Weakly Supervised Part-of-Speech Tagging for Morphologically-Rich, Resource-Scarce Languages

Kazi Saidul Hasan **Vincent Ng**

Human Language Technology Research Institute
University of Texas at Dallas

European Chapter of
the Association for Computational Linguistics, 2009

Unsupervised POS Tagging

POS-tag an unlabeled corpus given a POS lexicon, subject to the constraints imposed by the lexicon

Word	POS tag(s)
...	...
running	NN, JJ
sting	NN, NNP, VB
the	DT
...	...

Figure: A partial lexicon for English

Unsupervised POS Tagging: Common Approach

- Train an HMM (i.e., learn its parameters, θ, which consists of the tag-transition distributions and the output distributions) to maximize the likelihood of the unlabeled corpus using EM

A Simplified HMM for POS Tagging

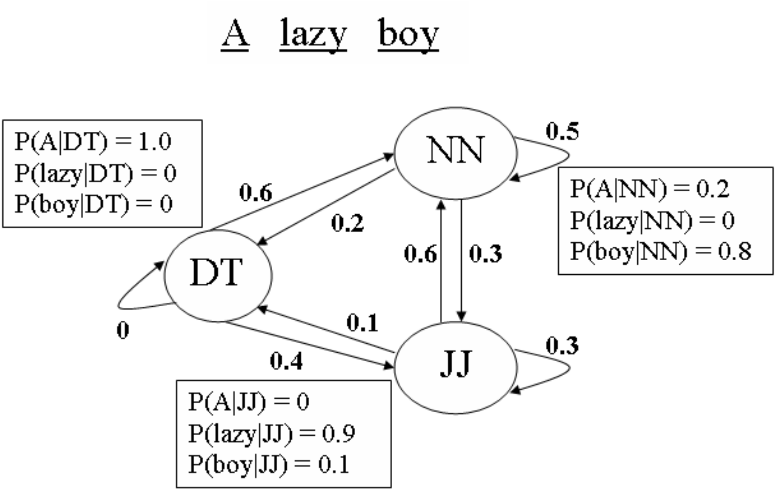

Figure: HMM Parameters

Problem with the Common Approach

- Tagging accuracy is sensitive to many factors (e.g., parameter initializations)

An Alternative to the Common Approach

Goldwater and Griffiths's (2007) nonparametric fully-Bayesian approach

- Adopts an HMM as the underlying model as before, but:
  1. integrates over all possible parameter values, rather than committing to a particular θ

     $$P(\mathbf{t}|\mathbf{w}) = \int P(\mathbf{t}|\mathbf{w}, \theta)P(\theta|\mathbf{w})d\theta$$

  2. favours the learning of skewed tag-transition and output distributions via the use of a prior, $P(\theta|\mathbf{w})$
- Performs inference using Gibbs sampling
- Still makes the usual (unrealistic) assumption that a perfect POS lexicon is available

Our Goals

1. Relax this unrealistic assumption by learning the lexicon automatically from a small set of tagged sentences
   - Many words do not appear in the relaxed lexicon
2. Propose two extensions to G&G's approach for tagging for morphologically-rich, resource-scarce languages
   - Use **Bengali** as our representative language

Extension 1: Induced Suffix Emission (IS)

Motivation

Suffixes are useful indicators of POS tags

A (somewhat naive) way of exploiting suffixes

1. Generate a list of induced suffixes from an unlabeled corpus (using Keshava and Pitler's (2006) algorithm)
2. Create a suffix-based POS lexicon by replacing each word in the original (i.e., word-based) POS lexicon with its suffix induced in Step 1
3. Have the HMM emit suffixes rather than words, subject to the constraints in the suffix-based POS lexicon

- Allows constraints to be imposed on unseen words

Extension 1: Induced Suffix Emission (IS) (contd.)

Potential problem: Over-generalization

Our solution: a hybrid approach

Emit a word if it is in the word-based POS lexicon, otherwise emit its suffix

Extension 2: Discriminative Prediction (DP)

Motivation

We can learn to exploit contextual information to tag a word from a set of POS-tagged sentences, L

Learn three types of probabilities from L:

1. $P(t_i|w_{i-2}, w_{i-1})$: probability of tag t_i following a word bigram
2. $P(t_i|w_{i-1})$: probability of tag t_i following a word
3. $P(t_i|w_i)$: probability of a word having tag t_i

Extension 2: Discriminative Prediction (DP) (contd.)

Apply the Discriminative Prediction Algorithm:

- **If** w_i is in L, assign t_i to w_i with $P(t_i|w_i)$
- **Else if** (w_{i-2}, w_{i-1}) is in L, assign t_i to w_i with $P(t_i|w_{i-2}, w_{i-1})$
- **Else if** w_{i-1} is in L, assign t_i to w_i with $P(t_i|w_{i-1})$
- **Else** obtain the tag using the Gibbs sampler

Evaluation

Goal

Evaluate our two extensions to G&G's tagging model using POS lexicons

- Corpus: Bengali dataset from IJCNLP-08 workshop, which comprises a 50K-token training set & a 30K-token test set
- Training set: for constructing POS lexicons
- Test set: for evaluating model accuracy
- Tagset: IIIT Hyderabad's POS tagset reduced to 15 tags
- Inference: running 5K iterations of the Gibbs sampler; hyperparameters learned by Metropolis-Hastings
- Lexicon: includes only the words and their tags that appear in the small set of POS-tagged sentences

Results

POS tagging models

- **BHMM (Baseline)**: G&G's fully-Bayesian tagging model
- **BHMM+IS**: BHMM with the induced suffix extension
- **BHMM+IS+DP**: BHMM with both extensions

Results (contd.)

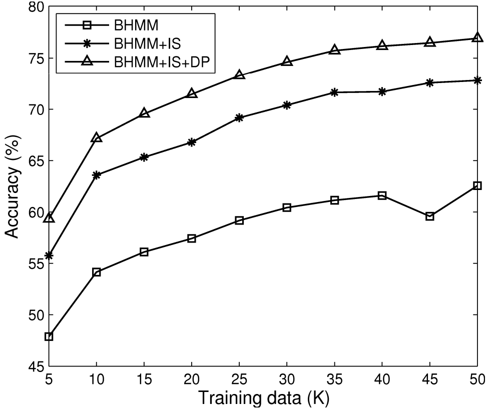

Figure: Learning curves of the POS tagging models

Summary

- Relaxed the unrealistic assumption by learning the lexicon automatically from a small set of tagged sentences
- Proposed two extensions to G&G's model for POS-tagging for morphologically-rich, resource-scarce languages that are effective in improving its performance
  1. Induced suffix emission
  2. Discriminative prediction