Introduction to Natural Language Processing

Sudha Bhingardive

IIT Bombay, Mumbai

Objective of the course

- Understanding of basics of Natural Language Processing
- Understanding Natural Language Processing using Python programming

Topics which will be covered

- Tokenization, Morphology, Stemming
- Regular expression
- POS Tagging
- Language Modeling
- Parsing
- Semantics, Representing and Understanding Meaning, Knowledge Representation
- NLP topics: Tex Summarization, Word Sense Disambiguation, Information Extraction, Sentiment Analysis,

Reference materials

- Main Text(s):
  - Natural Language Understanding: James Allan
  - Speech and NLP: Jurafsky and Martin
  - Foundations of Statistical NLP: Manning and Schutze
- Other References:
  - NLP a Paninian Perspective: Bharati, Cahitanya and Sangal
  - Statistical NLP: Charniak
- Journals
  - Computational Linguistics, Natural Language Engineering, AI, AI Magazine, IEEE SMC
- Conferences
  - ACL, EACL, COLING, MT Summit, EMNLP, IJCNLP, HLT, ICON, SIGIR, WWW, ICML, ECML

What is NLP?

- Branch of AI
- To give computers the ability to process human language
- 2 Goals
  - Science Goal: Understand the way language operates
  - Engineering Goal: Build systems that analyze and generate language; reduce the man machine gap

Ambiguity

- This is what makes NLP challenging
- The crux of the problem

Stages of processing

- Phonetics and phonology
- Morphology
- Lexical Analysis
- Syntactic Analysis
- Semantic Analysis
- Pragmatics
- Discourse

Phonetics

- Processing of speech
- Challenges
  - Homophones: *bank (finance)* vs. *bank (river bank)*
  - Near Homophones: *maatraa* vs. *maatra (hin)*
  - Word Boundary
    - *aajaayenge (aa jaayenge (will come)* or *aaj aayenge (will come today)*
    - *I got [ua]plate*
  - Disfluency: *ah, um, ahem etc.*

Morphology

- Word formation rules from *root* words
- Nouns: Plural (*boy-boys);* Gender marking (czar-czarina)
- Verbs: Tense (*stretch-stretched);* Aspect (*e.g. perfective sit-had sat*); Modality (e.g. *request khaanaa → khaaiie)*
- First crucial first step in NLP
- Languages rich in morphology: e.g., Dravidian, Hungarian, Turkish
- Languages poor in morphology: Chinese, English
- Languages with rich morphology have the advantage of easier processing at higher stages of processing

Lexical Analysis

- Essentially refers to dictionary access and obtaining the properties of the word

  e.g. dog

  noun (lexical property)

  take-'s'-in-plural (morph property)

  animate (semantic property)

  4-legged (-do-)

  carnivore (-do)

Challenge: *Lexical or word sense disambiguation*

Lexical Disambiguation

First step: *part of Speech Disambiguation*

- *Dog* as a *noun* (animal)
- *Dog* as a verb (*to pursue)*

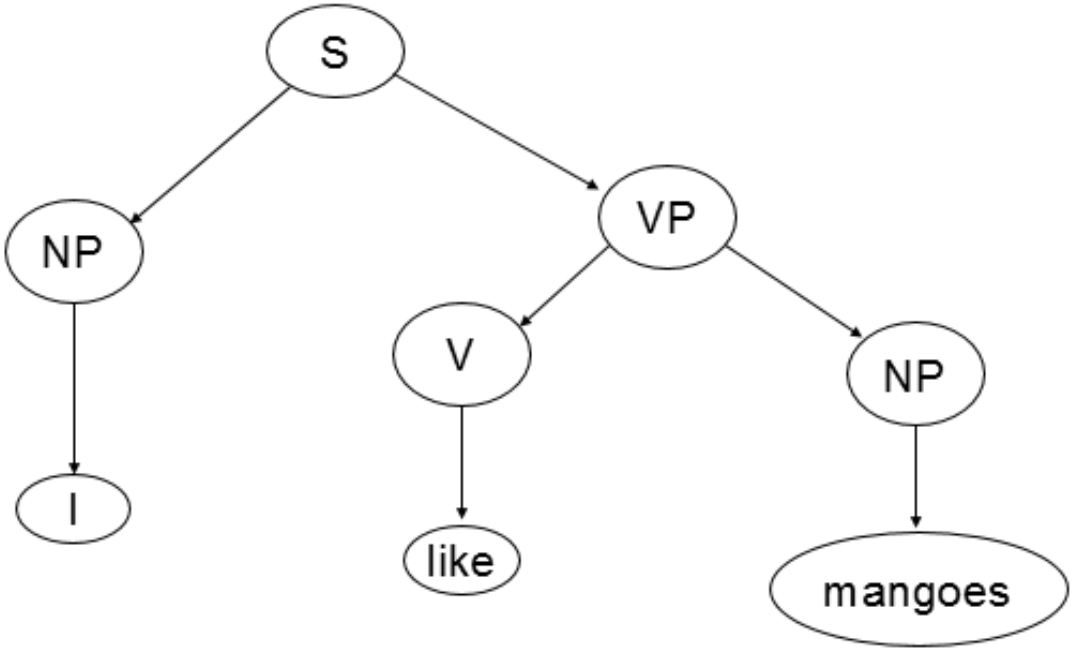
Sense Disambiguation

- *Dog (*as *animal)*
- *Dog (*as *a very detestable person)*

Syntax Processing Stage

- Structure Detection:

  I like mangoes



Syntax Processing Stage contd..

- Scope

  1.The old men and women were taken to safe locations

  (old men and women) vs. *((old men) and women)*

  2. No smoking areas will allow Hookas inside

- Preposition Phrase Attachment
  - *I saw the boy with a telescope*

    (who has the *telescope?*)
  - I saw the boy with the pony-tail

    (world knowledge: *pony-tail* cannot be an *instrument of seeing*)

Semantic Analysis

- Representation in terms of
  - Predicate calculus/Semantic Nets/Frames/Conceptual Dependencies and Scripts
- *John gave a book to Mary*
  - Give action: Agent: John, Object: Book, Recipient: Mary
- Challenge: ambiguity in semantic role labeling
  - (Hin) aapko mujhe mithaai khilaanii padegii (ambiguous in Marathi and Bengali too; not in Dravidian languages)

Pragmatics

- Very hard problem
- Model user intention
  - *Tourist (in a hurry, checking out of the hotel, motioning to the service boy): Boy, go upstairs and see if my sandals are under the divan. Do not be late. I just have 15 minutes to catch the train.*
  - *Boy (running upstairs and coming back panting): yes sir, they are there.*

Discourse

- Processing of *sequence* of sentences
- *Mother* to *John*:

  John go to school. It is open today. Should you bunk? Father will be very angry.
- Ambiguity of *open*
- *bunk* what?
- *Why will the father be angry?*
  - Complex chain of reasoning and application of world knowledge

  Ambiguity of *father*

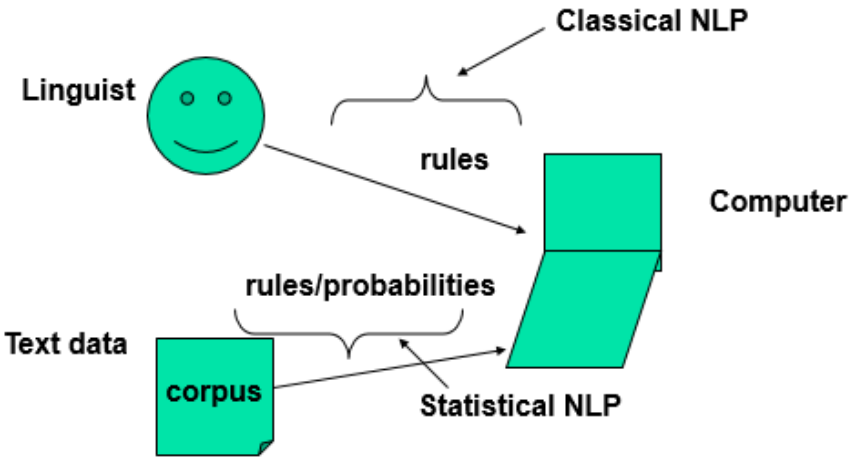
  father as *parent*

  or

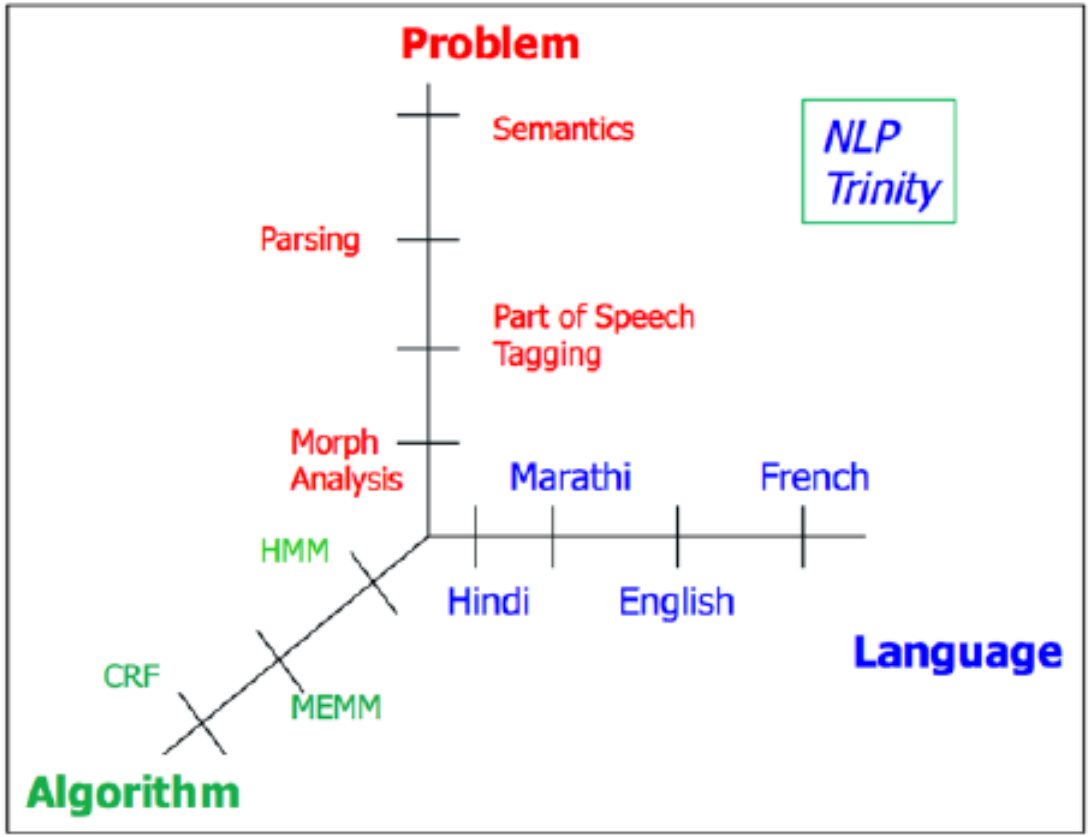
  father as *headmaster*

Two Views of NLP and the Associated Challenges

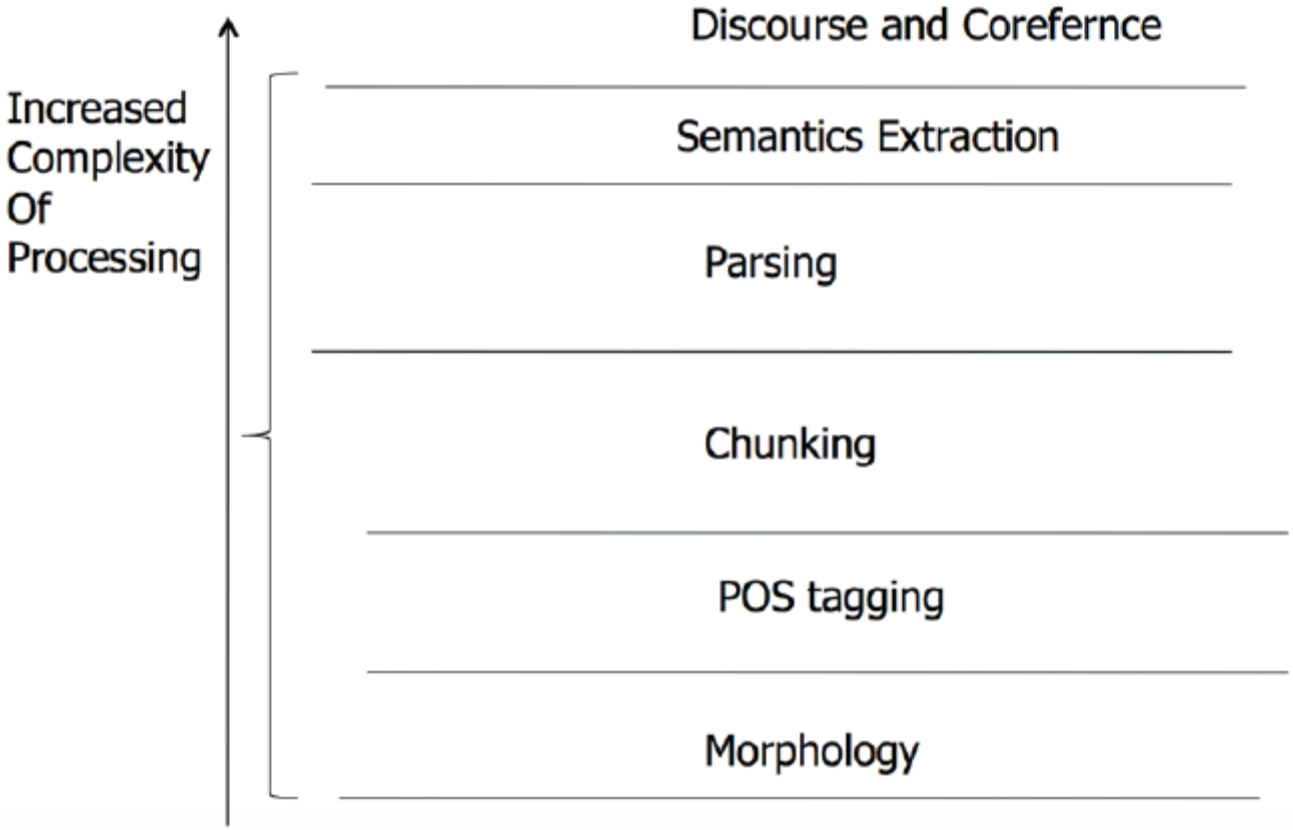
- Classical View:
  - Layered processing; Ambiguities
- Statistical/Machine Learning View:
  - Prediction



NLP Trinity


Problem
Semantics
NLP Trinity
Parsing
Part of Speech Tagging
Morph Analysis
Marathi
French
HMM
Hindi
English
Language
CRF
MEMM
Algorithm

NLP Layers


Increased Complexity Of Processing
Discourse and Corefernce
Semantics Extraction
Parsing
Chunking
POS tagging
Morphology

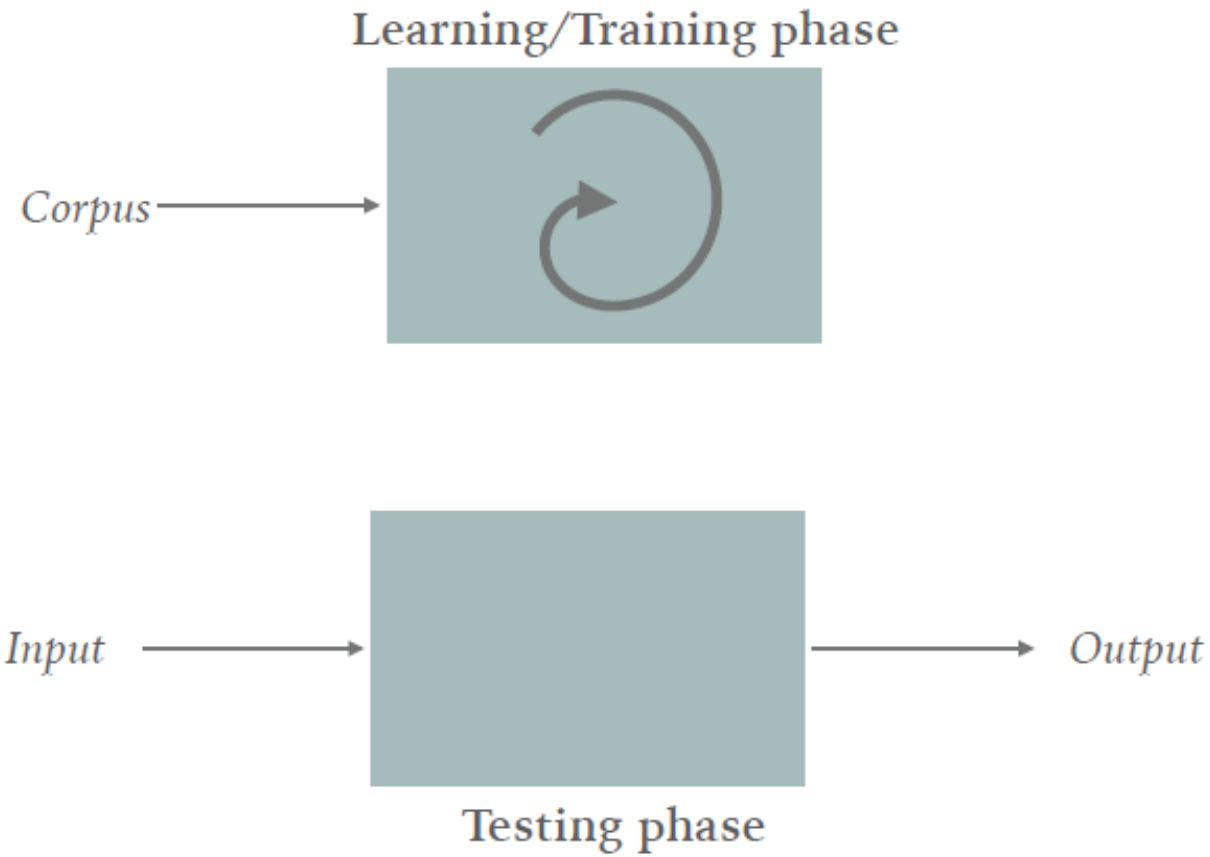
NLP Techniques

- Rule based approach
- Knowledge based approach
- Machine Learning based approach
  - Supervised approach
  - Unsupervised approach
  - Semi-supervised approach

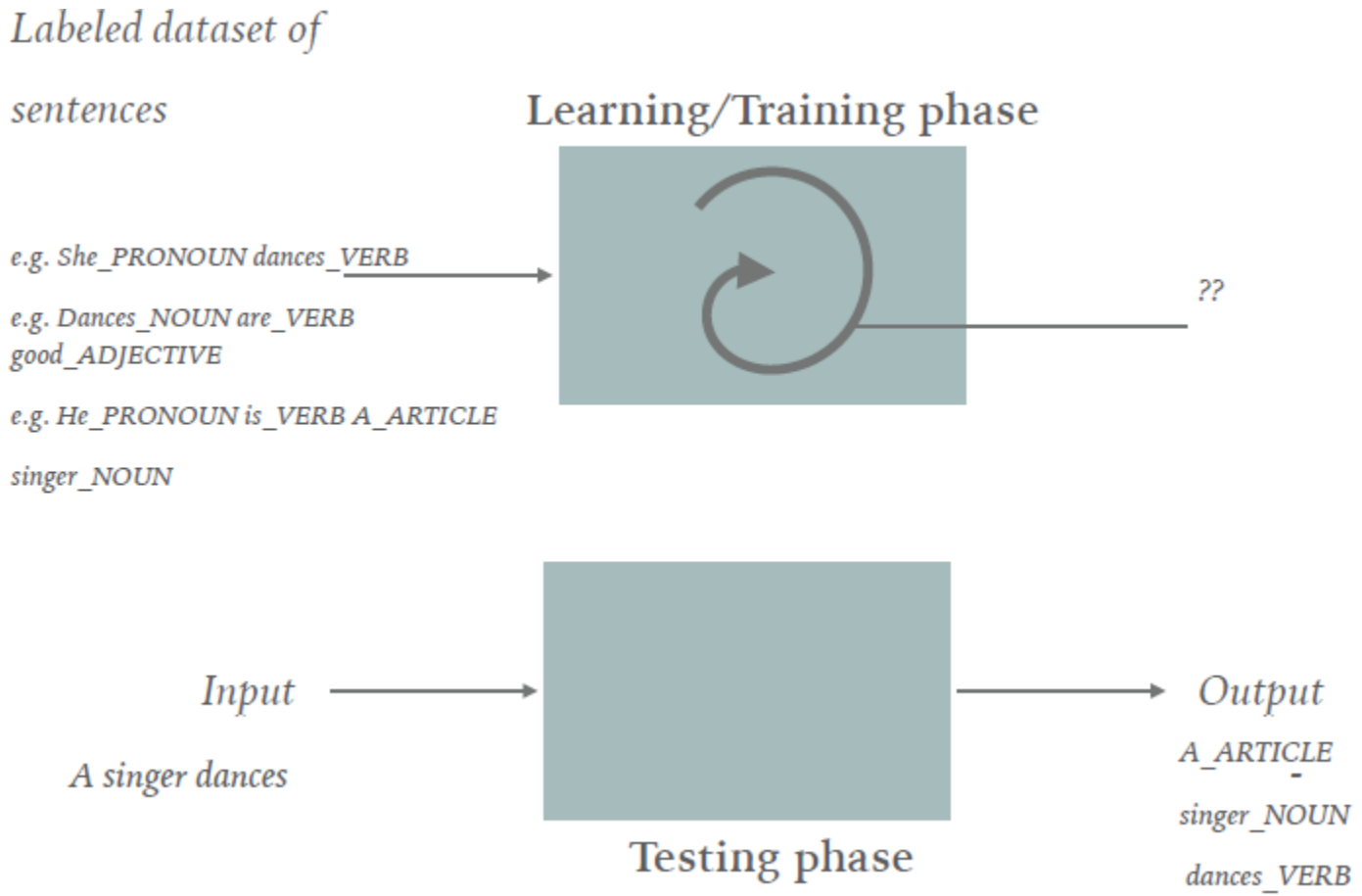
What is the output of NLP system?

- Option 1: A set of rules
- Option 2: A set of probability values

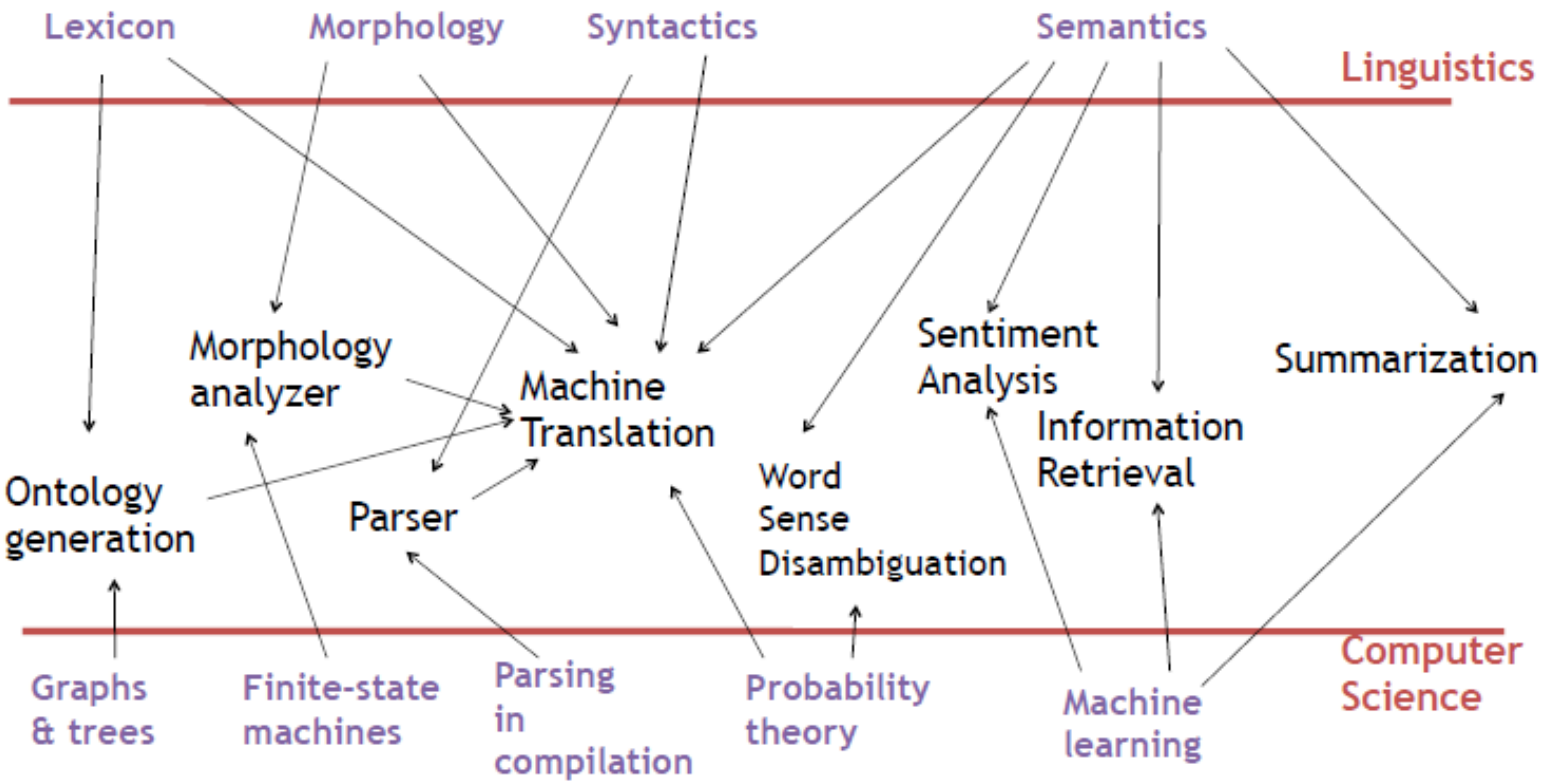
Basic building blocks



Example: Parts-Of-Speech tagging



NLP: At the confluence of Linguistics & computer science



Lexical Resources

- Backbone of every NLP tasks
- Computer do not understand natural languages, like English, Hindi, etc.
- We need to train them based on available resources

Corpus

- A collection of text, it can be a plain text document containing data or a structured XML file containing data along with its descriptors and headers
- Corpora - Plural form of corpus
- Examples:
  - Monolingual corpora
  - Comparable
  - Parallel
  - Annotated corpora: pos-tagged, sense-tagged

Corpus contd..

- Monolingual corpus

A boy is sitting in the kitchen

A boy is playing a tennis

Some men are watching tennis

A girl is holding a black book

Corpus contd..

- Comparable corpus
  - A collection of "similar" texts in different languages or in different varieties of a language.
  - A pair of corpora in two different languages, which come from the same domain.
  - An example would be a corpus of articles about football from English and Danish newspapers; or legal contracts in Spanish and Greek.

Corpus contd..

- Parallel corpus: the texts in one language (L1) are translations of texts in the other language (L2).

L1 (English)	L2 (Hindi)
A boy is sitting in the kitchen	एक लड़का रसोई में बैठा है
A boy is playing a tennis	एक लड़का एक टेनिस खेल रहा है
Some men are watching tennis	कुछ पुरुष टेनिस देख रहे हैं
A girl is holding a black book	एक लड़की एक काली पुस्तक पकड़े हुए है

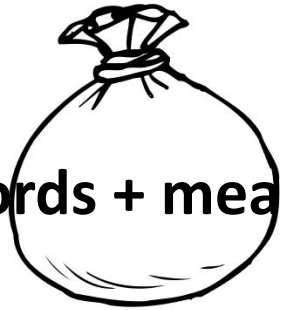
Corpus contd..

- Annotated corpus:
  - Corpus annotated with different label to each unit in the corpus
- Pos-tagged corpus
  - She_PRON eats_Verb rice_NOUN
- Sense-tagged corpus
  - The city_18406385 is_22579744 famous_41426596 for its majestic_41333338 forts_13350250 , palaces_13834381 which attract_21492358 tourists_110557758 from the world_19138104 .

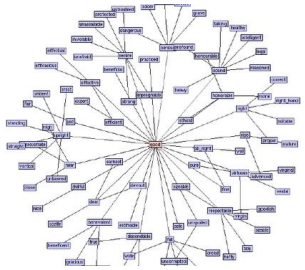
What is WordNet?

Dictionary

=



Words + meanings



WordNet

=

Words + meanings + Semantic and Lexical Relations

What is WordNet? Contd..

- A lexical knowledge database for a language
- Consists of synsets and lexico-semantic relations
- Categorizes synsets into four main parts-of-speech categories: nouns, adjectives, adverbs and verbs
- Monolingual WordNet
  - English, Hindi, Sanskrit
- Multilingual WordNet
  - IndoWordNet, EuroWordNet, BabelNet

WordNet Synset

- Each synset consist of:
  - Sense ID
  - Parts-of-speech category
  - Synset Members (Synonyms words)
  - Gloss or Concept Definition
  - Example Sentence

- Synset of a boy:

(10305010) (n) male child, boy (a youthful male person) "the baby was a boy"; "she made the boy brush his teeth every night"; "most soldiers are only boys in uniform"

WordNet Lexico-Semantic Relations

- Synonymy
- Antonymy
- Gradation
- Hypernymy / Hyponymy
- Meronymy / Holonymy
- Entailment
- Attribute
- Nominalization
- Ability Link
- Capability Link
- Function Link

Lexical Relations

- Relation between words
- Synonymy: relationship between words in a synset.
  - {plant, flora}, ‘plant’ and ‘flora’ are related through synonymy relation.
- Antonymy: relationship between words having an opposite meaning.
  - ‘day’ and ‘night’ are antonyms of each other.
- Gradation:
  - ‘morning’, ‘afternoon’, ‘evening’ are related through gradation relation

Semantic Relations

- Relation between synsets
- Hypernymy / Hyponymy: is-a-kind-of relation
  - ‘fruit’ is a hypernym of ‘mango’ and ‘mango’ is a hyponym of ‘fruit’.
- Meronymy / Holonymy: part-whole relation
  - ‘hand’ is a meronym of ‘body’ and ‘body’ is a holonym of ‘hand’

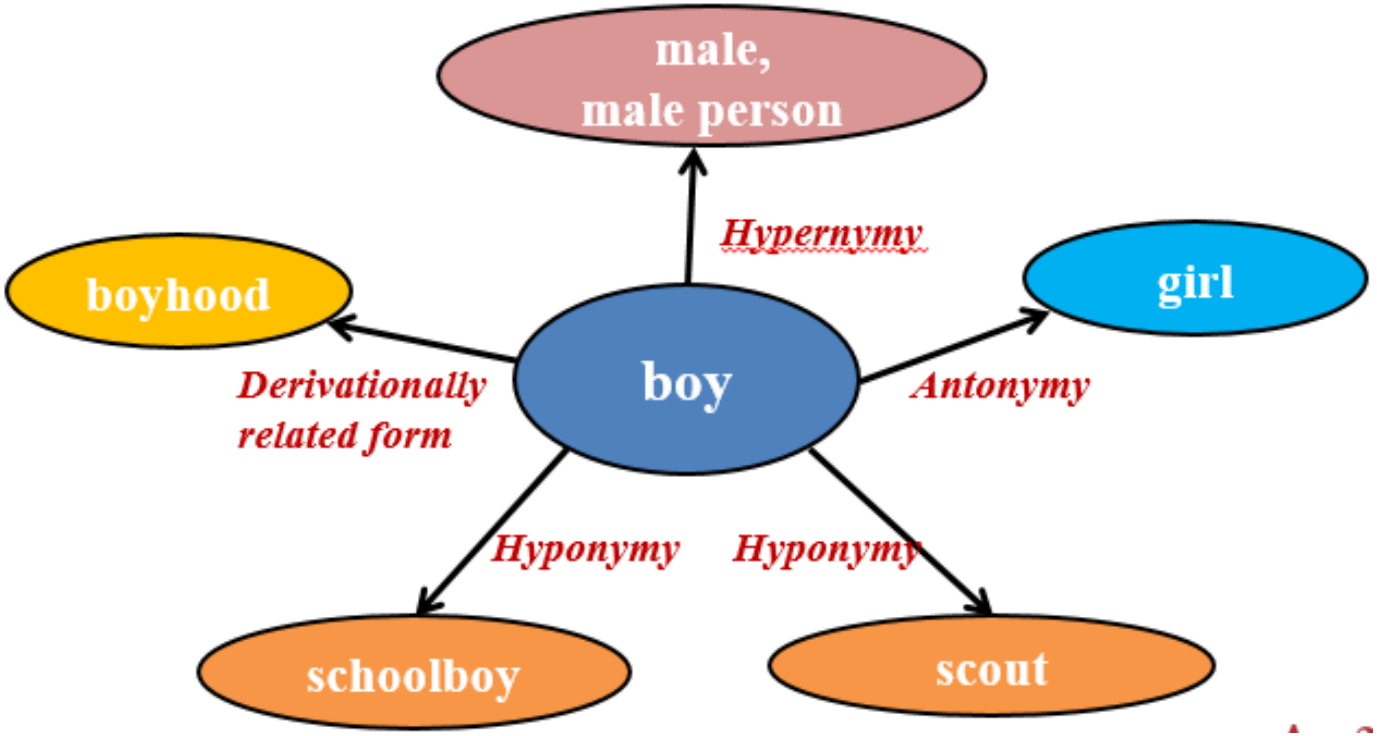
Semantic Relations contd..

- Entailment:
  - ‘snore’ entails ‘sleep’
- Attribute: relationship between noun and adjective synsets
  - ‘hot’ is a value of or attribute of ‘temperature’
- Nominalization: relationship between noun and verb synsets
  - ‘service’ nominalizes the verb ‘serve’

Semantic Relations contd..

- Ability Link: specifies the inherited features of a nominal concept
  - ‘animal’ and ‘walk’, ‘fish’ and ‘swim’
- Capability Link: relationship specifies the acquired features of a nominal concept
  - ‘person’ and ‘swim’
- Function Link: relationship specifies the function of a nominal concept
  - ‘vehicle’ and ‘move’ and ‘teacher’ and ‘teach’

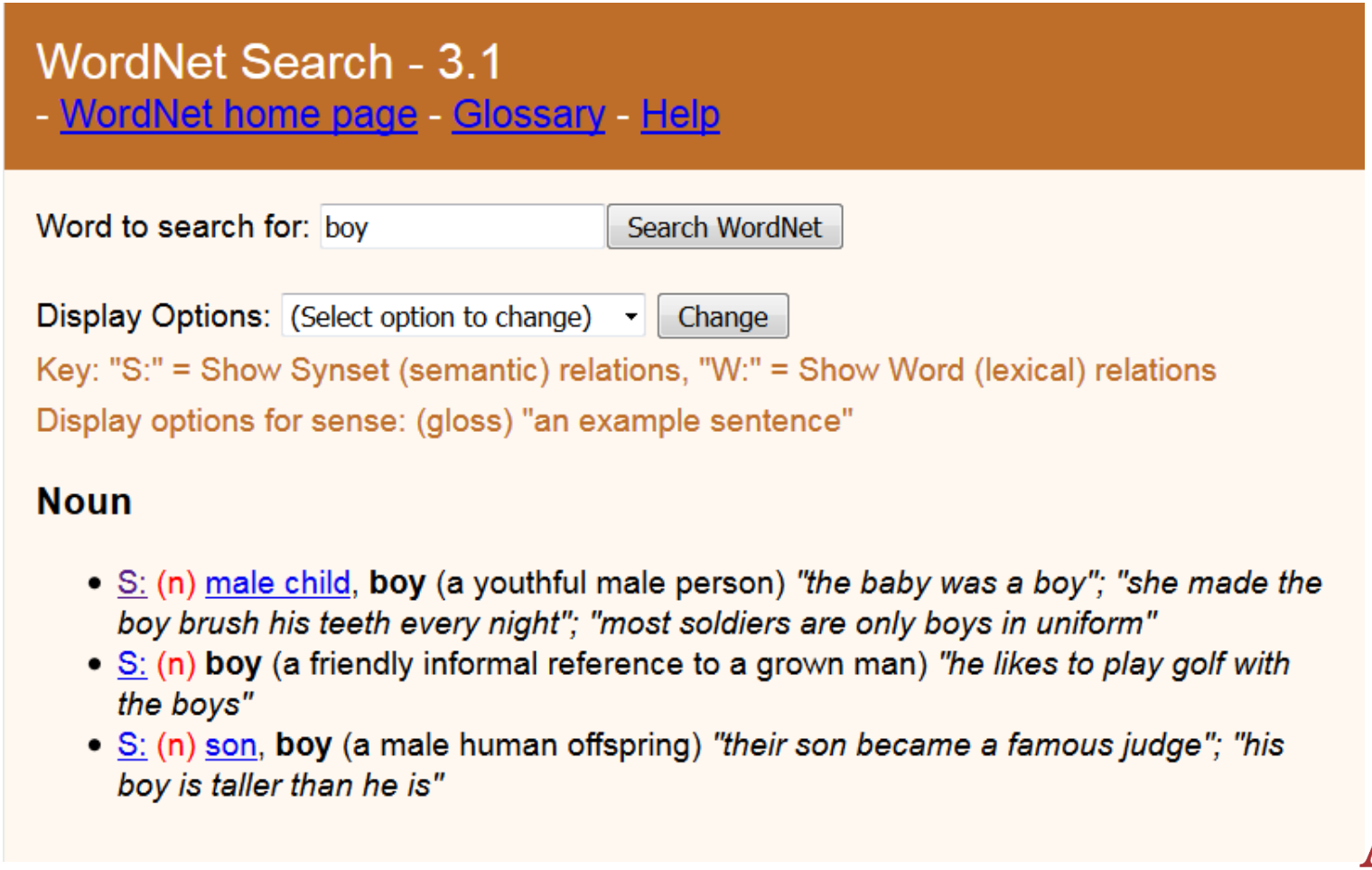
WordNet Lexico-Semantic Relations



Some important wordnets

- English WordNet (Fellbaum, 1998):
  - First semantic net created at Princeton University
- Hindi WordNet (Narayan et. al, 2002)
  - First Indian language Wordnet which is created from English WordNet using expansion approach at IIT Bombay
- IndoWordnet (Bhattacharyya, 2010)
  - A Multilingual Wordnet for 17 Indian Languages
- BabelNet (Navigli, 2010)
  - A very large, wide coverage multilingual semantic network
  - 271 languages, 14 million synsets, and about 745 million word senses
  - Obtained by automatic integration of Wikipedia (encyclopedic) and WordNet (lexicographic)

English WordNet Interface



English WordNet Interface contd..

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: boy Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

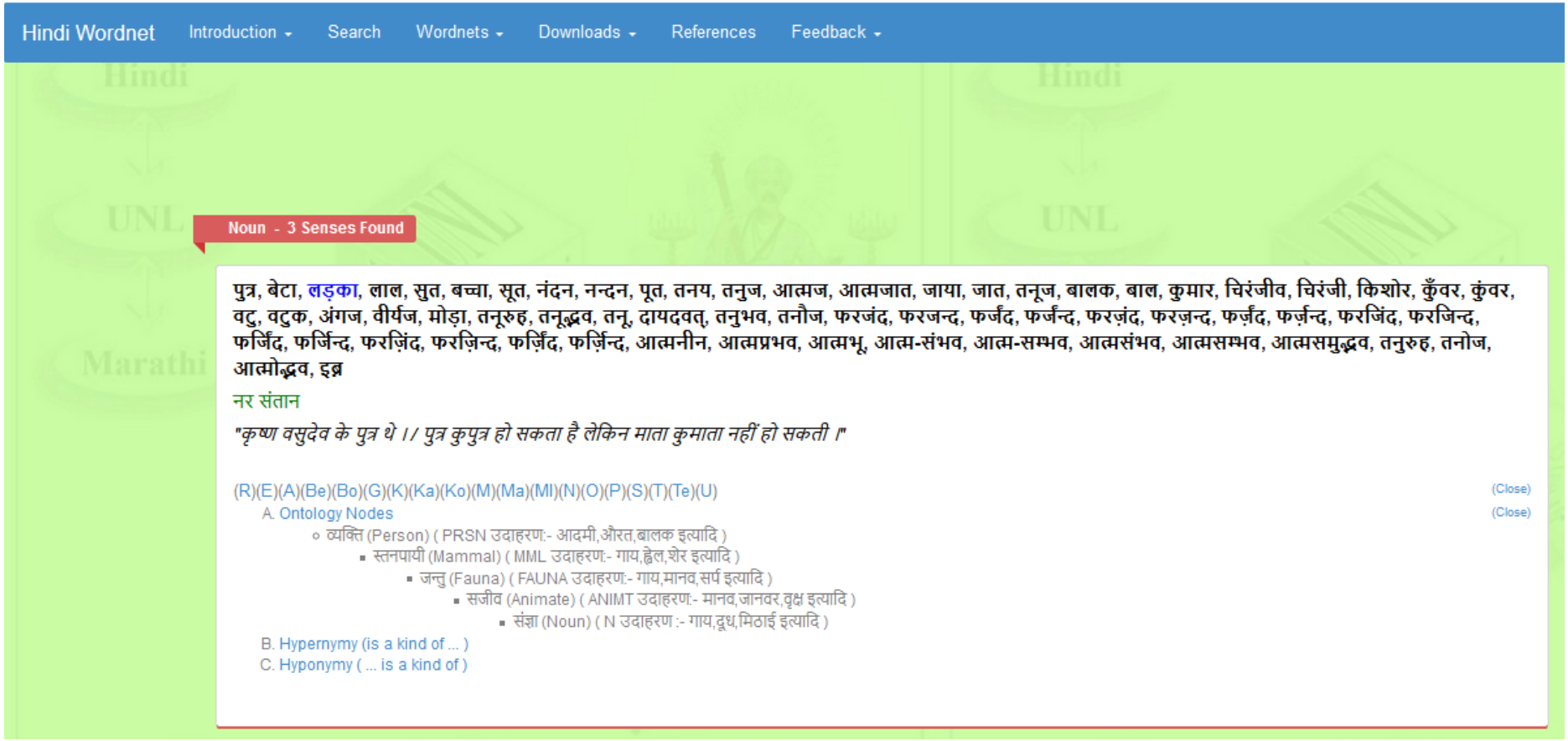
Display options for sense: (gloss) "an example sentence"

Noun

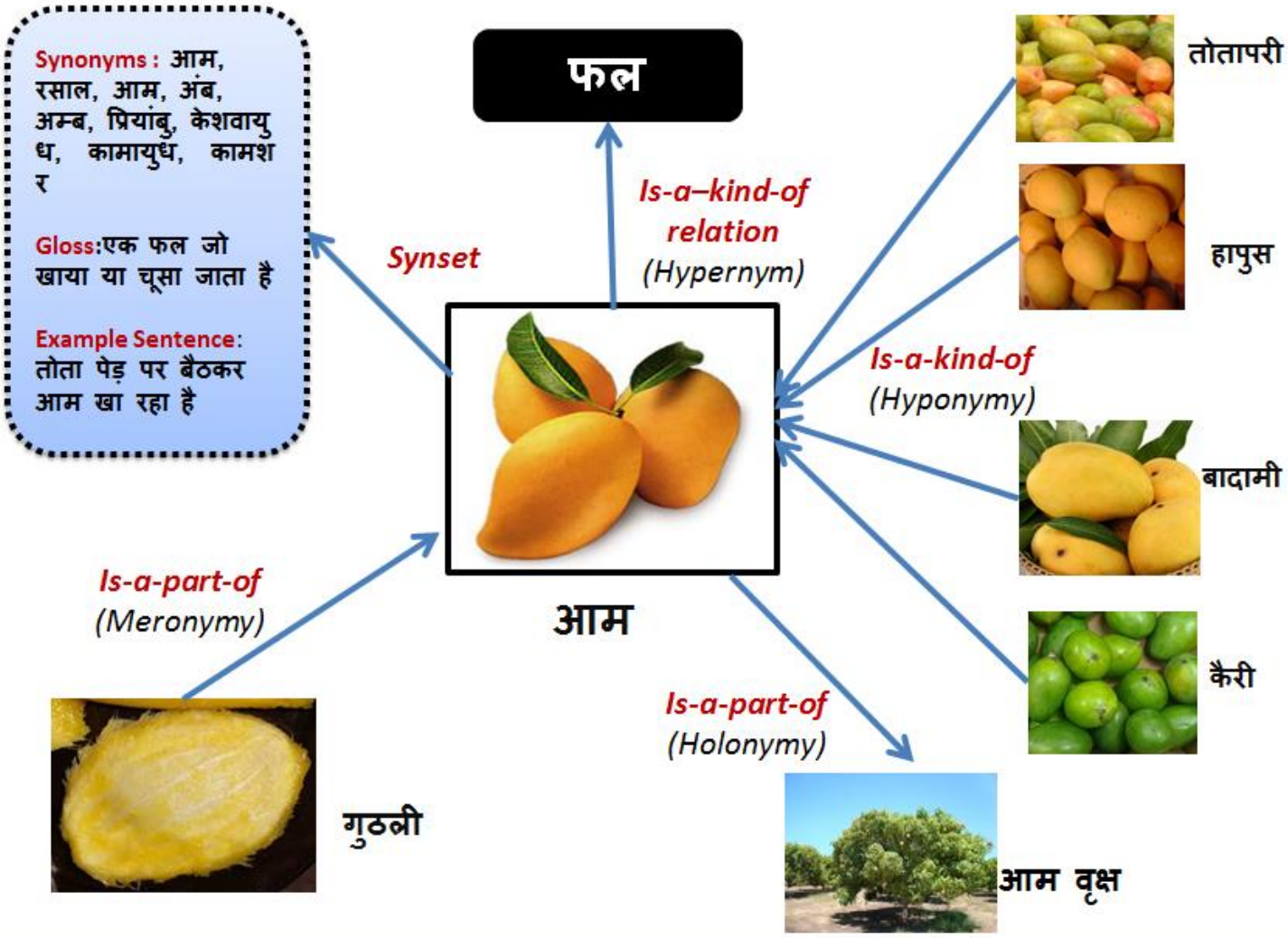
- S: (n) male child, **boy** (a youthful male person) *"the baby was a boy"; "she made the boy brush his teeth every night"; "most soldiers are only boys in uniform"*
  - *direct hyponym* / *full hyponym*
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) male, male person (a person who belongs to the sex that cannot have babies)
  - ***antonym***
    - W: (n) female child [Opposed to: male child] (a youthful female person) *"the baby was a girl"; "the girls were just learning to ride a tricycle"*
    - W: (n) girl [Opposed to: boy] (a youthful female person) *"the baby was a girl"; "the girls were just learning to ride a tricycle"*
  - *derivationally related form*

Hindi WordNet Interface

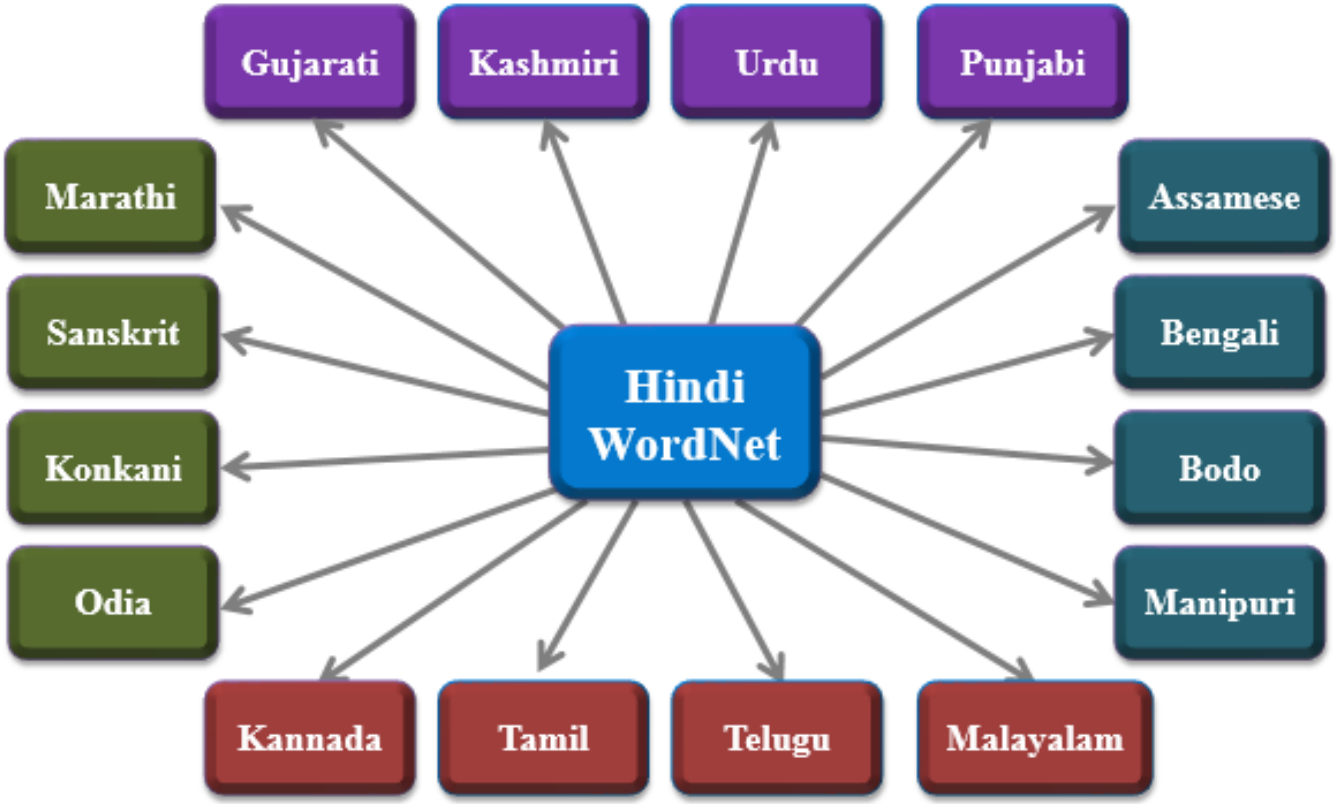
Hindi WordNet Interface contd..



Hindi WordNet Structure



IndoWordNet Structure



IndoWordNet Interface

IndoWordNet linked Synset

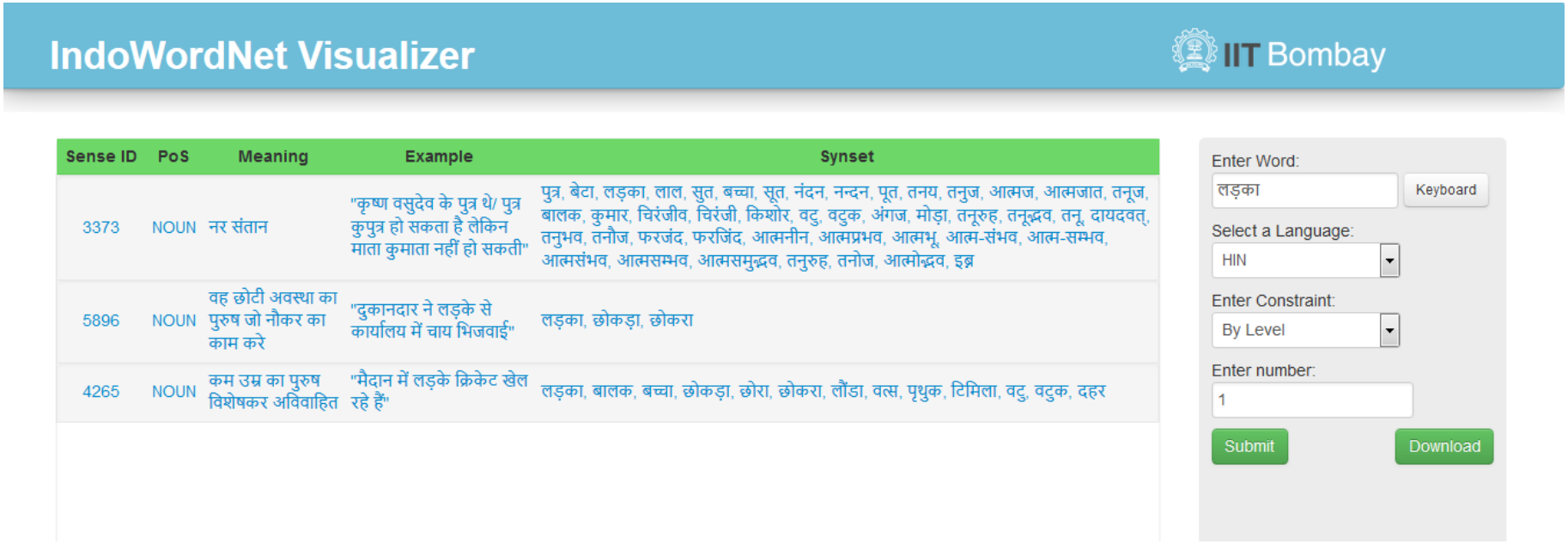
Bengali WordNet	Hindi WordNet	Marathi WordNet
(4265) (n)	(4265) (n)	(4265) (n)
ছেলে, বালক	लड़का, बालक, बाल, बच्चा, छोकड़ा, छोरा, छोकरा	मुलगा,पोरगा,पोर,पोरगे
কম বয়সের পুরুষ, বিশেষত অবিবাহিত	कम उम्र का पुरुष, विशेषकर अविवाहित	साधारणतः सोळा वर्षांखालील पुरुष व्यक्ती
"ময়দানে ছেলেরা ক্রিকেট খেলছে"	*"मैदान में लड़के क्रिकेट खेल रहे हैं ।"*	*"तो मुलगा खूपच हुशार आहे"*

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

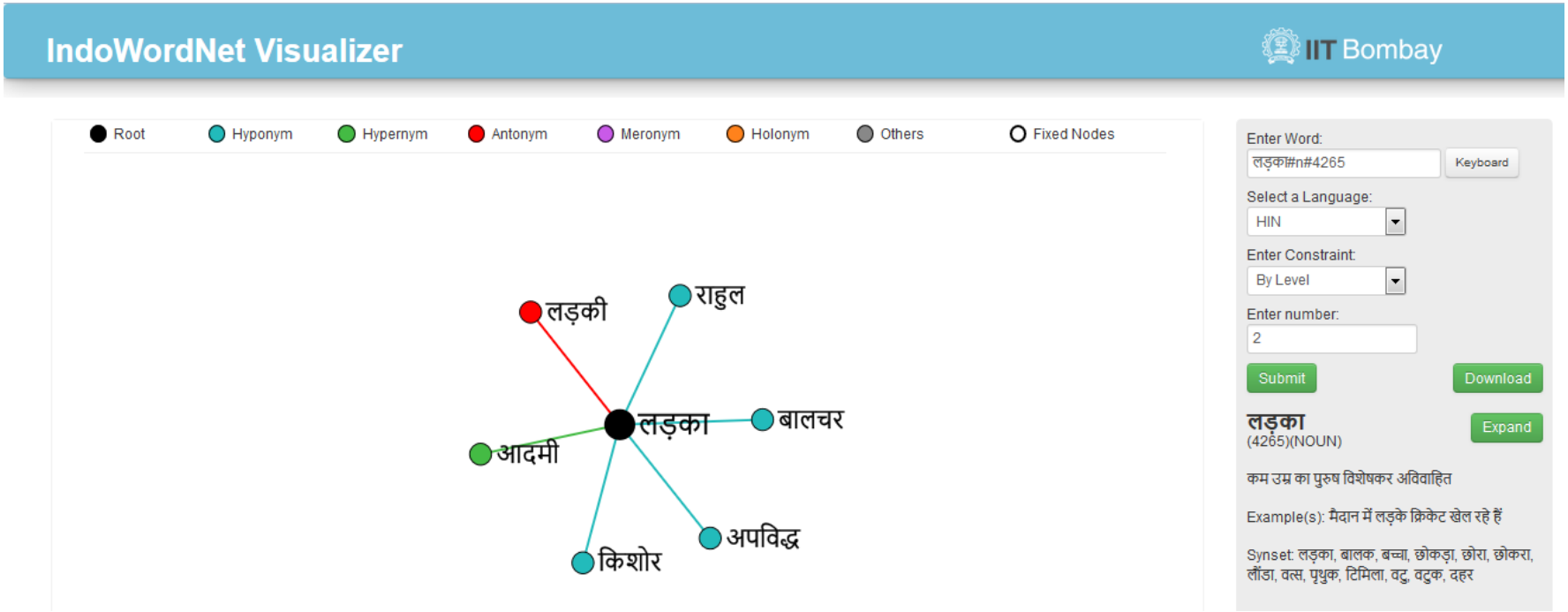
IndoWordNet Synset Statistics

	Noun	Verb	Adjective	Adverb	Total
Hindi	29664	3626	6313	534	40137
Assamese	9065	1676	3805	412	14958
Bengali	27281	2804	5815	445	36346
Bodo	8788	2296	4287	414	15785
Gujarati	26503	2805	5828	445	35599
Kannada	12765	3119	5988	170	22042
Kashmiri	21041	2660	5365	400	29469
Konkani	23144	3000	5744	482	32370
Malayalam	20071	3311	6257	501	30140
Manipuri	10156	2021	3806	332	16351
Marathi	23271	3146	5269	539	32226
Nepali	6748	1477	3227	261	11713
Odiya	27216	2418	5273	377	35284
Punjabi	23255	2836	5830	443	32364
Sanskrit	31476	1247	4004	265	36997
Tamil	16312	2803	5827	477	25419
Telugu	12078	2795	5776	442	21091
Urdu	22990	2801	5786	443	34280

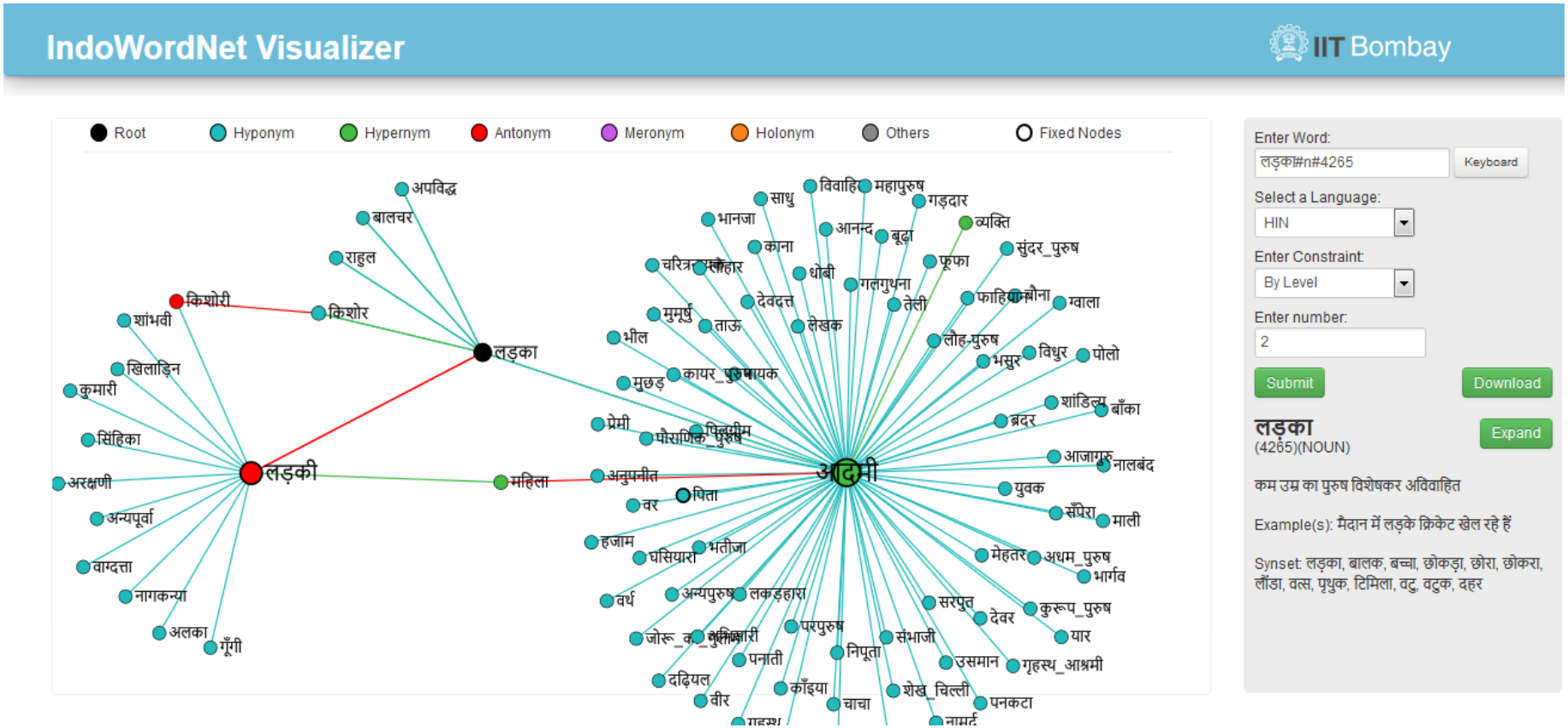
IndoWordNet Visualizer Interface



IndoWordNet Visualizer Interface

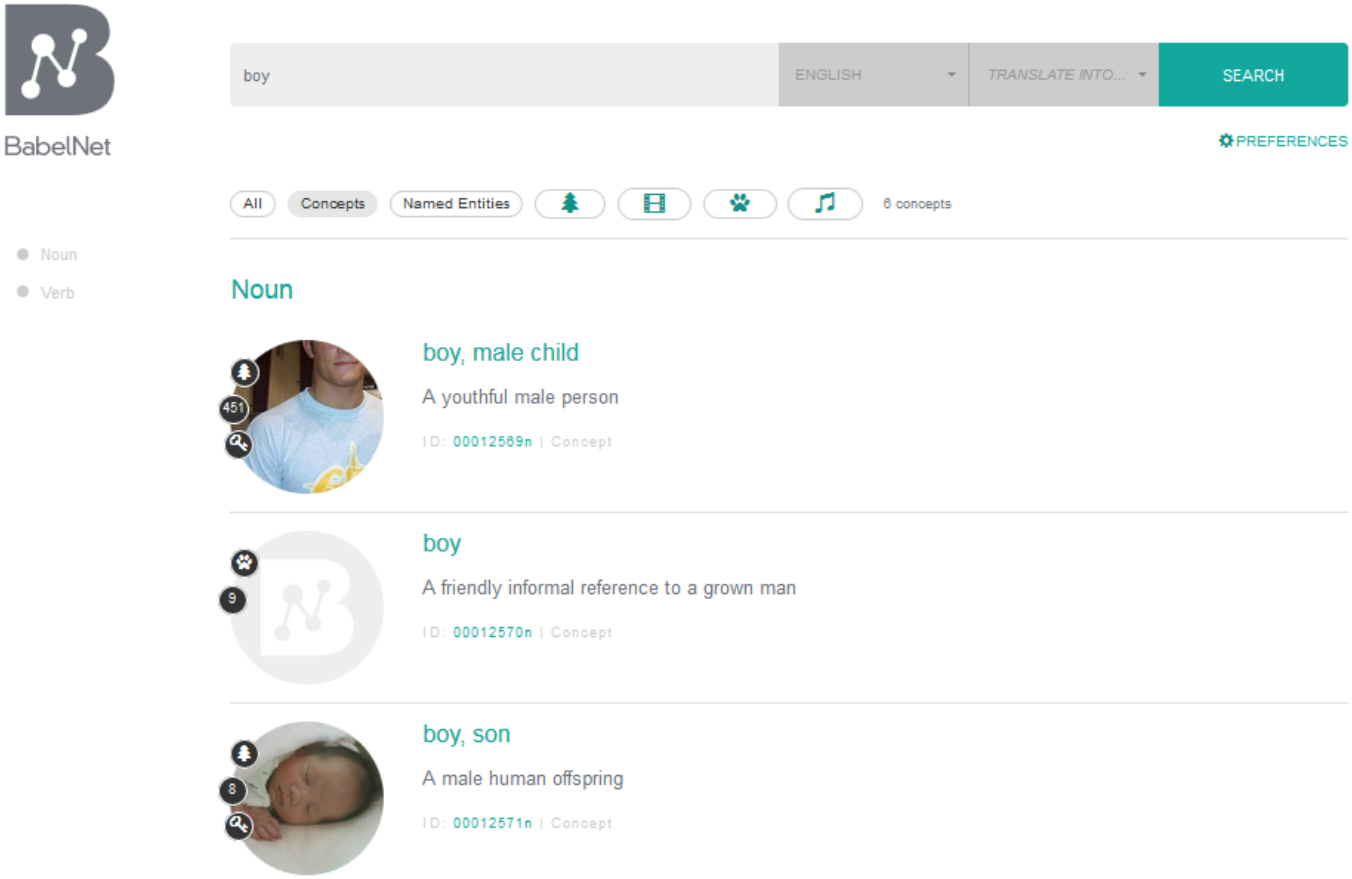


IndoWordNet Visualizer Interface



Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

BabelNet


BabelNet
boy
ENGLISH
TRANSLATE INTO...
SEARCH
PREFERENCES
All
Concepts
Named Entities
6 concepts
Noun
Verb
Noun
boy, male child
A youthful male person
ID: 00012569n | Concept
boy
A friendly informal reference to a grown man
ID: 00012570n | Concept
boy, son
A male human offspring
ID: 00012571n | Concept

Wordnets in the World

- The Global WordNet Organization gives access of wordnets in the world

  http://globalwordnet.org/wordnets-in-the-world/

- Albanian, Arabic, Spanish, Catalan, Basque, Italian, Bulgarian, Czech, Greek, Romanian, Serbian, Turkish, Chinese, Danish, Dutch, Estonian, French, German, Hungarian, Icelandic, Portuguese, Irish, Japanese, Korean, Kurdish, Latin, Macedonian, Norwegian, Persian, Polish, Russian, Swedish

WordNet Applications

- Machine Translation
- Word Sense Disambiguation
- Sentiment Analysis
- Information Retrieval
- MultiWord Expression Detection
- Document structuring and categorization
- Cognitive NLP

NLP Python libraries

- **NumPy** (Mathematical Computing, Advanced mathematical functionalities)
- **Matplotlib** (Numerical plotting library, useful in data analysis)
- **Scipy** (Library for scientific computation)
- **Scikit-learn** (Machine Learning/Data-mining library,
- **PIL** (Python library for Image Processing)
- **PySpeech** (Library for speech processing and text-to speech conversion)
- **XML/LXML** (XML Parsing and Processing)
- **NLTK** (Natural Language Processing)

And many more…

The Natural Language Toolkit (NLTK)

- Developed by Steven Bird and Co. at Stanford University (2006).
- Open source python modules, datasets and tutorials
- Papers:
  - Bird, Steven. "NLTK: the natural language toolkit." *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 2006.
  - Loper, Edward, and Steven Bird. "NLTK: The natural language toolkit." *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*. Association for Computational Linguistics, 2002.

Components of NLTK

- **Code**: corpus readers, tokenizers, stemmers, taggers, chunkers, parsers, wordnet, ... (50k lines of code)
- **Corpora**: >30 annotated data sets widely used in natural language processing (>300Mb data)
- **Documentation**: a 400-page book, articles, reviews, API documentation

Components of NLTK

- Code
  - Corpus Readers
  - Tokenizers
  - Stemmers
  - Taggers
  - Parsers
  - WordNet
  - Semantic Interpretation
  - Clusterers
  - Evaluation Metrics

Components of NLTK contd..

- **Corpus**
  - **Brown Corpus**
  - **Carnegie Mellon Pronouncing Dictionary**
  - CoNLL 2000 Chunking Corpus
  - Project Gutenberg Selections
  - NIST 1999 Information Extraction: Entity Recognition Corpus
  - US Presidential Inaugural Address Corpus
  - **Indian Language POS-Tagged Corpus**
  - Floresta Portuguese Treebank
  - Prepositional Phrase Attachment Corpus
  - **SENSEVAL 2 Corpus**
  - Sinica Treebank Corpus Sample
  - Universal Declaration of Human Rights Corpus
  - **Stopwords Corpus**
  - TIMIT Corpus Sample
  - Treebank Corpus Sample

Components of NLTK contd..

- Books:
  - Natural Language Processing with Python - *Steven Bird, Edward Loper, Ewan Klein*
  - Python Text Processing with NLTK 2.0 Cookbook – *Jacob Perkins*
- Included in NLTK:
  - Installation instructions
  - API Documentation: describes every module, interface, class, and method

NLTK Modules

NLP Tasks	NLTK Modules	Functionality
Accessing Corpora	nltk.corpus	Standardized interfaces to corpora and lexicons
String Processing	nltk.tokenize, nltk.stem	Tokenizers, sentence tokenizers, stemmers
Collection Discovery	nltk.collections	t-test, chi-squared, point-wise mutual information
POS Tagging	nltk.tag	n-gram, backoff, Brill, HMM, TnT
Chunking	nltk.chunk	Regular expression, n-gram, named entity
Parsing	nltk.parse	Chart, feature-based, unification, probabilistic, dependency
Classification	nltk.classify, nltk.cluster	Decision tree, maximum entropy, naive Bayes, EM, k-means
Semantic Interpretation	nltk.sem, nltk.inference	Lambda calculus, first-order logic, model checking
Evaluation Metrics	nltk.metrics	Precision, recall, agreement coefficients
Probability Estimation	nltk.probability	Frequency distributions, smoothed probability distributions
Applications	nltk.app	Graphical concordancer, parsers, WordNet browser
Linguistics fieldwork	nltk.toolbox	Manipulate data in SIL Toolbox format

Accessing corpus using NLTK

```
>>> from nltk.corpus import brown
>>> brown.words()
['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', ...

>>> from nltk.corpus import twitter_samples
>>> twitter_samples.fileids()
[u'negative_tweets.json', u'positive_tweets.json', u'tweets.20150430-223406.json']
```

Accessing wordnet synsets

```
>>> from nltk.corpus import wordnet
>>> from nltk.corpus import wordnet as wn

>>> wn.synsets('dog')
[Synset('dog.n.01'), Synset('frump.n.01'), Synset('dog.n.03'),
Synset('cad.n.01'), Synset('frank.n.02'), Synset('pawl.n.01'),
Synset('andiron.n.01'), Synset('chase.v.01')]

>>> wn.synsets('dog', pos=wn.VERB)
[Synset('chase.v.01')]
```

Accessing sense definition, lemma

```
>>> wn.synset('dog.n.01').definition()
a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds

>>> wn.synset('dog.n.01').lemmas()
[Lemma('dog.n.01.dog'), Lemma('dog.n.01.domestic_dog'), Lemma('dog.n.01.Canis_familiaris')]

>>> [str(lemma.name()) for lemma in wn.synset('dog.n.01').lemmas() ]
['dog', 'domestic_dog', 'Canis_familiaris']
```

Accessing semantic relations

```
>>> dog = wn.synset('dog.n.01')

>>> dog.hypernyms()
[Synset('canine.n.02'), Synset('domestic_animal.n.01')]

>>> dog.hyponyms()
[Synset('basenji.n.01'), Synset('corgi.n.01'), Synset('cur.n.01'),
Synset('dalmatian.n.02'), ...]
```

WordNet Similarities

```
>>> dog = wn.synset('dog.n.01')
>>> cat = wn.synset('cat.n.01')

>>> dog.path_similarity(cat)
0.2
>>> wn.lch_similarity(dog,cat)
2.028148247292856
```

** Path similarity:Return a score denoting how similar two word senses are, based on the shortest path that connects the senses in the is-a (hypernym/hypnoym) taxonomy

**Leacock-Chodorow Similarity: Return a score denoting how similar two word senses are, based on the shortest path that connects the senses (as above) and the maximum depth of the taxonomy in which the senses occur"

Access to the Open Multilingual WordNet

```
>>> sorted(wn.langs())
['als', 'arb', 'cat', 'cmn', 'dan', 'eng', 'eus', 'fas',
'fin', 'fra', 'fre', 'glg', 'heb', 'ind', 'ita', 'jpn', 'nno',
'nob', 'pol', 'por', 'spa', 'tha', 'zsm']

>>> wn.synsets(b'\xe7\x8a\xac'.decode('utf-8'), lang='jpn')
[Synset('dog.n.01'), Synset('spy.n.01')]

>>> wn.synset('dog.n.01').lemma_names('ita')
['cane', 'Canis_familiaris']
```