WordNet and Word Sense Disambiguation

Sudha Bhingardive
Department of CSE,
IIT Bombay

Supervisor
Prof. Pushpak Bhattacharyya

WordNet

Outline

- What is WordNet?
- WordNet Synset
- Principles used for Synset Creation
- WordNet Lexico-Semantic Relations
- Important WordNets: English, Hindi, IndoWordNet, BabelNet
- Applications

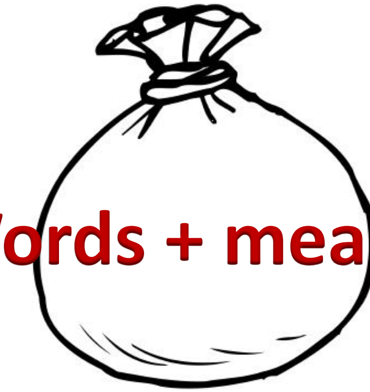
Outline

- What is WordNet?
- WordNet Synset
- Principles used for Synset Creation
- WordNet Lexico-Semantic Relations
- Important WordNets: English, Hindi, IndoWordNet, BabelNet
- Applications

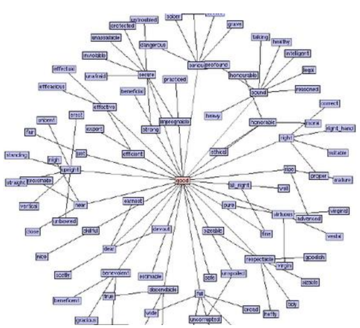
What is WordNet?

Dictionary

=

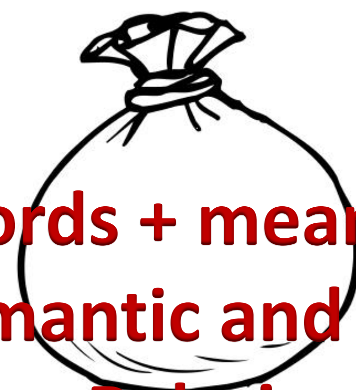


Words + meanings



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

Outline

- What is WordNet?
- WordNet Synset
- Principles used for Synset Creation
- WordNet Lexico-Semantic Relations
- Important WordNets: English, Hindi, IndoWordNet, BabelNet
- Applications

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"*

Outline

- What is WordNet?
- WordNet Synset
- Principles used for Synset Creation
- WordNet Lexico-Semantic Relations
- Important WordNets: English, Hindi, IndoWordNet, BabelNet
- Applications

Principles used for Synset Creation

- Minimality
  - The minimal set of words to make the concept unique
- Coverage
  - The maximal set of words ordered by frequency in the corpus to include all possible words standing for the sense.
- Replaceability
  - The example sentence should be such that the most frequent words in the synset can replace one another in the sentence without altering the sense.

Sysnet of bank:

depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*

Outline

- What is WordNet?
- WordNet Synset
- Principles used for Synset Creation
- WordNet Lexico-Semantic Relations
- Important WordNets: English, Hindi, IndoWordNet, BabelNet
- Applications

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’

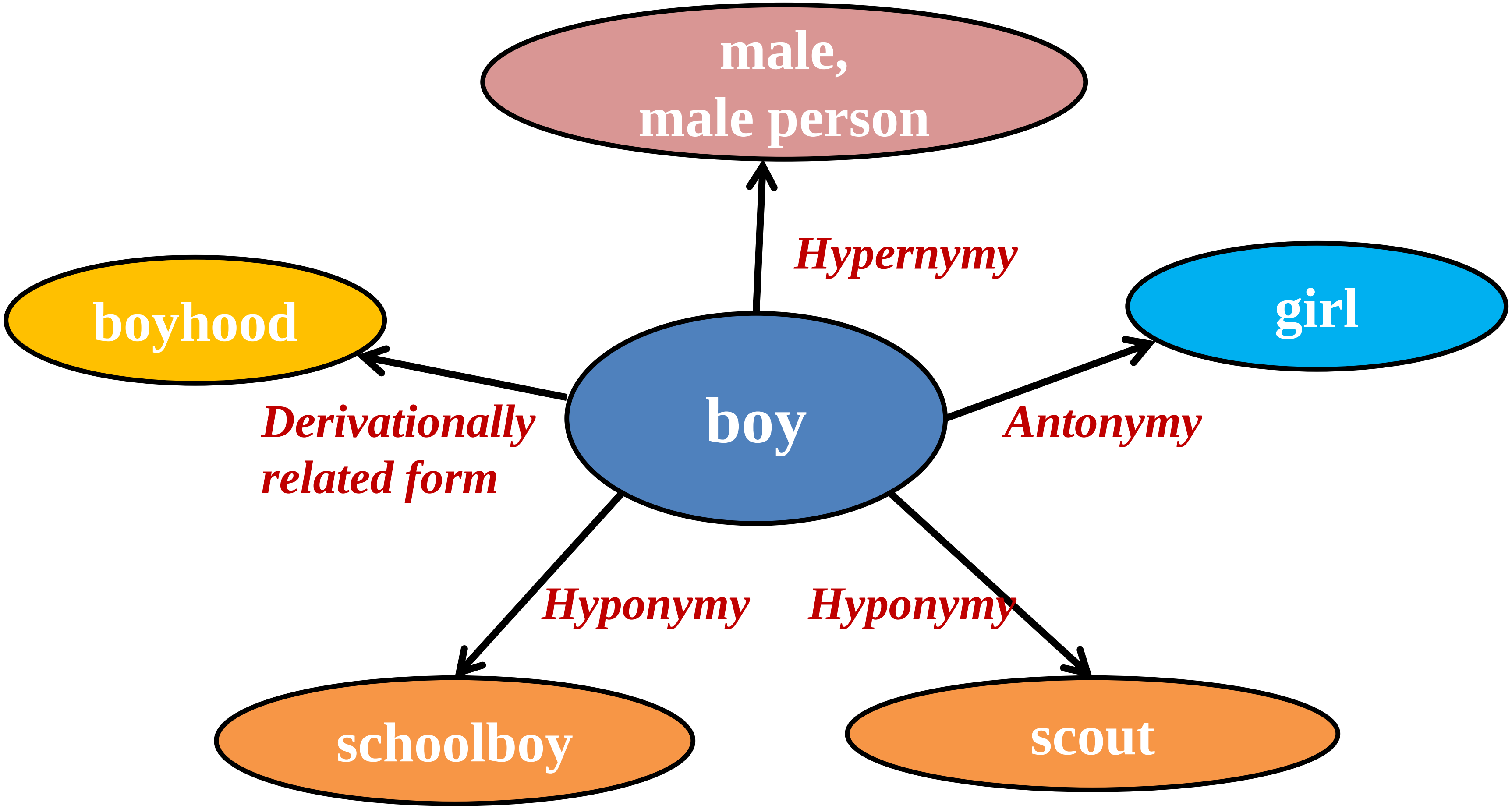
Semantic Relations contd..

- 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



Outline

- What is WordNet?
- WordNet Synset
- Principles used for Synset Creation
- WordNet Lexico-Semantic Relations
- Important WordNets: English, Hindi, IndoWordNet, BabelNet
- Applications

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

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"*
- S: (n) **boy** (a friendly informal reference to a grown man) *"he likes to play golf with the boys"*
- S: (n) son, **boy** (a male human offspring) *"their son became a famous judge"; "his boy is taller than he is"*

http://wordnetweb.princeton.edu/perl/webwn

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"

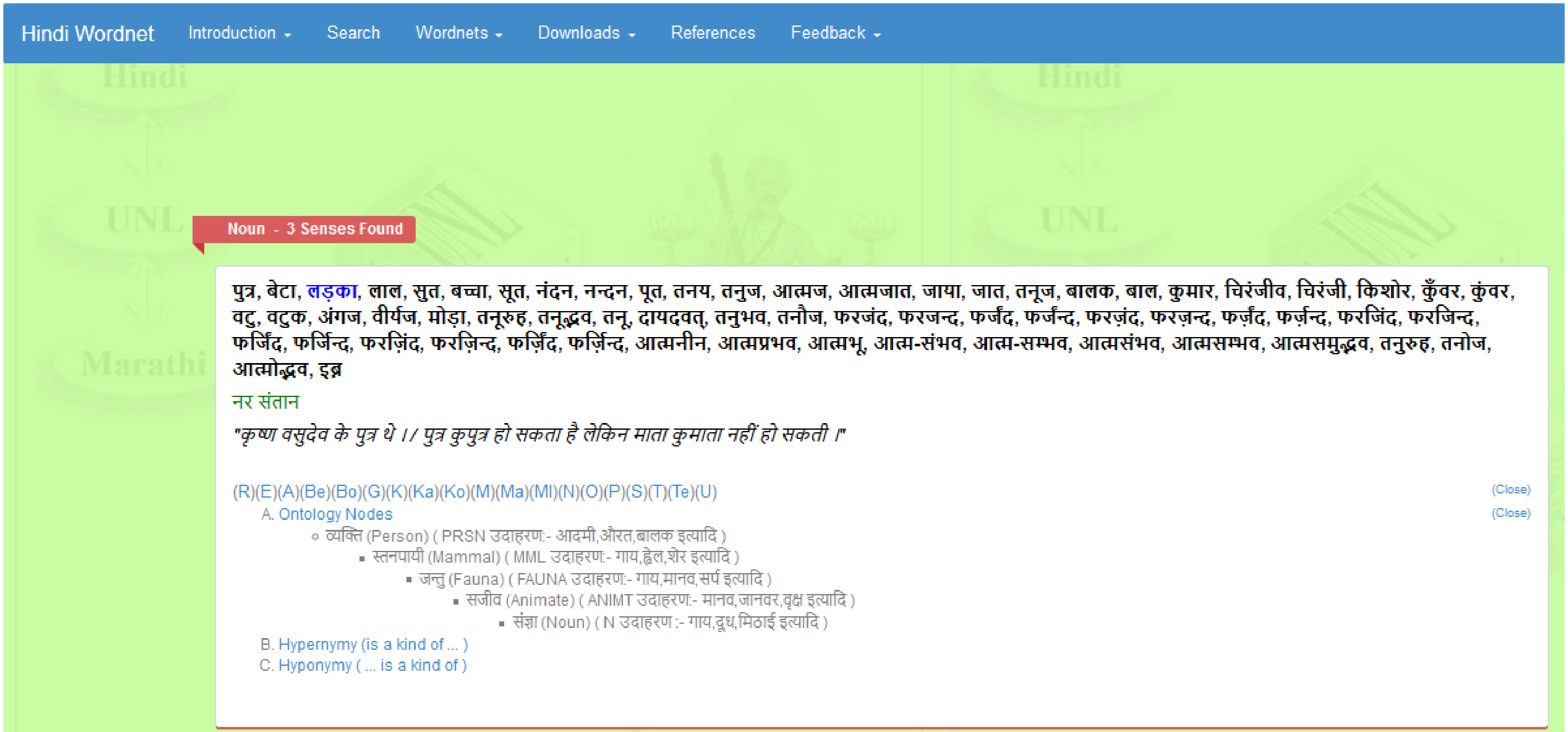
Noun

- 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"*

Hindi WordNet Interface

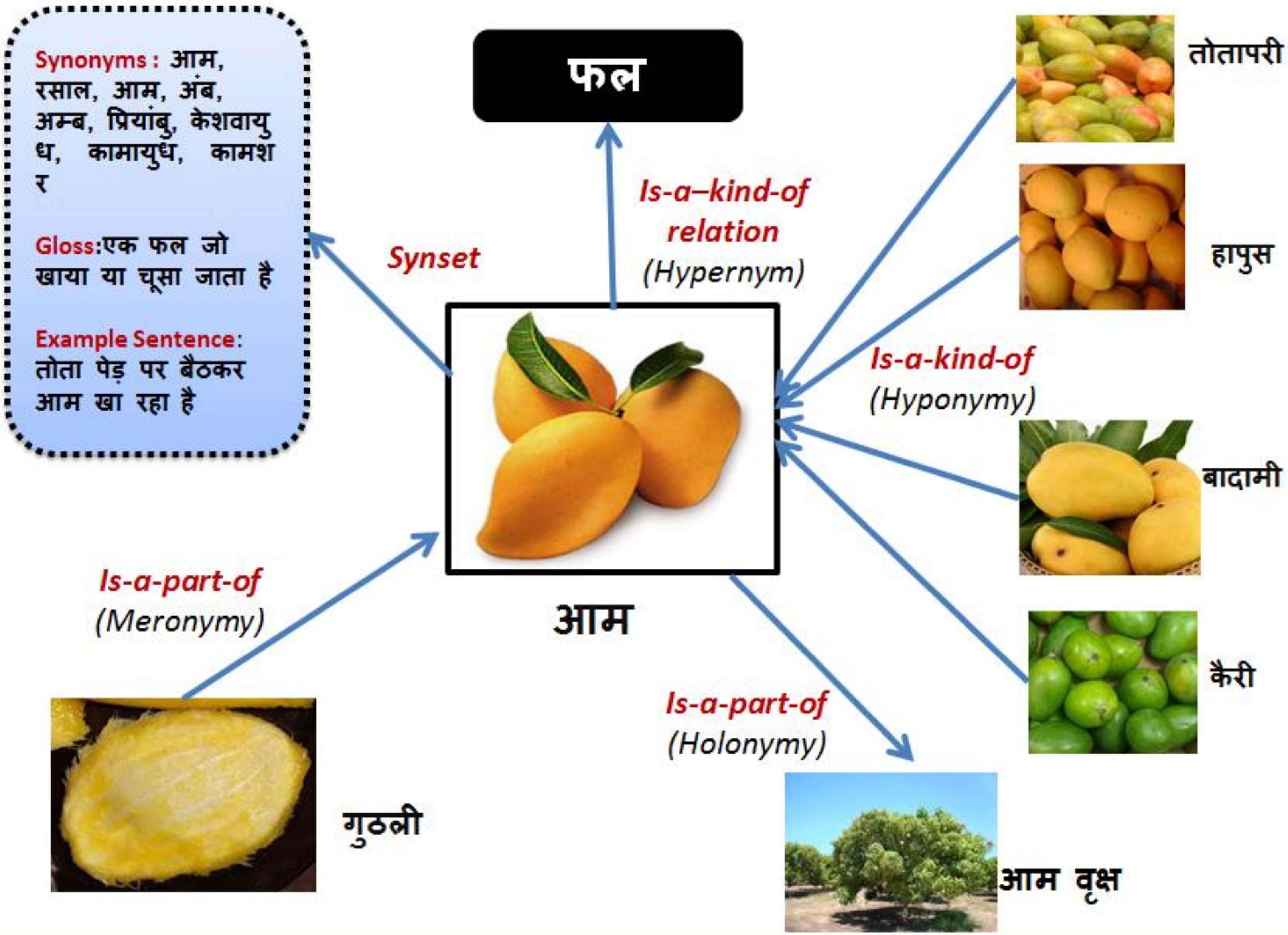
http://www.cfilt.iitb.ac.in/wordnet/webhwn/

Hindi WordNet Interface contd..



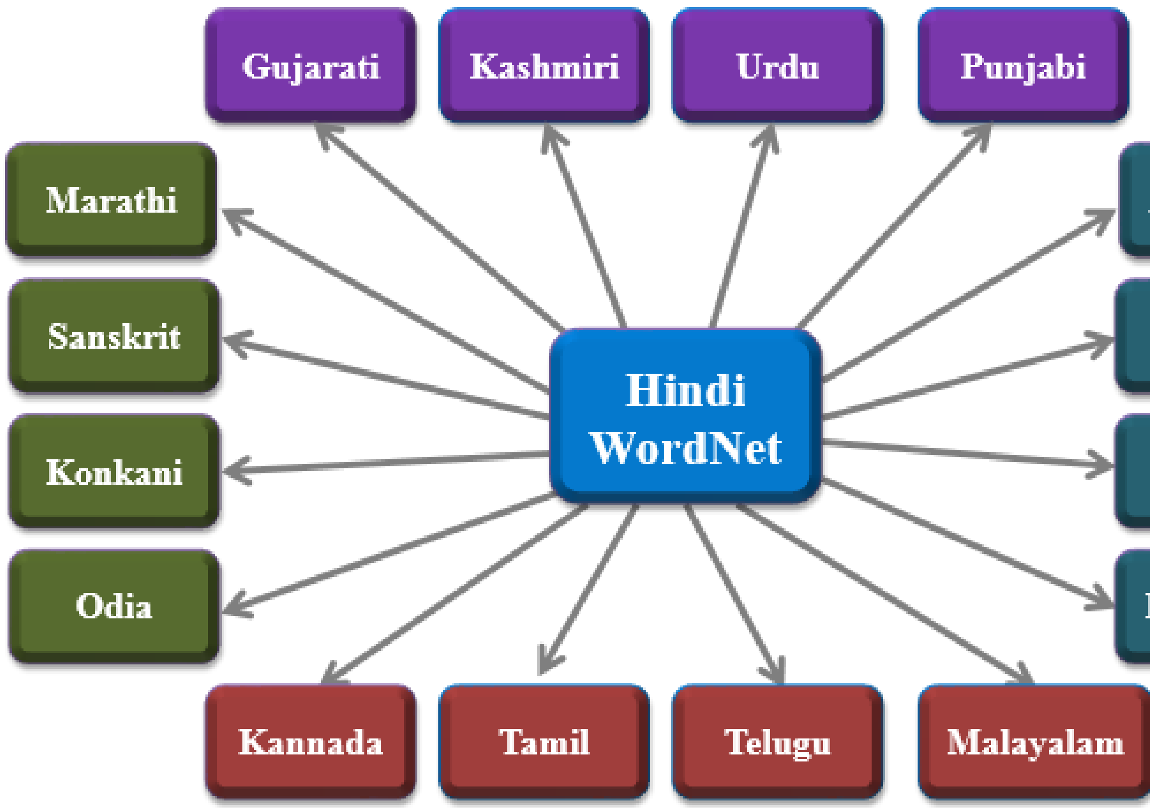
http://www.cfilt.iitb.ac.in/wordnet/webhwn/

Hindi WordNet Structure



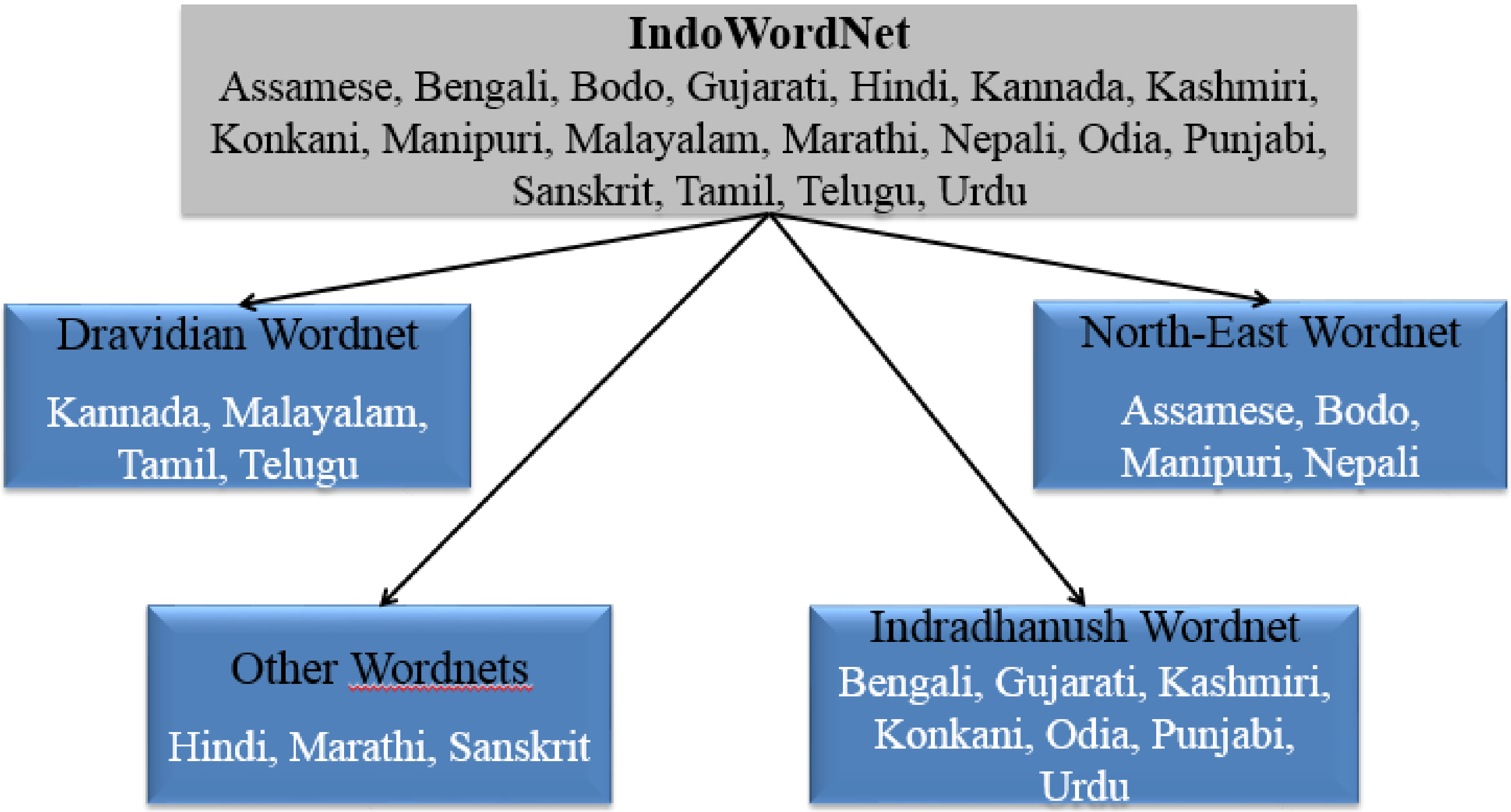
Hindi WordNet Mobile App

IndoWordNet



http://www.cfilt.iitb.ac.in/indowordnet/

IndoWordNet contd..



Institutes involved in creating IndoWordNet

- Indian Institute of Technology, Bombay – Hindi, Marathi, Sanskrit
- Goa University, Goa – Konkani
- Gauhati University, Guwahati – Assamese, Bodo
- University of Hyderabad, Hyderabad – Odia
- Jawaharlal Nehru University, New Delhi – Urdu
- Dharmsinh Desai University, Nadiad – Gujarati
- University of Kashmir, Srinagar – Kashmiri
- Punjabi University, Patiala – Punjabi
- Thapar University, Patiala – Punjabi
- Manipur University, Imphal – Manipuri
- Assam University, Silchar – Nepali
- Amrita Vishwa Vidyapeetham, Coimbatore – Malayalam
- University of Mysore, Mysore – Kannada
- Tamil University , Tanjavur – Tamil
- Dravidian University, Kuppam – Telugu

IndoWordNet Interface

http://www.cfilt.iitb.ac.in/indowordnet/

IndoWordNet linked Synset

(4265) (n)

ছেলে, বালক

কম বয়সের পুরুষ, বিশেষত অবিবাহিত

"ময়দানে ছেলেরা ক্রিকেট খেলছে"

Bengali WordNet

(4265) (n)

लड़का, बालक, बाल, बच्चा, छोकड़ा, छोरा, छोकरा

कम उम्र का पुरुष, विशेषकर अविवाहित

"मैदान में लड़के क्रिकेट खेल रहे हैं ।"

Hindi WordNet

(4265) (n)

मुलगा,पोरगा,पोर,पोरगे

साधारणतः सोळा वर्षांखालील पुरुष व्यक्ती

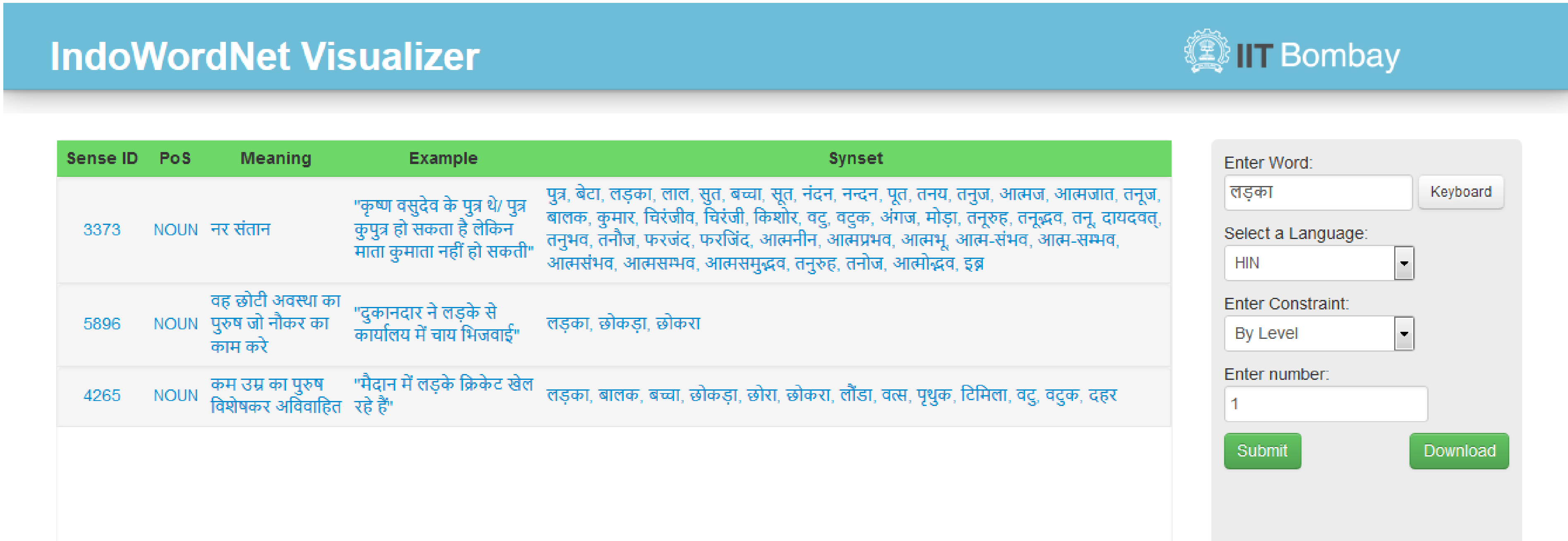
"तो मुलगा खूपच हुशार आहे"

Marathi WordNet

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

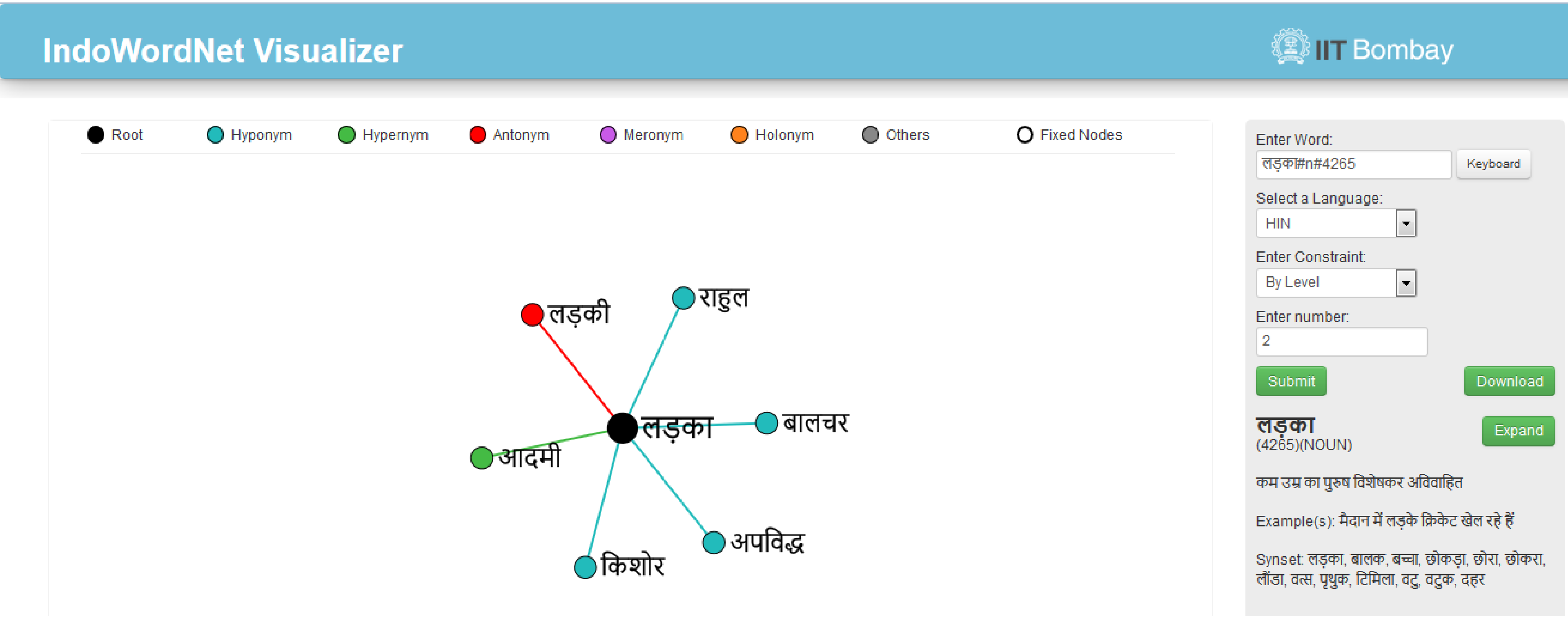
IndoWordNet Visualizer Interface



Sense ID	PoS	Meaning	Example	Synset
3373	NOUN	नर संतान	"कृष्ण वसुदेव के पुत्र थे/ पुत्र कुपुत्र हो सकता है लेकिन माता कुमाता नहीं हो सकती"	पुत्र, बेटा, लड़का, लाल, सुत, बच्चा, सूत, नंदन, नन्दन, पूत, तनय, तनुज, आत्मज, आत्मजात, तनूज, बालक, कुमार, चिरंजीव, चिरंजी, किशोर, वटु, वटुक, अंगज, मोड़ा, तनूरुह, तनूद्भव, तनू, दायदवत्, तनुभव, तनौज, फरजंद, फरजिंद, आत्मनीन, आत्मप्रभव, आत्मभू, आत्म-संभव, आत्म-सम्भव, आत्मसंभव, आत्मसम्भव, आत्मसमुद्भव, तनुरुह, तनोज, आत्मोद्भव, इब्न
5896	NOUN	वह छोटी अवस्था का पुरुष जो नौकर का काम करे	"दुकानदार ने लड़के से कार्यालय में चाय भिजवाई"	लड़का, छोकड़ा, छोकरा
4265	NOUN	कम उम्र का पुरुष विशेषकर अविवाहित	"मैदान में लड़के क्रिकेट खेल रहे हैं"	लड़का, बालक, बच्चा, छोकड़ा, छोरा, छोकरा, लौंडा, वत्स, पृथुक, टिमिला, वटु, वटुक, दहर

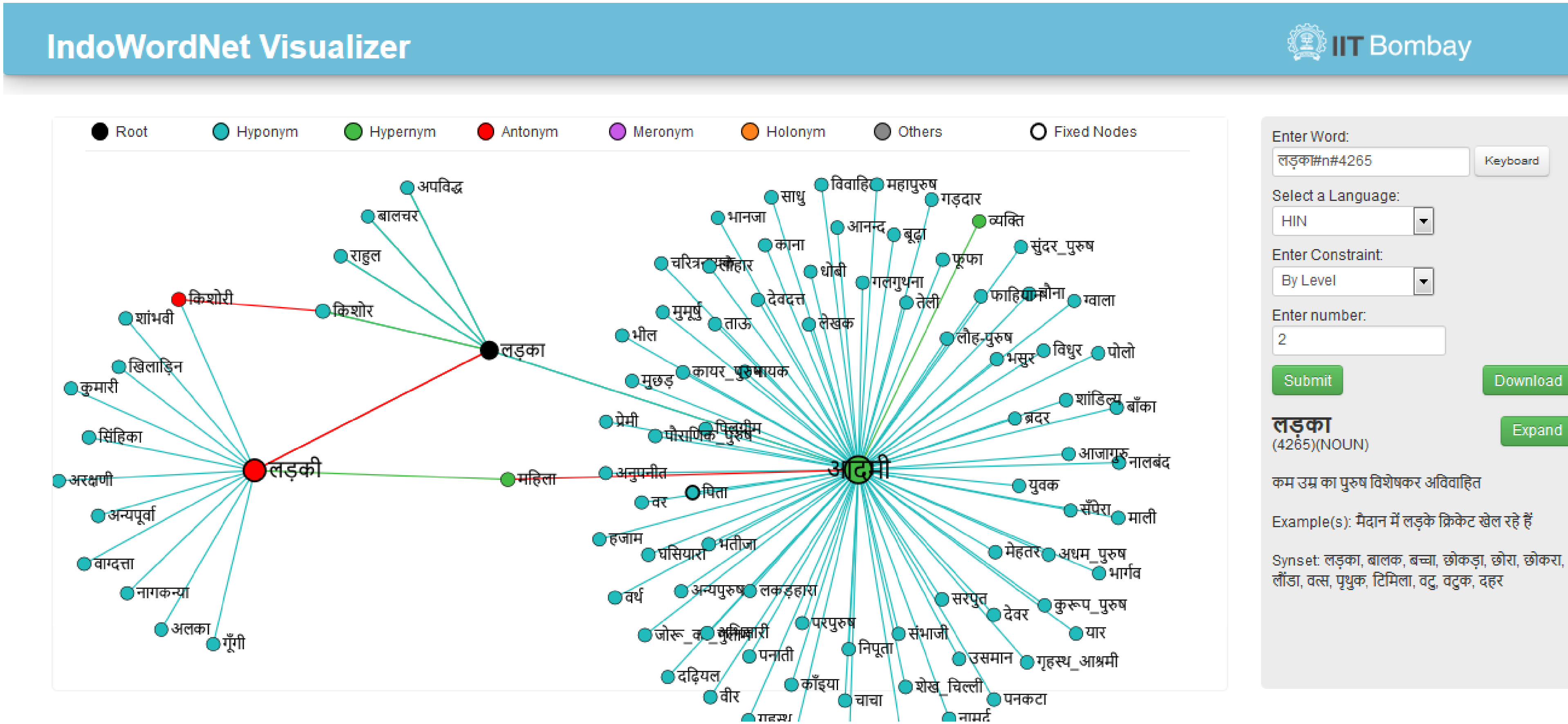
http://www.cfilt.iitb.ac.in/Drawgraph/input.html

IndoWordNet Visualizer contd..



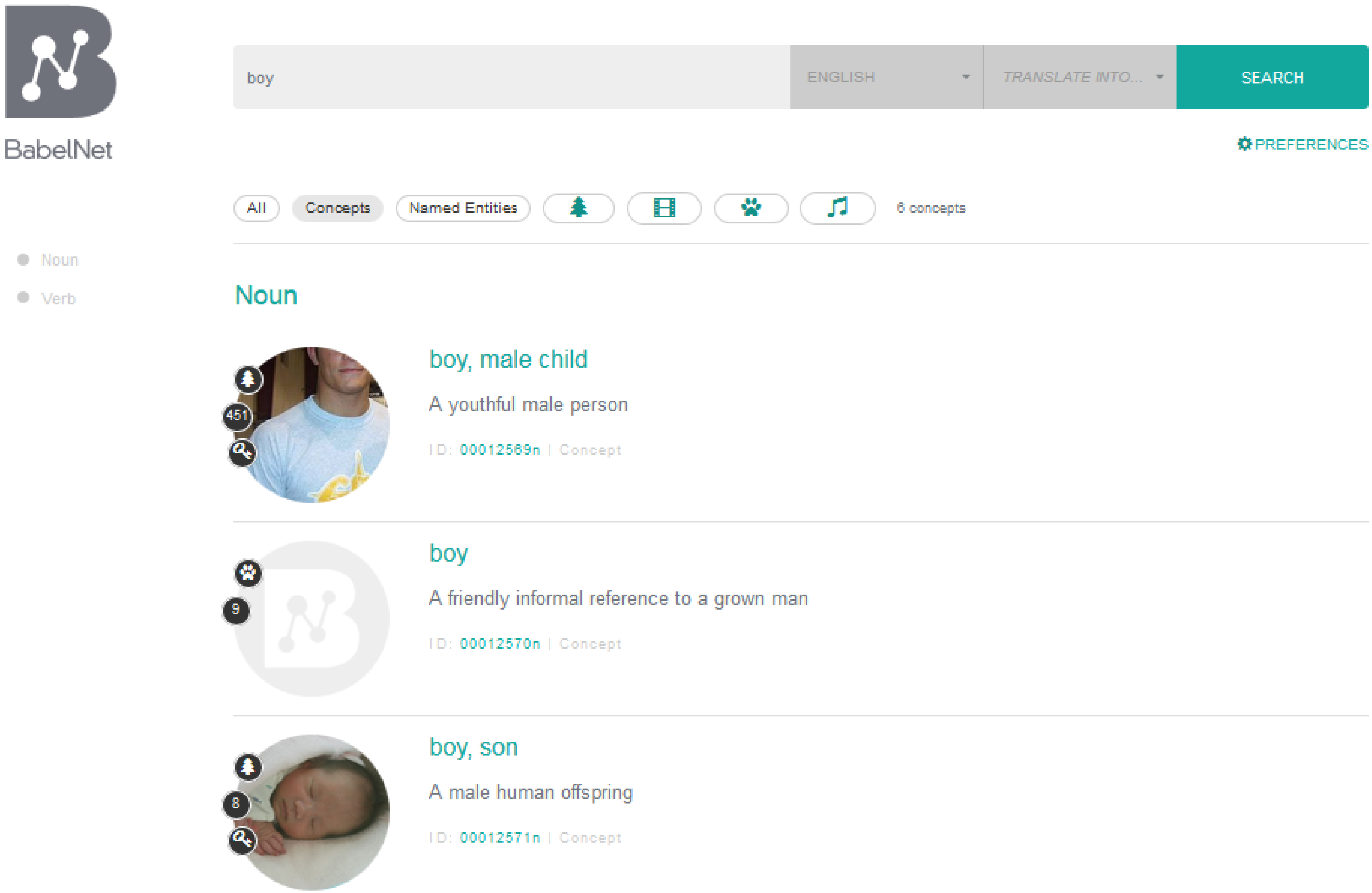
http://www.cfilt.iitb.ac.in/Drawgraph/input.html

IndoWordNet Visualizer contd..



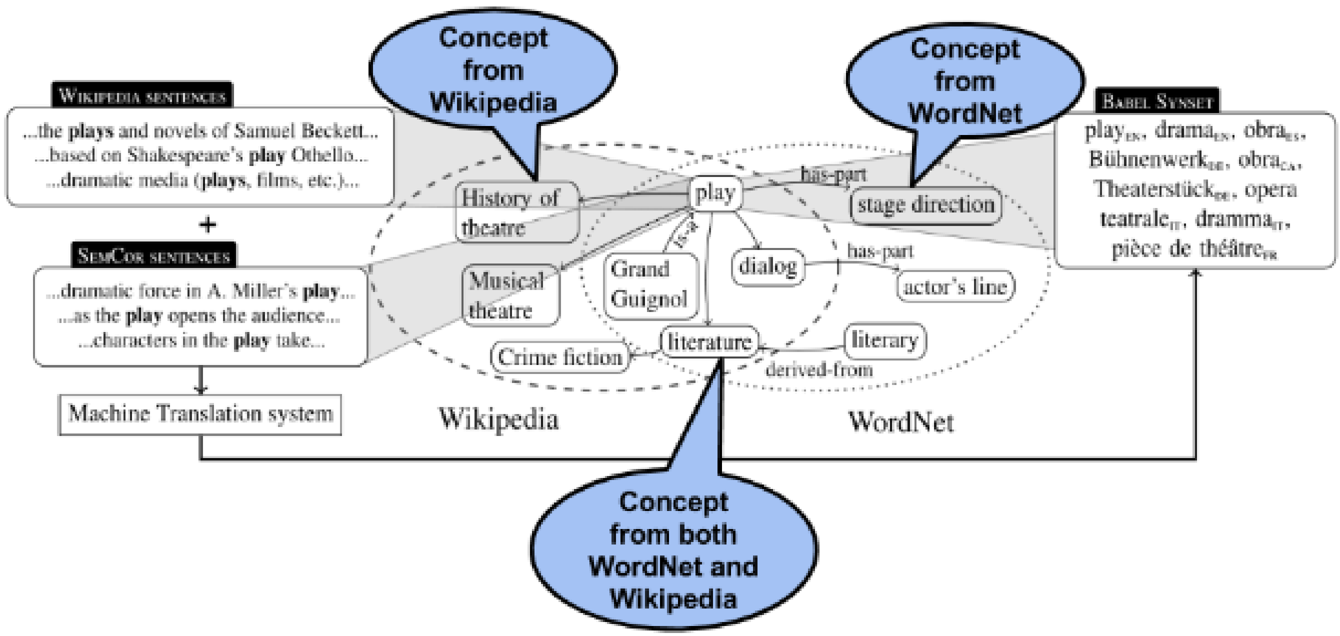
http://www.cfilt.iitb.ac.in/Drawgraph/input.html

BabelNet Interface



http://babelnet.org/

BabelNet Synset



http://babelnet.org/

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 API's and similarity tools

- English:
  - Java API: extJWNL , JAWS, JWNL
  - Python API: NLTK
  - WordNet::Similarity tool
- Hindi:
  - Java API: JHWNL
  - Python API
  - IndoWordNet::Similarity tool

Outline

- What is WordNet?
- WordNet Synset
- Principles used for Synset Creation
- Lexico-Semantic Relations
- Important WordNets: English, Hindi, IndoWordNet, BabelNet, EuroWordNet
- Applications

WordNet Applications

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

Word Sense Disambiguation

Outline

- Introduction
  - Ambiguity
  - WSD Definition
  - Position of WSD in NLP layers
- Motivation
- WSD block diagram
- Lexical Resources needed
  - Sense Repository
  - Sense Annotated Corpus
- WSD approaches
  - Knowledge based
  - Corpus based (Supervised, Unsupervised)
- Applications

Outline

- Introduction
  - Ambiguity
  - WSD Definition
  - Position of WSD in NLP layers
- Motivation
- WSD block diagram
- Lexical Resources needed
  - Sense Repository
  - Sense Annotated Corpus
- WSD approaches
  - Knowledge based
  - Corpus Based (Supervised, Unsupervised)
- Applications

Ambiguity

- A word, phrase or sentence is ambiguous if it has more than one meaning

- Structural ambiguity: due to the sentence structure
  - *A boy saw a man with a telescope* (English)
  - राम ने दौड़ते हुए शेर को देखा (Hindi)

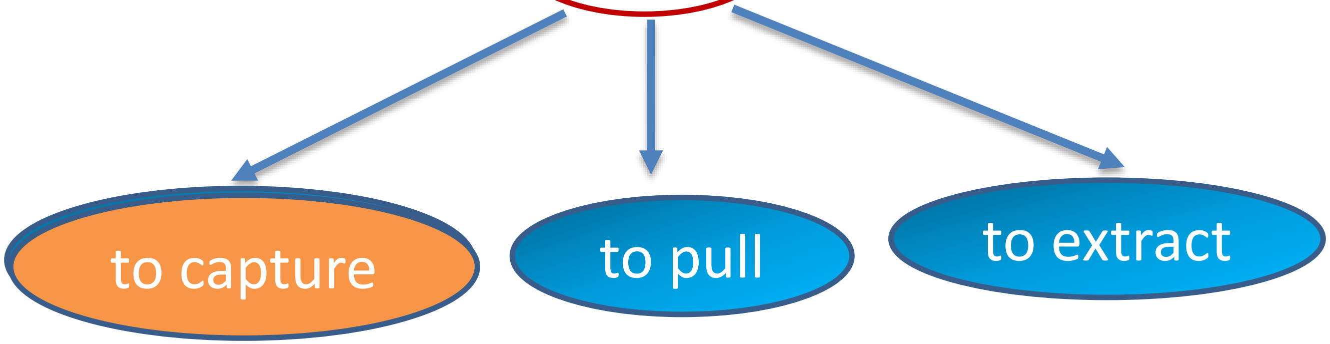
- Lexical ambiguity: due to polysemous words
  - *She put her glasses on the table* (English)
  - पड़ोसी ने हमारे घर में आग लगायी (Hindi)

WSD Definition

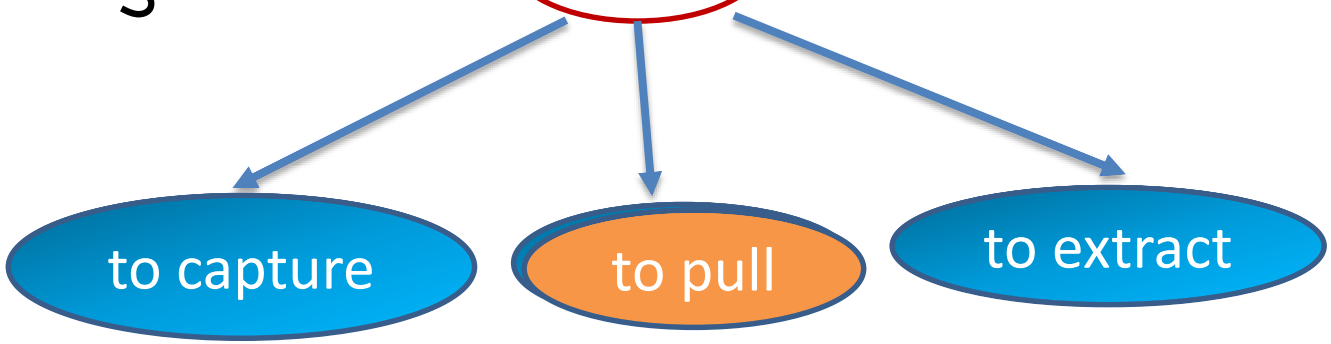
- Word Sense Disambiguation (WSD) is the problem of computationally determining the ‘sense’ or ‘meaning’ of a word in a particular context.

WSD Example

यहाँ फोटो खींचना मना है



वह कुएँ से रस्सी खींचती है



Why WSD is difficult?

- Sometimes human even fails to disambiguate

‘उसका हाथ मशीन के नीचे आ गया’

1. **हाथ**, बाज़ू, हस्त, बाँह, बाहु, बाजू, भुजा, कर, - कन्धे से पंजे तक का वह अंग जिससे चीज़ें पकड़ते और ... *...धीजी के हाथ बहुत लंबे थे । / भीम की भुजाओं में बहुत बल था।*
2. **हाथ**, कर, पंजा, पाणि - ... भाग *"उसका हाथ मशीन के नीचे आ गया ।"*
3. **हाथ**, हस्त, कर, पाणि - कोहनी से पंजे के सिरे तक का भाग *“दुर्घटना में उसका दाहिना हाथ टूट गया ।“*
4. **हाथ**, हस्त - चौबीस अंगुल की एक नाप य... सिरे तक की लंबाई की नाप ,*"इस वस्त्र की लंबाई दो हाथ है ...*
5. **हाथ** - ताश के खेल में एक दौर में गिरने वाले ... उसके बाद खेल से बाहर हो जाएँ,*"मेरे सात हाथ बन चुके हैं ।"*

Fine-grained senses

Coarse-grained senses

Why WSD is difficult? contd..

- From practical point of view, it is essential to make sense distinction according to the needs of the application

- Coarse grained senses – Information Retrieval, Information Extraction, Document Categorization, Machine Translation

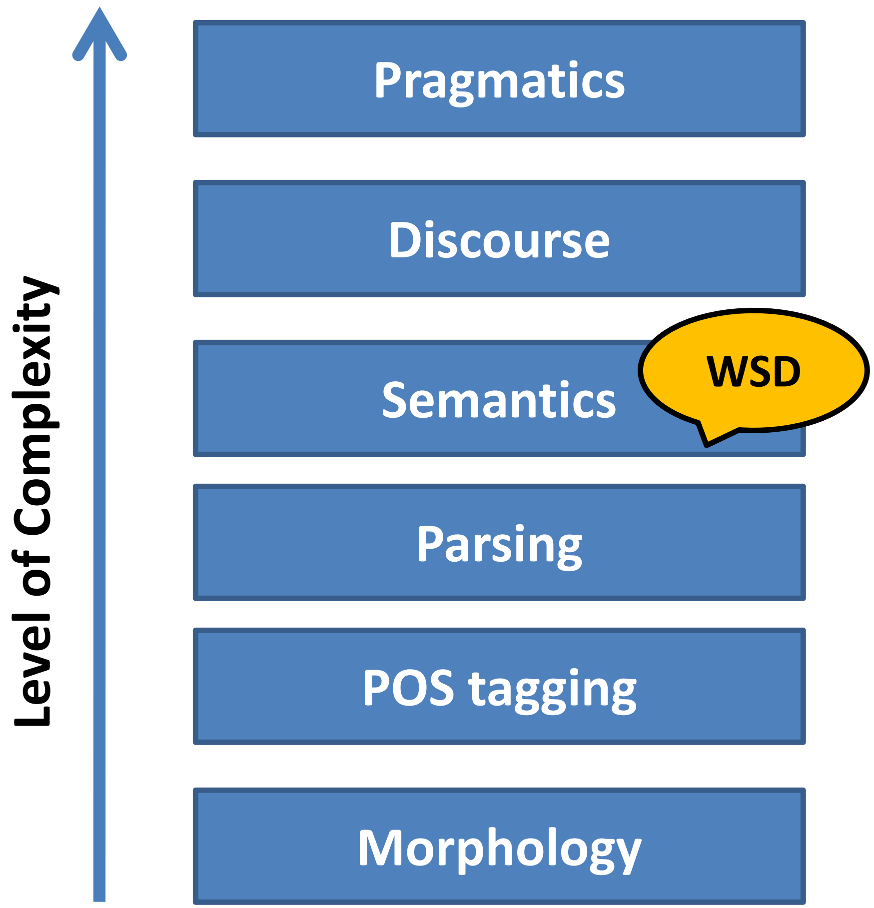
- Fine grained senses – Language Learning, Machine Translation of distant languages like Chinese-English

Why WSD is difficult? contd..

- Generally verbs are more polysemous as compared to other parts-of-speech

Verb	#Senses	Verb	#Senses	Verb	#Senses
निकलना	31	निकालना	29	लगना	26
लगाना	23	चढ़ना	22	उतरना	21
पड़ना	19	आना	19	छोड़ना	19
चढ़ाना	18	चलना	17	उठना	17
मिलना	17	देखना	16	बोलना	16
टूटना	16	खुलना	15	उड़ाना	15
उठाना	14	खोलना	14	छूटना	14
बनाना	14	लेना	13	रहना	13
जमना	12	बांधना	12	बैठना	12
खाना	12	काटना	12	बाँधना	12

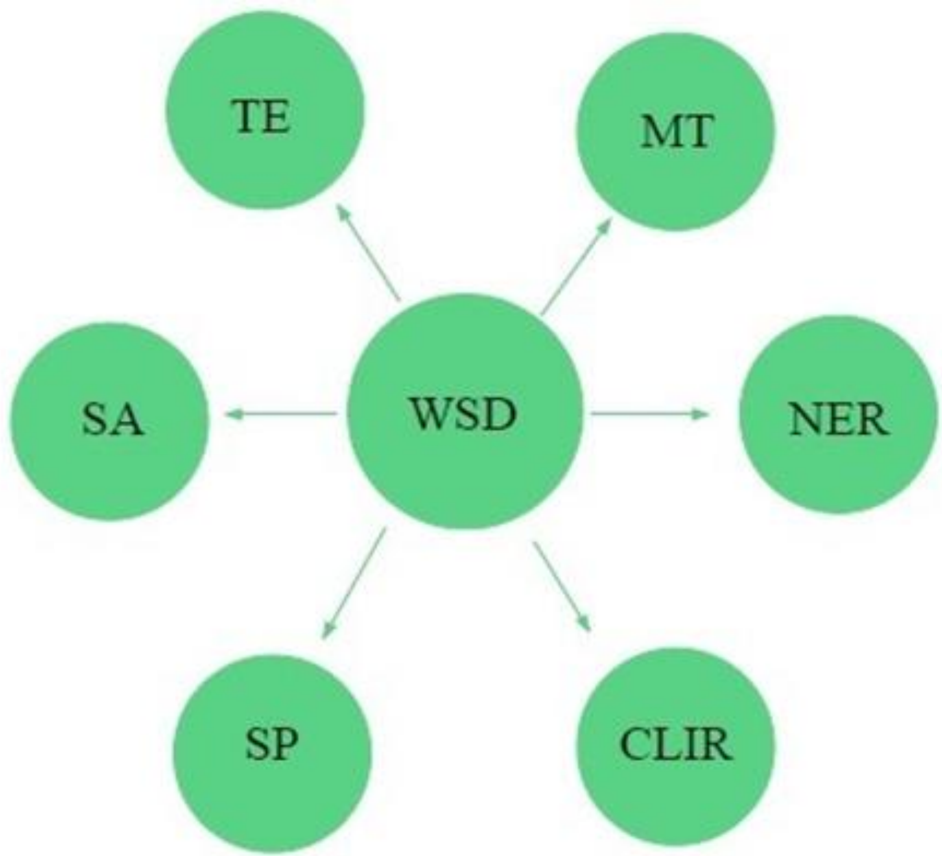
Position of WSD in NLP layers



Outline

- Introduction
  - Ambiguity
  - WSD Definition
  - Position of WSD in NLP layers
- Motivation
- WSD block diagram
- Lexical Resources needed
  - Sense Repository
  - Sense Annotated Corpus
- WSD approaches
  - Knowledge based
  - Corpus Based (Supervised, Unsupervised)
- Applications

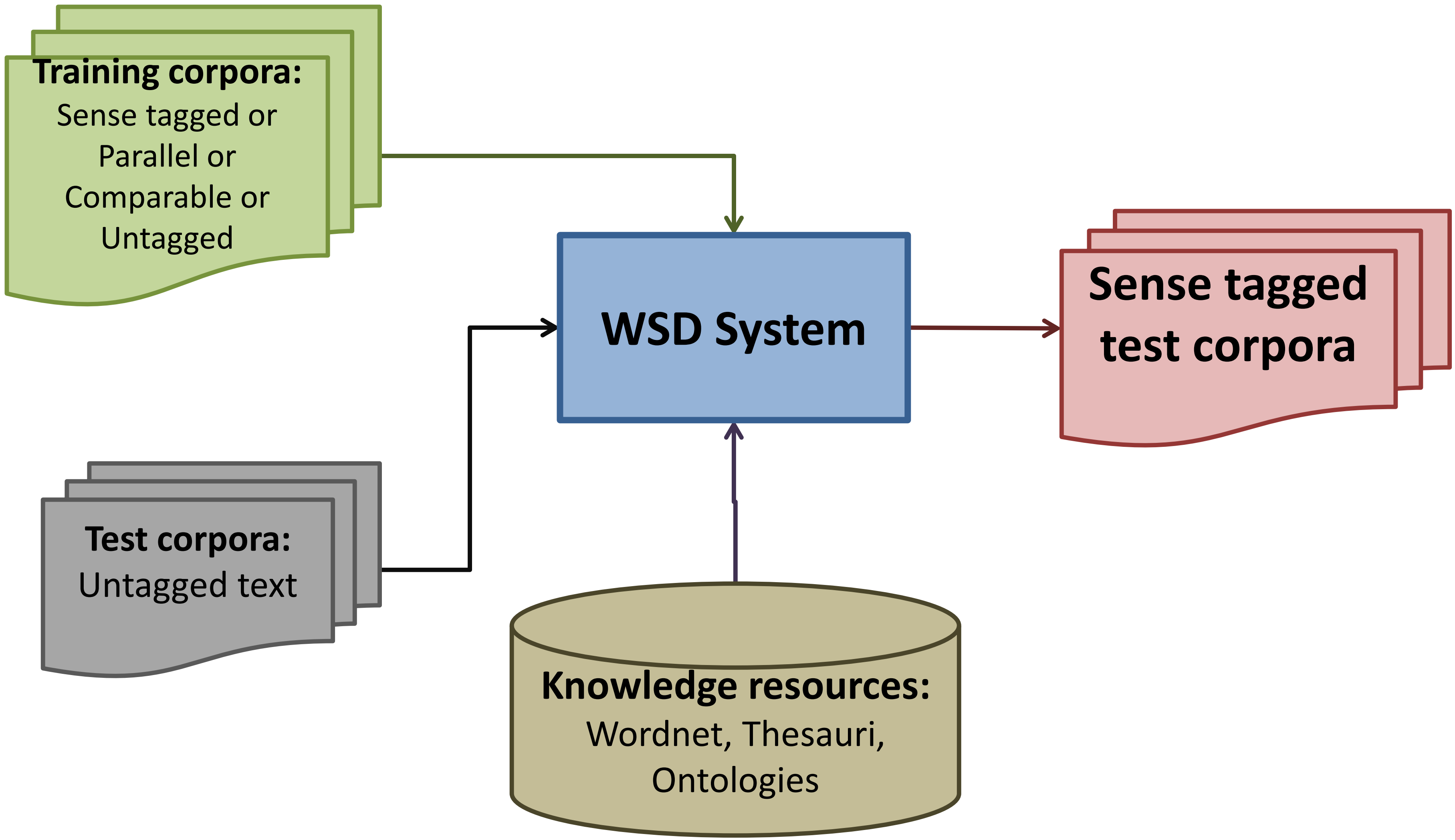
Motivation



Outline

- Introduction
  - Ambiguity
  - WSD Definition
  - Position of WSD in NLP layers
- Motivation
- WSD block diagram
- Lexical Resources needed
  - Sense Repository
  - Sense Annotated Corpus
- WSD approaches
  - Knowledge based
  - Corpus Based (Supervised, Unsupervised)
- Applications

Block diagram of WSD



Outline

- Introduction
  - Ambiguity
  - WSD Definition
  - Position of WSD in NLP layers
- Motivation
- WSD block diagram
- Lexical Resources needed
  - Sense Repository
  - Sense Annotated Corpus
- WSD approaches
  - Knowledge based
  - Corpus Based (Supervised, Unsupervised)
- Applications

Lexical Resources for WSD

- Sense Repository
  - Dictionary
  - Thesaurus
  - Wordnet
- Sense Annotated Corpus

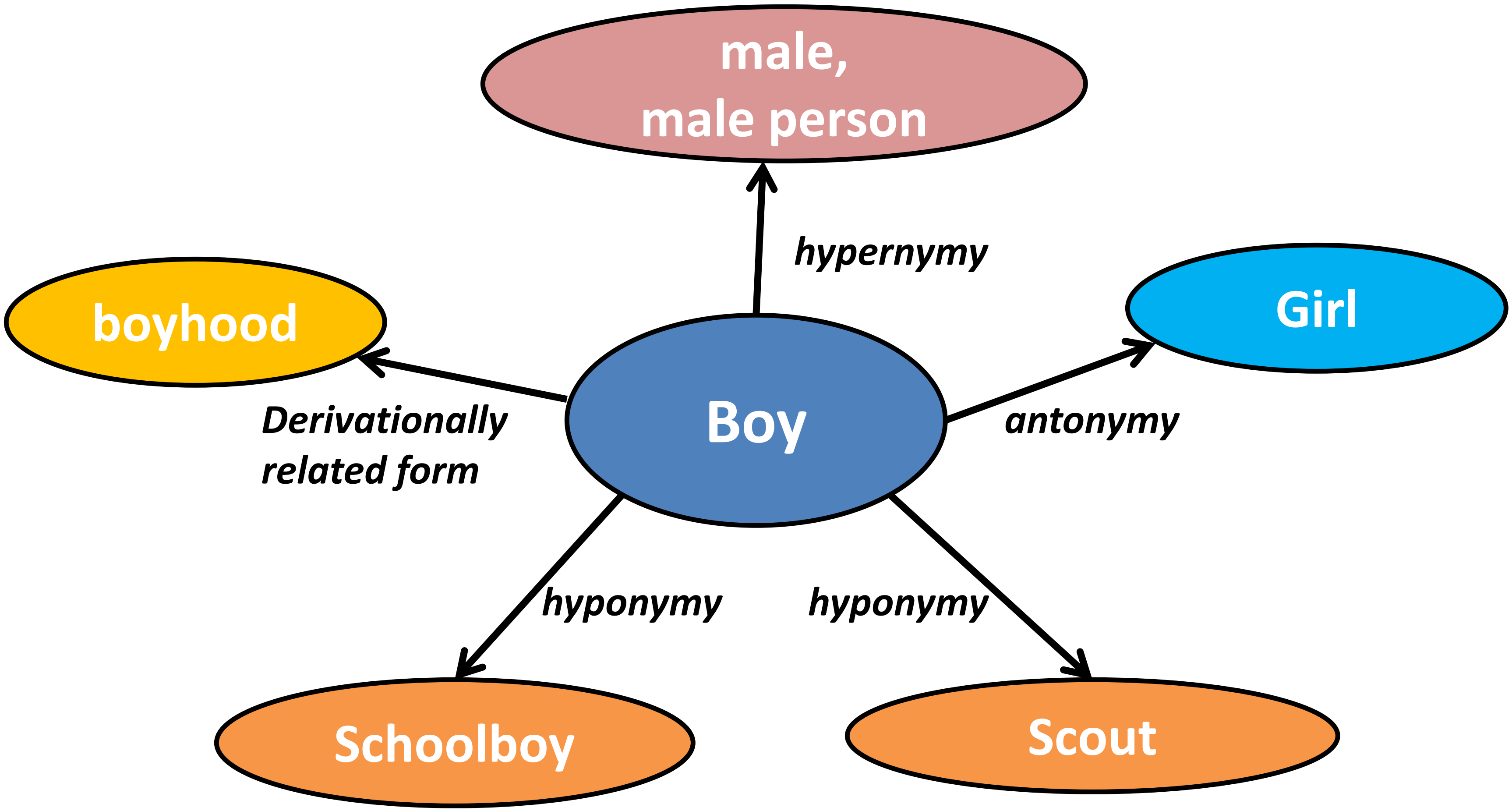
WordNet

- Lexical knowledge base
- Consists of synsets and semantic relations
- For example: Senses of ‘boy’ from WordNet
  - (10305010**)** 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"*

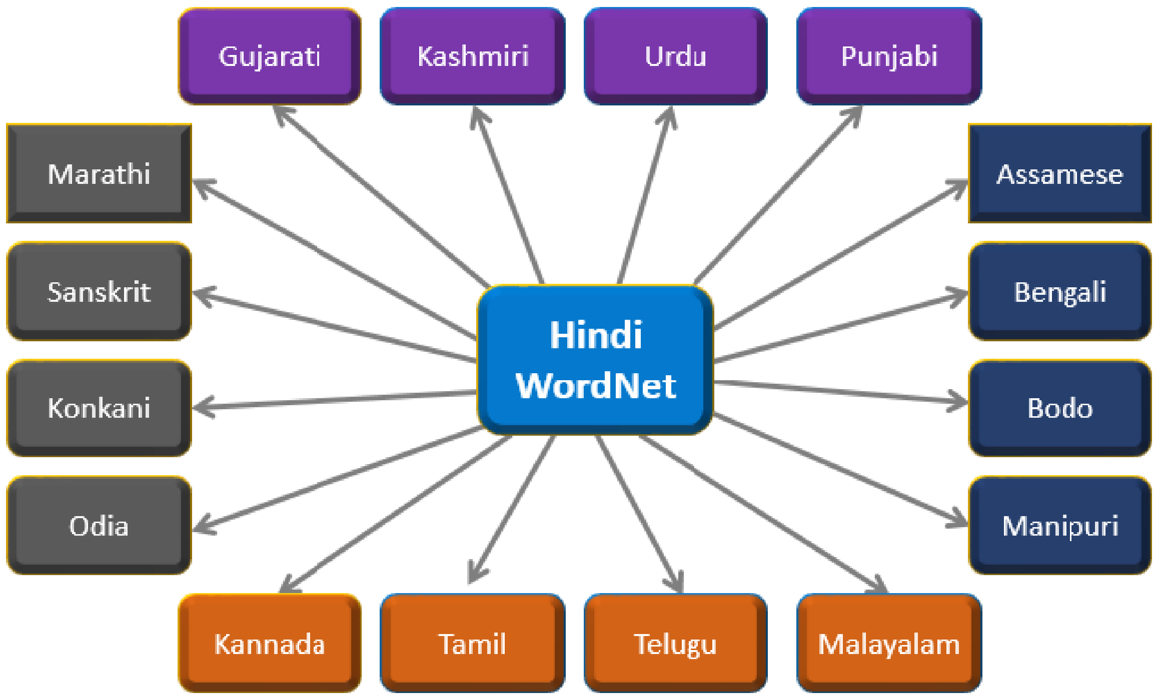
  - (09890332**)** S: (n) **boy** (a friendly informal reference to a grown man) *"he likes to play golf with the boys"*

  - (10643436**)** S: (n) **son**, **boy** (a male human offspring) *"their son became a famous judge"; "his boy is taller than he is"*

WordNet :Lexico-Semantic relations



IndoWordNet



IndoWordNet Structure

http://www.cfilt.iitb.ac.in/indowordnet/

IndoWordNet Synset

(4265) (n)

ছেলে, বালক

কম বয়সের পুরুষ, বিশেষত অবিবাহিত

"ময়দানে ছেলেরা ক্রিকেট খেলছে"

Bengali WordNet

(4265) (n)

लड़का, बालक, बाल, बच्चा, छोकड़ा, छोरा, छोकरा

कम उम्र का पुरुष, विशेषकर अविवाहित

"मैदान में लड़के क्रिकेट खेल रहे हैं ।"

Hindi WordNet

(4265) (n)

मुलगा,पोरगा,पोर,पोरगे

साधारणतः सोळा वर्षांखालील पुरुष व्यक्ती

"तो मुलगा खूपच हुशार आहे"

Marathi WordNet

Sense Annotated Corpus

- Corpus annotated with sense tags from wordnet
  - English corpus:
    - **SemCor Corpus, OntoNotes, DSO, Senseval , SemLink**
  - Indian language corpus:
    - **CFILT corpus (Hindi and Marathi Health-Tourism)**
  - Japanese corpus
    - **Jsemcor corpus**
  - Dutch corpus:
    - **DutchSemCor**
  - Spanish corpus:
    - **SpsemCor**

Sense Annotated Corpus contd..

CFILT corpus: (Hindi-health domain)

- व्यायाम_5939 शरीर_1961 को स्वस्थ_1831 और तन्दुरुस्त_1831 रखने_ में सहायता_3623 करता है
- दैनिक_6246 व्यायाम_5939 सबसे उत्कृष्ट_2360 लाभ_2751 प्रदान_1694 करते_ हैं
- स्वास्थ्य_8407 शारीरिक_9166 , मानसिक_2151 और सामाजिक_3540 सुख_3538 की एक_187 अवस्था_652 है
- इसमें केवल_4509 बीमारी_1423 की अनुपस्थिति_6745 से भी अधिक_2403 शामिल_10810 है

Outline

- Introduction
  - Ambiguity
  - WSD Definition
  - Position of WSD in NLP layers
- Motivation
- WSD block diagram
- Lexical Resources needed
  - Sense Repository
  - Sense Annotated Corpus
- WSD approaches
  - Knowledge based
  - Corpus Based (Supervised, Unsupervised)
- Applications

WSD approaches

- Knowledge-based WSD:
  - uses an explicit lexicon (machine readable dictionary (MRD), thesaurus) or ontology (e.g. WordNet).
- Corpus-based WSD: (Supervised & Unsupervised)
  - the relevant information about word senses is gathered from training on a large corpus.
- Hybrid approach:
  - combining aspects of both of the aforementioned methodologies

Knowledge-based WSD

Algorithm	Accuracy
WSD using Selectional Restrictions	44% on Brown Corpus
Lesk's algorithm	50-60% on short samples of *"Pride and Prejudice"* and some "*news stories*".
WSD using conceptual density	54% on Brown corpus.
WSD using Random Walk Algorithms	54% accuracy on SEMCOR corpus which has a baseline accuracy of 37%.
Walker's algorithm	50% when tested on 10 highly polysemous English words.

Simple Lesk Algorithm

- Example: **pine cone**

pine	1	kinds of evergreen tree with needle-shaped leaves
	2	waste away through sorrow or illness
cone	1	solid body which narrows to a point
	2	something of this shape whether solid or hollow
	3	fruit of certain evergreen trees

- Dictionary definitions of pine1 and cone3 literally overlap: “evergreen” + “tree”
- So “pine cone” must be pine1 + cone3

Simplified Lesk Algorithm

The **bank** can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

given the following two WordNet senses:

$bank^1$	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
$bank^2$	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

- Count words in the context (sentence) which are also in the Gloss or Example for 1 and 2;
- Choose the word-sense with most "overlap"

Corpus Based approaches

- A corpus-based approach extracts information on word senses from a large annotated data collection.
- Distributional information about an ambiguous word refers to the frequency distribution of its senses
- collocational or co-occurrence information
- part-of-speech
- …

Corpus Based approaches

- There are two possible approaches to corpus-based WSD systems:

  - **Supervised approaches**
    - use annotated training data
    - basically amount to a classification task

  - **Unsupervised algorithms**
    - applied to raw text material
    - annotated data is only needed for evaluation
    - correspond to a clustering task rather than a classification.

  - **Bootstrapping**
    - looks like supervised approaches
    - it needs only a few seeds instead of a large number of training examples

Supervised Approaches

Approach	Average Precision	Average Recall	Corpus	Average Baseline Accuracy
Naïve Bayes	64.13%	Not reported	Senseval3 – All Words Task	60.90%
Decision Lists	96%	Not applicable	Tested on a set of 12 highly polysemous English words	63.9%
Exemplar Based disambiguation (k-NN)	68.6%	Not reported	WSJ6 containing 191 content words	63.7%
SVM	72.4%	72.4%	Senseval 3 – Lexical sample task (Used for disambiguation of 57 words)	55.2%
Perceptron trained HMM	67.60	73.74%	Senseval3 – All Words Task	60.90%

Unsupervised approaches

- Supervised WSD performs well but needs sense tagged corpora
- Obtaining sense tagged corpora is costly in terms of time and money
- A high degree of language dependence and makes it difficult to apply them to a variety of languages
- Despite of the less accuracy, unsupervised approaches are chosen for their resource consciousness and robustness

Classification of Unsupervised WSD Methods

Strictly Unsupervised Methods

Distributional Methods

Translational Equivalence Methods

- Brown et al

Type Based

- Hyperlex
- Latent Semantic Indexing (LSA)
- Hyper Space Analogue to Language (HAL)
- Clustering by Committee (CBC)

Token Based

- Context Group Discrimination
- McQuitty's Similarity Analysis

Unsupervised Approaches

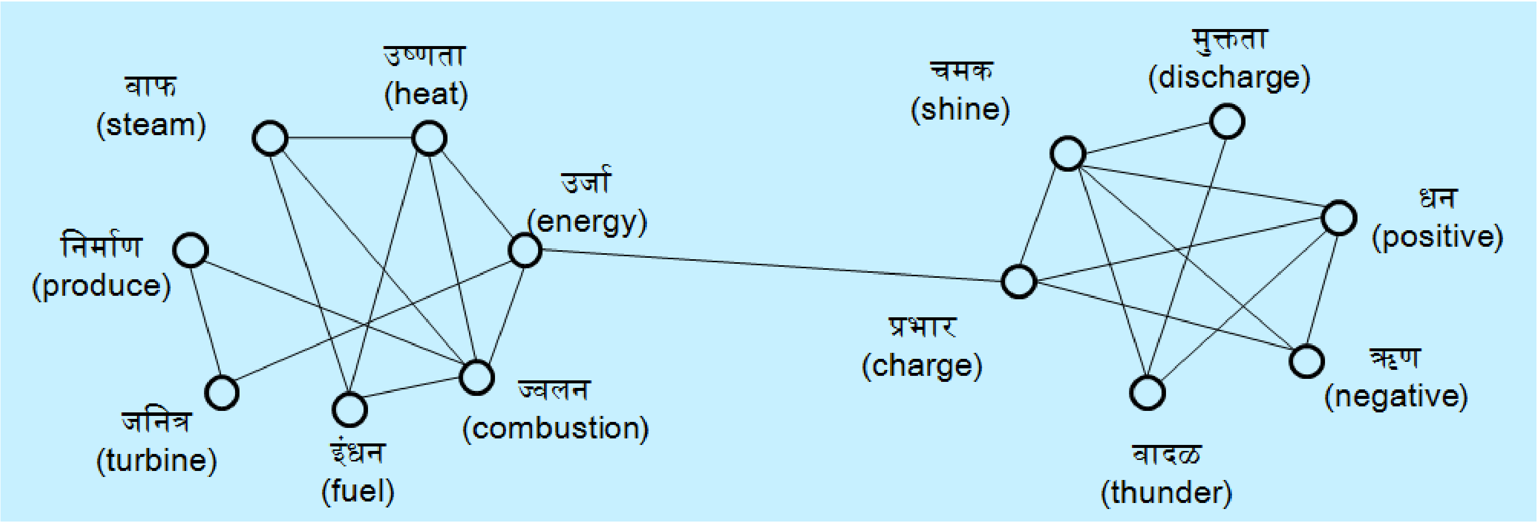
Approach	Precision	Average Recall	Corpus	Baseline
Lin's Algorithm	**68.5%.** The result was considered to be correct if the similarity between the predicted sense and actual sense was greater than 0.27	Not reported	**Trained using WSJ corpus containing 25 million words. Tested on 7 SemCor** files containing **2832** polysemous **nouns.**	**64.2%**
Hyperlex	**97%**	82% (words which were not tagged with confidence>threshold were left untagged)	Tested on a set of 10 highly polysemous French words	**73%**
WSD using Roget's Thesaurus categories	**92%** (average degree of polysemy was 3)	Not reported	Tested on a set of 12 highly polysemous English words	**Not reported**
WSD using parallel corpora	**SM: 62.4%** **CM: 67.2%**	SM: 61.6% CM: 65.1%	Trained using a English Spanish parallel corpus Tested using Senseval 2 – All Words task (only nouns were considered)	**Not reported**

Hyperlex (Veronis, 2004)

- Target word WSD developed for Information Retrieval applications
- Instead of using “dictionary defined senses” extract the “senses from the corpus” itself
- Works only for nouns and adjectives
- Co-occurrence graph is constructed for words which co-occur with the target word
- Words which are syntactically correlated are connected with edges
- Weight of an edge is determined by following formula:

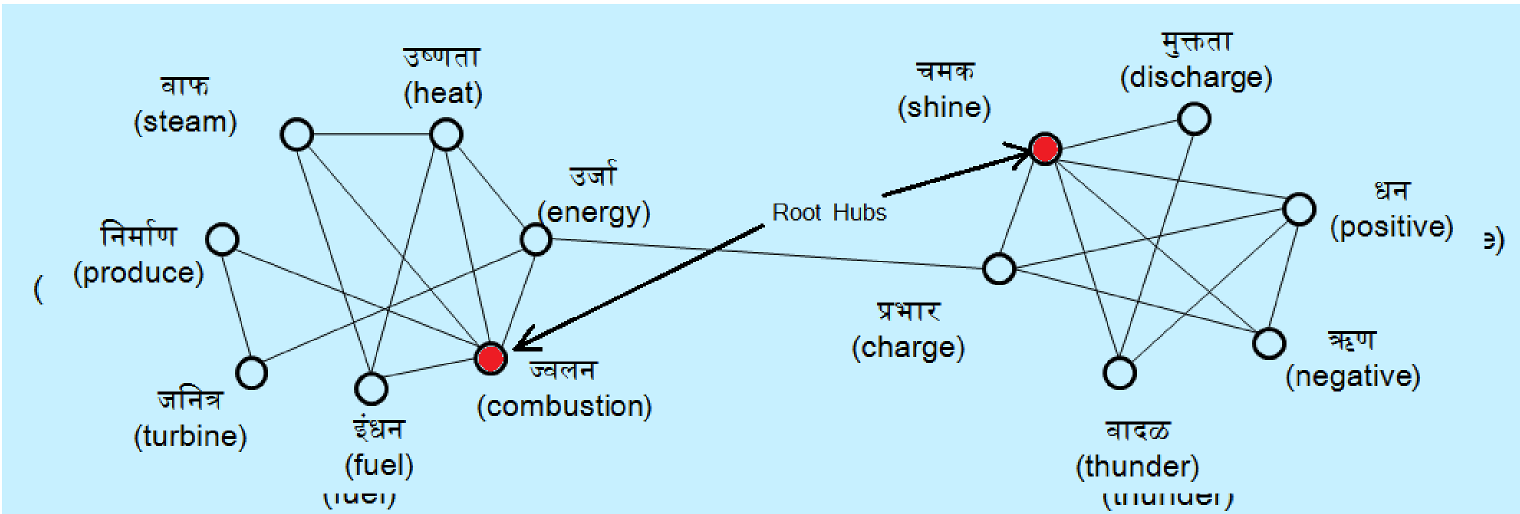
$$w_{AB} = 1 - \max\left(P(A|B), P(B|A)\right)$$

Example of co-occurrence graph



Co-occurrence graph for the word वीज (electricity/lightening)

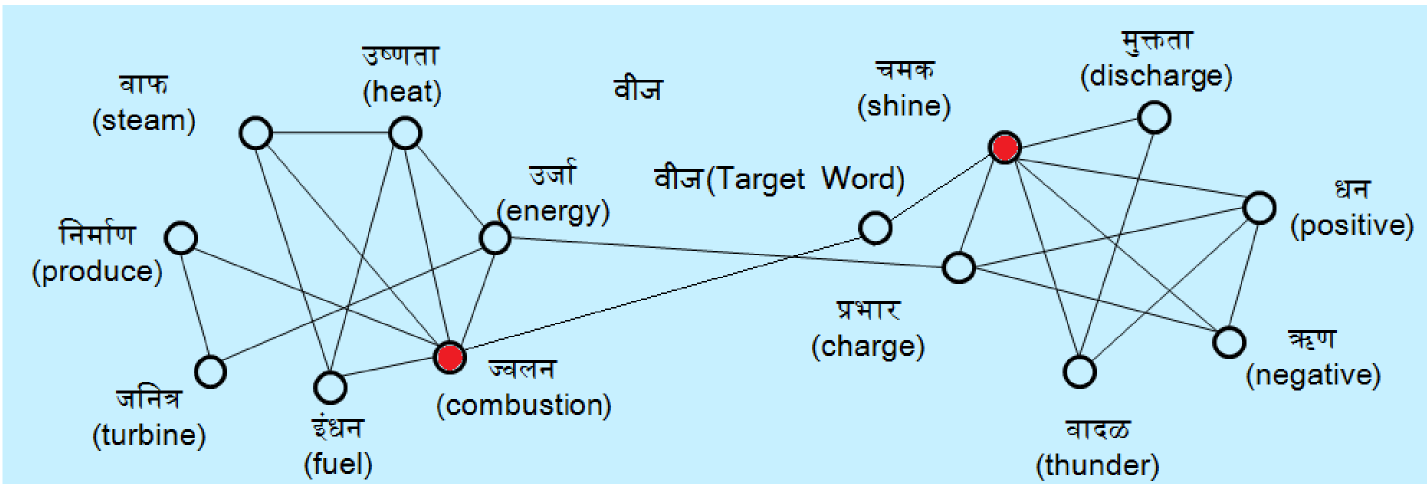
Root Hubs Detection



Co-occurrence graph for the word वीज (electricity/lightening)

- Root hubs are identified as the most connected nodes of each strongly connected component

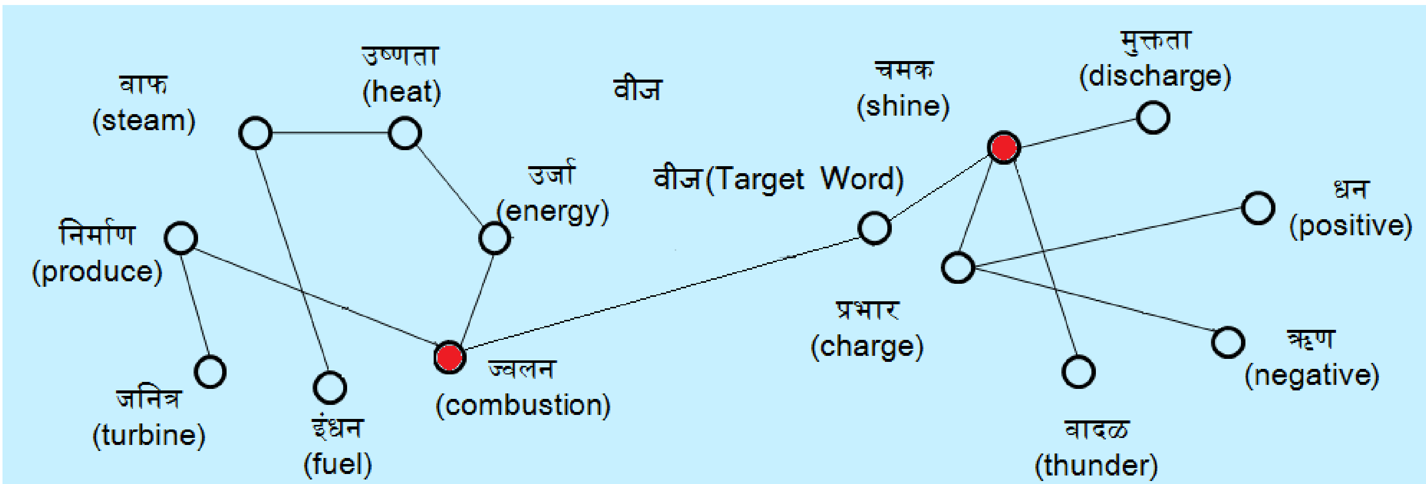
Target Word Added



Co-occurrence graph for the word वीज (electricity/lightening)

- Target word is added to the graph and connected to root hubs using edges of zero weight

Minimum Spanning Tree found



Co-occurrence graph for the word वीज (electricity/lightening)

Then score vector for each word is computed as follows:

$$S_i = \begin{cases} \dfrac{1}{1 + d(h_i, v)} & \text{if } v \in \text{component } i \\ 0 & \text{otherwise} \end{cases}$$

Where, *d(hi,v)* is the distance between the root hub *hi* and node *v*

Hyperlex contd..

- For the given occurrence of a target word, only words from its context take part in the scoring process
- The score vectors of all words are added for the given context
- The component with highest score becomes the winner sense
- Accuracy: 97% for 10 highly polysemous French words

Comparing WSD approaches

	Supervised	Semi-Supervised	Unsupervised	Knowledge based
Accuracy	high	moderate	low	low
Coverage	low	low	low	high
Need of tagged corpora	yes	Very few	no	no
Need of Knowledge resources	No	no	no	yes

Outline

- Introduction
  - Ambiguity
  - WSD Definition
  - Position of WSD in NLP layers
- Motivation
- WSD block diagram
- Lexical Resources needed
  - Sense Repository
  - Sense Annotated Corpus
- WSD approaches
  - Knowledge based
  - Corpus Based (Supervised, Unsupervised)
- Applications

WSD Applications

- Machine Translation
  - Translate "bill" from English to Spanish
  - Is it a "pico" or a "cuenta"?
  - Is it a bird jaw or an invoice?
- Information Retrieval
  - Find all Web Pages about "cricket"
  - The sport or the insect?
- Question Answering
  - What is George Miller's position on gun control?
  - The psychologist or US congressman?

WSD @ IIT Bombay

Unsupervised WSD approaches

- Approach 1:
  - Bilingual WSD using Expectation Maximization (EM) algorithm

  (Sudha Bhingardive, Samiulla Shaikh and Pushpak Bhattacharyya, Neighbor Help: Bilingual Unsupervised WSD Using Context, Association for Computational Linguistics (ACL) 2013, Sofia, Bulgaria, 4-9 August, 2013)

- Approach 2:
  - Most Frequent Sense Detection using Word vectors or embeddings

  (Sudha Bhingardive, Dhirendra Singh, Rudramurthy V, Hanumant Redkar and Pushpak Bhattacharyya, Unsupervised Multilinual Most Frequent Sense Detection using Word Embeddings, Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL) 2015, Denver, Colorado, USA, May 31 - June 5, 2015.)

Unsupervised WSD approaches

- Approach 1:
  - Bilingual WSD using Expectation Maximization (EM) algorithm

  (Sudha Bhingardive, Samiulla Shaikh and Pushpak Bhattacharyya, Neighbor Help: Bilingual Unsupervised WSD Using Context, Association for Computational Linguistics (ACL) 2013, Sofia, Bulgaria, 4-9 August, 2013)

- Approach 2:
  - Most Frequent Sense Detection using Word vectors or embeddings

  (Sudha Bhingardive, Dhirendra Singh, Rudramurthy V, Hanumant Redkar and Pushpak Bhattacharyya, Unsupervised Multilinual Most Frequent Sense Detection using Word Embeddings, Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL) 2015, Denver, Colorado, USA, May 31 - June 5, 2015.)

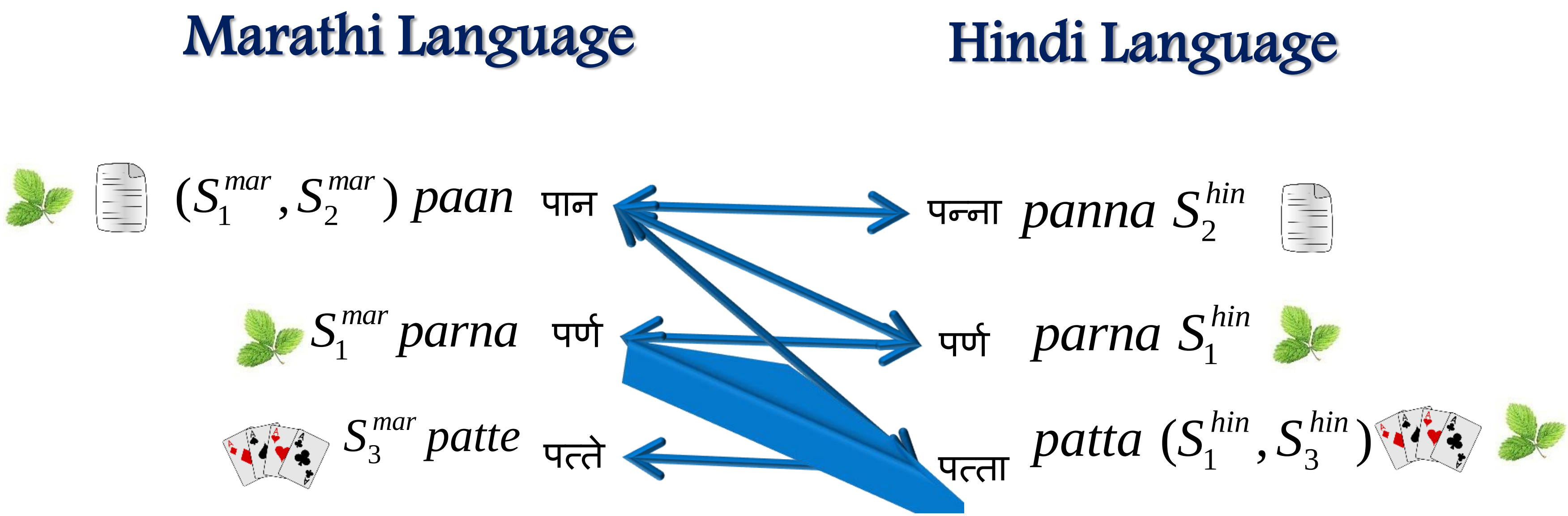
Problem Statement

- For a given untagged text of two languages perform word sense disambiguation using unsupervised technique

Overview of the approach

- Extension of Bilingual WSD (Khapra et al., 2011) by adding context
- Two resource scarce languages can help each other without the need of any sense tagged corpora in either languages.
- Approach uses untagged corpora and the aligned wordnets
- Approach relies on the key observation that sense distribution of any language remains same within a domain
- Context-based EM formulation is used for estimating the sense distribution
- An improvement of 17% - 35% in verb accuracy

Mode of Working



A bipartite graph of translation correspondences

Formulation

Marathi language

$$P(S_1^{mar} \mid paan) = \frac{\#(S_1^{mar}, paan)}{\#(S_1^{mar}, paan) + \#(S_2^{mar}, paan)}$$

Using Cross-links in Hindi:

$$P(S_1^{mar} \mid paan) = \frac{\#(S_1^{hin}, patta) + \#(S_1^{hin}, parna)}{\#(S_1^{hin}, patta) + \#(S_1^{hin}, parna) + \#(S_2^{hin}, panna)}$$

where,

$$\#(S_1^{hin}, patta) = P(S_1^{hin} \mid patta) * \#(patta)$$

Marathi language

$$P(S_1^{hin} \mid patta) = \frac{\#(S_1^{mar}, paan) + \#(S_1^{mar}, parna)}{\#(S_1^{mar}, paan) + \#(S_1^{mar}, parna) + \#(S_3^{mar}, patte)}$$

Formulation by Khapra et al., 2011

E- Step:

$$P(S^{L_1} | u) = \frac{\sum_{v} P(\pi_{L_2}(S^{L_1}) | v).\#(v)}{\sum_{S_i^{L_1}} \sum_{y} P(\pi_{L_2}(S_i^{L_1}) | y).\#(y)}$$

$s_i^{L_1} \in synsets_{L_1}(u)$

$v \in crosslinks_{L_2}(u, S^{L_1})$

$y \in crosslinks_{L_2}(u, S_i^{L_1})$

M- Step:

$$P(S^{L_2} | v) = \frac{\sum_{a} P(\pi_{L_1}(S^{L_2}) | u).\#(u)}{\sum_{S_i^{L_2}} \sum_{z} P(\pi_{L_1}(S_i^{L_2}) | z).\#(z)}$$

$s_i^{L_2} \in synsets_{L_2}(v)$

$u \in crosslinks_{L_1}(v, S^{L_2})$

$z \in crosslinks_{L_1}(v, S_i^{L_2})$

Two languages mutually help each other in estimating sense distribution

Adding Context

Basic formulation

$$P(S_1^{mar} \mid paan) = \frac{P(S_1^{hin} \mid patta) * \#(patta) + P(S_1^{hin} \mid parna) * \#(parna)}{P(S_1^{hin} \mid patta) * \#(patta) + P(S_1^{hin} \mid parna) * \#(parna) + P(S_3^{hin} \mid panna) * \#(panna)}$$

After adding the context

$$P(S_1^{mar} \mid paan, zaad) = \frac{\#(S_1^{hin} \mid patta, ped).\#(patta, ped) + \#(S_1^{hin} \mid parna, ped).\#(parna, ped)}{\#(S_1^{hin} \mid patta, ped).\#(patta, ped) + \#(S_1^{hin} \mid parna, ped).\#(parna, ped) + \#(S_3^{hin} \mid panna, ped).\#(panna, ped)}$$

Adding Semantic Relatedness

- Concurrence counts are unreliable
- They can make sense only if we have huge amount of corpora
- Semantic relatedness gives a good estimation of co-occurrence count.

New Formulation

After adding semantic relatedness

E-step:

$$P(S^{L_1}|u,a) = \frac{\sum_{v,b} P(\pi_{L_2}(S^{L_1})|v,b) \cdot \sigma(v,b)}{\sum_{S_i^{L_1}} \sum_{x,b} P(\pi_{L_2}(S_i^{L_1})|x,b) \cdot \sigma(x,b)}$$

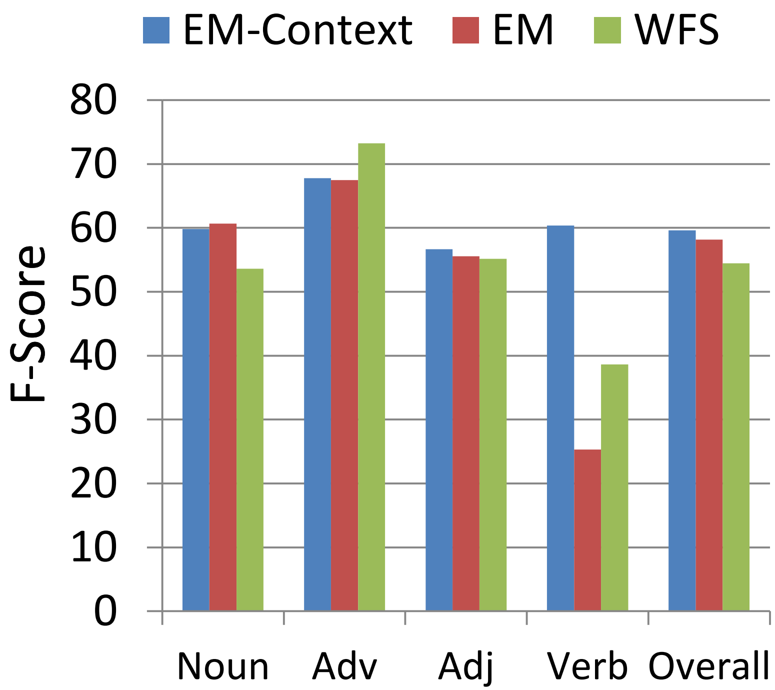
$$\begin{aligned} where, S_i^{L_1} &\in synsets_{L_1}(u) \\ a &\in context(u) \\ v &\in crosslinks_{L_2}(u, S^{L_1}) \\ b &\in crosslinks_{L_2}(a) \\ x &\in crosslinks_{L_2}(u, S_i^{L_1}) \end{aligned}$$

M-step:

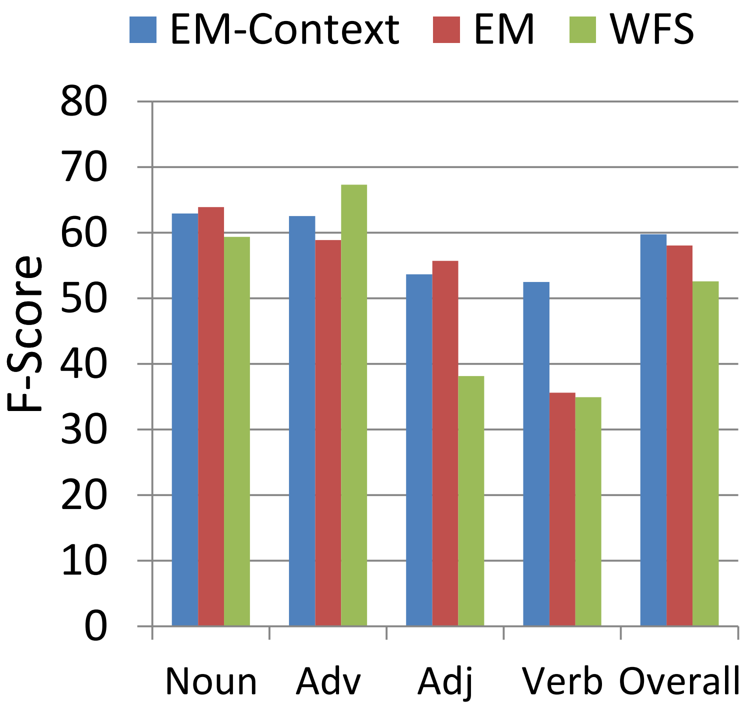
$$P(S^{L_2}|v,b) = \frac{\sum_{u,a} P(\pi_{L_1}(S^{L_2})|u,a) \cdot \sigma(u,a)}{\sum_{S_i^{L_2}} \sum_{y,b} P(\pi_{L_1}(S_i^{L_2})|y,a) \cdot \sigma(y,a)}$$

$$\begin{aligned} where, S_i^{L_2} &\in synsets_{L_2}(v) \\ b &\in context(v) \\ u &\in crosslinks_{L_1}(v, S^{L_2}) \\ a &\in crosslinks_{L_1}(b) \\ y &\in crosslinks_{L_1}(v, S_i^{L_2}) \end{aligned}$$

Results on Health domain

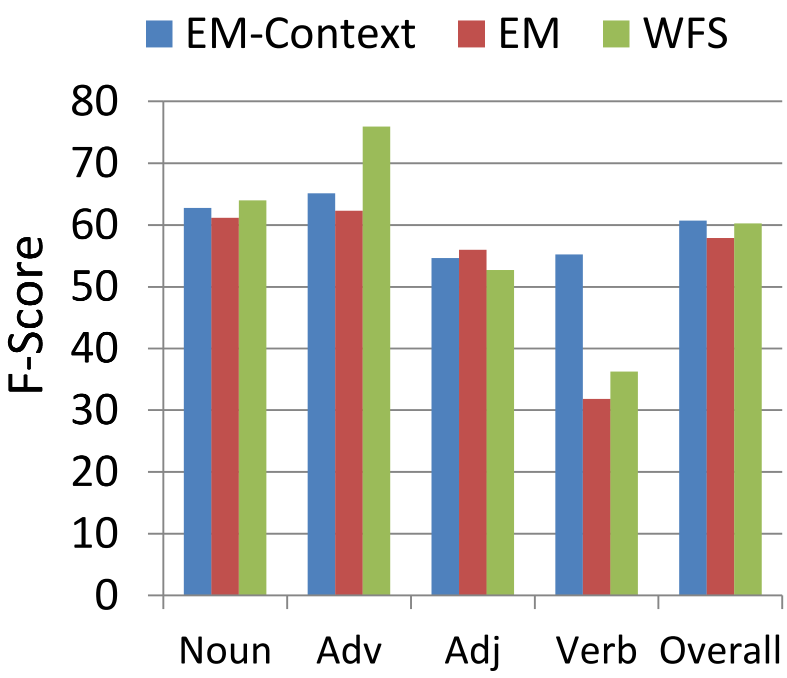


Hindi Health Domain

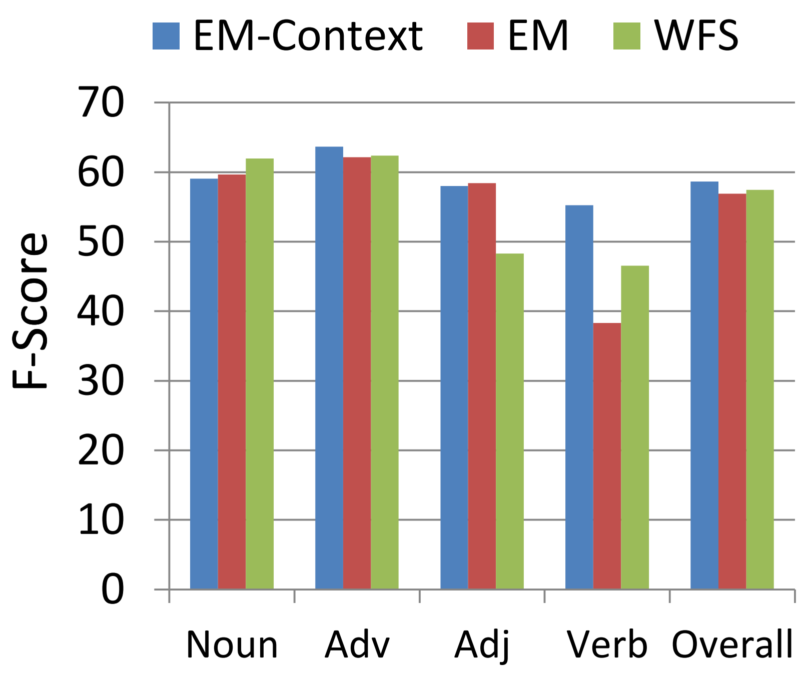


Marathi Health Domain

Results on Tourism domain



Hindi Tourism Domain



Marathi Tourism Domain

Error Analysis contd..

वे पत्ते खेल रहे हैं

(vaha patte khel rahe the)

(They are playing cards)

वे पेड़ के नीचे पत्ते खेल रहे हैं

(vaha ped ke niche patte khel rahe hai)

(They are playing cards below the tree)

Semantic structure of the sentence can help in such situations

Error Analysis

मैं बैग का फोटो निकाल रही हूँ ।
(mein bag kaa photo nikaal rahii hun)
(I am clicking the photo of bag)

मैं बैग से फोटो निकाल रही हूँ ।
(mein bag se photo nikaal rahii hun)
(I am taking the photo outside the bag)

Function words help in disambiguation, since they define semantic relations between two content words.

Error Analysis contd..

- We have considered single word crosslinks in our approach.
- Sometimes one word has multi-word crosslinks in another language.

अब → **आता, या वेळी, या वेळेस, ह्या वेळी , ह्या वेळेस**

(ab) (aata, ya veli, ya veles, hya veli, hya veles)

(Hindi) **(Marathi)**

Language properties also play an important role

Error Analysis contd..

- Resource related problems:
  - too fine grained HWN senses

ऊपर, **अधिक**, **ज्यादा**, **ज़्यादा**, **और** - अधिक या ज्यादा *"यह चीनी दस किलो से ऊपर है / भाजीवाले ने एक किलो सब्जी तौलने के बाद ऊपर से डाला"*

बहुत, **ख़ूब**, **खूब**, **भरपूर**, **बड़ा**, **ज्यादा**, **ज़्यादा**, **अधिक**, **काफ़ी**, **काफी**, **जमकर**, **डटकर**,**कड़ा** - अधिक मात्रा में *"आज वह बहुत हँसा"*

We should consider coarse-grained senses to increase accuracy

Bilingual WSD using Word Embeddings

- Word embeddings are used an approximation to the co-occurrence counts
- Verb accuracy improved by 8.5% for Marathi.
- Adjective accuracy improved by 7% for Hindi and 2.5% for Marathi.

WSD Algorithm	HIN-HEALTH					MAR-HEALTH				
	NOUN	ADV	ADJ	VERB	Overall	NOUN	ADV	ADJ	VERB	Overall
Combined	59.32	68.98	63.18	60.02	**60.94**	62.75	61.19	**56.22**	**60.99**	**61.30**
EM-C-DistSimi	59.59	69.20	**63.87**	55.73	61.09	63.09	61.82	55.60	43.69	58.92
EM-C-WnSimi	59.82	67.80	56.66	**60.38**	59.63	62.90	62.54	53.63	52.49	59.77
EM	**60.68**	67.48	55.54	25.29	58.16	**63.88**	58.88	55.71	35.60	58.03
WFS	53.49	**73.24**	55.16	38.64	54.46	59.35	**67.32**	38.12	34.91	52.57
RB	32.52	45.08	35.42	17.93	33.31	33.83	38.76	37.68	18.49	32.45

Unsupervised WSD approaches

- Approach 1:
  - Bilingual WSD using Expectation Maximization (EM) algorithm

  (Sudha Bhingardive, Samiulla Shaikh and Pushpak Bhattacharyya, Neighbor Help: Bilingual Unsupervised WSD Using Context, Association for Computational Linguistics (ACL) 2013, Sofia, Bulgaria, 4-9 August, 2013)

- Approach 2:
  - Most Frequent Sense Detection using Word vectors or embeddings

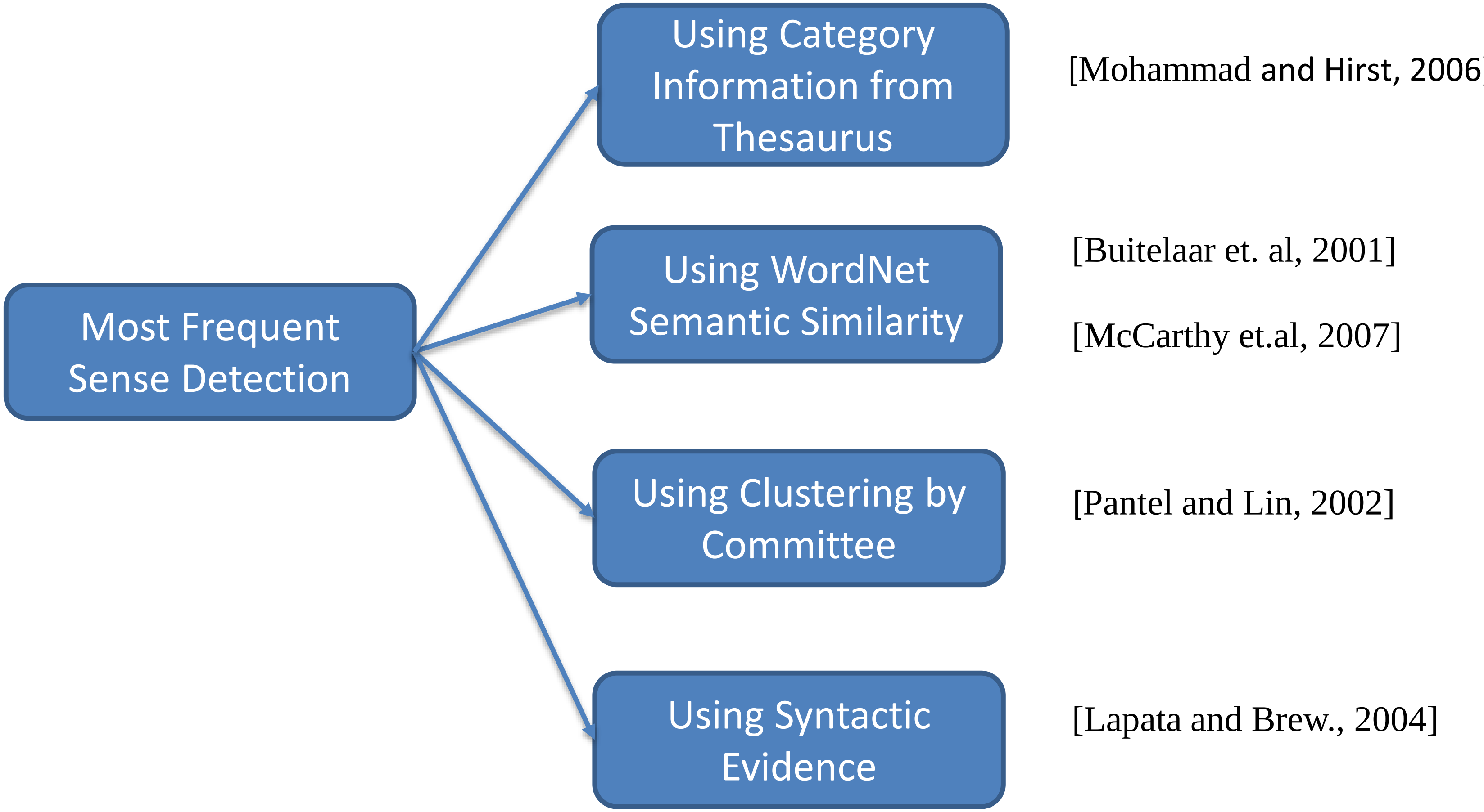
  (Sudha Bhingardive, Dhirendra Singh, Rudramurthy V, Hanumant Redkar and Pushpak Bhattacharyya, Unsupervised Multilinual Most Frequent Sense Detection using Word Embeddings, Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL) 2015, Denver, Colorado, USA, May 31 - June 5, 2015.)

Most Frequent Sense Detection

- **Problem Statement:**
  - For a given word, find the most frequent sense of a word using unsupervised technique

- **Motivation**:
  - The first sense heuristic is often used as a baseline for WSD systems
  - For WSD systems, it is hard to beat this baseline (5 out of 26 supervised approaches beat this baseline)
  - Manually tagging data is costly in terms of time and money
  - It would be useful to have a method of ranking senses directly from untagged data

Related Work



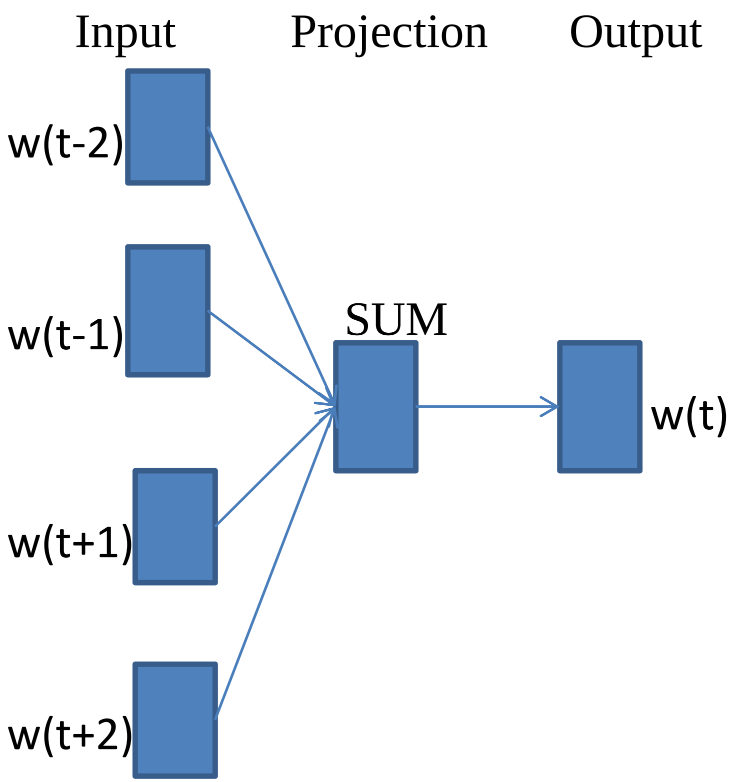
Our Approach [UMFS-WE]

- A unsupervised approach for MFS detection using word embeddings
- Word embedding of a word is compared with sense embeddings and the sense with highest similarity is considered as the most frequent sense
- Extendable and portable: Domain independent approach and easily portable to multiple languages

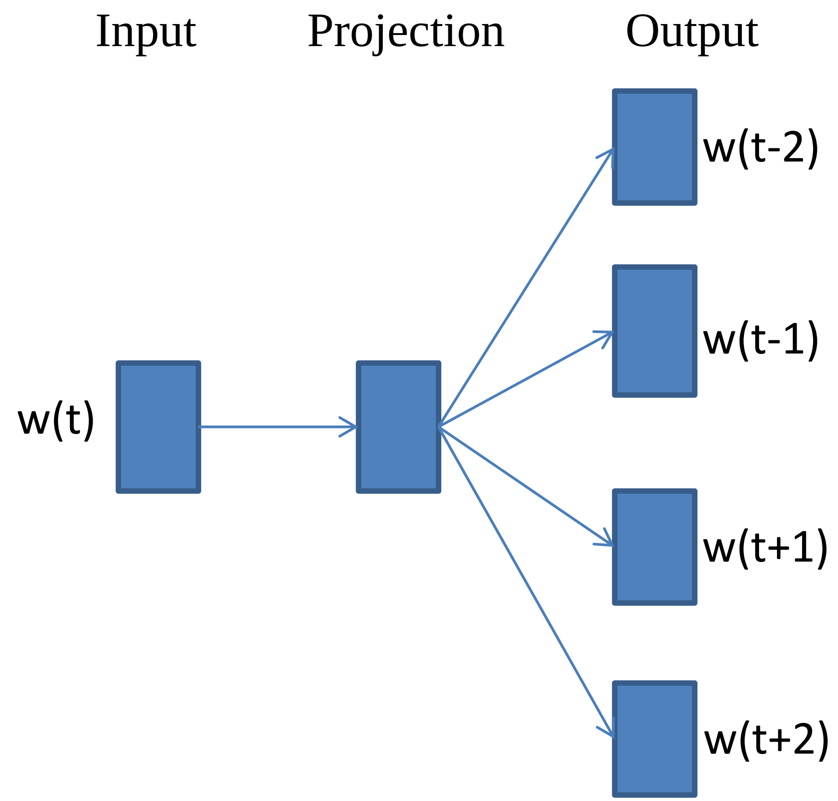
Word Embeddings

- Represent each word with low-dimensional real valued vector.
- Increasingly being used in variety of Natural Language Processing tasks
- **word2vec tool** (Mikolov et. al, 2013)
  - One of the most popular word embedding tool
  - Source code provided

Word Embeddings contd..



Continuous bag of words model (CBOW)



Skip-gram model

Word Embeddings contd..

- **word2vec tool** (Mikolov et. al, 2013)
  - It captures many linguistic regularities

Vector('king') – Vector('man') + Vector('woman') => Vector('queen')

Word Embeddings contd..

- Distributionally Similar words of फल (fala, fruit)

words	cosine similarity
फ़ल	0.840545
केला	0.705185
ल	0.688565
सीताफल	0.685993
पपीता	0.682171
सौन्दर्यवर्धक	0.677420
कन्दमूल	0.672466
अननास	0.655930
भाजियाँ	0.650811
आड़ू	0.650100

Sense Embeddings

- The **sense-bag** for the sense S_i is created as below,

$$\mathrm{SB}(S_i) = \{\mathrm{x} | \mathrm{x} \text{ - } \mathrm{Features}(S_i)\}$$

  – Features(S_i) - WordNet based features for sense S_i

- Sense embeddings are obtained by taking the average of word embeddings of each word in the sense-bag

$$vec(S_i) = \frac{\sum_{x \in SB(S_i)} vec(x)}{N}$$

  – S_i - ith sense of a word W
  – N - Number of words present in the sense-bag $SB(S_i)$

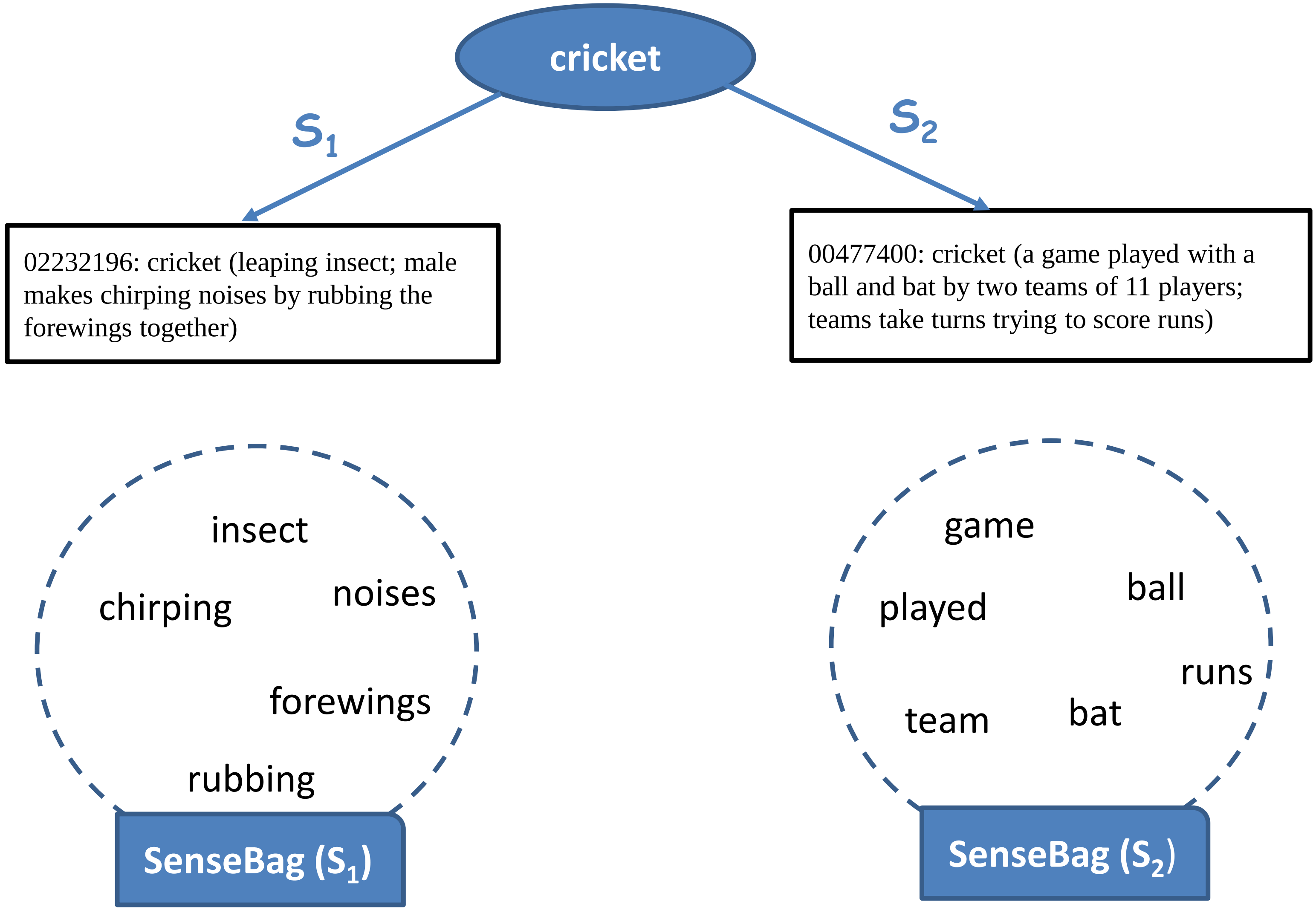
MFS Detection

- We treat the MFS identification problem as finding the closest cluster centroid (*i.e.*, sense embedding)
- Cosine similarity is used.
- Most frequent sense is obtained

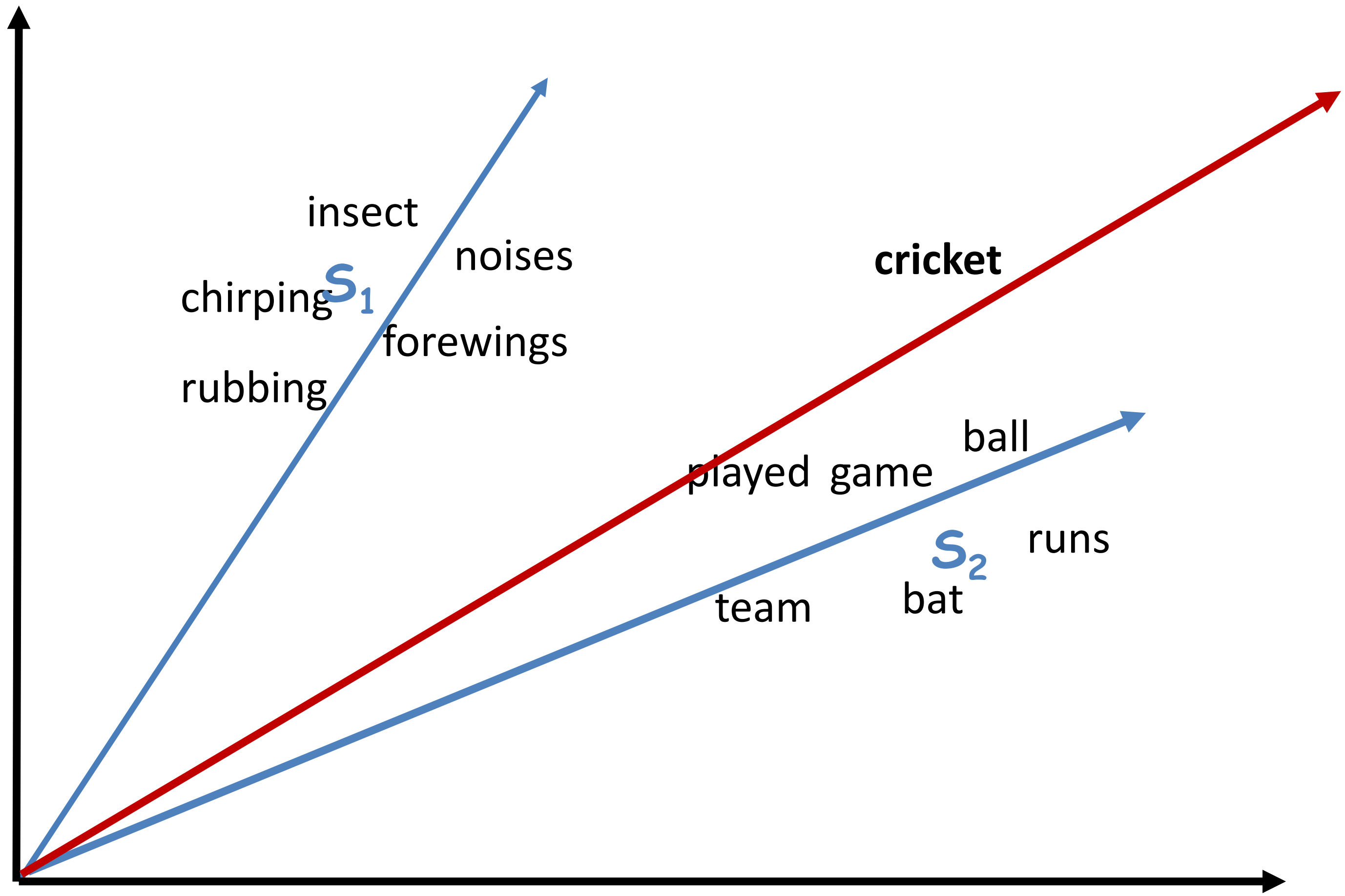
$$MFS_w = \underset{S_i}{\operatorname{argmax}} \cos(vec(W), vec(S_i))$$

  - $vec(W)$ - word embedding of a word W
  - S_i- i^{th} sense of word W
  - $vec(S_i)$ - sense embedding for S_i

MFS Detection



MFS Detection contd..



Experiments

A. Experiments on WSD
   1. Experiments on WSD using Skip-Gram model
      - Hindi (Newspaper)
      - English (SENSEVAL-2 and SENSEVAL-3)
   2. Experiments on WSD using different word vector models
   3. Comparing WSD results using different sense vector models
      - Retrofitting Sense Vector Model (English)
   4. Experiments on WSD for words which do not exists in SemCor

B. Experiments on selected words (34 polysemous words from SENSEVAL-2 corpus)
   1. Experiments using different word vector models
   2. Comparing results with various sizes of vector dimensions

Experiments

A. Experiments on WSD
   1. Experiments on WSD using Skip-Gram model
      - Hindi (Newspaper)
      - English (SENSEVAL-2 and SENSEVAL-3)

[A.1] Experiments on WSD using skip-gram model

- Training of word embeddings:
  - Hindi: Bojar (2014) corpus (44 M sentences)
  - English: Pre-trained Google-News word embeddings
- Datasets used for WSD:
  - Hindi: Newspaper dataset
  - English: SENSEVAL-2 and SENSEVAL-3
- Experiments are restricted to only polysemous nouns.

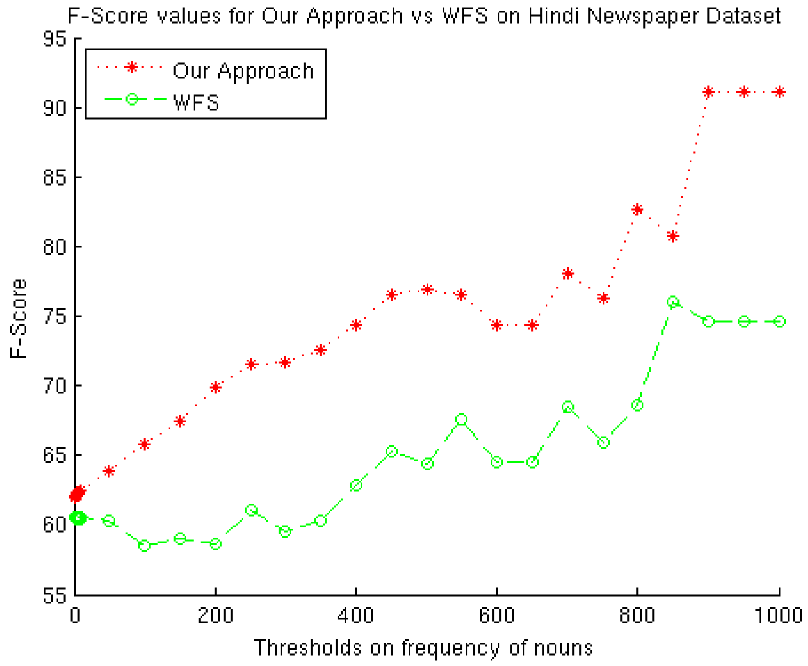
[A.1] Results on WSD

HINDI WSD	Newspaper dataset		
	Precision	Recall	F-Score
UMFS-WE	62.43	61.58	62.00
WFS	61.73	59.31	60.49

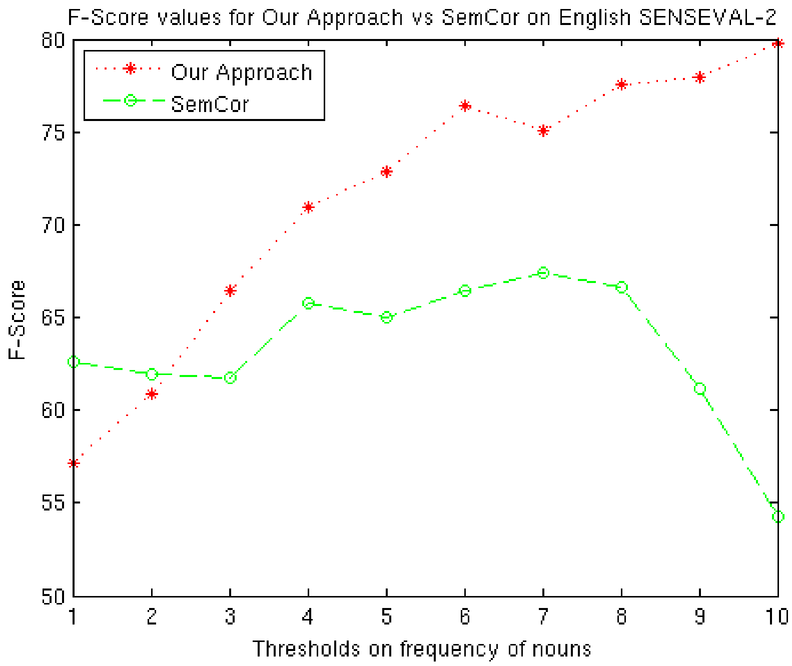
ENGLISH WSD	SENSEVAL-2 dataset			SENSEVAL-3 dataset		
	Precision	Recall	F-Score	Precision	Recall	F-Score
UMFS-WE	52.39	52.27	52.34	43.34	43.22	43.28
WFS	61.72	58.16	59.88	66.57	64.89	65.72

[A.1] Results on WSD contd..

- F-Score is also calculated for increasing thresholds on the frequency of nouns appearing in the corpus.



Hindi WSD



English WSD

[A.1] Results on WSD contd..

- WordNet feature selection for sense embeddings creation

Sense Vectors Using WordNet features	Precision	Recall	F-measure
SB	51.73	38.13	43.89
SB+GB	53.31	52.39	52.85
SB+GB+EB	56.61	55.84	56.22
SB+GB+EB+PSB	59.53	58.72	59.12
SB+GB+EB+PGB	60.57	59.75	60.16
SB+GB+EB+PEB	60.12	59.3	59.71
SB+GB+EB+PSB+PGB	57.59	56.81	57.19
SB+GB+EB+PSB+PEB	58.93	58.13	58.52
SB+GB+EB+PGB+PEB	**62.43**	**61.58**	**62**
SB+GB+EB+PSB+PGB+PEB	58.56	57.76	58.16

SB: Synset Bag
GB: Gloss Bag
EB: Example Bag
PSB: Parent Synset Bag
PGB: Parent Gloss Bag
PEB: Parent Example Bag

Table: Hindi WSD results using various WordNet features for Sense Embedding creation

Experiments

A. Experiments on WSD
   1. Experiments on WSD using Skip-Gram model
      - Hindi (Newspaper)
      - English (SENSEVAL-2 and SENSEVAL-3)
   2. Experiments on WSD using different word vector models

[A.2] Experiments on WSD using various Word Vector models

- We compared MFS results on various word vector models as listed below:

Word Vector Model	Dimensions
SkipGram-Google-News (Mikolov et. al, 2013)	300
Senna (Collobert et. al, 2011)	50
MetaOptimize (Turian et. al, 2010)	50
RNN (Mikolov et. al, 2011)	640
Glove (Pennington et. al, 2014)	300
Global Context (Huang et. al, 2013)	50
Multilingual (Faruqui et.al, 2014)	512
SkipGram-BNC (Mikolov et. al, 2013)	300
SkipGram-Brown (Mikolov et. al, 2013)	300

Table: Word Vector Models

[A.2] Experiments on WSD using various Word Vector models contd..

WordVector	Noun	Adj	Adv	Verb
SkipGram-Google-News	54.49	**50.56**	**47.66**	20.66
Senna	54.49	40.44	28.97	21.9
RNN	39.07	28.65	40.18	19.42
MetaOptimize	33.73	36.51	32.71	19.83
Glove	**54.69**	49.43	39.25	18.18
Global Context	48.3	32.02	31.77	20.66
SkipGram-BNC	53.03	48.87	39.25	**23.14**
SkipGram-Brown	30.29	48.87	27.10	13.29

Table: English WSD results for words with corpus frequency > 2

Experiments

A. Experiments on WSD
   1. Experiments on WSD using Skip-Gram model
      - Hindi (Newspaper)
      - English (SENSEVAL-2 and SENSEVAL-3)
   2. Experiments on WSD using different word vector models
   3. Comparing WSD results using different sense vector models
      - Retrofitting Sense Vector Model (Jauhar et al, 2015)

[A.3] Results on WSD

WordVector	SenseVector	Noun	Adj	Adv	Verb
SkipGram-Google-News	Our model	**58.87**	53.53	**46.34**	20.49
	Retrofitting	47.84	**57.57**	32.92	**21.73**
Senna	Our model	**61.29**	43.43	**21.95**	**24.22**
	Retrofitting	6.9	**68.68**	**21.95**	1.86
RNN	Our model	**42.2**	26.26	**40.24**	**21.11**
	Retrofitting	10.48	**62.62**	21.95	1.24
MetaOptimize	Our model	**37.9**	50.5	**31.7**	**18.01**
	Retrofitting	10.48	**62.62**	21.95	1.24
Glove	Our model	**58.33**	53.33	**39.02**	**17.39**
	Retrofitting	9.94	**62.62**	21.95	1.24
Global Context	Our model	53.22	37.37	**24.39**	19.25
	Retrofitting	12.36	**68.68**	21.95	1.24
SkipGram-Brown	Our model	29.31	60.6	**23.17**	11.42
	Retrofitting	11.49	**68.68**	21.95	1.26

Table: English WSD results for words with corpus frequency > 2

Experiments

A. Experiments on WSD
   1. Experiments on WSD using Skip-Gram model
      - Hindi (Newspaper)
      - English (SENSEVAL-2 and SENSEVAL-3)
   2. Experiments on WSD using different word vector models
   3. Comparing WSD results using different sense vector models
      - Retrofitting Sense Vector Model (English)
   4. Experiments on WSD for words which do not exists in SemCor

[A.4] English WSD results for SENSEVAL-2 words which do not exist in SemCor

Word Vector	F-score
SkipGram-Google-News	**84.12**
Senna	79.67
RNN	24.59
MetaOptimize	22.76
Glove	79.03
Global Context	28.09
Multilingual	35.48
SkipGram-BNC	68.29
SkipGram-BNC-Brown	74.79

proliferate, agreeable, bell_ringer, audacious, disco, delete, prestigious, option, peal, impaired, ringer, flatulent, unwashed, cervix, discordant, eloquently, carillon, full-blown, incompetence, stick_on, illiteracy, implicate, galvanize, retard, libel, obsession, altar, polyp, unintelligible, governance, bell_ringing.

Experiments

A. Experiments on WSD
   1. Experiments on WSD using Skip-Gram model
      - Hindi (Newspaper)
      - English (SENSEVAL-2 and SENSEVAL-3)
   2. Experiments on WSD using different word vector models
   3. Comparing WSD results using different sense vector models
      - Retrofitting Sense Vector Model (English)
   4. Experiments on WSD for words which do not exists in SemCor

B. Experiments on selected words (34 polysemous words from SENSEVAL-2 corpus)
   1. Experiments using different word vector models

[B.1] Experiments on selected words

- 34 polysemous nouns, where each one has atleast two senses and which have occurred at least twice in the SENSEVAL-2 dataset are chosen

Token	Senses	Token	Senses
church	4	individual	2
field	13	child	4
bell	10	risk	4
rope	2	eye	5
band	12	research	2
ringer	4	team	2
tower	3	version	6
group	3	copy	3
year	4	loss	8
vicar	3	colon	5
sort	4	leader	2
country	5	discovery	4
woman	4	education	6
cancer	5	performance	5
cell	7	school	7
type	6	pupil	3
growth	6	student	2

[B.1] MFS Results on selected words

Word Vectors	Accuracy
SkipGram-BNC	63.63
SkipGram-Brown	48.38
SkipGram-Google-News	60.6
Senna	57.57
Glove	**66.66**
Global Context	51.51
Metaoptimize	27.27
RNN	51.51
Multilingual	63.4

Table: English WSD results for selected words from SENSEVAL-2 dataset

Experiments

A. Experiments on WSD
   1. Experiments on WSD using Skip-Gram model
      - Hindi (Newspaper)
      - English (SENSEVAL-2 and SENSEVAL-3)
   2. Experiments on WSD using different word vector models
   3. Comparing WSD results using different sense vector models
      - Retrofitting Sense Vector Model (English)
   4. Experiments on WSD for words which do not exists in SemCor

B. Experiments on selected words (34 polysemous words from SENSEVAL-2 corpus)
   1. Experiments using different word vector models
   2. Comparing results with various sizes of vector dimensions

[B.2] Comparing MFS results with various sizes of vector dimensions

Word Vectors	Accuracy
SkipGram-BNC-1500	60.61
SkipGram-BNC-1000	60.61
SkipGram-BNC-500	66.67
SkipGram-BNC-400	**69.69**
SkipGram-BNC-300	63.64
SkipGram-BNC-200	60.61
SkipGram-BNC-100	48.49
SkipGram-BNC-50	51.52

MFS for Indian Languages

- *Polyglot*[1] word embeddings are used for obtaining MFS.
  - word embeddings are trained using Wikipedia data.

- Currently, system is working for *Marathi*, *Bengali*,*Gujarati*, *Sanskrit*, *Assamese*, *Bodo*, *Oriya*, *Kannada*, *Tamil*, *Telugu*, *Malayalam* and *Punjabi*.

- Due to lack of gold data, we could not evaluate results

- APIs are developed for finding the MFS for a word

[1]https://sites.google.com/site/rmyeid/projects/polyglot

MFS for using BabelNet

- MFS is calculated by using BabelNet as a sense repository.

- BabelNet covers 271 languages and is obtained from the automatic integration of: WordNet, Open Multilingual WordNet, Wikipedia, Omega Wiki, Wiktionary, Wikidata.

- System is working for *English*, *Russian*, *Italian*, *French*, *German*, and *Spanish*.

- Due to lack of gold data, we couldn't evaluate results for these language.

Conclusion

- WSD helps in solving ambiguity
- Bilingual WSD approach showed how two resource deprived languages help each other in WSD
- Unsupervised MFS approach showed that how word embeddings captures the MFS of a word
- Both the approaches are language independent
- They can be used in NLP applications

Publications Contributing to Thesis

- Sudha Bhingardive, Dhirendra Singh and Pushpak Bhattacharyya, Automatic Synset Ranking of Indian Language Wordnets using Word Embedding, *38th International Conference of the Linguistic Society of India* (**LSI**), Indian Institute of Technology Guwahati, Assam, 10-12 November, 2016.

- Sudha Bhingardive, Rajita Shukla, Jaya Saraswati, Laxmi Kashyap, Dhirendra Singh and Pushpak Bhattacharyya, **Synset Ranking of Hindi WordNet**, *Language Resources and Evaluation Conference* (**LREC**), Portorož, Slovenia, 23-28 May, 2016.

- Sudha Bhingardive, Hanumant Redkar, Prateek Sappadla, Dhirendra Singh and Pushpak Bhattacharyya**, IndoWordNet::Similarity - Computing Semantic Similarity and Relatedness using IndoWordNet**, *Global WordNet Conference* (GWC), Bucharest, Romania, 27-30 January, 2016.

- Harpreet Arora, Sudha Bhingardive and Pushpak Bhattacharyya, **Detecting Most Frequent Sense using Word Embeddings and BabelNet**, *Global WordNet Conference* (GWC), Bucharest, Romania, 27-30 January, 2016.

- Sudha Bhingardive, Dhirendra Singh, Rudramurthy V, Hanumant Redkar and Pushpak Bhattacharyya, **Unsupervised Most Frequent Sense Detection using Word Embeddings**, *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (**NAACL**), Denver, Colorado, USA, May 31 - June 5, 2015.

Publications Contributing to Thesis contd..

- Sudha Bhingardive, Dhirendra Singh and Pushpak Bhattacharyya, **Using Word Embeddings for Bilingual Unsupervised WSD**, *International Conference on Natural Language Processing* (**ICON**) 2015, Trivandrum, Kerala, India, 11-14 December, 2015.

- Devendra Singh Chaplot, Sudha Bhingardive and Pushpak Bhattacharyya, **IndoWordnet Visualizer: A Graphical User Interface for Browsing and Exploring Wordnets of Indian Languages**, *Global WordNet Conference* (**GWC**), Tartu, Estonia, 25-29 January, 2014

- Sudha Bhingardive, Ratish Puduppully, Dhirendra Singh and Pushpak Bhattacharyya, **Merging Verb Senses of Hindi WordNet using Word Embeddings**, *International Conference on Natural Language Processing* (**ICON**), Goa, India, 18-21 December, 2014.

- Sudha Bhingardive, Samiulla Shaikh and Pushpak Bhattacharyya, **Neighbor Help: Bilingual Unsupervised WSD Using Context**, *Association for Computational Linguistics* (**ACL**), Sofia, Bulgaria, 4-9 August, 2013.

- Book Chapter:

Word Sense Disambiguation Using IndoWordNet

Sudha Bhingardive and Pushpak Bhattacharyya, Book Title: "The WordNet in Indian Languages", Springer Science+Business Media, Singapore, 2016, N.S. Dash et al. (eds.), ISBN : 978-981-10-1907-4.

Other Publications

- Raksha Sharma, Sudha Bhingardive and Pushpak Bhattacharyya, **Meaning Matters: Senses of Words are More Informative than Words for Cross-domain Sentiment Analysis**, *International Conference on Natural Language Processing* (**ICON**), Indian Institute of Technology (Banaras Hindu University), Varanasi, India, 17-20 December, 2016.

- Dhirendra Singh, Sudha Bhingardive, Pushpak Bhattacharyya, **Multiword Expressions Dataset for Indian Languages**, *Language Resources and Evaluation Conference* (**LREC**), Portorož, Slovenia, 23-28 May, 2016.

- Hanumant Redkar, Sudha Bhingardive, Kevin Patel, Pushpak Bhattacharyya, **WWDS APIs: Application Program Manipulation of World WordNet Database Structure**, The *Association for the Advancement of Artificial Intelligence* (**AAAI**), Phoenix, USA, 12-17 February, 2016.

- Dhirendra Singh, Sudha Bhingardive and Pushpak Bhattacharyya, **Detection of Compound Nouns and Light Verb Constructions using IndoWordNet**, *Global WordNet Conference* (**GWC**), Bucharest, Romania, 27-30 January, 2016.

Other Publications contd..

- Hanumant Redkar, Sudha Bhingardive, Diptesh Kanojia, and Pushpak Bhattacharyya, **World WordNet Database Structure: An Efficient Schema for Storing Information of Wordnets of the World**, *The Association for the Advancement of Artificial Intelligence* (**AAAI**) 2015, Austin, Texas, 25-30 January 2015.

- Dhirendra Singh, Sudha Bhingardive, Kevin Patel and Pushpak Bhattacharyya, **Detecting Multiword Expression using Word Embeddings and WordNet**, *37th International Conference of the Linguistic Society of India* (**LSI-37**), Javaharlal Nehru University, Delhi, 15-17 October, 2015.

- Dhirendra Singh, Sudha Bhingardive, Kevin Patel and Pushpak Bhattacharyya**, Multiword Expressions detection for Hindi Language using Word Embeddings and WordNet-based Features**, *International Conference on Natural Language Processing* (**ICON**), Trivandrum, Kerala, India, 11-14 December, 2015.

- Sudha Bhingardive, Tanuja Ajotikar, Irawati Kulkarni, Malhar Kulkarni and Pushpak Bhattacharyya, **Semi-Automatic Extension of Sanskrit Wordnet using Bilingual Dictionary**, *Global WordNet Conference* (**GWC**), Tartu, Estonia, 25-29 January, 2014.

References

- Pushpak Bhattacharyya, IndoWordNet, Lexical Resources Engineering Conference (LREC), Malta, May, 2010.
- Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande and P. Bhattacharyya, An Experience in Building the Indo WordNet - a WordNet for Hindi, First International Conference on Global WordNet (GWC), Mysore, India, January 2002.
- R. Navigli and S. Ponzetto., BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artificial Intelligence, 193, Elsevier, 2012, pp. 217-250
- Paul Buitelaar and Bogdan Sacaleanu. 2001. “Ranking and selecting synsets by domain relevance”, NAACL 2001 Workshop.
- Xinxiong Chen, Zhiyuan Liu and Maosong Sun. 2014. “A Unified Model for Word Sense Representation and Disambiguation”, Proceedings of ACL 2014.

References

- Z. Harris. 1954. “Distributional structure”, Word 10(23):146-162.
- Tomas Mikolov, Chen Kai, Corrado Greg and Dean Jeffrey. 2013. “Efficient Estimation of Word Representations in Vector Space”, In Proceedings of Workshop at ICLR, 2013.
- Diana McCarthy, Rob Koeling, Julie Weeds and John Carroll. 2007. “Unsupervised Acquisition of Predominant Word Senses”, Computational Linguistics, 33 (4) pp 553-590.
- Ondrej Bojar, Diatka Vojtech, Rychly Pavel, Stranak Pavel, Suchomel Vıt, Tamchyna Ales and Zeman Daniel. 2014. “HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation”, LREC 2014.

References

- Mitesh Khapra, Salil Joshi and Pushpak Bhattacharyya, Help Me and I will Help You: A Bilingual Unsupervised Approach for Estimating Sense Distributions using Expectation Maximization, 5th International Conference on Natural Language Processing (IJCNLP 2011), Chiang Mai, Thailand, November 2011.

- Diab, M. and Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL ’02, Pages 255–262, Morristown, NJ, USA. Association for Computational Linguistics.

- Unsupervised word sense disambiguation using bilingual comparable corpora. In Proceedings of the 19th international conference on Computational linguistics - Volume 1,COLING ’02, pages 1– 7, Stroudsburg, PA, USA. Association for Computational Linguistics.

References

- Mihalcea, R., Tarau, P., and Figa, E. (2004). Pagerank on semantic networks, with application to word sense disambiguation. In Proceedings of Coling 2004, pages 1126–1132, Geneva, Switzerland. COLING.
- Pedersen, T. and Bruce, R. F. (1997). Distinguishing word senses in untagged text. CoRR, cmp-lg/9706008.
- Pedersen, T., Purandare, A., and Kulkarni, A. (2005). Name discrimination by clustering similar contexts. In Gelbukh, A. F., editor, CICLing, volume 3406 of Lecture Notes in Computer Science, pages 226–237. Springer.
- Veronis, J. (2004). Hyperlex: Lexical cartography for information retrieval. Comput. Speech Lang., 18(3).
- Navigli, R. and Lapata, M. (2007). Graph connectivity measures for unsupervised word sense disambiguation. In Veloso, M. M., editor, IJCAI, pages 1683–1688.

References

- Gale, W., Church, K., and Yarowsky, D. (1992). Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics.
- Ng, H. T., Wang, B., and Chan, Y. S. (2003). Exploiting parallel texts for word sense disambiguation: an empirical study. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL ’03, pages 455–462, Morristown, NJ, USA. Association for Computational Linguistics.
- Agirre, Eneko, and German Rigau. "Word sense disambiguation using conceptual density." Proceedings of the 16th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1996.
- Kulkarni, M., Dangarikar, C., Kulkarn, I., Nanda, A., and Bhattacharyya, P. “Introducing sanskrit wordnet”. In Conference on Global Wordnet(GWC), 2010.
- Navigli, R, “A quick tour of babelnet1.1”, In CICLing, 2013.

Thank You !!!

Contact Details

Sudha Bhingardive,

Dept. of CSE,

IIT Bombay, Mumbai

sudha@cse.iitb.ac.in

bhingardivesudha@gmail.com