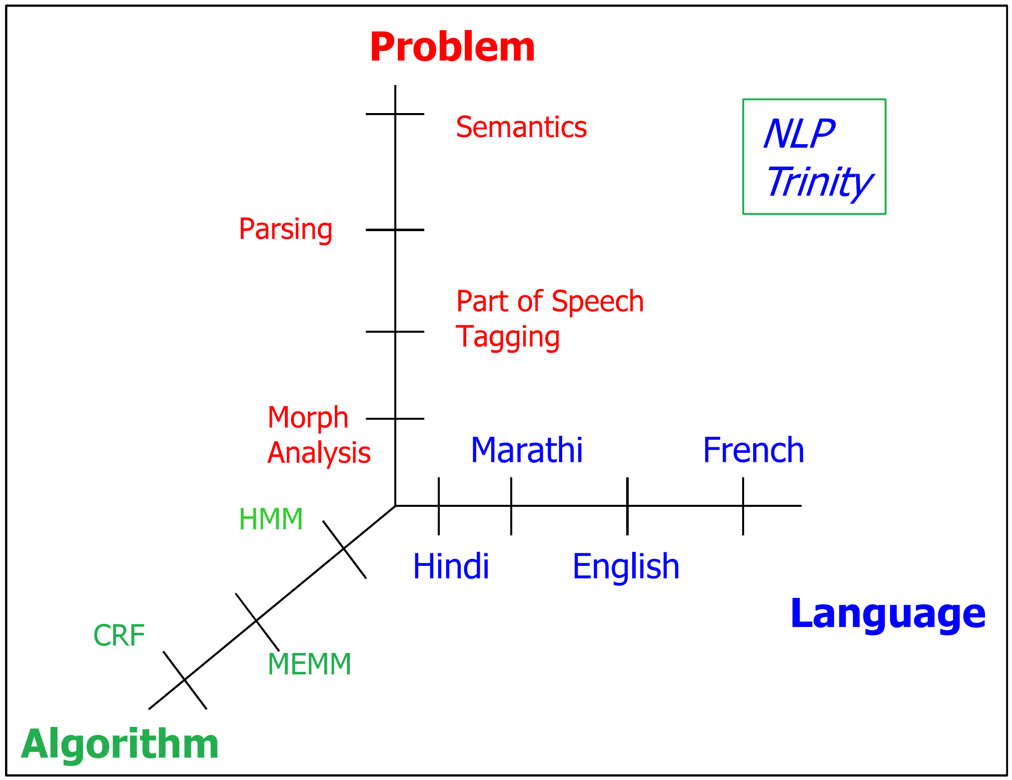
CS460/626 : Natural Language Processing/Speech, NLP and the Web

Lecture 25, 26:
Wordnet and Word Sense Disambiguation
(an overview first)

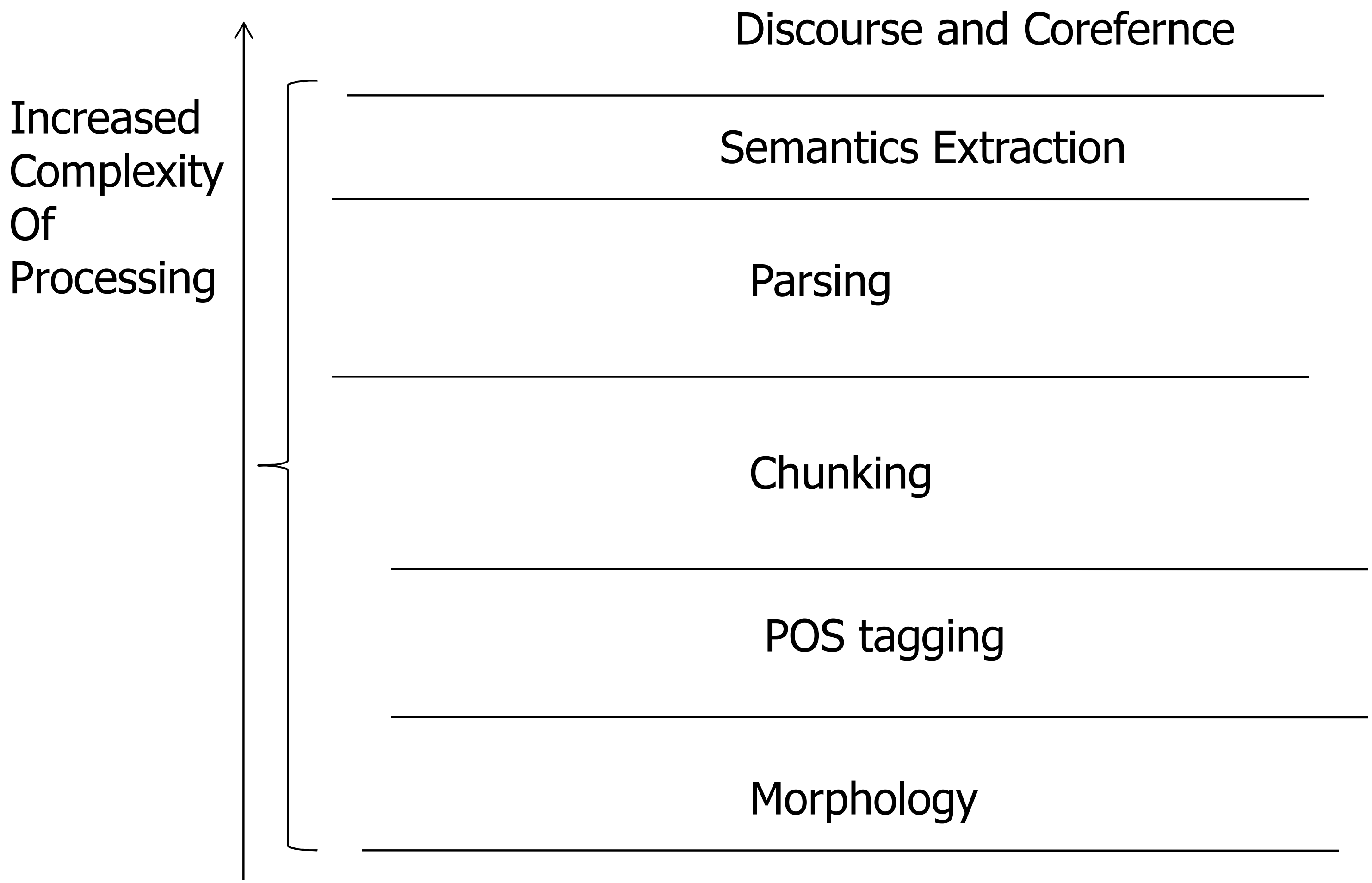
Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

15th and 18th Oct, 2012

NLP Trinity

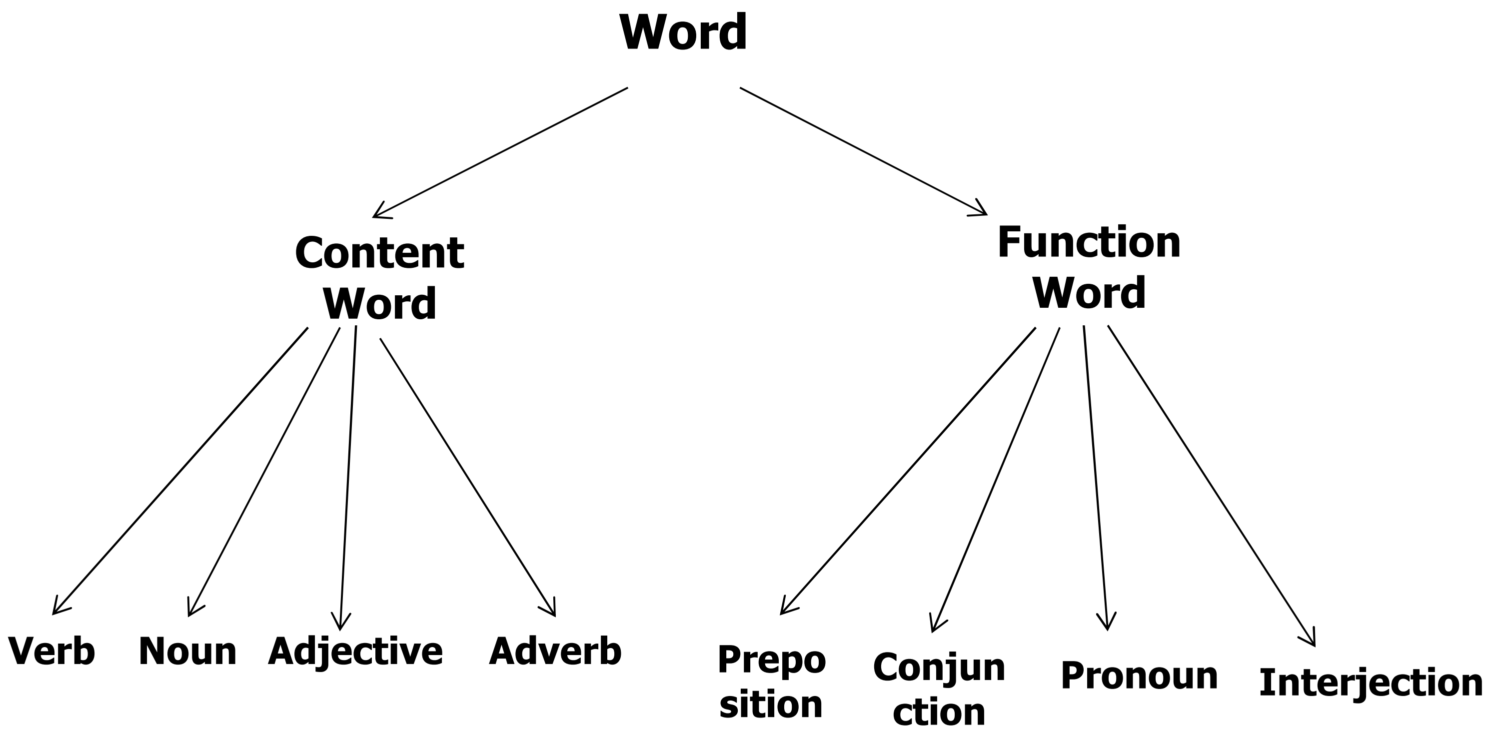

Problem
Semantics
NLP Trinity
Parsing
Part of Speech Tagging
Morph Analysis
Marathi
French
HMM
Hindi
English
Language
CRF
MEMM
Algorithm

NLP Layer


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

Background

Classification of Words



NLP: Thy Name is Disambiguation

- A word can have multiple meanings

and

- A meaning can have multiple words

Word with multiple meanings

Where there is a will,

Where there is a will,

There are hundreds of relatives

Where there is a will

There is a way

There are hundreds of relatives

A meaning can have multiple words

Proverb
“A cheat never prospers”

Proverb: “A cheat never prospers

but
can get rich faster”

WSD should be distinguished from structural ambiguity

- Correct groupings a must
  - …
- *Iran quake kills 87, 400 injured*
- *When it rains cats and dogs run for cover*

Should be distinguished from structural ambiguity

- Correct groupings a must
  - …
- *Iran quake kills 87, 400 injured*
- *When it rains, cats and dogs runs for cover*
- *When it rains cats and dogs, run for cover*

Groups of words (Multiwords) and names can be ambiguous

- *Broken guitar for sale, no strings attached (Pun)*
- *Washington voted Washington to power*
- *pujaa ne pujaa ke liye phul todaa*
- *(Pujaa plucked flowers for worship)*
- *(deep world knowledge) The use of a shin bone is to locate furniture in dark room*

Stages of processing

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

Example of WSD

- **Operation**, surgery, surgical operation, surgical procedure, surgical process -- (a medical procedure involving an incision with instruments; performed to repair damage or arrest disease in a living body; "they will schedule the operation as soon as an operating room is available"; "he died while undergoing surgery") TOPIC->(noun) surgery#1
- **Operation**, military operation -- (activity by a military or naval force (as a maneuver or campaign); "it was a joint operation of the navy and air force") TOPIC->(noun) military#1, armed forces#1, armed services#1, military machine#1, war machine#1
- **Operation** -- ((computer science) data processing in which the result is completely specified by a rule (especially the processing that results from a single instruction); "it can perform millions of operations per second") TOPIC->(noun) computer science#1, computing#1
- mathematical process, mathematical **operation**, **operation** -- ((mathematics) calculation by mathematical methods; "the problems at the end of the chapter demonstrated the mathematical processes involved in the derivation"; "they were learning the basic operations of arithmetic") TOPIC->(noun) mathematics#1, math#1, maths#1

IS WSD NEEDED IN LARGE APPLICATIONS?

Word ambiguity→topic drift in IR

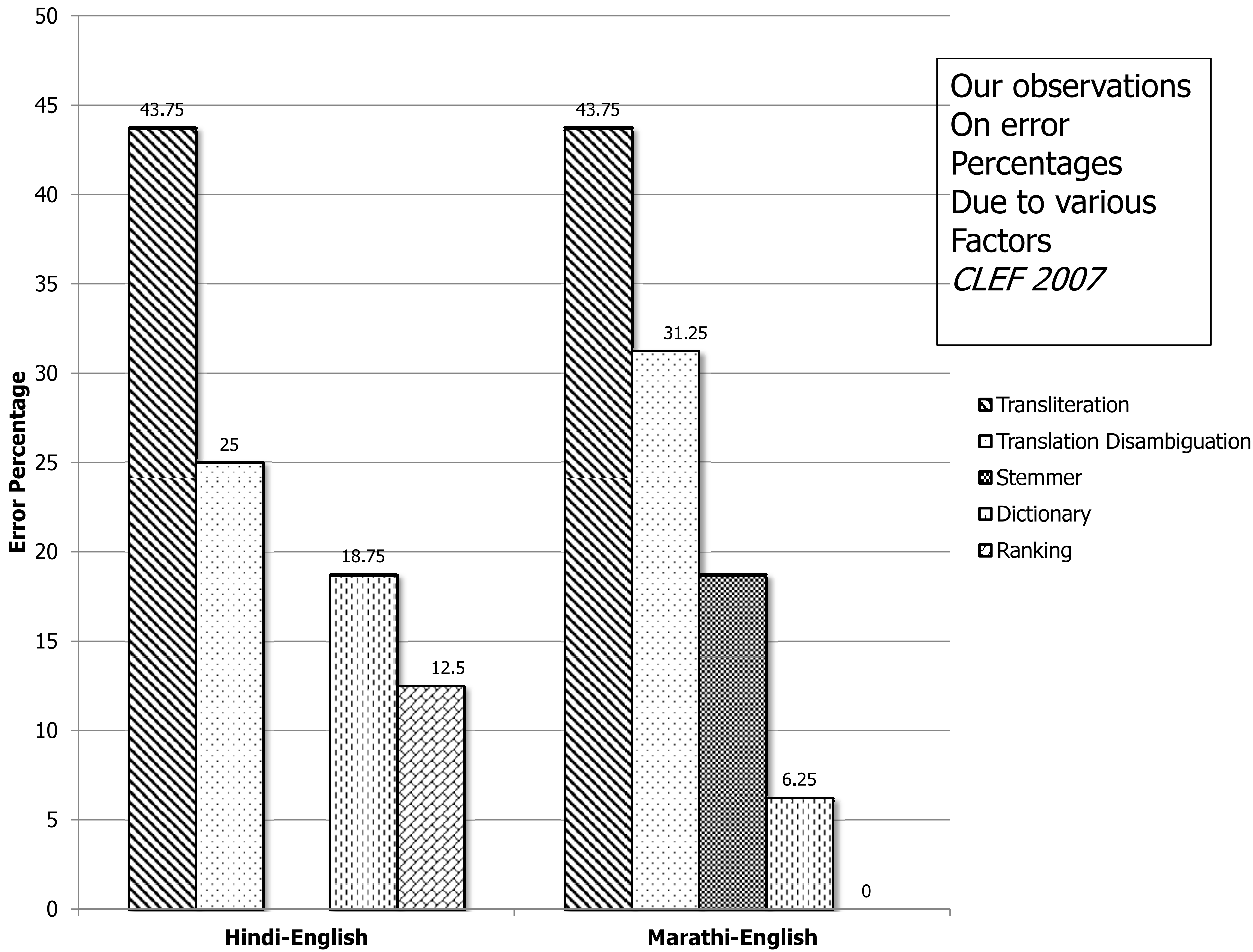
{case, container} *Drifted topic due to inapplicable sense!!!*

Query word:
*"Madrid bomb blast **case**"*

{case, suit, lawsuit}

Drifted topic due to expanded term!!!

{suit, apparel}



Our observations
On error
Percentages
Due to various
Factors
CLEF 2007
Error Percentage
50
45
40
35
30
25
20
15
10
5
0
43.75
25
18.75
12.5
43.75
31.25
6.25
0
Hindi-English
Marathi-English
Transliteration
Translation Disambiguation
Stemmer
Dictionary
Ranking

How about WSD and MT?

Zaheer Khan, the India fast bowler, has been ruled out of the remainder of the series against England.

He will return to India and will be replaced by left-arm seamer RP Singh.

Zaheer picked up a hamstring injury during the first Test at Lord's.

He had been withdrawn from the squad for India's recent Test series in the West Indies due to a right ankle injury.

भारत के तेज गेंदबाज, जहीर खान, इंग्लैंड के खिलाफ श्रृंखला के शेष के बाहर शासन किया गया है. ***(ruled in the administrative sense??)***

वह भारत लौटने और बाएँ हाथ के तेज गेंदबाज आरपी सिंह द्वारा प्रतिस्थापित किया जाएगा.

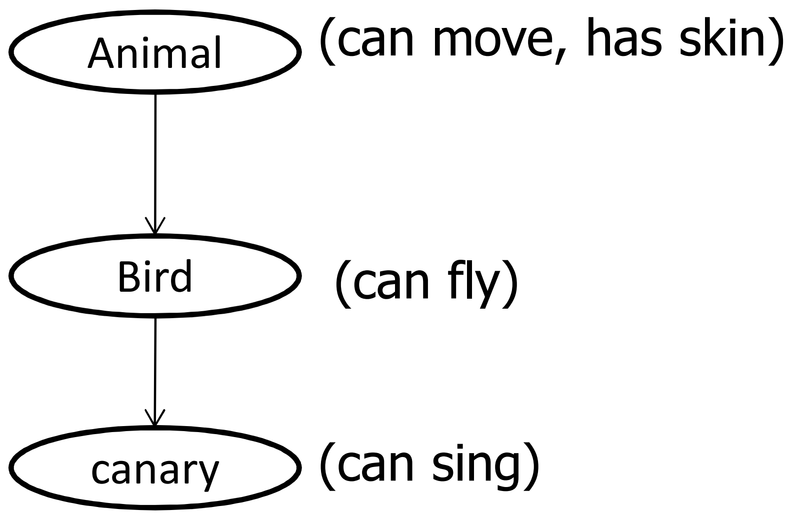
जहीर लॉर्ड्स में पहले टेस्ट के दौरान हैमस्ट्रिंग चोट उठाया. ***(lifted??)***

वह भारत की वेस्ट इंडीज में हाल ही में एक सही ***(correct??)*** *टखने की चोट के कारण टेस्ट श्रृंखला के लिए टीम से वापस ले लिया गया था.*

Wordnet

Psycholinguistic Theory

- Human lexical memory for nouns as a hierarchy.
- *Can canary sing? - Pretty fast response.*
- *Can canary fly? - Slower response.*
- *Does canary have skin? – Slowest response.*



Wordnet - a lexical reference system based on psycholinguistic theories of human lexical memory.

Essential Resource for WSD: ***Wordnet***

Word Meanings	Word Forms				
	F_1	**F_2**	**F_3**	**…**	**F_n**
M_1	(*depend*) $E_{1,1}$	(*bank*) $E_{1,2}$	(rely) $E_{1,3}$		
M_2		(*bank*) $E_{2,2}$		(*embankment*) $E_{2,\ldots}$	
M_3		(*bank*) $E_{3,2}$	$E_{3,3}$		
…				…	
M_m					$E_{m,n}$

Wordnet: History

- The first wordnet in the world was for English developed at Princeton over 15 years.
- The Eurowordnet- linked structure of European language wordnets was built in 1998 over 3 years with funding from the EC as a a mission mode project.
- Wordnets for Hindi and Marathi being built at IIT Bombay are amongst the first IL wordnets.
- All these are proposed to be linked into the **IndoWordnet** which eventually will be linked to the English and the Euro wordnets.

Basic Principle

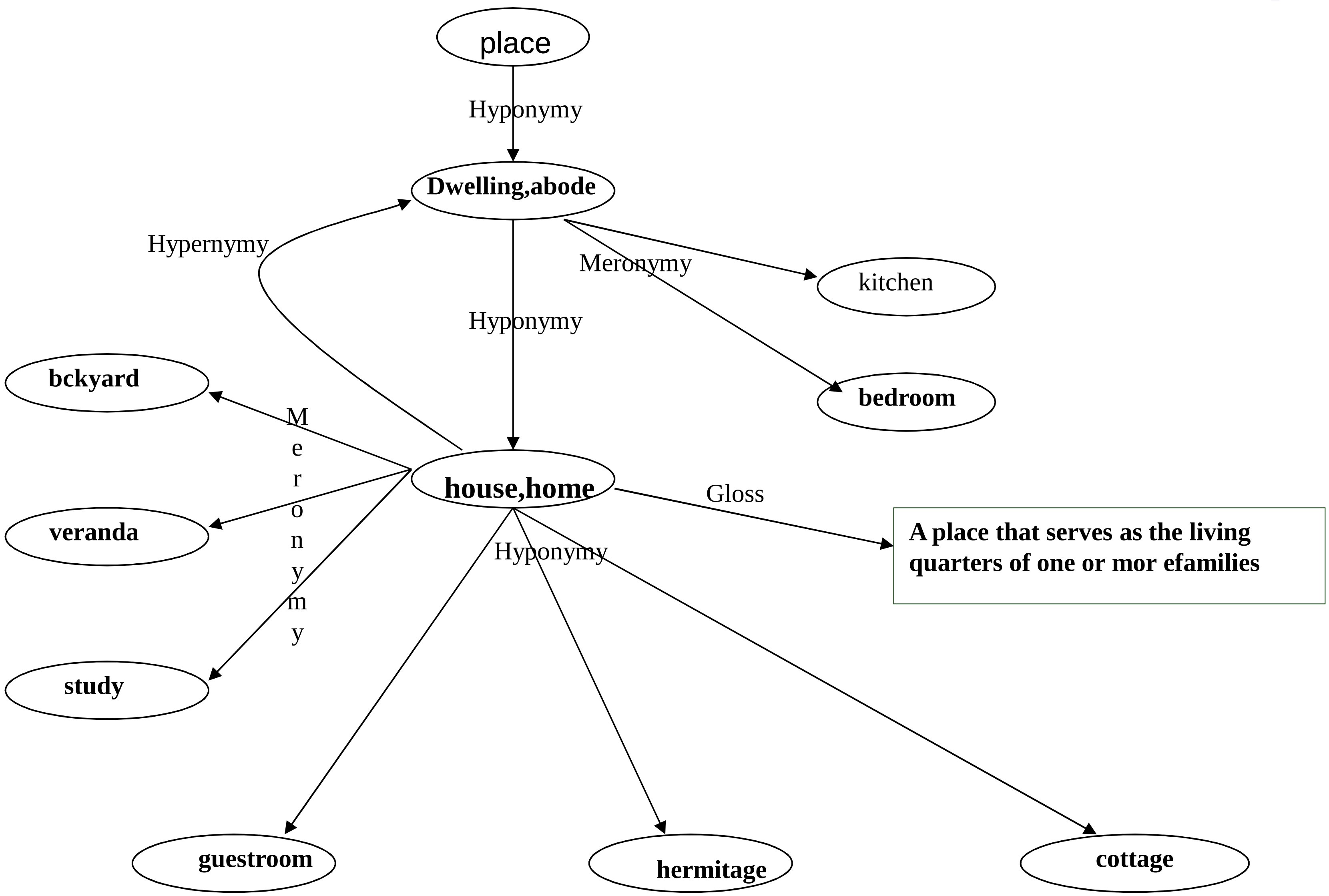
- Words in natural languages are polysemous.
- However, when synonymous words are put together, a unique meaning often emerges.
- Use is made of *Relational Semantics.*

Lexical and Semantic relations in wordnet

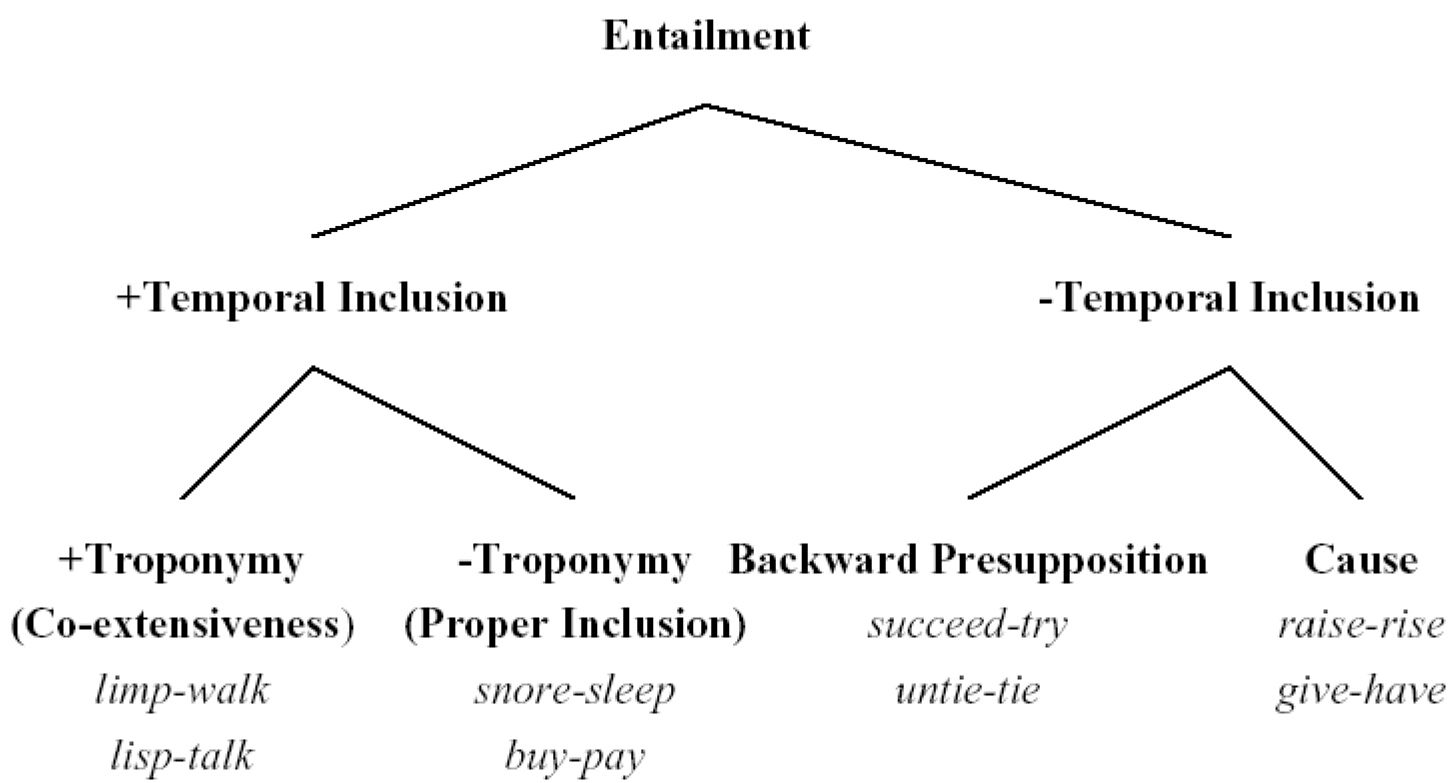
1. Synonymy
2. Hypernymy / Hyponymy
3. Antonymy
4. Meronymy / Holonymy
5. Gradation
6. Entailment
7. Troponymy

1, 3 and 5 are lexical (*word to word)*, rest are semantic (*synset to synset).*

WordNet Sub-Graph



Organization of verbs



Recent introductions in wordnet: Metonymy

- Container for contained
  - *The kettle boiled* (water)
- Possessor for possessed/attribute
  - *Where are you parked?* (car)
- Represented entity for representative
  - The government will announce new targets
- Whole for part
  - *I am going to fill up the car with petrol*

Metonymy *(contd)*

- Part for whole
  - *I noticed several new faces in the class*
- Place for institution
  - *London hosted the largest Olympic*

Question: Can you have part-part metonymy

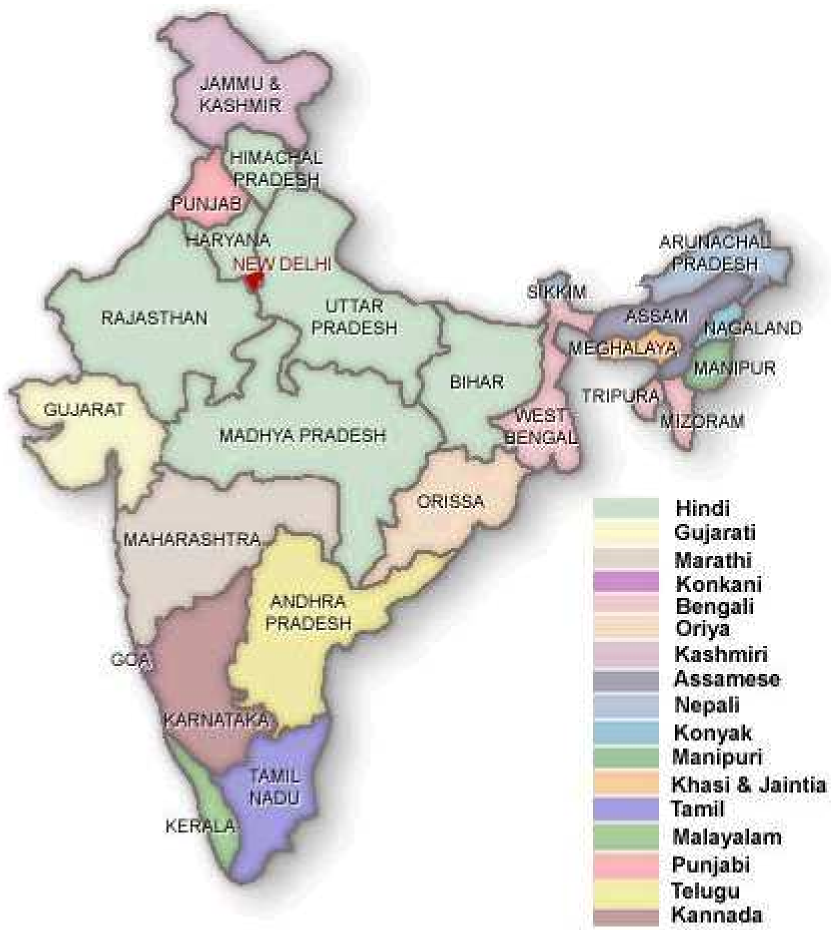
Purpose of Metonymy

- More idiomatic/natural way of expression
  - More natural to say *the kettle is boiling* as opposed to *the water in the kettle is boiling*
- Economy
  - *Room 23 is answering* (but not **is asleep*)
- Ease of access to referent
  - *He is in the phone book* (but not *on the back of my hand)
- Highlighting of associated relation
  - *The car in the front decided to turn right* (but not **to smoke a cigarette)*

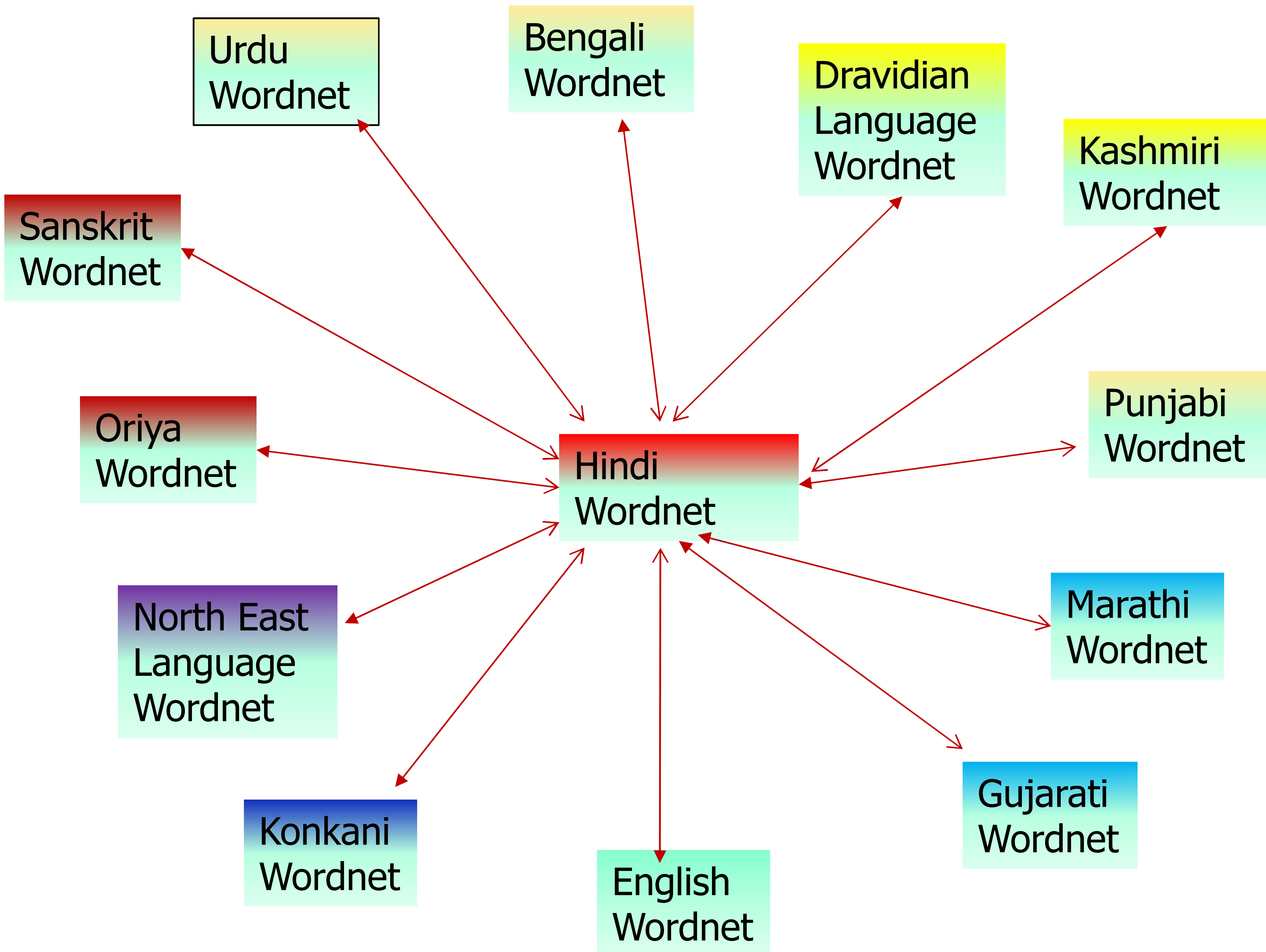
IndoWordNet

Linked Indian Language Wordnets

Linguistic Map of India



INDOWORDNET



Size of Indian Language wordnets (June, 2012) 1/2

Assamese 14958 Guahati University, Guahati, Assam

Bengali 23765 Indian Statistical Institute, Kolkata, West Bengal

Bodo 15785 Guahati University, Guahati, Assam

Gujarati 26580 Dharmsingh Desai University, Nadiad, Gujarat

Kannada 4408 Mysore University, Mysore, Karnataka

Kashmiri 23982 Kashmir University, Srinagar, Jammu and Kashmir

Konkani 25065 Goa University, Panji, Goa

Malayalam 8557 Amrita University, Coimbatore, Tamilnadu

Manipuri 16351 Manipur University, Imphal, Manipur

Marathi 24954 IIT Bombay, Mumbai, Maharastra

Size of Indian Language wordnets (June, 2012) 2/2

Nepali 11713 Assam University, Silchar, Assam

Oriya 31454 Hyderabad Central University, Hyderabad, Andhra Pradesh

Punjabi 22332 Thapar University and Punjabi University, Patiala, Punjab

Sanskrit 18980 IIT Bombay, Mumbai

Tamil 8607 Tamil University, Thanjavur, Tamilnadu

Telugu 14246 Dravidian University, Kuppam, Andhra Pradesh

Urdu 23071 Jawaharlal Nehru University, New Delhi

Categories of Synsets (1/2)

•**Universal**: Synsets which have an indigenous lexeme in all the languages *(e.g. Sun ,Earth).*

•**Pan Indian**: Synsets which have indigenous lexeme in all the Indian languages but no English equivalent *(e.g. Paapad).*

•**In-Family:** Synsets which have indigenous lexeme in the particular language family (*e.g.* the term for *Bhatija* in Dravidian languages).

Categories of Synsets (2/2)

•**Language specific**: Synsets which are unique to a language (*e.g. Bihu* in Assamese language)

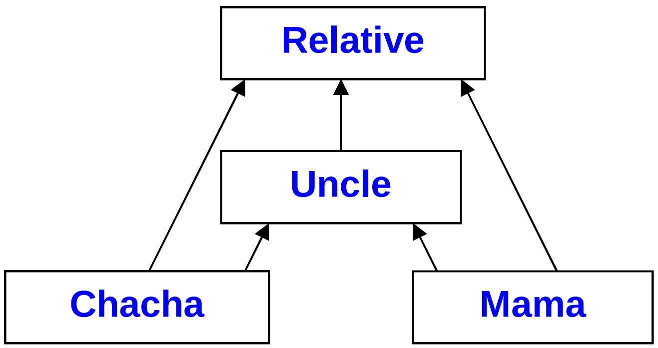
•**Rare**: Synsets which express technical terms (*e.g. ngram*).

•**Synthesized:** Synsets created in the language due to influence of another language (*e.g. Pizza*).

Expansion approach: linking is a subtle and difficult process

- To link or not to link
- While linking:
  - face lexical and semantic chasms
  - Syntactic divergences in the example sentences
    - Change of POS
    - Copula drop (Hindi→Bangla)

Linking kinship relations and fine grained concepts



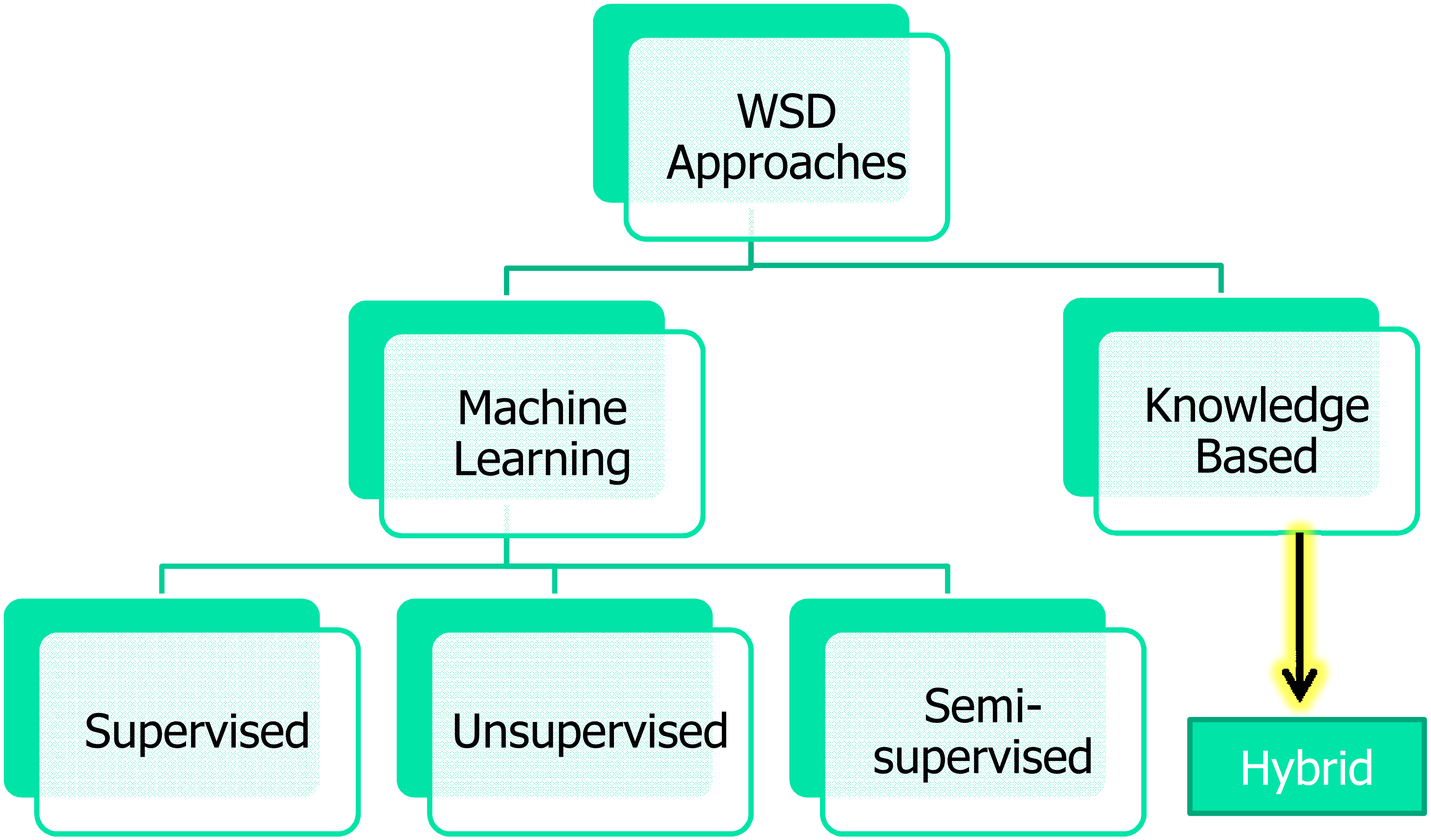
पानी direct आब

पानी hypernym त्रेश

Case of kashmiri

WSD techniques

Bird's eye view



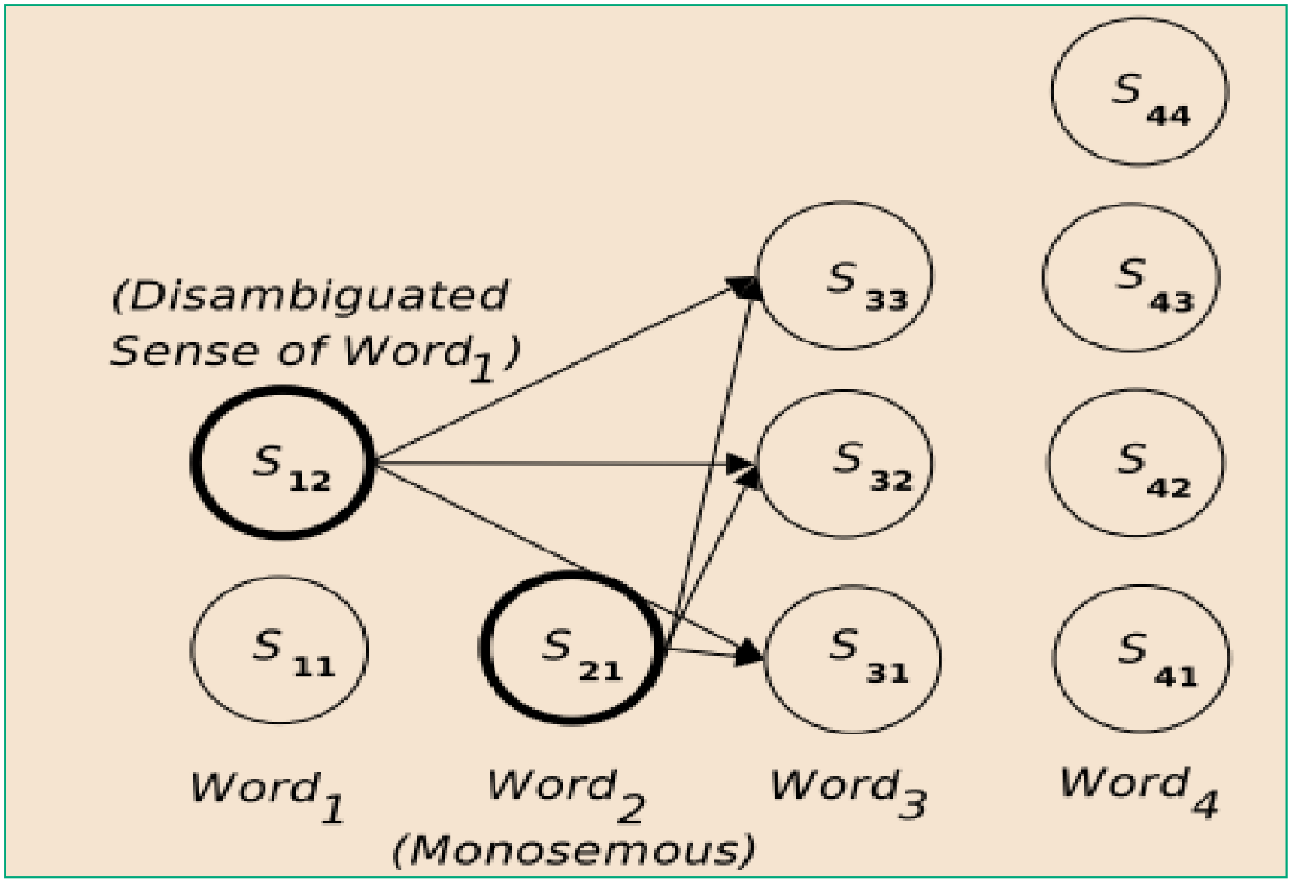
Multilingual resource constrained WSD

Long line of work…

- Mitesh Khapra, Salil Joshi and Pushpak Bhattacharyya, *It takes two to Tango: A Bilingual Unsupervised Approach for Estimating Sense Distributions using Expectation Maximization*, 5th International Joint Conference on Natural Language Processing (**IJCNLP 2011**), Chiang Mai, Thailand, November 2011.
- Mitesh Khapra, Salil Joshi, Arindam Chatterjee and Pushpak Bhattacharyya, *Together We Can: Bilingual Bootstrapping for WSD*, Annual Meeting of the Association of Computational Linguistics (**ACL 2011**), Oregon, USA, June 2011.
- Mitesh Khapra, Saurabh Sohoney, Anup Kulkarni and Pushpak Bhattacharyya, *Value for Money: Balancing Annotation Effort, Lexicon Building and Accuracy for Multilingual WSD*, Computational Linguistics Conference (**COLING 2010**), Beijing, China, August 2010.
- Mitesh Khapra, Anup Kulkarni, Saurabh Sohoney and Pushpak Bhattacharyya, *All Words Domain Adapted WSD: Finding a Middle Ground between Supervision and Unsupervision*, Conference of Association of Computational Linguistics (**ACL 2010**), Uppsala, Sweden, July 2010.
- Mitesh Khapra, Sapan Shah, Piyush Kedia and Pushpak Bhattacharyya, *Domain-Specific Word Sense Disambiguation Combining Corpus Based and Wordnet Based Parameters, 5th International Conference on Global Wordnet (**GWC2010**), Mumbai, Jan, 2010.*
- Mitesh Khapra, Sapan Shah, Piyush Kedia and Pushpak Bhattacharyya, *Projecting Parameters for Multilingual Word Sense Disambiguation, Empirical Methods in Natural Language Prfocessing (**EMNLP09**), Singapore, August, 2009.*
- Mitesh Khapra, Pushpak Bhattacharyya, Shashank Chauhan, Soumya Nair and Aditya Sharma, *Domain Specific Iterative Word Sense Disambiguation in a Multilingual Setting, International Conference on NLP (ICON08), Pune, India, December, 2008.*

Algorithm for multilingual, resource constrained WSD

Iterative WSD



Scoring function

$$S^* = \underset{i}{\operatorname{argmax}} \left(\theta_i * V_i + \sum_{j \in \mathrm{J}} W_{ij} * V_i * V_j \right)$$

$$\mathrm{J} = Set\ of\ disambiguated\ Words$$
$$\theta_i = BelongingnessToDominantConcept\ (S_i)$$
$$V_i = P(S_i \mid word)$$
$$W_{ij} = CorpusCooccurences\ (S_i, S_j)$$
$$where, {}^*\ 1/WNConceptualDistance(S_i, S_j)$$
$$*\ 1/WNSemanticGraphDistance(S_i, S_j)$$

Motivated by the Energy expression in Hopfield network

Neuron	→	*Synset*
Self-activation	→	*Corpus Sense Distribution*
Weight of connection between two neurons	→	*Weight as a function of corpus co-occurrence and Wordnet distance measures between synsets*

Iterative WSD

Algorithm 1: *performIterativeWSD(sentence)*

1. Tag all monosemous words in the sentence.
2. Iteratively disambiguate the remaining words in the sentence in increasing order of their degree of polysemy.
3. At each stage select that sense for a word which maximizes the score given by the Equation below

$$S^* = \underset{i}{\operatorname{argmax}} \left(\theta_i * V_i + \sum_{j \in J} W_{ij} * V_i * V_j \right)$$

Data

	English			Hindi		Marathi	
	Tourism	Health	SemCor	Tourism	Health	Tourism	Health
Noun	62636	53173	66194	62336	24089	45589	17477
Verb	30269	31382	84815	6386	1401	7879	3018
Adjective	25295	21091	24946	18949	8773	13107	4781
Adverb	7018	6421	11803	4860	2527	4036	1699
All	125218	112067	187758	92531	36790	70611	26975

#Polysemous words (tokens)

	English			Hindi		Marathi	
	Tourism			Tourism		Tourism	
Noun	25345	19400	17642	35812	18923	27386	11326
Verb	1413	1189	4467	3667	5109	2672	1473
Adjective	13318	9952	8969	28998	12138	16725	6087
Adverb	4449	5070	7704	13699	7152	5023	1868
All	44525	35611	38782	82176	43322	51806	20754

#monosemous words

	English Tourism			Hindi Tourism		Marathi Tourism	
Noun	4307	3185	5921	3020	1545	2269	1272
Verb	1804	1560	3135	303	120	334	239
Adjective	1738	1602	2559	778	539	663	431
Adverb	310	281	454	62	56	95	73
All	8159	6628	12069	4163	2260	3361	2015

#Polysemous unique words (types)

	English Tourism			Hindi Tourism		Marathi Tourism	
Noun	14.54	16.69	11.18	20.64	15.59	20.09	13.74
Verb	16.78	20.12	27.05	21.08	11.68	23.59	12.63
Adjective	14.55	13.17	9.75	24.36	16.28	19.77	11.09
Adverb	22.64	22.85	26.00	78.39	45.13	42.48	23.27
All	15.35	16.91	15.56	22.23	16.28	21.01	13.39

Token to Type ratio

	English Tourism	H	S	Hindi Tourism	H	Marathi Tourism	H
Noun	3.74	3.97	3.55	3.02	3.17	3.06	3.17
Verb	5.01	5.31	4.28	5.05	6.58	4.96	5.18
Adjective	3.47	3.57	3.26	2.66	2.75	2.60	2.72
Adverb	2.89	2.96	2.72	2.52	2.57	2.44	2.45
All	3.93	4.15	3.64	3.09	3.23	3.14	3.29

Average degree of WN polysemy

	English Tourism	H	S	Hindi Tourism	H	Marathi Tourism	H
Noun	1.68	1.57	1.90	1.61	1.51	1.64	1.50
Verb	2.06	1.99	2.44	2.26	1.65	1.84	1.62
Adjective	1.67	1.57	1.70	1.73	1.58	1.66	1.52
Adverb	1.81	1.75	1.79	2.13	2.05	1.77	1.70
All	1.77	1.68	1.99	1.69	1.54	1.67	1.53

Average degree of and corpus polysemy

Performance of different algorithms: monolingual WSD

Algorithms	Tourism			Health		
	P%	R%	F%	P%	R%	F%
IWSD	77.00	76.66	76.83	78.78	78.42	78.60
PPR	53.1	53.1	53.1	51.1	51.1	51.1
SVM	78.82	78.76	78.79	79.64	79.59	79.61
(McCarthy et al. 2007)	51.85	49.32	50.55	-	-	-
RB	25.50	25.50	25.50	24.61	24.61	24.61
WFS	62.15	62.15	62.15	64.67	64.67	64.67
MFS	77.60	75.20	76.38	79.43	76.98	78.19

WSD is costly![1]

WordNets

- Princeton Wordnet: ~80000 synsets: 30 man years
- Eurowordnet: 12 man years on the average for 12 languages
- Hindi wordnet: 24 man years
  - http://www.cfilt.iitb.ac.in/wordnet/webhwn/
- Indowordnet: getting created; 15 languages; 4 people on the average; in 1 year close to 15000 synsets done
- Scale of effort really huge
- Tricky too: when it comes to expanding from one wordnet to another

Machine Learning based WSD is costly![2]

Sense Annotated corpora for Machine Learning

- SemCor: ~200000 sense marked words
- SemEval/Senseval competition: to generate sense marked corpora
- Sense marked corpora created at IIT Bombay
  - http://www.cfilt.iitb.ac.in/wsd/annotated_corpus
  - English: Tourism (~170000), Health (~150000)
  - Hindi: Tourism (~170000), Health (~80000)
  - Marathi: Tourism (~120000), Health (~50000)
  - 12 man years for each <L,D> combination

Cost-accuracy trade off

High Accuracy

Supervised
(e.g., Ng and Lee, 1996;
Lee et. al., 2004)

High Cost

Low Accuracy

Unsupervised
(e.g., Agirre and Rigau, 1996;
McCarthy et al.,2004, Mihalcea,
2005)

Low Cost

This is the *dream!*

spread from one <L,D> combination to others

		Languages							
		Hindi	Marathi	Tamil	Telugu	..	..	..	Kannada
Domains	Tourism	X				..	..	..	
	Health		X			..	..	..	
	Finance					..	..	..	
	Sports					..	..	..	
	..	..	..	..	..	..	..	..	..
	..	..	..	..	..	..	..	..	..
	Politics					..	..	..	

Language Adaptation scenarios

Scenario	Annotated corpus in L_1	Annotated corpus in L_2	Synset aligned multilingual dictionary	Manual cross-linkages
Scenario 1	Sufficient	None	Yes	Yes
Scenario 2	Sufficient	None	Yes	No
Scenario 3	Sufficient	On demand	Yes	Varying amounts
Scenario 4	None	None	Yes	No
Scenario 5	Seed data	Seed data	Yes	No

Scenario 1: L_1 with annotated data L_2 with none

Projecting the sense: example (1/2)

S_1^{mar} = *the body part which connects the head to the rest of the body*
S_2^{mar} = *respect*

We are interested in estimating $P(S_1^{mar}|maan)$ and $P(S_2^{mar}|maan)$. It is also given that S_1^{hin} and S_2^{hin} are the synsets aligned to S_1^{mar} and S_2^{mar} in the MultiDict (*i.e.*, $\pi_{hin}(S_1^{mar}) = S_1^{hin}$ and $\pi_{hin}(S_2^{mar}) = S_2^{hin}$). The words in S_1^{hin} and S_2^{hin} are as given below:

S_1^{hin} = *gardan, galao, greeva, kandhar, halak*
S_2^{hin} = *pratishtha, aadar, izzat, sammaan, ...*

Further, according to the manual cross-linkages in the MultiDict, we have

$crosslink_{hin}(maan, S_1^{mar})$ = *gardan*
$crosslink_{hin}(maan, S_2^{mar})$ = *sammaan*

Projecting the sense: example (2/2)

Using the above information, we can estimate $P(S_1^{mar}|maan)$ as shown below,

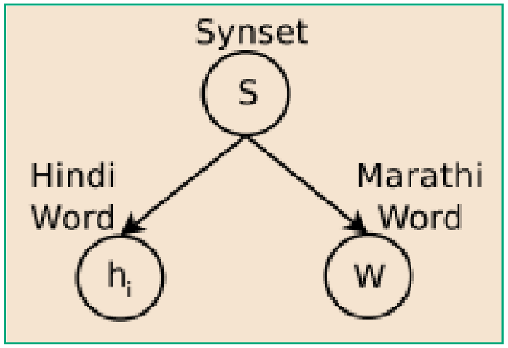
$$P(S_1^{mar}|maan) = \frac{\#(S_1^{mar}, maan)}{\#(S_1^{mar}, maan) + \#(S_2^{mar}, maan)}$$

Replacing these counts by the counts of the cross-linked words, we get

$$P(S_1^{mar}|maan) = \frac{\#(S_1^{hin}, gardan)}{\#(S_1^{hin}, gardan) + \#(S_2^{hin}, sammaan)}$$

$P(S_2^{mar}|maan)$ can be estimated similarly

Projecting with probabilistic cross linking: example (1/3)



$$E[\#(S_1^{mar}, maan)] = P(gardan|maan, S_1^{hin}) * \#(S_1^{hin}, gardan)$$
$$+ P(gala|maan, S_1^{hin}) * \#(S_1^{hin}, gala)$$
$$+ P(greeva|maan, S_1^{hin}) * \#(S_1^{hin}, greeva)$$
$$+ P(kandhar|maan, S_1^{hin}) * \#(S_1^{hin}, kandhar)$$
$$+ P(halak|maan, S_1^{hin}) * \#(S_1^{hin}, halak)$$

Projecting with probabilistic cross linking: example (3/3)

Once $E[\#(S_1^{mar}, maan)]$ and $E[\#(S_2^{mar}, maan)]$ have been estimated $P(S_1^{mar}|maan)$ can be estimated as follows,

$$\cancel{P(S_1^{mar}|maan) = \frac{\#(S_1^{mar}, maan)}{\#(S_1^{mar}, maan) + \#(S_2^{mar}, maan)}}$$

Replacing these counts by the expected counts, we get

$$P(S_1^{mar}|maan) = \frac{E[\#(S_1^{mar}, maan)]}{E[\#(S_1^{mar}, maan)] + E[\#(S_2^{mar}, maan)]}$$

$P(S_2^{mar}|maan)$ can be estimated similarly.

Validating sense projection

Sr. No	Marathi Word	Synset	P(S|word) as learnt from sense tagged Marathi corpus	P(S|word) as projected from sense tagged Hindi corpus
1	किंमत (kimat)	{ worth }	0.684	0.714
		{ price }	0.315	0.285
2	रस्ता (rasta)	{ roadway }	0.164	0.209
		{road, route}	0.835	0.770
3	ठिकाण (thikan)	{ land site, place}	0.962	0.878
		{ home }	0.037	0.12

For Hindi →Marathi

- *Average KL Divergence=0.29*
- *Spearman's Correlation Coefficient=0.77*

For Hindi →Bengali

- *Average KL Divergence=0.05*
- *Spearman's Correlation Coefficient=0.82*

There is a high degree of similarity between the distributions learnt using projection and those learnt from the self corpus.

Co-occurrence parameter Projection

Sr. No	Synset	Co-occurring Synset	P(co-occurrence) as learnt from sense tagged Marathi corpus	P(co-occurrence) as learnt from sense tagged Hindi corpus
1	{रोप, रोपटे} {small bush}	{झाड, वृक्ष, तरुवर, द्रुम, तरू, पादप} {tree}	0.125	0.125
2	{मेघ, अभ्र} {cloud}	{आकाश, आभाळ, अंबर} {sky}	0.167	0.154
3	{क्षेत्र, इलाक़ा, इलाका, भूखंड} {geographical area}	{यात्रा, सफ़र} {travel}	0.0019	0.0017

Within a domain, the statistics of co-occurrence of senses remain the same across languages.

Co-occurrence of the synsets {cloud} and {sky} is almost same in the Marathi and Hindi corpus.

IWSD with parameter projection (Marathi using Hindi)

Algorithms	Tourism			Health		
	P%	R%	F%	P%	R%	F%
MCL	73.36	68.83	71.02	75.86	66.6	70.93
PCL	68.57	67.93	68.25	65.75	64.53	65.14
IWSD-Self	78.36	77.77	78.07	78.15	75.91	77.01
WFS	57.15	57.15	57.15	55.55	55.55	55.55

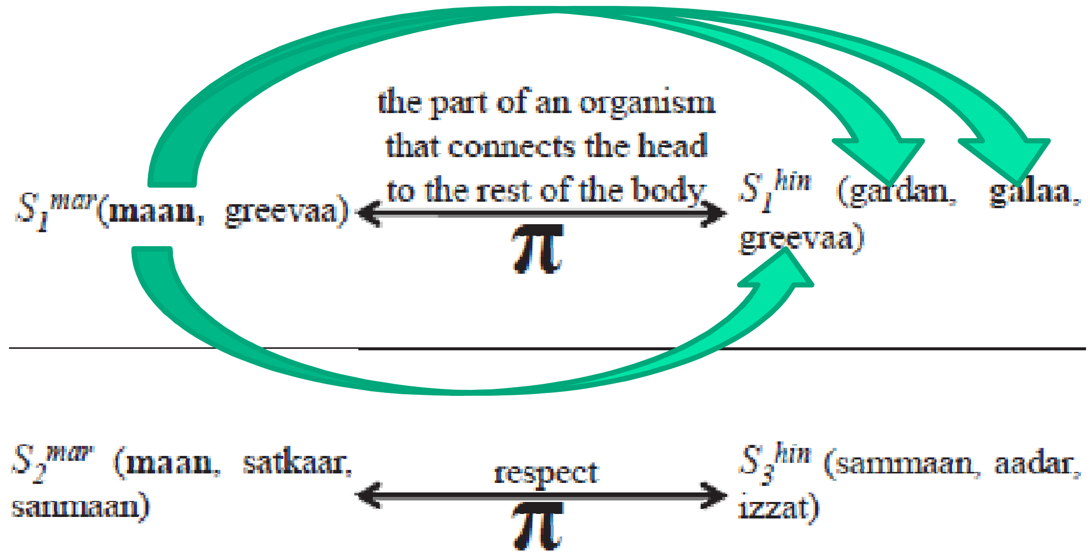
MCL-manual cross linked; PCL: probabilistic cross linked; IWSD=Self: IWSD with own language training data; WFS: wordnet first sense

Mitesh Khapra, Sapan Shah, Piyush Kedia and Pushpak Bhattacharyya, *Projecting Parameters for Multilingual Word Sense Disambiguation*, Empirical Methods in Natural Language Prfocessing (**EMNLP09**), Singapore, August, 2009.

Mitesh Khapra, Saurabh Sohoney, Anup Kulkarni and Pushpak Bhattacharyya, *Value for Money: Balancing Annotation Effort, Lexicon Building and Accuracy for Multilingual WSD*, Computational Linguistics Conference (**COLING 2010**), Beijing, China, August 2010.

Scenario 3: EM- both L_1 and L_2 with no annotated data

ESTIMATING SENSE DISTRIBUTIONS

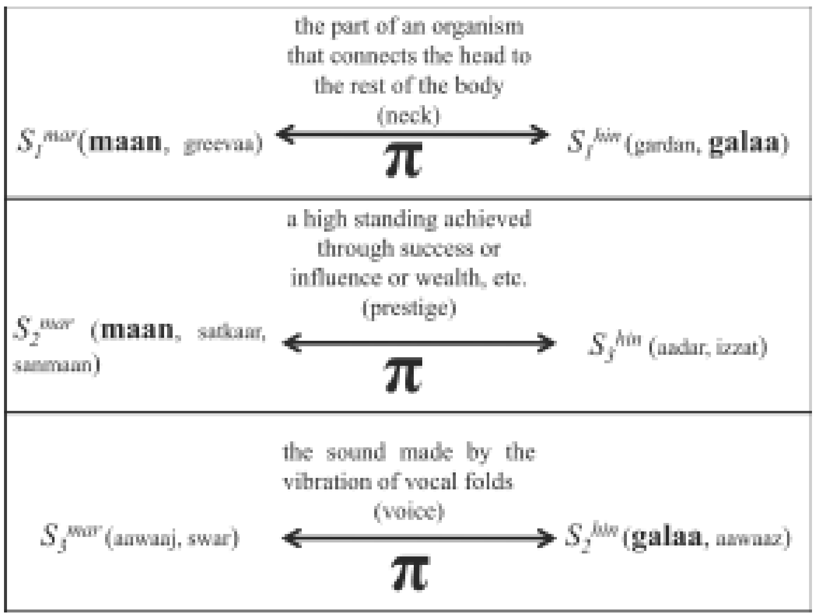


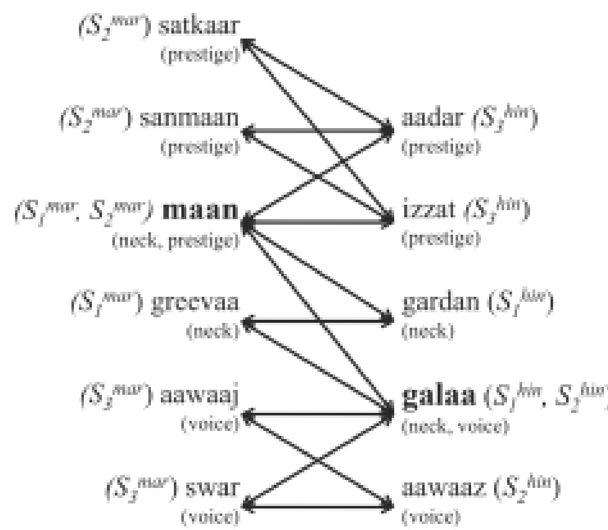
If sense tagged Marathi corpus were available, we could have estimated

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

But such a corpus is not available

Framework: Figure 1 and Figure 2





E-M steps

E-step

$$P(S_1^{mar}|maan) \approx \frac{P(S_1^{hin}|gardan)\cdot\#(gardan) + P(S_1^{hin}|galaa)\cdot\#(galaa)}{Z}$$

$$\text{where, } Z = P(S_1^{hin}|gardan)\cdot\#(gardan) + P(S_1^{hin}|galaa)\cdot\#(galaa) + P(S_3^{hin}|aadar)\cdot\#(aadar) + P(S_3^{hin}|izzat)\cdot\#(izzat)$$

M-step

$$P(S_1^{hin}|galaa) \approx \frac{P(S_1^{mar}|maan)\cdot\#(maan) + P(S_1^{mar}|greeva)\cdot\#(greeva)}{Z}$$

$$Z = P(S_1^{mar}|maan)\cdot\#(maan) + P(S_1^{mar}|greeva)\cdot\#(greeva) + P(S_3^{mar}|aawaaj)\cdot\#(aawaaj) + P(S_3^{mar}|swar)\cdot\#(swar)$$

where,

$S_1^{mar} = \pi_{hin}(S_1^{hin})$ (see Figure 1)

$S_3^{mar} = \pi_{mar}(S_2^{hin})$ (see Figure 1)

$(maan, greeva) \in translations_{mar}(galaa, S_1^{hin})$ (see Figure 2)

$(aawaaj, swar) \in translations_{mar}(galaa, S_2^{hin})$ (see Figure 2)

Points to note…

- Symmetric formulation
- *E* and *M* steps are identical except for the change in language
- Either can be treated as the E-step, making the other as the M-step
- A back-and-forth traversal over translation correspondences in the two languages
- Does not require parallel corpus – only in-domain corpus is needed

In General..

E-Step:

$$P(S_k^{L_1}|u) \approx \frac{\sum_{v} P(\pi_{L_2}(S_k^{L_1})|v) \cdot \#(v)}{\sum_{S_i^{L_1}} \sum_{y} P(\pi_{L_2}(S_i^{L_1})|y) \cdot \#(y)}$$

$$\begin{aligned} \textit{where, } S_k^{L_1}, S_i^{L_1} &\in synsets_{L_1}(u) \\ v &\in translations_{L_2}(u, S_k^{L_1}) \\ y &\in translations_{L_2}(u, S_i^{L_1}) \end{aligned}$$

M-Step:

$$P(S_j^{L_2}|v) \approx \frac{\sum_{a} P(\pi_{L_1}(S_j^{L_2})|a) \cdot \#(a)}{\sum_{S_i^{L_2}} \sum_{b} P(\pi_{L_1}(S_i^{L_2})|b) \cdot \#(b)}$$

$$\begin{aligned} \textit{where, } S_j^{L_2}, S_i^{L_2} &\in synsets_{L_2}(v) \\ a &\in translations_{L_1}(v, S_j^{L_2}) \\ b &\in translations_{L_1}(v, S_i^{L_2}) \end{aligned}$$

Experimental Setup

- Languages: Hindi, Marathi
- Domains: Tourism and Health (largest domain-specific sense tagged corpus)

Category	Polysemous words Tourism	Polysemous words Health	Monosemous words Tourism	Monosemous words Health
Noun	62336	24089	35811	18923
Verb	6386	1401	3667	5109
Adjective	18949	8773	28998	12138
Adverb	4860	2527	13699	7152
All	92531	36790	82175	43322

Table 2: Polysemous and Monosemous words per category in each domain for Hindi

Category	Polysemous words Tourism	Polysemous words Health	Monosemous words Tourism	Monosemous words Health
Noun	45589	17482	27386	11383
Verb	7879	3120	2672	1500
Adjective	13107	4788	16725	6032
Adverb	4036	1727	5023	1874
All	70611	27117	51806	20789

Table 3: Polysemous and Monosemous words per category in each domain for Marathi

Category	Avg. degree of wordnet polysemy for polysemous words: Tourism	Health
Noun	3.02	3.17
Verb	5.05	6.58
Adjective	2.66	2.75
Adverb	2.52	2.57
All	3.09	3.23

Table 4: Average degree of wordnet polysemy per category in the 2 domains for Hindi

Category	Avg. degree of wordnet polysemy for polysemous words: Tourism	Health
Noun	3.06	3.18
Verb	4.96	5.18
Adjective	2.60	2.72
Adverb	2.44	2.45
All	3.14	3.29

Table 5: Average degree of wordnet polysemy per category in the 2 domains for Marathi

Algorithms Being Compared

- EM (our approach)
- Personalized PageRank (Agirre and Soroa, 2009)
- State-of-the-art bilingual approach (using Mutual Information) (Kaji and Morimoto, 2002)
- Random Baseline
- Wordnet First sense baseline (supervised baseline)

Results

Algorithm	Average				
	N	R	A	V	O
WFS	60.00	68.64	52.39	39.65	57.29
EM	53.35	56.95	51.39	29.98	51.26
PPR	56.17	0.00	38.94	29.74	48.88
RB	34.74	44.32	39.38	17.21	34.79
MI	10.97	3.89	10.07	5.63	9.97

Average 4-fold cross validation results averaged over all Language-Domain pairs for all words

- Performs better than other state-of-the-art knowledge based and unsupervised approaches
- Does not beat the Wordnet First Sense Baseline which is a supervised baseline

Error Analysis – Non-Progressiveness estimation

- Some words have the same translations in the target language across senses
  - *saagar(hindi) ←→ samudra (marathi) ("large water body" as well as "limitless")*
- *Such words thus form a closed loop of translations*
- *In such cases the algorithm does not progress and gets stuck with the initial values*
- *Same is the case for some language specific words for which corresponding synsets were not available in the other language*
- *Such words accounted for 17-19% of the total words in the test corpus*

have problem of Non Progressive Estimation

Algorithm	Average				
	N	R	A	V	O
WFS	60.86	65.00	52.64	42.00	57.70
EM	57.78	61.28	54.16	31.87	54.98
PPR	58.03	0.00	40.91	30.58	50.42
RB	34.17	43.37	39.21	15.64	34.13
MI	9.62	4.69	8.96	4.17	8.78

- *Results are now closer to Wordnet First Sense Baseline*
- *For 2 out of the 4 language domain pairs the results are slightly better than WFS – remarkable for an unsupervised approach*

Conclusions (1/2)

- NLP is all about processing ambiguity, with WSD as a fundamental task
- Resource constraint and multilinguality brings additional challenge
- Wordnet: Great unifier of India (similar to *Adi Shankaracharya, Bollywood films…)*
- Getting linked with English WN; would like to link with Eurowordnet
- Application in MT, Search, Language teaching, e-commerce

Future work

- Closer study needed for familialy close languages
- Usage of language specific properties, in particular, **morphology**
- The projection idea can be used in other NLP problems like POS tagging and Parsing

URLs

- For resources

  www.cfilt.iitb.ac.in

- For publications

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

Thank you

Questions and comments?