Modelos probabilistas de lenguaje

Curso de procesamiento de lenguaje natural

Julio Waissman

Maestría en Tecnologías de la Información
UNaM/UNEE

9 de marzo de 2018

Modelo secuencial probabilista

1. Sea una secuencia de palabras (tokens)

$$w = (w_1, w_2, \ldots, w_k)$$

2. Se quiere modelar la probabilidad de la secuencia de palabras

$$\Pr(w) = \Pr(w_1, w_2, \ldots, w_k)$$

3. Si utilizamos la *regla de la cadena* (teorema)

$$\Pr(w) = \Pr(w_1) \Pr(w_2|w_1) \cdots \Pr(w_k|w_1, w_2, \ldots, w_{k-1})$$

4. Y aplicamos la *hipótesis de Markov* (heurística)

$$Pr(w_i|w_1, \ldots, w_{i-1}) = \Pr(w_i|w_{i-n+1}, \ldots, w_{i-1})$$

Modelo de lenguaje por bigramas

- Si asumimos que $n = 2$, entonces

$$\Pr(w) = \Pr(w_1)\Pr(w_2|w_1)\Pr(w_3|w_2)\cdots\Pr(w_k|w_{k-1})$$

- En forma general:

$$\Pr(w) = \prod_{i=1}^{k+1} \Pr(w_i|w_{i-1})$$

- Por ejemplo: $Pr(\text{la}, \text{cumbiera}, \text{intelectual}) =$

$$\Pr(\text{la})\Pr(\text{cumbiera}|\text{la})\Pr(\text{cumbiera}|\text{intelectual})$$

- Problemas con los inicios y los finales de una frase:

$$Pr(\text{<s>}, \text{la}, \text{cumbiera}, \text{intelectual}, \text{</s>}) =$$

$$\Pr(\text{<s>})\Pr(\text{la}|\text{<s>})\Pr(\text{cumbiera}|\text{la})\Pr(\text{cumbiera}|\text{intelectual})\Pr(\text{intelectual}|\text{</s>})$$

Modelo de lenguaje por n-gramas

Modelo de lenguaje por *n*-gramas

$$\Pr(w) = \prod_{i=1}^{k+1} \Pr(w_i | w_{i-n+1}, \ldots w_{i-1})$$

Donde los parámetros de la distribución se obtienen maximizando la verosimilitud logarítmica:

$$\log \Pr(w_{train}) = \sum_{i=1}^{N+1} \log \Pr(w_i | w_{i-n+1}, \ldots, w_{i-1})$$

¿Que podemos hacer con un modelo probabilista secuencial?

1. Estimar la cadena más plausible

$$w^* = \arg\max \Pr(w_i, \ldots, _k \,|w_1, \ldots, w_{i-1})$$

2. Calcular la probabilidad de una frase dada en dicho lenguaje

$$Pr(w_1, w_2, \ldots, w_k)$$

3. Estimar la probabilidad de una palabra dado un contexto

$$Pr(w_i|w_{i-r}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+n})$$

4. Recomendaciones en servicios de mensajería, correctores de ortografía, reconocimiento de voz, teclado telefono, etc.

Todo se trata de contar

- Sea $c(w_a, w_b, w_c)$ el número de veces que aparece el trigrama w_a, w_b, w_c en el corpus de entrenamiento.
- El corpus de entrenamiento puede a su vez estar dividido en documentos (frases).
- La funcion de cuenta se puede extenter a cualquier n-grama como $c(w_{a_1}, , w_{a_n})$.

Todo se trata de contar

- La probabilidad se calcula como

$$\Pr(w_c|w_a, w_b) = \frac{c(w_a, w_b, w_c)}{c(w_a, w_b)}$$

- Obtener los valores de $\Pr(w_c|w_a, w_b)$ para todos los trigramas (w_a, w_b, w_c) que se pueden encontrar a partir de las palabras de nuestro diccionario, es lo que se conoce como Modelo de lenguaje por trigramas o modelo de trigramas.
- Generalizable fácilmente a bigramas, cuatrigramas, etc. . .

Un ejemplo

Dos cuerpos frente a frente
son a veces dos olas
y la noche es océano.

Dos cuerpos frente a frente
son a veces dos piedras
y la noche desierto.

Dos cuerpos frente a frente
son a veces raíces
en la noche enlazadas.

Dos cuerpos frente a frente
son a veces navajas
y la noche relámpago.

Dos cuerpos frente a frente
son dos astros que caen
en un cielo vacío.

Otro ejemplo

¿Como medir la calidad del modelo?

- Sea $w_{test} = (w_1, \ldots, w_N)$ un *corpus de prueba*.

- La verosimilitud del modelo respecto al corpus de prueba esta dada por

$$\mathcal{L} = \Pr(w_{test}) = \prod_{i=1}^{k+1} \Pr(w_i | w_{i-n+1}, \ldots w_{i-1})$$

- La perplejidad del modelo respecto al corpus de prueba esta dada por

$\[P = 1_{\overline{\{Pr}}(w_{test})^{\frac{1}{N}}\}\}$

Mientras menor sea la perplejidad, mejor es el modelo

Problemas con la estimación de estos modelos

- ¿Que pasa si un n-grama dado $(w_1, \ldots, w_n)$ (obtenido a partir del vocabulario) no se encuentra en el corpus de entrenamiento?

  $$\Pr(w_1, \ldots, w_n) = 0$$

  y por lo tanto la probabilidad de cualquier secuencia con dicho n-grama será 0. (¡Y la perplejidad infinita!)
  - Problema del método de estimación
  - Métodos de suavizado

- ¿Que pasa si en el conjunto de prueba (o en la aplicación) existen palabras que no se encuentran en el vocabulario?
  - Palabras fuera de vocabulario (*OOV words*)
  - Uso de token espacial (<UNK>)
  - Necesidad de un *vocabulario cerrado*

- Temas fuera del alcance del curso

Etiquetado secuencial

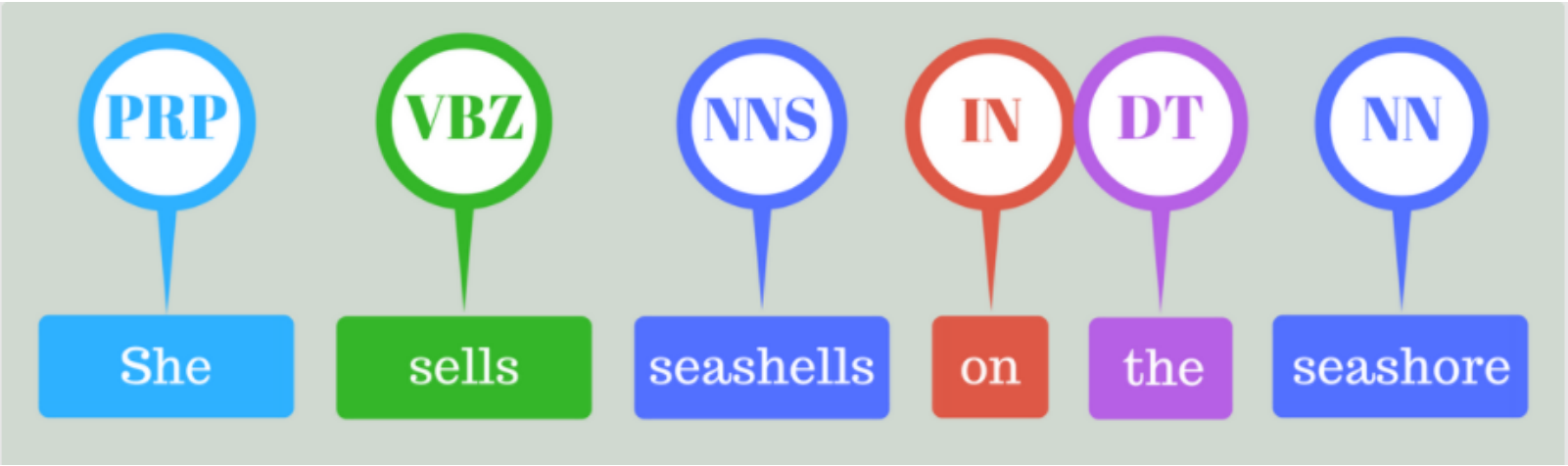
Problema

Dada una secuencia de tokens, inferir la secuencia *más probable* de etiquetas para cada uno de esos tokens

Dos Aplicaciones principales

- Etiquetado de *Parte del discurso* (*POS Tagging*)
- Reconocimiento de entidades (*NER*)

Etiquetado de parte del discurso



- Etiquetas estándar de acuerdo a *Universal dependencies*
- 17 etiquetas POS universales (todas usadas por español)

```
ADP - AUX - CCONJ - DET - NUM - PART - PRON - SCONJ # Cerrados
ADJ - ADV - INTJ - NOUN - PROPN - VERB # Abiertos
PUNCT - SYM - X # Otros
```

- 3 corpus etiquetados en español.

Reconocimiento de entidades

- Etiquetado de nombres propios que pertenecen a clases en el mundo real que tienen nombre propio:
  - Peronas (`PER`),
  - Organizaciones (`ORG`),
  - Localidades (`LOC`),...
- Tambien es común considerar como entidades con nombre a:
  - Fechas y horas
  - unidades, etc.
- Las entidades pueden ser tan específicas como sea necesario

Notación IOB para etiquetado

Típicamente codificado en xml

The	O	viridicatic	B-TRIVIAL
studies	O	acid	I-TRIVIAL
also	O	and	O
resulted	O	terrestric	B-TRIVIAL
in	O	acid,	I-TRIVIAL
the	O	found	O
identification	O	in	O
of	O	ethyl	B-SYSTEMATIC
two	O	acetate	I-SYSTEMATIC
known	O	and	O
tetronic	B-FAMILY	n-butanol	B-SYSTEMATIC
acids,	I-FAMILY	extracts.	O

El etiquetado de corpus es una tarea no trivial (realizada a mano).

Idea básica para etiquetado secuencial

- Sea $x = (x_1, x_2, \ldots, x_T)$ una secuencia de tokens (entradas)
- Sea $y = (y_1, y_2, \ldots, y_T)$ una secuencia de etiquetas (salidas)
- El problema es encontrar la *secuecia más probable de etiquetas* y dada la secuencia de tokens x, lo cual se puede hacer:
  - De forma *generativa*

$$\arg\max_y \Pr(x, y)$$

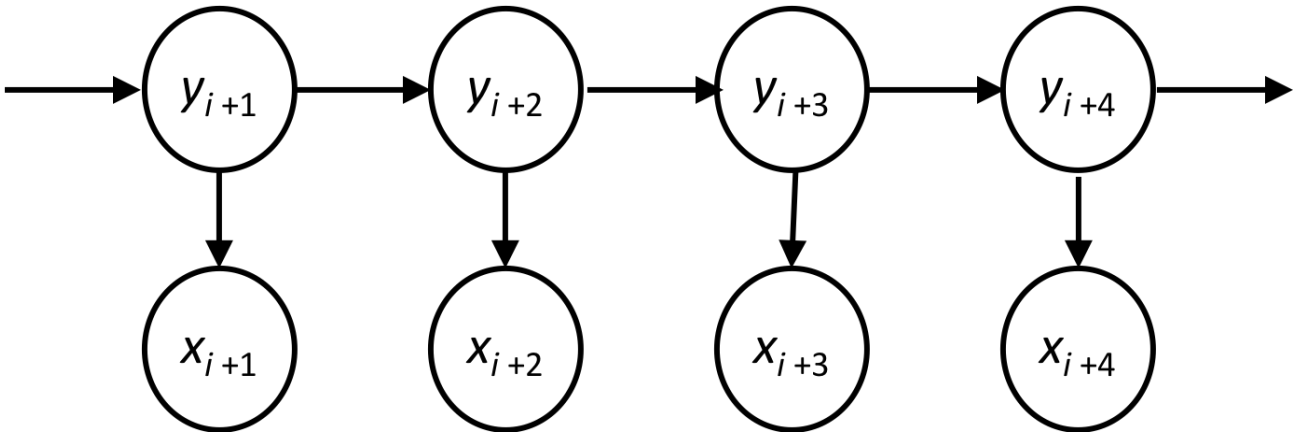
  - De forma *discriminativa*

$$\arg\max_y \Pr(y|x)$$

Solo vamos a establecer muy superficialmente los esquemas de los modelo principales

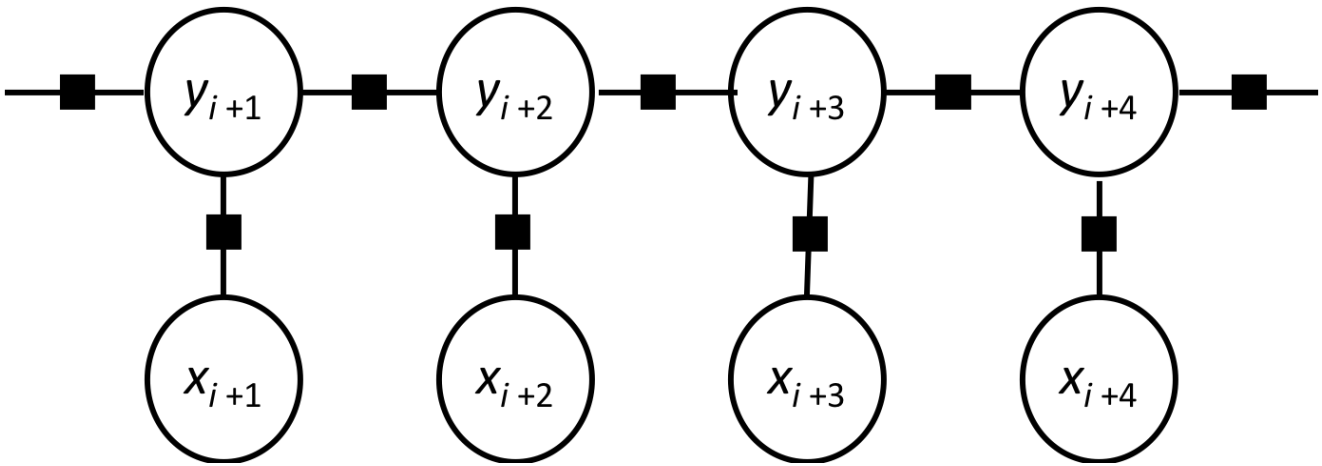
Esquema de etiquetado con HMM

$$\Pr(y, x) = \prod_{t=1}^{T} \Pr(y_t | y_{t-1}) \Pr(x_t | y_t)$$



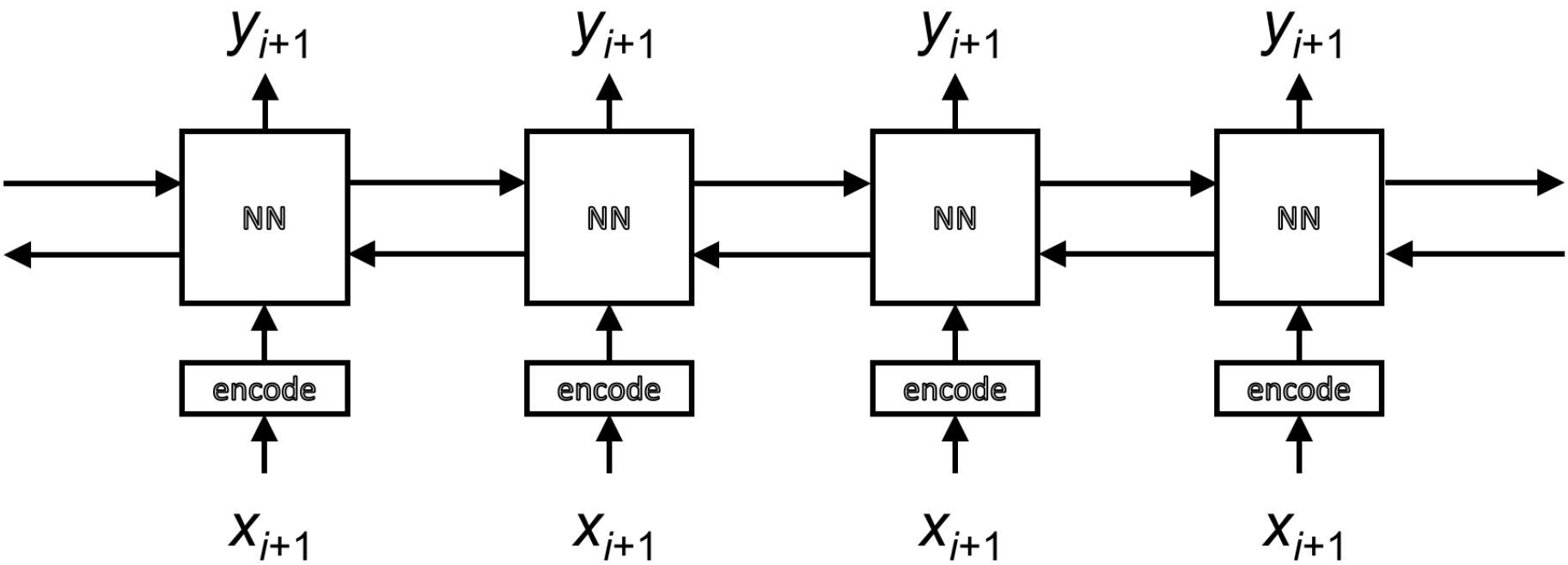
Esquema de etiquetado con CRF

$$\Pr(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} \exp\left(\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}) \sum_{s=1}^{S} \phi_s F_s(y_t, x_t) \right)$$



Esquema de etiquetado con RNN

Redes recurrentes bi-direccionales



Etiquetando con *spacy*