Indian Language Lexical Resources

Pushpak Bhattacharyya
Computer Science and Engg. Dept.
Indian Institute of Technology Bombay
The Genesis
Universal Networking Language

Common language for computers to express information written in natural language

• Large International Effort, IIT Bombay became a part in 1996

• Application:

  Electronic language to overcome language barrier

  Information Distribution System
The UNL System

UNL System Components:
- Internet Web Page Contents
- UNL Editor
- UNL Proxy
- UNL Viewer

Language Servers:
- Language Server A
  - UNL <-> Arabic
  - enconverter
  - deconverter
- Language Server B
  - UNL <-> Hindi
  - enconverter
  - deconverter

Languages Supported:
- Arabic
- Hindi
- Italian
- French
Long ago, in the city of Babylon, the people began to build a huge tower, which seemed about to reach the heavens.
Kinds of MT Systems
(point of entry from source to the target text)

- Deep understanding level
- Interlingual level
- Logico-semantic level
- Syntactico-functional level
- Syntagmatic level
- Morpho-syntactic level
- Graphemic level

- Direct translation
- Semi-direct translation
- Multilevel transfer
- Conceptual transfer
- Semantic transfer
- Syntactic transfer (surface)
- Syntactic transfer (deep)
- Ascending transfer
- Descending transfers

- Ontological interlingua
- Semantico-linguistic interlingua
- SPA-structures (semantic & predicate-argument)
- Multilevel description
- F-structures (functional)
- C-structures (constituent)
- Tagged text
- Text
Universal Word

- Vocabulary of the UNL
- A UW represents a concept
  1) Basic UW
  2) Restricted UW

ex) spring(icl>tool)
    spring(icl>season)
Languages Covered

• 6 UN official Languages.
  Arabic, Chinese, English, French, Spanish, Russian
• Other languages
  German, Greek, Hindi, Indonesian, Italian, Japanese, Korean, Mongol, Latvia, Portuguese, Thai
Lexicon

• Format of the Lexicon
   [HW] {ID} “UW” (ATTRIB1, ATTRIB2,….) <FLG,FRE,PRI>;

• Examples:
   [Bird] {} “bird (icl>animal>concrete volitional thing)”
   (N,ANI,SG,CONCRETE)<E,0,0>;
   [try] {} “try(icl>event)” (V,PRES,SIMPL,DISC)<E,0,0>;
   [from] {} “from” (PRE,INI,PLC,TIME,INEFF)<E,0,0>;
   [What] {} “what”(PRON,INTER,WH-WORD)<E,0,0>;}
From the English wordnet: Bird

- The noun bird has 5 senses (first 2 from tagged texts)
- 1. (31) bird -- (warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings)
- 2. (1) bird, fowl -- (the flesh of a bird or fowl (wild or domestic) used as food)
- 3. dame, doll, wench, skirt, chick, bird -- (informal terms for a (young) woman)
- 4. boo, hoot, Bronx cheer, hiss, raspberry, razzing, razz, snort, bird -- (a cry or noise made to express displeasure or contempt)
- 5. shuttlecock, bird, birdie, shuttle -- (badminton equipment consisting of a ball of cork or rubber with a crown of feathers)
- The verb bird has 1 sense (no senses from tagged texts)
- 1. bird, birdwatch -- (watch and study birds in their natural habitat)
Role of Semantic Knowledge

Mary eats noodles with a fork.  \( \text{ins}(\text{eat}(\text{icl}>\text{do}), \text{fork}(\text{icl}>\text{thing})) \)

Mary eats noodles with John.  \( \text{cag}(\text{eat}(\text{icl}>\text{do}), \text{John}) \)

Ram eats noodles with vegetables.  \( \text{cob}(\text{eat}(\text{icl}>\text{do}), \text{vegetable}(@\text{pl})) \)

A demon eats noodles with a goat.  \( \text{cob}?? \) Or \( \text{cag}?? \)

**Semantic Attributes are used to solve such PP-Attachment ambiguities and decide UNL relations.**

Example:

The words in above sentences are classified as follows:

*Fork is classified under inanimate things that can be used as instruments.*

*Mary and John are classified as animate things that are humans and Omnivores.*

*Demon is similarly classified as animate things that are Omnivores.*

*Noodles and vegetables are classified as inanimate things that are edible.*
Ontology for enriching Lexicon

- Add attributes to the lexicon by traversing the DAG

\[
\begin{align*}
\text{people} & \rightarrow \text{community} \\
\text{book} & \rightarrow \text{thing} \\
\text{boy} & \rightarrow \text{human} \\
\text{water} & \rightarrow \text{fluid}
\end{align*}
\]
Current potential of the System

The soldier went away to the totally deserted desert to desert the house in the desert.

[S]
mod(deserted(icl>vacant):11,total(icl>complete):0T)
aoj(deserted(icl>vacant):11,desert(icl>landscape):1A.@def)
plc(go(icl>event):0C.@entry.@past.@pred,away(icl>logical place):0H)
obj(desert(icl>do):1K.@present.@pred,house(icl>place):1V.@def)
plc(desert(icl>do):1K.@present.@pred,desert(icl>landscape):28.@def)
plt(go(icl>event):0C.@entry.@past.@pred,desert(icl>landscape):1A.@def)
pur(go(icl>event):0C.@entry.@past.@pred,desert(icl>do):1K.@present.@pred)
agt(go(icl>event):0C.@entry.@past.@pred,soldier(icl>human):04.@def)
[/S]
Hindi Wordnet at IIT Bombay

Current Team:
Pushpak Bhattacharyya, Prabhakar Pande, Laxmi Kashyap, Devendra Kairwan, Salil Joshi, Arun Karthikeyan, Prachur Goel

and many previous PhD, Masters and Bachelor Students and Research Staff
Great Language Diversity of India

[Map of India showing language distribution]
### Languages and the speaker population

<table>
<thead>
<tr>
<th>Language</th>
<th>Population (2001 census; rounded to most significant digit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindi</td>
<td>450, 000, 000</td>
</tr>
<tr>
<td>Marathi</td>
<td>72, 000, 000</td>
</tr>
<tr>
<td>Konkani</td>
<td>7, 000, 000</td>
</tr>
<tr>
<td>Sanskrit</td>
<td>6000</td>
</tr>
<tr>
<td>Nepali</td>
<td>13, 000, 000</td>
</tr>
</tbody>
</table>
## Languages and the speaker population (contd.)

<table>
<thead>
<tr>
<th>Language</th>
<th>Population (2001 census; rounded to most significant digit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kashmiri</td>
<td>5, 000, 000</td>
</tr>
<tr>
<td>Assamese</td>
<td>13, 000, 000</td>
</tr>
<tr>
<td>Tamil</td>
<td>60, 000, 000</td>
</tr>
<tr>
<td>Malayalam</td>
<td>33, 000, 000</td>
</tr>
<tr>
<td>Bodo</td>
<td>1, 000, 000</td>
</tr>
<tr>
<td>Manipuri</td>
<td>1, 000, 000</td>
</tr>
</tbody>
</table>
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

IIT Bombay Natural Language Processing Group heavily supported by Government and Industry
What is Hindi Wordnet

- Wordnet – A lexical database
- Hindi Wordnet Inspired by the English WordNet
- Built conceptually
- Synsets or the Synonymy Sets are the basic building blocks
- Different organizing principles for different syntactic categories
Example Entry in Hindi Wordnet

• Synset
  \{gaaya, gauu, gaiyaa, dhenu\}, Cow

• Gloss
  – Text definition
    (siingwaalaa eka shaakaahaarii maadaa choupaayaa)  
    (a horny, herbivorous, four-legged female animal)

  – Example sentence
    (hinduu loga gaaya ko go maataa kahate hain evam usakii puujaa karate hain)  
    (The Hindus considers cow as mother and worship it.)
Relations in Wordnet

- Synonymy
- Hypernymy / Hyponymy
- Antonymy
- Meronymy / Holonymy
- Gradation
- Entailment
- Troponymy
(gaaya , gauu) Cow

Four-legged animal

Hypernym

(shaakaahaarii) herbivorous

Attribute

(chaupaayaa, pashu) Four-legged animal

Hyponym

kaamadhenu A kind of cow

Hyponym

mainii gaaya A kind of cow

Antonym

(baila) Ox

Ability Verb

Gloss

(puunchh ) Tail

Meronym

(shaakaahaarii) herbivorous

A horny, herbivorous, four-legged female animal

Gloss

(paguraanaa) ruminate

Attribute

(thaana) udder

Meronym
## Statistics

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synsets</td>
<td>33500</td>
</tr>
<tr>
<td>Unique Words</td>
<td>80400</td>
</tr>
<tr>
<td>Related Synsets</td>
<td>33500</td>
</tr>
<tr>
<td>Hindi-English Linked Synsets</td>
<td>13000</td>
</tr>
<tr>
<td>Hits</td>
<td>260000</td>
</tr>
</tbody>
</table>

IIT Bombay
Impact, Use and Visibility of Hindi Wordnet

- Free download with API under GPL
- Available from LDC (linguistics data consortium), Upenn: topmost linguistic data repository in the worlds
- Commercial license purchased by Google for work on Indian language search engine
- Discussions on with Microsoft, Inforsys
- To be available from ELRA: language data repository of Europe
- Available from LDC-IL: LDC of India
Impact, Use and Visibility of created resources (continued)

- Daily reference form all over the world
- More than 2 Lakh hits so far since 2006
- More than 3000 downloads
- Pivot for wordnets of many Indian languages
- Base resource used by many researchers for IL work on translation, summarization, cross lingual search
Hindi Wordnet giving rise to other Indian Language wordnets
Linked wordnets

- Immense Lexical Resource
- Great benefits to machine translation, cross-lingual search
- Very useful for language teaching, pedagogy, comparative linguistics
- Akin to Eurowordnet, but critical differences due to typical Indian language characteristics
<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>((W_1, W_2, W_3, W_4, W_5, W_6))</td>
<td>((W_1, W_2, W_3, W_4, W_5, W_6))</td>
<td>((W_1, W_2, W_3))</td>
<td>((W_1, W_2, W_3))</td>
<td>((W_1, W_2, W_3))</td>
<td>((W_1, W_2, W_3))</td>
</tr>
<tr>
<td>(sun)</td>
<td>(सूर्य, सूरज, भाृन, भासुकर, परशाकर, दनिकर, अंधुमान, अंधुमाली)</td>
<td>(सूर्य, भाृन, दनिकर, अंधुकर, रवि, दनिश, दनिमणी)</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>(cub, lad, laddie, sonny, sonny boy)</td>
<td>( , , , , )</td>
<td>( , , )</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>(son, boy)</td>
<td>( , , , , , )</td>
<td>,</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
International Global Wordnet Conference, Jan 31-Feb 4, 10

A major International Event Granted to IIT Bombay Because of The success Of Hindi Wordnet
Conclusions

- It is important to represent words accurately and with rich information.
- Multifarious fundamental applications:
  1. Information Extraction
  2. Machine Translation on the internet
  3. Text Mining
  4. Cross Lingual Search

(PTO.)
Cross Lingual Search Portal

- www.clia.iitb.ac.in
- तरुणिती यात्रा
- Multiinstitute Nation wide consortium project (IITs, ISI, Jadavpur, Anna University, CDAC)
निरुपती यात्रा

शोध : ो साइटसाडी ओविश्वसनीय जाल

निकाल 1 ते 2 एकल 2 शोधा -- निरुपती यात्रा

दक्षिणी ४५४ | मनोगत
देवस्थान संस्थान मालमते तिरुपती पूजाही श्रीमंत आहे बदूधां!
"://www.manogat.com/node/13799 (सावतलेले) (सारांश)

विद्यादाहारी पर्यटक भवानी मंदिर
रूपती, शिर्कूत भाविकाला अगदी माफक दरात पोटेंट्र प्रसद दिला जातो।
"://bela.esakal.com/2009/01/28235429/mahalaxmi-temple--devotees-is.html (सावतलेले) (सारांश)

हिंदी | इंग्लिश | मराठी बंगाली | English
rupati

rupati Home tirupati Holy Theerthams tirupati Other Places tirupati Festivals tirupati Darshan Timings tirupati Hotel tirupati Restaurant tirupati...

http://www.tirupati.ind.in/ (साध्वदेवें) (सारांश)

rupati

tirupati festivals This Vaishnavite shrine is agog with excitement and activity during festivals such as Vaikunta Ekadashi, ... tirupati means 'Lord...