Annals of Library and Information Studies
Vol. 64, March 2017, pp. 37-43

Hindi language text search: a literature review

Pratibha Singh[a] and Aditya Tripathi[b]

[a] Ph.D student, [b]Associate Professor & Head of the Department, Department of Library and Information Science, Banaras Hindu University, Varanasi- 221005, E-mail:pratibhasinghbhu@gmail.com;aditya@bhu.ac.in

Received: 03 August 2016; revised: 11 December 2016; accepted: 06 February 2017

The literature review focuses on the major problems of Hindi text searching over the web. The review reveals the availability of a number of techniques and search engines that have been developed to facilitate Hindi text searching. Among many problems, a dominant one is when a text formed by combinatorial characters or words is searched.

Keywords: Conjunction; Combinatorial characters; Corpora; Lexicon; UNICODE

Introduction

Hindi, the national language of India is widely spoken in the country and is the most preferred language after English. Hindi not only has one of the richest vocabularies but it has an equally rich script, grammar, word collection, and parts of speech. Hindi originated from Sanskrit and the most spoken form of Hindi is "Khari Boli". The influence of other languages over Hindi can be seen clearly in its present form as it has accepted words from other languages like Urdu, Arabic, Prakrit, Farsi and English[1].

Several search engines such as Google, Yahoo, Raftaar, Hinkhoj, Khoj, and Guruji facilitate Hindi searching to satisfy the information needs of Hindi users[2].

A word can be considered as one of the basic building blocks for a language. Words carry meaning and thus information content of the text is mainly based on the words used. Hindi has a wide variety of forms and dialects. It also has different grammatical rules for the formation of words, but Hindi lacks the rules for formation of masculine and feminine genders[3]. It has various combinatorial characters as well as combinational words, which makes it difficult to account these characters and words into a computational system.

The study aims to present literature review on the developments in Hindi search on the Internet. Unlike English, Internet technologies for Hindi and other Indian languages are still in the evolutionary stage. This study is an attempt to highlight the evolving technologies in Hindi language computing.

Objectives of the study

- To present an overview of works in the field of Indian languages;
- To list the major projects and technologies which are intended to promote Hindi language;
- To identify the hindrances to the performance of Cross Language Information Retrieval (CLIR) and Multilingual Information Retrieval (MLIR) systems; and
- To identify the major issues regarding Hindi language searching.

Methodology

This study adopts a systematic literature review. It differs from traditional narrative reviews by being more systematic and explicit in the selection of the studies and by employing rigorous and reproducible methods of evaluation. The databases used for review

of literature are ERIC, Google scholar, and Web of Science. The keywords used for search are 'Hindi language searching', 'developing Hindi language search 'and 'Hindi text search'. The articles retrieved were reviewed and filtered. Total of 33 articles, organizational reports and websites were studied to identify common themes and key issues. In most cases, the articles cited are based on the project reports of the different organizations, institutions and groups. Different Hindi grammar books, dictionaries and internet resources have been followed for the collection of terms to develop corpora. Institutional websites were also reviewed to know about the current trends and types of research going on in various institutes on the present topic or related field.

Literature review

Most challenging task of a research study is to review the past literature to extract a functional methodology or working model for a research problem. Meredith[4] defined literature review as a summary of the existing literature by finding research focus, trends and issues. Fink[5] further modified the definition and defined literature as a "systematic, explicit and reproducible design for identifying, evaluating and interpreting the existing body of recorded documents". Brewerton and Millwards[6] defined a literature review as content analysis, where qualitative and quantitative techniques are used to find the structural and content criteria.

Literature review is an essential approach to "conceptualize research areas and survey and synthesize prior research"[7]. According to Kenneth Lyons[8], not to be confused with a book review, a literature review surveys scholarly articles, books and other sources (e.g. dissertations, conference proceedings) relevant to a particular issue, area of research, or theory providing a description, summary and critical evaluation of each work. The purpose of a literature review is to offer an overview of significant literature published on a topic. Systematic review is an efficient technique for hypothesis testing, for summarizing the results of existing studies and for assessing the consistency among previous studies. A systematic review should be based on some standards. There are four core principles suggested by experts for a literature review which are replicable, exclusive, aggregative and algorithmic[9].

In this study, literature relating to Hindi language text searching was collected, scanned and reviewed.

Language research in India

Language research in India dates back to 1970s and is mostly related to language translation. Machine translation activity in India is relatively young. The earliest efforts dated from mid 80's and early 90's. There are different Indian institutions that have played important roles in the development of machine translation. These include the Indian Institute of Technology-Kanpur, Computer & Information Science Department-University of Hyderabad, National Center for Software Technology (NCST)-Mumbai, Centre for Digital and Advanced Computing (CDAC)-Mumbai & Pune and Department of Official Languages; Ministry of Home Affairs; Government of India.

However, research on Indian Language Processing (ILP) tools, Indian Language Resources: Corpora, Lexical Resources and Dictionaries, and Web based search tools are of utmost importance if Indian languages have to make a mark on the internet.

Indian language processing tools

For writing a language over the Web, a writing platform is required including a standard for character representation, fonts, glyphs and a text editor. There are number of standards for character representation like American Standard Code for Information Interchange (ASCII), Indian Standard Code for Information Interchange (ISCII) and UNICODE which have most developed code table for scripts of different languages. These standards have a set of characters (upper and lower case alphabets), symbols, control codes etc. UNICODE is the most developed global standard for character representation covering almost all the scripts of all languages around the world. Indic script table of UNICODE is built on ISCII and UNICODE has a unique value for each character. Numbers of text editors are present in the present time like WordPad, Notepad, Microsoft office suit and Open office suit. All these tools are available with regular features like text editing, formatting and decoration but they are unable to count and sort the characters of the Indic script. UNICODE also lacks the sort order of characters for Indian languages. Counting of characters in case of combination

characters and hidden characters is also difficult for word processors in Hindi language[10].

Indian language resources

A good quality corpus is the basis for computational linguistics. Processing of a language is done over the corpora to understand the possible outcomes. A dictionary is required to understand a word or sentence, where terms are arranged with their parts of speech, gender, and form etc. for example English WorldNet. Lack of optical recognition software is also a reason behind the lack of good corpora. Chitrankan, a software developed by Centre for Digital and Advanced Computing (C-DAC), Pune promises to convert text into editable form through scanning, but the software has many issues and still has a long way to go[11].

Hindi also lacks rules regarding formation of words, for example there is no rule which identifies if a term is feminine or masculine or it is singular or plural. In some cases feminine and masculine are decided on the basis of ending alphabets of the terms. For example feminine is identified on the basis of vowels at the ending of terms for example मकई, लकड़ी etc. but गाय, किताब are also feminine hence, it is very difficult to identify the gender form of the term. The other major issue with the Hindi language search engine is word sense disambiguation. In Hindi, a single term have different meaning according to the context for example 'कलम' means कलम (pen) and तना (stem), 'लाल' meant for लाल रंग and it also meant for बेटा. It is very difficult to identify the context of the user's query. To resolve the word sense ambiguity a search system developed should have the ability to measure the level of ambiguity in the query and having the feature of Word Sense Disambiguity (WSD)[12].

Hindi also has number of terms, which are derived from the combination of words, and such words are broken into components during processing of text and after that search should be performed. For combinational words Hindi has well defined rules known as conjunctions (sandhi). Conjunctions are the formation of a new term by the combination of sounds of two terms next to each other. For example

विद्यालय = विद्या + आलय
देवेंद्र = देव + इंद्र

Translation lexicons plays vital role in the field of information retrieval. A good quality of translation lexicon determines the retrieval performance of information system. Sometimes translation lexicons lack good coverage of proper nouns and it turns out that names appear often in queries and constitute the largest class of out-of-vocabularies terms in cross language information system. Regular updating of lexicons is also a tough job and this will affect the performance of Hindi searching over the web[13.]

Web based search tools

The major aspects of internet are searching and search engines. In the present time, searching is not limited to locate information but it has become a big market for advertisement and business. At an early stage of its development, search engines (Veronica, Archie and Jug Head) only provided dictionary based search, but with the development of ICT and its impact over different areas of information, role of search engines have changed. A number of search engines like Google, Yahoo, and Bing are emerging as players of the present time. About 72.3% users of internet are non-English and this population is the target of search engine providers.

Search engine optimization technique, ranking and clustering of web pages are the strategies adopted to find out the possible market. To provide multilingual search facility is a big challenge for search engine providers who are targeting this multilingual user population. Considering Hindi alone, this language has several dialects and variations. It also has a loose grammar and phoneme methods. Hence, it is very difficult to mechanize Hindi searching. Like English, SoundEx technique has also evolved for Hindi language for searching homonymous as well as homophone terms. A number of search engines like Khoj.com, Raftaar.in, Guruji, Hinkhoj are specifically developed for Hindi searching are indexed with Hindi webpage. Google, Yahoo and Bing also support Hindi searching, but they are restricted to regular expressions. These search engines also facilitate combinatorial searching with Boolean operators. Lucene, an indexer which also provides Hindi searching by implementing UNICODE character search, is also limited to regular expressions only[14].

Stemming a technique of extracting the root word from a given word is another approach for Hindi

searching. Hindi language is influenced by many other languages like Urdu, Farsi, Sanskrit, English, Awadhi, Prakrit and Arabic etc. These languages have different rules for word formation hence it is very difficult to form a rule for word formation through Hindi grammar during formation of sentence. Stemming algorithm also fails in formulating methods for mechanization of Hindi searching. Searching a query in natural language is not as simple as English searching. Hindi language searching is very much different from English language searching due to lack of rules. To understand a sentence semantically by a machine is still a big challenge. Even in English language searching, it requires a deep level syntactic analysis, and in case of Hindi searching lots of experiments and exercises still remain[15]. Factors, which affect the searching of Hindi information on the web are given below[16].

- Morphology of Hindi language
- Phonetic nature of Hindi language
- Words synonyms
- Ambiguous words

In a large multi-lingual society like India, there is a great demand for translation of documents from one language to another. Most of the state governments work in the respective regional languages whereas the Union Government's official documents and reports are in bilingual form (Hindi/English). In order to have a proper communication there is a need to translate these documents and reports in the respective regional languages. A machine assisted translation system or a translator's workstation would increase the efficiency of the human translators. The market is largest for translation from English into Indian languages, primarily Hindi. Hence, it is no surprise that a majority of the Indian machine translation (MT) systems have been developed for English-Hindi translation. Machine translation activities in India are relatively young. The earliest efforts date from the mid 80s and early 90s[17].

A fully automatic machine translation system should have different modules such as morphological analyzer, part of speech tagger, chunker, named entity recognizer, word sense disambiguator, syntactic transfer module and target word generator. The different techniques used for translation differs in the number of modules used and also the way these modules are implemented. Both rule based and statistical approaches have been tried in the implementation of each of these modules. The machine translation systems developed for Indian languages have been discussed by Nair and Peter[18].

Mantra, a machine assisted translation tool, which translates government orders, notifications, circulars and legal documents from English to Hindi, was developed by Government of India in 1996. It is a domain specific tool based on Lexicalized Tree Adjoining Grammar (LTAG) formalism to represent English as well as Hindi grammar. The main aim of the tool is to provide translation tools to government agencies[19].

In 1995 Prof. R.M.K. Sinha developed ANUBHARTI at IIT Kanpur which is a machine translation tool based on hybridized example-based approach and was started in 2004[20.]

Prasad Pingali and Vasudeva Verma Language Technologies Research Center (LTRC) IIT, Hyderabad presented the experiments of LTRC in Cross-Language Evaluation Forum (CLEF) 2006. They focused on Hindi, Telugu and Afaan Oromo as query languages for retrieval from English document collection and contributed to Hindi and Telugu to English Cross Lingual Information Retrieval (CLIR) system with the experiment at CLEF[21].

In Cross Language Evaluation Forum (CLEF) 2007, Hindi to English and Marathi to English Cross Lingual Information Retrieval System were presented by Chinnakotla, Ranadive, Bhattacharyya and Damani of Department of CSE IIT, Bombay. They took a query translation based approach using bilingual dictionaries.The query term which is not found in the dictionary is transliterated using simple rule based approach which utilizes the corpus to return the ‘k’ closest English transliteration of the given Hindi or Marathi term. The resulting transliteration choices for each query term are disambiguated using an iterative page-rank style algorithm, which is based on term-term co-occurrence statistics, to produce the final translated query[22].

In CLEF 1, 2007, a cross-language retrieval system for English documents in response to queries in Hindi and Bengali was presented by Mandal, Dandapat, Banerjee, Gupta and Sarkar from Department of Computer Science and Engineering, IIT Kharagpur. In the presented system they followed the dictionary

based machine translation approach to generate the equivalent English query out of Hindi and Bengali[23].

Om – a transliteration scheme was developed by Ganapathyraju, Balakrishnan, Balakrishnan and Reddy of Language Technologies Institute, Carnegie Mellon University (Pittsburg) in association with Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore. This scheme uses ASCII characters and has an ability to exploit the phonetic nature of alphabets to represent the Indian language alphabets. Om is beneficial for those users who cannot read other Indian scripts except their mother tongue. A text editor was also developed in the research for Indian languages which integrates the Om input into a word processor such as Microsoft WinWord for many other Indian languages[24].

Om has more developed features than other transliteration tools which are designed and developed before Om. Some of key features of Om transliteration tool are as follows[25].

- Om scheme is common to all Indian languages; the display of the text can be converted between the supported languages by choosing it on the menu.
- Easy support for new languages
- Key in the input as we speak
- Uses lowercase English alphabets and some special characters
- Switches between the languages at the click of a mouse
- Saves the output in ASCII and Unicode format
- Exchange email in Indian languages.

Bhattacharyya of IIT Mumbai developed a machine translation system based on Universal Networking Language (UNL) which is a United Nations project for developing the interlingua for languages of world. UNL is used for translating English language to Hindi[26].

Das, Seetha, Kumar and Rana have investigated the impact of query expansion using Hindi WordNet in the context of English–Hindi CLIR system. WordNet is a lexical database, machine readable thesaurus of Hindi language. They have translated English query using Shabdanjali dictionary. The translated queries have been expanded using Hindi WordNet and nine query expansion strategies have been formulated. Run title field of topics were used for query formulation and expansion, in one run title + description field was use for formulation and expansion. The queries were translated then expanded and submitted to the retrieval system to retrieve documents from the Fire Hindi Test collection. Their observations suggest that simple query expansion using Hindi[27].

Anglabharti is a machine translation technology developed by IIT Kanpur under the leadership of Prof. R.M.K. Sinha. It is a rule based system which translates in Indian languages from English and has approximately 1750 rules, 54000 lexical words divided into 46 to 58 paradigms. It uses pseudo Interlingua named as Pseudo Lingua for Indian Language (PLIL) as an intermediate language[28].

Anusaaraka is fully-automatic general purpose high quality machine translation system which can translate any Indian language into another Indian language based on Panini Ashtadhyayi (grammar rules). It a Natural Language Processing (NLP) Research and Development project for Indian languages and English undertaken by (Chinmaya International Foundation). It was developed by the International Institute of Information Technology, IIIT Hyderabad and Department of Sanskrit Studies, University of Hyderabad[29].

C-DAC Kolkata developed a project named speech-to-speech MAT (Machine Aided Translation) based dialogue system from Hindi to Indian languages [Hindi-Bangla, Hindi-Punjabi, and Hindi-Malayalam] for education and tourism. Universal Speech Translation Advanced Research (USTAR) Consortium has developed a speech translation app which is speech to speech based and helps multiple users to communicate in different languages in real time either face to face or remotely. Application also contributes to overcome the barriers of languages as well as communication modalities. It helps users to communicate with the visually-impaired via spoken word or with the hearing impaired via text input. The application covers a total of 31 Asian and world languages.

A Multilingual Computer Lexicon has been also developed by C-DAC to provide and explain the terms and terminologies in the native language of those people who are not comfortable with English[30].

Table 1—Language Translation Systems at a glance

Year of development	Systems	Organization/ developers	Translation
1991	Anglabharti	IIT Kanpur, Prof. R.M.K. Sinha	English to Indian languages
1995	Anusaarka	IIT Kanpur, Prof. Rajeev Sangal	Indian language to other Indian language
1999	The Mantra (Machine assisted Translation tool)	CDAC, Mumbai	English text in Hindi specified domain
2002	English-Hindi translation system	Lata Gore	English to Hindi weather narration domain
2002	Vassaanubada	Kommaluri Vijayanand	Bilingual Bengali-Assamese
2003	UNL based English-Hindi translation system	IIT Bombay, Pushpak Bhattacharya	English to Hindi
2004	ANGLABHARTI II	IIT Kanpur	English to Indian Language
2004	The Matra system	(KBCS) (NCST), (CDAC) Mumbai	English to Hindi
2004	ANUBHARTI	IIT Kanpur, Prof. R.M.K. Sinha	Hindi to Indian language
2004	Shiva & Shakti	IIIT Hyderabad, IIS-B and Carnegie Mellon University	English to Hindi
2004	ANUBAAD	Jadavpur University, Kolkata, Dr. Sivaji Badhopadhyaya	English to Bengali
2004	Hinglish	R. Mahesh, K. Sinha and Anil Thakur	Hindi to English
2006	IBM English-Hindi	IBM India Research LAB	English to Hindi
2007	Punjabi to Hindi	Gurpreet Singh Josan, Punjabi University, Patiala	Word to word Punjabi-Hindi
2009	Sampark	Consortium of Institutions: IITs, IITs Ana University, IIS-B and CDAC	Indian Language to Indian language
2009	Hindi to Punjabi	Vishal Goyal, Punjabi University, Patiala	Word to word Hindi-Punjabi

C-DAC has developed plethora of tools to enable development of Indian language application with greater ease, below is a list to name a few:

- Intelligent Script Manager (ISM)
- Name Translation Tool from English to Indian Language
- Indian Language Software Development Kit
- iPlugin (web based development tool for Indian languages)

C-DAC Kolkata language processing team is working for the development of Bangla language focusing mainly translation of Bangla language to English and other Indian languages[31].

In India, number of translation systems has been developed which facilitates source language translation into target language. Table 1 presents an overview of the existing machine translation systems[32].

Conclusion

Developments, drawbacks, issues and scope of Hindi language in the computational system were studied based on literature review. It is found that a number of problems still exist in the area of translation involving the Hindi language. Many institutes are working on a number of projects for the development of Hindi language in the field of computers. However, searching for Indic languages using word formation techniques or morphological studies are yet to be undertaken.

References

1. Kumar S and Mansotra V, Query optimization: a solution for low recall problem in Hindi language information retrieval, *International Journal of Computer Applications*, 55 (17) (2012).
2. Maurya A, Goswami S, Tyagi AK, Kumar A and Singh M, Creating websites in Hindi, *DESIDOC Bulletin of Information Technology*, 22 (2),2002, 3-16
3. Kumar S and Mansotra V, Factors affecting the performance of Hindi language searching on web: an experimental study,

International Journal of Scientific and Engineering Research, 3(4) (2012) 1-15.

4. Meredith J, Theory building through conceptual methods, *International Journal of Operations and Production Management*, 13 (1993) 3.
5. Fink A, *Conducting Research Literature Reviews: From Paper to the Internet*, Sage, Los Angeles, C A, 1998.
6. Literature review. Available at http://www.emeraldinsight.-com/doi/full/10.1108/13598541211246549 (Accessed on 10 September 2014).
7. Webster J and Watson Rechard T. Analyzing the past to prepare for the future: Writing a literature review. Available at http://www.academia.edu/2831926/LITERATURE_REVI-EW_w_mmmm_) (Accessed on 8 September 2014).
8. Write a literature review. Available at http://guides.library.-ucsc.edu/write-a-literature-review (Accessed on 15 September 2014).
9. Denyer D and Tranfield D, *Producing a systematic review, The sage handbook of organizational research methods*, Sage publishing, p.p (671-689).
10. ISCII. Available at http://tdil.mit.gov.in/standards/iscii.aspx (Accessed on 15 September 2014).
11. Chitrankan. Available at http://cdac.in/index.aspx?id=mlc_g-ist_chitra/(Accessed on 18 January2015).
12. Rastogi P and Dwivedi SK, Performance comparison of word sense disambiguation algorithm ln Hindi language supporting search engines, *International Journal Of Computer Science*, 8 (2) (2011).
13. Saravanan K, Raghvendra U and Kumaran A, Improving cross-language information retrieval by transliteration mining and generation, multilingual system research, Microsoft Research India, Bangalore
14. Tripathi A, Problems and prospects of Hindi language search and text processing, *Annals of Library and Information Science*, 59 (2012), 219-222.
15. Kumar S and Mansotra V, An experimental analysis on the influence of English on Hindi language information retrieval, *International Journal of Computer Applications*, 41(11) (2012) 30-35.
16. Kumar S, Factors affecting the performance of Hindi language searching on Web. Available at http://www.ijser.org/researchpaper%5CFactors-Affecting-the-Performance-of-Hindi-Language-searching-on-web.pdf/ (Accessed on 18 January 2015).
17. Naskar S and Bandhopadhyay S, Use of machine translation in India: Current status. Available at http://www.mt-archive.info/MTS-2005-Naskar-2.pdf/ (Accessed on 20 December 2014).
18. Nair LR and Peter SD, Machine Translation System for Indian Languages available at http://research.ijcaonline.org/-volume39/number1/pxc3877014.pdf/ (Accessed on 15 January 2015).
19. MANTRA: Machine Assisted Translation available at http://www.cdac.in/html/aai/mantra.asp/ (Accessed on 15 January 2015).
20. Sinha RMK, Machine Translation. Available at https://sites.google.com/site/profrmksinha/research-projects/-machine-translation/ (Accessed on 2 December 2104).
21. Prasad P and Verma V, Hindi and Telugu to English cross language information retrieval at CLEF 2006, Language and Technology Research Centre, IIIT Hyderabad. Available at http://ceur-ws.org/Vol-1172/CLEF2006wn-adhoc-PingaliEt2006.pdf (Accessed on 17 March 2015).
22. Swapna N, Kumar HN and Padmaja RB, Information retrieval in Indian Languages: A case study on cross-lingual and multi-lingual, *International Journal of Research in Computer and Communication Technology* (2278-5841), 2012, 1.
23. Mandal D, Gupta M, Dandapat S, Banerjee P and Sarkar S, Bengali and Hindi to English CLIR evaluation, Department of Computer Science and Engineering, IIT Kharagpur. Available at http://web.cs.iastate.edu/~debasis/papers/clef-2007_final.pdf. (Accessed on 18 May 2016)
24. OM: One Tool for Many Indian Languages. Available at http://repository.ias.ac.in/64442/1/19-auth.pdf/ (Accessed on 5 January 2015).
25. Ganapathiraju M, Balakrishanan M, Balakrishanan N and Reddy R, OM: One tool for many Indian languages. Available at http://www.serc.iisc.ernet.in/~balki/papers/A05-1124.pdf (Accessed on 17 March 2015).
26. Center for Indian Language Technology, IIT Bombay. Available at http://www.cfilt.iitb.ac.in/(Accessed on 10 January 2015).
27. Das S, Seetha A, Kumar M and Rana J L, Post Translation Query Expansion Using Hindi Word Net For English –Hindi CLIR System, Department Of Computer Applications, MANIT, Bhopal, Dr. C.V Raman University Bilaspur and Department Of Computer Science & IT, SIRT, Bhopal, FIRE 2010, Feb 2010, Gandhinagar, Gujrat, India. Available at http://irsi.res.in/fire/paper_2010/sujaydas-manit-fire2010.pdf/ (Accessed on 6 January 2015).
28. AnglaBharti Mission. Available at http://www.cse.iitk.ac.in/-users/rmk/mission/mission.htm/(Accessed on 20 September 2014).
29. Anussarka. Available at http://www.chinfo.org/sankara/Brochures/Anusaaraka%20Brochure.pdf/ (Accessed on 5 January 2015).
30. C-DAC Indian Language for Media. Available at http://cdac.in/index.aspx?id=mc_ilm_indian_language_media (Accessed on 4 December 2014).
31. http://cdac.in/index.aspx?id=mc_ilf_indian_language_fcdt (Accessed on 4 December 2014).
32. Kumar S, An extensive literature review on CLIR and MT Activities in India. Available at http://www.ijser.org/researchpaper%5CAn-Extensive-Literature-Review-on-CLIR-and-MT-activities-in-India.pdf (Accessed on 12 September 2014).