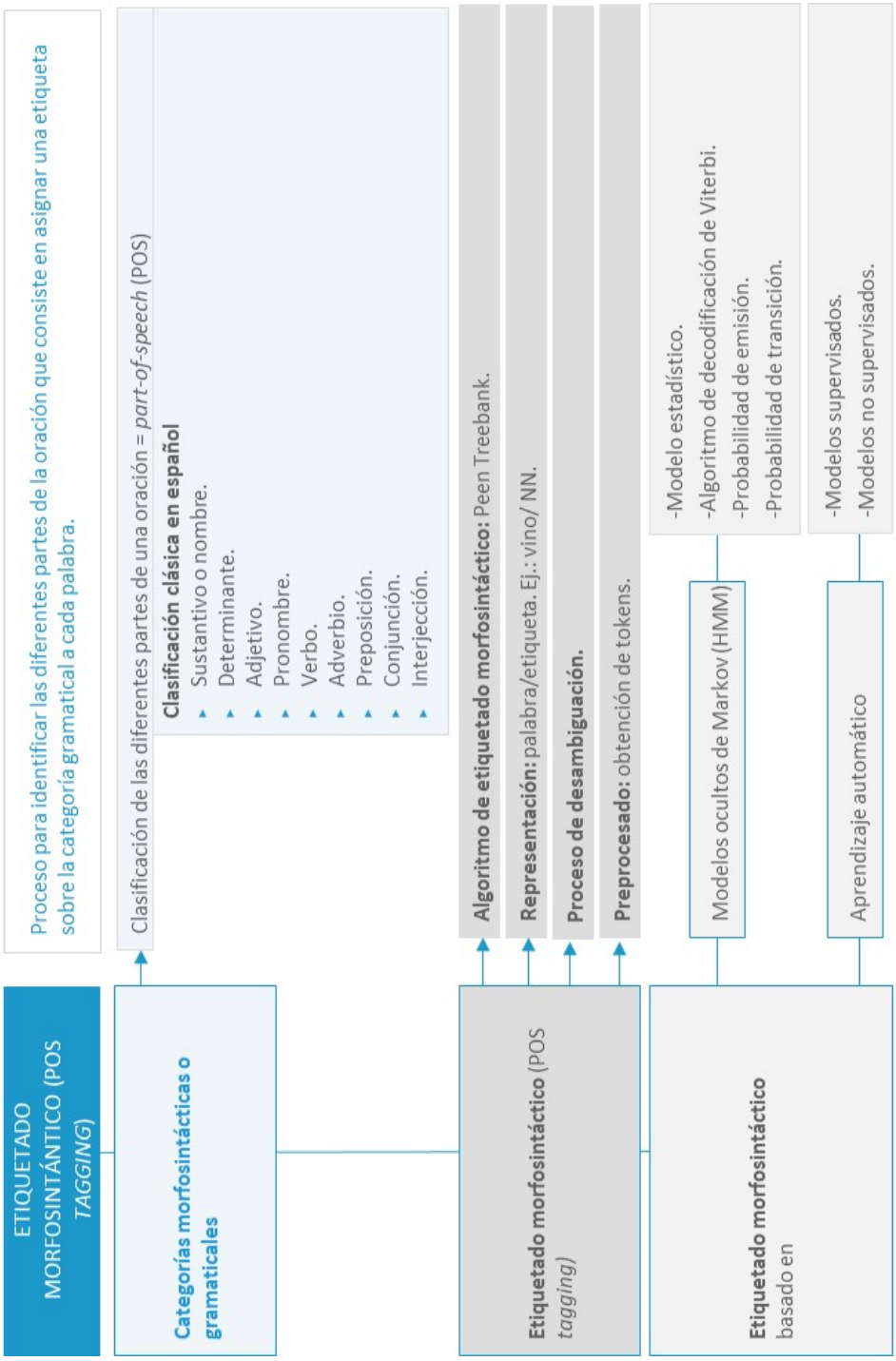
Procesamiento del Lenguaje Natural

Tema 3. Etiquetado morfosintáctico (POS tagging)

Índice

Esquema



3.1. Introducción y objetivos

A continuación, se estudiará el etiquetado morfosintáctico y cómo calcularlo, haciendo hincapié en los modelos ocultos de Markov *(Hidden Markov Models).*

Objetivos

- Identificar las diferentes categorías morfosintácticas o también llamadas gramaticales.
- Entender el funcionamiento y las características del etiquetado morfosintáctico.
- Aplicar un método estocástico basado en modelos ocultos de Markov (HMM) para realizar el etiquetado morfosintáctico.
- Describir diversos métodos basados en aprendizaje automático para realizar el etiquetado morfosintáctico.

3.2. Categorías morfosintácticas o gramaticales

La Real Academia Española (RAE, s. f.) en su diccionario de la lengua española define la palabra **morfosintaxis** como: «1. f. Ling. Parte de la gramática que integra la morfología y la sintaxis». Tal como se ha presentado en los temas anteriores:

> La morfología estudia la estructura de las palabras y la sintaxis, el modo en que se combinan las palabras. Por lo tanto, la morfosintaxis aúna la morfología y la sintaxis para determinar las diferentes partes de la oración, llamadas *part-of-speech* (POS) en inglés.

Las **categorías morfosintácticas** del lenguaje, que en español también se llaman categorías **gramaticales,** proporcionan una clasificación de las diferentes partes de la oración, es decir, una clasificación de las palabras según su tipo.

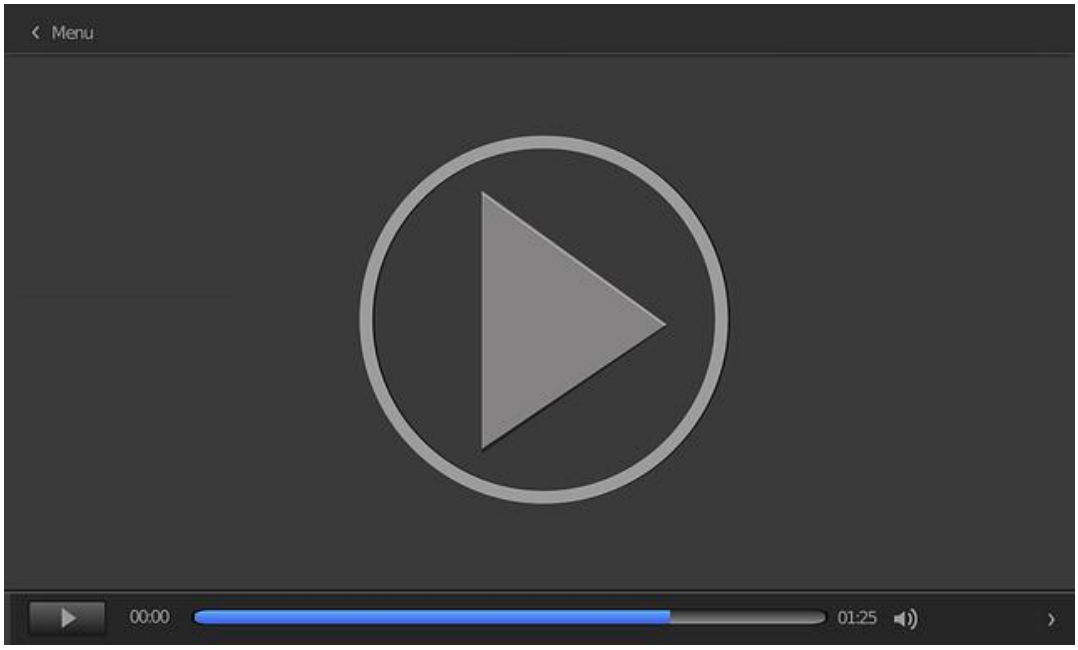
Las categorías gramaticales del español, según la clasificación clásica, son nueve: sustantivo o nombre, determinante, adjetivo, pronombre, verbo, adverbio, preposición, conjunción e interjección. Las definiciones de las diferentes clases de palabras o categorías gramaticales se presentan en la siguiente tabla.

Categorías morfosintácticas o gramaticales	
Sustantivo o nombre	Clase de palabras cuyos elementos poseen género y número, forman sintagmas nominales con diversas funciones sintácticas y designan entidades de diferente naturaleza.
Determinante	Clase de palabras cuyos elementos determinan al sustantivo o al grupo nominal y se sitúan generalmente en posición prenominal. El artículo definido y los demostrativos son determinantes.
Adjetivo	Clase de palabras cuyos elementos modifican a un sustantivo o se predican de él y denotan cualidades, propiedades y relaciones de diversa naturaleza. Ejemplos: inteligente, amplio, numérico, etc.
Pronombre	Clase de palabras cuyos elementos hacen las veces del sustantivo o del sintagma nominal y que se emplean para referirse a las personas, los animales o las cosas sin nombrarlos. Ej.: ella, esto, quién.
Verbo	Clase de palabras cuyos elementos pueden tener variación de persona, número, tiempo, modo y aspecto.
Adverbio	Clase de palabras cuyos elementos son invariables y tónicos, están dotados generalmente de significado léxico y modifican el significado de varias categorías, principalmente de un verbo, de un adjetivo, de una oración o de una palabra de la misma clase.
Preposición	Clase de palabras invariables cuyos elementos se caracterizan por introducir un término, generalmente nominal u oracional, con el que forman grupo sintáctico.
Conjunción	Clase de palabras invariables, generalmente átonas, cuyos elementos manifiestan relaciones de coordinación o subordinación entre palabras, grupos sintácticos u oraciones.
Interjección	Clase de palabras invariables, con cuyos elementos se forman enunciados exclamativos, que manifiestan impresiones, verbalizan sentimientos o realizan actos de habla apelativos.

Tabla 1. Definiciones de las diferentes categorías morfosintácticas. Fuente: elaboración propia adaptado de la Real Académica Española (RAE), s. f.

Ideas clave

En el vídeo *Morfosintaxis y partes de la oración* se presentará qué es la morfosintaxis, parte de la gramática que integra la morfología y la sintaxis.



03.01. Morfosintaxis y partes de la oración

Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=a53bd504-0455-43f9-91be-ac6b00c4c877

Conocer las partes de la oración (categorías morfosintácticas o gramaticales) es útil debido a la gran cantidad de información que brindan sobre una palabra y sus vecinos.

- Saber si una palabra es un sustantivo o un verbo nos dice mucho acerca de las palabras vecinas, por ejemplo, los sustantivos pueden ir precedidos de determinantes o seguidos de adjetivos.

- Y también sobre la estructura sintáctica, por ejemplo, los sustantivos son generalmente parte de los sintagmas nominales.

Por estas razones, el etiquetado morfosintáctico es un componente muy importante del análisis sintáctico que se va a estudiar a continuación.

3.3. Funcionamiento y características del etiquetado morfosintáctico

El etiquetado morfosintáctico, llamado *POS tagging (part-of-speech tagging)* en inglés, es el **proceso para identificar las diferentes partes de la oración** y consiste en **asignar una etiqueta *(tag)*** sobre la categoría gramatical a cada una de las palabras de un texto de entrada.

La entrada del algoritmo de etiquetado morfosintáctico es una secuencia de palabras y la salida del algoritmo es una secuencia de pares formados por la palabra y la correspondiente etiqueta indicando la categoría gramatical a la que pertenece dicha palabra. Ante esto:

- Existen diferentes conjuntos de etiquetas que se pueden utilizar en el análisis morfosintáctico.
- Algunos etiquetadores morfosintácticos permiten indicar a su entrada el conjunto de etiquetas que:
  - Identifican cada categoría gramatical.
  - Se van a utilizar en el proceso de etiquetado de cada parte de la oración.

Hoy en día, la mayoría de los algoritmos de procesamiento del lenguaje natural que procesan palabras en inglés utilizan el **Penn Treebank** (Marcus, Santorini y Marcinkiewicz, 1993).

El Penn Treebank es un conjunto de 45 etiquetas que identifican las diferentes partes de la oración, tal como se muestra en la siguiente imagen.

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	*+,%, &*
CD	cardinal number	*one, two*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb gerund	*eating*
JJ	adjective	*yellow*	VBN	verb past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, sing.	*IBM*	$	dollar sign	*$*
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

Figura 1. Etiquetas para las categorías gramaticales en el Penn Treebank. Fuente: Jurafsky y Martin, 2009.

En el etiquetado morfosintáctico se identifican las diferentes partes de la oración y se asigna una etiqueta a cada una de las palabras, tal y como se ha comentado anteriormente.

La forma más habitual para representar la salida del etiquetador morfosintáctico es colocar, después de cada palabra, la etiqueta para la categoría gramatical separada por una barra.

Por ejemplo, si el etiquetador morfosintáctico analiza la frase:

- Bebo un vaso del vino tinto.

Suponiendo que se utilizan las etiquetas para las categorías gramaticales definidas en el Penn Treebank, la salida sería:

```
bebo/VBP un/DT vaso/NN de/IN el/DT vino/NN tinto/JJ
```

Para la frase:

- Vino de un lugar lejano.

El etiquetado morfosintáctico sería:

```
vino/VBZ de/IN un/DT lugar/NN lejano/JJ
```

En estos dos ejemplos de etiquetado morfosintáctico se observa que la misma palabra «vino» se etiqueta de forma distinta para las dos frases.

En el primer ejemplo, la palabra «vino» pertenece a la categoría gramatical de los sustantivos o nombres (NN). Mientras que, en el segundo ejemplo, pertenece a la categoría gramatical de los verbos (VBZ).

Así, el etiquetado morfosintáctico realiza durante su funcionamiento un proceso de desambiguación: una palabra, que es ambigua y puede pertenecer a más de una categoría gramatical, se etiqueta correctamente según el contexto de la frase analizada.

Además, es importante notar que el algoritmo que realiza el etiquetado morfosintáctico ha separado la palabra «del» (que aparece en la frase «bebo un vaso del vino tinto») en las palabras *de* y *el* antes de etiquetarlas respectivamente como una preposición (IN) y un determinante (DT).

Ideas clave

Identificar que una palabra es una contracción y separarla en las dos que la constituyen forma parte del proceso previo de preprocesado de la oración, que permite separar la frase en las diferentes palabras.

La identificación de las palabras de una oración también es llamada proceso de obtención de los tokens, porque token es el nombre inglés para definir una cadena de caracteres que representa una palabra y se realiza siempre previamente al etiquetado morfosintáctico y a otras tareas de procesamiento del lenguaje natural.

3.4. Etiquetado morfosintáctico basado en modelos ocultos de Markov (HMM)

Una de las técnicas más utilizadas en el etiquetado morfosintáctico es los modelos ocultos de Markov o HMM (por sus siglas del inglés*, Hidden Markov Model)*. Esta técnica consiste en construir un modelo de lenguaje estadístico que se utiliza para obtener, a partir de una frase de entrada, la secuencia de etiquetas gramaticales que tiene mayor probabilidad.

Un modelo oculto de Markov es un modelo estadístico que se puede representar como una máquina de estados finitos, pero donde las transacciones entre estados son probabilísticas y no determinísticas. El objetivo es determinar los parámetros desconocidos (ocultos) a partir de los parámetros observables.

Por ejemplo, si hemos etiquetado una palabra como determinante, la próxima palabra será un nombre con un 40 % de probabilidad, un adjetivo con otro 40 % y un número el 20 % restante. Conociendo esta información, un sistema puede decidir que la palabra «vino» en la frase «el vino» es más probable que sea un nombre a que sea un verbo.

Para el etiquetado morfosintáctico, **los HMM son entrenados en un conjunto de datos totalmente etiquetados.** Esto es un conjunto de frases con cada palabra anotada con una etiqueta describiendo su categoría gramatical. A partir de los datos de entrenamiento, los HMM fijan estimaciones de máxima verosimilitud para cada uno de los estados y determina las diferentes probabilidades que rigen el modelo.

Para llevar a cabo el proceso de estimación de probabilidades se usa el algoritmo de **decodificación de Viterbi** (Forney, 1973).

Ideas clave

El objetivo de decodificación HMM es elegir la secuencia de etiquetas más probable dada la secuencia de observación

n palabras

w_1^n :

$$\hat{t}_1^n = \underset{t_1^n}{argmax}\ P\left(t_1^n | w_1^n\right)$$

Esta ecuación la podemos reescribir mediante el uso de la regla de Bayes de la siguiente forma:

$$\hat{t}_1^n = \underset{t_1^n}{argmax}\ \frac{P\left(w_1^n | t_1^n\right) P\left(t_1^n\right)}{P\left(w_1^n\right)}$$

También podemos simplificar esta ecuación eliminando el denominador $P\left(w_1^n\right)$:

$$\hat{t}_1^n = \underset{t_1^n}{argmax}\ P\left(w_1^n | t_1^n\right) P\left(t_1^n\right)$$

Los etiquetadores HMM hacen dos suposiciones que permiten simplificar estas ecuaciones aún más:

- La primera es que la probabilidad de aparición de una palabra depende solo de su propia etiqueta y es independiente de las palabras y etiquetas vecinas:

$$P\left(w_1^n | t_1^n\right) \approx \prod_{i=1}^{n} P\left(w_i^n | t_i^n\right)$$

- La segunda suposición, también llamada bigrama o digrama, es que la probabilidad de una etiqueta solo depende de la etiqueta anterior, en lugar de toda la secuencia de etiquetas:

$$Pt_i^n \approx \prod_{i=1}^{n} P\left(t_i | t_{i-1}\right)$$

Aplicando estas suposiciones a las ecuaciones anteriores terminamos con la siguiente ecuación para la secuencia de etiquetas más probable de un etiquetador bigrama, las cuales corresponden a la probabilidad de emisión y la probabilidad de transición de un HMM:

$$\hat{t}_1^n = \underset{t_1^n}{argmax}\ P\left(t_1^n|w_1^n\right) \approx \underset{t_1^n}{argmax} \prod_{i=1}^{n} \overbrace{P\left(w_i^n|t_i^n\right)}^{emission} \overbrace{P\left(t_i|t_{i-1}\right)}^{transition}$$

Ejemplos de probabilidad de transición y probabilidad de emisión

Veamos a través de un ejemplo cómo se calculan y se utilizan en una tarea de etiquetado estas probabilidades. En el etiquetado HMM, las probabilidades se estiman simplemente a partir de un corpus de entrenamiento etiquetado; para este ejemplo vamos a utilizar el **corpus etiquetado WSJ,** una colección de un millón de palabras que se publicaron en los artículos del Wall Street Journal (WSJ) en 1989 y que están anotadas utilizando las etiquetas morfosintácticas del Penn Treebank.

Un corpus lingüístico es una colección de textos representativos de una lengua que se utilizan para el análisis lingüístico. Los corpus pueden estar anotados o etiquetados de forma que las palabras que lo conforman presentan, además, algún tipo de información lingüística.

Accede al corpus a través del aula virtual o desde la siguiente dirección web: https://catalog.ldc.upenn.edu/LDC2000T43

Probabilidades de transición

Las probabilidades de transición de etiqueta $P\left(t_i|t_{i-1}\right)$ representan la probabilidad de una etiqueta dada la etiqueta anterior. Por ejemplo, los verbos modales (etiqueta MD) como *«can»* (poder) son muy

probablemente seguidos por un verbo en la forma base (etiqueta VB) como *«run»* (correr), por lo que espera que esta probabilidad sea alta.

La estimación de **máxima verosimilitud** de una probabilidad de transición se calcula por recuento de las veces que vemos la primera etiqueta en un corpus etiquetado y la frecuencia con que esta primera etiqueta es seguida por la segunda según la siguiente fórmula:

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1},t_i)}{C(t_{i-1})}$$

En el corpus WSJ, por ejemplo, los MD aparecen 13 124 veces, de las cuales son seguidos por un VB 10 471 veces, lo cual resulta en una estimación de máxima probabilidad de:

$$P(VB|MD) = \frac{C(MD,VB)}{C(MD)} = \frac{10471}{13124} = .80$$

Probabilidades de emisión

Las probabilidades de emisión $P\left(w_1^n|t_1^n\right)$ representan la probabilidad de que, dada una etiqueta (digamos MD), esta se asocie con una palabra concreta (digamos *«will»*). La estimación de **máxima verosimilitud** de la probabilidad de emisión en general se define como:

$$P(w_i|t_i) = \frac{C(t_i,w_i)}{C(t_i)}$$

De los 13 124 casos de MD en el corpus WSJ, estas se asocian o refieren a «will» en 4046 ocasiones, por tanto:

$$P(will|MD) = \frac{C(MD,will)}{C(MD)} = \frac{4046}{13124} = .31$$

Como aclaración, hemos de decir que esta probabilidad no se refiere a cuál es la etiqueta más probable para la palabra *«will»,* ya que esta sería la probabilidad *a*

posteriori

$P(MD \vee will)$. En su lugar,

$P(will \vee MD)$ responde a una pregunta ligeramente menos intuitiva, concretamente si vamos a generar una etiqueta MD, ¿qué probabilidades hay de que este MD sea «*will*»?

Los dos tipos de probabilidades mencionados anteriormente, la probabilidad de transición

$P(VB \vee MD)$ y la probabilidad de emisión

$P(will \vee MD)$, se corresponden al conjunto A de probabilidades de transición del HMM y al conjunto B de probabilidades de observación *B* del HMM.

La Figura 2 ilustra algunas de las probabilidades de transición A para tres estados en un etiquetador morfosintáctico HMM; el etiquetador completo tendría un estado para cada etiqueta. En esta imagen, las probabilidades de transición A se utilizan para calcular la probabilidad *a priori*.

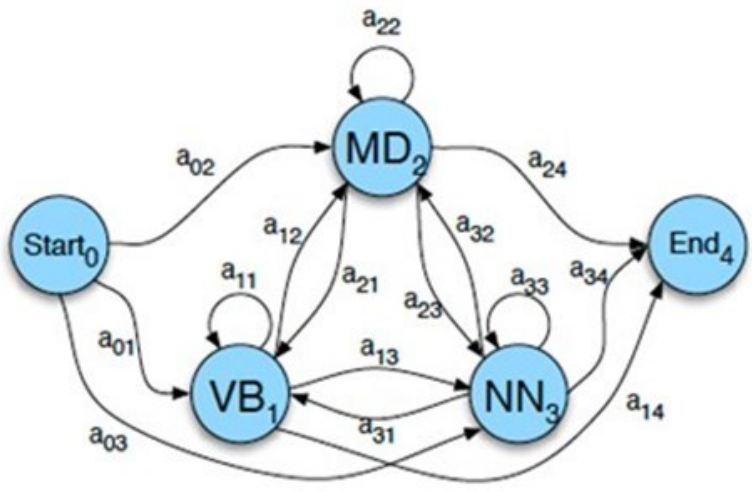


Figura 2. Cadena de Markov correspondiente a los estados ocultos del HMM. Fuente: Jurafsky y Martin, 2009.

La Figura 3 muestra otra vista de estos tres estados, pero centrándose en algunas de las probabilidades de observación B de cada palabra. Cada estado oculto está asociado con un vector de probabilidades para cada palabra en observación.

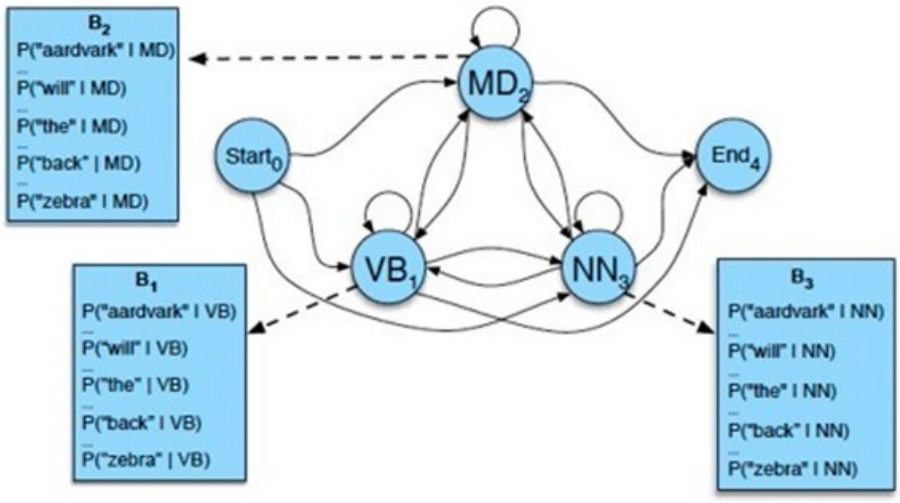
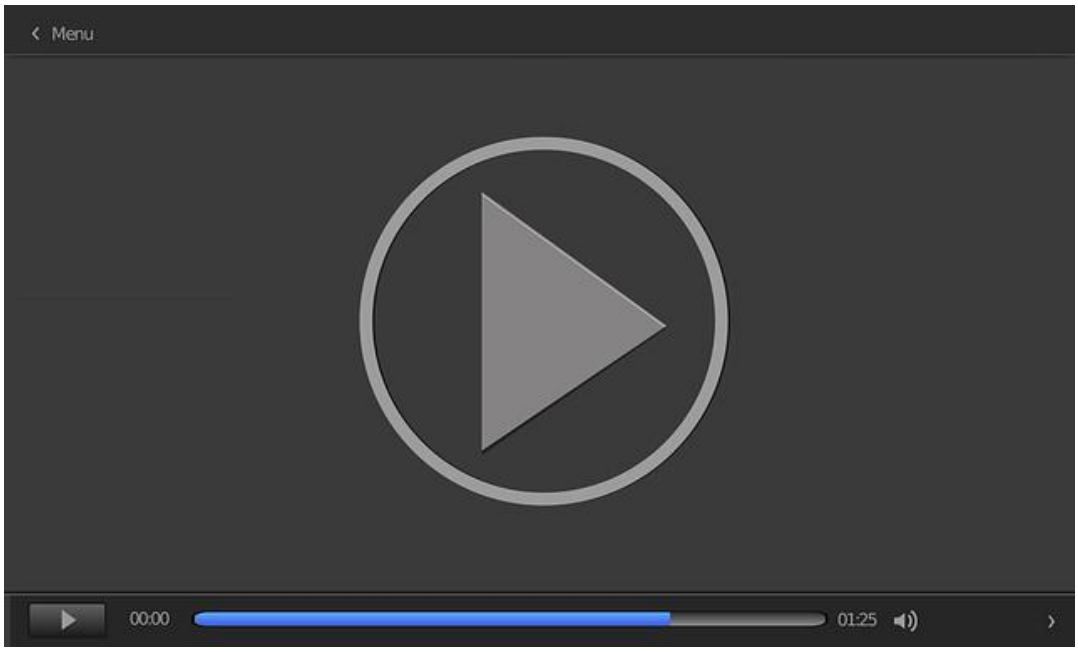


Figura 3. Probabilidades de observación B para el HMM de la Figura 2. Fuente: Jurafsky y Maritn, 2009.

Como indicábamos sobre esta última imagen, cada estado está asociado con un vector de probabilidades con una probabilidad para cada posible palabra en observación, a excepción de los estados no emisores de inicio y fin.

Ideas clave

En el vídeo *Construcción de un etiquetador morfosintáctico basado en un HMM bigrama a partir de un corpus etiquetado* se explicará cómo crear un etiquetador morfosintáctico a partir de un corpus etiquetado.



03.02. Construcción de un etiquetador morfosintáctico basado en un hmm bigrama a partir de un corpus etiquetado

Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=7b463f3e-fc41-4140-acd8-ac6b00a4b017

Finalmente, vamos a trabajar a través de un ejemplo de cálculo de la mejor secuencia de etiquetas que corresponde a la siguiente secuencia de palabras:

- Janet will back the Bill (en español, «Janet respaldará la ley»).

La secuencia de etiquetas correcta es:

```
Janet/NNP will/MD back/VB the/DT bill/NN
```

Sea el modelo HMM el definido por el conjunto A de probabilidades de transición (Figura 4), y el conjunto B de probabilidades de observación (Figura 5). Cada elemento

a_{ij} del conjunto A describe la probabilidad de transitar de un estado oculto

i (etiqueta

i) a otro estado oculto

j (etiqueta

j). Cada elemento

$b_i(o_t)$ describe la probabilidad de observar las palabras dadas las etiquetas.

	PNN	MD	VB	JJ	NN	RB	DT
<*S*>	0.2767	0,0006	0,0031	0,0453	0,0449	0,0510	0.2026
PNN	0.3777	0,0110	0,0009	0,0084	0,0584	0,0090	0,0025
MD	0,0008	0,0002	0.7968	0,0005	0,0008	0.1698	0,0041
VB	0,0322	0,0005	0,0050	0,0837	0,0615	0,0514	0.2231
JJ	0,0366	0,0004	0,0001	0,0733	0.4509	0,0036	0,0036
NN	0,0096	0,0176	0,0014	0,0086	0.1216	0,0177	0,0068
RB	0,0068	0,0102	0.1011	0.1012	0,0120	0,0728	0,0479
DT	0.1147	0,0021	0,0002	0.2157	0.4744	0,0102	0,0017

Figura 4. Conjunto A de probabilidades de transición

$P(t_i|t_{i-1})$ calculadas a partir del corpus WSJ. Fuente: Jurafsky y Martin, 2009.

En la imagen anterior vemos que cada fila representa el evento condicionante; por ejemplo:

$$P(VB \vee MD) = 0.7968.$$

La siguiente imagen (Figura 5) se obtiene a partir del recuento de apariciones de una palabra en el corpus. Así:

- La palabra «Janet» solo aparece como un nombre propio (NNP).
- La palabra *«will»* aparece en el corpus como tres categorías gramaticales diferentes, puede ser:
  - Un verbo modal (MD) para generar el futuro de los verbos en inglés.
  - Un verbo (VB) que significa «desear» en español.
  - O un nombre (NN) que significa «deseo» o «voluntad» en español.

	Janet	will	back	the	bill
PNN	0.000032	0	0	0.000048	0
MD	0	0.308431	0	0	0
VB	0	0.000028	0.000672	0	0.000028
JJ	0	0	0.000340	0.000097	0
NN	0	0.000200	0.000223	0.000006	0.002337
RB	0	0	0.010446	0	0
DT	0	0	0	0.506099	0

Figura 5. Conjunto B de probabilidades de observación calculadas a partir del corpus WSJ. Fuente: Jurafsky y Martin, 2009.

La Figura 6 muestra un esquema con las posibles etiquetas para el ejemplo anterior, así como la correcta secuencia de etiquetado final. Esta secuencia del etiquetado morfosintáctico se ha calculado aplicando el **algoritmo de Viterbi,** cuyo pseudocódigo se presenta en la Figura 7.

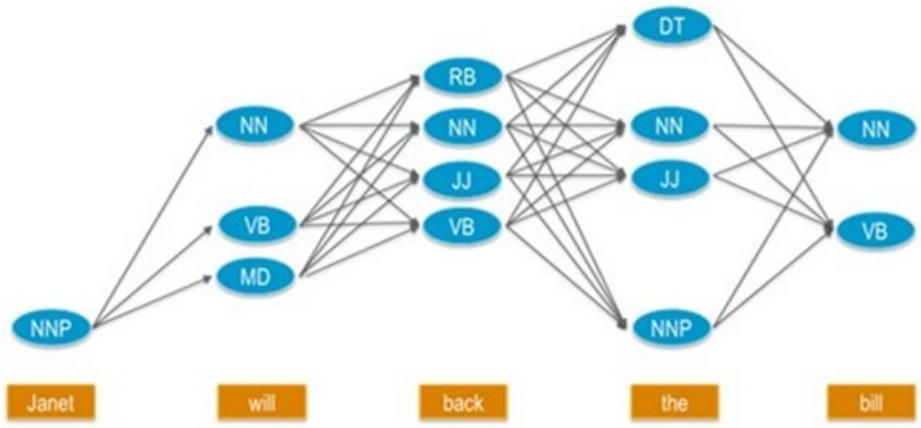


Figura 6. Diagrama de la tarea de etiquetado para la frase ejemplo. Fuente: Jurafsky y Martin, 2009.

function VITERBI(*observations* of len T,*state-graph* of len N) **returns** *best-path*

create a path probability matrix *viterbi*$[N+2,T]$
for each state s **from** 1 **to** N **do** ; initialization step
viterbi$[s,1] \leftarrow a_{0,s} * b_s(o_1)$
backpointer$[s,1] \leftarrow 0$
for each time step t **from** 2 **to** T **do** ; recursion step
for each state s **from** 1 **to** N **do**
viterbi$[s,t] \leftarrow \max_{s'=1}^{N}$ *viterbi*$[s',t-1] * a_{s',s} * b_s(o_t)$
backpointer$[s,t] \leftarrow \operatorname{argmax}_{s'=1}^{N}$ *viterbi*$[s',t-1] * a_{s',s}$
viterbi$[q_F,T] \leftarrow \max_{s=1}^{N}$ *viterbi*$[s,T] * a_{s,q_F}$; termination step
backpointer$[q_F,T] \leftarrow \operatorname{argmax}_{s=1}^{N}$ *viterbi*$[s,T] * a_{s,q_F}$; termination step
return the backtrace path by following backpointers to states back in time from *backpointer*$[q_F,T]$

Figura 7. Pseudocódigo del algoritmo de Viterbi para encontrar la secuencia óptima de *tags* en un etiquetador morfosintáctico basado en HMM. Fuente: Jurafsky y Martin, 2009.

El algoritmo de Viterbi crea una matriz de probabilidades con una columna para cada observación

t y una fila para cada estado

q_i de la máquina de estados finitos (o autómata finito) que representa el HMM. Para el ejemplo, el algoritmo crea N = 5 columnas de estado, la primera para la observación de la primera palabra «Janet», la segunda para «will» y así sucesivamente hasta completar las cinco palabras que conforman la frase, tal como

se muestra en la Figura 8.

Se empieza en la primera columna para establecer el valor de Viterbi en cada celda $v_t(i)$, que se calcula como el producto de la probabilidad de transición (hasta ese estado desde el estado inicial) por la probabilidad de observación (de la primera palabra). Entonces, se avanza columna a columna y, para cada estado en la columna 1, se calcula la probabilidad de transición a cada estado en la columna 2, y así sucesivamente.

Para cada estado q_i en el tiempo t, se calcula valor de Viterbi (representado como *viterbi[s,t]* en el pseudocódigo de la Figura 7) tomando el máximo sobre las extensiones de todas las rutas que conducen a la celda actual, según la siguiente ecuación:

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i)\, a_{ij} b_j(o_t)$$

Donde:

- $v_{t-1}(i)$ es la probabilidad de la ruta de Viterbi previa, es decir, la ruta para el tiempo anterior o en la observación $t-1$.
- a_{ij} es la probabilidad de transición del estado anterior q_i al estado actual q_j.
- $b_j(o_t)$ es la probabilidad de observación del símbolo o_t dado el estado actual q_j.

Entonces, cada celda de la primera columna donde aparece la palabra «Janet» se calcula multiplicando la probabilidad de Viterbi previa en el estado de inicio q_0 , que es $v_0(0) = 1.0$ por la probabilidad de transición desde el estado de inicio q_0 hasta la etiqueta para esa celda, por ejemplo:

$P(NNP \vee start) = 0.2767$ para la celda en la que la etiqueta es NNP

Y por la probabilidad de observación de la palabra «Janet» dada la etiqueta de esa celda, por ejemplo:

$P(Janet \vee NNP) = 0.000032$ para la celda en la que la etiqueta es NNP.

Por lo tanto:

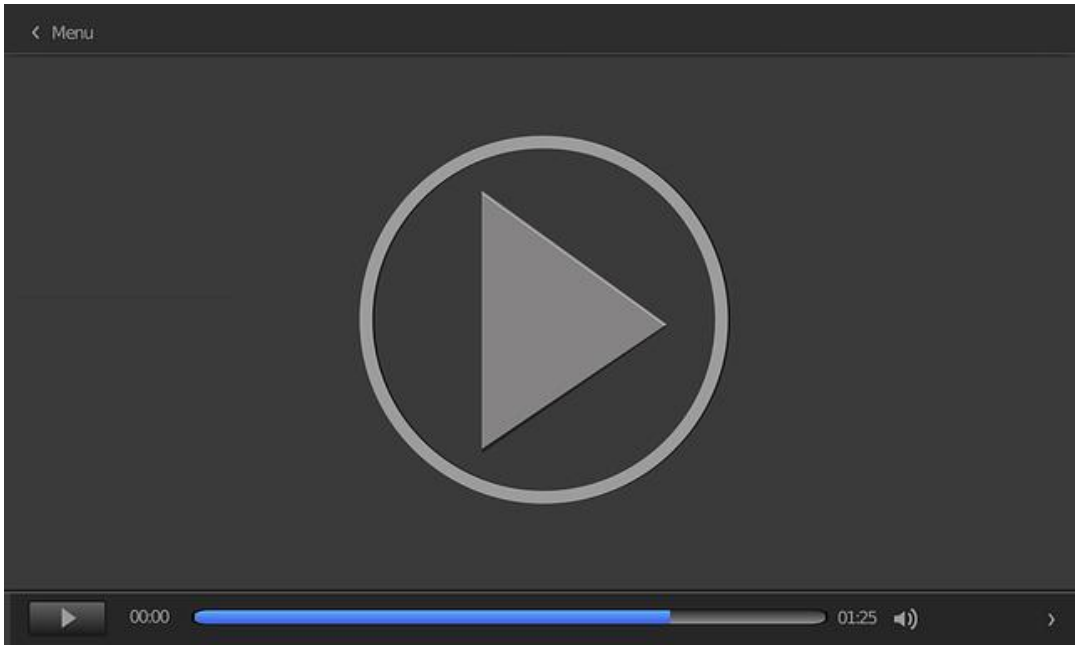
- $v_1(1) = 1.0 \cdot 0.2767 \cdot 0.000032 = 0.000009$ para la celda en la que la etiqueta es NNP para la columna donde la palabra es «Janet».
- El resto de las celdas en esta columna son cero, ya que la palabra «Janet» no puede tener asociada ninguna de las otras etiquetas gramaticales.

A continuación, cada celda en la columna *«will»* se actualiza con la ruta de probabilidad máxima desde la columna anterior. En la Figura 8 se muestran los valores para las celdas MD, VB y NN. Cada celda obtiene los siete valores de la columna anterior multiplicados por la probabilidad de transición apropiada; se toma el valor máximo, $v_1(1) \cdot P(MD \mid NNP)$, que proviene del estado NNP en la columna anterior y este valor se multiplica por la probabilidad de observación del símbolo *«will»* dada la etiqueta correspondiente a la celda en cuestión. Por ejemplo, para la celda MD el valor de Viterbi es:

$$v_2(2) = v_1(1) \cdot P(MD \vee NNP) \cdot P(will \vee MD) = 0.00000002772$$

Se continúa aplicando el algoritmo columna a columna hasta llegar al final.

En el vídeo *Creación de la matriz de probabilidades de la ruta de Viterbi* se verá cómo realizar la matriz de Viterbi para llevar a cabo el etiquetado morfosintáctico de una oración.



03.03. Creación de la matriz de probabilidades de la ruta de Viterbi

Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=5a01bda1-9564-443e-be54-ac6b00c6a1ed

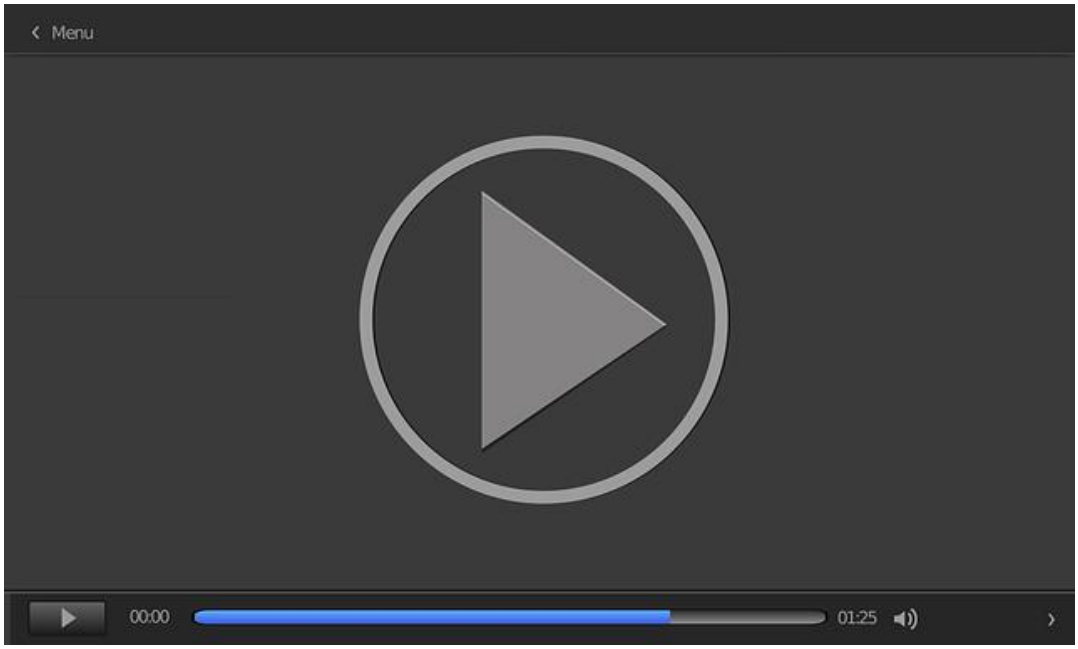
Para la probabilidad máxima de Viterbi, se traza la inversa para obtener la ruta concreta que ha llevado a ese valor y que se corresponde con la mejor secuencia de etiquetas: NNP MD VB DT NN.

Entonces, esta secuencia de etiquetas se corresponde con el etiquetado morfosintáctico de la frase:

```
Janet/NNP will/MD back/VB the/DT bill/NN
```

Ideas clave

En el vídeo *Obtención de la ruta de Viterbi con máxima probabilidad* se verá cómo obtener la ruta de Viterbi, lo que significa que cuando apliquemos el logaritmo seamos capaces de decodificar y obtener la ruta más probable que nos dé el mejor etiquetado morfosintáctico de una oración.



03.04. Obtención de la ruta de Viterbi con máxima probabilidad

Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=150bec3e-c01f-48ef-810f-ac6b00c1e1ac

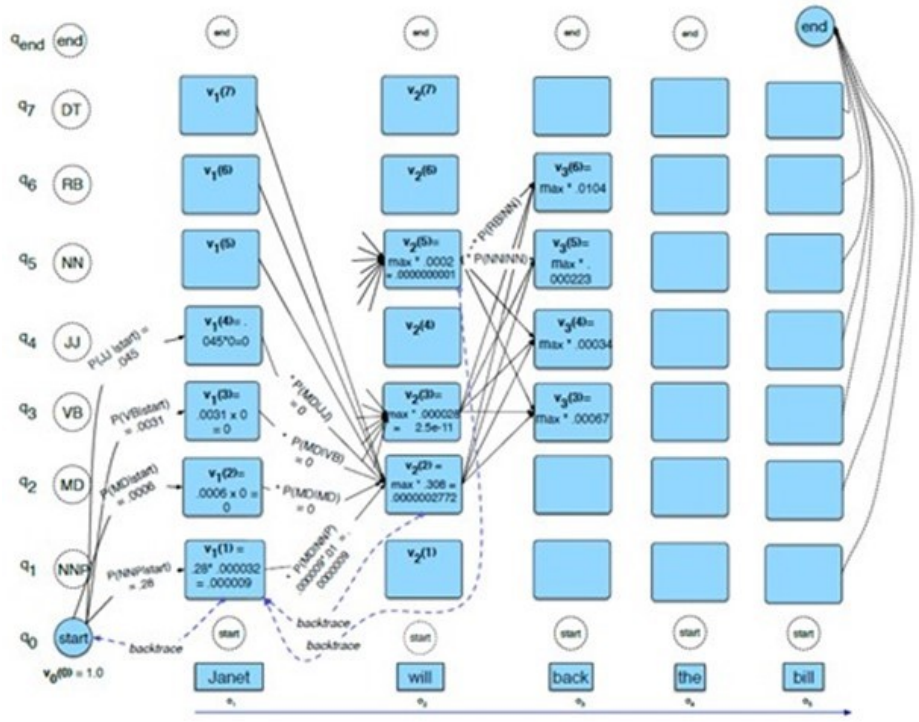


Figura 8. Matriz para la aplicación del algoritmo de Viterbi en el etiquetado morfosintáctico de la frase «Janet will back the bill». Fuente: Jurafsky y Martin, 2009

3.5. Etiquetado morfosintáctico basado en aprendizaje automático

Los modelos más vanguardistas de etiquetado morfosintáctico utilizan diversas técnicas de aprendizaje automático tanto supervisado como no supervisado. Ejemplos de modelos supervisados son, entre otros:

- El algoritmo perceptrón (Collins, 2002).
- El modelo logaritmo lineal bidireccional (Toutanova, Klein, Manning y Singer, 2003).
- Las máquinas de vectores soporte (Giménez y Márquez, 2004).

En estos casos de aprendizaje supervisado, el modelo recibe como entrada una serie de parámetros extraídos del texto, y tiene como salida la predicción del POS tag de una palabra en concreto. A modo de ejemplo, en (Giménez y Márquez, 2004) se propone el uso de las siguientes variables de entrada:

word features	$w_{-3}, w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}, w_{+3}$
POS features	$p_{-3}, p_{-2}, p_{-1}, p_0, p_{+1}, p_{+2}, p_{+3}$
ambiguity classes	a_0, a_1, a_2, a_3
may_be's	m_0, m_1, m_2, m_3
word bigrams	$(w_{-2}, w_{-1}), (w_{-1}, w_{+1}), (w_{-1}, w_0)$ $(w_0, w_{+1}), (w_{+1}, w_{+2})$
POS bigrams	$(p_{-2}, p_{-1}), (p_{-1}, a_{+1}), (a_{+1}, a_{+2})$
word trigrams	$(w_{-2}, w_{-1}, w_0), (w_{-2}, w_{-1}, w_{+1}),$ $(w_{-1}, w_0, w_{+1}), (w_{-1}, w_{+1}, w_{+2}),$ (w_0, w_{+1}, w_{+2})
POS trigrams	$(p_{-2}, p_{-1}, a_{+0}), (p_{-2}, p_{-1}, a_{+1}),$ $(p_{-1}, a_0, a_{+1}), (p_{-1}, a_{+1}, a_{+2})$
sentence_info	punctuation ('.', '?', '!')
prefixes	$s_1, s_1 s_2, s_1 s_2 s_3, s_1 s_2 s_3 s_4$
suffixes	$s_n, s_{n-1} s_n, s_{n-2} s_{n-1} s_n, s_{n-3} s_{n-2} s_{n-1} s_n$
binary word features	intial Upper Case, all Upper Case, no initial Capital Letter(s), all Lower Case, contains a (period / number / hyphen ...)
word length	integer

Figura 9. Variables de entrada para predecir los POS tags con un modelo de aprendizaje supervisado.

Fuente: Giménez y Márquez, 2004.

Como se puede ver, se generan variables de entrada que hacen referencia tanto a la palabra en sí sobre la que se quiere predecir el POS tag, como a palabras de su contexto, tanto previas o posteriores. Se incluyen también variables que hacen referencia a los posibles POS tags que tiene la palabra en cuestión y las palabras previas o posteriores.

No obstante, tanto estos como los algoritmos presentados en el tema dependen en gran medida del dominio de datos de entrenamiento, así como el etiquetado marcado por los expertos.

Algunos trabajos recientes en etiquetado morfosintáctico se centran en buscar alternativas que permitan relajar estas condiciones, es el caso de los algoritmos no supervisados que etiquetan clústers de palabras en clases morfosintácticas (Christodoulopoulos, Goldwater, y Steedman, 2010) (Sirts, Eisenstein, Elsner, y Goldwater, 2014).

Muchos algoritmos se basan en la combinación de datos etiquetados con datos no etiquetados, por ejemplo, usando **coentrenamiento** (Søgaard, 2010). La asignación de etiquetas a texto de muy distintos tipos como, por ejemplo, aquellos provenientes de *Twitter,* puede requerir la adición de nuevas etiquetas para las direcciones URL (URL), nombre de usuario menciona (USR), *retweets* (RT), y *hashtags* (HT). La normalización de las palabras no estándar y técnicas *bootstrapping* para emplear datos no supervisados (Derczynski et al., 2013).

En estos ejemplos se lleva a cabo un **etiquetado por clasificación**, donde el objetivo es, dada una secuencia, predecir una a una las etiquetas de sus palabras, considerando información de la propia palabra y del contexto. Esto se refleja en la siguiente ecuación, donde f() es la función que relaciona las variables de entrada para una determinada palabra con la salida predicha con la estructura (*token, POS tag*):

$$f(w = \textit{they can fish}, m = 2) = (\textit{can}, V)$$

Finalmente, además de estos modelos de aprendizaje automático, existen otros algoritmos estadísticos muy populares para tareas de POS tagging, como es el caso del algoritmo Conditional Random Field (CRF) (Lafferty, 2001). También existen métodos basados en la generación de reglas, como es el caso del algoritmo de Brill para POS tagging (Brill, 1992).

3.6. Named Entity Recognition

Además de poder obtener el POS Tag de una palabra mediante el uso de algoritmos estadísticos o técnicas de aprendizaje automático, también es posible identificar de entre los distintos sustantivos a qué tipo de entidad nombrada hace referencia (a una persona, a una organización o a una ubicación, por ejemplo).

Esta tarea se denomina **NER *(named-entity recognition*),** y con ella se suelen identificar dentro de un texto categorías como las siguientes (a modo de ejemplo, ya que se pueden usar más categorías):

- PER: categoría de personas, como por ejemplo el nombre de alguien (ej.: Frodo Baggins)
- GPE: categoría para identificar países, ciudades o estados (ej.: Madrid).
- LOC: categoría para identificar ubicaciones concretas (ej.: Vesubio).
- ORG: categoría para identificar organizaciones o empresas (ej.: Microsoft).
- MONEY: categoría para identificar referencias a dinero (ej.: $ 5).
- DATE: categoría para identificar fechas (ej.: 05/02/2021 o «jueves»).

De esta manera, una frase en la que se ha llevado a cabo un NER daría como resultado:

[PER Miguel López], de [ORG Microsoft], ha ido a una conferencia a [GPE Londres].

Existen distintas maneras de representar las NE (*named entity*) dentro de una frase, donde aparecen representaciones para, incluso, diferenciar qué token representa el comienzo de una NE, cuáles están en medio y cuáles están al final para casos en los que varios tokens estén asociados a una misma NE.

Words	IO Label	BIO Label	BIOES Label
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O

Figura 10. Distintas representaciones de las NE. Fuente: Jurafksy y Martin, 2009.

De igual manera que ocurre con los algoritmos de POS tagging, la identificación de las NER se puede realizar:

- Con **diccionarios** donde se tengan ya identificadas palabras (o combinaciones de palabras) junto con su NE.
- Con **sistemas basados en reglas** que identifiquen las NE en base a ciertos patrones.
- Con **modelos estadísticos o de aprendizaje automático,** que en base a un entrenamiento previo y a una serie de variables que modelen el contexto de las palabras, puedan predecir qué token o tokens hacen referencia a una NER.

3.7. Referencias bibliográficas

Brill, E. (1992). *A simple rule-based part of speech tagger.* PENNSYLVANIA UNIV PHILADELPHIA DEPT OF COMPUTER AND INFORMATION SCIENCE.

Christodoulopoulos, C., Goldwater, S. y Steedman, M. (2010). Two decades of unsupervised POS induction: How far have we come? En *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 575-584). Association for Computational Linguistics.

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1-8). Association for Computational Linguistics.

Derczynski, L., Ritter, A., Clark, S. y Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. En *Proceedings of Recent Advances in Natural Language Processing* (pp. 198-206). Association for Computational Linguistics.

Eisenstein, J. (2019). Introduction to Natural Language Processing (pp. 145-148). MIT Press Ltd.

Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE, 61*(3), 268-278.

Giménez, J. y Márquez, L. (2004). SVMTool: A general POS tagger generator based on Support Vector Machines. En *Proceedings of the 4th International Conference on Language Resources and Evaluation.* LREC.

Jurafsky, D. y Martin, J. H. (2009*). Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition and Computational Linguistics.* Prentice-Hall.

Marcus, M. P., Santorini, B. y Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics, 19*(2), 313-330.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. En *Proceedings of the Eighteenth International Conference on Machine Learning,* pages 282–289

RAE. (s. f.). Morfosintaxis. En *Diccionario de la lengua española* (actualización de la 23ª ed.). http://dle.rae.es/?id=PpC0akB

Sirts, K., Eisenstein, J., Elsner, M., y Goldwater, S. (2014). POS induction with distributional and morphological information using a distance-dependent Chinese restaurant process. En *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 265-271). Baltimore, Estados Unidos. http://www.aclweb.org/anthology/P14-2044

Søgaard, A. (2010). Simple semi-supervised training of part-of-speech taggers. En *Proceedings of the ACL 2010 Conference Short Papers* (pp. 205-208). Uppsala, Suecia: Association for Computational Linguistics.

Toutanova, K., Klein, D., Manning, C. D. y Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *En Proceedings of HLT-NAACL* (pp. 173-180). Association for Computational Linguistics.

Etiquetado morfosintáctico basado en aprendizaje no supervisado

Christodoulopoulos, C., Goldwater, S. y Steedman, M. (2010). Two decades of unsupervised POS induction: How far have we come? En *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 575-584). Association for Computational Linguistics. https://homepages.inf.ed.ac.uk/sgwater/papers/emnlp10-20yrsPOS.pdf

El artículo presenta la evaluación de siete algoritmos que realizan el etiquetado morfosintáctico basándose en técnicas de aprendizaje no supervisado. Sorprendentemente, se muestra que los primeros algoritmos que aparecieron hace ya más de veinticinco años proporcionan mejores resultados que otros más recientes.

Analizador morfológico automático de la Biblioteca Virtual Miguel de Cervantes

Analizador morfológico automático. http://data.cervantesvirtual.com/analizador-sintactico-automatico

Esta aplicación de la Biblioteca Virtual Miguel de Cervantes permite introducir un texto en español y realizar el análisis morfosintáctico de forma automática, por lo que se logra identificar la categoría gramatical de cada palabra.

1. Indica las afirmaciones correctas sobre la morfosintaxis:

A. Determina las diferentes partes de la oración.

B. Estudia la estructura de las palabras.

C. Parte de la gramática que aúna la morfología y la sintaxis.

D. Estudia el modo en que se combinan las palabras.

2. La clasificación de las diferentes palabras según su tipo o clase se conocen como:

A. Categorías morfosintácticas.

B. Categorías gramaticales.

C. Partes de la oración.

D. *Part-of-speech* (POS) en inglés.

3. Indica las afirmaciones correctas sobre la probabilidad de transición del HMM utilizado en el etiquetado morfosintáctico:

A. Representa la probabilidad de la etiqueta anterior dada una etiqueta.

B. Representa la probabilidad de una etiqueta dada la etiqueta anterior.

C. La estimación de máxima verosimilitud de una probabilidad de transición se calcula como la división entre el recuento de las veces que vemos la primera etiqueta en un corpus etiquetado entre la frecuencia con que la primera etiqueta es seguida por la segunda.

D. La estimación de máxima verosimilitud de una probabilidad de transición se calcula como la división entre el recuento de las veces que vemos la primera etiqueta seguida por la segunda en un corpus etiquetado entre la frecuencia con que aparece la primera etiqueta.

4. Indica las afirmaciones correctas sobre el Penn Treebank:

A. Se utiliza en el etiquetado morfosintáctico en inglés.

B. Consiste en un conjunto de etiquetas gramaticales.

C. Etiqueta los signos de puntuación de la frase.

D. Etiqueta todos los verbos con una misma etiqueta.

5. Indica las afirmaciones correctas sobre el etiquetado morfosintáctico:

A. Su salida es la secuencia de etiquetas de las categorías gramaticales.

B. Se llama POS *tagging (part-of-speech* tagging en inglés).

C. Asigna etiquetas sobre la categoría gramatical a cada una de las palabras de la oración.

D. Su entrada solo es una secuencia de palabras.

6. Indica las afirmaciones correctas sobre un modelo oculto de Markov (HMM):

A. Es una máquina de estados finitos.

B. Es un modelo estadístico.

C. Su objetivo es determinar parámetros desconocidos a partir de parámetros observables.

D. Es un autómata finito donde las transacciones entre estados son probabilísticas.

7. Indica las afirmaciones correctas sobre un etiquetador morfosintáctico basado en HMM:

A. Modelo de lenguaje estadístico que permite obtener la secuencia de etiquetas gramaticales que tenga mayor probabilidad para una frase.

B. Se entrena con un conjunto de frases en la que cada palabra está anotada con una etiqueta describiendo su categoría gramatical.

C. Se fijan las estimaciones de máxima probabilidad para cada una de las condiciones de la máquina de estados finitos a partir de los datos de entrenamiento.

D. Se usa el algoritmo de decodificación de Viterbi para estimar las probabilidades.

8. Indica las afirmaciones correctas sobre el algoritmo implementado por un etiquetador morfosintáctico:

A. No se ve afectado por problemas de ambigüedad gramatical de las palabras.

B. Aplica un proceso de desambiguación.

C. Etiqueta directamente la oración sin realizar ningún procesado previo.

D. Aplica un proceso de obtención de tokens.

9. Indica las afirmaciones correctas sobre la probabilidad de emisión del HMM utilizado en el etiquetado morfosintáctico:

A. Permite identificar la palabra más probable para una etiqueta dada.

B. Representa la probabilidad de que, dada una palabra, esta se asocie a una etiqueta concreta.

C. Representa la probabilidad de que, dada una etiqueta, esta esté asociada a una palabra concreta.

D. Permite identificar la etiqueta más probable para una palabra dada.

10. Indica las afirmaciones correctas sobre la aplicación del algoritmo de Viterbi para obtener la secuencia de etiquetas más probables en un etiquetador morfosintáctico HMM:

A. Cada columna en la matriz de probabilidades corresponde a una palabra en la frase a analizar y se llama observación.

B. Cada celda en una columna de la matriz de probabilidades corresponde a un estado de la máquina de estados finitos y se corresponde a una etiqueta morfosintáctica.

C. La probabilidad de Viterbi de una ruta se calcula como la multiplicación de probabilidad de la ruta de Viterbi previa por la probabilidad de transición del estado anterior al estado actual por la probabilidad de observación del símbolo (etiqueta morfosintáctica), dado el estado actual.

D. La secuencia de etiquetas morfosintáctica correctas se obtiene de trazar la rutan inversa que ha llevado a obtener el valor de Viterbi máximo en el estado final.