Annals of Library and Information Studies
Vol. 59, December 2012, pp. 219-222

Problems and prospects of Hindi language search and text processing

Aditya Tripathi

Assistant Professor, Department of Library & Information Science, Banaras Hindu University, Varanasi – 221005. Email: adeethtripaathi@gmail.com, aditya@bhu.ac.in

Hindi has evolved to its' present form almost a thousand years before. Today it has become a major language of the world and requires representation over Web as an expressive language for literary people and those who love to communicate in Hindi. Unicode has enabled to read and write over the Web but Hindi text processing and retrieval are some major issues which should be given considerable attention. The present paper draws attention of researchers to undertake the challenging areas of Hindi language processing and invites them to undertake research projects.

Keywords: Text processing, Hindi, Language search, Indian language

Introduction

Hindi is a spoken language of several of the Indians as well as people from abroad. There are many who can't understand any other language except Hindi. Like any other language, Hindi should be represented and supported over Web with services and tools. Unicode has enabled to read and write over Web but one of the major obstacles is support by existing search engines for effective retrieval. Google and others do support Hindi search but that is only limited to pattern matching of literals[1]. Searching on the Internet is important to locate the documents, digital objects or services. Domain specific intelligent agents are playing an important role in tourism, academics, social networking and so on. In a multilingual world it is necessary to develop such agents which not only retrieve precise information but also understand language of users[2].

The whole scene warrants a look into the language research in India with relation to character representation, word and sentence formations. The area of Indian languages is quite diversified and includes all Indian languages, however, the present study looks only at the Hindi language.

Indian languages can be divided into four families of languages namely, the Austric (Austro-Asiatic) or 'Eastern' family, the Sino-Tibetan (Tibetan-Chinese) family, the Dravidian family and the Indo-Aryan (Indo-European) family. In terms of total number of speakers, Hindi ranks third after English and Chinese[3]. Hindi is not restricted to any particular province and has wider audience. Hindi evolved from Sanskrit as an offshoot of Prakrit languages through several dialectical changes and the modern form has emerged in the popular form of Khari Boli only around 1000 AD, as it is assumed. Hindi has been influenced and enriched by Dravidian, Turkish, Farsi, Arabic, Portugese, English and so on. Hindi is often confused with Urdu (word *Urd* means language of camp) which was spoken within camps of soldiers during Mughal period[4]. Hindi is written in Devnagari or 'Nagari' script which is phonetic script that is why unlike English, Hindi is pronounced as it is written. The language has two phonemes; vowel and consonant. There are 13 vowels and 33 consonants in the language, in general.

Language research in India dates back to 1970s and is mostly related to language translation. However, the state of affairs has changed and many factors were identified in 1990s. These are:

- Indian Language Processing (ILP) Tools
- Indian Language Resources: Corpora, Lexical Resources, Dictionaries
- Web based search tools

The mentioned three areas are of utmost importance if Indian languages have to make the mark on Internet.

Indian Language Processing (ILP) tools

In order to represent a language over the Web, a writing platform is required which includes an editor, a standard for character representation, fonts or more precisely glyphs. Unicode, which is a multi-byte code, is a global standard for character representation covering all the world scripts except the obsolete or dead scripts[5]. The Indic script table of Unicode is built on Indian Script Code for Information Interchange (ISCII), which is a one byte code. In good old ISCII standard, switches were used to change the scripts whereas, Unicode is devoid of this lacuna[6] and each character is represented with a unique value. Today there are a number of editors available to create formatted as well as unformatted text, like Notepad, Wordpad, Gedit, Microsoft Office Suite, Open Office Suite and so on. The regular features like, formatting the text and decorations are there with all the tools for Hindi however, sorting and counting the characters are missing. The sort program sorts the character by it's Unicode value, unfortunately in case of Indian script, sort order cannot be rendered with Unicode values of characters. There are several combinatorial characters that are derivative of two or more than two characters placed at two different places in Unicode table, for example, the character KSH (क्ष) is achieved by the combination of characters KA (क) and SHA (ष). The character KSH (क्ष) is placed nearly at the bottom of the table however, with regular sort program of a word processor it would be sorted and place immediately after character KA (क)[5] i.e. the top of the table.

Another difficulty is reflected at the time of character count. For combination characters, word-processors can't count the number of character and counts even the hidden characters which are not counted otherwise in a written Devnagari script.

For example,

त्र → त + ् + र

The cited example suggests that there are three characters in the word processor however, in a writing system of Devnagari it is considered as one character only. Though it sounds trivial but while processing the character strings such cases make significant differences.

Indian Language Resources: Corpora, Lexical Resources, Dictionaries

Computational linguistic studies can be conducted only when a good number of Corpora exist. Corpora is a written text or list of words on which processing can be done to understand the possible outcomes. It tests the experiments and builds the level of confidence. In order to understand a word or a sentence, one requires a dictionary where each word is given with its part of speech, gender forms and so on, more like the English Wordnet[7]. This project has been taken up and is yet to be completed[8]. Another reason for not having good corpora is lack of OCR (Optical Character Recognition) software. Centre for Digital and Advanced Computing (CDAC), Pune has produced software, namely, Chitrankan which promises to scan Devnagari text and converts it to editable form[9]. Chitrankan has still long way to go and different font faces have to be incorporated for effective character recognition.

One of major trouble with Hindi is lack of rules with regard to formation of words. The simple example that can be cited is identification of masculine and feminine form. There is no set of rule(s) which suggests that a word is masculine or feminine form. Further, it is also difficult to identify singularity and plurality of words. Though a common understanding or rules do exist but very frequently these rules are violated and cannot be accommodated as it is, in any computational system. For example, word लड़की, खिड़की, ढ़फली, बकरी etc. are feminine, and an understanding can be made that words with vowel EE ending are feminine but गाय, किताब, कमीज़, भैंस etc. are also feminine, these are equally common used words. Hence, it becomes very difficult to distinguish the gender forms. Once gender forms are decided, singularity and plurality can be established by already laid rules in Hindi linguistics though it is not very consistently followed all over.

There are many words in Hindi which are derived due to the combination of words. While searching or processing the text such words are to be broken into components and then the search should be performed. However, Hindi has well laid rules for such kind of word formation known as Sandhis and Samasas. Sandhi is a conjugation of a word, where each word is pronounced as it is, except the conjugation point, for example, उद्धर्वोन्मुख is formed with the combination of words उद्धर्व and उन्मुख. In Samasa, both the words are conjugated and their pronunciation also remains same, for example, विक्रमादित्य is combination of words विक्रम and आदित्य. In principal, it looks easy to mechanize construction and dissociation of words with Sandhis and Samasas. A detailed study into would provide better insight and further challenges[10].

Web based search tools

Searching and search agents are key features of World Wide Web (WWW). It will be very hard to imagine Internet without search engines. Started from plain directory based search (Archie, Veronica and JugHead), searching has become a big market for business. Not only helping to find the information or resources over Web, it has become a medium for marketing and advertisement. The effect of dynamism can be seen through the amount of search companies sold, bought and slowed to halt (due to hard competition) in last two decades. The reining players of the day are, Google, Yahoo and Bing competing neck to neck[11].

On the Internet, 72.3% users are non-English users[12]. These users involve content creators, searchers, readers, customers of services/products and often the common man. There have been 121,000,000 Internet users in India with 10.2% Internet penetration in the country. The search companies do target this 72.3% population as this is considered as the major consumer base for all the productions of the world. The strategies to pick-up the market involves, clustering the webpages, ranking of webpages and other Search Engine Optimization (SEO) techniques.

Targeting this multilingual population with a multilingual facility of search is a major challenge. Considering Hindi alone, the language has several dialects and variations. Unlike English, Hindi has loose grammar and phoneme methods. Hence, it is difficult to mechanize the search process. In English, the search strategies involve use of soundex which is further improved by metaphone algorithms. Soundex algorithms are used to search words with same pronunciation but different spellings and meanings i.e. homonyms and homophones or otherwise similar kind of pronunciation but not actually homonyms and homophones. Soundex is further improved by use of metaphones which is now in generation of metaphone 2 or double metaphones and metaphone 3[13]. For English, metaphone algorithm works out well with the accuracy of almost 99% however in other languages it is still being worked out.

As it is pointed out, for other languages these algorithms are yet to evolve and Hindi is no exception. There are search engines which specifically index Hindi webpages like, Khoj.com[14], Raftaar.in[15]and Guruji[16]. However, Google and other search engines like Yahoo and Bing also support Hindi but it is restricted to regular expressions. Combinatorial search with Boolean operators can also be achieved using these search engines. Indexers like Lucene though have implemented Unicode character search but that also is limited to regular expressions only[17].

Another approach of searching the words in Hindi is stemming. Stemming means extracting the root word from a given word. It is assumed that once the root word is derived, all variations from the root can be searched at once. Since, Hindi borrowed words from Farsi, Awadhi, Sanskrit, Persian, Panjabi, Oriya, Dravidian, Turkish, Arabic, Portuguese, English etc. and words from different languages follow the rules of respective languages for word formation, hence, it is very difficult to form the rules for word formation in Hindi language though grammar of Hindi is fairly consistent when it comes to forming a sentence. With this context it is very difficult to formulate methods to mechanize the search using stemming algorithm.

Searching a word in English sounds a simple task but answering the queries in natural language still needs maturity. Several search engines appeared in due course claiming to understand natural language and answering them, ASKJEEVES.COM (now ask.com[18]) was one such example, which claimed once upon a time that it can handle queries in natural language but failed to do so even with simple queries. The strategy these search engines used is to cull out key words from the given sentence and perform search. Understanding semantics of a sentence in English is a far cry even today and requires deep level syntactic analysis of sentences. And because of this simple reason machine translation has not been achieved. Condition of Hindi or any other language of the world is no different from English.

Conclusion

India has officially 22 recognized languages. Having a lead software workforce nation we are yet to address the issues of Indian languages and scripts in ICT. There is still lot more to be done towards developing applications, services and tools in our national languages. CDAC has done quite a commendable job and is still working in the direction but initiatives should also come from other research institutions. A collaborative and uniform approach should be adopted towards allocating funds for research project which would encourage others to take part in language research. An emphasis should be given to identify research pockets and they should be duly funded.

Lack of corpora and vastness of Hindi language pose problems to develop suitable algorithm for search. The issues which appear trivial with English language computing can pose serious problem in case of Hindi or any other Indic language. An emphasis should also be given to the projects for developing search algorithms for Hindi, developing an OCR tool, improving word processing tools, translation tools.

References

1 Google Search Engine, Available at http://www.google.com (accessed on 15.7.2012)

2 Tanaka-Ishii K and Nakagawa H, A multilingual usage consultation tool based on internet searching: more than a search engine, less than QA in Proceedings of the 14th international conference on World Wide Web, 2005,363-371 available at http://doi.acm.org/10.1145/1060745.1060800

3 Shukla S, Hindi, Encyclopaedia of Language and Linguistics 2nd Edition. Brown, Keith (ed.) V 5. 2006, 303-305.

4 The Urdu Language, Available at http://www.theurdulanguage.com/History.htm (accessed on 15.7.2012)

5 The Unicode Standard, Version 5.0 (Boston, MA, Addison-Wesley, 2007. ISBN 0-321-48091-0

6 ISCII, Available at http://tdil.mit.gov.in/Standards/ISCII.aspx (accessed on 15.7.2012)

7 Wordnet: A lexical database for English, Available at http://wordnet.princeton.edu/ (accessed on 15.7.2012)

8 Hindi Wordnet , Available at http://www.cfilt.iitb.ac.in/wordnet/webhwn/(accessed on 15.7.2012)

9 CDAC:Gist-Products, Available at http://pune.cdac.in/html/gist/products/chitra.aspx (accessed on 15.7.2012)

10 Madangopal This Hindi and Dev Nagri. (Metropolitan Book: New Delhi) 1953, 328p.

11 Search Engine History.com, Available at http://www.searchenginehistory.com/(accessed on 15.7.2012)

12 Top Ten Internet Languages - World Internet Statistics, Available at http://www.internetworldstats.com/stats7.htm (accessed on 15.7.2012)

13 Metaphone, Available at http://en.wikipedia.org/wiki/Metaphone(accessed on 15.7.2012)

14 Khoj - Indian Local Search Engine, Available at http://www.khoj.com/(accessed on 15.7.2012)

15 Raftaar - Hindi Search Engine India - (रफ़्तार - हिन्दी खोज यंत्र) हिन्दी सर्च इंजन भारत।Web Search, Available at http://raftaar.in/ (accessed on 15.7.2012)

16 guruji: the Indian search engine, Available at http://www.guruji.com/ (accessed on 15.7.2012)

17 Apache Lucene, Available at http://lucene.apache.org/ (accessed on 15.7.2012)

18 Ask.com - What's Your Question? Available at http://www.ask.com/(accessed on 15.7.2012)