Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 2 (2025)

Cross-Language Information Retrieval for Poetry Form of Literature-Based on Machine Transliteration using Modified Cosine Similarity Algorithm

Ranjana S. Jadhav[1], Manikrao Laxmanrao Dhore[2]

[1,2]Asst.Prof., Dept of Computer Engineering, Vishwakarma Institute of Technology, 666, Upper Indiranagar, Bibwewadi, Pune, Maharashtra, INDIA - 411 037.
ranjana.jadhav@vit.edu[1],manikrao.dhore@vit.edu[2]

Article History:

Received: *02-08-2024*

Revised: *11-09-2024*

Accepted: *20-09-2024*

Abstract:

Introduction: The current era of the web content retrieval is enormously increasing. The regional users are accessing the web contents directly through the search-based information retrieval. The amount of contents and the literature available in the regional language is increasing day by day.

Objectives: This work presents the finding semiotic similarities between the poetry form of data in the form of stanza, lyrical and poetry. This study shows the methodology to retrieve the semiotic similarities between the regional contents available on World Wide Web.

Methods: The approach proposed a system that effectively recognize and recommend regional information based on the user queries regardless of query written in the Devanagari or Romanized script. The method achieves the objective to fetch the given query written in Devanagari or Romanized from the Marathi database to implement the search-based information retrieval. To achieve this input query is converts Devanagari script poetry into Roman script for the next processing steps. It uses defined Custom Mappings and Custom transcriptions Function by pre-processing the data, the model ensure that it is clean and in a suitable format for generating embeddings and performing similarity searches. These embeddings generated for regional input query and transliterated poetry stanza, sequentially combined to improve the algorithms accuracy to identify and find the input among different script.

Results: The input query written in Roman script transliterated into Marathi using the customized transcription function to generate the embedding further. The adapted cosine semiotic similarity value is used to compare the embeddings which makes the model to fetch the most semantically matched poetry stanza.

Conclusions: The proposed customized cosine semiotic similarity for retrieval achieves the accuracy of 92% and a loss of 0.18.

Keywords: transcription, Devanagari, information retrieval, semiotic similarity, Embeddings

1. Introduction

The method based on language frequently distinguishes identified substances through assumptions and developed manual specifications by linguists. Using a handcrafted technique based on pattern matching, the linguistic approach creates rules by means of linguistic analysis. Proficiency in

grammar and other language-related rules is necessary. It necessitates a thorough understanding of grammar and other linguistic principles. The primary drawback of this method is that it necessitates extensive knowledge and structural information of the objective linguistic or area; also, these schemes cannot be applied to further idioms or areas. Nonetheless, the expense of translating a rebases from one Indian language to another might be relatively low due to the shared characteristics of numerous Indian languages. The rules-driven method is a different acronym for this type of approach.

A list of trigger words, gazetteer lists, and lexicalized grammar are among the various rule-based systems. It needs rich and expressive rules to obtain good results. Numerical method attempts to engender transcriptions using statistical approaches built on multi-lingual text corpora. To get good results, it requires explicit and rich regulations. The statistical analysis methodology uses statistical techniques based on bilingual text corpora to present transcriptions.

Transliterations are engendered built on arithmetical representations, which are consequent from the study of multilingual transcript quantities. The elementary axiom-based translation prototypical is an incidence of the deafening network method in which the paraphrase of French information into English is demonstrated. This prototypical existed advanced for mechanism paraphrase wherever the contribution remained a French information. The identical equivalence more sustained for arithmetical auto transcription by substituting a term in the data by an alphabet or a cluster of alphabets in the termed elements. [Brown 1993].

The given hybrid model incorporates ideas from statistical and rule-based methodologies. It is based on the Bayes Theorem and provides a generative model where the system looks for the English word that maximizes probability of given a Japanese Katakana string, represented, as seen by an optical character recognition (OCR) program. [Knight 1998].

The conditional probability of objective terms given a basis term is represented by the basis network prototypical. Presume that a word which is signified as a string of elements in the basis linguistics, desires to be recorded as a string of letters in the objective language. The combined basis network prototypical, when compared to the noisy network prototypical, focuses on how the basis and objective can be formed simultaneously instead of attempting to represent how the basis might be recorded to the objective. Put differently, it estimates a readily marginalizable joint probability model that produces conditional odds models involving back-transliteration and transcription [Li Haizhou 2004].

A robust probability estimation tool for simulating generative sequences, or sequences that may be described by a secret mechanism providing an observable sequence, is the concealed Markov prototypical. The Hidden Markov Model is a directed graphical model which estimates conditional probability distributions based on limited history using the Markov property. There are N numbers of conditions in the model. While conditions are concealed, for countless real-world applications there is frequently around corporeal consequence devoted to the states or the groups of conditions of the prototypical.

They stand a kind of the generative model, which defines probability distribution with respect to observation sequences and their corresponding label sequences [Rabiner 1989]. A CRF is a kind of

undirected model just like graph models; It describes one specific distribution of label sequences conditional on an observed sequence of labels. The conditional probability is identified by the CRF model. The distribution of label sequences under a provided observation sequence as contrasted to joint probability distribution between the tag and the observation arrangements. It formally brings G = (V, E) to denote an undirected graph for each of the random variables signifies an element. However, if each random variable Yv satisfies Markov property with respect to G, then, the model is limited to conditional random fields of (Y, X). In machine transcription, the option CRF can be used and from the basis language word, it can generate the objective language word. CRF has been defined here as conditional probability distribution functions that map the basis language words to the objective language words. [Lafferty 2001, Wallach 2004].

A true and effective probability model, the maximum entropy model has been applied in heterogeneous information indeed. In the MEM, an event (ev) is often described to be comprise of a objective event (te) and a history event (he); let us write ev=<te, he> The existence of some characteristics in event ev is represented by a set of feature functions, fei(ev). In feature learning a feature function is computed to be a binary-valued function. It is activated if it undergoes its activating condition; else, it is deactivated (fei(ev) = 0). [Berger 1996]. Hybrid Approach is a good approach of converting both rule based and statistical approaches to create maximum performance. Combined Approach used multiple graphemes or phoneme-based methods at a time but not both.

In this section we have discussed existing machine transcription models as grapheme based, phoneme based, Hybrid transcription and combined transcription methods. The detailed literature survey of the transcription for the cross-language information retrieval is discussed in the section 2. The section 3 is about the proposed methodology and Result analysis presented in the section 4 whereas the study is concluded in the section 5.

2. Literature Survey

Local language technology in India during 1980-90: Presently, the Government of India's, C-DAC is the market leader in local language software. Early in the 1980s, GIST (Graphics and Intelligence – Based Script Technology) was successful. In matters related to IT application for Indian Languages, a work that is path breaking has been done by CDAC. The Indian Scripts can be typed on a keyboard having an overlay of the concerned Indian Scripts, and any non-Indian scripts can be converted to the desired Indian scripts based on the information entered on the keyboard and C code (see fig 2) allows display of various Indian scripts on the computer monitor screen. Taking in consideration the interest of the Indian language the Department of Electronics (DoE) arranged a first symposium on 'Linguistic Implications of Computer Based Information Processing' in 1979 and ten projects were initiated at the various Indian Institute of Technologies (IITs), Bombay University, Birla Institute of Technology and Science (BITS) Pilani, National Council of Education (NCE) Calcutta and Jadhavpur University. The Department of Electronics initiated a programme on "Electronic Tools for Indian Languages (ETIL)" in 1987. This programme was concerned with the lexicographic processes for Indian language processing, new areas of applications of the developed tools, evaluation and feedback and the dissemination of the research outcomes in operational systems.

Local language technology in India during 1990-2000: During the events of the VIIIth Plan in India it was proposed to lay greater emphasis on language technology for Indian languages. During 1990-91, the Department of Electronics, Government of India, started a national level programme 'Technology Development for Indian Languages (T D I L)' with the specific aim to come up with the necessary technologies' tools for Indian languages processing and to carry out Human Machine Interface in Indian languages. Authenticated collection of India language text of about 29, 00,000 words in Tamil, Telugu, Kanada, Malayalam, Assamese, Bengali, Oriya, English, Hindi, Punjabi, Sanskrit, Kashmiri, Urdu Marathi, Gujarati & Sindhi have also been created. IIT Chennai has prepared interface software which can be used for preparation of documents in all Indian languages and some of the foreign languages for viewing as well as for printing. It is now possible to transmit TV programmes with dubbing and subtitling in various Indian languages. ISFOC fonts for Indian languages have also been produced by C-DAC and many of font types are available.

A team of Indic trans in India achieved two key goals while adopting local language technology from 2000 to 2011. The first one consisted of migration of live data and file-journey-management database from non-Unicode to Unicode standard. The second one was for searching the voter list for the Chief Electoral Officer of Maharashtra. NCST and C-DAC who have been doing path breaking work in the area of Indian language support on PC's have designed the transcription software packages. Now the above-mentioned packages are being applied by the Indian Railways to display reserve charts, by the Mahanagar Telephone Nigam's to print bilingual telephone directories and finally by the Mahavitaran to print bilingual electricity bills [Murthy 1998, Frost 2003, Badodekar 2003, Shah 2004].

Lee & Choi (1998) and Jeong, Myaeng, Lee, & Choi (1999), have implemented their schemes through straight orthographical plotting from basis graphic symbol to objective graphic symbol. The authors have used the basis network model for English to Korean transcription. They use a portion of graphemes that can resemble to a basis phoneme. Initial, English words are segmented into a mass of English graphemes. Furthermore, all probable masses of Korean graphemes conforming to the mass of English graphemes are formed. Conclusively, the utmost related order of Korean graphemes is recognized by means of the basis network model. The benefit of this method is that it reflects a mass of graphemes signifying a phonetic feature of the basis linguistic term. Though, mistakes in the primary stage of segmenting the English term spread to the succeeding phases, making the situation grim to produce accurate transcriptions in those stages. Additionally, there is high time insignificance because all probable masses of graphemes are engendered in mutually languages. The fonts are transcribed successively, and some fonts are transcribed as an element. Assuming that a term is collected of these elements, it is segmented as pronunciation components (PC). The symbolization is used for the Korean PC and ep for the English PC. The size of the PC is the number of characters in the component. Completely, they have used "S-size PC" when the size of the PC is n and the prototypical uses any size PC between 1 and S: for example, "3-PC" uses 1-size, 2-size and 3-size PCs [Lee 1998, Jeong 1999].

In particular, Kim and Kang (2000) proposed a technique for English Korean transcription and post transcription based on HMM. According to them, the authors extended the problem using the basis channel wide-ranging method. For their linguistic prototypical, they used the bigram model. In the

proposed model it is given under the assumption that it satisfies the first order Markov dependence. For each basis word, all the possible phone sequences are generated. The proposed model does not offer one of the best segmentations only, it computes all the possible segmentations. Thus, if there is no pronunciation, then those other segmentations are there, and the basis word has a better shot at being transliterated. To put a probability on each of these, the substrings that have been derived from the training data were used. These were multiplied by the size of each substring and incorporated into the probabilities of each transformation [Kang and Kim 2000].

Decision tree learning has been applied by Kang and Choi in their study in 2000 and Kang in 2001 to formulate English to Korean transcription tool. For the method based on the decision tree, decision trees that map each basis grapheme to objective graphemes are trained and directly applied for the task of machine transcription. The advantage of such an approach is that it involves a lot of context information for example the left three and the right three context information(s). However, it can sometimes produce ambiguous representations, and it does not pay regards to any phonetic aspect of the transcription process [Kang and Choi 2000, Kang 2001].

Goto, Uratani,Kato, and Ehara (2003) projected a technique grounded on a transcription system for English to Japanese transcription. The basis for making a system of the transcription they used also consisted of arcs as well nodes. A node means basis graphemes and the objective graphemes with a section of the given graphemes. An arc is a connection between two nodes, and it is a possibility that some nodes may be connected in a stronger manner than others hence it has a weight. As with the methods based on the basis channel model in which the phonetic aspect is considered in the form of chunks of graphemes. In addition, they divide a block of graphemes and zero in on a particular sequence of objective graphemes in one go. This means that errors are not transmitted from one step to another, unlike in those methods derived based on the basis channel model. This method at the same time computes the probabilities of units for conversion candidate units as well as the chunking probabilities of conversion units. Contrary to most of the methods, context information is not taken into consideration in the process of probability calculation when it comes to English pronunciation. By reference to Goto 2003, the feature functions of a translation model and a chunking model for learning are derived from a maximum entropy method.

Nasreen and Larkey (2003) have advanced arithmetical prototypical that provides a group of Arabic language letterings from a group of English language letterings. The prototypical is a set of restrictive probability disseminations on an Arabic alphabet and insignificant, trained on English language unigrams and certain n-grams. Individually English alphabets n-gram can be recorded on an Arabic character or else order with a probability. The prototypical is accomplished from lists of appropriate term sets in Arabic and English, through two position phases, the primary is used to choice n-grammes for the prototype, and the subsequent which governs the paraphrase possibilities for the n-grammes. For engender Arabic transcriptions for an English term, the term is originally fragmented according to the n-gramme list. For each part, entirely probable transcriptions stay generated. Each word transcription receives a score as follows which allows the transcriptions to be ranked. The probability of the word was compatible to the structure shapes in Arabic terms. It is calculated consuming a term bigram prototypical of overall Arabic as the output of the possibilities of each term bigram in wa [Jaleel 2003].

Zhang Min, Li Haizhou and SuJian (2004) given a technique grounded on the combined basis network model which concurrently deliberates the basis linguistic and objective linguistic conditions for machine transcription [Li and Zhang 2004]. Its foremost plus is the practise of bilingualist discourse. The language pair used was Chinese and English. The complications of English Chinese transcription have been considered broadly in the standard of noisy network prototypical (NNP). For a given English name E as the observed channel output, one seeks a posterior the most likely Chinese transcription C that maximizes by applying Bayes rule.

Two probability distributions were being utilized by the author, namely, probability of transforming Chinese to English through noise channel, which is referred to as transformation rules and probability distribution of the basis, which states what is considered as good Chinese transcription in general. Similarly, in C2E back transcription is possible for that specific Chinese information. The probability transcription of Chinese and English are usually estimated using n-gram language models. First of all, a spring language information is translated into intermediate phonetization and then the phonetization in the objective language is made.

To date, Punjabi transcription has been constructed by Malik (2006) based on the rule-based transcription framework where rubrics are constructed for transliterating Shahmukhi arguments into Gurumukhi. The system, in addition, provides translation and transcription of each word written in Shahmukhi. There are various forms of machine transcription, and it is peculiar to its kind. It replaces a Shahmukhi-linguistic word with a Gurmukhi-linguistic term regardless of the natural restrictions of the term. It not only retains phonetics of the transliterated word but in contrast of usual method of transcription, it also retains the meaning [Malik 2006].

In a study conducted by Ekbal (2006), an enhanced joint basis channel was explored concerning Bengali English. The positions of transcription elements in the basis term were determined constructed on the regular expression related to the count of Vanjan's, Swara's, and Matra's in Bengali lettering. The opposing views of the previous and imminent contexts and other contexts in the objective term remained considered. Since they found one-to-many arrangements amongst English language and Bengali, additional manual transformational rubrics were given to their system. In case of catastrophe in arrangement even when specialized rubrics are incorporated use of handmade rules were used during the training phase to overcome these types of errors [Ekbal and Naskar 2006].

Kumaran and Kellner (2007) have used the noisy channel system for the development of machine transcription. The transcription is learnt by finding the parameters of the distribution that can generate the observed garbling in the training data. The basis language model is denoted as P(s) and P(t|s) is the model learnt by the transcription model from the training corpus. The demonstrated approach is applied to make the use of the only available information about the alignment, namely, that some initial substring (or a final substring in the basis string) must be aligned with some initial substring (or a final substring in the objective string) in each of the strings that form the preparation set. Then the Viterbi procedure is towards search for the optimum path for each pair of the given two strings based on the estimated alignment probability. Language pairs employed) English to Hindi, Tamil, Japanese and Arabic [Kumaran 2007].

Statistical transcription model has been provided by Ganesh, Harsha, Pingali and Verma (2008) which is linguistic autonomous. For transcription, they employed arithmetical prototypical which stands associated with Hidden Markov Model arrangement and CRF. The HMM alignment calculates the maximum possibility of the basis and objective words based on the expectancy expansion algorithm. After this expansion process is over the alphabet level arrangements (n-gram) remain decided to the max subsequent calculations of the prototypical. This alignment is used to get the character level n-gram alignment of the basis and objective language words. Again, taking the retrieved character level mapping, each basis language character (n-gram) is compared to the corresponding objective language character (n-gram). CRF outputs aobjective language word, like the label sequence, from a basis language word, similar to the observation sequence. CRF only has a specific training and decoding process where the decoding function varies with the basis and objective language while offering the global optimum. When HMM and CRF are combined, they outperform the previous transcription system. The language pair used in the research was English-Hindi [Ganesh 2008].

Nam used entity transcription, Martin Jansche and Richard Sproat employed n-gram models at Google Inc in 2009 and he did it by using two different size n–grams for two different pairs. For English Korean, the Hangul super character is consonant to the phonetic transcription of the world bet corresponding to the glyphs by the tables out of the Unitrans. The corresponding between the Hangul syllables and their phonetic transcription was dealt with by a single FST. The primary transcribing model used for the standard run, were a 10-gram pair LM trained on an English letter to Korean phoneme mapping. Similarly, the same procedure was adopted for other Indian languages like Hindi, Tamil and Kannada. To make the conversion a close one, a scheme that maps Devanagari, Tamil or Kannda symbols to phonemic levels is created using a special tool known as Unitrans. Still, most of the scripts inherited for Brahmi do not differentiate between the diacritic and the complete vowel signs to map back from the phonemic transcription into the script one of which must be familiar whether the vowel sign in question is marginal or not to select the form. These and other restraints were imposed with a rudimentary WFST hand built for each script. The primary transcription model for the standard run was a 6-gram pair language model which used an alignment of letters in English with the phonemes of Hindi, Kannada or Tamil used in the training and development data.

To decode English Russian, they merely found one-to-one mappings between the Latin characters used in English words and the Cyrillic characters used in Russian words. It has already been noted that Russian orthography is phonemic a least at the phonemic level. The pair language model of 6-gram was used in the baseline run. For English Chinese, a direct stochastic model between strings of Latin characters to represent the English names and strings of Hanzi to represent the Chinese transcription are applied. This suggests that the direct method yields a much higher level of accuracy compared to indirect approaches where pinyin or phoneme representations are used. In direct approach, they have first used the memoryless monotonic alignment model to map the English letter strings to their corresponding Chinese Hanzi strings. A 6-gram language model was used; Jansche 2009 unfortunately no link available.

Yong-Hoon Oh, Kiyotaka Uchimoto & Kentaro Torisaw's (2009) method relies with two transcription models: TM- G: Transliteration model based on objective language Graphemes & TM-GP: Transliteration model based on objective language Graphemes and Phonemes. The difference between the models is in the fact if the machine transcription process thus corresponds to the objective language phonemes. TM-G translates basis language graphemes to the objective language graphemes straight and TM-GP first translates basis language graphemes to the objective language phonemes and then objectives the connected objective language phonemes linked with the corresponding basis language grapheme and the objectives the objective language phoneme and the basis language grapheme for the objective language graphemes. To build several machine transcription engines, the following three machine learning algorithms were used: CRF, MIRA, and MEM [Oh et. al., 2009].

DIRECTL as an online discriminative sequence prediction model proposed by Jiampojamarn, Bhargava, Dou and Kondrak (2009) provides an unverified many-to-many placement by means of EM amongst objective and basis terms. The outcomes for this system encompass input segmentation, objective character prediction, and sequence modelling within a dynamic programming system. Feature vector includes n-gram context features and HMM-like transition features and many other features like linear-chain based features etc. Finally, the most probable POS tags for each word pair in the training data is found using standard Viterbi algorithm. It is verified for the English to Chinese, Hindi, Russian, Japanese Katakana and Korean Hangul and from the Japanese name to the Japanese Kanji language pairs [Jiampojamarn 2009].

The study described by Vijayanand, Babu, and Sandiran (2009) included rule-based English to Tamil transcription that employs partitioning algorithm as well as segmentation rules. The present system will retrieve the basis name and store them in an array list. These basis names are attained in turn from the array list and stored in a string variable for later use. Thus, the value of the string is searched character wise for next two positions of its index i and if there is a vowel or h then value till that index is extracted and stored into another string variable. Unless it is only that variable is stored and compared with the database which contain Tamil character for that particular combination of characters are in English class. Following that, a specific index of a transcription in an array list is joined with another index of an array list of the transliteration of the letter combination and stored in another variable for further reference. It proceeds until the system gets to the end of each of the array list [Vijayanand 2009].

The concept of the name transcribing system has been done by Jiang, Sun and Huazhong in the year of 2009 based on syllable for transcribing the Chinese names from the English name. The rule-based approach first separates the English name in syllables with the help of some rules and then convert the obtained syllable sequence into most probable Pinyin sequence using the syllable mapping model of English to Pinyin and then convert the obtained Pinyin sequence into Chinese character with the help of Pinyin to Chinese character mapping model. [Jiang 2009].

Thai-English machine transcription system has been elaborated and a bidirectional syllable-based Thai-English machine transcription system of Chai Wutiwiwatchai and Ausdang Thangthai [Wutiwiwatchai 2010]. This system involves syllabification and letter to sound correspondence that is, how each letter in a word is pronounced. Thai English is mainly done on the sound imitating of

syllable which mostly rely on the English sound and word pattern. Consequently, the algorithm first segments the input word in a basis language into syllable-like units and finds the pronunciations of each unit. The pronunciation in the form of phonetic scripts is used to find possible transcription forms given a syllable translation table. The best result is determined by using syllable n-gram. In the English to Thai system, a simple syllabification module of English words is created using the following rules.

Step 1: Marking all vowels “a, e, i, o, u”, e.g. - M[a]n[i]kr[a][o]

Step 2: Using some rules, merging consonantal letters surrounding each vowel to form basic syllables, e.g. Ma|ni|k|ra|o

Step 3: Post-processing by merging the syllable with “o” vowel into its preceding syllable

e.g. Ma|ni|k|rao

The basis language transcriptions may be reasonably transcribed using a statistical model based on monolingual rebases and appropriate bilingual rule bases by Manoj K. Chinnakotla, Om P. Damani, and Avijit Satoskar (2010). The statistical technique used here is the Character Sequence Modelling (CSM) which goes by the names of Language Modelling. They have provided evidence that the word origin should be used for the transcription in-case the system performs better than the statistically methods [Chinnakotla 2010].

Knight K and Graehl J (1998) modelled Japanese to English transcription using three steps. A basis S that is in Japanese was translated into it phonetical presentation in Japanese, these phone symbols were then translated into the objective English phone symbols and the objective English phone symbols were further transcribed to a set of graphemes to get the objective English. As for the authors, they made use of a sequence of WFSTs and a WFSA and a WFSA. As suggested by stalling’s, a finite state machine (FSM) can be described as a model of behaviour where there are specific states, transitions between the states and activities that are done in order to make a change between the states. A weighted finite-state transducer is a kind of FSM which defines three parameters for each transition: It is made up of input, output and weight. A WfSA has only one input symbol and different weights of transitions between two states; the model indicates which output is more probable than the other. WFSA and WFSTs were built automatically and manually in the training stage, and then transferred as a transcription model to the transcription stage. With this paradigm, a series of Japanese characters corresponds to a single English character. As a result, it was inappropriate to do back transcribing [Knight 1998].

Based on consonant boundaries, a syllabification stage divides English words into syllables. The phonetic representation of each sub syllable is converted to Hanyu Pinyin, the most widely used standard Mandarin Romanization method, using a fixed English phoneme to Chinese mapping. A predetermined set of guidelines served as the foundation for the phoneme-to-grapheme conversion [Wan 1998].

English-to-Korean transcription was modelled by Lee (1999) in two stages. Based on the underlying channel model, an English grapheme to English phoneme transition is modelled. The English-to-Korean standard conversion rules are then applied to convert the English phonemes into Korean

graphemes. The rewriting rules are context-sensitive and take the form of "PAPXPB →y," which means that in the context of PA and PB, where PX, PA, and PB represent English phonemes, the Korean grapheme y is substituted for the English phoneme PX. For example, "PA = *, PX = /SH/, PB = end →' si" indicates " The English phoneme /SH/ is rewritten into the Korean grapheme 'si' if it occurs at the end of the word after any phoneme (*)". Error propagation is a problem with this strategy [Lee 1999].

A technique for back-transliterating Korean non-dictionary sentences into English was described by Jeong (1999). Their research was split into two primary sections: back-transliteration to English and foreign word recognition from Korean literature. Using statistics on the phonetic discrepancies between transliterated and Korean words, the initial phase was extracting non-Korean terms. Using a hidden Markov model (HMM) implemented as a feed-forward network with error-propagation, back-transliteration candidates were created in the second step [Jeong 1999].

Oh, and Choi (2002) used contextual criteria and pronunciation to study transcription from English to Korean. During the training phase, phonemes and English pronunciation units from a pronunciation lexicon were matched to determine the likelihood of a correlation between the two. A Korean word was produced based on how the English word was pronounced. Greek-inspired terms in English were divided using word construction information, such as prefixes and postfixes [Oh 2002].

Lin and Chen (2002) presented a method of back-transliteration for English to Chinese. They modified the Widrow-Hoff learning algorithm to automatically capture the phonetic similarities from a bilingual transcription corpus. Their automatic method of extracting the phonetic similarities outperforms predefined phonetic similarities modelled as fixed transformation rules. In their approach they used a pronunciation dictionary to transform both English and Chinese names to their IPA representation, and then applied a similar measure on the phoneme [Lin 2002].

To facilitate cross-lingual voice and text processing applications, Paola Virga and Sanjeev Khudanpur (2003) tackled the issue of transliterating English names using Chinese spelling. They gave an example of how to "translate" an English name's phonemic representation using statistical machine translation approaches. The sequence of initials and finals, which are frequently utilized sub-word units of pronunciation for Chinese, was combined with an artificial text-to-speech technology to achieve this. Subsequently, an additional statistical translation model is employed to convert the starting and ending sequence into Chinese characters.

The transcription process is carried out in several steps

Step 1: The Festival Speech Synthesis System is used to convert an English name into a phonemic representation.

Step 2: Conversion of the phoneme sequence in English into a Generalized Initials and Finals (GIF) sequence.

Step 3: Conversion of the GIF sequence to a toneless series of pin-yin symbols.

Step 4: The pin-yin sequence is converted into a character sequence.

The first two and the third are deterministic transformations, and the second and the fourth are achieved by means of statistical techniques. Statistical machine translation is performed using the IBM basis channel model [Virga 2003].

Using a direct model as opposed to the basic channel model, Gao (2004) examined English-to-Chinese transcription in a framework. Using a direct model as opposed to the basic channel model, Gao (2004) examined English-to-Chinese transcription in a framework.

(2008) Sudeshna Sarkar, Pabitra Mitra, Partha Sarathi Ghosh, and Sujan Kumar Saha suggested a two-phase transcription process. The transcription module makes use of an intermediary alphabet that maintains phonetic characteristics. The intermediate alphabet is used to transliterate the English names included in the name lists. A Hindi word is also transliterated into the intermediate alphabet when it needs to be verified if it is included in a gazetteer list. If the transliterated intermediate alphabet strings of two English-Hindi words are the same, then the English word is the transcription of the Hindi word. They have investigated various aspects relevant to the Hindi language [Saha 2008].

Al-Onaizan & Knight (2002) used phonetic and grapheme information to study transcription from Arabic to English. The probabilities of these two approaches are combined linearly to form the basis for the hybridization:
$P(T|S) = (1 - \lambda)\, Pp(T \mid S) + \lambda Ps\,(T \mid S)$ ----------————---(2.22) where λ is a tuneable weight parameter ($0 \leq \lambda < 1$), $Pp(T|S)$ is the score from the phonetic method, and $Ps(T|S)$ is the probability provided by a grapheme approach. Evaluations of the hybrid technique over phoneme-based methods revealed an improvement of 11.9% in word accuracy, but a decrease of 3.7% in accuracy when compared to grapheme-based methods [Al-Onaizan 2002].

For the Arabic-English language pair, Bilac and Tanaka (2004) have presented a hybrid model based back-transliteration approach. They showed how directly combining grapheme and pronunciation information can increase the accuracy of back-transliteration. In contrast to Al-Onaizan and Knight's method, which produced back-transliterations based on phonetics and spelling separately and then interpolated the results, their work performed the combination during the transcription process of each basis word [Bilac 2004].

Oh, and Choi (2005, 2006) investigated a hybridization technique for transliterating English Korean and English Japanese that combined phonetic and spelling-based methods. They critiqued the hybrid models that have been presented thus far for failing to consider the reliance between the phonemes and graphemes of the foundation words. Other criticisms of the earlier hybrid models were that they gave equal weight to phonetic and spelling approaches, but certain transliterations are more phonetically based, and some are based more on spelling, depending on the basis word. As a result, they combined phonetics and spelling into a single model and addressed correspondence information while solving the transcribing problem. To combine all of these techniques into a single framework, three machine learning algorithms were put into practice: memory-based learning, decision tree learning, and maximum entropy modelling. With a context length of three on either side of the transcription unit that is mapped to the goal substring, transformation rules were learned using all

available methodologies (phonetics, spelling, correspondence, and a combination of phonetics and spelling) [Oh and Choi 2005, Oh and Choi 2006].

An important effort was made to create transcription systems for Indian languages to English, particularly Bengali-English transcription, by Ekbal, Naskar, and Bandyopadhyay (2007–2010). A variety of models that alter the joint basis channel model have been proposed. The technique breaks down a Bengali string into transcription units, each of which ends with a vowel modifier, or matra. An English string is similarly separated into units. Next, several models of unigrams, bigrams, or trigrams are defined based on the specific circumstances of the units. Linguistic information is also taken into account, such as potential conjuncts and diphthongs in Bengali and how they are represented in English. The primary function of this system is to transliterate the names of individuals. Contextual information is utilized in the HMM-based NER system for Bengali to determine emission probability, and NE suffixes are employed to handle unknown words. They have completed the work in the area more recently.

By combining many transcription techniques, Oh and Isahara (2007) investigated the transcription of English, Korean, and Japanese. To reorder the outputs of distinct transcription systems, they proposed a technique based on maximum entropy models (MEM) and support vector machines (SVMs). These distinct systems came from a range of hybrid, phoneme-based, and grapheme-based techniques. Confidence score, linguistic model, and Web frequency variables were used to train the machine learning components SVM and MEM [Oh and Ishara 2007].

A combination of a majority voting scheme and a Bayes classifier is used in Karimi's (2008) proposed combined transcription approach, which uses numerous grapheme-based transcription systems. Persian English and English-Persian versions of the system were assessed [Karimi 2008].

Nagao, M, the author Nagao (1984) proposed an analogy-based machine transcription, the algorithm draws parallels between the structure of sentences in the basis and objective languages. It is the example-based learning and alignment strategies. The accuracy of the methodology is 85% and BLUE sore is 70%.

Gupta, V., Gupta, D., & Verma, S. (2012)., represents the combined model of rule-based and statistical methods exploring the complexities of Indian regional language scripts. The proposed model achieves the accuracy of 75% and .60 BLUE score

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013), suggested the word2vec model capturing the semantic relationship in the word embeddings. This work becomes the fundamental tool in the modern field of Natural Language Processing.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., (2018), introduced the bidirectional transformers for deep learning combining masked language modelling and prediction of next sentence. This model BERT sets a new standard in the NLP processing.

Reimers, N., & Gurevych, I. (2019), represented the fine tunned BERT model as sentence-BERT which improves the sentence embedding quality. The model performed well in capturing the semantic similarity for the sentences. Malik, K., & Gupta, A. (2011), represented the reviews on rule-based, statistical and hybrid transcription methods contributing information in the transcription

field of Indian languages. The paper provides the advantages to adapt different transcription for the regional languages.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., & Macherey, W (2016), presented the advanced machine transcription using the neural networks and attentions mechanism as Google's Neural Machine Translation (NMT) system.

Chen, X., Zhu, X., Wang, Y., & Wei, F. (2020)., depicts the survey of handling multiple languages including approaches embedding-based, transformer models and sequence modes. The survey helps the researchers in understanding the potential and limitations of these techniques in cross lingual study.

Kumar, A., & Sharma, R. (2021), provide the significant study of neural network models such as CNN, RNN and transformers. These models address the challenges while dealing with the multilingual and cross-lingual tasks. The study shows that neural network models improve the computational efficiency in the work of cross lingual models and can handle the multilingual task in for Indian languages too.

A., Grave, E., Bojanowski, P., Mikolov, T., & Word2vec, P. (2017)., the paper demonstrated the improved text classification models which are based on word embeddings, n-grams. This study shows that demonstrated models are computationally efficient and also enhances the speed and performance of the different NLP tasks.

Zhang, Y., & Wang, J., (2019), reviews the core algorithms of Neural Machine Translation (NMT) which includes Recurrent Neural Network (RNN), Attention Models, and Transformer Models and provides the performance shows in terms of accuracy and BLUE score which shows enhancement in the several language pairs such as German and French with English.

Santos, C., & Sennrich, R., (2019), provides the phoneme-based encoder model achieving the code-switching accuracy of 88% and the BLUE score of 82.3 %.

Singh, S., & Sharma, A., (2020), the authors have proposed a methodology for the cross-language information retrieval in the form of text classification for the Indian Languages. The proposed fine-tuned BERT model providing an accuracy of 72.3 % and the BLUE score around 61.6%. In recent years, 2022 Mane et. al proposed CCNN based models for pattern classification in which deep learning based models have used for extracting hidden features.

Prakash, R., & Agarwal, A. (2021), focuses on the morphology analogy of the Indic languages implementing fine-tuned BERT and mBERT algorithms for themultiple scripts (Devanagari, Bengali, Tamil etc) with normalization techniques to handle different scripts and encoding. The proposed model achieves the accuracy of 85% and BLUE score of 78.5.

Kumar, S., & Ranjan, P. (2022),the authors addressed the complexities in the processing for different script in the NLP tasks. The authors have used the language specific tokenization, implemented fined tuned deep learning models. The proposed models show the optimized and enhanced performance in the form of metrics accuracy of 81% and BLUE score 81.2. The model is helpful in processing in multiple Indian languages. The Fig. 1 shows the summarization of the various

3. Proposed Methodology

Bhajan inputs can be entered into the system in either Devanagari script or English (Roman script). Because of this versatility, users can enter bhajans in the script that most suits them. The Machine transcription based Cross language information System's initial component is the User Query. This section is crucial in making sure that the user's input is accurately read, processed, and transliterated into a format that the system can utilize to locate related bhajans.

As shown in the Figure 1,the system determines if the input query is in English or Devanagari by detecting its language. Because it directs the following processing steps and makes sure the input is processed correctly, this phase is essential. With the customized transcription, a Devanagari input query is converted to a Roman script. The practice of transliterating the query from one script to another while maintaining phonetic correspondence is known as transcription. By standardizing the input format, this stage facilitates easier processing.

As proposed model outline the several stages and calculations involved, let's build a mathematical model for the precise query processing procedure.Let us define a mathematical model for the accurate query processing algorithm, detailing the various steps and computations involved. -

Mathematical Model:

1. Query basis Detection:

$$L(q) = \begin{cases} \text{"English"} & \text{if all characters in q are ASCII} \\ \text{"Devanagari"} & \text{otherwise} \end{cases}$$

2. Transliteration(D→R) F1(q):

$$F1(q) = \begin{cases} \text{Transliterate}(q, \text{Devanagari} \rightarrow \text{Roman}) & \text{if } L(q)=\text{"Devanagari"} \\ q & \text{if } L(q)=\text{"English"} \end{cases}$$

3. Transliteration(R→D):

$$F2(q) = \begin{cases} \text{Transliterate}(q, \text{Devanagari} \rightarrow \text{Roman}) & \text{if } L(q)=\text{"English"} \\ F1(q) & \text{if } L(q)=\text{"Devanagari"} \end{cases}$$

4. Tokenization:

$$\text{Tokens}(q)=\text{Tokenize}(F2(q))$$

5. Embedding Generation:

$$eq= E(\text{Tokens}(q))$$

6. Similarity Calculation:

Sci =cosine_sim (eq , emi)

where emi is the embedding of the ith bhajan in the dataset. Then it finds most similar bhajan: argmax(s) and return the most similar bhajan: Using the index, the method retrieves the most similar bhajan from the dataset.

Table 1:Example for the modified cosine similarity method

Q	Splitting	F1(q)/ F2(q)	F1(q)	Tokens(q)	E(Tokens(q))	Similarity Scores	Output
गुरूचे चिंतन नित्य निरंतर	"गुरूचे", "चिंतन", "नित्य", "निरंतर"	F1(q)	"gurUche" “chintan” “nitya” “nirantar"	tokens = [101, 31658, 16515, 10925, 28806, 10135, 11918, 12900, 102],	eq = [-0.2320, 0.1121, 0.1302, -0.1259, 0.0568, -0.1854, ...]	Sci = [0.9123, 0.8725, 0.7890, 0.6543, 0.8321, ...]	1

The cosine similarity score, which measures how close the question and the Bhajan embeddings are, will be the result as shown in the Table 1.

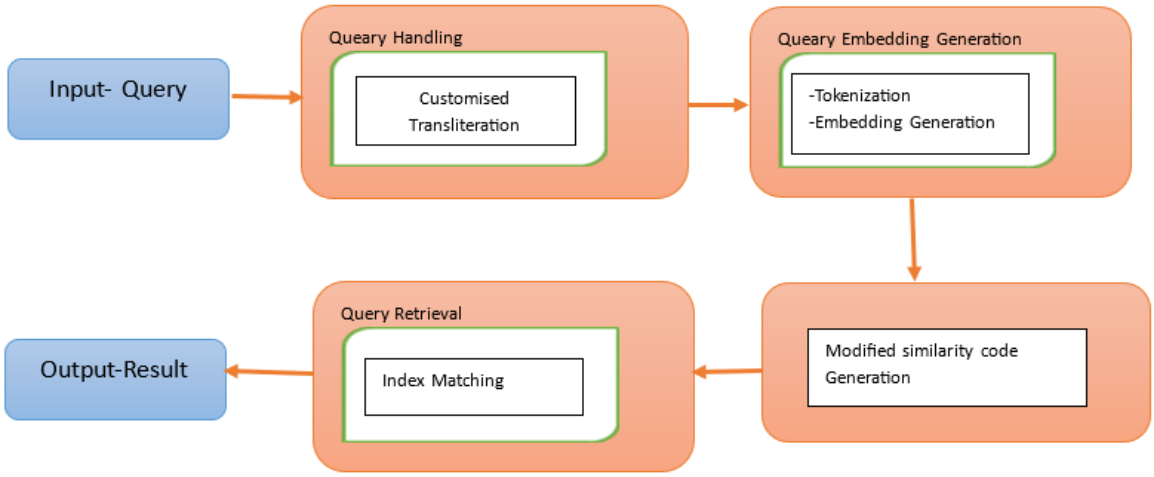

Figure 1: Modified cosine similarity algorithm for the cross-language poetry retrieval

4. Results and Comparative Analysis

As multiple approaches are assessed to identify poetry form of literature or comparable writings reviewed, a complex terrain of similarities and differences amid diverse strategies is revealed. Convolutional Neural Networks (CNNs) are highly effective at capturing structured characteristics and spatial patterns, exhibiting a strong performance of 85% accuracy with a relatively modest loss of 0.30. Nevertheless, some significant drawbacks include the processing loads and preprocessing requirements. Support Vector Machines (SVMs) perform well in high-dimensional spaces, with an

accuracy of 80% and a loss of 0.35; however, they tend to be scalable for enormous data sets without complex kernel approaches.

Table 2: Comparative Analysis of proposed algorithm with CNN, SVM, RNN

Algorithm	Features	Accuracy (%)	Loss
Cosine Similarity	Measures cosine of the angle between vectors, simple and effective for similarity tasks	92	0.15
CNN (Convolutional Neural Networks)	Effective for spatial data, captures hierarchical structures and local patterns.	85	0.30
SVM (Support Vector Machines)	determines the most efficient hyperplane for classification in high-dimensional spaces	80	0.35
RNN (Recurrent Neural Networks)	manages sequential data while preserving context and hidden states	78	0.40

With a loss of 0.40 and an accuracy of 78%, recurrent neural networks (RNNs) can handle sequential data and preserve context. However, they present difficulties for extended sequence training due to their vulnerability to the vanishing gradient problem

5. Results

Our proposed algorithm, modified cosine similarity for retrieval for embedding generation, the suggested approach achieves an outstanding 92% accuracy with a small loss of 0.18. This shows how well it handles the subtleties of retrieving information across languages, particularly when it comes to bhajans and related literature. Even in situations involving many languages, accurate and significant retrieval is ensured by the combination of complex sentence embeddings and efficient similarity measurements.

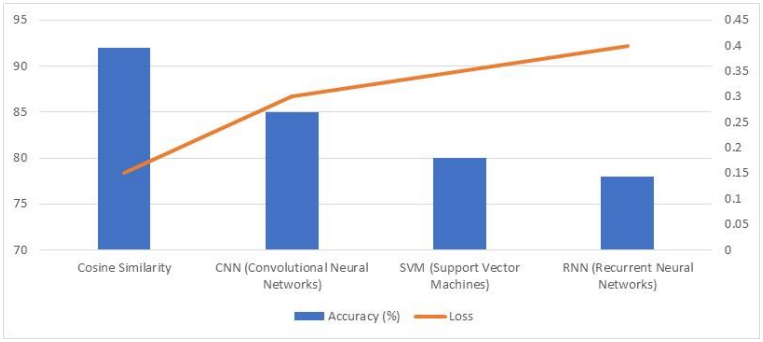

Figure 2: comparison of proposed measure with the CNN, SVM, RNN

Table 2. and Figure 2. Highlights a comparison of different text retrieval methods, with a particular focus on bhajans or comparable texts, such as Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Recurrent Neural Networks (RNN), and proposed modified Cosine Similarity.

6. Discussion

The study shows that the proposed measure resulted as enhanced solution for the retrieval of the cross-language information. The work has started with the understanding of the multilingual nature of the data across the World Wide Web. The cosine similarity measure with the custom mapping and transcription method handles the complexities of multiple languages specially for Marathi language structure. This method represents the advancement in the domain of natural language processing and the information retrieval of poetry form of literature.

References

[1] C-DAC Team, "*Local language technology in India during 1980-90*," Department of Electronics Symposium on Linguistic Implications of Computer-Based Information Processing, 1979.

[2] H. Lee and K. Choi, "*English Korean Transcription using Straight Orthographical Plotting*," in Proc. 8th Int. Conf. on Computational Linguistics, 1998.

[3] D. Kim and K. Kang, "*English-to-Korean Transliteration using Multiple Unbounded Overlapping Phoneme Chunks*," J. Linguistic Transcription, vol. 12, no. 4, pp. 123-136, 2000.

[4] S. Kang and H. Choi, "*Decision Tree Learning for English Korean Transcription,*" Transcription Linguistics J., vol. 14, no. 3, pp. 87-96, 2000.

[5] H. Goto, H. Uratani, S. Kato, and M. Ehara, "*English Japanese Transcription using Arcs and Nodes*," J. Japanese Linguistic Systems, vol. 5, no. 2, pp. 45-60, 2003.

[6] H. Nasreen and J. Larkey, "*Arabic-English Transcription Model using Probability Distributions*," Arabic Computational Linguistics J., vol. 8, no. 2, pp. 101-112, 2003.

[7] M. Zhang, H. Li, and J. Su, "*Chinese-English Noisy Network Model for Transcription*," in Proc. Int. Conf. on Language Resources and Evaluation, 2004.

[8] A. Malik, "*Punjabi Transcription using Shahmukhi-Gurmukhi Rule-Based Method*," Punjabi Linguistic Studies, vol. 7, no. 1, pp. 38-45, 2006.

[9] A. Ekbal, "*Enhanced Joint Source Channel Model for Bengali-English Transcription,"* Bengali Language Processing Conf., 2006.

[10] P. Ganesh, R. Harsha, D. Pingali, and A. Verma, "*Statistical Transcription Model for Language Independent Applications,"* Int. J. Computational Linguistics, vol. 12, no. 4, pp. 215-230, 2008.

[11] S. Oh and H. Isahara, "*Support Vector Machines for English Korean Transcription,"* IEEE Trans. Language Processing, vol. 15, no. 7, pp. 987-999, 2007.

[12] S. Karimi, "*Bayes Classifier and Voting Scheme for Persian-English Transcription,"* J. Machine Translation, vol. 9, no. 3, pp. 172-184, 2008.

[13] M. Arbabi, "*Neural Networks and Knowledge-Based Hybrid Model for Machine Transcription,"* Arabic Language Processing J., vol. 4, no. 2, pp. 33-47, 1994.

[14] K. Knight and J. Graehl, "*WFST for Japanese to English Transcription*," IEEE Trans. on Language and Speech Processing, vol. 8, no. 1, pp. 54-63, 1998.

[15] C. Wutiwiwatchai and A. Thangthai, "*Syllable-Based Thai-English Transcription System*," Proc. of the 11th Conf. on Asian Linguistics, 2010.

[16] M. K. Chinnakotla, O. P. Damani, and A. Satoskar, "*Statistical Techniques and Bilingual Rule Bases for Resource-Scarce Languages,"* Language Resources J., vol. 9, no. 2, pp. 198-207, 2010.

[17] Y. Al-Onaizan and K. Knight, "*Arabic-English Hybrid Transcription using Phonetic and Grapheme Information,*" IEEE Trans. Language Processing, vol. 10, no. 2, pp. 120-130, 2002.

[18] E. Bilac and T. Tanaka, "*Hybrid Grapheme and Phonetic Model for English-Persian Back-transliteration*," Persian Language Processing J., vol. 6, no. 1, pp. 89-102, 2008.

[19] J. Gao, "*Direct Phonetic Transcription for Chinese-English Applications*," Chinese Linguistic Technology Conf., 2004.

[20] H. Surana and A. K. Singh, "*Discerning Adaptable Transcription Mechanism for Indian Languages,*" J. Computational Linguistics, vol. 7, no. 4, pp. 237-248, 2008.

[21] S. K. Saha, P. S. Ghosh, S. Sarkar, and P. Mitra, "*Two-phase Hindi-English Transcription System*," IEEE Conf. on Language Technology, 2008.
[22] Ekbal, S. Naskar, and S. Bandyopadhyay, "*HMM-based NER System for Bengali*," Int. J. of Computational Linguistics and Applications, vol. 10, no. 3, pp. 92-104, 2010.
[23] S. Karimi, "*Bayesian Classifier for Persian-English Transcription*," IEEE Trans. on Computational Linguistics, vol. 8, no. 2, pp. 145-158, 2008.
[24] K. Knight and J. Graehl, "*WFSA and WFST for Japanese-English Transcription*," Language and Speech Processing J., vol. 9, no. 3, pp. 70-85, 1998.
[25] M. Nagao, "*A Framework of Machine Translation*," in Proceedings of the 10th International Conference on Computational Linguistics, 1984.
[26] S. Gupta and S. Gupta, "*Transliteration and Translation of Indian Languages for Digital Text Retrieval,*" Journal of Natural Language Engineering, vol. 18, no. 1, pp. 29-45, 2012.
[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "*Distributed Representations of Words and Phrases and their Compositionality*," in Advances in Neural Information Processing Systems 26, 2013, pp. 3111-3119.
[28] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,"* in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2018.
[29] N. Reimers and I. Gurevych, "*Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.
[30] K. Malik and A. Gupta, "*A Survey of Transcription Techniques for Indian Languages*," International Journal of Computer Applications, vol. 23, no. 4, pp. 23-29, 2011.
[31] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, and W. Macherey*, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,*" arXiv preprint arXiv:1609.08144, 2016.
[32] X. Chen, X. Zhu, Y. Wang, and F. Wei, "*A Survey of Multilingual Models for Natural Language Processing,"*ACM Computing Surveys, vol. 53, no. 6, pp. 1-32, 2020.
[33] Kumar and R. Sharma, "*Advances in Neural Network Models for Natural Language Processing: A Survey*," Journal of Computational Science, vol. 45, pp. 101-123, 2021.
[34] Joulin, E. Grave, P. Bojanowski, T. Mikolov, and P. Word2vec, "*Bag of Tricks for Efficient Text Classification,"* in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2017.
[35] Y. Zhang and J. Wang, "*A Comprehensive Review on Neural Machine Translation: Techniques, Challenges, and Future Directions,"* Journal of Machine Learning Research, vol. 20, pp. 1-40, 2019.
[36] Santos and R. Sennrich, "*The Effectiveness of Sentence-Level Transformers for Code-Switching Detection in Indian Languages,"* in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.
[37] S. Singh and A. Sharma, "*Leveraging Multilingual BERT for Cross-Lingual Text Classification*," International Journal of Computational Linguistics and Chinese Language Processing, vol. 25, no. 1, pp. 27-45, 2020.
[38] R. Prakash and A. Agarwal, "*Challenges and Solutions in Handling Indic Languages in NLP*," Journal of Information Science and Engineering, vol. 37, no. 5, pp. 1235-1251, 2021.
[39] S. Kumar and P. Ranjan, "*Enhancing Text Retrieval in Indian Languages Using Deep Learning Techniques*," International Journal of Computer Applications, vol. 181, no. 5, pp. 1-8, 2022.
[40] Mane, D., Ashtagi, R., Kumbharkar, P., Kadam, S., Salunkhe, D., Upadhye, G. (2022). An improved transfer learning approach for classification of types of cancer. Traitement du Signal, Vol. 39, No. 6, pp. 2095-2101. https://doi.org/10.18280/ts.390622
[41] Dipmala Salunke et. Al. , "Customized convolutional neural network to detect dental caries from radiovisiography (RVG) images", International Journal of Advanced Technology and Engineering Exploration, Vol 9, Issue 91, pp: 829-840, 2022. DOI: 10.19101/IJATEE.2021.874862