ISSN: 2455 – 5428
International Journal of Current Research and Modern Education
Impact Factor 6.725, Special Issue, January - 2017
International Conference on Smart Approaches in Computer Science Research Arena
On 5th January 2017 Organized By
Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu

DEVELOPMENT IN INFORMATION RETRIEVAL FOR HINDI LANGUAGE

C. Mohana
Assistant Professor, Department of Hindi, Sri Sarada College for women (Autonomous), Salem, Tamilnadu

Cite This Article: C. Mohana, "Development in Information Retrieval for Hindi Language", International Journal of Current Research and Modern Education, Special Issue, January, Page Number 40-42, 2017.

Abstract:

India is a highly multilingual country with 22 constitutionally recognized languages. Besides these, hundreds of other languages are used in India, each one with a number of dialects. The officially recognized languages are Hindi, Bengali, Punjabi, Marathi, Gujarati, Oriya, Sindhi, Assamese, Nepali, Urdu, Sanskrit, Tamil, Telugu, Kannada, Malayalam, Kashmiri, Manipuri, Konkani, Maithali, Santhali, Bodo, and Dogari. Hindi written in the Devanagari alphabet is India's official national language and has the most speakers, estimated to be more than 500 million. I present an overview of the historical development of the modern Indicscripts' writing system, their mechanization and adaptation to computing, and I examine how this facilitated development of Indian language processing. I concentrate primarily on the Devanagari script and the Hindi language as these are the most popular on the subcontinent

Introduction:

The Indian Constitution mentions 18 languages as languages of India. Each language has its own literature, comprising great novels, drama and poetry. It is indeed these differences that make India an interesting country. India has over 1683languages and dialects and an estimated 850 languages in daily use. All communities have their own culture, rooted from their language and all languages have their literature. Hindi is the third most widely-spoken language in the world (after English and Mandarin): an estimated 500-600 million people speak the language. A direct descendant of Sanskrit through Prakrit and Apabhramsha, Hindi belongs to the Indo-Aryan group of languages, a subset of the Indo-European family. It has been influenced and enriched by Persian, Turkish, Farsi, Arabic, Portuguese, and English. Hindi is broadly identical with Urdu, the official language of Pakistan, and is closely related to Bengali, Punjabi and Guajarati. A good knowledge of Hindi is therefore likely to be useful to anyone having an interest in the countries of South Asia or in the numerous South Asian communities of the world. Devanagari consists of 11 vowels and 33 consonants, and is written from left toright. The general appearance of the Devanagari script is that of letters 'hanging from a line'. This 'line', also found in many other South Asian scripts, is actually a part of most of the letters and is drawn as the writing proceeds. The script has no capital letters

Characteristics of Hindi Language and Devanagari Script:

Hindi is written using the Devanagari script. Devanagari is also used to write other languages, such as Nepali and Marathi, and is the most common script used to write Sanskrit. Several other languages have scripts which are related to Devanagari, such as Bengali, Punjabi, and Gujarati. Devanagari consists of 11 vowels and 33 consonants, and is written from left to right.

Unicode and Devanagari:

They are all alphabets in which most symbols stand for a consonant plus aninherent vowel (usually the sound /a/). In the Unicode Standard, this sign is denoted by the Sanskrit word virzma. In some languages another designation is preferred. In Hindi, for example, the word Hal refers to the character itself, and halant refers to the consonant that has its inherent vowel suppressed; in Tamil, the word pukki is used. Kharoshthi, written from right to left, was supplanted by Brahmi and its derivatives. There are said to be some 200 different scripts deriving from it. By the eleventh century, the modern script known as Devanagari was in ascendancy in India proper as the major script of Sanskrit literature. The major official scripts of India proper, including Devanagari, are all encoded according to a common plan, so that comparable characters are in the same order and relative location. This structural arrangement, which facilitates transliteration to some degree, is based on the Indian national standard (ISCII) encoding for these scripts, and makes use of a virama.

Standards:

The Devanagari block of the Unicode Standard is based on ISCII-1988 (Indian Script Code for Information Interchange). The ISCII standard of 1988 differs from and is an update of earlier ISCII standards issued in 1983 and 1986. The Unicode Standard encodes Devanagari characters in the same relative positions as those coded in positions A0–F416 in the ISCII-1988 standard. The same character code layout is followed for eight other Indic scripts in the Unicode Standard. In November 1991, at the time The Unicode Standard, Version 1.0, was published, the Bureau of Indian Standards published a new version of ISCII in Indian Standard (IS) 13194:1991

Design of English-Hindi Based Cross Language Information Retrieval System:

The following are the steps of the system. Initially the query will be taken as input from the user the query can be in English or Hindi. The query if in Hindi will be converted to English with the help of bilingual dictionary. The problem of ambiguity will be handled using multiple selection technique. The documents will be retrieved using cosine similarity. The relevant documents will be shown to user if user is not satisfied query expansion option will be their where the pseudo relevance feedback and co-occurring term technique will be combined to restructure the query. After which the final retrieval will be done.

Conversion of Hindi Words to English Words:

For the conversion of Hindi words to English words, a dictionary database has been built. Hindi language consists of words which can have different meanings in English. Thus both the entries are stored in dictionary and while retrieving the

ISSN: 2455 – 5428
International Journal of Current Research and Modern Education
Impact Factor 6.725, Special Issue, January - 2017
International Conference on Smart Approaches in Computer Science Research Arena
On 5th January 2017 Organized By
Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu

documents both type of documents will be retrieved. To further improve the result query expansion option will be provided. Below is the example of how the dictionary will look.

English Hindi Dictionary:

English Word	Hindi Word	Hindi English Word
Computer	lax.kd	sanganak
Study	i<kbZdjuk	adhyayan
Structure	lajpuk	sanrachana
Processing	çlaLdj.k	prasanskaran
Engineering	vfHk;kaf=dh	abhiyaantrikee

Indian Languages on Internet:

Rise of Hindi, Urdu and other Indian languages on the Web, has lead millions of non-English speaking Indians to discover uses of the Internet in their daily lives. They are sending and receiving e-mails, searching for information, reading e-papers, blogging and launching Web sites in their own languages. Two American IT companies, Microsoft and Google, have played a big role in making this possible. The new character encoding from Unicode Consortiuma nonprofitin California whose members include Google, IBM, Oracle, Microsoft, Sun Microsystems, Yahoo and the Government of India proved to be a boon for Indian languages. Microsoft incorporated the Hindi Unicode font, Mangal, in its operating system in 2001. Since then, the Hindi Unicode support has been a part of all subsequent up gradations of Microsoft'soperating systems. The earlier system could incorporate only 127 characters, which is not enough for the Hindi Devanagariscript. The Unicode system can incorporate up to 65,000 characters. As most computers in India use Microsoft's operating system, it ensured that the Hindi font was available to most of them as they upgraded the operating software. In 2004, the Hindi version of Microsoft Office 2003, which included Word, Excel, PowerPoint and Outlook, was launched. Now the Hindi version ofMicrosoft Office 2007 is also available.

Impact and Uses of Hindi WordNet:

- ✓ Free download with API under *GPL*
- ✓ Available from LDC (linguistics data consortium), Upend
- ✓ To be available from ELRA: Language Data Repository of Europe
- ✓ Available from LDC-IL: LDC of India
- ✓ Daily reference from all over the world
- ✓ More than 281000 hits so far since 2006
- ✓ More than 5000 downloads
- ✓ Pivot for WordNet's of many Indian languages
- ✓ Base resource used by many researchers for ILwork on translation, summarization, cross lingual search
- ✓ Commercial license acquired by major search engines companies

Development of Language Corpora in Indian Language:

Central Institute of Indian Languages (CIIL) is a nodal agency for development of Indian Language Corpora. It has co-coordinated with various Indian agencies and Universities for developing more than 45 million corpora in Scheduled Languages of India which is also a part of TDIL program. Enabling Minority Language Engineering (EMILLE) program provides the corpora, architecture and tool for Asian languages. It has a monolingual corpus which contains approximately 96,157,000 words and a parallel corpus consists of 200,000 words of text in English which helps in the translation of Bengali, Hindi, Punjabi and others languages. C-DAC Noida has developed the parallel text corpus Gyan-Nidhi for 12 Indian languages (Hindi, Punjabi, Gujarati, Marathi, Tamil, Telugu, Kannada, Nepali, Oriya, Malayalam, Bangla, and Assamese).

Machine Translation in India:

Although Translation in India is old, Machine Translation is comparatively young. Earlier efforts in this field have been noticed since 1980, involving different prominent Institutions such as IIT Kanpur, University ofHyderabad, NCST Mumbai and CDAC Pune. During late 1990 many new projects initiated by IIT Mumbai, IIIT Hyderabad, AU-KBC Centre, Chennai and Jabalpur University, Kolkata were undertaken. TDIL has started a consortium mode project since April 2008, for building computational tools and Sanskrit-Hindi MT under the leadership of Amba Kulkarni (University of Hyderabad). The goal of this Project is to build children's stories using multimedia and e-learning content.

Mantra:

Machine Assisted Translation Tool (Mantra) is a brain child of Indian Government during 1996 for translation of Government orders, notifications, circulars and legal documents from English to Hindi. The main goal was toprovide the translation tools to government agencies. Mantra software is available in all forms such as desktop, network and web based. It is based on LexicalizedTree Adjoining Grammar (LTAG) formalism to represent the English as well as the Hindi grammar. Initially, it was domain specific such as Personal Administration, specifically Gazette Notifications, Office Orders, Office Memorandums and Circulars, graduallythe domains were expanded. At present, it also covers domains like Banking, Transportation and Agriculture etc. Earlier Mantra technology was only for English to Hindi translation but currently it is also available for English to other Indian Languages such as Gujarati, Bengali and Telugu. MANTRA-Rajyasabha is a system for translating the parliament proceedings such as papers to be laid on the Table [PLOT], Bulletin Part-I, Bulletin Part- II, List of Business [LOB] and Synopsis. Rajya Sabha Secretariat of Rajya Sabha (the upper house of the Parliament of India) provides funds for updating the MANTRA – Rajyasabha system.

ISSN: 2455 – 5428
International Journal of Current Research and Modern Education
Impact Factor 6.725, Special Issue, January - 2017
International Conference on Smart Approaches in Computer Science Research Arena
On 5th January 2017 Organized By
Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu

Tamil-Hindi MAT System:

K B Chandrasekhar Research Centre of Anna University, Chennai hasdeveloped the machine-aided Tamil to Hindi translation system. The translationsystem is based on Anusaaraka Machine Translation System and follows lexicontranslation approach. It also has small sets of transfer rules. Users can access the system at
http://www.aukbc.org/research_areas/nlp/demo/mat/

Conclusion:

India being a multilingual country had already recognized the potential of multilingual computing and some of the programs to build competency were initiated more than a decade ago. Since then slowly and steadily many research, development and application oriented activities have been built up at government, public and private organizations. This has resulted in creating awareness about the use of computers in the areas of language analysis, understanding and processing. Preparatory work for building corpora of contemporary text have led to the development of potential applications like morphological analyzers, spell checkers etc. Defining and refining standards, development of operating systems, human machine interfaces, Internet tools and technologies, machine-aided translations and speech related efforts are some of the major thrust areas identified for attention in the near future. Standardization of terminology for use in regional languages is also receiving considerable attention. The challenges ahead require cooperative efforts in the many upcoming areas such as automatic translations of web-based information, search engines, multimedia content generation and refinement of human machine interfaces. It is well recognized that these efforts need to be accelerated particularly to meet the objective of deeper and wider penetration of IT in the country.

References:

1. India (2012). National policy on information and communication technology (ICT) in school education. New Delhi: Department of School Education and Literacy Ministry of Human Resource Development, Government of India.
2. Devanagari through the Ages, pub. No. 8/67,
3. Central Hindi Directorate, New Delhi, 1967.
4. Om Vikas, “Language Technology Development in India”. Ministry of Information Technology, New Delhi, India.
5. http://tdil.mit.gov.in/newsletter1.htm
6. http://www.cdacindia.com//
7. Improving Performance of Hindi-English based Cross Language Information Retrieval using Selective Documents Technique and Query Expansion Aditi Agrawal1, Dr. A. J. Agrawal
8. www.cfilt.iitb.ac.in.
9. http://www.aukbc.org/research_areas/nlp/demo/mat/