Natural Language Processing

Diptesh Kanojia, Pushpak Bhattacharyya
IITB-Monash Research Academy, IIT Bombay, Monash University
Center For Indian Language Technology (CFILT)

Original Slide Credits: Prof. Pushpak

Prof. Pushpak Bhattacharyya

Professor, IIT Bombay (on lien)

Director, IIT Patna

www.cfilt.iitb.ac.in, www.cse.iitb.ac.in/~pb

People

Associated faculty:
3 CSE + 1 HSS
Students:
PhD : Graduated-22; Ongoing-13
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

Natural Language Processing (NLP)

NLP is a way for computers to analyze, understand, and derive meaning from human language in a smart and useful way.

By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.

It lies under the purview of Artificial Intelligence (AI) which is an area of study in Computer Science.

AI is so ‘everywhere’!

How does it relate to "Data Science" ?

NLP is at the crux of data science and they are related because both of them utilize Machine/Deep Learning algorithms for specific purposes.

NLP = building systems that can understand language $\subsetneq$ AI

ML/DL = building systems that can learn from experience $\subsetneq$ AI

NLP $\cap$ ML/DL = building systems that can learn how to understand language.

How can it be used ?

NLP algorithms are typically based on machine learning algorithms.

Instead of hand-coding large sets of rules, NLP can rely on machine learning to automatically learn these rules by analyzing a set of examples (i.e. a large corpus, like a book, down to a collection of sentences), and making a statical inference.

As a general rule, the more data analyzed, the more accurate the model will be.

But, “Overfitting be bad!”

NLP Applications and Related Sub-Areas

Machine Translation

Information Retrieval

Sentiment Analysis

Fighting Spam

Information Extraction

Text Summarization

Question Answering

Sarcasm Detection

Noun Compound Interpretation

Cognitive NLP

Natural Language Generation

Computational Phylogenetics

Sense Disambiguation

Explainability of Neural Networks

WordNets

Essay Grading

Cognate Detection

Textual Entailment

Emoji Analysis

Speech Recognition

Text-to-Speech

NLP Applications and Related Sub-Areas

Machine Translation

Information Retrieval

Sentiment Analysis

Fighting Spam

Information Extraction

Text Summarization

Question Answering

Sarcasm Detection

Noun Compound Interpretation

Cognitive NLP

Natural Language Generation

Computational Phylogenetics

Sense Disambiguation

Explainability of Neural Networks

WordNets

Essay Grading

Cognate Detection

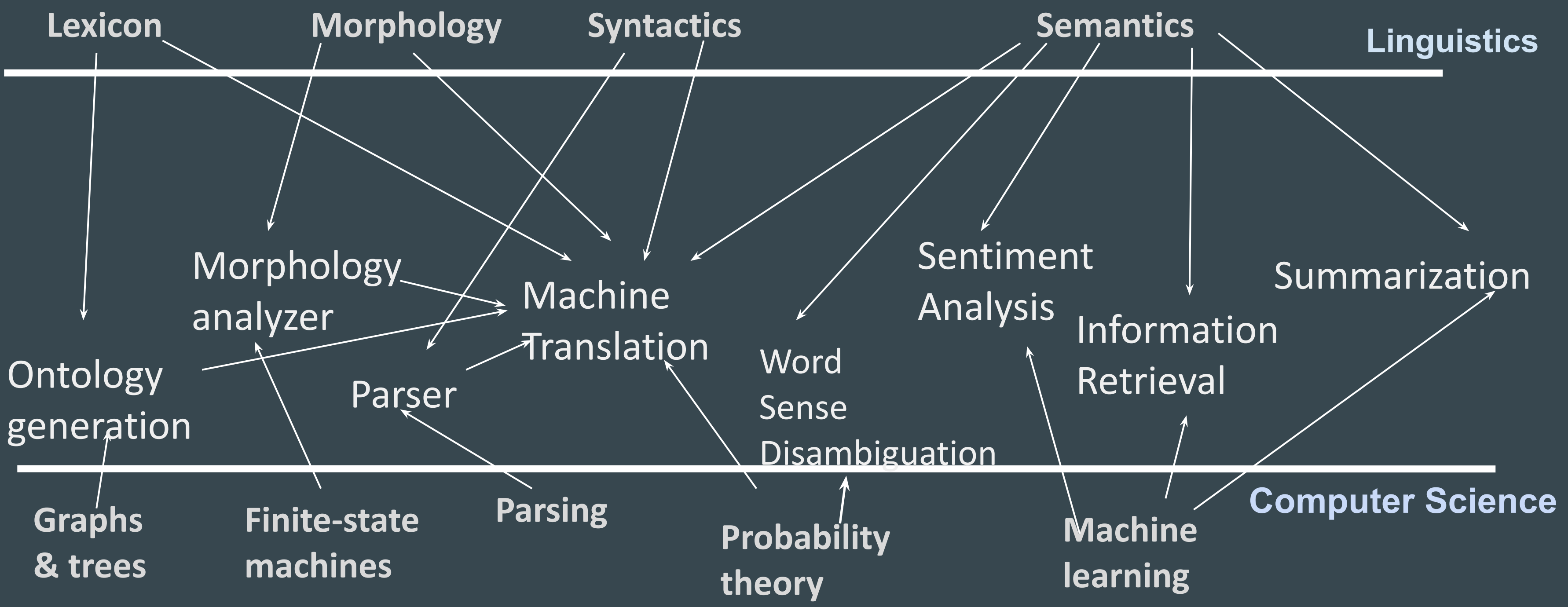
Textual Entailment

Emoji Analysis

Speech Recognition

Text-to-Speech

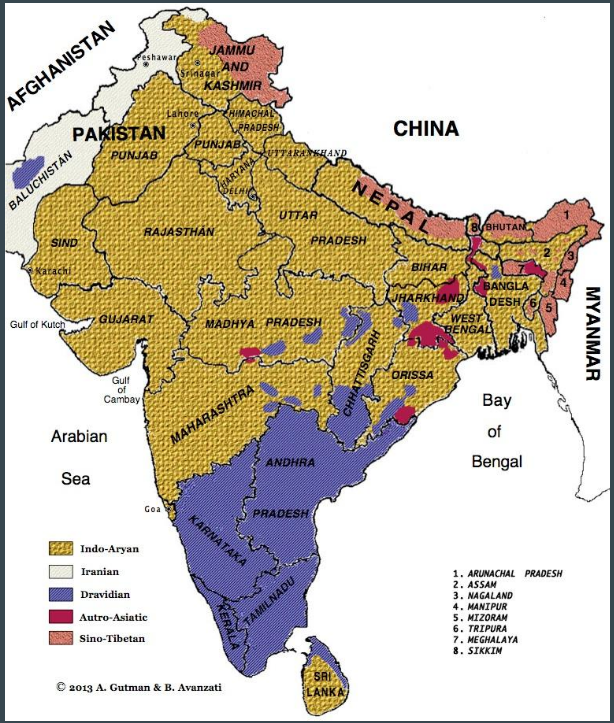
NLP: At the confluence of linguistics & computer science



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 language



Key features of Indian languages

- Word order: Subject-Object-Verb

  हम ओसाका से क्योटो तक ट्रैन मे आये (Hindi)

  we osaka+from kyoto+to train+in came

  We came from Osaka to Kyoto in a train

- Morphologically rich

  आम्ही ओसाकापासून क्योटोपर्यंत ट्रैनमध्ये आलो (Marathi)

  we osaka+from kyoto+to train+in came

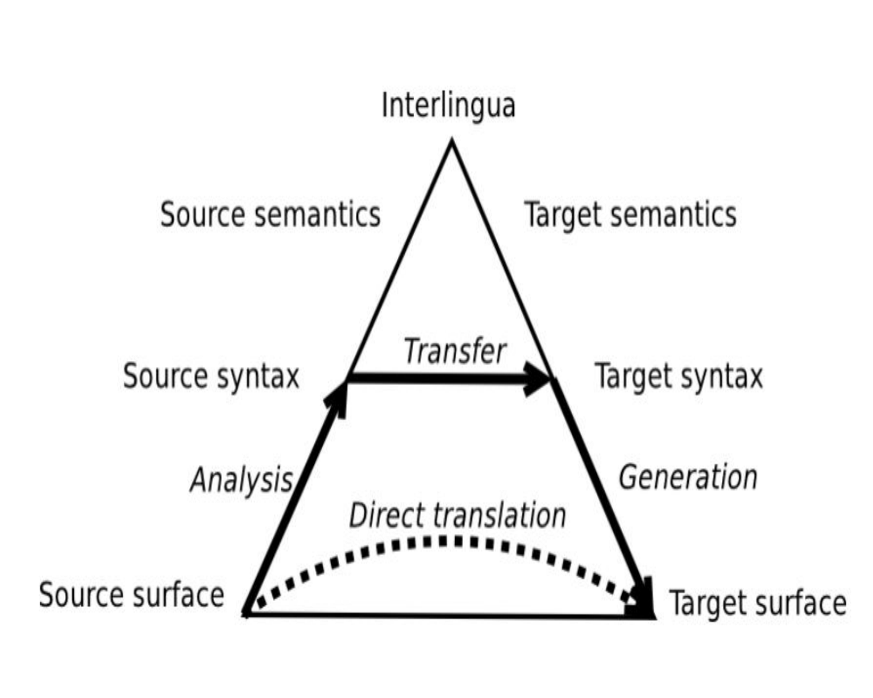
Key Research Areas

- Machine Translation
- Sentiment Analysis
- Information Retrieval
- Lexical Semantics
- Information Extraction
- Cognitive NLP

Machine Translation

Machine Translation : An Overview

- Machine Translation (MT) among Indian languages
    - English → Indian Languages
    - Indian Languages → English
    - Between Indian Languages
- Paradigms
    - Statistical & Neural MT
    - Interlingua-based MT
    - Transfer-based MT



Statistical and Neural MT

- Translation & Transliteration among related languages:
  Scaling Statistical MT systems to a large number of languages with high accuracy and less resources
  - Relatedness of languages and its utility to SMT (NAACL 2016 Tutorial)
  - Investigation of subword units of translation: Orthographic Syllable and BPE (EMNLP 2016, VarDial 2016, IJCNLP 2017, SCLeM 2017/2018)
  - Comparative study of pan-Indian translation (LREC'14)
  - Reuse of resources, leveraging similarities (LREC'14, ICON'14, NAACL'15)
  - Unsupervised transliteration using phonetic & contextual information (CoNLL 2016)

Statistical and Neural MT: a bit more

- Exploring Multilingual learning in Neural MT paradigm
    - Multilingual transliteration and translation between related languages
    - Pivot Translation (IJCNLP 2017, ICON'14)
- 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)
    - Role of Morphology Injection in SMT: A case study for Indian Languages (TALLIP 2017)

Statistical and Neural MT: a ‘byte’ more

- Pivot-based SMT: Addressing language divergence issues
  - Multiple assisting languages (NAACL’15)
  - Addressing word order & morphological richness (ICON’15)
- 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)

Spoken Language Translation

- Imagine Donald Trump calling Kim Jong Un, Trump speaks in English and Kim Jong Un speaks in Korean

- Uses two broad areas :
  - **ASR** - Automatic Speech recognition
  - **MT** - Machine Translation

- **Aim** :
  ASR techniques - speech to text
  MT techniques - text to text
  TTS technique - text to speech

Spoken Language Translation: NMT Step

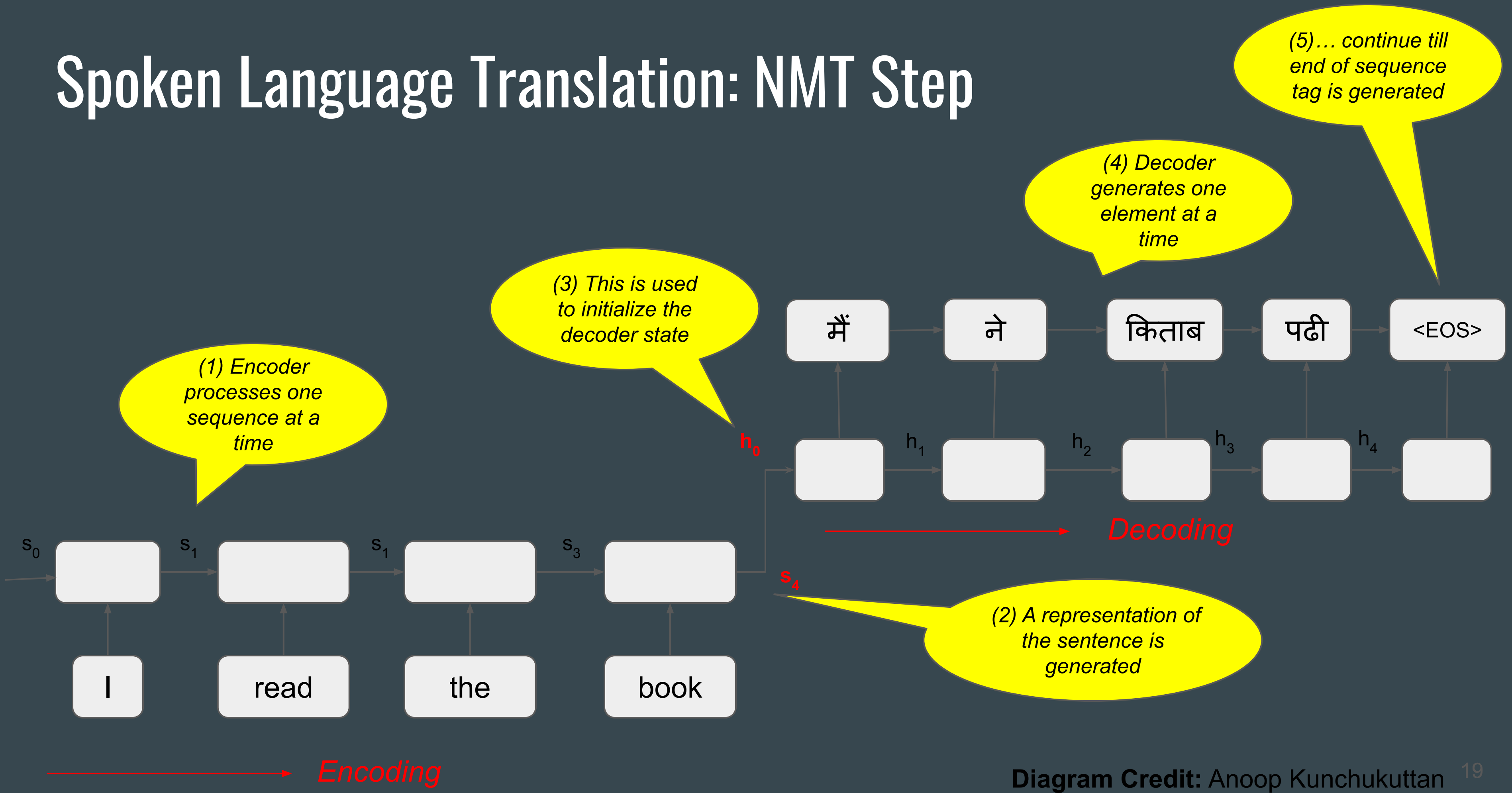
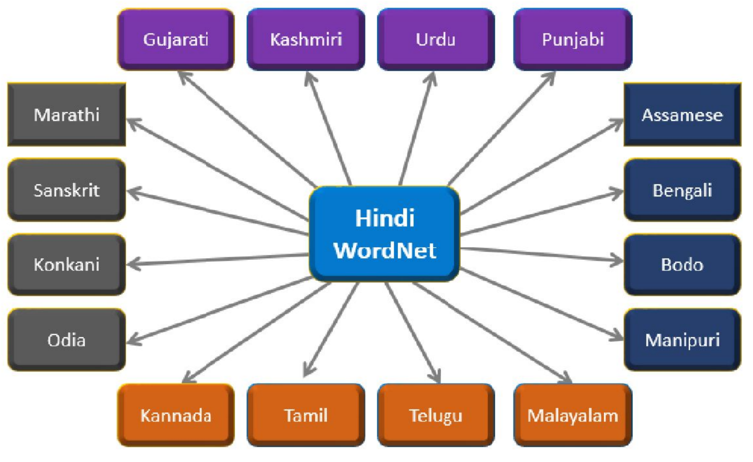


Diagram Credit: Anoop Kunchukuttan

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 Structure

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 Embeddings (NAACL 2015, GWC 2018)
  - Word embedding of a word is compared with sense embeddings to get the predominant sense of word
  - One can use the deep neural networks based embeddings to come up with the predominant sense.
  - Automatic synset ranking can be done by using the same approach

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

Lower Bounds on Dimensions of Word Embeddings

(IJCNLP 2017)

- Usual range for number of word embedding dimensions : 50 - 300
- Many smartphone companies want to build an app which can internally use word embeddings
- Memory limit for apps often in MBs
- Natural thought process: decrease dimensions
  - To what value? 100? 50? 20?
- Depends on the entities we want to place in the space and the corpus

Sentiment Analysis

Sentiment Analysis: An Overview

Lexicon Generation

- Augment polarity to Wordnet adjectives
- Creation of the earliest Wordnet based sentiment lexicon for Indian language
- A lexicon that rates words with a synset differently

Statistical Approaches

- Classifiers that use word senses as features instead of words
- Using word senses to bridge cross-lingual gap
- Hybrid approaches for cross-domain SA

Special challenges

- Thwarting is when a part of sentence reverses the polarity of majority of preceding portion
- Sarcasm is the use of words of one polarity to imply another

Computational Sarcasm

Definition: Computational approaches to sarcasm

‘This phone is awesome. Use it as a paperweight.’ OR ‘I loooovvvee Nicki Minaj!’

- Computational Sarcasm
  - Sarcasm Generation
    - An open-source chatbot that responds sarcastically
    - An emotion tracking engine
  - Sarcasm Detection
    - Detection using incongruity within text
    - Detection using author’s historical text
  - Sarcasm Studies in Humans
    - Sentiment understanding using eye-tracking
    - Sarcasm understanding using eye-tracking

https://github.com/adityajo/ComputationalSarcasm

Sarcasm Suite

Home | Refresh | About Sarcasm Suite | Sarcasm Generator | Sarcasm Detection

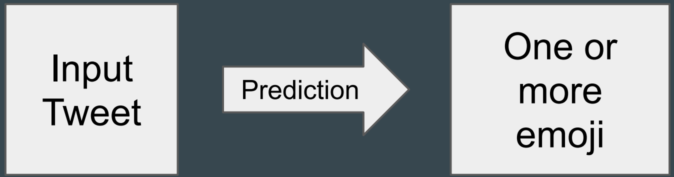
This portal is a demonstration of all projects created as a part of PhD Research by: Aditya Joshi.
Aditya is advised by Dr. Pushpak Bhattacharyya at IIT Bombay, India and Dr. Mark Carman at Monash University, Australia, as a part of the PhD programme at IITB-Monash Research Academy.

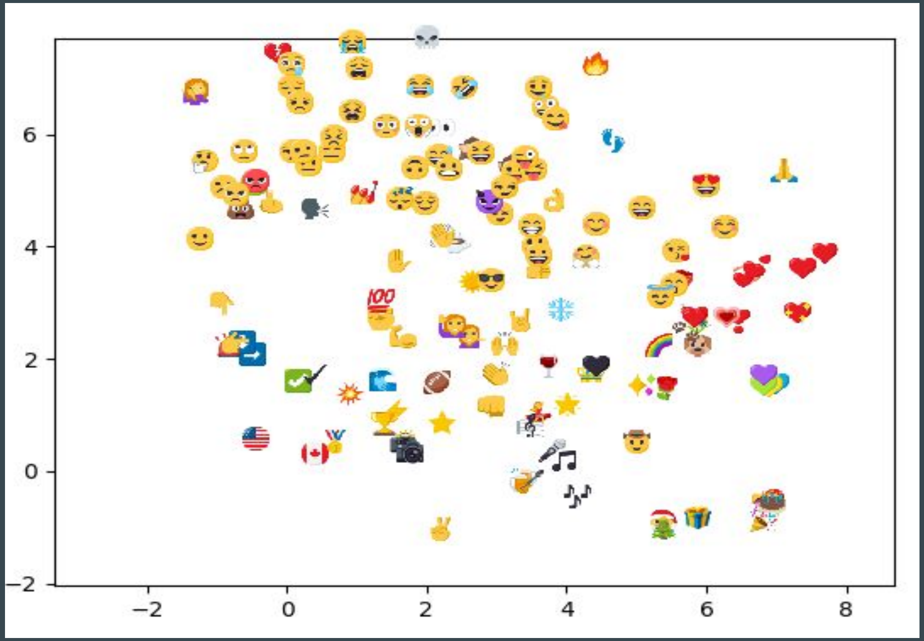
This demonstration was created by Diptesh Kanojia and Aditya Joshi, as a demonstration submission to AAAI 2017.
Please cite appropriate papers, if you use derivates of this portal, including but not limited to screenshots or outputs.

http://www.cfilt.iitb.ac.in/sarcasm-suite/

An Automatic Emoji Recommendation System

- The objective of our automatic recommendation system is to predict one or more relevant emojis for a given input tweet





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/

Investigated Problems in Cognitive NLP

Labeling data with cognitive information for supervised learning

Cognitive Information Used for Feature Engineering

Collecting cognitive information during text Annotation

- Predicting Translation Complexity (ACL 2013)
- Measuring Sentiment Annotation Complexity (ACL 2014)
- Scanpath Complexity: Personalizing Reading/Annotation Effort (AAAI 2017)

- Subjectivity Extraction for Sentiment Analysis (WASSA 2014)
- Predicting Readers' Sarcasm Understandability (AAAI 2016)
- Leveraging Cognitive Features for Sentiment Analysis and Sarcasm Detection (ACL-CONLL 2016)

Information Extraction

Coreference Resolution for Noisy Text

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.

Feature Engineering

- Explore features specific to noisy text
- Dependency parse based features found more useful for noisy text

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

Computational Phylogenetics

- Find evolutionary ties between old manuscripts
  - Analyze the underlying challenges
  - Study word etymology, and relate to the available versions of the manuscripts
- Find (understand) relationships between an ancestral sequence and its descendants
  - Evolution of family of sequences
- Estimate time of divergence between a group of manuscripts

Introduction to Phylogenetics

- The Computational purview of our research problem complies of developing new methods for phylogeny estimation, and analysis.
    - or using the currently available methods to analyze the ‘text’ data and prepare a critical edition of the said text.
- The phylogenetic tree construction can be done via various methods viz. Distance method, Bayesian Inference, Maximum likelihood etc.; eventual aim is to be able to construct a phylogenetic tree depicting the hierarchy and the timeline of the evolution of the text.
- Despite of the availability of various methods, there is no guarantee to be able to do so with high probability under reasonable conditions, some which do, they vary considerably in their requirements (Warnow et al., 2001).

The problem is...

- Multiple versions of the same ‘text’ are available due to manual copying and modification in due time.

- Different versions are prone to various errors such as typographical errors / missing portions / additional comments.

Image Source: http://sanskrit.lib.cam.ac.uk/category/saiva-literature

The possible solutions are…

- Despite multiple variants, many can be clubbed into a ‘clade’ or a ‘family’ of variants from a common ancestor.
- Bayesian inferencing using the probabilities of ‘parenthood’ or ‘descendance’ associated with variants and the families of variants.
- Heuristic search and optimization methods are used in combination with tree-scoring functions to identify a reasonably good tree that fits the data.
- Advanced methods use the optimality criterion of maximum likelihood, often within a Bayesian Framework.

Cognate Identification

Cognate Identification is the problem of finding sets of words or word pairs which are related to each other etymologically. They have a history with each other.

- They may or may not carry the same meaning. Given that with time, the same word may change the meaning slightly.

Languages tend to change with time, and derive a lot from each other.

- English - French -- father - père
- Bengali - Hindi
  - हाजार - हज़ार (haajaar - hazaar) meaning thousand
  - जीबन - जीवन (jeeban - jeevan) meaning life

Cognate Identification: False Friends

- Bengali - Hindi
  - ओभिमान - अभिमान (Obhimaan - Abhimaan) meaning holding a grudge and pride respectively.
- English - Spanish -- Vase - Vaso meaning holder for flowers and drinking glass respectively.
- English - French -- Pretend - prétendre where the french word means claim instead of the English sense which means to present something which is not true.
- **False Friends have the same origin but do not have the same meaning. They are still cognate words.**

Cognate Identification: False Cognates

- English - Greek -- ache - akhos; both mean pain
- English - Hindi/Sanskrit - saint - sant; both mean a person who has an exceptional degree of holiness
- English - English (Same Language) -- Marshal - Martial; Orthographically similar but different origin, and different meaning.

The Cognate Matrix

Meaning \ Origin	Same	Different
	True Cognates	**False Cognates**
Same	**Father – Père** (En – Fr) **हज़ार – হাজার** (Hi – Bn) (hazaar – hajaar) (both meaning “thousand”) **जीवन – জীবন** (Hi – Bn) (Jeevan – jeeban) (both meaning “life”) **Celebrate – Celebrar** (En – Es) (both meaning the “action of celebrating”)	***ache – ákhos*** (En – El) (both meaning “pain”) **Saint – Sant** (En – Sa) (both meaning “a holy person”) ***feu - Feuer*** (Fr – De) (both meaning “fire”) ***ciao – chào*** (It – Vi) (both meaning “hello/goodbye”)
	False Friends	**Non Cognates**
Different	***friend - frände*** (En – Sv) (meaning “friend” and “Relative” respectively) ***Friend – frænde*** (En - Da) (meaning “friend” and “Relative” respectively) **Vase – Vaso** (En - Es) (“flowers holder” and “glass of water”) **अभिमान – অভিমান** (Hi - Bn) (obhimaan – abhimaan) (both meaning the “action of celebrating”)	**sentences - palabras** (En – Es) **enemy – bạn** (En – Vi) **comma – kochać** (En - Pl) **Bank – bank** (En - En) (When both mean differently – context wise)

False Friends’ Detection

Definition - False friends are word pairs which pose a challenge to NLP tasks of Cognate Detection and Machine Translation since they share similar spelling but mean completely different (For e.g., “gift” in German means “Poison” in English).

Please note that True Cognates spell and mean the same across languages.

Previous studies use lexical similarity and corpus-based measures.

But no notion of Semantic Similarity, which is essential in determining a false friend word pair!

Approaches

Baseline - Combine lexical similarity computation approaches like Normalized Edit Distance, Cosine Similarity, and Jaccard Index.

Our Approach - Use Cross-lingual Word Embeddings (CLWE)

How!?

Build a common space which projects the embeddings of two different monolingual embeddings. Use a simple linear projection for that matter - but a common space is necessary.

Similarity from CLWE

One can easily compute cosine similarity between two vectors.

Is a better measure available? - Angular Cosine Similarity / Distance.

Compute the similarities between each pair.

Done? Really?

Context Plays an important role

You can build your dataset from either a knowledge graph like a Linked Wordnet discussed earlier.

OR

You can use corpus to find lexically similar words (in-domain)

But in both the cases, if you have the context of the word - that should essentially help you determine what ‘sense’ is a word used in.

Create a Bag-of-words of the context available from either dataset.

Two Scores Problem!

Score 1 - similarity between the word-pair (from neural embeddings)

Score 2 - similarity between the contexts (from neural embeddings)

Can you suggest a method which can learn on both the scores and classify a pair to be a false friend, or for that matter a cognate pair?

This is where you use your neural models - to learn the threshold to be applied on a score.

Some Results

	LP	Baselines						Our Approach					
		OSA			Castro *et. al.*(2018)			WEA (100 dim.)			WEA (50 dim.)		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
WNData (Dataset 1)	**Hi-Bn**	0.86	0.34	0.49	0.61	0.55	0.58	0.95	0.89	**0.92**	0.92	0.87	0.90
	Hi-Gu	0.36	0.51	0.42	0.64	0.58	0.61	0.91	0.69	0.79	0.93	0.95	**0.94**
	Hi-Mr	0.39	0.3	0.34	0.32	0.42	0.36	0.9	0.68	0.77	0.92	0.93	**0.92**
	Hi-Pa	0.28	0.65	0.39	0.58	0.49	0.53	0.98	0.77	0.86	0.99	0.97	**0.98**
	Hi-Sa	0.12	0.35	0.18	0.59	0.33	0.42	0.63	0.84	0.72	0.63	0.99	**0.77**
	Hi-Ml	0.10	0.63	0.18	0.41	0.34	0.37	0.66	0.63	0.65	0.71	0.86	**0.78**
	Hi-Ta	0.04	0.79	0.07	0.17	0.28	0.21	0.38	0.38	0.38	0.36	0.66	**0.47**
	Hi-Te	0.07	0.66	0.14	0.39	0.52	0.45	0.43	0.55	0.48	0.47	0.84	**0.61**
	Hi-Ne	0.35	0.42	0.38	0.55	0.49	0.52	0.88	0.64	0.74	0.90	0.96	**0.93**
CData (Dataset 2)	**Hi-Bn**	0.66	0.29	0.40	0.55	0.35	0.43	0.85	0.6	**0.70**	0.84	0.53	0.65
	Hi-Gu	0.32	0.48	0.38	0.49	0.65	0.56	0.71	0.62	0.66	0.73	0.65	**0.69**
	Hi-Mr	0.29	0.22	0.25	0.29	0.38	0.33	0.69	0.61	0.65	0.76	0.62	**0.68**
	Hi-Pa	0.22	0.57	0.32	0.61	0.55	0.58	0.71	0.71	**0.71**	0.74	0.69	**0.71**
	Hi-Sa	0.09	0.28	0.14	0.52	0.41	0.46	0.55	0.56	0.55	0.55	0.6	**0.57**
	Hi-Ml	0.10	0.54	0.17	0.31	0.39	0.35	0.65	0.52	0.58	0.65	0.59	**0.62**
	Hi-Ta	0.07	0.69	0.13	0.27	0.18	0.22	0.28	0.21	0.24	0.26	0.39	**0.31**
	Hi-Te	0.09	0.58	0.16	0.49	0.32	0.39	0.61	0.54	0.57	0.63	0.58	**0.60**
	Hi-Ne	0.31	0.38	0.34	0.52	0.59	0.55	0.75	0.56	0.64	0.79	0.59	**0.68**

Table 4: Results: Precision (P), Recall (R) and F-Scores (F) for all language pairs (LP) when Orthographic Similarity based baseline Approach (OSA), Word Embeddings based (WEA) Approaches, and Castro, Bonanata, and Rosá (2018)'s approach are evaluated against Gold data.

Cognate Detection Problem - slightly different approach!

When trying to detect lexically similar words, use the baseline measures.

When getting into semantics - use cross-lingual word embeddings like we did for the last problem.

But can we learn the threshold using a classification model?

Various Neural Models

CNNs - Convolutional neural networks have been used for various NLP tasks where a character sequence is in play. But is the character sequence only thing we are interested in? On top of that CNNs penalize heavily when the order of characters changes.

BiLSTMs - Bidirectional Long Short Term Memory(s) have been used in important sequence to sequence learning tasks like machine translation.

Feed Forward - sounds too simple?

Proposed Architecture

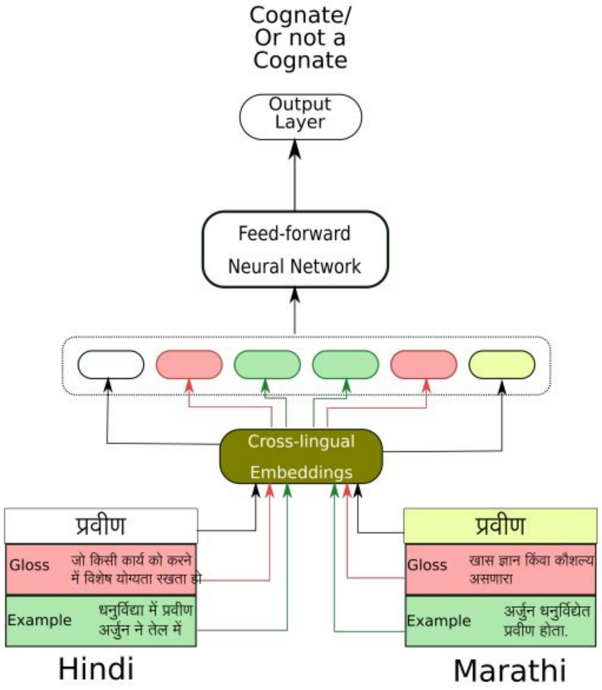


Figure 1: Architecture of the Proposed Deep Learning Model for Cognate Detection

Some more results!

LP	Baseline			Cross-lingual embeddings						Neural Networks w/ Cross-lingual embeddings								
	OSA			WEA100			WEA50			CNN			Bi-LSTM			FFNN		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Hi-Bn	0.39	0.33	0.36	0.91	0.48	0.63	0.67	0.74	0.68	0.58	0.76	0.66	0.63	0.74	0.67	0.69	0.73	**0.71**
Hi-Gu	0.41	0.16	0.23	0.93	0.57	0.71	0.75	0.79	0.76	0.76	0.81	0.77	0.74	0.79	0.76	0.80	0.79	**0.80**
Hi-Mr	0.47	0.21	0.29	0.96	0.46	0.62	0.71	0.76	0.72	0.60	0.77	0.68	0.60	0.76	0.67	0.72	0.76	**0.73**
Hi-Pa	0.29	0.07	0.11	0.98	0.48	0.64	0.74	0.79	0.73	0.62	0.78	0.70	0.67	0.78	0.67	0.75	0.79	**0.76**
Hi-Sa	0.41	0.17	0.24	0.71	0.70	0.70	0.71	0.77	0.72	0.66	0.77	0.70	0.68	0.77	0.70	0.74	0.78	**0.75**
Hi-Ml	0.26	0.30	0.28	0.61	0.54	0.57	0.61	0.65	0.60	0.61	0.65	0.60	0.58	0.64	0.57	0.64	0.67	**0.65**
Hi-Ta	0.24	0.17	0.20	0.49	0.50	0.49	0.54	0.51	0.50	0.52	0.48	0.45	0.54	0.50	0.47	0.57	0.55	**0.55**
Hi-Te	0.20	0.14	0.16	0.64	0.65	0.64	0.64	0.70	0.63	0.60	0.71	0.59	0.62	0.70	0.61	0.66	0.69	**0.67**
Hi-Ne	0.42	0.18	0.25	0.85	0.55	0.67	0.75	0.81	0.75	0.67	0.82	0.74	0.79	0.82	0.74	0.77	0.81	**0.78**

Table 3: WNData Precision (P), Recall (R) and F-Scores (F) when baseline (OSA), word embeddings (WEA50, WEA100), neural networks (FFNN, RNN) based classification approaches are evaluated against gold data.

Cognate Identification: Future!

Work published at GWC 2019. Ongoing work and submissions to ACL 2020 underway.

We have already identified the possible corpus we could work with and the experiments are already underway.

Some of the initial experiments have also been submitted for review in a conference. Fingers crossed!

Final Slide: No More!

Boring!

Boring!

Boring!

But, I really do hope you enjoyed some parts of it, if not all!

Thank you!

Questions?

I would be happy to answer any and all.

References

www.cfilt.iitb.ac.in

www.cse.iitb.ac.in/~pb

www.cse.iitb.ac.in/~diptesh