Language Technologies for Indian Languages

Pushpak Bhattacharyya
Department of Computer Science and Engineering
IIT Bombay
Perspective
Resource Richness: how many members of the world’s language families have them?

A list of language families

- Indo-European, Dravidian, and minor European languages
- Afro-Asiatic and Caucasian languages
- Nilo-Saharan, Kordofanian, and Khoisan languages
- Niger-Congo languages
- More Niger-Congo languages, including Bantu
- Uralic, and Altaic, and Miao-Yao, and Tai, and Austro-Asiatic, and other Asian languages
- Sino-Tibetan languages
- Austronesian languages
- North American Indian languages - Eskimo, Na-dené, Algic, Keres, Siouan, Caddoan, Iroquoian, Kiowa-Tanoan "Hokan", isolates
- Mesoamerican Indian languages - "Penutian", Uto-Aztecan, Oto-Manguean, Macro-Chibchan, Paezan Yanomaman
- South American Indian languages - "Andean", "Equatorial", Tupi-Cariban, Macro-Otomakoan, Guamo-Chapacuran, Macro-Arawakan, Bora-Witotoan, Macro-Walkurúan, Macro-Panoan, Macro-Ge, isolates
- Indo-Pacific languages
- Australian languages
Living Languages: *how many have richly annotated corpora?*
*Ans: Less than 10*

<table>
<thead>
<tr>
<th>Continent</th>
<th>No of languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>2092</td>
</tr>
<tr>
<td>Americas</td>
<td>1002</td>
</tr>
<tr>
<td>Asia</td>
<td>2269</td>
</tr>
<tr>
<td>Europe</td>
<td>239</td>
</tr>
<tr>
<td>Pacific</td>
<td>1310</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6912</strong></td>
</tr>
</tbody>
</table>
But Rich Language Properties like Morphology? Many languages have them

Rich Annotation

English

French, Chinese, German, Spanish and a few others

Rich Morphology

Many languages

English, Chinese
Roadmap

• Indian Linguistics and NLP Scenario
• NLP activities at IIT Bombay
• Tools and Resources developed at IIT Bombay for language processing
  – Shallow Parsing
  – Machine Translation
  – Wordnets
Indian Linguistic and NLP Scenario
Great Linguistic Diversity

- Major streams
  - Indo European
  - Dravidian
  - Sino Tibetan
  - Austro-Asiatic

- Some languages are ranked within 20 in the world in terms of the populations speaking them
  - Hindi and Urdu: 5th (~500 million)
  - Bangla: 7th
  - Marathi 14th
3 Language Formula

- Every state has to implement
  - Hindi
  - The state language (Marathi, Gujarathi, Bengali etc.)
  - English

- Big translation requirement, e.g., during the financial year ends
Major Language Processing Initiatives

• Mostly from the Government: Ministry of IT, Ministry of Human Resource Development, Department of Science and Technology

• Recently great drive from the industry: NLP efforts with Indian language in focus
  – Google
  – Microsoft
  – IBM Research Lab
  – Yahoo
  – TCS
Technology Development in Indian Languages (TDIL)

- Started by the Ministry of IT in 2000
- 13 resource center across the country
- Responsibility for two languages: one major and one minor
- For example,
  - IIT Bombay: Marathi and Konkani
  - IIT Kanpur: Hindi and Nepali
  - ISI Kolkata: Bangla and Santhaali
  - Anna University: Tamil
Achievements of TDIL

- Localization in most major languages, through the mediation of Center for Development of Advanced Computing (CDAC)
- Office products, True type fonts
- Machine Translation Systems and Supports
- Lexical Resources
Achievements in TDIL: *MT* systems and Supports

- **Transfer based:** *Anglabharati* (IIT Kanpur), *Shakti* (IIIT Hyderabad), Tamil-Hindi system (Anna University KBC, Chennai)
- **Interlingua based:** *Universal Networking Language (UNL)* based (IIT Bombay; part of an UN initiative with 15 other countries)
Achievements in TDIL: Lexical Resources

Wordnets: Hindi and Marathi (IIT Bombay)

- Ontologies: Tamil concept hierarchy (Tanjavur University, AU-KBC)

- Semantically rich lexicons: IIT Kanpur, IIITH, IIT Bombay

- Corpora: Central Institute of Indian Languages (CIIL)

- Web Content: All 13 centers, Gujarathi content is exhaustive and of good quality
Achievements in TDIL: Shallow Processing

- **Part of Speech Tagger:** Hindi (IIITH, IIT Bombay) and Marathi (IIT Bombay)

- **Chunker:** Hindi (IIITH, IIT Bombay), Marathi (IIT Bombay)

- **Parser:** Nobody has started; but PoS tagger and chunker are the precursors
Achievements in TDIL: Signal Level Processing

- **Speech Processing**: CDAC Noida, Tata Institute of Fundamental Research, IIT Madras (Tamil speech processing engine deployed in Chennai Railway Station)

- **OCR**: ISI Kolkata (Bangla and Hindi; further developed and marketed by CDAC), University of Hyderabad (Telegu), Baroda University (Gujarathi)
Recent Initiatives

- **NLP Association of India:** 2 years old: *recently efforts are on making tools and resources freely available on the website of NLPAI*
- **LDC-IL** (like the *Linguistic Data Consortium at UPenn*)
  - Approved by the planning commission
- **National Knowledge Commission:** special drive on translation (human and machine)
Recent Initiatives contd

- India wide advertisement for CFP on *MT, CLIR, Speech and OCR*

- Consortia set up already for IL-IL MT, E-IL MT and CLIA

- SAALP: South Asian Association for Language Processing (formed with SAARC countries)
Industry Scenario: *English*

- How to use NLP to increase the search engine performance (*precision, recall, speed*)

- **Google, Rediff, Yahoo, IRL, Microsoft**: all have search engine, IR, IE R & D projects outsourced from USA and being carried out in India.
Industry Scenario: Indian Language

• English-Hindi MT is regarded as critical

• IBM Research lab has massive English Hindi Parallel Corpora (news domain)
  – Statistical Machine Translation

• Microsoft India at Bangalore has opened a Multilingual Computing Division

• Google and Yahoo India is actively pursuing IL search engine
NLP at IIT Bombay
IIT Bombay’s effort on MT and accessory systems
Centre for Indian Language Technology (CfILT)
The Center

• Part of Computer Science and Engineering Department, IIT Bombay
• Research in language processing
• Resource building
• Linguistics and computer science
Research Staff

- Language: 13
- Computation: 03
- Associated Students: 06-07 each year
- Research Scholars: 05
- Associated faculty: 06

*about 30 associated members at any point of time*
Indian Language Lexical Resources and Tools

• Lexical Resources
  – Verb Knowledge Base
  – Hindi Wordnet
  – Marathi Wordnet

• Tools
  – Hindi POS Tagger
  – Marathi POS Tagger
  – Semantically Relatable Sequences: Intermediary to Semantics
  – Hindi Generation
  – Translation and Search in a QA-Forum
POS Tagging of Indian Languages
Introduction

• POS tagging is the process of identifying lexical category of a word in a sentence on the basis of its context

Input: राम खेल रहा है (raam khel raha haai)
raam play <verb string of continuity>

Output: राम_NNP खेल_VM रहा_VAUX है_VAUX

• Wide applications
  – Machine translation, Information etc.
The Bigger Picture

- Parsing
- Chunking
- POS Tagging
- Morphological Analysis
- Stemming
- Lexical Resources: Wordnet, Dictionaries
POS Tagging for Indian Languages
(Singh, Gupta, Sinha, Bhattacharyya, ACL 2006)

- Accurate POS taggers not available for Indian languages
- Indian languages are morphologically rich. Hence,
  - First step is to analyze the language
  - Tools for harnessing morphological information are needed
- Stochastic Methods cannot be used
  - Non-availability of tagged corpora
Challenges in POS Tagging for Indian Languages

- Large number of word forms making Stochastic techniques expensive and unreliable.
- Difficulty in identifying morpheme boundary as large number of morphemes are fused together.
- Complexity due to -
  - Inter-POS ambiguity
  - Free word ordered structure
  - Complex morphology of Indian Languages
POS Tagger: System Architecture

Input Sentence

Tokeniser

Token (Word)

Lexicon

Suffix Lists

S-R rules

Derivational rules

Spelling variations
Heuristics

Stemmer

[ Stem + Category + Suffix ]

Morpheme Table

Analysis Table

Morphological Analyser

Morpheme Identification

Morpheme Analysis

Tag Generator

[Word+ Disambiguated Category+ Analysis]

Tagset

VG Identification

Rules

Disambiguator

VG Identification

Disambiguation Engine

Disambiguation Rules

Output Sentence

Word Level Phase

Group Level Phase

(Complete Sentence)
Intermediate Tools

- Intermediate tools for initial processing
  - Stemmer
    - Identifies suffixes and stem
  - Morphological Analyser:
    - Analyses suffixes provided by stemmer
    - Provides category and grammatical feature information
Results

• Results for Hindi:
  – Accuracy: 94.24
  – Trained on a corpus of 20,000 words
  – Corpus Domain: News
  – Source: http://www.bbc.co.uk/hindi/

• Results for Marathi:
  – Accuracy: 85.23
  – Trained on a corpus of 10,000 words
  – Corpus Domain: News
  – Source: http://www.e-sakal.com (newspaper)
Application in MT: Hindi generation

Singh, Dalal, Vachani, Bhattacharyya and Damani, MT Summit 07
Schematic of the generation system
Example of output stages

<table>
<thead>
<tr>
<th>Module</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original English Sentence</td>
<td>For this, you contact the farmers of Manchar region or of Khatav taluka</td>
</tr>
<tr>
<td>UNL Expression</td>
<td>See Sentence 4 and Figure 1</td>
</tr>
<tr>
<td>Lexeme Selection</td>
<td>संपक कसान् यह आप क्षेत्र तालुके मंचर खटाव contact farmer this you region taluka manchar khatav</td>
</tr>
<tr>
<td>Case Identification</td>
<td>संपक कसान् यह आप क्षेत्र तालुके मंचर खटाव contact farmer this you region taluka manchar khatav</td>
</tr>
<tr>
<td>Morphology Generation</td>
<td>संपक कीलिए कसान्त यह आप क्षेत्र contact @imperative farmer @pl this you region taluka मंचर खटाव taluka manchar Khatav</td>
</tr>
<tr>
<td>Function Word Insertion</td>
<td>संपक कीलिए कसान्त यह आप क्षेत्र contact farmers this for you region या taluka के मंचर खटाव or taluka of Manchar Khatav</td>
</tr>
<tr>
<td>Syntax Planning</td>
<td>इसके लए आप मंचर क्षेत्र या खटाव This for you manchar region or khatav taluka के कसान्त को संपक कीलिए</td>
</tr>
</tbody>
</table>
Evaluation

- 901 sentences from agricultural domain
- Actual questions and problem statements by farmers
- Translated into English
- Converted into UNL graphs (semi automatic)
- UNL-to-Hindi generation applied
Human evaluators: score the generated sentences

**Fluency of the given translation is:**

(4) Perfect: Good grammar  
(3) Fair: Easy-to-understand but flawed grammar  
(2) Acceptable: Broken - understandable with effort  
(1) Nonsense: Incomprehensible

**Adequacy: How much meaning of the reference sentence is conveyed in the translation:**

(4) All: No loss of meaning  
(3) Most: Most of the meaning is conveyed  
(2) Some: Some of the meaning is conveyed  
(1) None: Hardly any meaning is conveyed
BLEU score computed

- One reference sentence per UNL graph
- More reference sentence creation in progress

Bleu score = 0.41

Correlation with fluency = 0.59
Multilingual Wordnets for Indian Languages
Wordnet work at IIT Bombay

- [http://www.cfilt.iitb.ac.in](http://www.cfilt.iitb.ac.in)
- Follow the design principle(s) of the Princeton Wordnet for English paying particular attention to language specific phenomena (such as complex predicates)
- **Hindi Wordnet**
  - Total Number of Synsets: 23,067
  - Total Number of Unique Words: 48,725
- **Marathi Wordnet**
  - Total Number of Synsets: 11,908
  - Total Number of Unique Words: **18,093**
# Status of other WN's

<table>
<thead>
<tr>
<th></th>
<th>Total Number of Synsets</th>
<th>Total Words</th>
<th>Unique Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordNet (2.1)</td>
<td>117597</td>
<td>155327</td>
<td></td>
</tr>
<tr>
<td>GermaNet (2004)</td>
<td>53312</td>
<td>76563</td>
<td></td>
</tr>
<tr>
<td>Multi Word Net (1.39)</td>
<td>32,700</td>
<td>58,000</td>
<td></td>
</tr>
</tbody>
</table>
HWN and MWN Created Using Different Principles

(Tatsam, i.e., Sanskrit words borrowed as such: very often)

HWN entry:
{peR, vriksh, paadap, drum, taru, viTap, ruuksh, ruukh, adhrip, taruvar} ‘tree’
jaR,tanaa, shaakhaa, tathaa pattiyo se yukt bahuvarshiya vanaspati ‘perennial woody plant having root, stem, branches and leaves’
peR manushya ke lie bahut hi upayogii hai ‘trees are useful to men’

MWN entry:
{jhaaR, vriksh, taruvar, drum, taruu, paadap} ‘tree’
mule, khoR, phaanghaa, pane ityaadiinii yukt asaa vanaspativishesh ‘perennial woody plant having root, stem, branches and leaves’
tii damuun jhaadacyaa saavlit baslii ‘Being tired/exhausted she sat under the shadow of the tree’
Lexico-semantic relations in wordnet

Semantic relations in WordNet:
• Synonymy
• Hypernymy / Hyponymy
• Antonymy
• Meronymy / Holonymy
• Gradation
• Entailment
• Troponymy
झाड, वृक्ष, तरु

RH

ग्लोस

झाडे पयावरण शुद्ध करण्याचे काम करतात।
Cross Part of Speech Linkages
(important for word sense disambiguation)

• **Links between nouns and verbs:**
  - **Ability link** specifies the features inherited by a nominal concept
    - \{machlii, macchii, matsya, miin, maahii\} ‘fish’ → \{tairnaa, pairnaa, paunrnnaa\} ‘swim’
  - **Capability link** specifies features acquired by a nominal concept
    - \{vyakti, maanas\} ‘person’ → \{tairnaa, pairnaa, paunrnnaa\} ‘swim’
  - **Function link** specifies function(s) associated with a nominal concept
    - \{adhyaapak, shikshak\} ‘teacher’ → \{paRhanaa, shikshaa denaa\} ‘teach’

• **Links between nouns and adjectives:** indicate typical properties of a noun
  - \{sher\} ‘tiger’ → \{maansaahaarii\} ‘carnivorous’.

• **Links between morphologically derived forms**
  - \{bhaaratiiyataa\} ‘indianness’ is derived from \{bhaaratiiya\} ‘Indian’ and is linked to it.
Hindi WN: just recently made free
Towards Multilingual Indo-WN

• Through Relation Borrowing (illustrated through HWN and MWN)
• When the meaning is found in both Hindi and Marathi: This is the most common case, since Hindi and Marathi are sister languages
• When the meaning is found in Hindi but not in Marathi: Relation borrowing is not possible
  – For instance, {दादा [daadaa, grandfather], बाबा [baabaa, grandfather], आजा [aajaa, grandfather], दहा [daddaa, grandfather], पितामह [pitaamaha, grandfather], प्रपिता [prapitaa, grandfather]} are words in Hindi for paternal grandfather. There are no equivalents in Marathi.
• When the meaning is not found in Hindi but is found in Marathi: The relations must be set up manually
  – For example, {गुढीपाडवा [gudhipaadva, new year], वर्षप्रतिपदा [varshpratipadaa, new year]} are words in Marathi which do not have any equivalents in Hindi.
Hindi Verb Knowledge Base (HVKB)

Chakrabarti, Sarma and Bhattacharyya, Lexical Resources Engineering
Journal (accepted)

calanaa ‘move’
(icl>act(agt>person))
ve loga dhiire dhiire chal rahe hai. ‘They are moving slowly’.
(gaman karnaa) ‘to move’
Frame:NP1; NP1_NOM
[VINT, VOA, VOA-BACT]
→ caRhanaa ‘climb’
(icl>move(>act)(agt>person)
ve loga dhiire dhiire chaRha rahe hai. ‘They are climbing slowly.’
upar ki or jaanaa ‘to move upwards’
Frame:NP1; NP1_NOM
[VINT, VOA, VOA-BACT]
MT and Cross lingual IR efforts in India
Dictionaries are based on WN

- Every word in every language linked with each other through Hindi Synsets
Proposed Standard

<table>
<thead>
<tr>
<th>Senses</th>
<th>Hindi</th>
<th>Marathi</th>
<th>Bangali</th>
<th>Oriya</th>
<th>Tamil</th>
</tr>
</thead>
<tbody>
<tr>
<td>(W₁, W₂, W₃, W₄, W₅, W₆)</td>
<td>(W₁, W₂, W₃, W₄, W₅, W₆)</td>
<td>(W₁, W₂, W₃)</td>
<td>(W₁, W₂, W₃)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(sun)</td>
<td>(सूर्य, सुरज, भानु, भास्कर, प्रभाकर, दिनकर, अंधुमान, अंधुमाली)</td>
<td>(सूर्य, भानु, दिवाकर, भास्कर, रीव, दितेश, दिनमणी)</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>(cub, lad, laddie, sonny, sonny boy)</td>
<td>(लडका, बालक, बच्चा, छोकर, छोर, छोकरा, लैडा)</td>
<td>(मुलगा, पोर, पार, पोर, पार)</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>(son, boy)</td>
<td>पुत्र, बेटा, लडका, बाल, सुत, बच्चा, नंदन, पून, चिरजीव, चिरजी,</td>
<td>मुलगा, पुत्र, लेक, चिरजीय, तनय,</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>
Studies on wordnets: small world properties

Ramanand, Ukey, Singh and Bhattacharyya, Mapping and Structural Analysis of Multilingual Wordnets, IEEE Data Engineering Bulletin, 30(1), March 2007
Human social graph

• How many links connect you to Adam Gilchrist/Nelson Mandela/Tim Berners-Lee?
• Sociological studies show that the diameter of human social graph is less than 10 (very, very small!)
• The concept of “Six Degrees of Separation”
Distance to Nelson Mandela

N≤10

Me → P2 → P3 → ... → PN-1 → Mandela
Graphs and Measures

• Measures
  – Average Shortest Path Length
  – Clustering Coefficient
  – Degree Distribution
• Random Graphs: low Avg. Shortest Path
• Regular Graphs: high Avg. Shortest Path
• Small World Graphs: low Avg. Shortest Path
Cluster Coefficient

- Measures what fraction of neighbours of a node are related to each other

- **Cluster Coefficient** $C_i$ for a node $i$ (with degree $k_i$) of a directed graph:

$$C_i = \frac{|E(\Gamma_i)|}{2 \times \left(\frac{k_i}{2}\right)}$$

where $\Gamma_i$ is the subgraph made of $i$ and its neighbours, $|E(\Gamma_i)|$ is the number of edges of the subgraph, and $2 \times \left(\frac{k_i}{2}\right)$ is the total number of possible edges in $\Gamma_i$.

- Random Graphs: low CC ($CC << 1$)
- Regular Graphs: high CC ($0.4 <= CC <= 0.7$)
- Small World Graphs: high CC ($0.4 <= CC <= 0.7$)
But why study Small Worlds for NLP?

- Seen in language organization
- Seen in Wordnets!!
Degree Distribution in Wordnets

- Exponents observed:
  - English WordNet (Nouns): -2.063
  - Hindi WordNet: -2.592
  - Marathi WordNet: -2.841
High-degree nodes in Wordnets

- Eng. WordNet (Nouns):
  - (city, metropolis, urban center): 664, (law, jurisprudence): 611
  - (person, individual, someone, somebody, mortal, soul): 400

- Hindi WordNet:
  - (vyaktii, maanas, shaks, shakhs, ba.ndaa (person)): 607
  - (karm, karanii, kaam, kaarya, krtya, kaarvaaii, kaarvaahii (action)): 524

- Marathi WordNet:
  - (vyaktii, maaNus, isama, manushya, paTThaa, paThyaa (person)): 626
  - (karm, krtii, kriyaa, kaam kaarya, krtya (action)): 546
Cluster Coefficient in Wordnets

- Wordnet Avg. Cluster Coefficient:
  - English WN (Nouns): 0.526
  - Hindi WN: 0.268
  - Marathi WN: 0.358
Average Shortest Path Length in Wordnets

- Average Shortest Path Lengths observed:
  - English WordNet (Nouns): 8.878
  - Hindi: 4.378
  - Marathi: 4.255
Concluding Remarks
Resource disadvantage: Tackle the problem at the grass root level: *a suggested program*

- Construct exhaustive and high quality morph analyser, exploiting morphological richness- if the language is endowed with such richness
- Construct excellent POS Tagger
- Construct very good chunker
- Construct good parser
Our Experience in a Multi Lingual Setting

• There is a strong indication that we will need much less annotated corpora for Marathi than for Hindi.

• A simple (almost naïve 😊) example with two tags, say TINF (to infinitive) and GER (gerund)
  
  I need to go (to infinitive)
  going is impossible now (gerund)
Observation 1/3

- Annotation richness: depends on availability of resources- **funds, linguistic expertise, people, time**
- Morphological richness: a language either has it or does not have
- Is a rich body of linguistic work available, at least for morphology?
- If yes (e.g. for Sanskrit: Panini’s Astadhyayi- a masterpiece of morphology work), exploit it to the fullest
Observation 2/3

- A happy situation in the East for many languages
- Rich morphology
- Excellent linguistic tradition
- One could do this:

  \[ MA \rightarrow POS\ Tagger \rightarrow Chunker \rightarrow Parser \]
Observation 3/3

• Lets look around in the world of SNLP algorithms
• Aren’t they clamoring for *more and more features*?
• MEMM, CRF: don’t they say they would benefit from features in large number, judiciously chosen?
• Where will these features come from?
• One source is **Morphology**
Conclusions

• Indian NLP emerging as one of the most active in the world
• Lexical Networks like wordnets are crucial
• So are high accuracy failure resilient tools like POS taggers
• Multilinguality emerging as a norm rather than a fashion
  – Methods needed to tackle the challenges
• Invariances in multilingual computation and resources form an interesting study
URLs

• For resources
  www.cfilt.iitb.ac.in

• For publications
  www.cse.iitb.ac.in/~pb
Merci