
Universidad
Tecmilenio®

Universidad Tecmilenio

Procesamiento de lenguaje natural y visión computacional

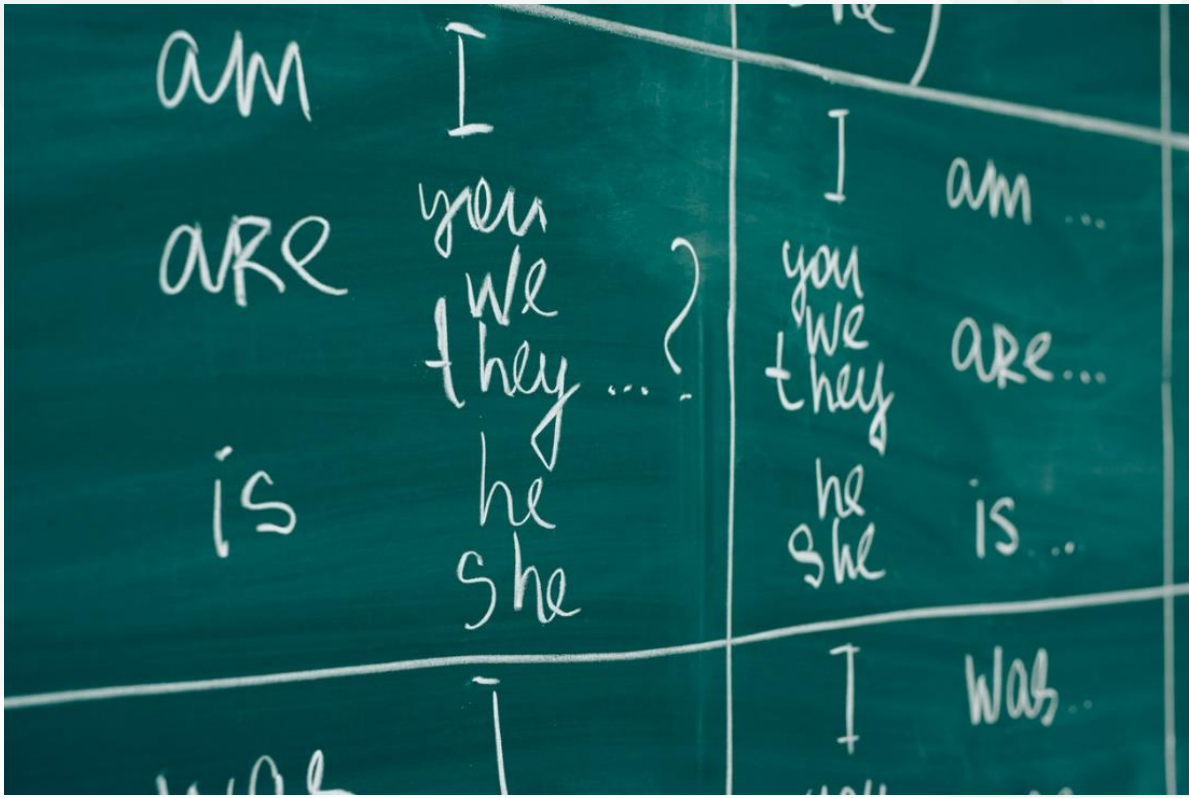
Etiquetas de secuencias

Introducción

El etiquetado de las partes de una oración es un proceso popular dentro del procesamiento del lenguaje natural y es utilizado para categorizar palabras en un texto. Una etiqueta POS, por tanto, indica la parte gramatical de una palabra y son extremadamente útiles, ya que brindan una señal lingüística sobre cómo se usa una palabra dentro del alcance de una frase, oración o documento.

En este tema conocerás lo siguiente:

- El etiquetado POS.
- Uso de las etiquetas POS.



Explicación

El etiquetado POS requiere previamente de *tokenizar* las palabras, incluyendo signos de puntuación. El etiquetado es también una tarea de desambiguación, ya que el significado de las palabras es confuso y estas pueden ser etiquetadas en más de una parte de la oración.

Uno de los corpus con etiquetado POS más utilizados es el *Penn Treebank* para el idioma inglés (Marcus, Santorini y Marcinkiewicz, 1993). Las 36 etiquetas que lo conforman son las siguientes:

1. CC Coordinating conjunction
2. CD Cardinal number
3. DT Determiner
4. EX Existential there
5. FW Foreign word
6. IN Preposition or subordinating conjunction
7. JJ Adjective
8. JJR Adjective, comparative
9. JJS Adjective, superlative
10. LS List item marker
11. MD Modal
12. NN Noun, singular or mass
13. NNS Noun, plural
14. NP Proper noun, singular
15. NPS Proper noun, plural
16. PDT Predeterminer
17. POS Possessive ending
18. PP Personal pronoun
19. PP$ Possessive pronoun
20. RB Adverb
21. RBR Adverb, comparative
22. RBS Adverb, superlative
23. RP Particle
24. SYM Symbol
25. TO to
26. UH Interjection
27. VB Verb, base form
28. VBD Verb, past tense
29. VBG Verb, gerund or present participle
30. VBN Verb, past participle
31. VBP Verb, non-third person singular present
32. VBZ Verb, third person singular present
33. WDT Wh-determiner
34. WP Wh-pronoun
35. WP$ Possessive wh-pronoun
36. WRB Wh-adverb

Explicación

El objetivo del etiquetado de secuencias es asignar etiquetas a palabras, específicamente, asignar etiquetas discretas a elementos discretos en una secuencia. Las etiquetas discretas son categorías gramaticales.
Un ejemplo de un etiquetado POS para la oración “La acción de gran valor se desplomó durante el transcurso de la semana”, utilizando las etiquetas de Penn Treebank sería:

```
La/DT  acción/NN  de/IN  gran valor/JJ  se-desplomó/VBD
durante/IN  el/DT  transcurso/JJ  de/IN  la/DT  semana/NN
```

Los pasos para etiquetar la secuencia de palabras son los siguientes:
Tokenizar el texto.
Aplicar etiquetas POS.

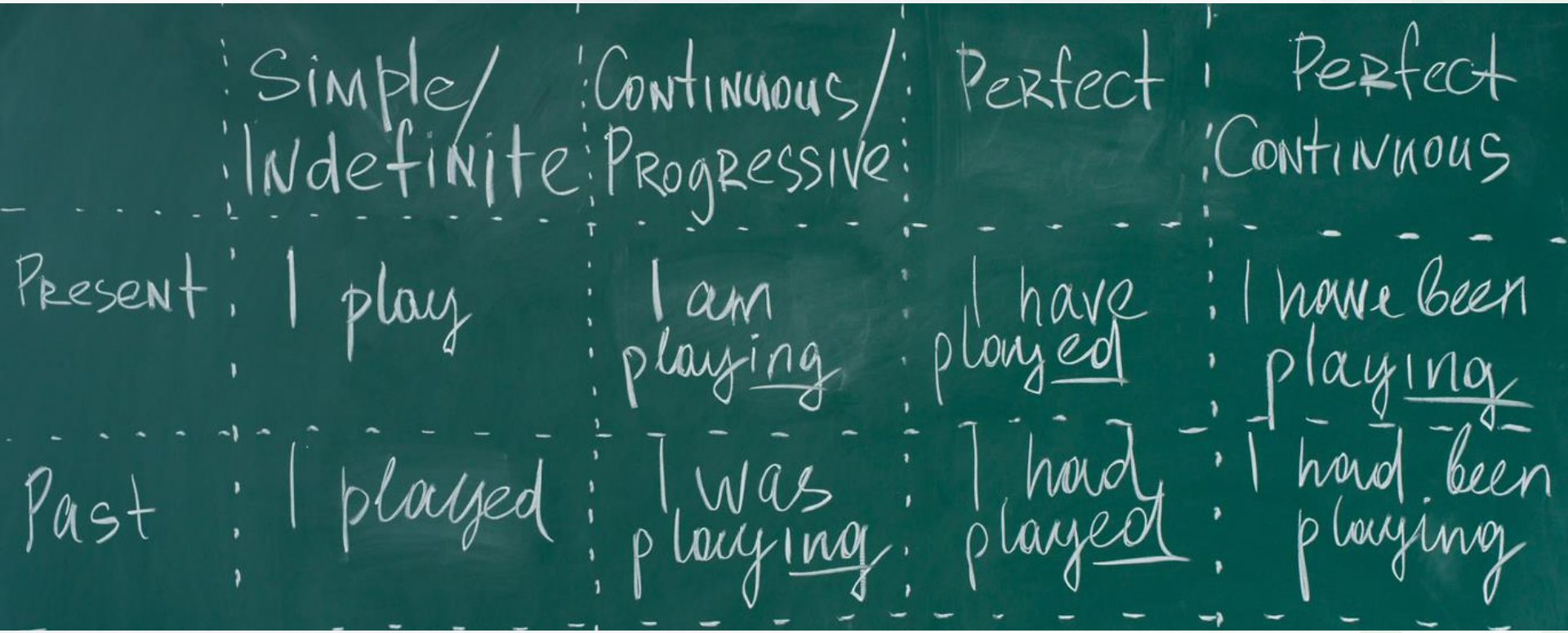
Actividad

01 Realiza ejercicios prácticos que incluyan la identificación de distintas categorías gramaticales en un texto como el Corpus Brown.

02 Investiga las etiquetas para español como UAM y RST treebank, entre otros.

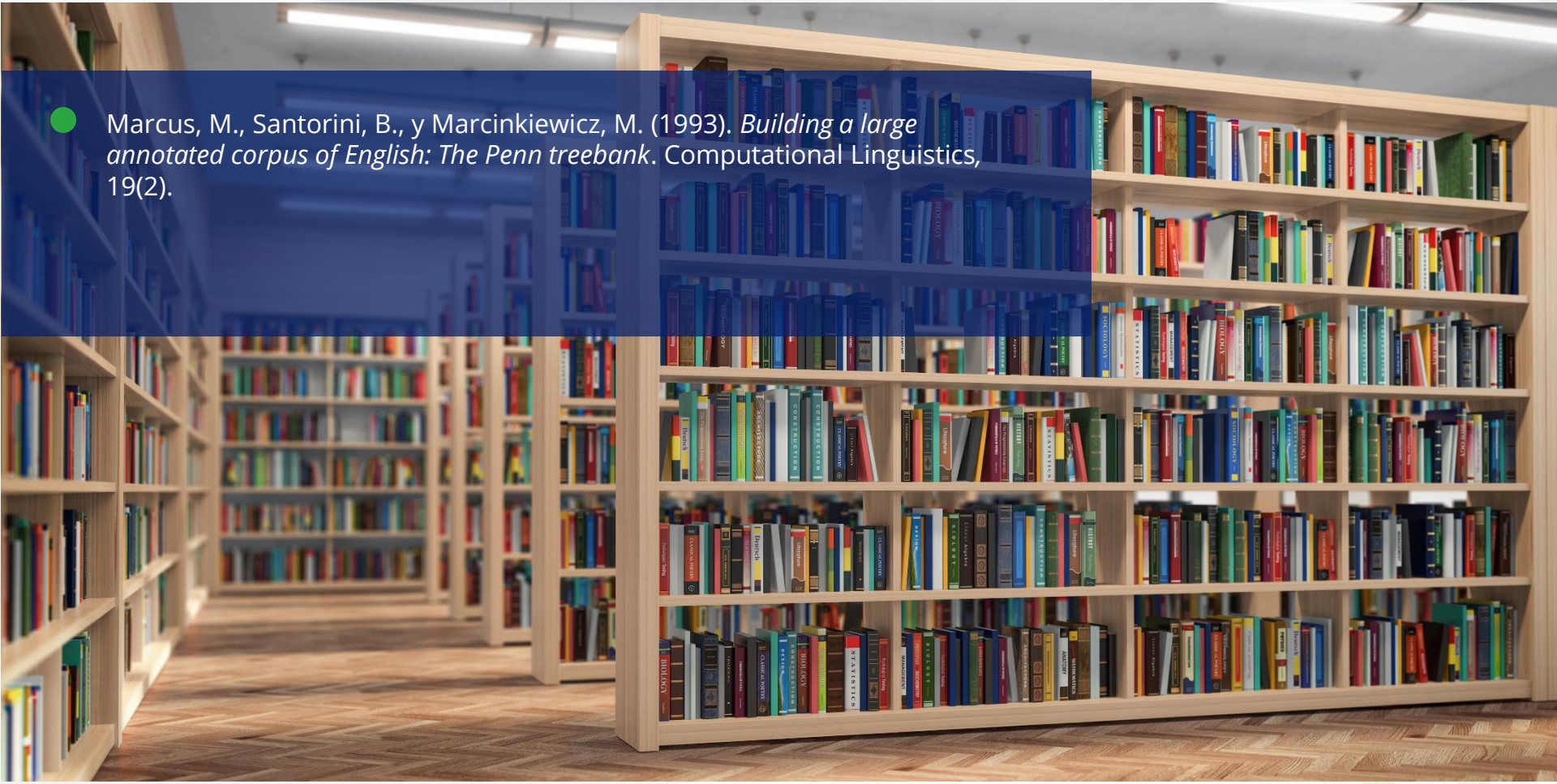
Cierre

Las palabras se pueden agrupar en clases, como sustantivos, verbos, adjetivos y adverbios. Las clases se conocen como categorías gramaticales o partes de una oración y el proceso que asigna estas categorías a un texto se conoce como etiquetado POS. Es posible encontrar distintas variedades de etiquetas POS, ya sea con base en unigramas, expresiones regulares o n-gramas, incluso palabras etiquetadas en función de la pronunciación o utilizando atributos morfológicos.



Bibliografía

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



Universidad Tecmilenio

Procesamiento de lenguaje natural y visión computacional

Modelo oculto de Markov

Introducción

El modelo oculto de Markov (HMM, por sus siglas en inglés) es una técnica estocástica utilizada para crear etiquetas POS; es un modelo ampliamente utilizado en otro tipo de aplicaciones en aprendizaje por refuerzo, reconocimiento de patrones, bioinformática y otras.

En este tema conocerás:
Las técnicas utilizadas para el etiquetado POS.
El etiquetado POS con el modelo oculto de Markov.
El uso del algoritmo de Viterbi en el proceso de etiquetado.

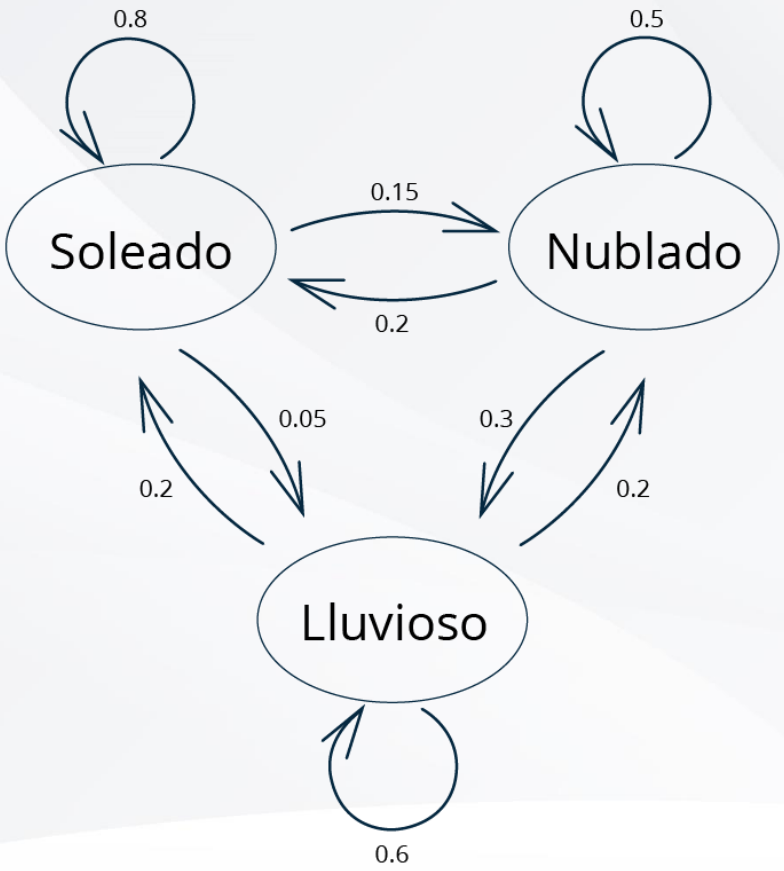
Explicación

Para comprender el modelo oculto de Markov es de utilidad recordar la **propiedad de Markov** utilizada en la teoría de probabilidad y estadística. Esta propiedad sugiere que la distribución de una variable aleatoria en el futuro depende únicamente de su distribución en el estado actual y ninguno de los estados anteriores tiene ningún impacto en los estados futuros, es decir, un proceso Markoviano carece de memoria.
Con la finalidad de calcular la probabilidad de la cadena de Markov o modelo de Markov se utiliza la expresión:

$$P(q_1, \dots, q_n) = \prod_{i=1}^{n} P(q_i | q_{i-1})$$

Explicación

En los modelos ocultos de Markov, la ocurrencia de cada estado está asociada a una distribución de probabilidad y la transición entre estados, se asocia a una probabilidad de transición de estados, es decir, la probabilidad de pasar de un estado a otro.



Explicación

Los componentes del modelo oculto de Markov son los siguientes:

• Conjunto de estados, S. Donde $|S| = N$ y N es el número total de estados. El conjunto de estados de un etiquetador POS está formado por las etiquetas.

• Conjunto de observaciones posibles, O. El conjunto de observaciones posibles de un etiquetador POS está formado por las palabras.

• Estado inicial, S_0. El estado inicial de un etiquetador POS es el inicio de la oración.

• Matriz de probabilidades de transición, A. En un etiquetador POS un ejemplo de la matriz A sería la probabilidad de que la palabra actual tenga una etiqueta de verbo dado que la etiqueta de la palabra previa fue un sustantivo.

• Matriz de probabilidades de salida por cada estado, llamadas también de emisión, B. En un etiquetador POS un ejemplo de la matriz B sería la probabilidad de que la palabra sea Pedro, dado que la etiqueta es un sustantivo.

• Vector de probabilidades iniciales, π.

Formalmente

$$\lambda = \{A, B, \pi\}$$

Actividad

1. Descarga el corpus "*Brown*" de NLTK y explóralo. No olvides importar las librerías y módulos requeridos para la actividad.
2. Explora el conjunto de etiquetas de *Penn Treebank*. Utiliza la instrucción nltk.help.upenn_tagset()
3. Explora las categorías del corpus Brown.
4. Utiliza la instrucción *help(nltk.tag.hmm.HiddenMarkovModelTagger)* para conocer el funcionamiento de la clase HiddenMarkovModelTagger.
5. Entrena y evalúa un etiquetador HMM con datos de entrenamiento y prueba respectivamente.

Cierre

Existen diferentes técnicas para realizar el etiquetado POS, siendo el modelo oculto de Markov una de las más utilizadas. Este tipo de modelo se entrena a partir de un conjunto de datos previamente etiquetados y puede utilizar el algoritmo de Viterbi como decodificador, es decir, un proceso capaz de encontrar la secuencia de etiquetas más probable dada una oración.