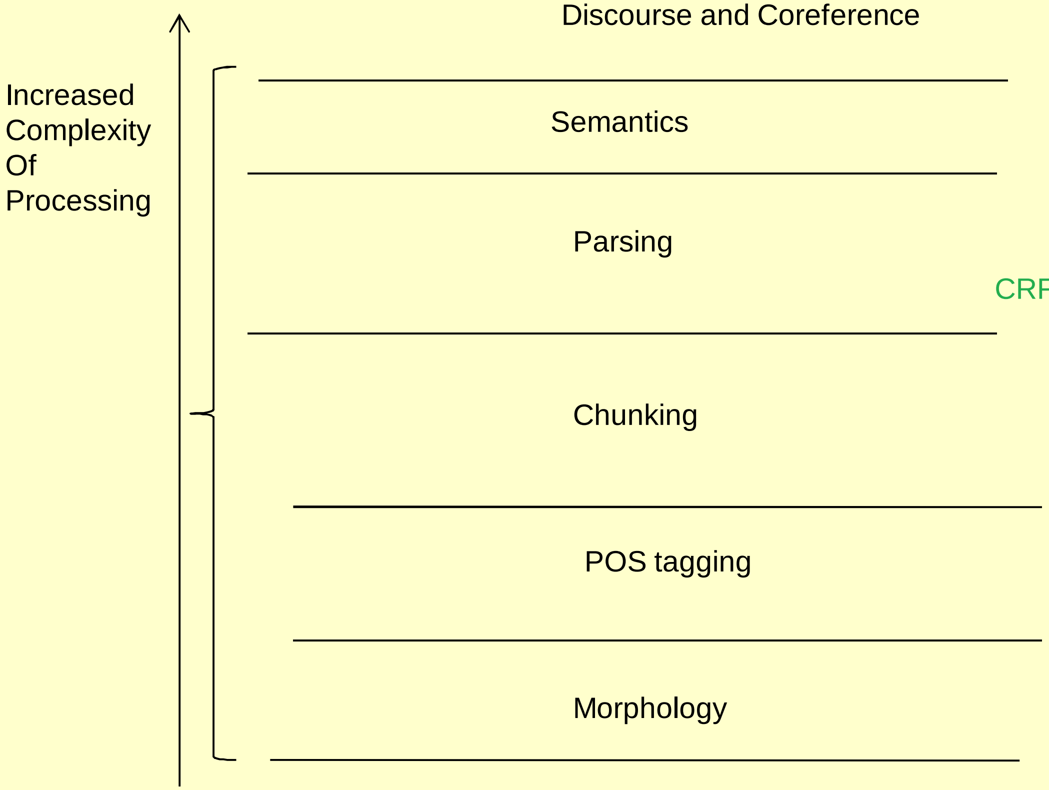
# NLP and ML: Points of Synergy and Divergence

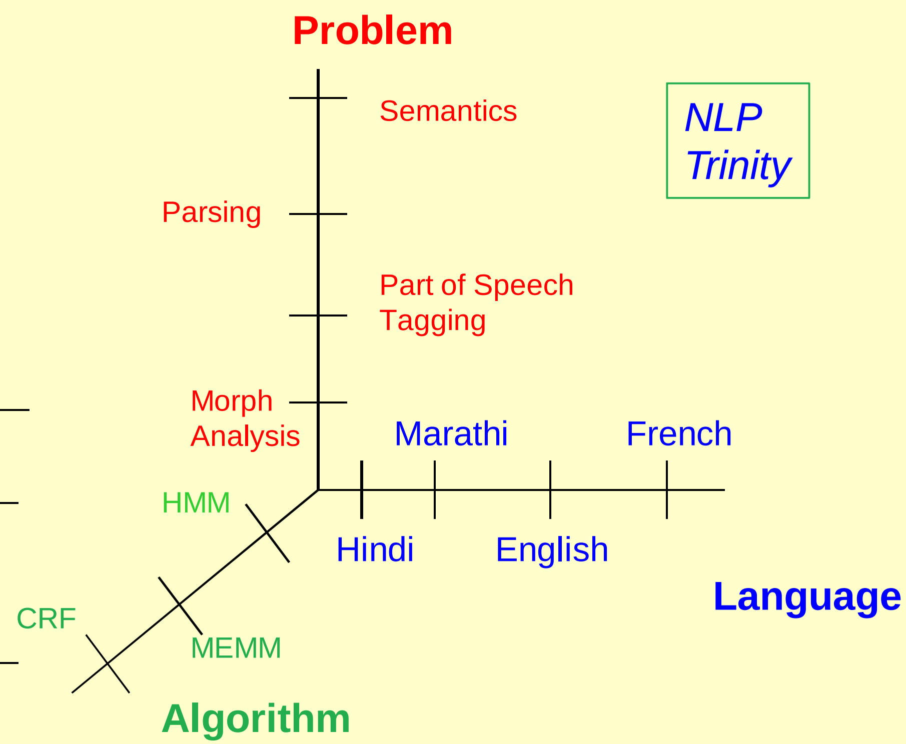
Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay
www.cse.iitb.ac.in/~pb

*(IWML workshop, IIT Kanpur, 2nd July, 2013)*

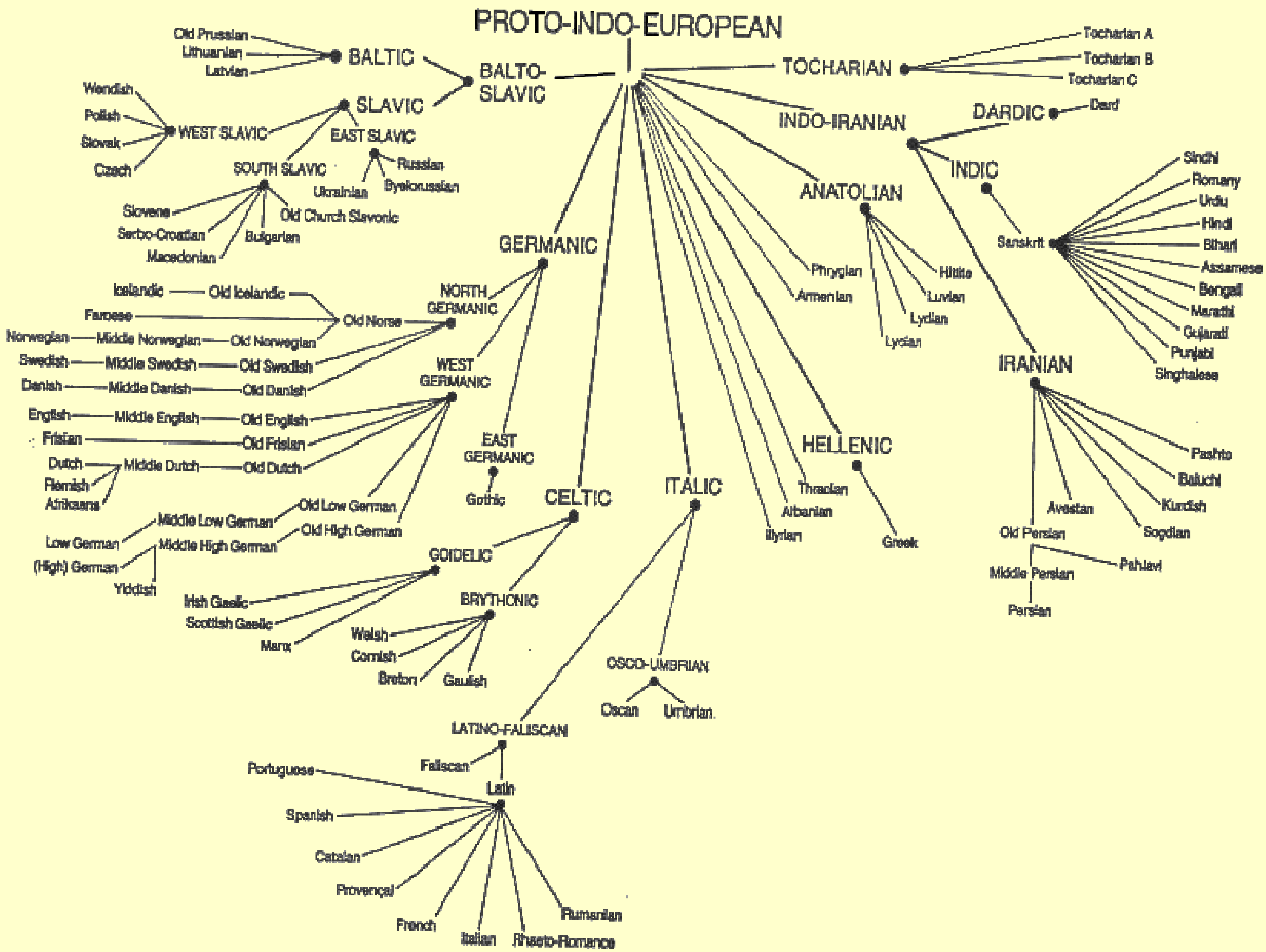
# Perspective

# NLP: a useful view



# Language Typology



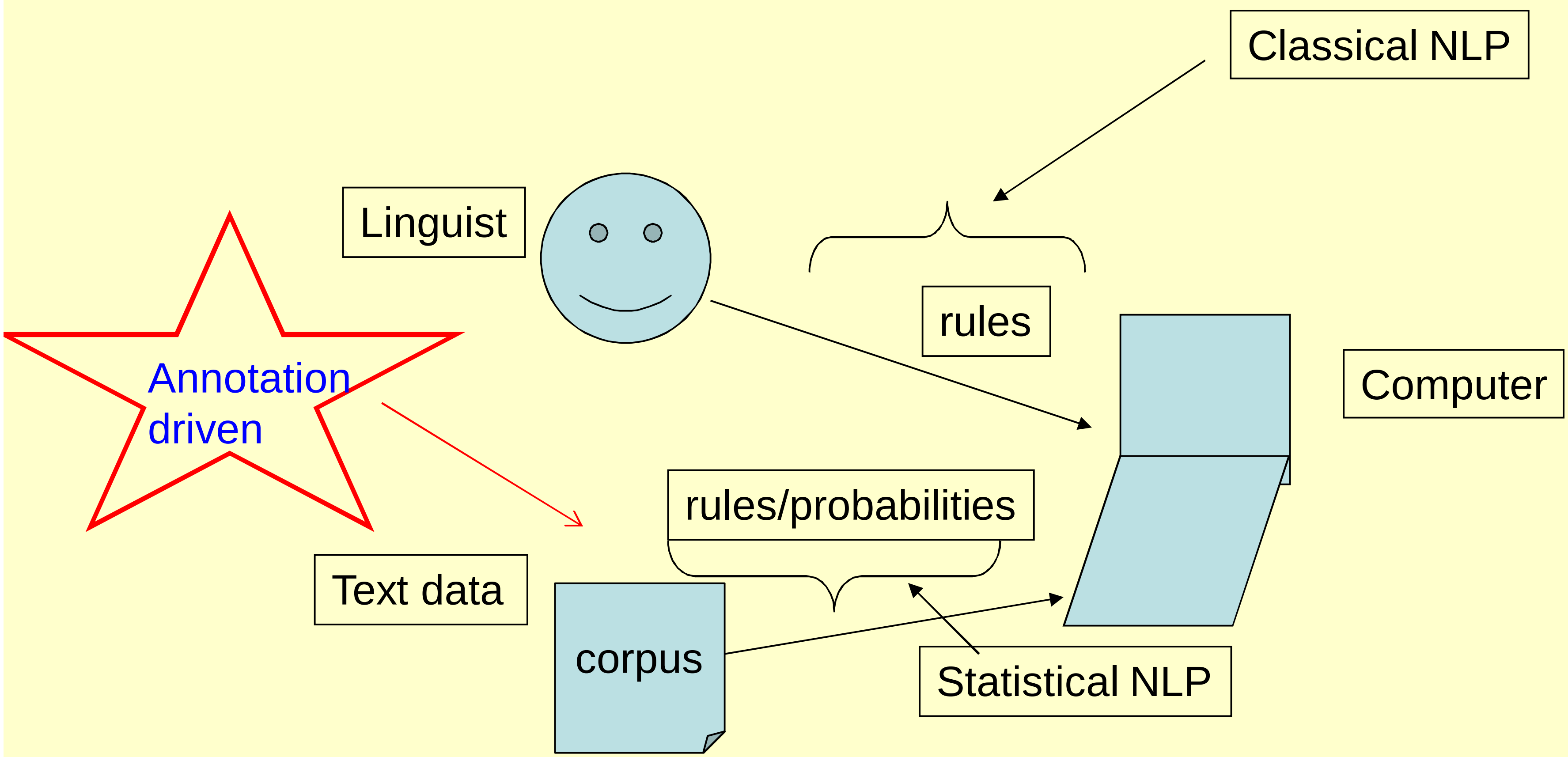
# Languages differ in expressing thoughts: Agglutination

- Finnish*:* “istahtaisinkohan”
- English: "I wonder if I should sit down for a while"

Analysis:

- ist + "sit", verb stem
- ahta + verb derivation morpheme, "to do something for a while"
- isi + conditional affix
- n + 1st person singular suffix
- ko + question particle
- han a particle for things like reminder (with declaratives) or "softening" (with questions and imperatives)

# Two approaches to NLP: Knowledge Based and ML based



# Empiricism vs. Rationalism

- Ken Church, “A Pendulum Swung  too Far”, LILT, 2011
  - Availability of huge amount of data: what to do with it?
  - 1950s: Empiricism (Shannon, Skinner, Firth, Harris)
  - 1970s: Rationalism (Chomsky, Minsky)
  - 1990s: Empiricism (IBM Speech Group, AT & T)
  - 2010s: Return of Rationalism?

***Resource generation will play a vital role in this revival of rationalism***

# Roadmap

- Perspective (done)
- Annotation
- Cooperative WSD
- Thwarting in sentiment analysis link
- Eye tracking based WSD link
- Multiword expressions link
- Conclusions

# Annotation

# Definition

(*Eduard Hovy, ACL 2010, tutorial on annotation)*

- Annotation (‘tagging’) is the process of adding new information into raw data by human annotators.

- Typical annotation steps:
  – Decide which fragment of the data to annotate
  – Add to that fragment a specific bit of information
  – chosen from a fixed set of options

# Example of annotation: sense marking

एक_4187 नए शोध_1138 के अनुसार_3123 जिन लोगों_1189 का सामाजिक_43540 जीवन_125623
व्यस्त_48029 होता है उनके दिमाग_16168 के एक_4187
हिस्से_120425 में अधिक_42403 जगह_113368 होती है।

(According to a new research, those people who have a busy social life, have larger space in a part of their brain).

नेचर न्यूरोसाइंस में छपे एक_4187 शोध_1138 के अनुसार_3123 कई_4118 लोगों_1189 के दिमाग_16168
के स्कैन से पता_11431 चला कि दिमाग_16168 का एक_4187 हिस्सा_120425 एमिगडाला सामाजिक_43540
व्यस्तताओं_1438 के साथ_328602 सामंजस्य_166
के लिए थोड़ा_38861 बढ़_25368 जाता है। यह शोध_1138 58 लोगों_1189 पर किया गया जिसमें उनकी
उम्र_13159 और दिमाग_16168 की साइज़ के आँकड़े_128065
लिए गए। अमरीकी_413405 टीम_14077 ने पाया_227806 कि जिन लोगों_1189 की सोशल नेटवर्किंग
अधिक_42403 है उनके दिमाग_16168 का एमिगडाला
वाला हिस्सा_120425 बाकी_130137 लोगों_1189 की तुलना_में_38220 अधिक_42403 बड़ा_426602 है।
दिमाग_16168 का एमिगडाला वाला हिस्सा_120425
भावनाओं_1912 और मानसिक_42151 स्थिति_1652 से जुड़ा हुआ माना_212436 जाता है।

# Ambiguity of लोगों (People)

- **लोग**, **जन**, **लोक**, **जनमानस**, **पब्लिक** - एक से अधिक व्यक्ति *"लोगों के हित में काम करना चाहिए"*
  - **(English synset) multitude, masses, mass, hoi_polloi, people, the_great_unwashed - the common people generally *"separate the warriors from the mass" "power to the people"***
- **दुनिया**, **दुनियाँ**, **संसार**, **विश्व**, **जगत**, **जहाँ**, **जहान**, **ज़माना**, **जमाना**, **लोक**, **दुनियावाले**, **दुनियाँवाले**, **लोग** - संसार में रहने वाले लोग *"महात्मा गाँधी का सम्मान पूरी दुनिया करती है / मैं इस दुनिया की परवाह नहीं करता / आज की दुनिया पैसे के पीछे भाग रही है"*
  - **(English synset) populace, public, world - people in general considered as a whole *"he is a hero in the eyes of the public”***

# Sense Marked corpora in Marathi

१४व्या शतकापासून_110076 ही इमारत_11502 कायदेविषयक_46868 व्यवसायासाठी_196 वापरली_29601 जात आहे.
गिल्डहॉलच्या नवीन_43064 कला_11642 दालनाचे_151743 बांधकाम_123565 आता_311083 पूर्ण_46726 झाले आहे.
एकच उणीवेची गोष्ट_1923 म्हणजे मधून छेदत_253701 जाणारा गर्दीचा_15499 मुख्य_451582 रस्ता_15828
हॅम्पस्टेड हीथच्या जवळ_3373 हाय गेट हिल आहे, व तिच्या माथ्यावर_11120 हाय गेट हे प्रसन्न_42949 खेडे_153030 आहे.
गिल्डहॉल हे लंडन_123879 शहराचे_13871 नागरी_46348 प्रशासनाचे_11009 प्रमुख_451582 कार्यालय_13980 आहे.
ह्या गावात_14696 पुरातन_41661 मूल्यवान वस्तूंची_1923 अनेक_4118 कुतूहलपूर्ण दुकाने_16187 व कॅफे आहेत.
एकच उणीवेची गोष्ट_1923 म्हणजे मधून छेदत_253701 जाणारा गर्दीचा_15499 मुख्य_451582 रस्ता_15828

*Snapshot of a Marathi sense tagged paragraph*

# Structural annotation

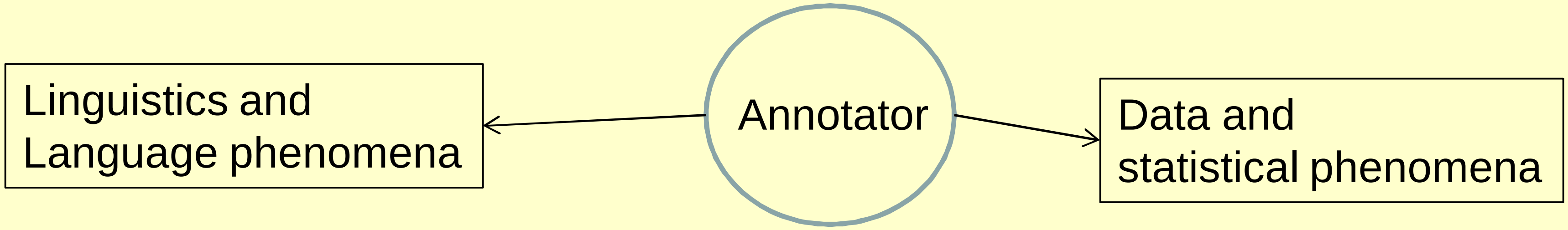
Raw Text: "My dog also likes eating sausage."

```
(ROOT
      (S
         (NP
             (PRP$ My) (NN dog))
          (ADVP (RB also))
         (VP (VBZ likes)
             (S (VP (VBG eating)
                 (NP (NN sausage))))) (. .)))
```

**poss(dog-2, My-1)**
**nsubj(likes-4, dog-2)**
**advmod(likes-4, also-3)**
**root(ROOT-0, likes-4)**
**xcomp(likes-4, eating-5)**
**dobj(eating-5, sausage-6)**

# Good annotators and good annotation designers are rare to find

- An annotator has to understand BOTH language phenomena and the data
- An annotation designer has to understand BOTH linguistics and statistics!



# Penn tag set

| | | | | | |
|---|---|---|---|---|---|
| CC | Coord Conjuncn | *and,but,or* | NN | Noun, sing. or mass | *dog* |
| CD | Cardinal number | *one,two* | NNS | Noun, plural | *dogs* |
| DT | Determiner | *the,some* | NNP | Proper noun, sing. | *Edinburgh* |
| EX | Existential there | *there* | NNPS | Proper noun, plural | *Orkneys* |
| FW | Foreign Word | *mon dieu* | PDT | Predeterminer | *all, both* |
| IN | Preposition | *of,in,by* | POS | Possessive ending | *'s* |
| JJ | Adjective | *big* | PP | Personal pronoun | *I,you,she* |
| JJR | Adj., comparative | *bigger* | PP$ | Possessive pronoun | *my,one's* |
| JJS | Adj., superlative | *biggest* | RB | Adverb | *quickly* |
| LS | List item marker | *1,One* | RBR | Adverb, comparative | *faster* |
| MD | Modal | *can,should* | RBS | Adverb, superlative | *fastest* |

# Penn Tagset cntd.

| | |
|---|---|
| VB | Verb, base form subsumes imperatives, infinitives and subjunctives |
| VBD | Verb, past tense includes the conditional form of the verb to be |
| VBG | Verb, gerund or persent participle |
| VBN | Verb, past participle |
| VBP | Verb, non-3rd person singular present |
| VBZ | Verb, 3rd person singular present |
| TO | *to* |

Language Phenomena

To

1. *I want to dance*
2. *I went to dance*
3. *I went to dance parties*

NNS & VBZ

1. Most English nouns can
2. act as verbs
3. Noun plurals have the
4. Same for as 3pS verbs

Christopher D. Manning. 2011. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011, Proceedings, Part I*. Lecture Notes in Computer Science 6608, pp. 171--189.

# Indian Language Tag set: Noun

| Sl. No | Category | | | Label | Annotation Convention** | Examples |
|---|---|---|---|---|---|---|
| | Top level | Subtype (level 1) | Subtype (level 2) | | | |
| **1** | **Noun** | | | **N** | **N** | ladakaa, raajaa, kitaaba |
| 1.1 | | Common | | NN | N__NN | kitaaba, kalama, cashmaa |
| 1.2 | | Proper | | NNP | N__NNP | Mohan, ravi, rashmi |
| 1.4 | | Nloc | | NST | N__NST | Uupara, niice, aage, |

# Scale of effort involved in annotation (1/2)

- Penn Treebank
  - Total effort: *8 million words, 20-25 man years (5 persons for 4-5 years)*
- Ontonotes: Annotate 300K words per year (*1 person per year*)
  - news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows,
  - with structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference)
  - in English, Chinese, and Arabic
- Prague Discourse Treebank (Czeck): 500,000 words, 20-25 man years (*4-5 persons for 5 years*)

# Scale of effort in annotation (2/2)

**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)
  - 6 man years for each <L,D> combination (3 persons for 2 years)

# Serious world wide effort on leveraging multiliguality

- Greg Durrett, Adam Pauls, and Dan Klein, *Syntactic Transfer Using Bilingual Lexicon*, EMNLP-CoNLL, 2012
- Balamurali A.R., Aditya Joshi and Pushpak Bhattacharyya, *Cross-Lingual Sentiment Analysis for Indian Languages using Wordent Synsets*, COLING 2012
- Dipanjan Das and Slav Petrov, *Unsupervised Part of Speech Tagging with Bilingual Graph-Based Projections,* ACL, 2011
- Benjamin Snyder, Tahira Naseem, and Regina Barzilay, *Unsupervised multilingual grammar induction*, ACL-IJCNLP, 2009

# Cooperative Word Sense Disambiguation

# Definition: WSD

- Given a context:
  - Get “meaning”s of
    - *a set of words (targetted wsd)*
    - or all words (*all words wsd*)
- The “Meaning” is usually given by the id of senses in a sense repository
  - usually the wordnet

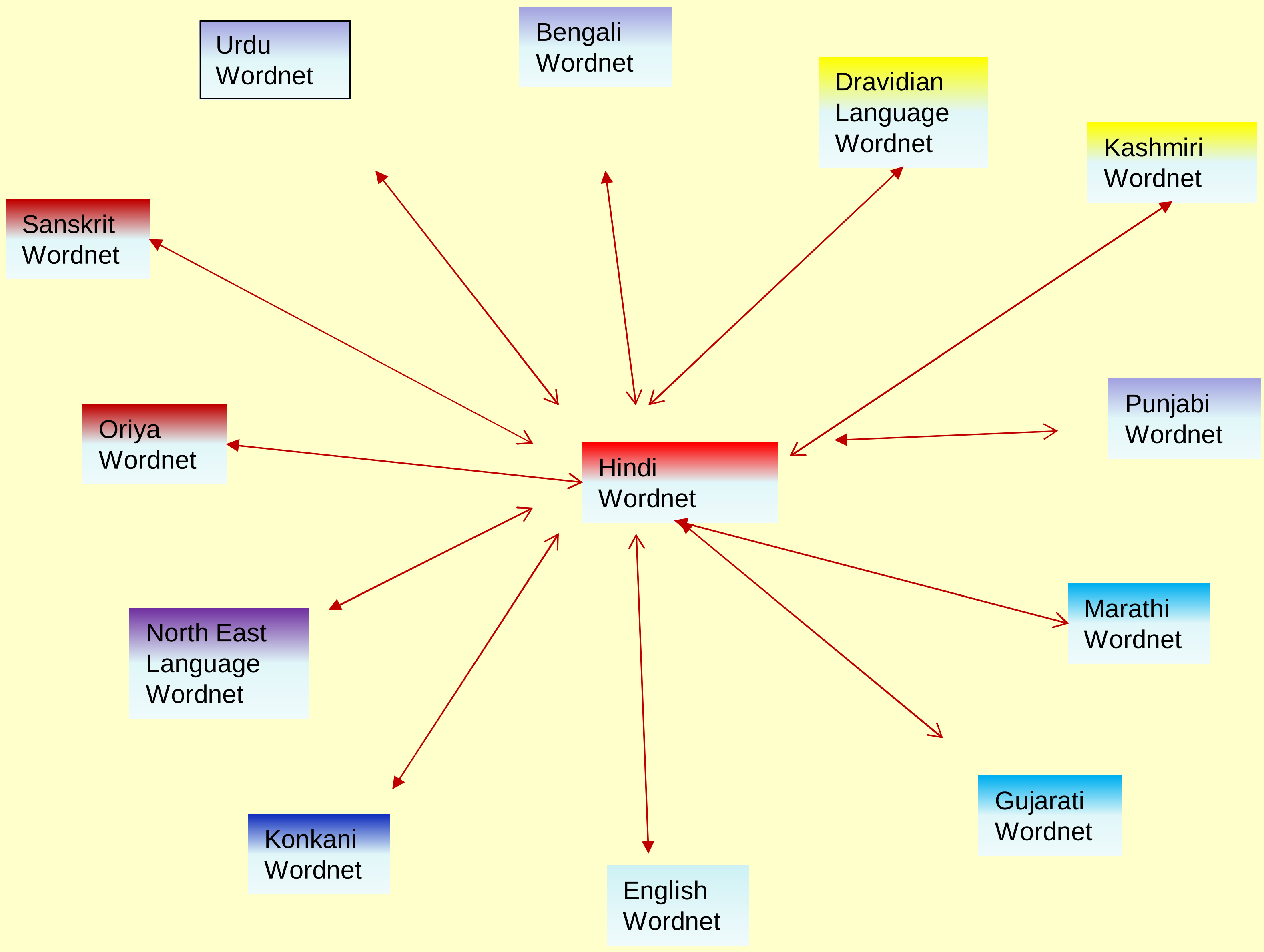
# Example: “*operation*” (from Princeton Wordnet)

- **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

- 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

WSD for ALL Indian languages:
Critical resource: INDOWORDNET

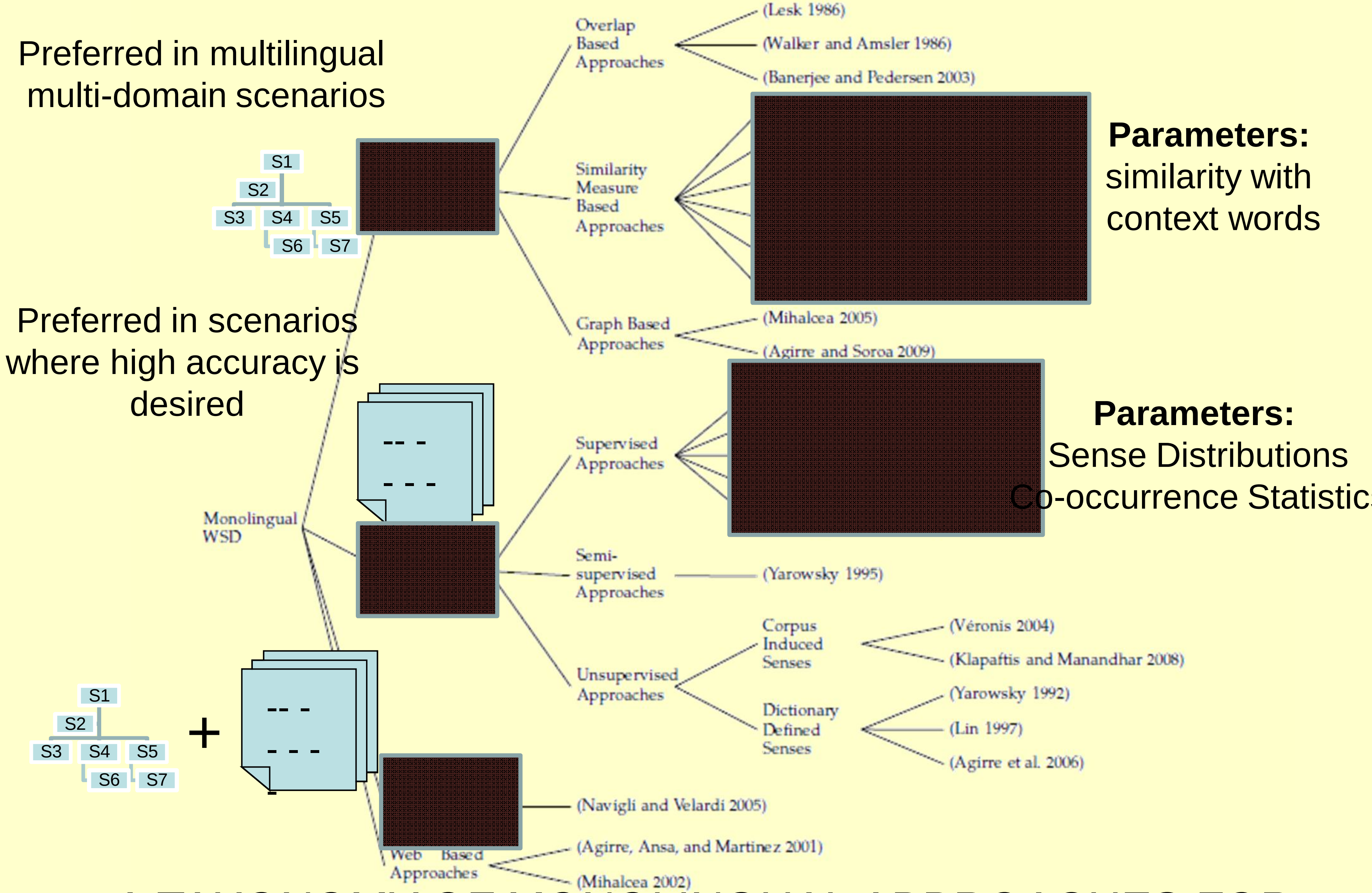


# Language-Domain GRID

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

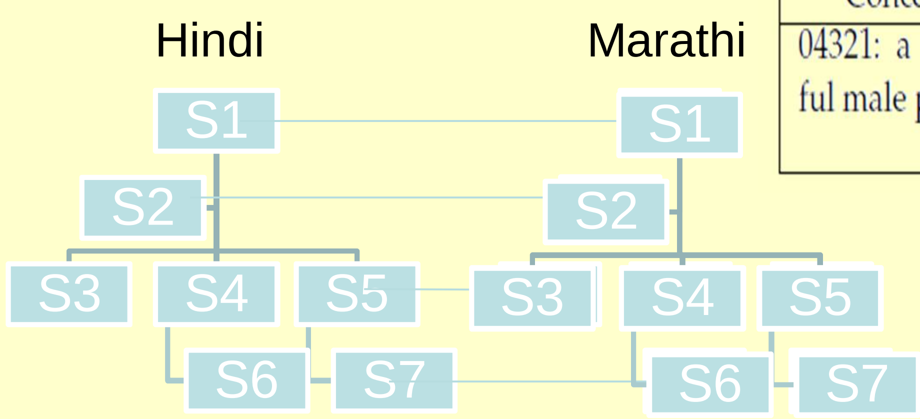
A grid of languages v/s domains. Each cell represents a language-domain pair. The Xs indicate the cells for which data is available.

***Ideal Goal:*** *Given sufficient resources for one cell in the grid we should be able to cater to all the cells in the grid*




# *A TAXONOMY OF MONOLINGUAL APPROACHES FOR WSD*

# Synset Based Multilingual Dictionary

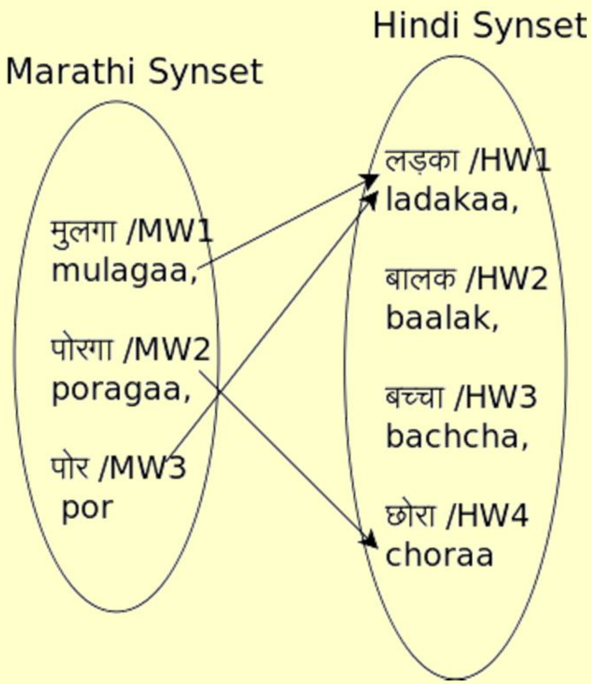



| Concepts | L1 (English) | L2 (Hindi) | L3 (Marathi) |
|---|---|---|---|
| 04321: a youthful male person | {malechild, boy} | {लड़का *(ladkaa)*, बालक *(baalak)*, बच्चा *(bachchaa)*} | {मुलगा *(mulgaa)*, पोरगा *(porgaa)*, पोर *(por)*} |

**A sample entry from the *MultiDict***

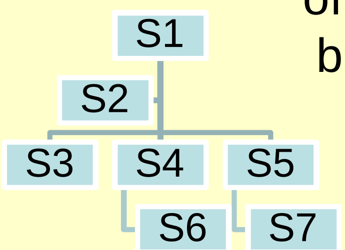
- Expansion approach for creating wordnets [Mohanty et. al., 2008]
- Instead of creating from scratch link to the synsets of existing wordnet
- Relations get borrowed from existing wordnet

# Cross Linkages Between Synset Members




- Captures native speakers intuition
- Wherever the word *ladkaa* appears in Hindi one would expect to see the word *mulgaa* in Marathi
- For this work we do not use these manual cross linkages as they have a cost associated with them
- Instead we assume that every word in the Hindi synset is a translation of a word in the corresponding Marathi synset

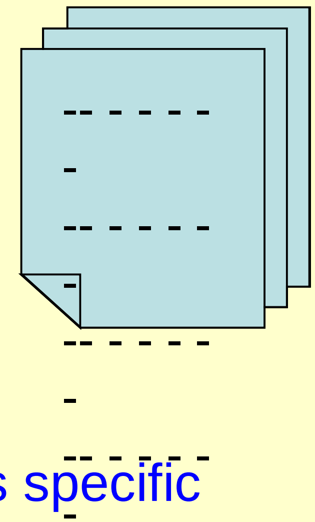
# Summary: two critical Resources Needed For WSD




**Wordnet:** A repository of senses and relations between senses

- Senses serve as class labels
- Similarity metrics defined on wordnet relations can contribute to a scoring function for ranking senses (sea::river)
- Sole guiding factor for Knowledge based approaches

**Annotated Corpus:** Words are manually Annotated with their context-appropriate sense



- Capture corpus specific behavior
- **Sense distributions**
- Co-occurrence statistics

# *Balancing Resources – 5 scenarios*

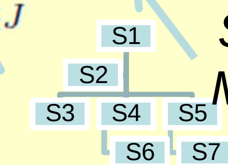
| | ***Annotated Corpus in L1*** | ***Aligned Wordnets*** | | ***Annotated Corpus in L2*** |
|---|---|---|---|---|
| **Scenario 1** | ✓ | ✓ | | ✘ |
| **Scenario 2** | ✓ | ✓ | | ✘ |
| **Scenario 3** | ✓ | ✓ | | ***Varies*** |
| **Scenario 4** | ✘ | ✓ | | ✘ |
| **Scenario 5** | ***Seed*** | ✓ | | ***Seed*** |

# *Iterative Word Sense Disambiguation*

$$S^* = \arg\max_i(\lambda * \theta \text{ [redacted] } + (1-\lambda) * \sum_{j \in J} \text{ [redacted] } V_i * V_j)$$

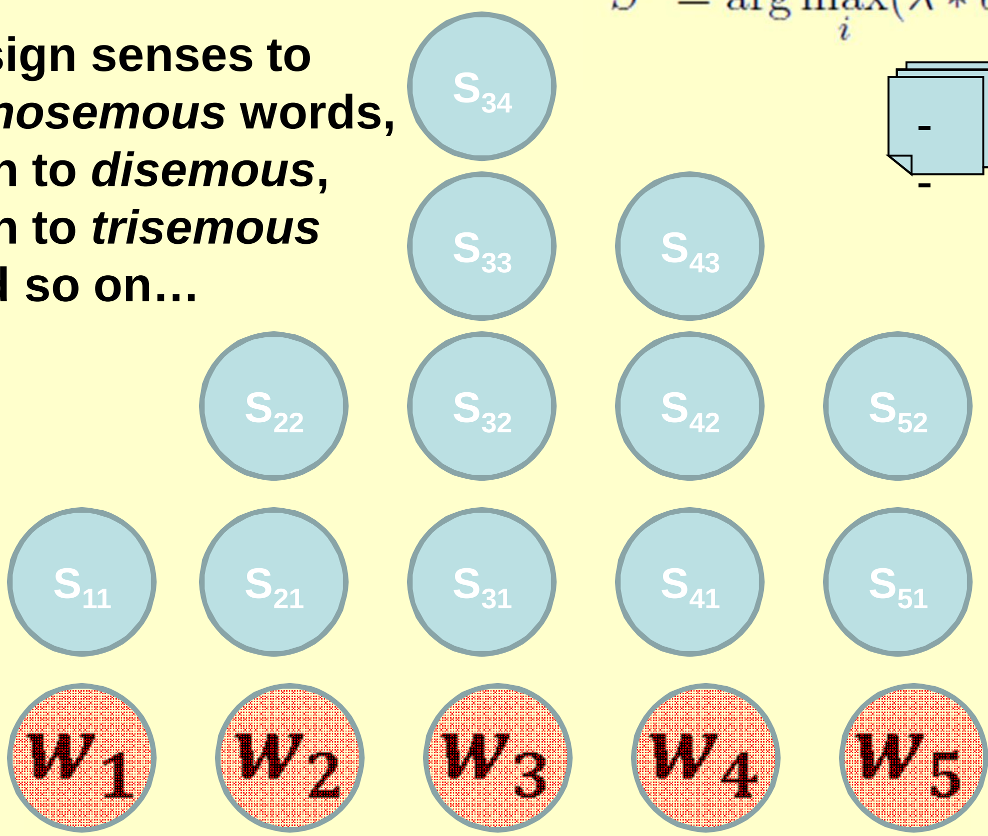
*Sense Distributions*

*Similarity Measures*




*Previously Disambiguated Words*

**Assign senses to *Monosemous* words, then to *disemous*, then to *trisemous* And so on…**




**Iterative *Disambiguation* (IWSD)**

Mitesh Khapra, Sapan Shah, Piyush Kedia and Pushpak Bhattacharyya,
•*Projecting Parameters for Multilingual Word Sense Disambiguation*, (**EMNLP09**)

# Which parameters are important for WSD

- **Sense distributions are the most important parameters for WSD**

- Other parameters do not contribute much

# Unsupervised WSD

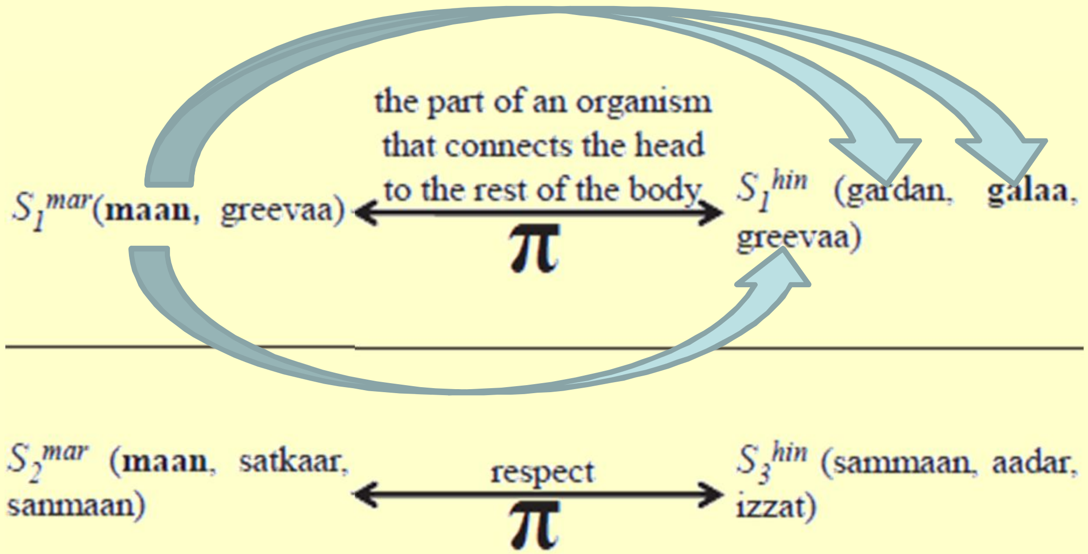
## *(No annotation!)*

Khapra, Joshi and Bhattacharyya, IJCNLP 2011

# Hypothesis

- Sense distributions across languages is invariant!!
  - Number of times a sense appears in a language is uniform across languages!
  - E.g., number of times the sense of “sun” appears in any language through “sun” and its synonyms remains the same!

# ESTIMATING SENSE DISTRIBUTIONS

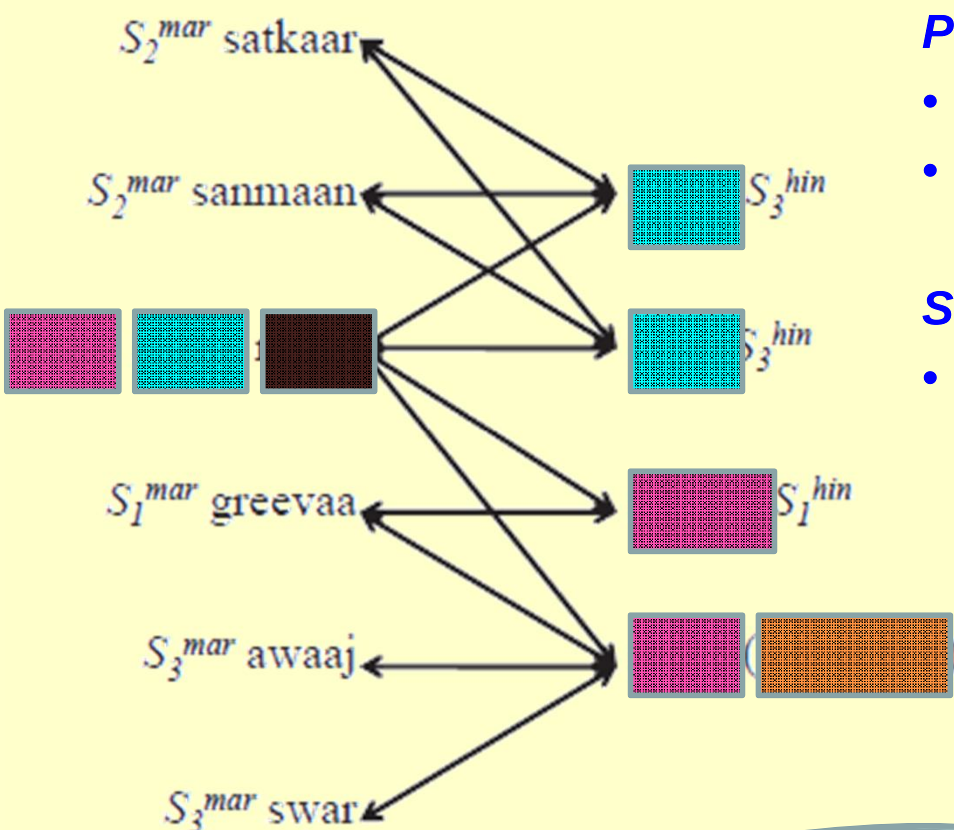



If sense tagged Marathi corpus were available, we could have estimated

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

But such a corpus is not available

# EM for estimating sense distributions




***Problem:***

- *galaa itself is ambiguous*
- *Its raw count cannot be used as it is*

***Solution:***

- *Its count should be weighted by* $P(S_1^{hin}|gala)$

***E-Step***

$$P(S_1^{mar}|maan) = \frac{\ldots\#(gardan) + \ldots\#(gala)}{\ldots\#(gardan) + \ldots\#(gala) + \ldots\#(aadar) + \ldots\#(izzat)}$$

***M-Step***

$$P(S_1^{hin}|gala) = \frac{P(S_1^{mar}|maan)\cdot\#(maan) + P(S_1^{mar}|greeva)\cdot\#(greeva)}{P(S_1^{mar}|maan)\cdot\#(maan) + P(S_1^{mar}|greeva)\cdot\#(greeva) + P(S_3^{mar}|awaaj)\cdot\#(awaaj) + P(S_3^{mar}|swar)\cdot\#(swar)}$$

# Results & Discussions

| Algorithms | Tourism | | | Health | | |
|---|---|---|---|---|---|---|
| | P% | R% | F% | P% | R% | F% |

[redacted]

Our values

Manual Cross Linkages

Probabilistic Cross Linkages

Skyline - self training data is available

Wordnet first sense baseline

S-O-T-A Knowledge Based Approach

S-O-T-A Unsupervised Approach

- Performance of projection using manual cross linkages is within 7% of Self-Training
- Performance of projection using probabilistic cross linkages is within 10-12% of Self-Training – remarkable since no additional cost incurred in target language
- Both MCL and PCL give 10-14% improvement over Wordnet First Sense Baseline
- *Not prudent to stick to knowledge based and unsupervised approaches – they come nowhere close to MCL or PCL*

# Adding context to the EM based approach

Bhingardive, Shaikh and Bhattacharyya, ACL 2013.

# Context as a bag of words

- We treat the context as a bag of words
- We assume that every context word independently affects the sense of the target word.
- Hence,

$$P(S \mid w, C) = \prod_{c_i \in C} P(S \mid w, c_i)$$

*where*,

$S$ is one of the candidate synsets of $w$,

$C$ is the sentential context,

$c_i$ is a word belonging to C.

# Adding context

## Basic EM 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

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

# The Formulation

- **The E-Step:**

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

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

$a \in context(u)$

$b \in crosslinks_{L_2}(a)$

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

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

- **The M-Step:**

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

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

$b \in context(v)$

$a \in crosslinks_{L_1}(a)$

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

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

# Exact co occurences: rare to find

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

# Add semantic relatedness

## Instead of:

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

## Use:

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

*where,*

$\sigma(v,b)$ represents the semantic relatedness between the senses through which \`$u$' and \`$a$' were translated to \`$v$' and \`$b$' respectively.

# Semantic Relatedness

Inverse distance relatedness is used. It is one of the simplest path based measures.

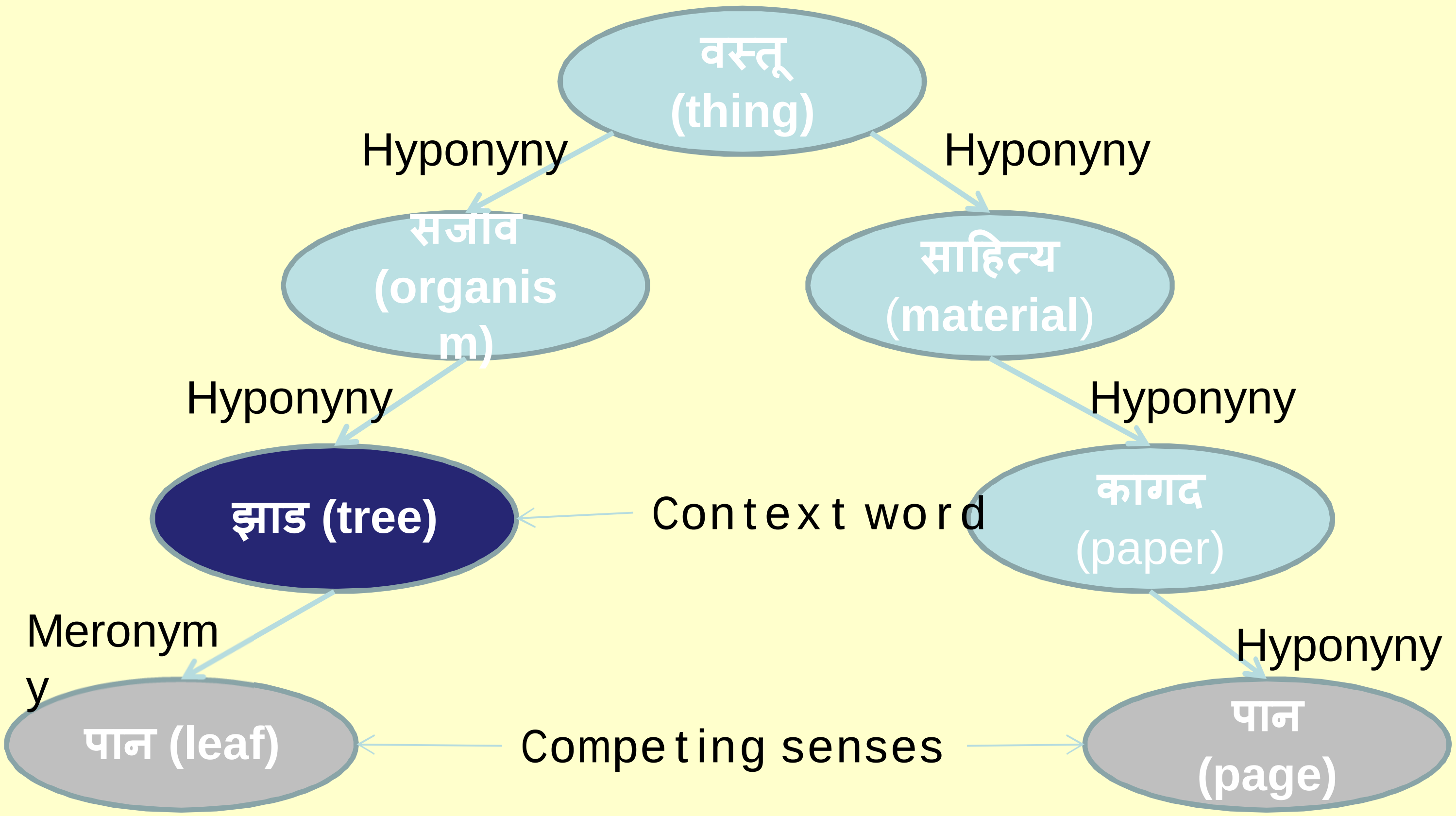
$$S.R. = \frac{1}{1 + d(c_1, c_2)}$$

*where,*

$d(c_1, c_2)$ is the shortest distance

between $c_1$ and $c_2$ in wordnet.

# Semantic Relatedness contd…




Distance =**1**, S.R.= 1/(1+1)=**0.5**

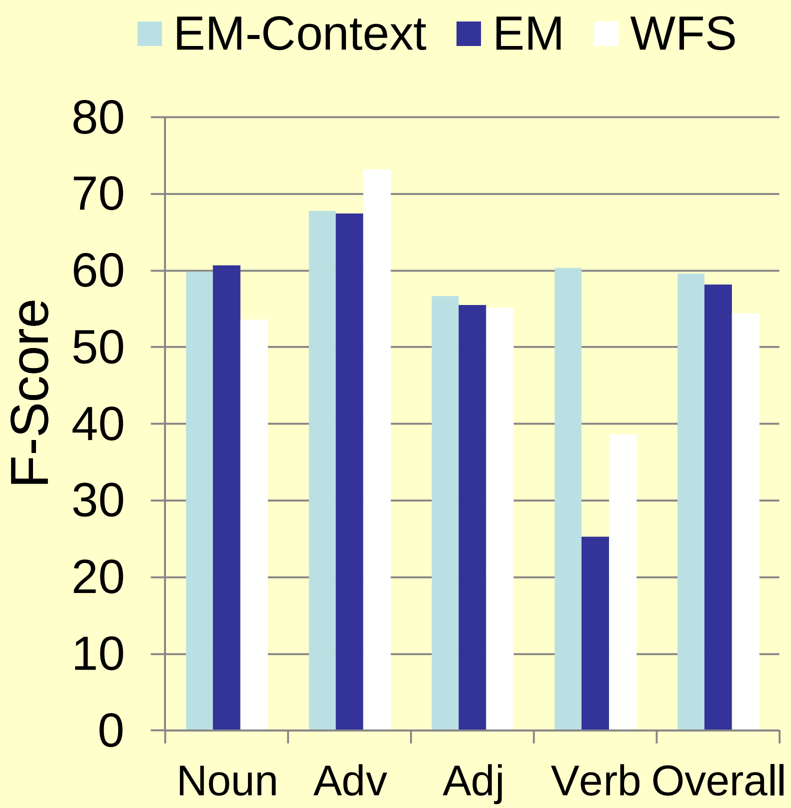
Distance =**5**, S.R.= 1/(1+5)=**0.16**

# RESULTS

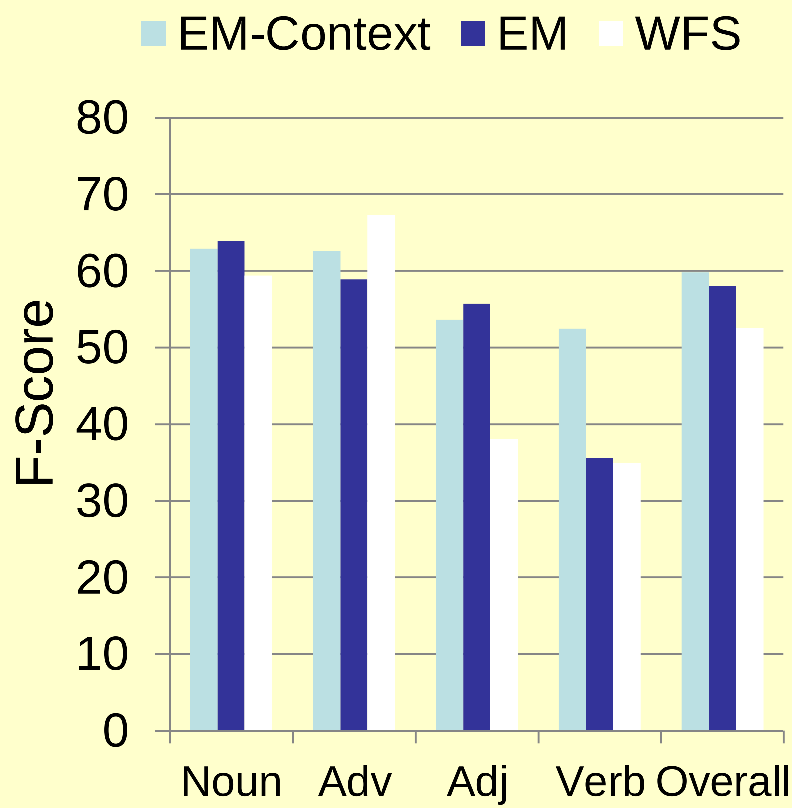
## EM-Context vs EM

# Results

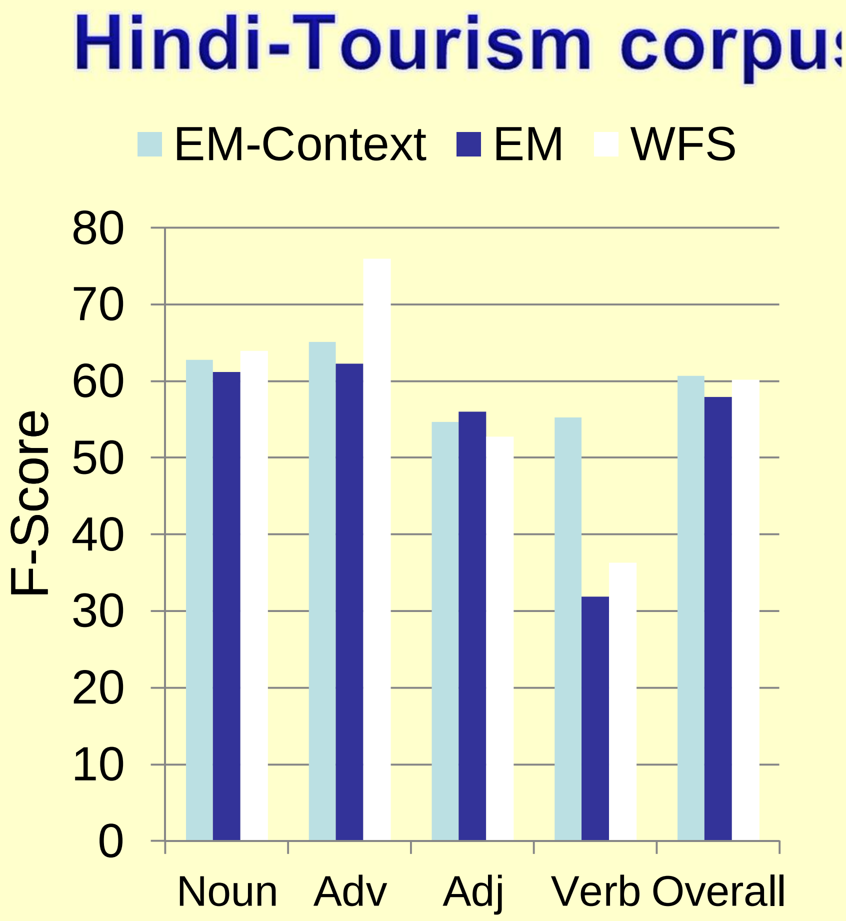
## Hindi-Health corpus

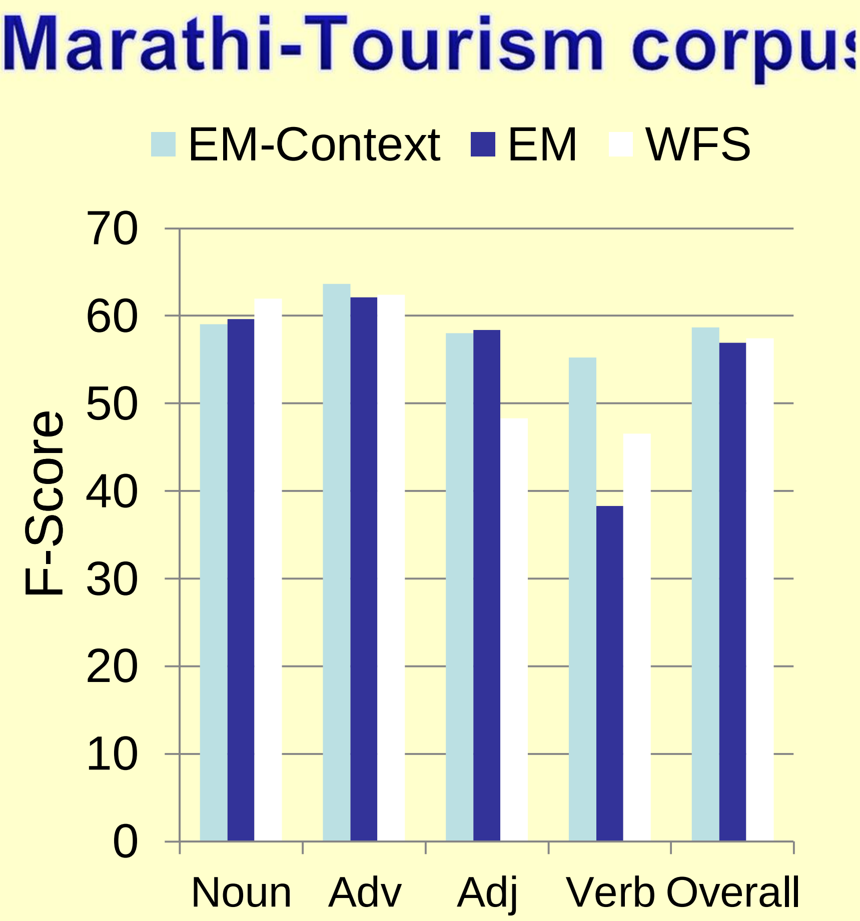



## Marathi-Health corpus



# Results contd…



# Error analysis

## Context as a bag of words

They were playing cards

Vaha patte khel rahe the.

वह पत्ते खेल रहे थे ।

Endorses the ‘cards’ sense

**Strongly related context words**

Endorses the ‘leaf’ sense

Endorses the ‘cards’ sense

वह पेड़ के नीचे पत्ते खेल रहे थे ।

Vaha ped ke neeche patte khel rahe the.

They were playing cards below the tree.

**Semantic structure of the sentence can help in such situations**

# Semantic roles (UNL representation)

<sentence>
They play cards under the tree.
</sentence>
<iitb>
agt (*play*(icl>act, equ>play):2.@present.@entry,
*They*(icl>pronoun):1 )
obj (*play*(icl>act, equ>play):2.@present.@entry,
*card*(icl>game>thing):3.@pl )
plc (*play*(icl>act, equ>play):2.@present.@entry,
*tree*(icl>woody_plant>thing):6.@def.@under )
</iitb>}

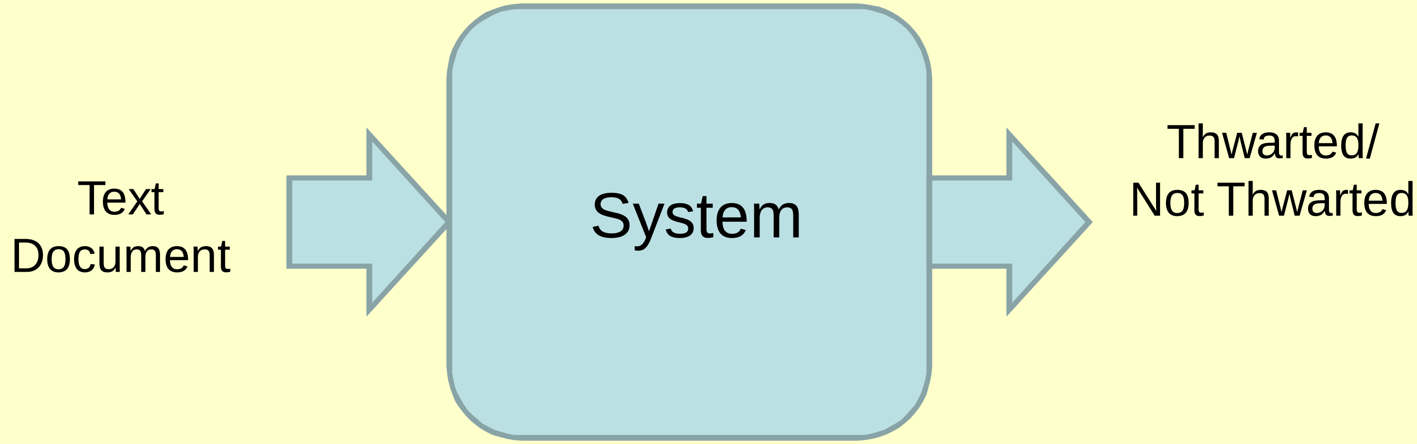
back

# Detecting Turnarounds in Sentiment Analysis: Thwarting

*Ramteke, Malu, Bhattacharyya, Nath, ACL 2013*

# Problem definition

- To detect Thwarting in text




Thwarted

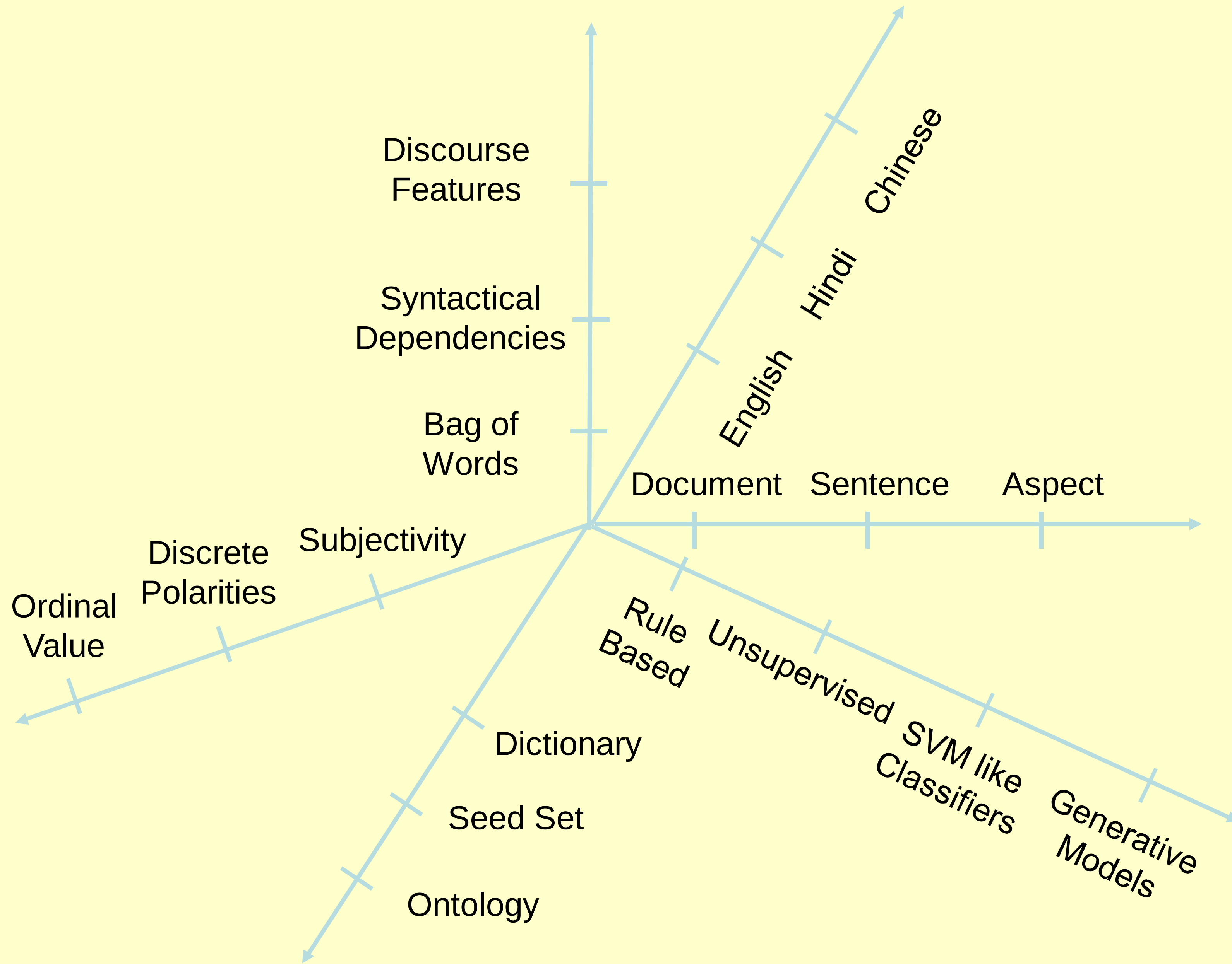
***The actors performed well. The music was enthralling. The direction was good. But, I still did not like the movie.***

Not Thwarted

***This camera has everything that you need. A Superb lens, an amazing picture quality and a long battery life. I love it.***

# Definitions

- **Sentiment Analysis**: The task of identifying if a certain piece of text contains any opinion, emotion or other forms of affective content.
- **Sentiment Polarity:** The sentiment exhibited by the document, sentence or word. It can be positive, negative or an ordinal value between the two.
- **Thwarting:** The scenario where a minority of a document's content determines its polarity.



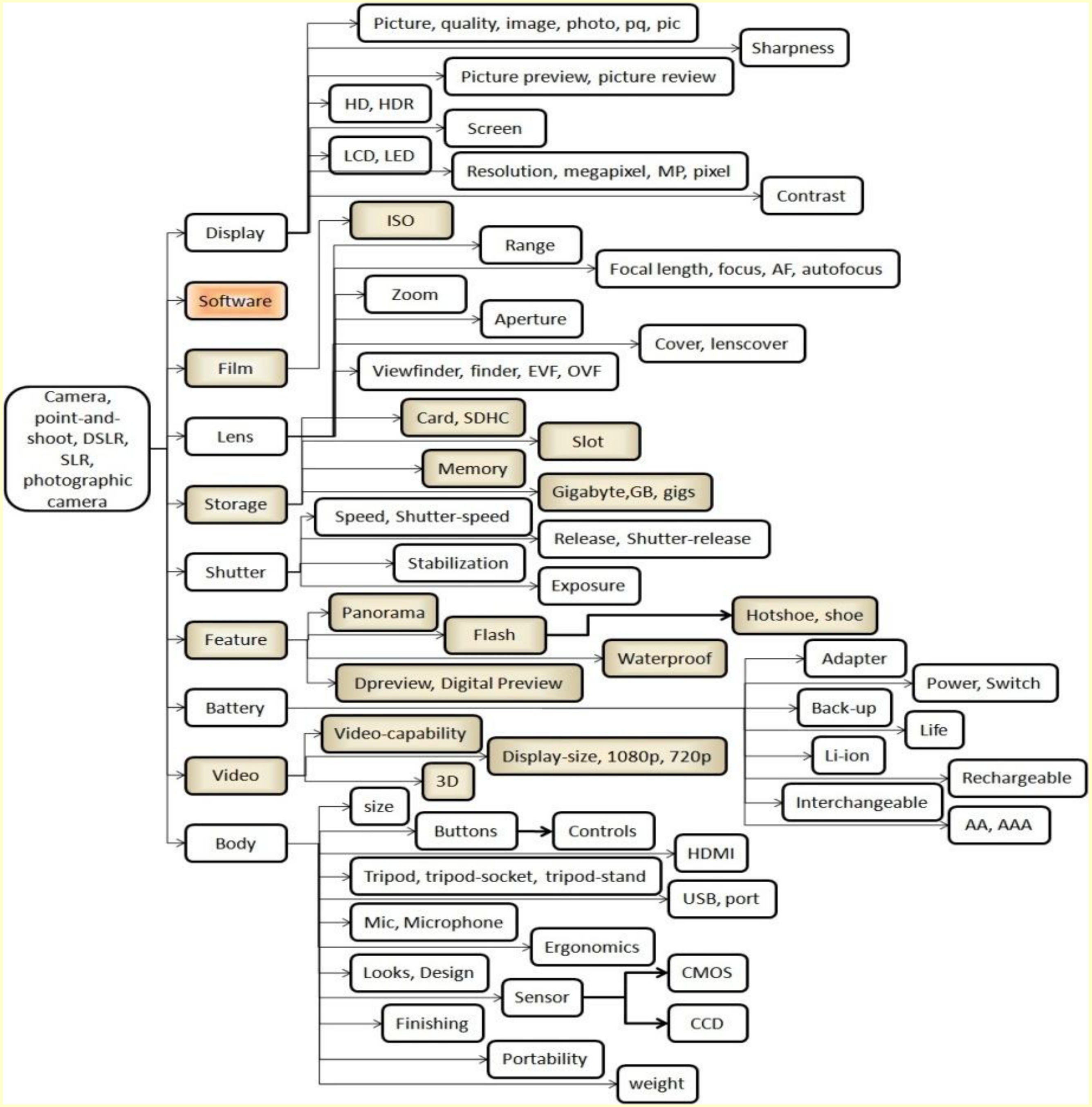

Dimensions of Sentiment Analysis

# Handling Data Skew

- Thwarting is a rare phenomenon and thus faces data skew
- Approaches to handling data skew in other tasks
  - Tao *et al.* (2006)
  - Hido *et al.* (2008)
  - Provost *et al.* (1999)
  - Viola *et al.* (2001)

# Domain Ontology

- Need for a weighting of entities related to a domain
- **Domain Ontology**: Aspects (entity parts) arranged in the form of a hierarchy
- An ontology naturally gives such weighting
  - Each level has a weight




**Camera Ontology**

# Basic idea

*From the perspective of the domain ontology, the sentiment towards the overall product or towards some critical feature mentioned near the root of the ontology should be opposite to the sentiment towards features near the leaves.*

# An Ex

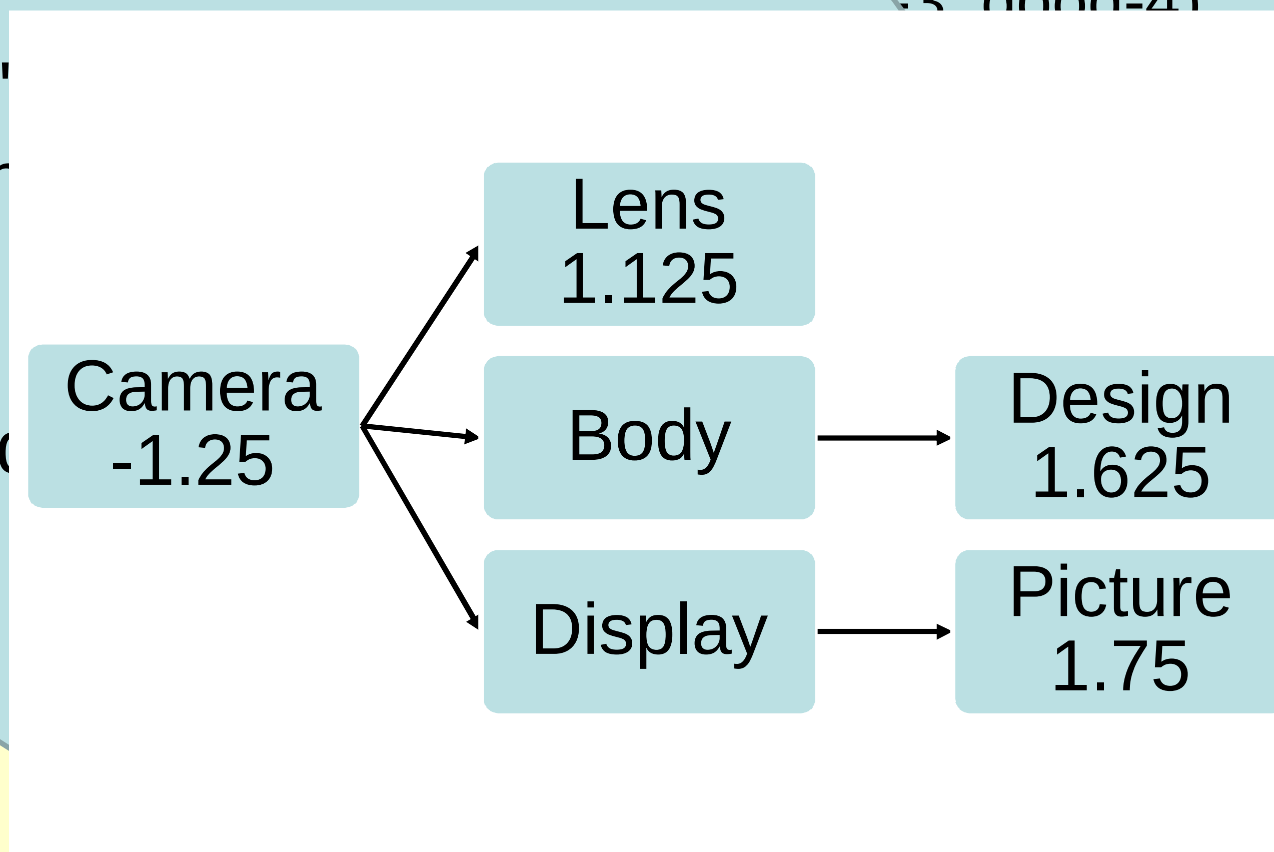
dobj(love-2, design-5)

mpressive-4, lens-2)

-3, pictures-2)

design

Th



# Results

| Level Weights | Precision | Recall | F1 Score |
|---|---|---|---|
| (4,3,2,1) | 0.01179 | 0.3125 | 0.02272 |
| (8,4,2,1) | 0.01182 | 0.3125 | 0.02277 |
| (20,15,10,5) | 0.01179 | 0.3125 | 0.02272 |
| (10,8,6,4) | 0.01179 | 0.3125 | 0.02272 |

The **Best AUC** for the experiments was found out to be **56.3%**

A Random Classifier is expected to have an AUC of **50%**

# Observations

- Need more principled approach to find weights
- Different Weight for nodes on the same level
  - Body and Video Capability
    - Individual tastes, not so critical
  - Lens or the Battery
    - More critical feature
- Learn Weights from corpus

# ML Approach to Tackle Thwarting

# Step 1: Extracting Weights

- Let the polarities of domain aspects in a review be represented by $A_1, A_2 \ \dots \ A_N$.
- Let the weights corresponding to each of these domain aspects be represented by $W_1, W_2 \ \dots \ W_N$.
- Let the overall polarity of the document be *P*.
- $P = \sum_i A_i * W_i$
- Also Minimize Hinge loss max(0,1 – $P.W^T.A$)

# Modifications

- Intuition: Lower level nodes influence higher level node polarities
  - Percolate polarity of child to parent
- Three types of Percolation
  - No percolation
  - Complete Percolation
  - Controlled Percolation
- Prior Bias towards weights

# Step 2: Representing Reviews

We then extract a vector of values

$$V_1 \,, V_2 \;\; \ldots \; V_M$$

from each review.

Each $V_i$ represents a weighted aspect polarity value.

# Step 3: Extracting features

1. Document polarity
2. Number of flips of sign (i.e. from positive to negative and vice versa) normalized by the number of terms in the sequence
3. The Maximum and the Minimum values in a sequence
4. The length of the longest positive contiguous subsequence
5. The length of the longest negative contiguous subsequence
6. The mean of the values

# Step 3: Extracting Features (contd.)

6. Total number of positive values in the sequence
7. Total number of negative values in the sequence
8. The first and the last value in the sequence
9. The variance of the moving averages
10. The difference in the averages of the longest positive and longest negative contiguous subsequences

# An Example

"I love the sleek design. The lens is impressive. The pictures look good but, somehow this camera disappoints me. I do not recommend it."

313, -0.05

| | Value |
|---|---|
| | -1 |
| | 3 |
| | 31325 |
| | .05 |
| | 1 |
| | 1 |
| | 0.003940625 |
| e | 2 |
| nce | 2 |
| Thw | 0.0091 |
| | -0.05 |
| Thw verages | 0 |
| The difference in ... ages of LPCS and LNCS | 0.081325 |

# Experiments

- Setup:
  - Dataset by Malu (2012)
  - We crawled[1] an additional 1000 reviews out of which 24 reviews were Thwarted
  - Camera domain
  - 2198 reviews   60 thwarted
  - Ontology for domain specific features
  - Data is skewed so weighing of classes employed
- Inter annotator Agreement
- Classification experiments
  - 10 fold cross validation
- Ablation Test

1. Reviews crawled from www.epinions.com

# Results: Inter annotator Agreement

- Cohen's kappa : 0.7317
- Agreement of 70% for the thwarted class
- Agreement of 98% for the non-thwarted
- Identifying thwarting is difficult even for humans

# Results: Classification - 1

| | Loss Type | |
|---|---|---|
| **Percolation Type** | Linear | Hinge |
| No percolation | **68.9** | 65.6 |
| Controlled | 66.89 | 62.39 |
| Complete | 67.65 | 63.43 |

Table 5.2: Results for non negative weights with prior

| | Loss Type | |
|---|---|---|
| **Percolation Type** | Linear | Hinge |
| No percolation | **69.01** | 67.42 |
| Controlled | 65.09 | 62.16 |
| Complete | 62.77 | 60.94 |

Table 5.3: Results for non negative weights without prior

# Results: Classification - 2

| | Loss Type | |
|---|---|---|
| **Percolation Type** | Linear | Hinge |
| No percolation | 73.87 | 70.12 |
| Controlled | **81.05** | 77.17 |
| Complete | 63.85 | 60.94 |

Table 5.4: Results for unconstrained weights without prior

| | Loss Type | |
|---|---|---|
| **Percolation Type** | Linear | Hinge |
| No percolation | 73.99 | 70.56 |
| Controlled | **78.47** | 72.03 |
| Complete | 62.88 | 61.36 |

Table 5.5: Results for unconstrained weights with prior

# Results: Ablation Test

| Feature Removed | Loss in AUC |
|---|---|
| Document Polarity | 10.01% |
| Number of flips of sign | 2.13% |
| The Maximum value in a sequence | 1.24% |
| The Minimum value in a sequence | 1.0% |
| The length of the longest positive contiguous subsequence | 1.2% |
| The length of the longest negative contiguous subsequence | 0.9% |
| The mean of the values | 2.0% |
| Total number of positive values in the sequence | 1.2% |
| Total number of negative values in the sequence | 1.0% |
| The first value in the sequence | 0.5% |
| The last value in the sequence | 1.1% |
| The variance of the moving averages | 5.0% |
| The difference in the averages of LPCS and LNCS | 3.0% |

# String Kernels based Model

- Convert the sequence of weighted polarities into a string
  - 0.0091, -0.0061, 0.0313, -0.05 → p n p n
- Five classes for polarities
  - Highly negative
  - Slightly negative
  - Zero
  - Slightly positive
  - Highly positive
- Determined using mean and 2 standard deviations on both sides
- N-grams as features

# Experiments and Results

- Same Dataset
- Weights from the optimal configuration
  - Unconstrained weights, without prior and controlled percolation
- AUC of **68.42**

# Observations and insights

- **Ontology** guides a rule based approach to thwarting detection, and also provides features for SVM based learning systems
- Percolating polarities is beneficial
- The Machine Learning based system scores over the rule based system by 25 %

back

# Eye Tracking based Sense annotation for the purpose of building a sense discrimination net

Salil Joshi, Diptesh Kanojia and Pushpak Bhattacharyya,

IIT Bombay

(NAACL 2013, Atlanta, 11 June, 2013)

# Insights from our earlier work (crowd sourced WSD)*

Humans need Context for Annotation

Tagging without context is often erroneous, and also a cognitive load due to uncertainty

In supervised WSD, machines rely primarily on prior sense distribution probability

Machines seem to be able to do best with just P(S/W); context per se does not seem important

“***A Study of the Sense Annotation Process: Man v/s Machine***” published in GWC 2012

# Questions

## Human Cognition in Sense Annotation

- What are the cognitive sub-processes associated with the human sense annotation task?

## Lexicographer’s Difficulty

- Which classes of words are more difficult to disambiguate and why?

# Eye-tracking

## Fixation

- Eye pause at a certain spot
- First data point
- *Where* someone is focusing, *for how long* and possibly *why*

## Saccades

- Second data point
- Eye gaze movement from one position to another

## Scan Path

- Combination of fixations and saccades

# Techniques for eye-tracking

*Most comfortable technique to measure gaze based on infrared light*

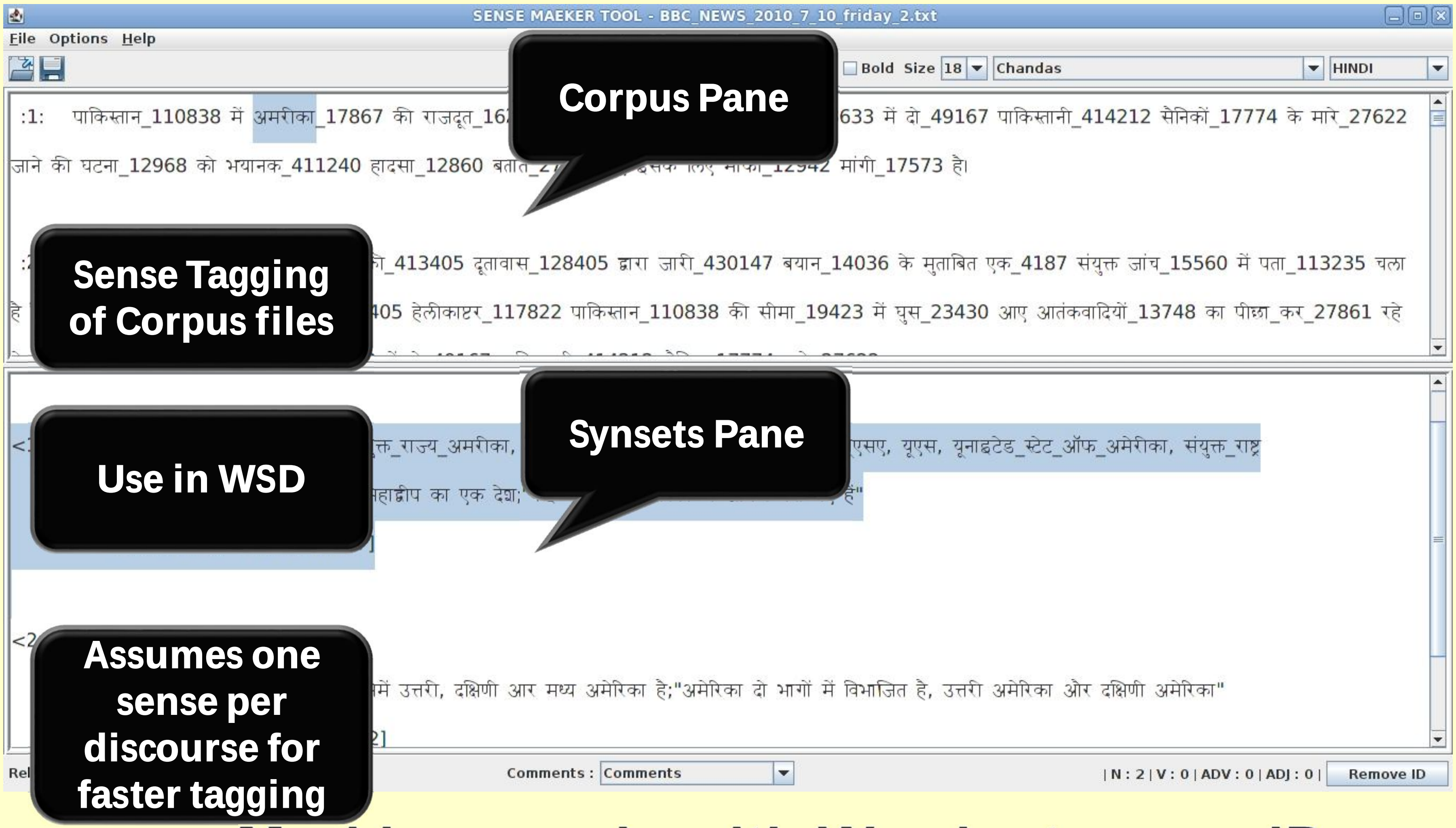
*A bit more complicated way to measure gaze using electric potential around the eye.*

*The eye tracking glasses are used for broad range of mobile eye tracking studies.*

*The ergonomic chin rest eye tracking device for high speed and accurate measurements with a large visual field.*

Image courtesy: www.smivision.com

# Sense marker tool

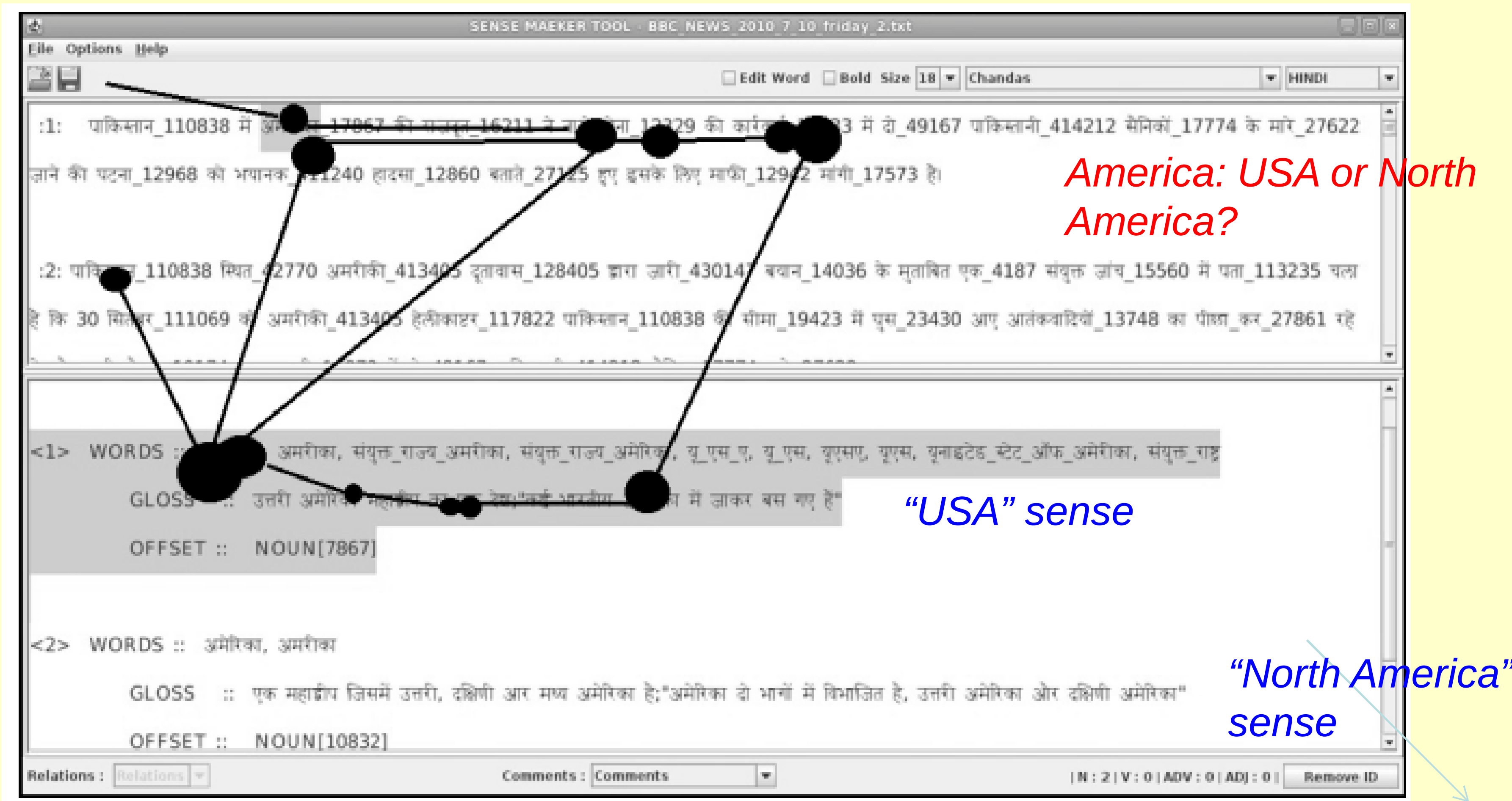



Marking words with Wordnet sense IDs

# Facts and Figures

- 2000 words used for experimentation
- Analysis done on data for open class words (nouns, verbs, adverbs and adjectives)
- Data from 6 lexicographers (3 skilled, 3 unskilled) collected
- Annotators used Sense-marker tool for tagging the word senses
- Gaze patterns analyzed

# An example of eye movement during sense marking




Note: Ball size indicative of the fixation time; lines are saccades

# Cognitive sub-processes in sense annotation*

Hypothesis Building: During annotation, the lexicographer makes initial hypotheses regarding meaning and domain of a word ($T_{hypo}$)

Clue-word Searching: Consequently he/she looks for contextual clues around the word to narrow down on 1 or at most 2 of the initial hypotheses ($T_{clue}$)

Gloss Matching: The lexicographer then scans the wordnet candidate senses of the word for synset words and gloss to map their hypothesis to one of the senses ($T_{gloss}$)

$$T_{total} = T_{hypo} + T_{clue} + T_{gloss}$$

***as discussed with the lexicographers, arguably our Most important contributionn**

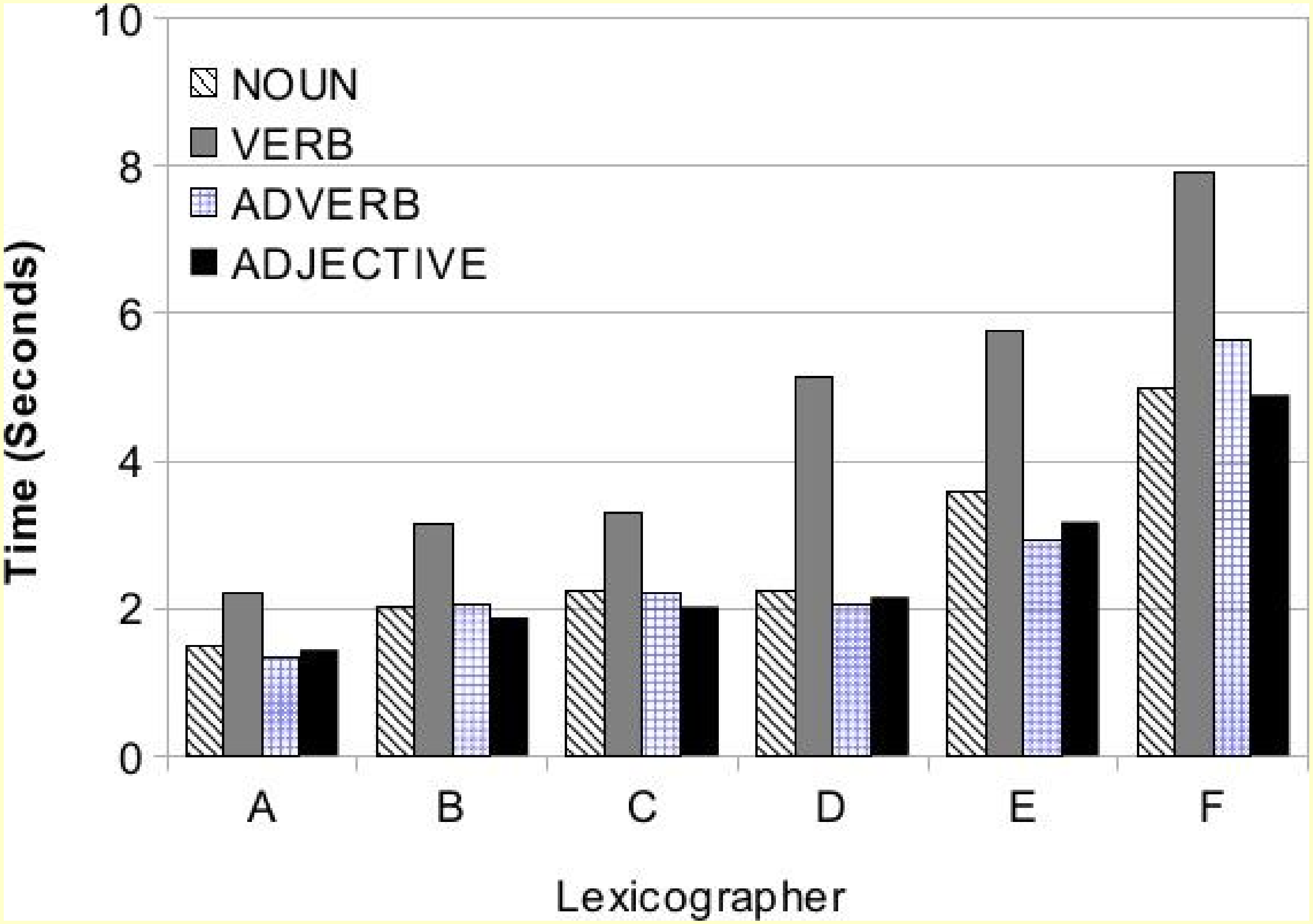
| Lexicographer | Time Taken (seconds) | | | |
|---|---|---|---|---|
| | $T_{hypo}$ | $T_{clue}$ | $T_{gloss}$ | $T_{total}$ |
| Skilled | 0.33 | 0.74 | 1.16 | 2.24 |
| Unskilled | 0.74 | 1.56 | 4.44 | 6.75 |

Time variations between skilled and unskilled lexicographers

| Word | Degree of polysemy | Unskilled lexicographers (seconds) | | | | Skilled lexicographers !(seconds) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $T_{hypo}$ | $T_{clue}$ | $T_{gloss}$ | $T_{total}$ | $T_{hypo}$ | $T_{clue}$ | $T_{gloss}$ | $T_{total}$ |
| लाना (laana – to bring) | 4 | 0.63 | 0.8 | **5.2** | 6.63 | 0.31 | 1.2 | **1.82** | 3.3 |
| करना (karanaa – to do) | 22 | 0.9 | 1.42 | 2.2 | 4.53 | 0.5 | 0.64 | 1.14 | 2.24 |
| जताना (jataanaa – to express) | 4 | 0.7 | 2.45 | **5.93** | 9.09 | 0.25 | 0.39 | **0.62** | 1.19 |

Time taken for verbs by lexicographers (examples)

# Results : time taken for different POS categories




Time taken for different POS categories for skilled (A-C) and unskilled (D-F) lexicographers

# Ontological statistics (verbs)

| Ontology | Average of Time Taken | No. of words |
|---|---|---|
| घटनासूचक (Event) | 1.870816444 | 11 |
| अनैच्छिक क्रिया (Verbs of Non-volition) | 2.59201 | 1 |
| अवस्थासूचक क्रिया(Verb of State) | 4.403871355 | 77 |
| शारीरिक कार्यसूचक bodily action | 4.97281795 | 40 |
| कर्मसूचक क्रिया (Verb of Action) | 5.376058091 | 11 |
| प्रेरणार्थक क्रिया (causative verb) | 5.635743 | 5 |
| संप्रेषणसूचक (Communication) | 5.895843818 | 11 |
| अधिकारसूचक (Possession) | 6.00231725 | 9 |
| परिवर्तनसूचक (Change) | 6.517663706 | 17 |
| विनाशसूचक (Destruction) | 8.7992645 | 3 |
| होना क्रिया (Verb of Occur) | 12.06406657 | 7 |
| भौतिक अवस्थासूचक (Physical State) | 13.4773335 | 2 |
| निरंतरतासूचक क्रिया (Verbs of Continuity) | 17.896006 | 2 |
| कार्यसूचक (Act) | 20.2321495 | 2 |
| मानसिक अवस्थासूचक (Mental State) | 74.698983 | 1 |
| **Grand Total** | **5.896812948** | **199** |

# Discussions

## Cognitive sub-processes for Sense Annotation

- Three stages: Hypothesis building, clue-word searching and gloss matching

## Skilled v/s unskilled lexicographers

- Unskilled $T_{gloss} >> T_{clue}$
- Skilled $T_{gloss} \sim T_{clue, ;}$ latch on to the POS quickly

## Maximum annotation time for verbs

- High degree of polysemy
- Senses are fine-grained
- In some cases the hypothesis does not match the candidate senses

## Adverbs and Adjectives

- Annotation time comparable to nouns
- Adjective and adverbs' proximity to the noun helps

# Observations

- ✓Sense annotation process can be divided into 3 stages: Hypothesis building ($T_{hypo}$), Clue-word searching ($T_{clue}$) and gloss matching ($T_{gloss}$)
- ✓The theory can be verified by analyzing the gaze patterns
- ✓Skilled lexicographers annotate the words faster
    - ✓have knowledge about the senses of a word (significantly reducing the time $T_{gloss}$)
- ✓Verbs take the highest time among the POS categories given the high degree of polysemy and lack of exact senses
- ✓Adverbs and adjectives are easier to annotate given their position near a verb or a noun
- ✓Automating the process of identifying the clue-words from the gaze patterns can lead to building a rich *discrimination-net*

back

# Multiword Expressions

About half the lexical items in most languages
are multiwords!

# Multi-Word Expressions (MWE)

- Necessary Condition
  - Word sequence separated by space/delimiter
- Sufficient Conditions
  - Non-compositionality of meaning
  - Fixity of expression
    - In lexical items
    - In structure and order

# Examples – Necessary condition

- Non-MWE example:
  – Marathi: सरकार हक्काबक्का झाले
  – Roman: sarakAra HakkAbakkA JZAle
  – Meaning: government was surprised
- MWE example:
  – Hindi: गरीब नवाज़
  – Roman: garIba navAjZa
  – Meaning: who nourishes poor

# Examples - Sufficient conditions ( Non-compositionality of meaning)

- Konkani: पोटांत चाबता
- Roman: poTAMta cAbatA
- Meaning: to feel jealous

- Telugu: చెట్టు కింది కి ప్లీడరు
- Roman: ceVttu kiMXa pLIdaru
- Meaning: an idle person

- Bangla: মাটির মানুষ
- Roman: mAtira mAnuSa
- Meaning: a simple person/son of the soil

# Examples – Sufficient conditions (Fixity of expression)

## In lexical items

- Hindi
  - usane muJe gAlI dI
  - *usane muJe galI pradAna kI
- Bangla
  - jabajjIbana karadaMda
  - *jIbanabhara karadaMda
  - *jabajjIbana jela
- English (1)
  - life imprisonment
  - *lifelong imprisonment
- English (2)
  - Many thanks
  - *Plenty of thanks

# Examples – Sufficient conditions (In structure and order)

- English example
  - kicked the bucket (died)
  - the bucket was kicked

    (not passivizable in the sense of dying)
- Hindi example
  - उम्र क़ैद
  - umra kEda (life imprisonment)
  - umra bhara kEda

# MW task (NLP + ML)

NLP

ML

| | *String + Morph* | *POS* | *POS+ WN* | *POS + List* | *Chun k-ing* | *Parsing* |
|---|---|---|---|---|---|---|
| ***Rules*** | Onomaetopic Redupli-cation<br><br>*(tik tik, chham chham)* | Non-Onomaetopic Redupli-cation<br><br>*(ghar ghar)* | Non-redup (Syn, Anto, Hypo)<br><br>*(raat din, dhan doulat)* | | | Non-contiguous something |
| ***Statistical*** | | Collloctions or fixed expressions<br><br>*(many thanks)* | | Conjunct verb (verbalizer list), Compund verb (verctor verb list) *(salaha dena, has uthama)* | | Non-contiguous Complex Predicate |

***Idioms will be list morph + look up***

# Summary

- Co-operative WSD
  - Good linguistics (high quality linked wordnets) + Good ML (novel EM formulation)
- Thwarting (difficult sentiment analysis problem)
  - Good NLP (ontology) + good ML (string kernels?)
- Crowd sourced discrimination net for WSD (sense clues)
  - Cognition study through eye tracking, leading to very useful resource for ML

# Conclusions

- Both Linguistics and Computation needed: **Linguistics is the eye, Computation the body**
- It is possible to leverage the resources created for one language in another
- Language phenomenon → Formalization → Hypothesis formation → Experimentation → Interpretation (Natural Science like flavor)
- Theory=NLP, Technique=ML

# URLS

(publications) http://www.cse.iitb.ac.in/~pb

(resources) http://www.cfilt.iitb.ac.in

# Thank you

Questions?

# Challenge of POS tagging

*Example from Indian Language*

# Tagging of *jo*, *vaha*, *kaun* and their inflected forms in Hindi and their equivalents in multiple languages

# DEM and PRON labels

- ***Jo_DEM*** *ladakaa kal aayaa thaa, vaha cricket acchhaa khel letaa hai*
- ***Jo_PRON*** *kal aayaa thaa, vaha cricket acchhaa khel letaa hai*

# Disambiguation rule-1

- ***If***
  - ***Jo is followed by noun***
- ***Then***
  - ***DEM***
- ***Else***
  - ***…***

# False Negative

- When there is arbitrary amount of text between the *jo* and the noun
- *Jo_???* ***bhaagtaa huaa, haftaa huaa, rotaa huaa, chennai academy a koching lenevaalaa*** *ladakaa kal aayaa thaa, vaha cricket acchhaa khel letaa hai*

# False Positive

- *Jo_DEM* (wrong!) ***duniyadarii samajhkar chaltaa hai, …***
- *Jo_DEM/PRON? manushya manushyoM ke biich ristoM naatoM ko samajhkar chaltaa hai, …* (ambiguous)

# False Positive for Bengali

- *Je_DEM* (wrong!) ***bhaalobaasaa paay,*** *sei bhaalobaasaa dite paare*

  *(one who gets love can give love)*
- *Je_DEM* (right!) *bhaalobaasa tumi kalpanaa korchho, taa e jagat e sambhab nay*

  (*the love that you imagine exits, is impossible in this world)*

# Will fail

- In the similar situation for
  - *Jis, jin, vaha, us, un*
- All these forms add to corpus count

# Disambiguation rule-2

- ***If***
  - ***Jo is oblique (attached with ne, ko, se etc. attached)***
- ***Then***
  - ***It is PRON***
- ***Else***
  - ***<other tests>***

# Will fail (false positive)

- In case of languages that demand agreement between *jo*-form and the noun it qualifies
- *E.g.* Sanskrit
- *Yasya_PRON* (wrong!) *baalakasya aananam drshtyaa…* (*jis ladake kaa muha dekhkar*)
- *Yasya_PRON* (wrong!) *kamaniyasya baalakasya aananam drshtyaa…*

# Will also fail for

- Rules that depend on the whether the noun following *jo/vaha/kaun or its form* is oblique or not
- Because the case marker can be far from the noun
- *<vaha or its form>  ladakii  jise piliya kii bimaarii ho gayiii thii* ***ko*** *…*
- **Needs discussions across languages**

*DEM vs. PRON cannot be disambiguated*

*IN GENERAL*

*At the level of the POS tagger*

*i.e.*

*Cannot assume parsing*

*Cannot assume semantics*

# POS critical for Parsing: Stanford Parser output

**Your query**
*My dog also likes eating sausage.*

**Tagging**
My/PRP$ dog/NN also/RB Likes/VBZ eating/VBG sausage/NN ./.

**Parse**
(ROOT (S (NP (PRP$ My) (NN dog)) (ADVP (RB also)) (VP (VBZ likes) (S (VP (VBG eating) (NP (NN sausage))))) (. .)))

**Typed dependencies**
poss(dog-2, My-1) nsubj(likes-4, dog-2) advmod(likes-4, also-3) root(ROOT-0, likes-4) xcomp(likes-4, eating-5) dobj(eating-5, sausage-6)