

Unsupervised Part-of-Speech Tagging: A Fully Bayesian Approach

KAZI SAIDUL HASAN

Ph.D. Student

Department of Computer Science

University of Texas at Dallas

saidul@hlt.utdallas.edu

January 06, 2009

Hidden Markov Models

Elements

- N = Number of states in the model (i.e. number of tags in the tagset)
- Σ = Set of output symbols (i.e. set of words)
- 3 types of parameters

Hidden Markov Models

Elements (cont')

- π_i = Probability of starting at state i ($i = 1, \dots, N$)
 - Vector of N elements
- a_{ij} = Probability of moving from state i to state j
 - Vector of N^2 elements
- $b_i(o)$ = Probability of emitting symbol o ($o \in \Sigma$) from state i
 - Vector of $N|\Sigma|$ elements

Unsupervised HMM Training

Expectation-Maximization (EM)

- Initialize the parameters to some arbitrary values
- Repeat until convergence:
 - E-step: compute the expected values of the unknown variables (tags) based on the current parameter estimates
 - M-step: re-compute the parameter values as a maximum likelihood estimate given the values of the unknown variables computed in the E-step

Unsupervised HMM Training

Expectation-Maximization (EM) (cont')

- This approach gives a single set of parameters
- But let's consider an example where having parameters with fixed values leads to a counterintuitive outcome

The Coin Example

- Problem: whether a given coin is biased or fair
- We've to decide by observing a number of coin flips
- Let's have a model parameter θ i.e. the probability of heads
- Standard way of computing the MAP estimate gives $\theta = n_H / T$ where $n_H = \#$ of heads and $T = \#$ of flips
- As a result, the underlying model suggests that the coin is biased for any sequence of flips that does not contain *exactly* 50% heads

The Bayesian Approach

Integrating over parameter values

- Now, why does integrating over θ help in such cases?
- Since it ensures that the probability of the coin being biased is high only when the sequence of coin flips has $\geq 80\%$ (or $\leq 20\%$) heads
- This obviously is a more sensible prediction
- For details, read Goldwater and Griffiths' (ACL, 2007) paper

The Bayesian Approach

Linguistically Appropriate Priors

- Another advantage of the Bayesian approach
- Linguistic structures have sparse distributions (e.g., one POS tag, say a determiner, is followed by a few other tags such as noun, adjective etc.)
- Integrating over θ permits the use of priors that favor sparse distributions

The Bayesian Approach

Linguistically Appropriate Priors (cont')

- θ includes a set of transition and output distributions
- Each distribution is a multinomial
- For a multinomial with K outcomes, the natural choice for the prior is a K -dimensional Dirichlet distribution
- We assume that all K parameters of the Dirichlet distribution are equal to α (symmetric Dirichlet)
- When $\alpha < 1$, high probabilities are assigned to sparse multinomial distributions

The Bayesian Approach Model

- $$P(t_i | \mathbf{t}_{-i}, \alpha) = \frac{n(t_{i-2}, t_{i-1}, t_i) + \alpha}{n(t_{i-2}, t_{i-1}) + T\alpha}$$
- $$P(w_i | t_i, \mathbf{t}_{-i}, \mathbf{w}_{-i}, \beta) = \frac{n(t_i, w_i) + \beta}{n(t_i) + W(t_i)\beta}$$

$n(x)$ denotes frequency of x . t_i and w_i are the i -th tag and word respectively. α and β are the parameters (hyperparameters) of the Dirichlet distribution. $W(t_i)$ denotes the number of possible words for tag t_i . T denotes the tagset size.

The Bayesian Approach

The Sampling Distribution for t_i

■ $P(t_i | \mathbf{t}_{-i}, \mathbf{w}, \alpha, \beta) =$

$$\frac{n(t_i, w_i) + \beta}{n(t_i) + W(t_i)\beta} \cdot \frac{n(t_{i-2}, t_{i-1}, t_i) + \alpha}{n(t_{i-2}, t_{i-1}) + T\alpha} \cdot$$

$$\frac{n(t_{i-1}, t_i, t_{i+1}) + I(t_{i-2} = t_{i-1} = t_i = t_{i+1}) + \alpha}{n(t_{i-1}, t_i) + I(t_{i-2} = t_{i-1} = t_i) + T\alpha} \cdot$$

$$\frac{n(t_i, t_{i+1}, t_{i+2}) + I(t_{i-2} = t_i = t_{i+2}, t_{i-1} = t_{i+1}) + I(t_{i-1} = t_i = t_{i+1} = t_{i+2}) + \alpha}{n(t_i, t_{i+1}) + I(t_{i-2} = t_i, t_{i-1} = t_{i+1}) + I(t_{i-1} = t_i = t_{i+1}) + T\alpha}$$

The Bayesian Approach

Inference

- Gibbs sampling (Geman and Geman, 1984)
- Random initialization of tags
- Iterative re-sampling
- Exchangeability
- Simulated annealing

The Bayesian Approach

Experimental Setup

- Tag dictionary (Merialdo, 1994), another type of *supervision*
- Data taken from WSJ treebank
- 20K iterations
- Temperature drops from 2.0 to 0.08
- Average over 5 runs
- Baseline – MLHMM

The Bayesian Approach

Fixed Hyperparameter Experiments

- Hyperparameter (α and β) values are varied manually
- Best score (86.8%) is achieved when $\alpha=0.003$ and $\beta=1.0$
- It beats MLHMM (74.5%) by a big margin
- The sparseness of the transition probability matrix is evident as $\alpha < 1$ gives the best score
- For β , the effects are somewhat smaller since the sparseness depends on the tag (content words vs. function words)

The Bayesian Approach

Metropolis-Hastings Update

- Manual option of hyperparameter selection is time consuming and needs a lot of luck
- The smart thing to do is to select these values automatically
- We can add priors over hyperparameters in the model
- And perform a single Metropolis-Hastings update (Gilks et al., 1996) after each iteration to accept new hyperparameter values

The Bayesian Approach

Inferred Hyperparameter Experiments

- 2 experiments conducted by –
 - varying corpus size
 - varying dictionary knowledge
- 2 Bayesian models -
 - BHMM1 – a single β for all tags
 - BHMM2 – a separate β for each tag class
- Same old baseline - MLHMM

The Bayesian Approach

Variation of Information (VI)

- A new evaluation measure
- It is the sum of the amount of information lost in moving from the gold standard to the found clustering for a set of data points
- VI can be computed alongside tagging accuracy and in practice, VI is more informative

The Bayesian Approach

Comparison: MLHMM, BHMM1, & BHMM2

- BHMM1 and 2 outperform MLHMM in almost all cases
- They perform better when less evidence is available
- VI scores give more insight into the systems
- When ambiguity is greater, BHMMs are less *confused* compared to MLHMM
- BHMM2 achieves the best VI score all the time

The Bayesian Approach

Proposed Extensions

- Unsupervised Setting
 - Pseudo-Suffix Emission
- Weakly Supervised Setting
 - Pseudo-Suffix Emission
 - Probabilistic Initialization
 - Discriminative Prediction

The Bayesian Approach

Pseudo-Suffix Emission

- Pseudo-suffixes are extracted by taking the last i characters from a word (in our case, $i = 1$ to 4)
- Motivation – Bengali’s inflectional morphology, even if we rely on pseudo-suffixes
- The idea is to emit 4 pseudo-suffixes alongside a word from a state
- This is done by adding 4 new output distributions to the trigram HMM model

The Bayesian Approach

Probabilistic Initialization

- Implemented in the weakly supervised setting when there is a small amount of labeled data
- The tag dictionary is formed from the labeled data
- For each of the tags of a word, we compute the probability of that tag (given that word) from the labeled data
- Instead of random initialization, we initialize the tags according to the probability distribution over the candidate tags of the word

The Bayesian Approach

Discriminative Prediction

Input: w_i, w_{i-1}, w_{i-2}, V

Output: t_i

if $w_i \in V$ **then**

$w_i =$ Tag drawn from the distribution of w_i 's candidate tags

else if $w_{i-1}, w_{i-2} \in V$ **then**

$w_i =$ Tag drawn from the distribution of the tags following $\langle w_{i-2}, w_{i-1} \rangle$

else if $w_{i-1} \in V$ **then**

$w_i =$ Tag drawn from the distribution of the tags following w_{i-1}

else

$w_i =$ Tag obtained using the Bayesian inference equation

end if

Thank you