Brief Introduction to CFILT

- Natural Language Processing @IIT Bombay started in 1996
- Work started with support from United Nations University, Tokyo for Universal Networking Language
- The Center was established in 2000
- Many faculty members & Ph.D, M.Tech, B.Tech students and linguists associated with the lab
Prof. Pushpak Bhattacharyya
Vijay and Sita Vashee Chair Professor
(pb@cse.iitb.ac.in)

www.cfilt.iitb.ac.in
Publications: www.cse.iitb.ac.in/~pb

People
Associated faculty:
3 CSE + 1 HSS

Students:
PhD: Graduated-14; Ongoing-17
MTech (so far): 120+
Btech (so far): 60+
Linguists & staff: ~13

Research & outreach
Publications in top NLP & AI conferences: ACL, NAACL, AAAI, EMNLP, COLING, WWW, ECML
Organizing major international conferences (COLING 2012)

Collaboration
Sponsorship: Ministry of IT, DST, Yahoo, IBM, Microsoft, Xerox, AOL, United Nations, Elsevier, Accenture, TCS
Associations with universities
(Copenhagen, Grenoble, Kyoto etc.)
Collaborations with many Indian universities
Projects

• Consortium of Indian universities, R&D units
  – Cross Lingual Information Access

• Elsevier
  – Text Mining, Information Extraction, Document Clustering

• Tata Consultancy Services:
  – Next Generation Information Extraction, Hybrid Statistical & Semantic Information Extraction

• Indian Language Corpora Initiative:
  – Consortium of Indian universities, R&D units

• EZDI Technologies
  – Health Technologies

• Accenture
  – Deep Semantics Question Answering
NLP: At the confluence of linguistics & computer science

Lexicon

Morphology
Morphology analyzer

Syntactics
Parser

Semantics
Word Sense Disambiguation

Sentiment Analysis
Information Retrieval

Summarization

Machine Translation

Graphs & trees
Finite-state machines

Probability theory
Machine learning

Ontology generation
Parsing in compilation

Linguistics

Computer Science

Linguistics is the eye and computation the body
Multilinguality is a key theme

- **5+1 language families**
  - Indo-Aryan (74% population)
  - Dravidian (24%)
  - Austro-Asiatic (1.2%)
  - Tibeto-Burman (0.6%)
  - Andaman languages (2 families?)
  - + English (West-Germanic)

- **22 scheduled languages**

- **11 languages with more than 25 million speakers**
  - 29 languages with more than 1 million speakers
  - Only India has 2 languages (+English) in the world’s 10 most spoken languages
  - 7-8 Indian languages in the top 20 most spoken languages
Key features of Indian languages

- **Word order: Subject-Object-Verb**

  हम ओसाका से क्योटो तक ट्रेन में आये (hindi)
  We came from Osaka to Kyoto in a train

- **Morphologically rich**

  आम्ही ओसाकापासून क्योटोपर्यंत ट्रेनमध्ये आलो (marathi)
  We came from Osaka to Kyoto in a train
Key Research Areas

- Machine Translation
- Sentiment Analysis
- Information Retrieval
- Lexical Semantics
- Information Extraction
- Cognitive NLP
Machine Translation
MT@IITB: Overview

Translation among Indian languages

- English → Indian languages
- Indian languages → English
- Between Indian languages

Paradigms

- Interlingua-based MT
- Transfer-based MT
- Statistical MT
Statistical MT (1)

● Phrase-based SMT: Incorporating linguistic knowledge
  ○ Source Reordering: En-IL, IL-En, various representations (IJCNLP’08)
  ○ Factor-based: Dependency parse information for generating case markers correctly (ACL’09)
  ○ Handling morphologically rich languages: unsupervised segmentation (ICON’14)
  ○ Post-ordering: Mainly for IL-En translation (ICON’15)

● Translation & Transliteration among related languages: Scaling Statistical MT systems to a large number of languages with high accuracy and less resources
  ○ Relatedness of languages
  ○ Comparative study of pan-Indian translation (LREC’14)
  ○ Reuse of resources, leveraging similarities (LREC’14, ICON’14, NAACL’15)
  ○ Unsupervised transliteration and translation (NAACL’16-under review)
Statistical MT (2)

- **Pivot-based SMT:** *Addressing language divergence issues*
  - Multiple assisting languages (NAACL’15)
  - Addressing word order (ICON’15)
  - Addressing morphological richness (ICON’15)
  - Combining character-based and phrase-based SMT

- **MT Evaluation:** *Incorporate semantics and address rich morphology*
  - Analysis of BLEU (ICON’07)
  - METEOR for Indic languages (LREC’14)
  - Textual entailment for evaluation (WMT’14)

- **Crowdsourcing:** *Exploring quality control issues*
  - Translation & transliteration resources with crowdsourcing (LREC’14)
  - Translation crowdsourcing pipeline (ACL’13)

*Shata-Anuvaadak MT System:* [http://www.cfilt.iitb.ac.in/indic-translator/](http://www.cfilt.iitb.ac.in/indic-translator/)
UNL-based English ➔ Hindi Translation System

(JMT, 2001)
Indian Language MT Project (ILMT)

- Translation between Indian languages
- Transfer based MT system
- Every language vertical develops analyzers and synthesizers
- Analysis up to shallow parsing
- Morphological analysis has an important role

Sampark MT system

http://sampark.iiit.ac.in/sampark/web/index.php/content
Sampark Architecture
Lexical Semantics
IndoWordNet

(LREC 2010, GWC 2002, GWC 2010)

• Linked lexical knowledge base of wordnets of various Indian languages
• Each wordnet is composed of synsets and semantic relations
• It covers 17 Indian languages linked to English WordNet
• Built using expansion approach
• Upto 40k synsets per language

• IndoWordNet: http://www.cfilt.iitb.ac.in/indowordnet/
• Hindi: http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php
• Marathi: http://www.cfilt.iitb.ac.in/wordnet/webmwn/wn.php
• Sanskrit: http://www.cfilt.iitb.ac.in/wordnet/webswn/wn.php
Activities related to IndoWordNet

Data Creation

• Hindi-English synset mapping
• Sense-annotated corpus creation
• Bilingual dictionary creation
• Synset Linking
• Synset Ranking
• Mapping images with synsets

Tools

• Developing WordNet related tools
• Semi-automatic expansion of wordnets
• Developing mobile applications and browser extensions
Word Sense Disambiguation

• Unsupervised approaches (*IJCNLP 2011, ACL 2013*)
  • Bilingual WSD using EM algorithm
  • Resource deprived languages help each other (*ACL 2011*)

• WSD using Word Embedding (*NAACL 2015*)
  • Word embedding of a word is compared with sense embeddings to get the predominant sense of word
  • Automatic synset ranking can be done by using the same approach
Multilinguality & Lexical Knowledge Network

*(ACL 2013, GWC 2014)*

- Design of Multilingual lexical resource
  - Linking of lexicons of 23 languages, Universal Word Dictionary, SUMO upper ontology
  - Revised Lexical Markup Framework to represent the resource

- Enrichment of Multilingual lexical resource
  - Unsupervised learning methods to automatically extract multilingual lexical taxonomies
  - Automatic domain assignment to lexical entries
# Enriching & Creating NLP resources using Deep Learning

## Enriching existing resources

- Automatic linking of synsets
  - Within a language specific wordnet
  - Cross-lingual
- Refining pretrained vector repositories
  - Detection and removal of non-specific vectors
  - Estimating task specific approximate representation for out-of-vocabulary words

## Creating new resources

- Creating vector representations of complex lexical entities such as
  - Synsets
  - Phrases
  - Sentences
  - Question/Answer pairs
- Investigating compositional and non-compositional methods of creating vectors
Information Retrieval
Crawled and Indexed Web Pages

Target Language Index in English

Hindi Query

CLIR Engine

Language Resources

Target Information in English

Result Snippets in Hindi

Ranked List of Results

Supports 9 languages: Hindi, Marathi, Punjabi, Oriya, Bengali, Tamil, Telugu, Gujarati and Assamese
Cross Lingual Search for Indian Languages

**Query Expansion**
Multilingual Pseudo-relevance feedback (*ACL 2010, SIGIR 2010*)
Structure Cognizant PRF (*IJCNLP 2013*)

**Query Transliteration**
Character Sequence Modelling (*TALIP 2010*)
Using Orthographic syllables of Indic scripts

**Crawling**
Conservative focussed crawling under resource constraints (*ICON 2015*)
Information Extraction
Indian language IE tools
- resource constraints
- multilinguality
POS, NER, Chunkers
*(ACL 2006, COLING 2010)*

Relation Extraction

Co-reference resolution
*(RANLP 2015)*

Textual Entailment
*(ICON 2013)*

Noun Compound Interpretation

Making sense of data
Multilingual Named Entity Recognition Using Deep Learning

*(CICLING 2016)*

- Deep Learning techniques do Feature Learning
- Word embeddings combined with Deep Learning have given comparable results with existing state-of-the-art feature engineered systems
- Use Deep Learning to learn language independent features
- Named Entities should have a common representation across languages
Sri Ragam is the asampoorna mela equivalent of K Priya acc to MD’s school. Thyagaraja gave life to K.Priya with his excellent compos, where as MD never touched this raga. In Sri ragam we have plenty of compos by the trinity incl the famous Endaro Sri Ranjani is a lovely janya of K Priya with plenty of compos by both T & MD.

**Nature of Text**

- Content related to **specific domain** of Indian classical music.
- Contain **short discourse of text and is noisy**

**Feature Engineering**

- Explore features specific to noisy text
- Dependency parse based features found more useful for noisy text
Hybrid Approach for Coreference resolution *(RANLP 2015, CICLING 2016)*

- Hybrid architecture for co-reference resolution (rule-based + machine learning)
- Hand-engineered Bayesian Network based model for small training sets
Relation Extraction

*(ICON 2013, CICLING 2016)*

- End-to-end relation extraction
  - Identifying *entity mentions* along with their types
  - Recognizing *semantic relations* among entity mention pairs
- Joint learning and inference of the two problems
- Markov Logic Networks for joint inference
Relation Extraction from medical text

Extract relation between medical entities (medical device, procedure, problems, drugs) from given clinical documents

- **otic drops** for an **ear discomfort**, for the next five days, two drops q.i.d.,
- **norco** 10/325 mg two q. 4 h p.r.n. for **pain** while awake.

- **otic drops** medicine_for **ear discomfort**
- **norco** medicine_for **pain**

- Exploring rich feature design using syntactic and dependency information
- Explore representation learning with convolutional neural networks
Noun Compound Interpretation

- **Noun compound**: “sequence of two or more nouns that act as a single noun”
  - **Example**: apple pie, student protest, colon cancer, colon cancer symptoms, etc.
- **Interpretation**: “identifying relations between nouns in a noun compound.”
  - **Labeling** “apple pie” Made-Of
  - **Paraphrasing** “apple pie”: “a pie made of apple”, or “a pie with apple flavor”
- **Motivation**: (Translation)
  - ENG: “Honey Singh became the latest victim of celebrity death hoax.”
  - HIN: “हनी सिंह प्रसिद्ध व्यक्ति की मौत के बारे में अफवाह के ताजा शिकार बने।”
- **Problem**:
  - “Given a noun+noun compound, assign an abstract label (relationship between two nouns)”
  - Set of abstract relations are defined by Tratz and Hovy (2010).
- **Challenges**: Highly productive, no clue from the context, and pragmatic influence
Sentiment Analysis
SA Research @ IITB

**Lexicon Generation**
1) Augment polarity to Wordnet adjectives
2) Creation of the earliest Wordnet based sentiment lexicon for Indian language
3) A lexicon that rates words with a synset differently

**Statistical Approaches**
1) Classifiers that use word senses as features instead of words
2) Using word senses to bridge cross-lingual gap
3) Hybrid approaches for cross-domain SA

**Special challenges**
1) Thwarting is when a part of sentence reverses the polarity of majority of preceding portion
2) Sarcasm is the use of words of one polarity to imply another
Detecting Granularity in Words: for the Purpose of Sentiment Analysis

- Many hidden properties of words other than being positive or negative which can lead to enrichment of existing sentiment analysis systems.
- Identifying these properties in polar words for different applications in sentiment analysis.

<table>
<thead>
<tr>
<th>Polar Word</th>
<th>Properties</th>
<th>Applications in SA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Domain dependence for polarity</td>
<td>In-domain SA</td>
</tr>
<tr>
<td></td>
<td>Domain dependence for significance</td>
<td>Cross-domain SA</td>
</tr>
<tr>
<td></td>
<td>Intensity within a semantic category</td>
<td>Star-rating Prediction</td>
</tr>
<tr>
<td></td>
<td>Intensity within a sense</td>
<td>Intensity in SentiWordNet</td>
</tr>
</tbody>
</table>

*(IJCNLP 2013, EMNLP 2015)*
**Definition:** Computational approaches to sarcasm

*This phone is awesome. Use it as a paperweight.*

*I loooovvvee Nicki Minaj!*
Emotion Analysis from Text

Hierarchical classification for emotion analysis

Leverage hierarchy of relations between emotion labels to improve emotion analysis using Hierarchical Naive Bayes

Emotion Analysis in Narratives and Discourses

Model as a sequence labelling problem
Sentiment Analysis and Deep Learning

- Models explored for sentiment analysis:
  - Convolutional Neural Networks (CNN)
  - Long Short-term Memory (LSTM) networks

- For sentiment classification tasks like:
  - Positive/negative/neutral sentiment detection
  - Aspect Classification

- On different types of data like:
  - Movie Reviews in languages like English and Hindi
  - Social Media texts like tweets
Cognitive NLP
Cognitive NLP

Linguistic tasks: Translation, SA, Sarcasm Detection, WSD

Understanding the phenomenon by analyzing Cognitive Information through **Eye-tracking/EEG/Key-logging**

Translating the insights into better NLP/NLU algorithms/systems.

Labeling data with cognitive information for supervised learning

Better feature-engineering for NLP systems

http://www.cfilt.iitb.ac.in/cognitive-nlp/
Some problems being investigated

- Labeling data with cognitive information for supervised learning
- Collecting cognitive information during text Annotation
- Predicting Translation Complexity (ACL 2013)
- Measuring Sentiment Annotation Complexity (ACL 2014)
- Subjectivity Extraction for Sentiment Analysis (WASSA 2014)
- Predicting Readers’ Sarcasm Understandability (AAAI 2016)
- Cognitive Information Used for Feature Engineering
Education Technology
Automatic Essay Grading

(QATS 2016)

Score various aspects of the essay, like language complexity, word usage, organization, coherence, etc. to generate an overall score to check the overall quality of the essay.

- Text complexity calculation and its effect on quality of the essay.
- Extraction of words / phrases and estimating their contribution to the quality of an essay.
- Eye-tracking to evaluate organization, coherence and cohesion of the essay.
Automated Grammatical Error Correction

- Addressing Class Imbalance in grammatical error correction *(ICON 2015)*
- Adapting Methods in Machine Translation to grammar correction *(CoNLL 2014)*
- Addressing Subject-Verb Agreement errors *(CoNLL 2013)*