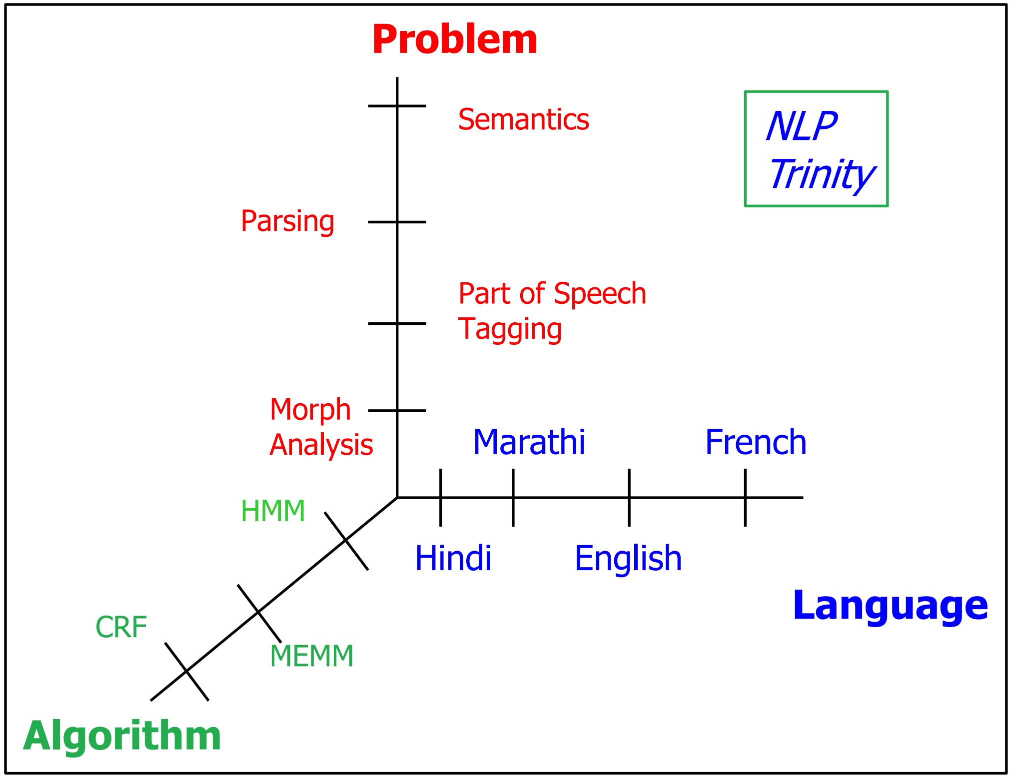

Problem
Semantics
NLP
Trinity
Parsing
Part of Speech
Tagging
Morph
Analysis
Marathi
French
HMM
Hindi
English
Language
CRF
MEMM
Algorithm

What is common amongst these problems

- Fitting *k* clusters to a set of N points
- Fitting *L* lines to a set of points in 2-dim plane
- Tossing two coins and getting the probabilities of heads from each from the observations
- A tourist asking for direction from a person in a country where the inhabitants only lie or speak the truth
- Getting the arc transition probabilities in a probabilistic FSM
- WSD from comparable corpora of two languages in unsupervised setting
- Fitting Gaussian distributions to a set of points

Maximum Likelihood considerations

EM: What is it?

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

EM ("how to know when you do not completely know")

29th August, 2012

Parameter estimation: an exercise in maximization

- Problem:- Given N_h no of heads obtained out of N trials, what is probability of obtaining head?
  - In case of one coin
    - Let probabilty of obtaining head = P_H

      This implies probability of obtaining exactly N_h successes (heads) out of N trials (tosses)

$$f(p_h) = \underset{N_h}{\overset{N}{C}} \times p_h{}^{N_h} \times (1 - p_h)^{N - N_h}$$

Most "likely" value of P_H

- To obtain the most likely value of P_H, we take *ln* of the above equation and differentiate wrt P_H

$$g(P_h) = \ln f(P_h) = \ln \binom{N}{N_h} + N_h \ln P_h + (N - N_h)\ln(1 - P_h)$$

$$\frac{d}{dP_h} g(P_h) = \frac{N_h}{P_h} - \frac{N - N_h}{1 - P_h} = 0$$

$$\Rightarrow P_h = \frac{N_h}{N}$$

Value of P_H in absence of any information

- Suppose we know nothing about the properties of a coin then what can we say about probability of head ? We have to use the entropy *E*:
  - Let P_H be the probability of head
  - Let P_T be the probability of head

$$P_H + P_T = 1 \qquad \text{----(1)}$$

$$E = -P_H \log_2 P_H - P_T \log_2 P_T$$

Entropy

- Entropy is defined as sum of the multiplication of probability and log of probability with – sign.It is the instrument to deal with uncertainity.
- So best we can do is to maximize the entropy.Maximize E subject to the eq (1) and get the value of p_H .

Finding P_H and P_T

$$f(P_H, P_T) = -P_H \log_2 P_H - P_T \log_2 P_T - \lambda(P_H + P_T - 1)$$

$$\frac{\delta F}{\delta \lambda} = P_H + P_T - 1 = 0 \qquad 1$$

$$\frac{\delta F}{S\, P_H} = -k \ln P_H - k - \lambda = 0 \qquad 2$$

$$\frac{\delta F}{S\, P_T} = -k \ln P_T - k - \lambda = 0 \qquad 3$$

From 2 and 3

$$-k \ln P_H - k - \lambda = -k \ln P_T - k - \lambda$$

$$\therefore P_H = P_T \qquad 4$$

From 4 and 1

$$P_H = P_T = \frac{1}{2}$$

A deeper look at EM

- Problem: two coins are tossed, randomly picking a coin at a time. The number of trials is N, number of heads is N_H and number of tails is N_T.
- How can one estimate the following probabilities:
  - p: prob. Of choosing coin_1
  - p_1: prob. Of head from coin_1
  - p_2: prob. Of head from coin_2

Expectation Maximization (1 Coin Toss)

- Toss 1 coin
  - K = Number of heads
  - N = Number of trials
- X = observation of tosses

  $= <x_1>, <x_2>, <x_3> \ldots <x_n>$ - each can take values 0 or 1
- p = probability of Head

  $$= \frac{1}{N}\sum_{i=1}^{N} x_i$$

  (as per MLE – maximizes probability of observed data)

Expectation Maximization (1 Coin Toss)

- $Y = <x_1, z_1>, <x_2, z_2>, <x_3, z_3> \ldots <x_i, z_i> \ldots <x_n, z_n>$
  - x_i = 1 for Head
  - = 0 for Tail
  - z_i = indicator function
  - = 1 if the observation comes from the coin
  - In this case, $z_i = 1 \; \forall i$
- $$P = \frac{1}{N} \sum_{i=1}^{N} x_i z_i$$

Expectation Maximization (2 coin toss)

- $X = <x_1>, <x_2>, <x_3> \ldots <x_i> \ldots <x_n>$
- $Y = <x_1, z_{11}, z_{12}>, <x_2, z_{21}, z_{22}>, <x_3, z_{31}, z_{32}> \ldots <x_i, z_{i1}, z_{i2}> \ldots <x_n, z_{n1}, z_{n2}>$
  - x_i = 1 for Head
  - = 0 for Tail
  - z_{i1} = 1 if the observation comes from coin 1 else 0
  - z_{i2} = 1 if the observation comes from coin 2 else 0
  - only 1 of z_{i1} and z_{i2} can be 1
  - x_i is observed while z_{i1} and z_{i2} is unobserved

Expectation Maximization (2 coin toss)

- Parameters of the setting
  - p_1 = probability of Head for coin 1
  - p_2 = probabilily of Head for coin 2
  - p = probability of choosing for coin 1 for the toss
- Express p, p_1 and p_2 in terms of observed and unobserved data

$$p_1 = \frac{\sum_{i=1}^{N} x_i z_{i1}}{\sum_{i=1}^{N} z_{i1}} \qquad p_2 = \frac{\sum_{i=1}^{N} x_i z_{i2}}{\sum_{i=1}^{N} z_{i2}} \qquad p = \frac{\sum_{i=1}^{N} z_{i1}}{\sum_{i=1}^{N} (z_{i1} + z_{i2})} = \frac{\sum_{i=1}^{N} z_{i1}}{N}$$

Expectation Maximization trick

- Replace z_{i1} and z_{i1} in p, p_1, p_2 with $E(z_{i1})$ and $E(z_{i2})$
  - z_{i1} : event of $x=x_i$ given that observation is from coin 1
  - $E(z_{i1})$ = expectation of z_{i1}

$$E(z_{i1}) = P(coin = coin1 \mid x = x_i)$$

$$= \frac{P(coin = coin1)P(x = x_i \mid coin = coin1)}{P(x = x_i)}$$

$$= \frac{P(coin = coin1)P(x = x_i \mid coin = coin1)}{P(coin = coin1)P(x = x_i \mid coin = coin1) + P(coin = coin2)P(x = x_i \mid coin = coin2)}$$

$$= \frac{p.p_1}{p.p_1 + (1-p).p_2}$$

Summary

- $X = <x_1>, <x_2>, <x_3> \ldots <x_i> \ldots <x_n>$
- $Y = <x_1, z_{11}, z_{12}>, <x_2, z_{21}, z_{22}>, <x_3, z_{31}, z_{32}> \ldots <x_i, z_{i1}, z_{i2}> \ldots <x_n, z_{n1}, z_{n2}>$

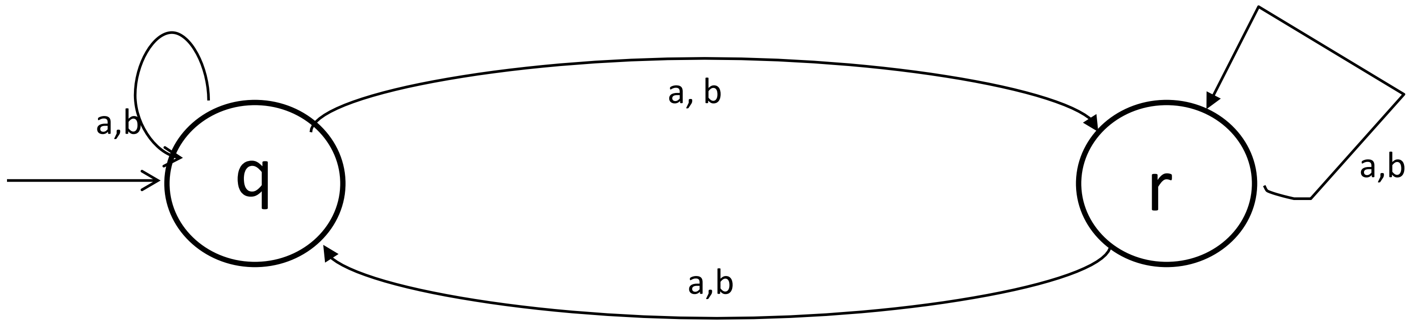
$$\left.\begin{aligned} p_1 = \frac{\sum_{i=1}^{N} x_i z_{i1}}{\sum_{i=1}^{N} z_{i1}} \qquad p_2 = \frac{\sum_{i=1}^{N} x_i z_{i2}}{\sum_{i=1}^{N} z_{i2}} \qquad p = \frac{\sum_{i=1}^{N} z_{i1}}{\sum_{i=1}^{N} (z_{i1} + z_{i2})} = \frac{\sum_{i=1}^{N} z_{i1}}{N} \end{aligned}\right\} \text{M step}$$

$$\left.\begin{aligned} E(z_{i1}) = \frac{p.p_1}{p.p_1 + (1-p).p_2} \qquad E(z_{i2}) = \frac{(1-p).p_2}{p.p_1 + (1-p).p_2} \end{aligned}\right\} \text{E step}$$

Observations

- Any EM problem has observed and unobserved data
- Nature of distribution
  - two coins follow two different binomial distributions
- Oscillation between E and M
  - convergence to local maxima or minima guaranteed
  - greedy algorithm

EM: Baum-Welch algorithm: counts



String = abb aaa bbb aaa

Sequence of states with respect to input symbols

o/p seq

$$\rightarrow q \xrightarrow{a} r \xrightarrow{b} q \xrightarrow{b} q \xrightarrow{a} r \xrightarrow{a} q \xrightarrow{a} r \xrightarrow{b} q \xrightarrow{b} q \xrightarrow{b} q \xrightarrow{a} r \xrightarrow{a} q \xrightarrow{a} r$$

State seq

Calculating probabilities from table

$$P(q \xrightarrow{a} r) = 5/8$$

$$P(q \xrightarrow{b} r) = 3/8$$

$$P(s^i \xrightarrow{w_k} s^j) = \frac{c(s^i \xrightarrow{w_k} s^j)}{\sum_{l=1}^{T} \sum_{m=1}^{A} c(s^i \xrightarrow{w_m} s^l)}$$

T=#states

A=#alphabet symbols

Table of counts

Src	**Dest**	**O/P**	**Count**
q	r	a	5
q	q	b	3
r	q	a	3
r	q	b	2

Now if we have a non-deterministic transitions then multiple state seq possible for the given o/p seq (ref. to previous slide's feature). Our aim is to find expected count through this.

Interplay Between Two Equations

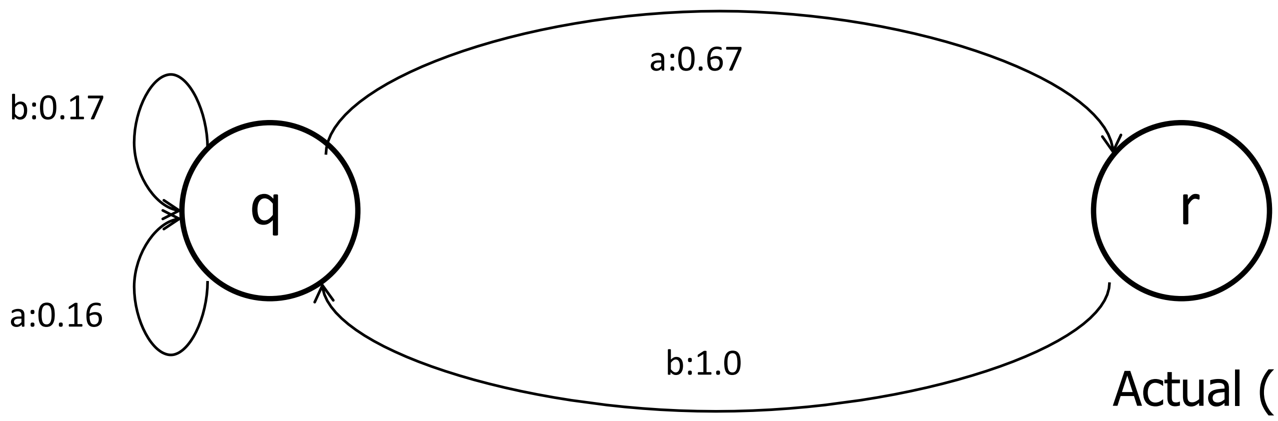
$$P(s^i \xrightarrow{W_k} s^j) = \frac{c(s^i \xrightarrow{W_k} s^j)}{\sum_{l=0}^{T} \sum_{m=0}^{A} c(s^i \xrightarrow{W_m} s^l)}$$

$$C(s^i \xrightarrow{W_k} s^j) =$$

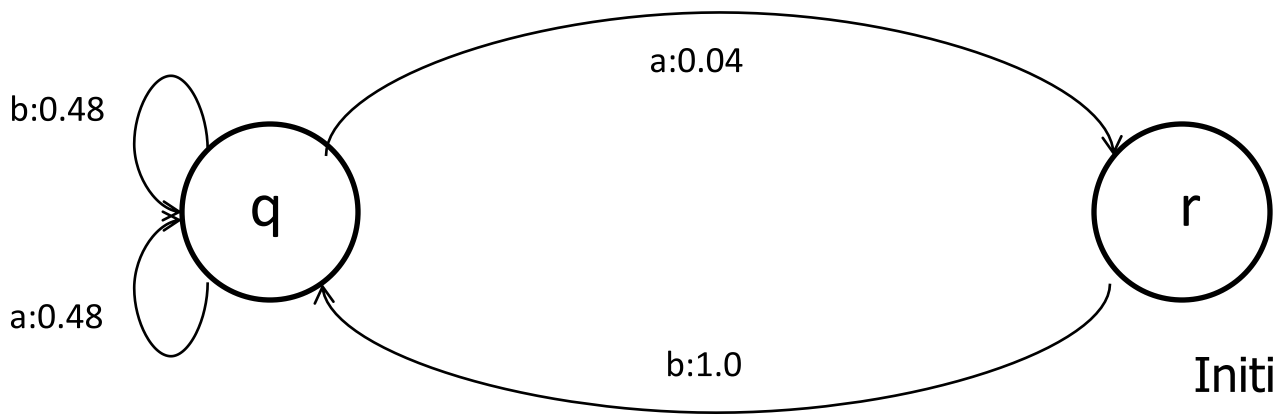
$$\sum_{s_{0,n+1}} P(S_{0,n+1} \mid W_{0,n}) \times n(s^i \xrightarrow{W_k} s^j, S_{0,n+1}, w_{0,n})$$

No. of times the transitions $s^i \xrightarrow{w_k} s^j$ occurs in the string

Illustration



Actual (Desired) HMM



Initial guess

One run of Baum-Welch algorithm: *string ababb*

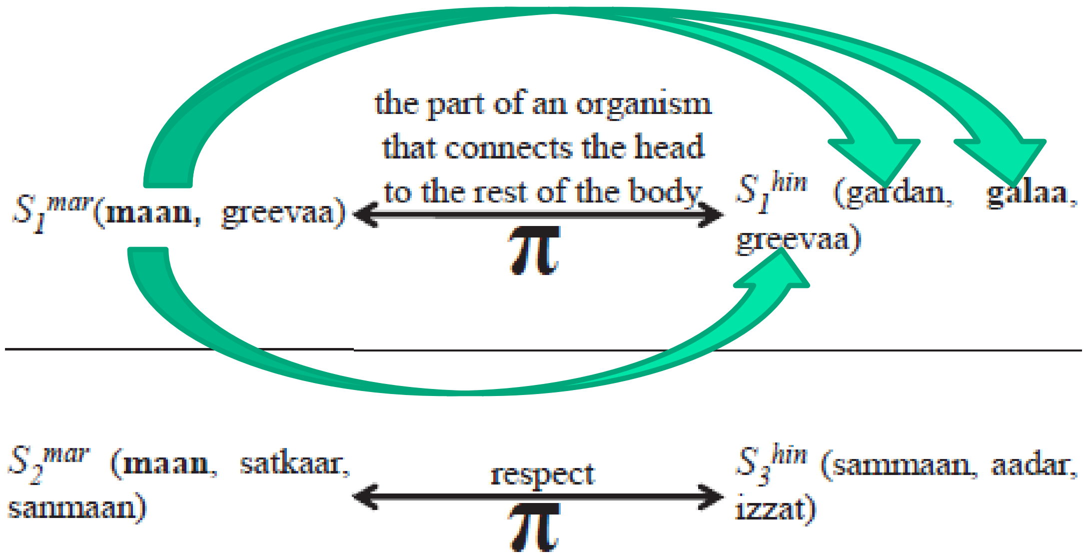
$\in\rightarrow a$	$a\rightarrow b$	$b\rightarrow a$	$a\rightarrow b$	$b\rightarrow b$	$b\rightarrow\in$	P(path)	$q\xrightarrow{a}r$	$r\xrightarrow{b}q$	$q\xrightarrow{a}q$	$q\xrightarrow{b}q$
q	r	q	r	q	q	0.00077	0.00154	0.00154	0	0.00077
q	r	q	q	q	q	0.00442	0.00442	0.00442	0.00442	0.00884
q	q	q	r	q	q	0.00442	0.00442	0.00442	0.00442	0.00884
q	q	q	q	q	q	0.02548	0.0	0.000	0.05096	0.07644
Rounded Total →						0.035	0.01	0.01	0.06	0.095
New Probabilities (P) →							0.06 =(0.01/(0.01+0.06+0.095)	1.0	0.36	0.581

State sequences

* ε is considered as starting and ending symbol of the input sequence string. Through multiple iterations the probability values will converge.

EM based unsupervised Approach

ESTIMATING SENSE DISTRIBUTIONS

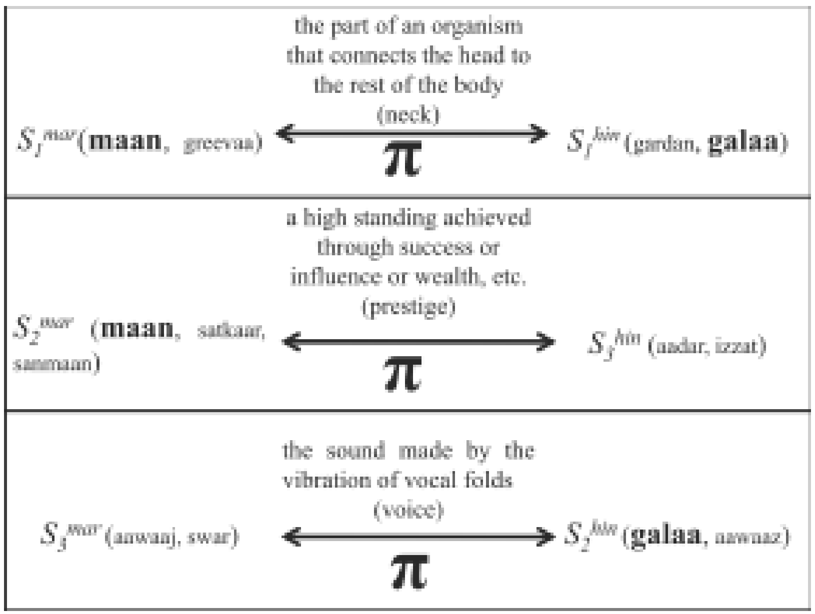


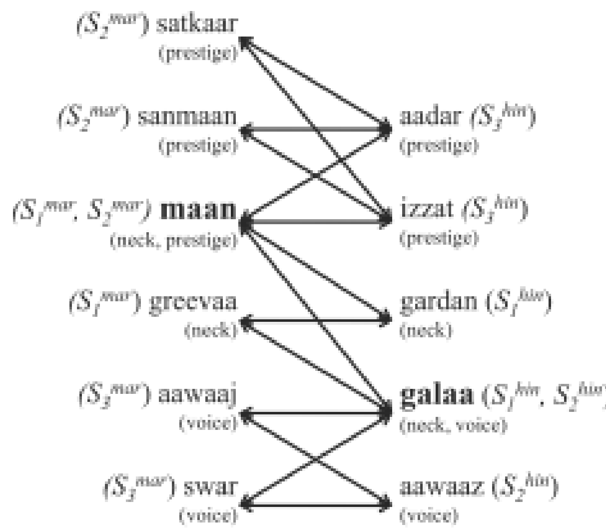
If sense tagged Marathi corpus were available, we could have estimated

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

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:

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

$$\text{where, } S_k^{L_1}, S_i^{L_1} \in synsets_{L_1}(u)$$

$$v \in translations_{L_2}(u, S_k^{L_1})$$

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

M-Step:

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

$$\text{where, } S_j^{L_2}, S_i^{L_2} \in synsets_{L_2}(v)$$

$$a \in translations_{L_1}(v, S_j^{L_2})$$

$$b \in translations_{L_1}(v, S_i^{L_2})$$

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	Avg. degree of wordnet polysemy for polysemous words 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	Avg. degree of wordnet polysemy for polysemous words 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