Natural Language Processing
Prof. Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

Lecture - 22
Natural Language Processing and Information Retrieval

So far we have looked at hidden markup model part of speech tagging with hidden markup model the task of sequence leveling, where the words of a sentence are given different kinds of levels. Now, today we would like to start a very interesting topic namely the relationship between two extremely important and large fields which deal with the textural information namely information retrieval and natural language processing.

So, if you see the title of today lecture - lecture number 22, that is natural language processing and information retrieval. So, it happens that both the fields are predominantly concerned with text of course music retrieval, video retrieval, image retrieval, and so on are very, very important fields in information retrieval. The fact remains that predominantly people look for textual information on the web and the science and the technology and art of information retrieval is concerned with how to retrieve and present textural information. Natural language processing also is concerned with text, it makes use of linguistic makes use of structure of sentences, and the meaning of sentences to see how language is produced, and how language is understood by human beings.

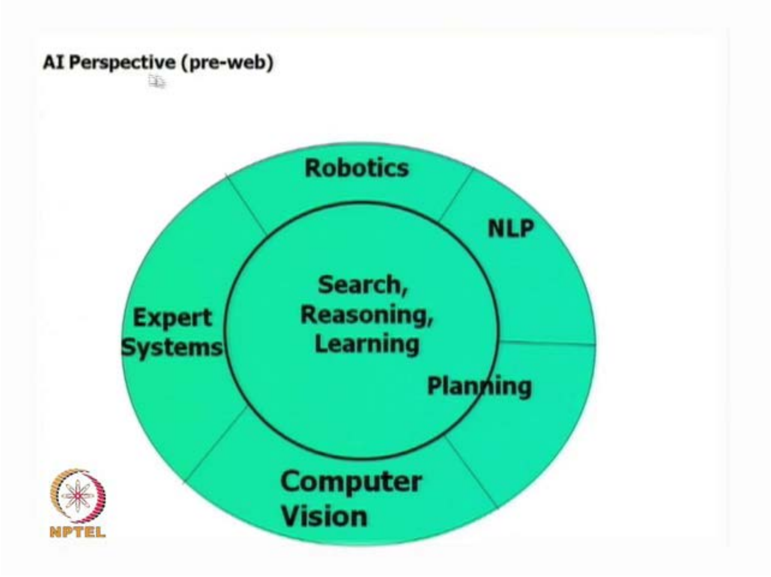
So, both the fields are concerned with text and expression of language in the form of textural data. So, it is very natural that the fields both the fields would have lot of things to do with each other, there should be a kind of synergy between natural language processing and information retrieval. However, if one looks at the history of information retrieval it started with the task of retrieving information in specific information basis. For example, in library information system how would one obtain a book given some specifications for the book like the author topic subject and so on, sometimes in the form of a very dry and precise information like the catalogue number of the book.

However, the information now is larger resident on the World Wide Web there is a tremendous amount of information and large volume of information large diversity of information on the web. One would like to retrieve this information from the web therefore; information retrieval in the context of the web has become far more important than just specific information search or call for search.

So, information retrieval is dealing with obtaining information from the web predominantly, but also could be from specific information basis the natural language processing on the other hand is concerned with the language data. If one looks at the history of the interaction of both the fields there is this intuitive feel that the natural language processing is definitely useful to information retrieval which is concerned obtaining textual information. Natural language processing can also benefit for from information retrieval because for certain language phenomena. A lot of textural information can be presented to the natural language processing system, so both the fields should be helpful to each other to say the least.

However, one sees that information retrieval and N L P in spite of their intuitive relationship and have maintained a respectful distance from each other information retrieval. Researchers have gone ahead with engineering of large systems without really bothering about the nitty, gritties and fineness of language processing language processing. On the other hand has made use of the web to obtain evidences of the language phenomena, but that is all many different techniques of information retrieval have been looked at with interest, but it is doubtful that n l p has made use of lots of i r techniques.

(Refer Slide Time: 05:50)



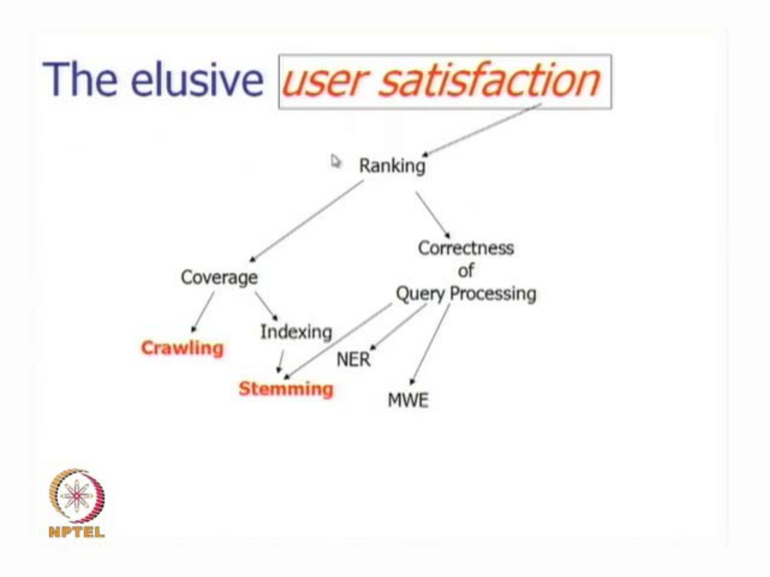
AI Perspective (pre-web)

So, today we would like to discuss these issues in more detail and we will take a perspective view to see where N L P are placed with respect to each other. If we take the a i perspective which is pre web, then there is this diagram of a i which is a classic. Actually, there are these whole areas of artificial intelligence like search reasoning learning some amount of planning which are used by these areas of a i namely robotics natural language processing planning computer vision expert system.

The areas on the outer circle are those that make a i useful for human life computer vision for example, is useful along with robotics for mechanical performance of task involving navigation, picking up of objects and so on and so forth. Language processing is useful for creating natural language interfaces to computational systems. So, all these essentially make use of these areas of a i in the inner circle namely search reasoning learning and planning, so this was pre web.

Post web, one definitely sees a strong presence of information retrieval which the a i community recognizes as important to a i. So, i r also makes use of important concepts in a i which are core do not necessarily the techniques of a i. So, there are specialized i r systems which try to do reasoning while processing the query and retrieving the document machine learning has contributed a lot to information retrieval especially on the question of learning to rank. Now, post web a i when this classical diagram is presented cannot ignore information retrieval.

(Refer Slide Time: 08:11)



The elusive *user satisfaction*

So, now where does language and artificial intelligence come in to play when we discuss information retrieval so this is a diagram which is concerned with the satisfaction of user information need. So, when a user presents a query to the web and retrieves information in terms of documents, then the user is actually presenting his or her information need to the search engine. Now, user satisfaction is completely a function of ranking of the documents, the documents which are retrieved corresponding to a query,

So, ranking essentially means the placement of highly relevant documents towards the top of the list of documents retrieved. So, when a query is presented for example, trekking in Darjeeling or tourism in Taj Mahal when such queries are presented a list of U R Ls are retrieved and the user would like to see that relevant documents are towards the top of the list. So, this particular concern comes within the preview of what is called ranking the documents are ranked according to the relevance.

Ranking on the other hand is a function of two important things, one is coverage after all what is ranked is a list of URLs or web addresses containing information for something to be ranked for a list of information items to be ranked the information. After all it has to be present from wherever it is retrieved, so this is the coverage issue and it is concerned with coverage of information. The more the coverage of information better is the possibility of obtaining relevant information.

Therefore, ranking is a function of coverage the other thing that ranking is a function of is the correctness of query processing when a query is presented the query is not matched directly against the documents. That would lead to what is known as the data sparsity problem, the query as such in the same form the form in which the user has presented his information need. It need not be present in the document itself the parts of the query would be present in the document, and then the document would be retrieved.

So, query processing is concerned with breaking the query into two important parts remove to suffixes. So, that the root word or the essential word is obtained for example, plurals plural markings like E S S are removed the singular root form is obtained. So, those are the functions of query processing it may be necessary to detect that there is a proper noun in the query Shahjahan, Taj Mahal etcetera. This is the task of query processing and the better the accuracy of query processing the better is the ranking of documents.

So, this is shown here query processing is a function of stemming that means bringing the words in to their root or cardinal form N E R means named entity recognition, M W E means multi work expressions multi work expressions like golf club. Then frozen expressions like cut your coat according to your cloth and so on. Coverage is a function of crawling that means retrieving or storing the documents from the web in the private machine depositor.

So, this is known as crawling how well one retrieves the pages one picks up the pages from the web and how frequently one does this is very important for the coverage. Once these documents have been obtained from the web, they have to be indexed there is a very important process called indexing this establishes mapping of important words to the documents which contains that. So, indexing also requires stemming to bring the words into their root form of a cardinal form.

So, this is a very important picture where user’s information need is shown to be a function of many different processes. What is the role of language processing artificial intelligence etcetera in these satisfaction of user information need where in the diagram does one fit the a i and n l p concerns. So, one thing is that when documents are retrieved and their words are stemmed, they often require language processing accurate language processing at the level of words. For example, a Marathi word like [FL] near the house or in front of the house here the most important entity in [FL] is [FL].

So, it is important to strip off [FL] and [FL] from [FL] and get [FL] which is then indexed and the meaning of indexing is already said is that the word. Now, will point to all to its documents which contain [FL], similarly detection of named entities detection of multi words are natural language processing concerns. Of course, they can be done without linguistic purely through machine learning statistics, but beyond the level one sees that language issues or language properties have to be used for higher accuracy in this processes.

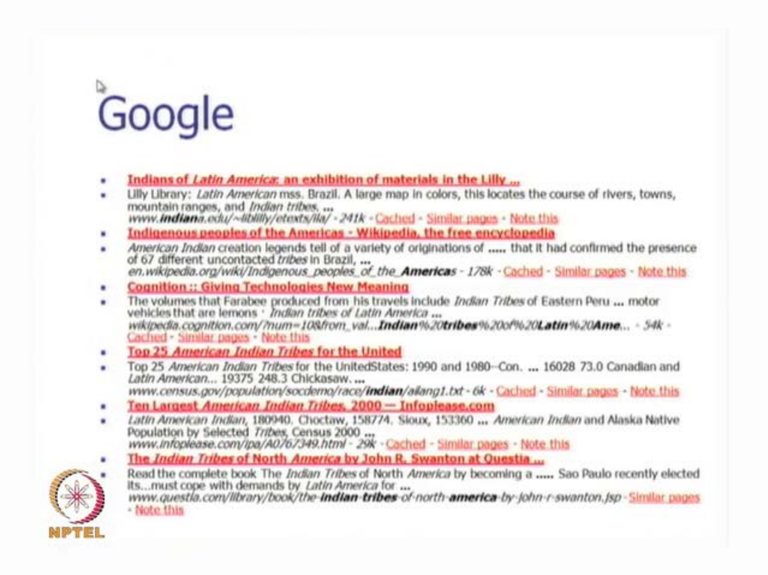
The use of artificial intelligence is in the fact that there is a human user which who is presenting the information need. When user has a mind of his own the human user is a cognitive entity and cognitive signs cognitive psychology modeling of the user, these things are very much concerns of artificial intelligence. So, these diagram therefore, establishes a sort of platform from where a i language processing could be useful to information retrieval.

(Refer Slide Time: 16:12)

Query: Indian Tribes in Latin America

Proceeding further, we take some example suppose there is a query Indian tribes in Latin America, so this is a piece of query presumably the person's. The query maker's information need is to find out things about Indian tribes in Latin America. Latin America is the countries consisting Brazil Argentina Cuba Mexico and so on. One would like to know the existence means of livelihood properties etcetera of Indian tribes in Latin America.

(Refer Slide Time: 16:55)



Google

- **Indians of *Latin America*: an exhibition of materials in the Lilly ...**
- Lilly Library: *Latin American* mss. Brazil. A large map in colors, this locates the course of rivers, towns, mountain ranges, and *Indian tribes*. ...
  *www.**indiana**.edu/~liblilly/etexts/ila/* - *241k* - Cached - Similar pages - Note this
- **Indigenous peoples of the Americas - Wikipedia, the free encyclopedia**
- *American Indian* creation legends tell of a variety of originations of ..... that it had confirmed the presence of 67 different uncontacted *tribes* in Brazil, ...
  *en.wikipedia.org/wiki/Indigenous_peoples_of_the_**America**s* - *178k* - Cached - Similar pages - Note this
- **Cognition :: Giving Technologies New Meaning**
- The volumes that Farabee produced from his travels include *Indian Tribes* of Eastern Peru ... motor vehicles that are lemons · *Indian tribes of Latin America* ...
  *wikipedia.cognition.com/?num=10&from_val...**Indian**%20**tribes**%20of%20**Latin**%20**Ame**...* - *54k* - Cached - Similar pages - Note this
- **Top 25 *American Indian Tribes* for the United**
- Top 25 *American Indian Tribes* for the UnitedStates: 1990 and 1980--Con. ... 16028 73.0 Canadian and *Latin American*... 19375 248.3 Chickasaw. ...
  *www.census.gov/population/socdemo/race/**indian**/ailang1.txt* - *6k* - Cached - Similar pages - Note this
- **Ten Largest *American Indian Tribes*, 2000 — Infoplease.com**
- *Latin American Indian*, 180940. Choctaw, 158774. Sioux, 153360 ... *American Indian* and Alaska Native Population by Selected *Tribes*, Census 2000 ...
  www.infoplease.com/ipa/A0767349.html - *29k* - Cached - Similar pages - Note this
- **The *Indian Tribes* of North *America* by John R. Swanton at Questia ...**
- Read the complete book The *Indian Tribes* of North *America* by becoming a ..... Sao Paulo recently elected its...must cope with demands by *Latin America* for ...
  *www.questia.com/library/book/the-**indian**-**tribes**-of-north-**america**-by-john-r-swanton.jsp* - Similar pages - Note this

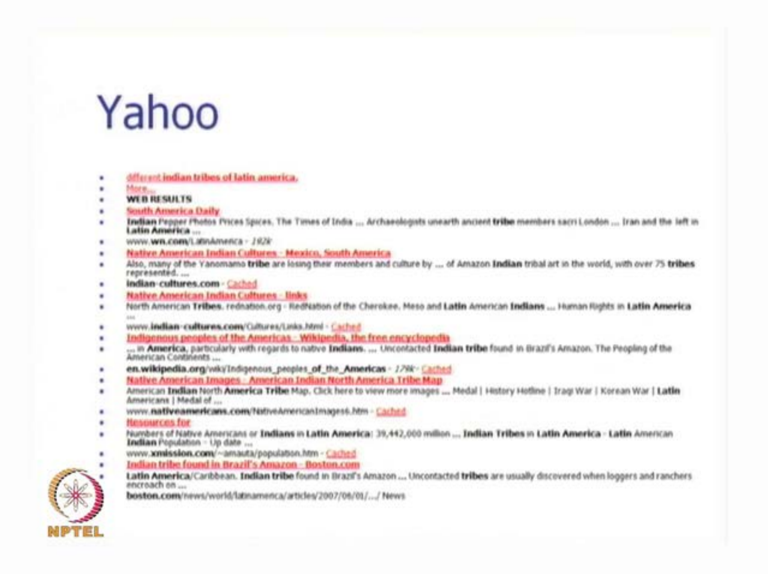
So, when this query is presented to google search for example, one sees that this is the U R L which is retrieved Indians of Latin America and exhibition of materials in the Lili etcetera. So, Lili library is in Brazil which is a large map of colors this locates the course of rivers towns mountain ranges Indian tribes etcetera. This is not exactly what the user has asked for this is about a particular library, whereas, the user wants to know about the Indians of Latin America indigenous people of Americas.

This is from the Wikipedia, it is Wikipedia document and here is a small snippet of information which gives a glimpse of what is there to come inside the U R L itself. So, American Indian creation legion tells of a variety of originations of that it had confirmed the presence of 67 different un contacted tribes in Brazil. So, this seems to be relevant to the query the first document which was retrieved was not really relevant cognition.

This is another URL the volume that Farabi produced from his tribal includes Indian tribes of eastern Peru motor vehicles that are lemons etcetera. This also does not really seem relevant it of course mentions a book that somebody called Farabi wrote. There Indian tribes of eastern a very specific region of Latin America is mentioned, but the U R L is not about the habits livelihood origination etcetera of tribes in Latin America.

Next U R L is top twenty five American Indian tribes for the United States top twenty five American Indian tribes for the United States 1900 and 1980 etcetera. So, this is also a book ten largest Americans Indian tribes info please dot come, so this is another U R L. These also really does not satisfy, the information need of the user the Indian tribes of North America by John R Swanton at Questa. This is also a notification of the existence of a book or an advertisement, so as far as relevance is concerned one would think that the second U R L was relevant first U R L was less relevant.

(Refer Slide Time: 19:54)



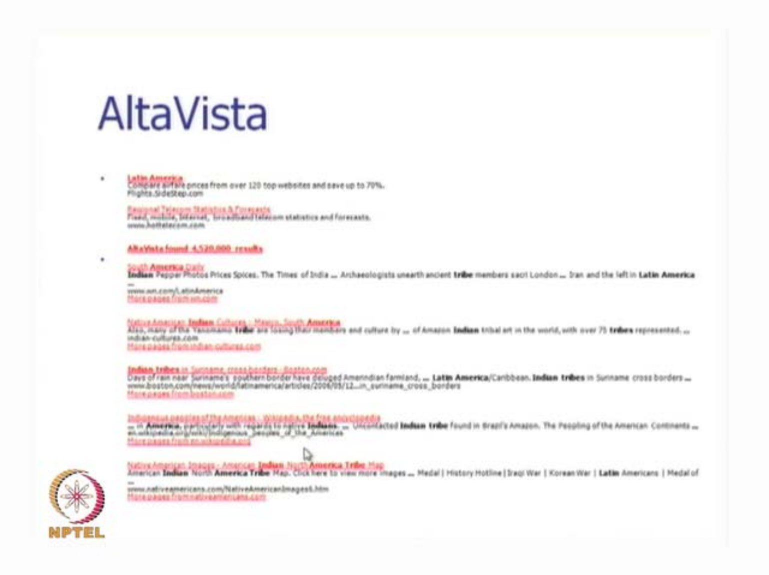
Yahoo

- different indian tribes of latin america.
- More...
- WEB RESULTS
- South America Daily
- Indian Pepper Photos Prices Spices. The Times of India ... Archaeologists unearth ancient tribe members sacri London ... Iran and the left in Latin America ...
- www.wn.com/LatinAmerica - 192k
- Native American Indian Cultures - Mexico, South America
- Also, many of the Yanomamo tribe are losing their members and culture by ... of Amazon Indian tribal art in the world, with over 75 tribes represented. ...
- indian-cultures.com - Cached
- Native American Indian Cultures - links
- North American Tribes. rednation.org - RedNation of the Cherokee. Meso and Latin American Indians ... Human Rights in Latin America ...
- www.indian-cultures.com/Cultures/Links.html - Cached
- Indigenous peoples of the Americas - Wikipedia, the free encyclopedia
- ... in America, particularly with regards to native Indians. ... Uncontacted Indian tribe found in Brazil's Amazon. The Peopling of the American Continents ...
- en.wikipedia.org/wiki/Indigenous_peoples_of_the_Americas - 179k - Cached
- Native American Images - American Indian North America Tribe Map
- American Indian North America Tribe Map. Click here to view more images ... Medal | History Hotline | Iraqi War | Korean War | Latin Americans | Medal of ...
- www.nativeamericans.com/NativeAmericanImages6.htm - Cached
- Resources for
- Numbers of Native Americans or Indians in Latin America: 39,442,000 million ... Indian Tribes in Latin America - Latin American Indian Population - Up date ...
- www.xmission.com/~amauta/population.htm - Cached
- Indian tribe found in Brazil's Amazon - Boston.com
- Latin America/Caribbean. Indian tribe found in Brazil's Amazon ... Uncontacted tribes are usually discovered when loggers and ranchers encroach on ...
- boston.com/news/world/latinamerica/articles/2007/06/01/.../ News

What is the story for yahoo, when this query is given to yahoo, Indian tribes in Latin America? So, it brings up these documents south America daily Indian paper photos spices by spices times of India archeologist unearth ancient tribe members circa London. It talks about some Indian tribe members, but not really of Latin America, it talks about some ancient tribe members, but not really of Latin America, Native American Indian cultures. Also many of the Yanomomo tribe are losing their members and culture by etcetera amazon Indian tribal art in the world with over seventy five types.

So, these is talking about a particular tribe existing in Mexico it is still narrow in its scope of information native American Indian cultures native north American tribes etcetera. So, this brings in information on Native American Indian tribes and their culture not really that of Latin America or not wholly that of Latin America Indigenous people of America. So, this is about the indigenous people in North America Native American images this is about tribal map resources for Indian tribe found in Brazil’s amazon etcetera.

(Refer Slide Time: 21:32)



AltaVista

- Latin America
  Compare airfare prices from over 120 top websites and save up to 70%.
  Flights.SideStep.com

  Regional Telecom Statistics & Forecasts
  Fixed, mobile, Internet, broadband telecom statistics and forecasts.
  www.hottelecom.com

  AltaVista found 4,520,000 results
- South America Daily
  Indian Pepper Photos Prices Spices. The Times of India ... Archaeologists unearth ancient **tribe** members sacl London ... Iran and the left in **Latin America** ...
  www.sn.com/LatinAmerica
  More pages from sn.com

  Native American **Indian** Cultures : Mexico, South **America**
  Also, many of the Yanomamo **tribe** are losing their members and culture by ... of Amazon **Indian** tribal art in the world, with over 75 **tribes** represented. ...
  indian-cultures.com
  More pages from indian-cultures.com

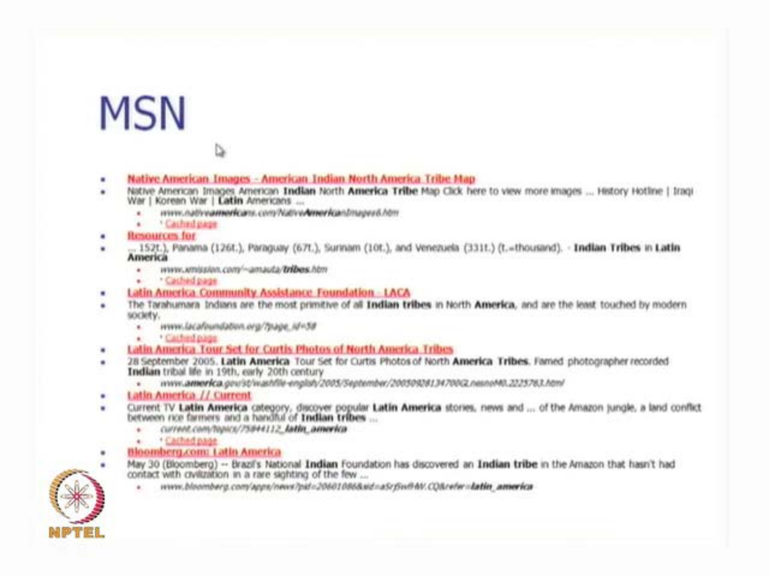
  Indian tribes in Suriname cross borders - Boston.com
  Days of rain near Suriname's southern border have deluged Amerindian farmland. ... **Latin America**/Caribbean. **Indian tribes** in Suriname cross borders ...
  www.boston.com/news/world/latinamerica/articles/2006/05/12...in_suriname_cross_borders
  More pages from boston.com

  Indigenous peoples of the Americas - Wikipedia, the free encyclopedia
  ... in **America**, particularly with regards to native **Indians**. ... Uncontacted **Indian tribe** found in Brazil's Amazon. The Peopling of the American Continents ...
  en.wikipedia.org/wiki/Indigenous_peoples_of_the_Americas
  More pages from en.wikipedia.org

  Native American Images - American **Indian** North **America Tribe** Map
  American **Indian** North **America Tribe** Map. Click here to view more images ... Medal | History Hotline | Iraqi War | Korean War | **Latin** Americans | Medal of ...
  www.nativeamericans.com/NativeAmericanImages6.htm
  More pages from nativeamericans.com

So, here we see some relevant documents towards the end of the list AltaVista is an old search engine. Here we again see these documents finding a place pretty much towards the top Native American Indian cultures has come up which was low in the list of documents that yahoo presented. Indian tribes in Surinam etcetera and again the information is specific to a particular region.

(Refer Slide Time: 22:04)



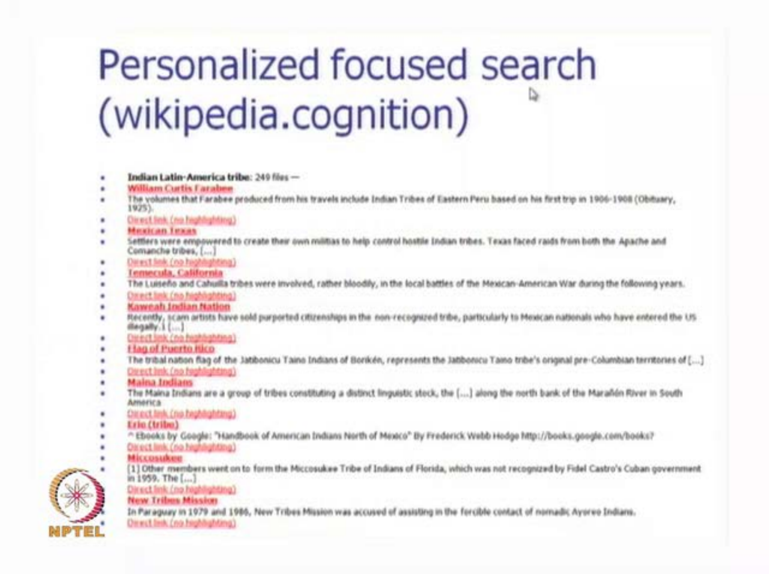
MSN

- Native American Images - American Indian North America Tribe Map
- Native American Images American **Indian** North **America Tribe** Map Click here to view more images ... History Hotline | Iraqi War | Korean War | **Latin** Americans ...
  - *www.nativeamericans.com/NativeAmericanImages6.htm*
  - Cached page
- Resources for
- ... 152t.), Panama (126t.), Paraguay (67t.), Surinam (10t.), and Venezuela (331t.) (t.=thousand). · **Indian Tribes** in **Latin America**
  - *www.emission.com/~amauta/tribes.htm*
  - Cached page
- Latin America Community Assistance Foundation - LACA
- The Tarahumara Indians are the most primitive of all **Indian tribes** in North **America**, and are the least touched by modern society.
  - *www.lacafoundation.org/?page_id=58*
  - Cached page
- Latin America Tour Set for Curtis Photos of North America Tribes
- 28 September 2005. **Latin America** Tour Set for Curtis Photos of North **America Tribes**. Famed photographer recorded **Indian** tribal life in 19th, early 20th century
  - *www.america.gov/st/washfile-english/2005/September/20050928134700GLnesnoM0.2225763.html*
- Latin America // Current
- Current TV **Latin America** category, discover popular **Latin America** stories, news and ... of the Amazon jungle, a land conflict between rice farmers and a handful of **Indian tribes** ...
  - *current.com/topics/75844112_latin_america*
  - Cached page
- Bloomberg.com: Latin America
- May 30 (Bloomberg) -- Brazil's National **Indian** Foundation has discovered an **Indian tribe** in the Amazon that hasn't had contact with civilization in a rare sighting of the few ...
  - *www.bloomberg.com/apps/news?pid=20601086&sid=aSrf5wfHAV.CQ&refer=latin_america*

MSN is another search engine, so these talks about the first document it brings up talks about Native American images India North American tribe and so on. Resources for is the next URL, so Indian tribes in Latin America brings information about these countries which are which is not really fully relevant, but at the same time can be used for useful purposes Latin America community assistance foundation. So, this describes some Indian tribes in Latin America Latin America tour set for Curtis photos of North America tribes. Here, the North American tribes are discussed, but not really those of Latin America and so on and so forth.

So, these examples shows that if we take a query and give it to a search Indian, then it is not necessary to that the search engine retrieves important documents relevant documents from the web to be given to the user. So, the users information need is satisfied that is not true different search engines have different levels of performance.

(Refer Slide Time: 23:32)



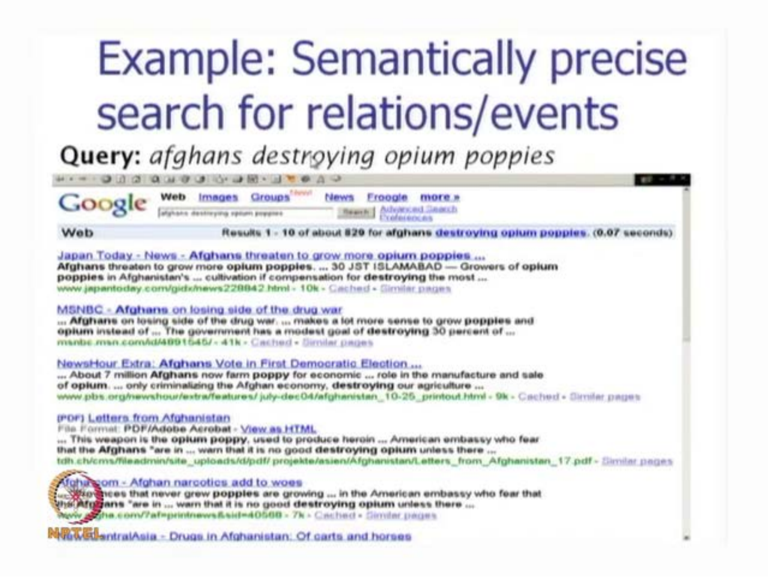
Personalized focused search (wikipedia.cognition)

- **Indian Latin-America tribe**: 249 files —
- William Curtis Farabee
- The volumes that Farabee produced from his travels include Indian Tribes of Eastern Peru based on his first trip in 1906-1908 (Obituary, 1925).
- Direct link (no highlighting)
- Mexican Texas
- Settlers were empowered to create their own militias to help control hostile Indian tribes. Texas faced raids from both the Apache and Comanche tribes, [...]
- Direct link (no highlighting)
- Temecula, California
- The Luiseño and Cahuilla tribes were involved, rather bloodily, in the local battles of the Mexican-American War during the following years.
- Direct link (no highlighting)
- Kaweah Indian Nation
- Recently, scam artists have sold purported citizenships in the non-recognized tribe, particularly to Mexican nationals who have entered the US illegally.1 [...]
- Direct link (no highlighting)
- Flag of Puerto Rico
- The tribal nation flag of the Jatibonicu Taino Indians of Borikén, represents the Jatibonicu Taino tribe's original pre-Columbian territories of [...]
- Direct link (no highlighting)
- Maina Indians
- The Maina Indians are a group of tribes constituting a distinct linguistic stock, the [...] along the north bank of the Marañón River in South America
- Direct link (no highlighting)
- Erie (tribe)
- ^ Ebooks by Google: "Handbook of American Indians North of Mexico" By Frederick Webb Hodge http://books.google.com/books?
- Direct link (no highlighting)
- Miccosukee
- [1] Other members went on to form the Miccosukee Tribe of Indians of Florida, which was not recognized by Fidel Castro's Cuban government in 1959. The [...]
- Direct link (no highlighting)
- New Tribes Mission
- In Paraguay in 1979 and 1986, New Tribes Mission was accused of assisting in the forcible contact of nomadic Ayoreo Indians.
- Direct link (no highlighting)

Now, we move onto something called personalized focus search where the accuracy of retrieval is much more and one hopes to see very relevant documents populating the list of U R L s which is which are retrieved from the web. So, Indian Latin American tribe is given as the query, again we see that the documents which are retrieved like William, Curtis, Farabi. This is information about William, Farabi’s book on tribal in Indian tribes in Latin America Mexican Texas settlers were empowered to create their own militias to help control hostile Indian tribes. Texas faced raised from both the apache and Comanche tribes, so this is about Texas facing problem from certain tribes like apache and Comanche.

This particular U R L itmecula, telephonia, the Luis Seno and Cahulia tribes were believed rather Bazili in the local battles of the Mexican American war during the following years, so these also really is not presenting information on tribes in Latin America.

(Refer Slide Time: 25:08)



Example: Semantically precise search for relations/events

Query: *afghans destroying opium poppies*

Results 1 - 10 of about 829 for afghans destroying opium poppies. (0.07 seconds)

Japan Today - News - Afghans threaten to grow more opium poppies ...
Afghans threaten to grow more opium poppies. ... 30 JST ISLAMABAD — Growers of opium poppies in Afghanistan's ... cultivation if compensation for destroying the most ...
www.japantoday.com/gidx/news228842.html - 10k - Cached - Similar pages

MSNBC - Afghans on losing side of the drug war
... Afghans on losing side of the drug war. ... makes a lot more sense to grow poppies and opium instead of ... The government has a modest goal of destroying 30 percent of ...
msnbc.msn.com/id/4891545/ - 41k - Cached - Similar pages

NewsHour Extra: Afghans Vote in First Democratic Election ...
... About 7 million Afghans now farm poppy for economic ... role in the manufacture and sale of opium. ... only criminalizing the Afghan economy, destroying our agriculture ...
www.pbs.org/newshour/extra/features/ july-dec04/afghanistan_10-25_printout.html - 9k - Cached - Similar pages

[PDF] Letters from Afghanistan
File Format: PDF/Adobe Acrobat - View as HTML
... This weapon is the opium poppy, used to produce heroin ... American embassy who fear that the Afghans "are in ... warn that it is no good destroying opium unless there ...
tdh.ch/cms/fileadmin/site_uploads/d/pdf/ projekte/asien/Afghanistan/Letters_from_Afghanistan_17.pdf - Similar pages

Afgha.com - Afghan narcotics add to woes
... provinces that never grew poppies are growing ... in the American embassy who fear that the Afghans "are in ... warn that it is no good destroying opium unless there ...
www.afgha.com/?af=printnews&sid=40568 - 7k - Cached - Similar pages

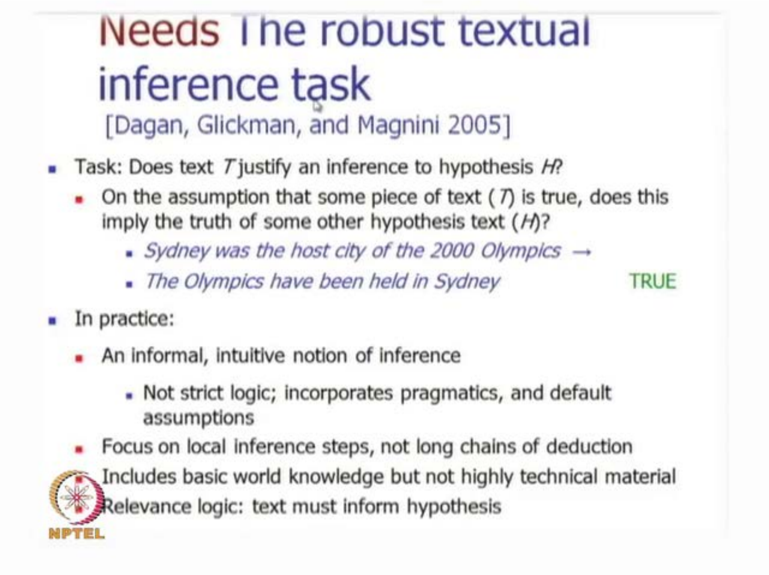
NewsCentralAsia - Drugs in Afghanistan: Of carts and horses

So, now one could make queries which are semantically precise and one could then search for relations and events. So, for example, there is a query Afghans destroying opium puppies so this is about afghans the people of Afghanistan and their act of destroying the opium puppies opium is a very important export from Afghanistan and the national economy depends on opium. So, Afghans destroying is an important or a surprising piece of information, so let see what is retrieved in response to this query. So, Japan today news afghan threaten to grow more opium puppies so you see the query was Afghans destroying opium puppies and what is retrieved is a document which exactly talks about the opposite information.

Here, afghans threaten to grow more opium puppies they are not destroying opium puppies this particular URL afghans are losing sight of the drug war makes a lot more sense to grow puppies and opium instead. So, here again the URL essentially discusses growing more puppies rather than destroying it news hour extra afghans vote for on first democratic election. So, Afghans now farm puppy for economic role in the manufacture and here also the URL is about Afghans voting for democratic election.

Incidentally, the issue of growing or destroying puppy comes up there is a p d f file from where again the snippet is picked out and presented the weapon is the weapon opium puppy used to produce heroine American embassy who fear that the afghans are in warn. It is no good destroying opium unless etcetera, so again we can see a URL which is not talking directly about the act of destruction of opium puppies well. So, semantically precise search for relations and events would have solved this problem because Afghans are actually destroying opium puppies whereas all the URLs almost all the URLs talked about growing more puppies.

(Refer Slide Time: 27:37)



Needs The robust textual inference task

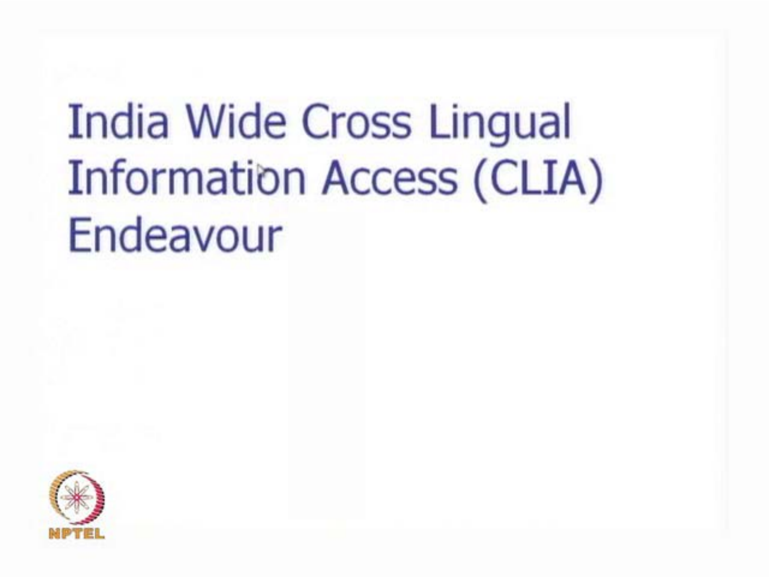
[Dagan, Glickman, and Magnini 2005]

- Task: Does text *T* justify an inference to hypothesis *H*?
  - On the assumption that some piece of text (*T*) is true, does this imply the truth of some other hypothesis text (*H*)?
    - *Sydney was the host city of the 2000 Olympics* →
    - *The Olympics have been held in Sydney* TRUE
- In practice:
  - An informal, intuitive notion of inference
    - Not strict logic; incorporates pragmatics, and default assumptions
  - Focus on local inference steps, not long chains of deduction
  - Includes basic world knowledge but not highly technical material
  - Relevance logic: text must inform hypothesis

Now, these few examples illustrate the need for robust textual information inference task. So, if we look at this task does text t justifying inference to hypothesis h. On the assumption that some piece of text is true does this imply that truth of some other hypothesis text h Sydney was the 2000 Olympics the Olympics have been held in Sydney. So, here is an example of this Sydney was the host city of the 2000 Olympics the Olympics have been held in Sydney which is a true entitlement. So, the question is it that the hypothesis text follows from the text which is given before.

So, in practice an informal intuitive notion of differencing is used which is strictly not based on logic, it incorporates pragmatics and default assumption the focus is on local differencing steps. Long chains of deduction include basic world knowledge, but not highly technical material relevance logic issued text must inform hypothesis. So, the discussion on this particular slide is that there is a need for inferring a piece of text from another and the task of information retrieval can be looked upon as drawing an inference from the query as far as the documents retrieved are concerned. So, we can look up and the retrieved document as being entailed by the query so it becomes an inference problem can the document be retrieved from the query alright we proceed further.

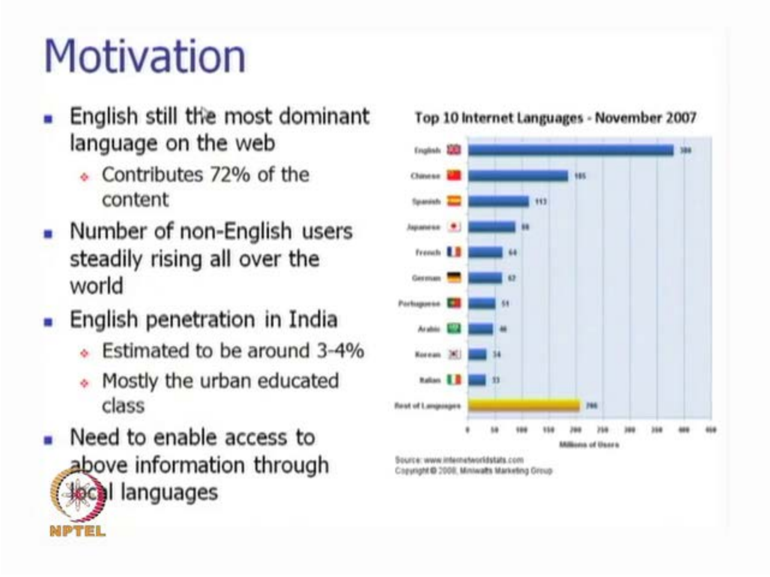
(Refer Slide Time: 29:35)



India Wide Cross Lingual Information Access (CLIA) Endeavour

We do a bit of digression and we describe very quickly, the large project on cross lingual information access these India wide endeavor on cross lingual information access.

(Refer Slide Time: 29:53)



Motivation

- English still the most dominant language on the web
  - Contributes 72% of the content
- Number of non-English users steadily rising all over the world
- English penetration in India
  - Estimated to be around 3-4%
  - Mostly the urban educated class
- Need to enable access to above information through local languages

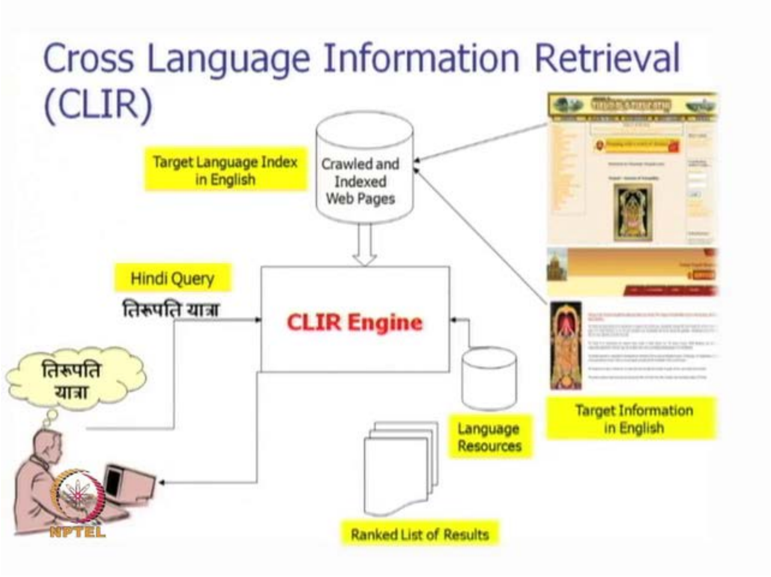
Top 10 Internet Languages - November 2007

Source: www.internetworldstats.com
Copyright © 2008, Miniwatts Marketing Group

Now, the motivation of course of search activity is well known not everybody would be able to use English to satisfy his or her information need. English is still the most dominant language on the web contributes to 72 percent of the content of the web. So, as it is shown here millions of users English has about 388 million users followed by Chinese the number 185 Spanish number 183 Japanese 88 French 64 German 67 Portuguese 51 Arabic 46 Korean 34 and Italian 33. So, these are some top players in terms of language on the web, and we can see that Indian languages still do not figure anywhere for the anywhere for this internet usage.

So, it is therefore, quite considerable that it should be possible for non English speakers non English users also to be able to call out information from the web by making non English queries the number of non English users is steadily rising all over the world. English penetration in India is estimated to be around 3 to 4 percent may be at most 5 percent and that too mostly in the urban educated classes, so there is a need to enable access to information on the web through local languages.

(Refer Slide Time: 31:40)


Cross Language Information Retrieval (CLIR)

So, with that important aim in mind government of India started cross lingual information retrieval project and much cityhood are involved in this project with IIT Bombay leading the effort so here is a schematic of what goes on and the user can very easily appreciate the tasks involved. For example there is a need in the mind of this user about knowing about Tirupati yatra this is a query in Hindi which means visit to Tirupati which is a holy shrine in India with lord Venketeshwara being the resident deity and millions of people visit that shrine. So, presumably the person wants to know the details of Tirupati yatra as to how he could he or she could perform this journey where they could stay you know accommodation facilities food facilities the ways to reach that place and so on.

So, all these are in the mind of the user, but he does not frame a long and precise query on the other hand he just writes on the search box Tirupati yatra. It is a Hindi query which now goes to the C L I R engine the C L I R engine actually converts it into Hindi into English or retains it as such.

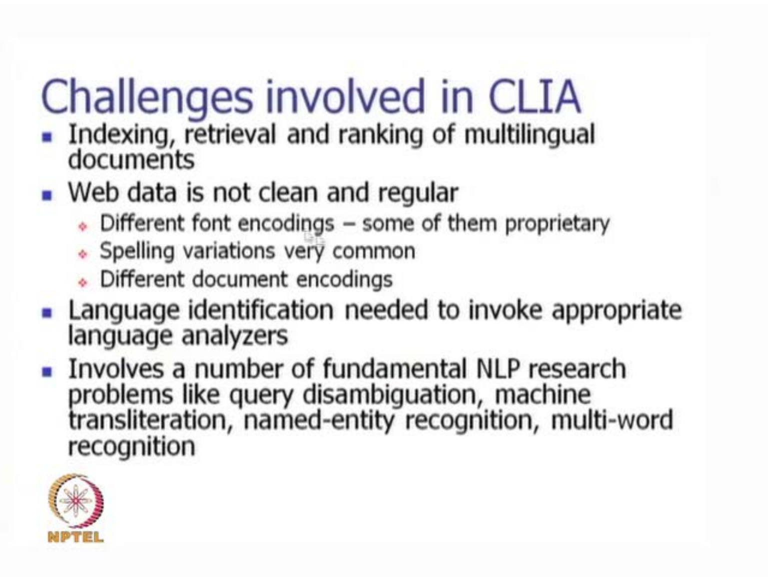
So, when the query is processed by the C L I R engine the relevant documents are retrieved as shown here target information is in English and there is a ranked list of results which the query has brought. So, the way the information flows is that the query comes in to the C L I R engine and is translated subjected to all kinds of ambiguity and other language constructs the web has been crawled which I showed some time back.

Crawl means whether its contents are brought over in to the personal computer, so web is crawled and the index web pages are also there.

So, relevant web pages have been indexed with the key word, so target language index in Hindi so given the Tirupati yatra it first finds out whether the information is contained in any document in Hindi or not. So, that leads to what is called the monolingual retrieval monolingual i r the query is in Hindi and the document which is retrieved is also in Hindi. What is more frequently done is that the language query is given and using language resources using many such modules the query is converted into English. Then the English page is matched again the content of the web in English, so target language index in English so the query was in Hindi.

We have retrieved documents in English this shows cross lingual information retrieval however the information has to be presented observable by the user. Therefore, the English pages may have to translated back to translate into Hindi or the essential information from this should be available in Hindi for the user to appreciate the answer. So, this for example, query for example, retrieves a document containing this sentences [FL], so there is a piece of information which the web has thrown at the user alright the result is obtained in Hindi and this is the snippet.

(Refer Slide Time: 36:17)



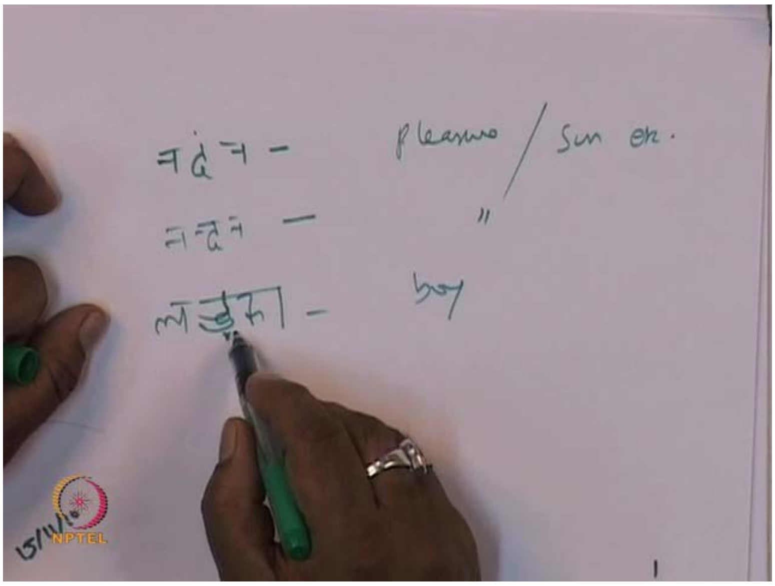
Challenges involved in CLIA

- Indexing, retrieval and ranking of multilingual documents
- Web data is not clean and regular
  - Different font encodings – some of them proprietary
  - Spelling variations very common
  - Different document encodings
- Language identification needed to invoke appropriate language analyzers
- Involves a number of fundamental NLP research problems like query disambiguation, machine transliteration, named-entity recognition, multi-word recognition

So, the challenges which are involved in the cross lingual information access are the following indexing retrieval and ranking of multilingual documents. All these documents should have been crawled up priory from the web and they should have been indexed to be available to the search engine. Ranking of multiple multi lingual documents should be there, so the user can find relevant documents towards the top of the list. So, web data is not clear and regular different font and coding some of them proprietary most of the Indian language content is available in multiple fonts which is a non standard way of writing the text depositing the text.

The problem there is that if the queries encoding is in one font and the documents encoding is in another the matching will not succeed and the document cannot be retrieved. So, these are very important problem most Indian languages have their own different sets of fonts. So, all fonts have to be standardized which is typically u t f 8 font spelling variations are very common, for example one could write Nandan with half na or Nandan with a [FL] on the top.

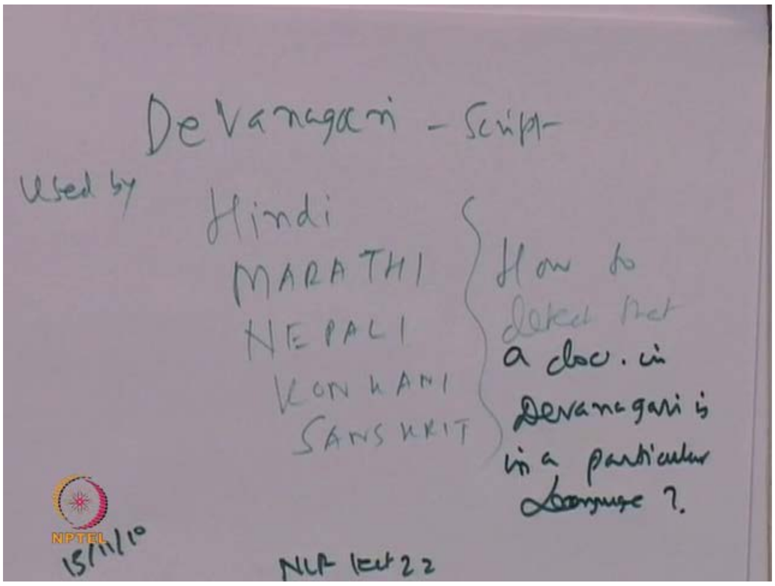
(Refer Slide Time: 37:48)



So, I just write to show you one could write Nandan this way which means pleasure or son etcetera or one could write half na da na. So, this is Nandan again meaning the same thing it is not uncommon to say [FL] which means a boy to be written without this dot. Many times it so happens that this dot is not given, however people can make out that this is [FL] meaning boy.

So, spelling variations are quite common in Hindi and in number of languages documents themselves are encoded in many different ways. So, all these challenges have to be made before cross lingual search can happen many times. It is important to do language identification to invoke appropriate language analyzers, so if the cross lingual information retrieval engine is such that it also does the job of identifying the language in which the information should be presented. Corresponding to the query, then that would require language analyzer, so is a particular document in Hindi or Marathi, so we also know.

(Refer Slide Time: 39:20)



Devanagari - Script

Used by

Hindi
MARATHI
NEPALI
KONKANI
SANSKRIT

How to detect that a doc. in Devanagari is in a particular Language ?

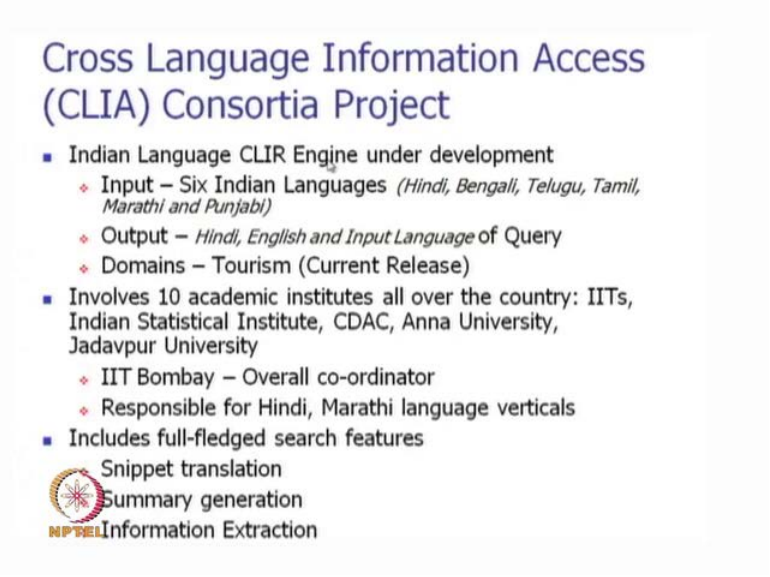
NPTEL

15/11/10

NLP lect 22

For example, that I write it down Devanagari as a script is used by Hindi used by Hindi, Marathi, Nepali, Konkani and Sanskrit. So, the problem with this is that how to detect that how to detect that a document in Devanagari is in a particular language this is the question how to detect that a document in Devanagari is in a particular language. These are language identification problem, so the language analysis involves number of fundamental NLP research problems like query disambiguation machine transliteration named entity recognition multi word recognition and so on and so forth. So, we will see examples of all this when we deal with actual queries but all these tasks are deep and difficult NLP tasks.

(Refer Slide Time: 40:41)



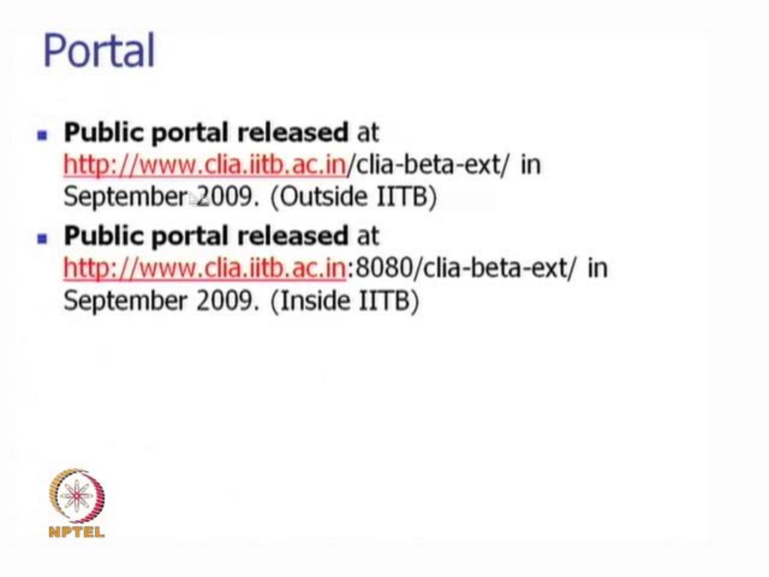
Cross Language Information Access (CLIA) Consortia Project

- Indian Language CLIR Engine under development
  - Input – Six Indian Languages *(Hindi, Bengali, Telugu, Tamil, Marathi and Punjabi)*
  - Output – *Hindi, English and Input Language* of Query
  - Domains – Tourism (Current Release)
- Involves 10 academic institutes all over the country: IITs, Indian Statistical Institute, CDAC, Anna University, Jadavpur University
  - IIT Bombay – Overall co-ordinator
  - Responsible for Hindi, Marathi language verticals
- Includes full-fledged search features
  - Snippet translation
  - Summary generation
  - Information Extraction

Now, the cross lingual information access project is being done in a consortium mode across the country this is under development the input could be a query in any of these six Indian languages namely Hindi, Bengali, Telugu, Tamil, Marathi and Punjabi. The output is in Hindi or English or the input language of query, that means three kinds of outputs are possible either in the language of the query or in Hindi or in English. The domains have been released the domain in which this task is to be implemented that domain is tourism.

So, one might wonder what is the role of domain in this search engine the point is that the queries are, let say in tourism domain then it is important to get the documents from that particular domain. If the resources and tools are built with the domain in mind, then the accuracy is also high, so it involves 10 academic institutes all over the country IITs Indian statistical institute, CDAC, Anna University, Jodhpur university and so on. IIT Bombay is the overall coordinator which is responsible for Hindi, Marathi language vertical it includes full fledged search features like snippet translation, summary generation, and information extraction in addition to searching.

(Refer Slide Time: 42:21)

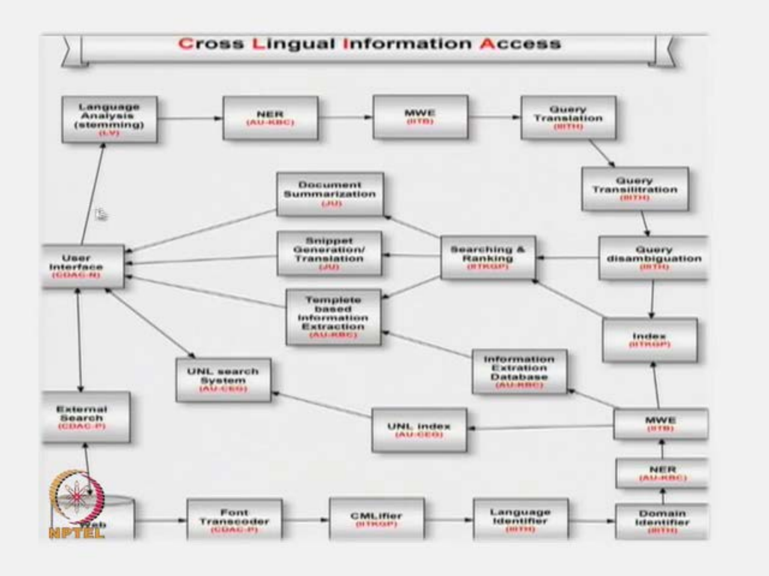


Portal

- **Public portal released** at http://www.clia.iitb.ac.in/clia-beta-ext/ in September 2009. (Outside IITB)
- **Public portal released** at http://www.clia.iitb.ac.in:8080/clia-beta-ext/ in September 2009. (Inside IITB)

The portal can be seen here in this U R L public portal is also released.

(Refer Slide Time: 42:30)



Cross Lingual Information Access

Language Analysis (stemming) (LV)
NER (AU-KBC)
MWE (IITB)
Query Translation (IIITH)
Query Transliteration (IIITH)
Document Summarization (JU)
Snippet Generation/ Translation (JU)
Searching & Ranking (IITKGP)
Query disambiguation (IIITH)
User Interface (CDAC-N)
Templete based Information Extraction (AU-KBC)
Index (IITKGP)
Information Extration Database (AU-KBC)
UNL search System (AU-CEG)
External Search (CDAC-P)
UNL index (AU-CEG)
MWE (IITB)
NER (AU-KBC)
Web
Font Transcoder (CDAC-P)
CMLifier (IITKGP)
Language Identifier (IIITH)
Domain Identifier (IIITH)

Now, the processing of the cross lingual information access system can be understood from these block diagram there is what is called the online processing and the offline processing. So, when a query is presented first it goes through language analysis namely stemming that means getting the root words or the un suffixed words from the sentence then there is named entity recognition identify all the proper nouns. So, below each task the name of the institute is given and Anna University K B C is supposed to implement this named entity recognizer multi word expressions have to be detected.

There, this word always hang together and they have to be accurately processed and documents containing them have to be retrieved query translation is the task of converting the query into the target languages form. Query translation is also is to be done which converts proper names into the target language form query disambiguation is required because the words in a query typically have multiple meanings. The meaning that is most relevant to the current context if the context is available, then disambiguation has to be done indexing has to be done by which one gets a list of words.

Maps them to the documents which contain them, so this is a line of processing on the query and the document. Searching and ranking is done once the query is given of the documents then comes the task of processing the output and presenting. When the document is retrieved U R L will be opened by the user, but before that one should give a short snippet of what is contained in the document this is the snippet generation.

Now, snippet has to be translated because the query was in a different language template based information extraction can be done. For example, if the document is about Jaipur which is retrieved then the user may quickly want to know what are the important places for sightseeing hospitals police station etcetera the document may have to be summarized and translated and shown to the user.

So, these particular path of language analysis N E R, M W E query translation query transliteration query disambiguation indexing index accessing searching document summarization snippet template. Then giving it to the user these are online processing and they are done with respect to the query that is presented to the user.

This particular task has to be supported by a tremendous amount of work offline, so from the web the documents are obtained fonts are Trans coded then something called. C M litigation is done that is intelligent tags sophisticated tags are placed in the document language identification is done for the document domain identification is done. We mentioned tourism named entity marquee was detected from there also and these information is converted into index.

There is another path called UML based search which is advanced search making use of semantic relations we can talk about that later. So, we have discussed today the relationship between information retrieval and NLP and also described the crossed lingual project going on in the country we will continue from here.