

Introducción a Natural Language Processing

Modelos de Secuencia

Ronald Cárdenas Acosta

Setiembre, 2016

Outline

- 1 Definición
- 2 Notación
- 3 Modelos de secuencia generativos
 - Hidden Markov Models
 - Estimación de Parámetros
 - Inferencia de una secuencia
- 4 Modelos de Secuencia Discriminativos
 - Definición
 - Features o Características
- 5 Análisis de errores
- 6 Aplicación: NER

Predicción Estructurada y Modelos de Secuencia

- Usados en escenarios de Predicción Estructurada, en el que el modelo inferirá una estructura determinada.
- Las muestras pueden presentar dependencia espacial o temporal.
- Un modelo de secuencia modela una estructura (una cadena, árbol, etc).
- Estos modelos se usan muchas aplicaciones de NLP:
 - Tagging
 - Named Entity Recognition
 - Part-of-Speech Tagging
 - Shallow Parsing (Chunking)
 - Syntactic and Dependency Parsing
 - Machine Translation

Part-of-Speech Tagging

INPUT:

Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT:

Profits/**N** soared/**V** at/**P** Boeing/**N** Co./**N** ,/, easily/**ADV** topping/**V** forecasts/**N** on/**P** Wall/**N** Street/**N** ,/, as/**P** their/**POSS** CEO/**N** Alan/**N** Mulally/**N** announced/**V** first/**ADJ** quarter/**N** results/**N** ./.

- N: sustantivo
- V: verbo
- P: preposición
- Adv: adverbio
- Adj: adjetivo

Named Entity Recognition

Según informó el (Departamento)_{ORG} que dirige el consejero (Inaxio Oliveri)_{PER}, los representantes de la (Consejería)_{ORG} y de las universidades del (País Vasco)_{LOC}, (Deusto)_{ORG} y (Mondragón)_{ORG} estudiaron los nuevos retos de estos centros educativos.

- PER: persona
- ORG: organización/empresa
- LOC: lugar

Notación

Sea el par de entrenamiento: $x^i, y^i = (x_1, \dots, x_L), (y_1, \dots, y_L)$

| | |
|---------------------------|---|
| $V = w_1, \dots, w_{ V }$ | Vocabulario, o conjunto de variables observadas |
| $S = s_1, \dots, s_{ S }$ | Conjunto de etiquetas a predecir, o estados |
| L | Longitud de secuencia |
| $x = x_1, \dots, x_L$ | Secuencia de observaciones |
| $y = y_1, \dots, y_L$ | Secuencia de estados |

Outline

- 1 Definición
- 2 Notación
- 3 Modelos de secuencia generativos
 - Hidden Markov Models
 - Estimación de Parámetros
 - Inferencia de una secuencia
- 4 Modelos de Secuencia Discriminativos
 - Definición
 - Features o Características
- 5 Análisis de errores
- 6 Aplicación: NER

Hidden Markov Model

- Es un modelo generativo.
- Es uno de los modelos probabilísticos más comunes.
- Es un tipo especial de *Probabilistic Graphical Model*, con estructura lineal o secuencial.
- Distingue las variables entre:
 - *variables observadas* ($x = x_1 \dots x_L$) o observaciones
 - *variables no observadas* ($y = y_1 \dots y_L$) o estados
- Consideración importante: las observaciones son independientes dados los estados que las generaron.

Hidden Markov Model

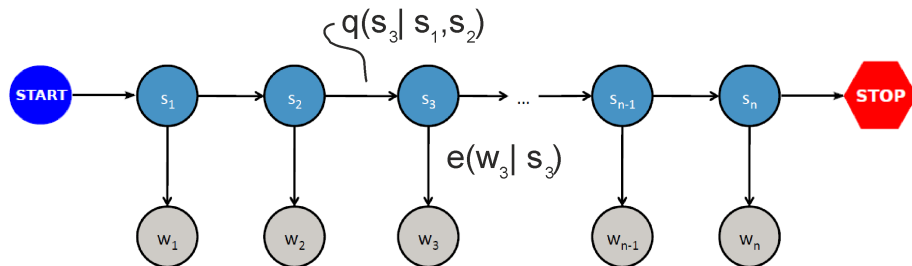


Figure: Modelo gráfico de un HMM.

Hidden Markov Model: Formulación

Sea el par de entrenamiento $x = [x_1, \dots, x_L], y = [y_1, \dots, y_L]$

$$P(x_1, \dots, x_L, y_1, \dots, y_L) = \prod_{j=1}^L q(y_j | y_1, \dots, y_{j-1}) \cdot e(w_j | y_j)$$

Donde:

- $q(y_j | y_1, \dots, y_{j-1})$: Probabilidad de transición de estado
- $e(w_j | y_j)$: Probabilidad de emisión de observación

Hidden Markov Model: Formulación

Para una cadena de Markov de primer grado (Bigram HMM):

$$P(x_1, \dots, x_L, y_1, \dots, y_L) = \prod_{j=1}^L q(y_j | y_{j-1}) \cdot e(w_j | y_j)$$

Consideraciones:

- El estado actual y_j solo depende del anterior y_{j-1} .
- La prob. de transición de un estado s_a a un estado s_b es independiente de la posición en la secuencia.

$$q(y_2 = s_a | y_1 = s_b) == q(y_5 = s_a | y_4 = s_b)$$

Outline

- 1 Definición
- 2 Notación
- 3 Modelos de secuencia generativos
 - Hidden Markov Models
 - Estimación de Parámetros
 - Inferencia de una secuencia
- 4 Modelos de Secuencia Discriminativos
 - Definición
 - Features o Características
- 5 Análisis de errores
- 6 Aplicación: NER

HMM: Estimación de Parámetros

- Sea el conjunto de todos los parámetros θ
- El modelo HMM es entrenado para maximizar la probabilidad logarítmica de la data:

$$\arg \max_{\theta} \sum_{i=1}^N \log P_{\theta}(X = x^i, Y = y^i) \quad (1)$$

HMM: Estimación de Parámetros

- Las probabilidades se aproximan por *Maximum Likelihood Estimate*, es decir, por cuentas.

$$q(s_j | s_{j-1}) = \frac{\text{count}(s_{j-1}, s_j)}{\sum_{k=1}^{|S|} \text{count}(s_k, s_j)}$$
$$e(w_j | s_j) = \frac{\text{count}(w_j, s_j)}{\sum_{v=1}^{|V|} \text{count}(w_v, s_j)}$$

Outline

- 1 Definición
- 2 Notación
- 3 Modelos de secuencia generativos**
 - Hidden Markov Models
 - Estimación de Parámetros
 - **Inferencia de una secuencia**
- 4 Modelos de Secuencia Discriminativos
 - Definición
 - Features o Características
- 5 Análisis de errores
- 6 Aplicación: NER

Inferencia de una secuencia

- Dada la secuencia de observación $x = x_1, \dots, x_L$, se busca la secuencia \hat{y} óptima.
- La inferencia de una secuencia se conoce como **decoding**.
- Existen dos enfoques principales, ambos basados en programación dinámica.

Inferencia de una secuencia

Posterior Decoding o Minimum risk decoding

Para cada posición j de la secuencia

$$\hat{y}_j = \arg \max_{s \in S} P(y_j = s | x_1, \dots, x_L)$$

- Minimiza la probabilidad de error para cada estado y_j , uno a la vez.
- No se garantiza que la secuencia final $\hat{y} = \hat{y}_1, \dots, \hat{y}_L$ sea válida, es decir, puede haber una prob. de transición q igual a cero.
- Algoritmo conocido como *Forward-Backward*

Inferencia de una secuencia

Viterbi Decoding

Consiste en hallar la secuencia de estados globalmente óptima.

$$\hat{y} = \arg \max_{y=y_1 \dots y_L} P(x = x_1, \dots, x_L; y = y_1, \dots, y_L)$$

Utiliza explícitamente las consideraciones de independencia del HMM.
Algoritmo más utilizado en Predicción Estructurada en general.

Outline

- 1 Definición
- 2 Notación
- 3 Modelos de secuencia generativos
 - Hidden Markov Models
 - Estimación de Parámetros
 - Inferencia de una secuencia
- 4 Modelos de Secuencia Discriminativos**
 - Definición**
 - Features o Características
- 5 Análisis de errores
- 6 Aplicación: NER

Clasificadores Secuenciales Discriminativos

Se busca hallar la secuencia óptima \hat{y}

$$\begin{aligned}\hat{y} &= \arg \max_{y_1, \dots, y_L} P(y|x) \\ &= \arg \max_{y_1, \dots, y_L} w \cdot f(x, y) \\ &= \arg \max_{y_1, \dots, y_L} \sum_{j=1}^L w \cdot f_{trans}(j, x_j, y_j, y_{j-1}) + w \cdot f_{emis}(x_j, y_j)\end{aligned}$$

Donde:

$$P(y|x) = \frac{1}{Z(x, y)} \exp\left(\sum_{j=1}^L w \cdot f_{trans}(j, x_j, y_j, y_{j-1}) + w \cdot f_{emis}(x_j, y_j)\right)$$

$Z(x, y)$ es la función partición

Outline

- 1 Definición
- 2 Notación
- 3 Modelos de secuencia generativos
 - Hidden Markov Models
 - Estimación de Parámetros
 - Inferencia de una secuencia
- 4 Modelos de Secuencia Discriminativos**
 - Definición
 - Features o Características**
- 5 Análisis de errores
- 6 Aplicación: NER

Features Básicas

Las features o características binarias se pueden dividir en dos grupos, de manera que imiten los parámetros de un HMM:

$f_{transicion}(j, x_j, y_j, y_{j-1})$ y $f_{emision}(x_j, y_j)$

| Condición | Nombre |
|----------------------------------|-------------------------------|
| $y_j = s_k \wedge j = 1$ | Feature de transición inicial |
| $y_j = s_k \wedge y_{j-1} = s_l$ | Feature de transición |
| $y_j = s_k \wedge j = L$ | Feature de transición final |
| $x_j = w_j \wedge y_j = s_k$ | Feature de emisión |

Features Extendidas

Se pueden definir features que:

- capturen morfología de la palabra: sufijos y prefijos
- dependan arbitrariamente de la secuencia entera de observación (p.e. x_{j-1}, x_j, x_{j+1})
- Información ortográfica: primera letra mayúscula, todo mayúscula, etc
- capturen la semántica de la palabra (p.e. id del cluster al que pertenece)
- Entre muchas más.

Features Extendidas

| Condición | Nombre |
|--|--------------------------------|
| $y_j = s_k \wedge j = 1$ | Feature de trans. inicial |
| $y_j = s_k \wedge y_{j-1} = s_l$ | Feature de transición |
| $y_j = s_k \wedge j = L$ | Feature de trans. final |
| $x_j = w_j \wedge y_j = s_k$ | Feature de emisión básica |
| $x_j = w_j \wedge w_j[0 : p] \forall p \in [1, 2, 3] \wedge y_j = s_k$ | Feature prefijos |
| $x_j = w_j \wedge w_j[L - p : L] \forall p \in [1, 2, 3] \wedge y_j = s_k$ | Feature sufijos |
| $x_j = w_j \wedge w_j$ es todo mayuscula $\wedge y_j = s_k$ | Feature ortografica: mayúscula |
| $x_j = w_j \wedge w_j$ tiene un dígito $\wedge y_j = s_k$ | Feature ortografica: dígitos |

Matriz de transición (Bigram HMM)

- Permite visualizar la distribución de probabilidad de transición
- Solo aplicable a HMM de primer grado (bigramas).
- Cada columna es el estado previo.
- Cada fila es el estado actual

Matriz de transición (Bigram HMM)

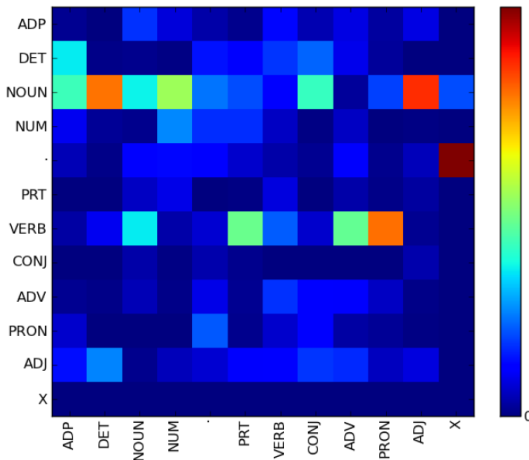


Figure: Matriz de transición de un modelo entrenado de POS Tagging. Columnas son estado anterior; filas, estado actual.

Matriz de Confusión

Permite visualizar los errores de clasificación por clase.

| | | Predicted class | |
|--------------|----------|----------------------|----------------------|
| | | <i>P</i> | <i>N</i> |
| Actual Class | <i>P</i> | True Positives (TP) | False Negatives (FN) |
| | <i>N</i> | False Positives (FP) | True Negatives (TN) |

Matriz de Confusión

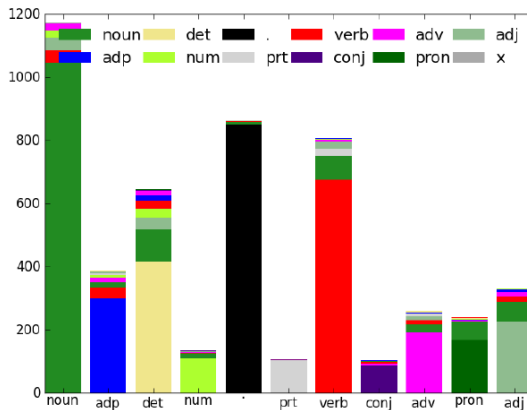


Figure: Representación en barras. Cada barra corresponde a una clase predecida y sus componentes a las clases verdaderas.

Named Entity Recognition

- **Entidad:** grupo de palabras que nombran una persona, lugar, organización, una fecha, número telefónico, etc
- **BIO-format / IOB-format:**
 - B-PER: palabra inicial de entidad Persona
 - I-PER: palabra miembro de entidad Persona
 - O: la palabra no es ninguna entidad

Named Entity Recognition

O B-person I-person I-person O O O O
 With Commander Chris Ferguson at the helm ,
 B-space-shuttle O O O B-place I-place I-place O
 Atlantis touched down at Kennedy Space Center .

Figure: Ejemplo de texto anotado en formato BIO con entidades Persona, Lugar, y nombre de Nave.