Hidden Markov Models and Conditional Random Fields
Lecture 13: October 23, 2013

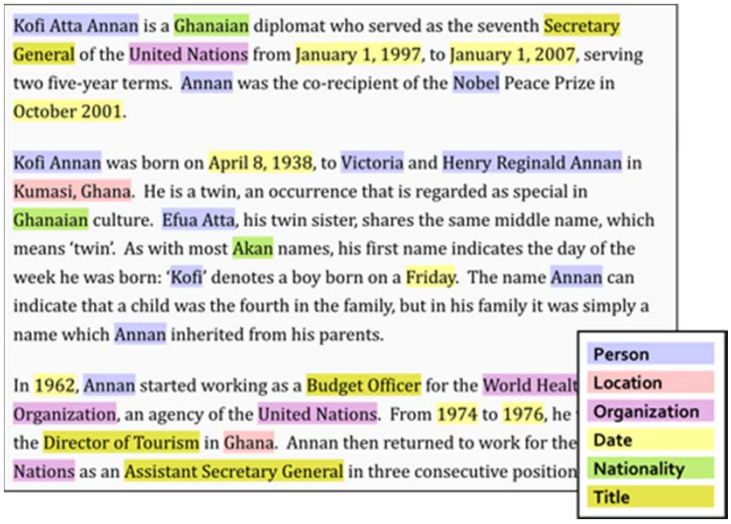
CS886-2 Natural Language Understanding

University of Waterloo

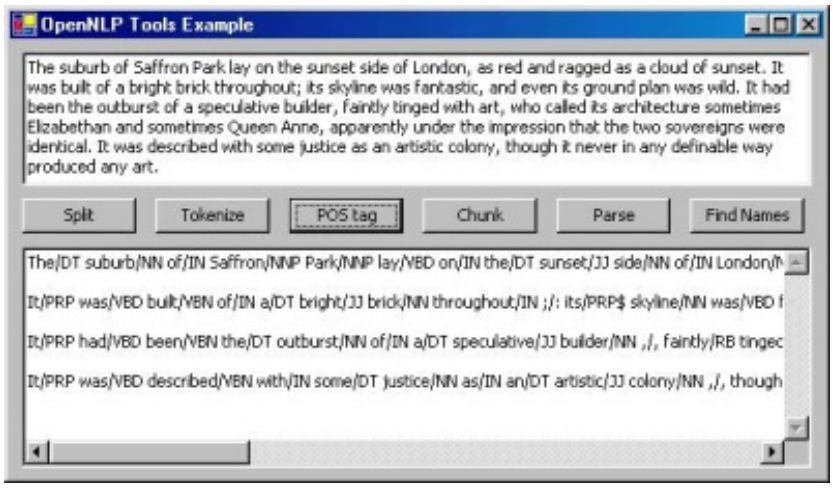
Sequence Data

- So far, we assumed that the data instances are classified independently
  - More precisely, we assumed that the data is iid (identically and independently distributed)
    - E.g., entity recognition, part-of-speech tagging, speech recognition
- Language data naturally forms a sequence where adjacent labels are correlated
  - Text: sequence of characters/words
  - Speech: sequence of audio frames/phonemes

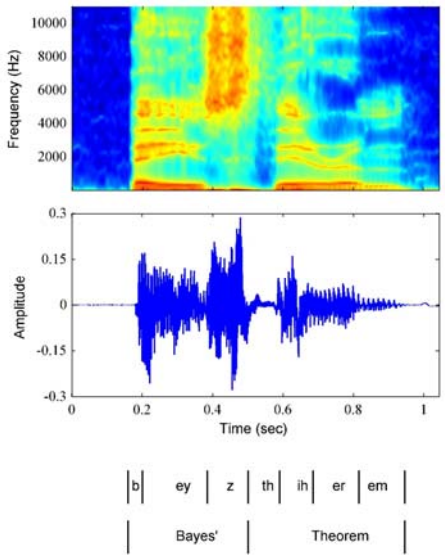
Named Entity Recognition



Part-of-speech Tagging



Speech Recognition



Classification

- Extension of some classification models for sequence data

	Independent classification	Correlated classification
Generative models	Naïve Bayes model	**Hidden Markov Model**
Discriminative models	Logistic Regression	**Conditional Random Field**

Hidden Markov Model

- Graphical Model

- Parameterization
  - Transition distribution:
  - Emission distribution:
- Joint distribution:

Assumptions

- **Stationary Process**: transition and emission distributions are identical at each time step

$$\Pr(X_t|Y_t) = \Pr(X_{t+1}|Y_{t+1}) \quad \forall t$$
$$\Pr(Y_t|Y_{t-1}) = \Pr(Y_{t+1}|Y_t) \quad \forall t$$

- **Markovian Process**: next state is independent of previous states given the current state

$$\Pr(Y_{t+1}|Y_t, Y_{t-1}, \ldots, Y_0) = \Pr(Y_{t+1}|Y_t) \quad \forall t$$

Conditional Random Field

- Graphical Model

- Parameterization
  - Transition potential:
  - Emission potential:
- Conditional distribution:

Assumptions

- **Stationary Process**: transition and emission distributions are identical at each time step

$$f(X_t, Y_t) = f(X_{t+1}, Y_{t+1}) \ \forall t$$
$$f(Y_t, Y_{t-1}) = f(Y_{t+1}, Y_t) \ \forall t$$

- **Markovian Process**: next state is independent of previous states given the current state

$$\Pr(Y_{t+1}|Y_t, Y_{t-1}, \dots, Y_0) = \Pr(Y_{t+1}|Y_t) \quad \forall t$$

Text Tagging

- Hidden Markov Model:
  - $\boldsymbol{Y}$**:** term tag (i.e., entity class, part-of-speech tag)
  - $\boldsymbol{X}$**:** features related to a term
  - $\boldsymbol{f(y_{t-1}, y_t)}$**:** correlation between adjacent tags
  - $\boldsymbol{f(x_t, y_t)}$: correlation between the featres and the tag of a term
- **Joint classification:** $\text{argmax}_{y_{1:t}} \Pr(y_{1:t}|x_{1:t})$?

Inference in sequence models

- Four common tasks:
  - **Monitoring:** $\Pr(y_t|x_{1..t})$
  - **Prediction:** $\Pr(y_{t+k}|x_{1..t})$
  - **Hindsight:** $\Pr(y_k|x_{1..t})$ where $k < t$
  - **Most likely explanation:**
    $$argmax_{y_0,\ldots,y_t} \Pr(y_{0..t}|x_{1..t})$$
- What algorithms should we use?

Hindsight

- $\Pr(y_k|x_{1..t})$ for $k < t$: distribution over a past state given observations
- Example: determine entity type of a word in a sentence

- computation:

$\Pr(y_k|x_{1..t}) \propto \Pr(y_k, x_{k+1..t}|x_{1..k})$ by conditioning

$= \Pr(y_k|x_{1..k}) \Pr(x_{k+1..t}|y_k)$ by chain rule

Forward-backward algorithm

1. Compute $\Pr(y_k|x_{1..k})$ by forward computation

For $i = 1$ to k do

$\Pr(y_i|x_{1..i}) \propto f(x_i, y_i) \sum_{y_{i-1}} f(y_{i-1}, y_i) \Pr(y_{i-1}|x_{1..i-1})$

End

2. Compute $\Pr(x_{k+1..t}|y_k)$ by backward computation

For $j = t$ downto k do

$\Pr(x_{j+1..t}|y_j) = \sum_{y_{j+1}} f(y_j, y_{j+1}) f(x_{j+1}, y_{j+1}) \Pr(x_{j+2..t}|y_{j+1})$

End

3. $\Pr(y_k|x_{k+1..t}) \propto \Pr(y_k|x_{1..k}) \Pr(x_{k+1..t}|y_k)$

- Linear complexity in t

Most likely explanation

- $argmax_{y_{0..t}} \Pr(y_{0..t}|x_{1..t})$: most likely state sequence given observations
- Example: speech recognition
- Computation:

$$\max_{y_{0..t}} \Pr(y_{0..t}|x_{1..t}) \propto \max_{y_t} f(x_t, y_t) \max_{y_{0..t-1}} Pr(y_{0..t}|x_{0..t-1})$$

- Recursive computation:

$$\max_{y_{0..i-1}} \Pr(y_{0..i}|x_{0..i-1}) \propto$$
$$\max_{y_{i-1}} f(y_{i-1}, y_i) f(x_{i-1}, y_{i-1}) \max_{y_{0..i-2}} \Pr(y_{0..i-1}|x_{0..i-2})$$

Viterbi Algorithm

1. Compute $\max_{y_{0..t}} \Pr(y_{0..t}|x_{1..t})$ by dynamic programming

$$\max_{y_0} \Pr(y_{0..1}) \propto \max_{y_0} f(y_0, y_1)$$

For $i = 1$ to $t - 1$ do

$$\max_{y_{0..i}} \Pr(y_{0..i+1}|x_{0..i}) \propto$$
$$\max_{y_i} f(y_i, y_{i+1}) f(x_i, y_i) \max_{y_{0..i-1}} \Pr(y_{0..i}|x_{0..i-1})$$

End

$$\max_{y_{0..t}} \Pr(y_{0..t}|x_{1..t}) \propto \max_{y_t} f(x_t, y_t) \max_{y_{0..t-1}} \Pr(y_{0..t}|x_{0..t-1})$$

- Linear complexity in t