme is phonetic in nature, the characters corresponds to the actual sound that is being spoken. Thus a single transliteration scheme is used for all the Indian languages, as they share the same set of sounds.
3. Each character (corresponding to a phone/sound) should not more than three letters length.
4. There is a minimal use of punctuation marks in the composition of a character
5. It can easily be extended to incorporate other languages like European, Middle Eastern etc.

In order to key-in data using IT3 notation and the Unicode characters, we make use of a simple transliteration editor [2]. Any new language can also be added to this editor with minimal efforts. The editor currently supports about six Indian Languages and three foreign languages – Arabic, Persian and Urdu. Transliteration editors are essential to key-in the particular language scripts into the computer using QWERTY keyboard. In a Digital Library, a transliteration editor is used in the entry of meta-data of books and sometimes the complete content of the books too. It is also being used in the Universal Dictionary Project to create dictionaries for Indian languages.

III. MULTI LINGUAL SEARCH IN A DIGITAL LIBRARY

In a digital library with multilingual content, the users often are interested in searching and viewing books in various languages. However, some or most of the languages might not be familiar to the user but the user might be interested to search and view the book. Multilingual search plays an important role helping the users of a digitial library to search for content in various language using a language familiar to the user. Once the desired book or articles are obtained, the user can directly read it or translate it to the language comfortable to the user.

Multilingual information retrieval [3],[4],[5] is a well explored area in the area of Information Retrieval. It is typically done by first translating the query to the language in which the documents are stored. After the translation of the query, the problem now reduces to that of a search problem where the task is to find the documents relevant to the query. Broadly, it can be said that the task has been seen as a translation followed by retrieval approach. For purposes of translation, existing bilingual dictionaries [6] or machine translation systems were used. Also, parallel corpus which consists of source and target language sentence pairs was mined for extracting bilingual dictionaries. These dictionaries where then used in the translation.

Recently, other corpus based approaches based on parallel corpora have been proposed using language modeling techniques which gained attention [7],[8],[9]. However, due to unavailability of parallel corpora for many languages, we couldn't employ the approaches. We aim for a simple translation using the minimal linguistic resources in the form of dictionaries available effectively.

A. Our Approach

We perform multilingual information retrieval in the digital library by employing the translation followed by retrieval approach using a dictionary for translation. The architecture of our multilingual retrieval framework is shown in Fig.1. The search query from the user is taken in the transliteration IT3 scheme for non-English languages and in ASCII for English and other European languages. The search query is translated depending on the language chosen. The translated query is then passed on to the retrieval engine which retrieves the documents and the results are presented to the user. In the subsections below, we describe in detail the translation

process and the retrieval engine in our multi lingual information retrieval framework.

Translation

When the user enters a query, he also selects the language in which the query is posed. The user can optionally select the language in which he seeks the results. The query is then translated to the corresponding language. If the user does not select a particular language, the query is translated to all the available languages in the digital library. Translation of the query is done by looking up the dictionary. For this purpose, we have used Universal Dictionary described in Section II.

Search Engine

The search engine performs the monolingual retrieval of the relevant documents related to the translated query in that respective language. We use Lucene[10], an open source search engine for the same. Any retrieval engine consists of two important phase the indexing and retrieval.

Indexing: In the indexing phase, all the documents in the digital library are converted to a form understandable by the retrieval engine and stored. All the documents in a language are indexed and an index is created labeled by the language. In the same way, an index is created for each language labeled by the respective language. Lucene allows for performing the indexing incrementally which gives us the flexibility of gradually adding content whenever available to the index. The content of the books is present in ASCII scheme for English and European languages and in IT3 for Indian and other non-English languages supported by our Digital Library of India project.

Retrieving: In the retrieval phase, pages matching the translated query are retrieved. When the user has not selected the language he wishes to see there are multiple translated queries one in each language in the digital library. For each query, the corresponding index is searched and the results are retrieved. The independent sets of results from the queries are merged into a single list before presenting them to the user.

IV. Multilingual Reading Assistant

In the previous section, we have discussed the implementation of a multi lingual search solution for digital libraries that helps in retrieval of content of different languages using a query from a particular language. However, in order to be of greater use for end-users of a multilingual digital library it is more important for a user to be able to read through a particular book in a different language and still get the information that he needs. Every user is very comfortable in one particular language, say his 'mother language' or a primary language and also knowledgeable in other foreign or secondary languages. However, while reading a book in a foreign language he might still be faced with problems of language understanding. At this point he might want to know the translation of the word or phrase or sentence in that book to his primary language. Most of us face this issue even while reading an English book and we take the help of an English dictionary in this regard. The same is true and can often be seen while attempting to read a book in a foreign language other than the primary language of the reader.

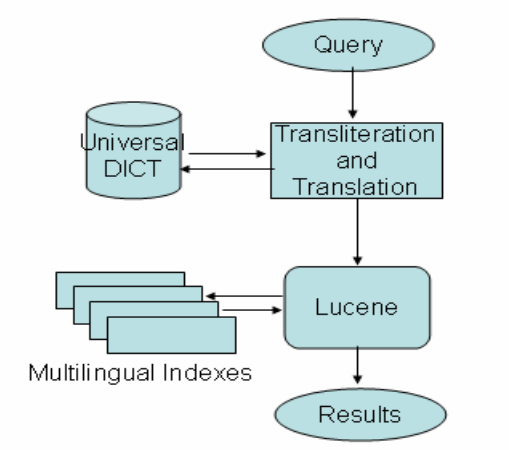

Fig 1: Multilingual Information Retrieval framework

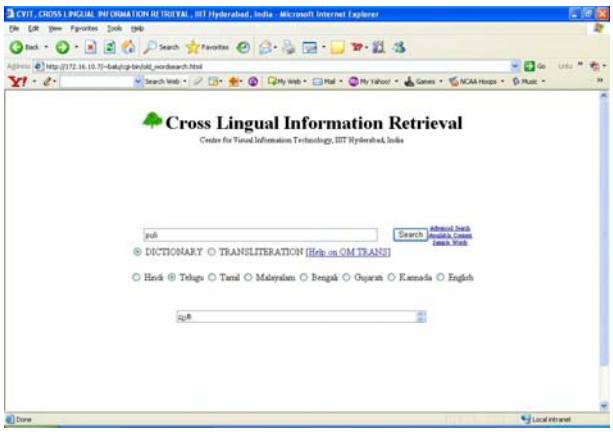

Fig 2: A Multilingual Retrieval system deployed for Digital Library of India

In this section we discuss a tool called the Reading Assistant, which helps a user read a book from a different language. Two primary sub components of

such a tool are the one that performs the translation of the requested word, phrase or a sentence and the interface that not only displays the translation but also helps a user provide feedback whenever unsatisfied with the translation. In the subsections below, we first describe the state of art of translation and how we perform the translation for our purposes in a digital library. We also discuss the architecture and implementation of the reading assistant application.

A. Translation

A Machine Translation system translates an input sentence given in a particular language to another language. Machine translation has been pursued for over 50 years now with a very vast literature. Broadly two different approaches have been pursued in the area of Machine translation. One is a knowledge rich approach with defined grammars and other linguistic resources for translation, also called rule based machine translation (RBMT) or knowledge based translation [11]. A second approach is one that uses huge parallel corpora in the translation process. Recently such corpus based approaches to machine translation have received wide focus. They are namely Example Based Machine Translation (EBMT) [12] and Statistical Machine Translation (SMT) [13].

Example based machine translation in its pure sense uses a parallel text corpus consisting of source and target sentences to obtain the translation. Given an input sentence, translation examples from the corpus that are best matched to the input are retrieved and adjusted to obtain the translation. Thus, the translation unit used in EBMT approaches is a complete sentence, providing a larger context for the generation of an appropriate translation. On the other hand, SMT approaches employing IBM models [13] translate an input sentence by the combination of word transfer ie (probability that the target language words or phrases generates the source words or phrases respectively) and word re-ordering (using language models of the target language).

B. Problems adopting earlier approaches

The earlier approaches to machine translation are not directly applicable for translation in the reading assistant. The problems in adopting the same are described in this section.

1. Most of the approaches proposed earlier in machine translation literature operate vertically focusing on a pair of languages as opposed to operating horizontally, catering for translations between a number of languages.
2. Most of the approaches proposed earlier in the literature make use of linguistic resources in the form of dictionaries, rules and some times a lot of manual effort. These resources cannot be assumed in our case given the large number of languages we are operating on.
3. On the other hand, the corpus based approaches rely on the availability of large amounts of corpus for learning of the translations between the pairs of languages. Even this is very difficult in this case. In a digital library like the Digital Library of India, there are books from about 18 different majority spoken Indian languages. Many of these may not still have the content in text form due to lack of OCRs for these languages, but parts of these books have been manually data entered and are available for people to read.
4. Due to heavy processing typically done in machine translation systems, they could take a long time for the translation. However, in our application, this might lead to frustration of the user. Hence, speed is also an essential feature in our case.
5. Since the reading assistant aids and assists the user in reading a book, customization of the translations should be possible.
6. The current approaches to machine translation aim at a perfect and precise translation.

To summarize, we need a simple, reliable and quick translation system and capable of adapting to the user's feedback. A rough or approximate translation given by the system suffices the need. Also given a digital library with N languages, if we have to cater to the need of all end-users and assist them in reading the content, we will be requiring N X N machine translation systems. We currently do not have such machine translation systems.

C. Our Approach

Our approach to translation is a simple one making use of the Universal Dictionary described in Section II and a phrasal dictionary. The user might request for the translation of a word or a phrase or a sentence. A word translation is performed by simply looking up the word in the Universal Dictionary. Phrase level translation is done by looking for phrasal translations in the Universal Dictionary, or an already existing dictionary of phrases or idioms for that particular pair of languages. If found, the translations found are given. Otherwise, translation is done by word to word translation of all the words in the given phrase.

We also provide APIs to plug existing machine translation systems and benefit from them. To translate a sentence we make use of an automatic machine translation system if one exists for that pair of languages. Machine translation systems for various pairs of languages do not still exist and may not produce acceptable translations even if existing. Therefore in the current implementation, whenever a machine translation output is not present for the sentence, or any unit larger than a phrase we consider it simply as a bag of words and perform a dictionary based translation. The goal of our translation is to aid and assist the user in understanding the document, but not to provide a perfect translation of the text. Our use of machine translation is strictly confined to assisting a user in reading and not educating the user in a different language.

D. The Reading Assistant:

Reading Assistant is a tool that assists a user in reading a book in a particular language possibly unfamiliar to the user. The reading assistant assists the user in reading a book by providing translations for one more words or phrases as mentioned in earlier sub sections.

The tool consists of the two main sub components. One is the interface that needs to be less distracting to the user and should blend well into his book reading activity. The other is a translation server which performs translation of the requested word, phrase or sentence. The architecture of the reading assistant is shown in Figure 3.

The user can start reading the textual content of a book and when faced with a word or phrase or a sub-sentential fragment that he does not understand, he could simply select the particular text and request for a translation. The multilingual translation server, present on our Digital Library servers, processes the request and returns the result in the requested language. The translation server makes uses of three primary resources. First is a machine translation system if it exists for the particular pair of languages. Second is the Universal Dictionary which provides a good enough translation for the sub-sentence or phrases or words. There is a third database of user specific entries collected as part of his feedback. This database carries primary importance and can override the translations from the other two resources.

We have deployed the translation server on the Digital Library of India project. The deployed translation server provides a translation using the approach discussed in previous section. The reading assistant service has been incorporated into the existing book reading interfaces of the project. If the user is unhappy with the translation result, he can provide feedback by entering the right meaning of his request. As mentioned, such requests are stored in the user specific entries database and are used to provide customized translations to the user in the future.

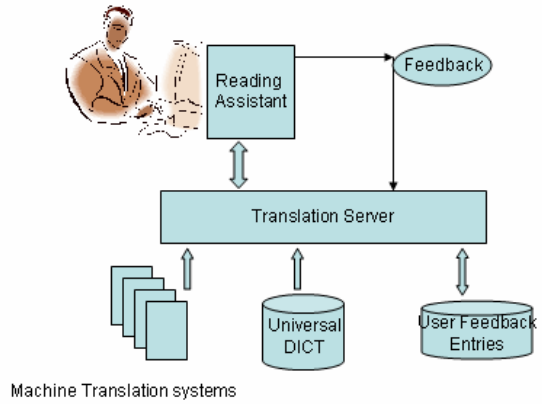

Fig. 3 Framework for Multilingual Reading Assistant

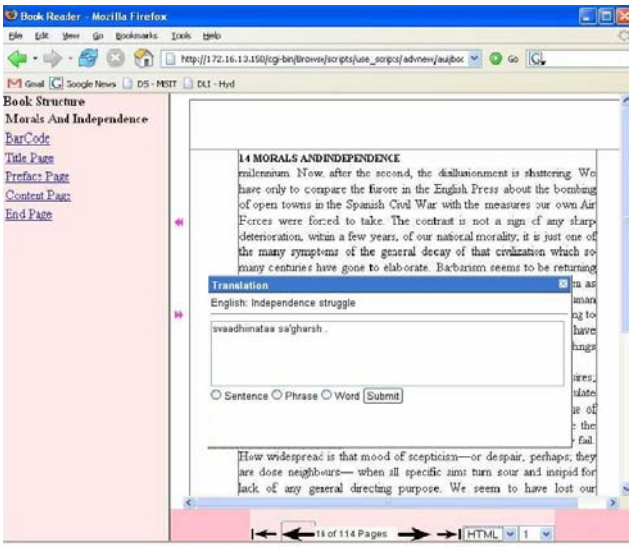

Fig.4 Multilingual reading assistant integrated into a digital library book reading interface

V. CONCLUSION

In this paper we have discussed Multilingual Information Retrieval and Translation as two language research technologies that could play a vital role in the success of multilingual digital libraries. We described our approach and implementation of a Multilingual

Information Retrieval system that helps users identify multilingual content. We also discussed a Reading Assistant service that assists users of a digital library in reading multilingual content via translation technology. We discussed how both these approaches that depend extensively on a Universal Dictionary, have been deployed on a real world digital library project – Digital Library of India to provide true multilingual access.

REFERENCES

[1] Vamshi Ambati, N.Balakrishnan, Raj Reddy, Lakshmi Pratha, C. V. Jawahar, "The Digital Library of India Project: Process, Policies and Architecture", *Proceedings of 2nd International Conference on Digital Libraries*, 2006.

[2] Lavanya Prahallad, Kishore Prahallad and Madhavi GanapathiRaju, "A Simple Approach for Building Transliteration Editors for Indian Languages", *Proceedings of 1st ICUDL*, 2005.

[3] Salton G , "Experiments in multi-lingual information retrieval", *Information Processing Letters*, pp. 6-11, 1973.

[4] Gregor Erbach, Gunter Neumann, and Hans Uszkoriet, "Mulinex multilingual indexing, navigation and indexing editing extensions for the world-wide web", *AAAI Symposium on Cross Language Text and Speech Retrieval*, 1997.

[5] Christian Fluhr, "Multilingual information retrieval", *Survey of the state of the art in Human Language Technology*, pp 391-405, 1995.

[6] Lisa Ballesteros and Bruce Croft, "Dictionary methods for cross-lingual information retrieval", *Proceedings. of 7th International DEXA Conference on Database and Expert Systems Applications*, pp. 791-801, 1996.

[7] Wessel Kraaij, Jian-Yun Nie, and Michel Simard, "Embedding web-based statistical translation models in cross-language information retrieval", *Computational Linguistics*., vol. 29 no. 3, pp 381–419, 2003.

[8] Victor Lavrenko, Martin Choquette, and W. Bruce, "Croft. Cross-lingual relevance models", *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '02)* , pp 175–182, 2002.

[9] J. Xu, R. Weischedel, and C. Nguyen, " Evaluating a probabilistic model for cross-lingual information retrieval", *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01)*, pp 105–110, 2001.

[10] http://lucene.apache.org, 2006

[11] Nyberg and Mitamura, "Controlled Language and Knowledge-Based Machine Translation: Principles and Practice", *Proceedings of the First International Workshop on Controlled Language Applications (CLAW '96)*, 1996.

[12] Makoto Nagao. "A framework of a mechanical translation between japanese and english by analogy principle", .Artificial and Human Intelligence, pages 173–180, 1984.

[13] Peter F. Brown, Stephen A. Della Pietra, Vi cent J. Della Pietra, and Robert L. Mercer. "The mathematics of statistical machine translation: Parameter estimation". *Computational Linguistics*, vol. 19 no. 2, 1993.