Issues in Machine Translation of Indian Languages for Information Retrieval

Margi Patel
Military College of Telecommunication Engineering
Mhow, India
margi.patel22@gmail.com

Brijendra Kumar Joshi
Military College of Telecommunication Engineering
Mhow, India
brijendrajoshi@yahoo.com

Abstract— Natural languages differ from one geographical location to the other. In India, there are 22 official languages [1]. Many documents have been digitized as a result of the advancement of information technology. Machine Translation Systems (MTS), as well as information retrieval systems, are required in order to retrieve any information from an existing digital document in any natural language. This paper describes some of the most important domains in information retrieval where Machine Translation (MT) is essential, such as Cross-Lingual Information Retrieval (CLIR) and Multi-Lingual Information Retrieval (MLIR).

Keywords-Machine Translation; CLIR; MLIR; MTS

I. Introduction

The automatic translation of text from one language to another is referred to as Machine Translation [3]. Machine Translation System (MTS) is an application of artificial intelligence in Natural Language Processing (NLP). The language of the input text is referred to as the source language, while the language of the output text is referred to as the target language. These days MTS is an arising area of study for scientists in India. India is multilingual country. Indian government utilizes Hindi or English language as a correspondence medium while different states of India utilize their local language as a correspondence medium [4]. There is a major interest for record transformation starting with one language into the other language. The English language is generally utilized in all fields. So MTSs are required for interpretation of local language to English language or vice- a-versa.

The act of storing, finding, and retrieving information from a database that fits a user's request is known as Information Retrieval (IR) [5]. Since non English material (Hindi, Gujarati, etc.) is rapidly expanding, the digital world is no longer monolingual. The capacity to obtain information in different languages is becoming important in an increasingly globalized economy. In the digital age, the multiplicity of languages is becoming a barrier to understanding and familiarity. As a result, IR has become a critical field of study in recent years. It has been discovered that when users receive services in their native language, they are more likely to accept and use them.

One of the most significant challenges in CLIR and MLIR is identifying relevant material for a query issued in the user's native language. As the World Wide Web expands, so does the amount of material available in languages other than English on the internet. In recent years, there has been a tremendous rise in the availability of non-English content on the internet. All important government institutions, newspapers, and publishing firms created websites in Hindi or Gujarati or any native language [6]. National boundaries are becoming less important in terms of commerce and information sharing as a result of globalization. Hindi is the world's third most commonly spoken language. Gujarati is also the most frequently spoken language in Gujarat. India is diverse in terms of languages, and just 12% of the population is familiar with the English language [7]. IR in languages such as Hindi, Gujarati, English etc. is gaining popularity. Google now supports transliteration in 14 languages namely Arabic, Bengali, Farsi (Persian), Greek, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Punjabi, Tamil, Telugu and Urdu [8]. Society gains the benefit of allowing users to access information in their native language and retrieve information in the same language without knowing which language the information is stored in the database through the MLIR process, making it a very effective research area. The IR system may help people in various areas, including agriculture, rural health, education, national resource planning, crisis management, information kiosks, and others.

In terms of IR development a lot of work is being done. Other relevant fields of research are also being pursued, such as NLP, MT, and so on. For IR, scholars have considered a variety of regional languages. Government organizations such as TDIL (Technology Development for Indian Languages) have also made major contributions to the standardization of Indian languages on the web [9].

II. LITERATURE REVIEW

Extra information is available online in the form of text, audio, video, and other media. This source will provide users with important information. IR is the act of obtaining essential documents or information from the content of a data collection such as the World Wide Web [10]. A web search engine is an IR system that searches the World Wide Web for information. IR involves a number of components such as:

Crawl: Download and save documents from the Internet.

Index: Searched documents are indexed.

Query: It is the user's input.

Rank: The system creates a list of documents that are rated based on their relevance to the query.

The availability of information on the internet is expanding in a variety of formats and languages. Though English used to dominate the web, it currently accounts for fewer than half of all documents on the internet. Because of the popularity and the availability of online information sources, there will be a greater demand for CLIR systems. CLIR is the search for documents written in a language other than the language used to clearly explain a query. This allows users to browse a collection of documents in a variety of languages to obtain valuable information in a usable format, even when the target language has little or no linguistic ability. CLIR is critical in countries like India, where a substantial proportion of the population does not speak English and so does not have access to the huge store of knowledge available on the Internet.

There are various areas where MT systems can be coupled with IR systems:

A. Books:

The books can be identified in order to accommodate to the representation of various domains. The key goal in picking a book is to select from a wide range of domains so that the corpus may cover a substantial portion of the vocabulary and not lose out on domain-specific words.

B. Magazines:

Magazine corpus often comprises a variety of texts such as cuisine, health, movies, fiction, current news, and so on.

C. Newspaper:

The newspaper corpus is made out of current text. Political news, editorials, sports news, and so on may be included in the text. The news articles include a wide range of topics, including religious views, scientific principles, and other concepts that must be communicated to the general public. As a result, it is assumed that the authors would have captured these domains in a clear and understandable manner. Such articles make good use of terminology, have a clear linguistic structure, and use effective phraseology. To entice readers, newspaper stories may employ colloquial, non-standard words or jargons. The words chosen must be descriptive and reflect the author's feelings and attitude toward the occurrences.

Following are the categorization of the contents that a common man may be interested in:

A. Science and Technology:

The science and technology domain comprises text extracts from numerous scientific publications, magazine articles, journal articles, and so on. These are also known as knowledge texts. The linguistic structure and word use differ from everyday language. The subject of the text is generally worldwide, therefore terminology of this domain will include the most borrowed terms.

B. Social Sciences:

Since language is a tool for the development and preservation of human society, language in the social sciences category correlates the linguistic aspects of the dynamic society. Human development and reformation are occurring in many community contexts, therefore all social knowledge and reality may be portrayed in this literature genre.

C. Official Document

The language used in official documents is highly standardized, clear, straightforward, and structurally changed. Official documents are designed to communicate about some activity, inquiry, or proceedings of some assembly.

Following are the subdomains recognized under Official Document, see Figure 1.

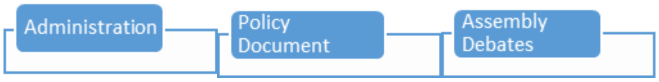

Figure 1: Subcategories of Official Documents

D. Aesthetics:

This category includes subdomains from Literature and Fine Arts. The text samples are taken from literary works. It is used to record literary phrases. Aesthetics content is compiled from several works. The text is most likely a typical description text of some kind. It exemplifies the linguistic style of the time period from which the work is drawn. It is a snippet from a piece of creative writing. It consists of fictional stories, essays on various themes, and so on. These articles are primarily the writer's personal expressions. It captures the writer's flow of words in the literary work.

The subdomains recognized under Aesthetics are given in Figure 2

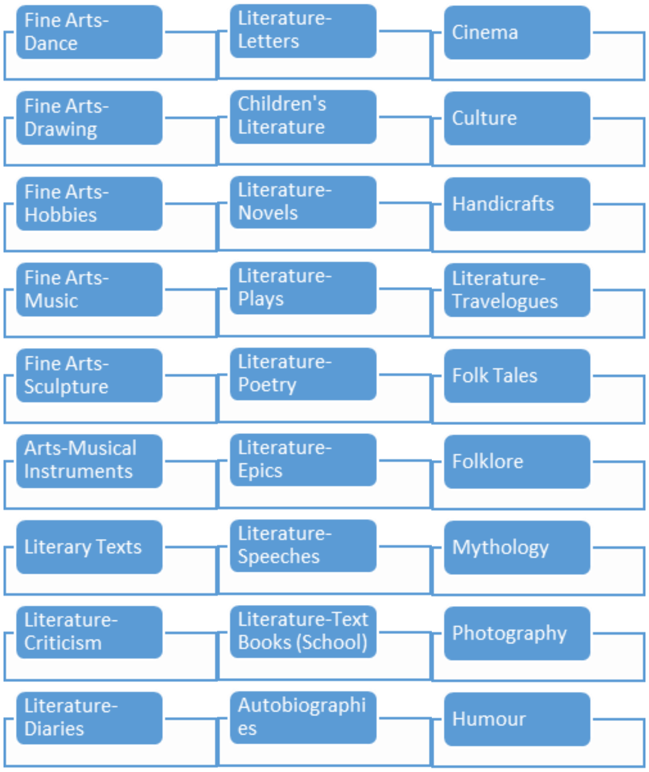

Figure 2: Subcategories of Aesthetics

E. Commerce:

The commerce is an important component of society. It exists and functions in collaboration with a variety of societal groups, including customers, suppliers, rivals, banks and financial institutions, government agencies, and labor unions.

The subdomains recognized under Commerce are shown in Figure 3.

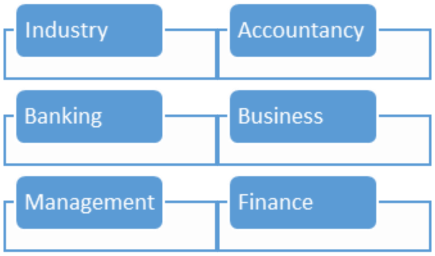

Figure 3: Subcategories of Commerce

F. *Mass Media*

For many individuals all around the globe, media is an essential part of their daily lives, both at work and at home. The text of this domain is a current character. Political news, editorials, or sports news may be included in the text. The newspaper is the primary source of the Mass Media text category. It contains terms that are often used in everyday life. Exposition, argument, description, and narrative are all structural elements of mass media language. It comprises many forms of writings. It consists of structures with various patterns, vocabulary, and styles. Everything is written in a language that everyone can connect to and comprehend. Some of the media prints take the shape of a dialogue or a series of questions and replies. This data often include an interviewer and an interviewee. They are generally conversations. The interviewee might be a celebrity or a well-known figure from the world of films, politics, or other fields. The words used in this type of literature are typically more personal and straightforward.

The subdomains recognized under Mass Media are depicted in Figure 4.

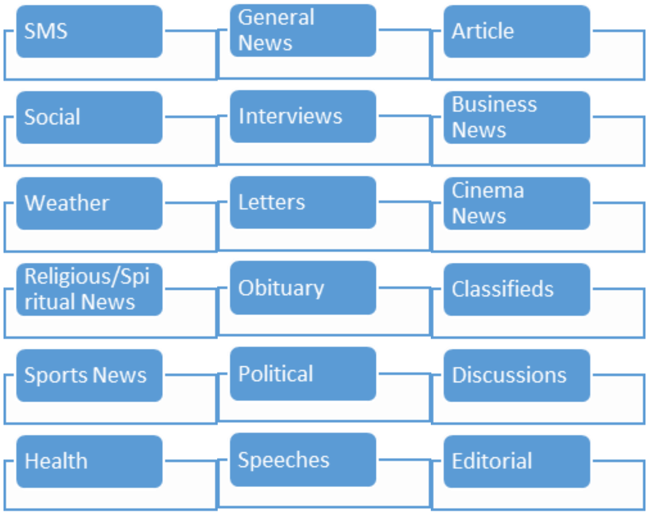

Figure 4: Subcategories of Mass Media

III. CHALLENGES IN DEVELOPING CROSS-LINGUAL INFORMATION RETRIEVAL

Following are the challenges in developing a CLIR system:

1. Conflict in translation:
More than one translation may be available when translating from source to target language. It is difficult to choose a suitable translation. For example, the word હાર (haar) has two connotations; defeat and necklace.

2. Phrase recognition and translation:
It is difficult to recognize phrases in a limited context and translate them as a whole unit rather than as separate words.

3. Transliterate/translate a word:
There are several confusing names that must be transliterated rather than translated.

For example, મૌલિક (Maulik) in Gujarati refers to a person’s name as well as Original. It is difficult to detect these situations solely on the given context.

4. Transliteration errors:
Errors in transliteration may result in the erroneous term being returned in the target language.

5. Dictionary coverage
When translating with a bi-lingual dictionary, dictionary completeness is an important criterion for measuring system performance.

6. Font:
Many web documents do not use the Unicode character set. Before they can be processed and stored, these documents must be converted to Unicode format.

7. An analysis of morphology (differs for each language)

8. Words that are not in the vocabulary
New terms are added into the language that the system may or may not recognize.

IV. CONCLUSION

Many domains where MT may be utilized for IR are described in this paper. Cross-lingual and multi-lingual IR introduce new paradigms for finding texts in many languages throughout the world, and this may serve as the foundation for searching not just between two languages but also across several languages. For many years, MT has been a thriving research sector in artificial intelligence and IR systems. Since natural languages are very complex, MT is a difficult task. As languages are evolutionary in nature, it is impossible to assume that one technique would be sufficient to handle the translation process. A few of the difficulties encountered in CLIR have also been highlighted.

V. REFERENCES

[1] https://indianexpress.com/article/india/more-than-19500-mother-tongues-spoken-in-india-census-5241056/ accessed on 24 July 2021.

[2] Jatin C. Modh , Dr. Jatinder Kumar R. Saini, “Study of Machine Translation Systems and Techniques for Indian Languages”,International Conference on The Journey of Indian Languages: Perpectives on Culture and Society, 14-15 October 2017,jointly organized by Dr.Babasaheb Ambedkar Open University and Indira Gandhi National Open University, 14-15 Oct. 2017, Ahemdabad, Gujarat.

[3] J. Hutchins and H. Somers, “An introduction to Machine Translation”, Academic Press, 1992, London.

[4] Jatin C. Modh , Dr. Jatinder Kumar R. Saini, “A Study Of Machine Translation Approaches For Gujarati Language”, International Journal of Advanced Research in Computer Science, Volume 9, Issue 1, January-February 2018, pp 285-288.

[5] Pothula Sujatha and P. Dhavachelvan. “A Review on the Cross and Multilingual Information Retrieval”, International Journal of Web & Semantic Technology (IJWesT) Vol 2, Issue 4, October 2011, pp 115-124.

[6] Mangala Madankara, Dr. M .B. Chandakb, Nekita Chavhanc, “Information Retrieval System and Machine Translation: A Review”, International Conference on Information Security and Privacy (ICISP2015), 11-12 December 2015, Nagpur.

[7] http://bweducation.businessworld.in/article/The-Importance-Of-English-Language-In-A-Country-As-Diverse-As-India-Where-Language-Can-Be-A-Common-Ground/02-10-2018-161270/ accessed on 24 July 2021.

[8] https://www.financialexpress.com/archive/google-transliterates-14-indian-languages-into-their-script/572706/ accessed on 24 July 2021.

[9] M. S. Madankar, “A Review on Information Retrieval in Indian Multilingual Languages”, International Journal of Advanced Research in Computer Science and Software Engineering. Volume 5, Issue 3, March 2015, pp 48-52.

[10] https://www.cfilt.iitb.ac.in/resources/surveys/Swapnil-Cross-lingual-Information-Retrieval.pdf, accessed on 24 July 2021.

[11] https://data.ldcil.org/upload/Overview/RawTextOverview.pdf, accessed on 24 July 2021.

AUTHORS PROFILE

Margi Patel is pursuing her PhD. Her research center is Military College of Telecommunication Engineering, MHOW (MP), India. She did her M.E. with specialization in Software Engineering from IET, DAVV. She joined IIST, Indore as an Assistant Professor in CSE Department in July 2006. Her research interest is in Natural Language Processing and Machine Learning.

Dr Brijendra Kumar Joshi is associated as a Professor of Electronics & Telecommunication and Computer Engineering at Military College of Telecommunication Engineering, MHOW (MP), India. He obtained BE in Electronics and Telecommunication Engineering from Govt Engg College, Jabalpur; ME in Computer Science and Engineering from IISc, Banglore, PhD in Electronics and Telecommunication Engineering from Rani Durgavati University, Jabalpur, and MTech in Digital Communication from MANIT, Bhopal. He has more than 37 years of teaching experience. His research interests are programming languages, compiler design, digital communications, mobile ad-hoc and wireless sensor networks, image processing, software engineering and formal methods. He has number of research publications to his credit. He has supervised six Ph D thesis and currently supervising nine research scholars. He has authored two books on Data Structures and Algorithms in C/C++ published by Tata McGraw-Hill, New Delhi.