HMMs como modelos sustitutos: Part Of Speech (POS) Tagging

28 de noviembre de 2025

1. Etiquetado de secuencias

El **etiquetado de secuencias** consiste en asignar una etiqueta a cada elemento de una secuencia, manteniendo la **dependencia** entre etiquetas vecinas. En PLN, la secuencia típica es una oración tokenizada $x_{1:T} = (x_1, \ldots, x_T)$ (palabras) y se busca predecir una secuencia de etiquetas $y_{1:T} = (y_1, \ldots, y_T)$.

Formalmente, el objetivo es encontrar la secuencia más probable:

$$\hat{y}_{1:T} \;=\; \arg\max_{y_{1:T}} \; P(y_{1:T} \mid x_{1:T})\,.$$

A diferencia de clasificar cada palabra de forma independiente, en etiquetado de secuencias se aprovecha que ciertas etiquetas tienden a seguir a otras (p. ej., un determinante suele preceder a un sustantivo), lo cual permite desambiguar usando el contexto.

1.1. Etiquetado de secuencias para POS y entidades nombradas

- **Etiquetado morfosintáctico (POS):** asigna categorías como `NOUN`, `VERB`, `ADJ`, etc., lo que aporta señales sobre estructura y significado.
- **Reconocimiento de entidades nombradas (NER):** identifica y clasifica menciones como `PERSON`, `LOCATION`, `ORGANIZATION`, entre otras.

Estas tareas se benefician de modelar dependencias entre etiquetas vecinas (contexto). Clásicamente se abordan con modelos generativos como **HMM** y modelos discriminativos como **CRF**; enfoques modernos usan redes neuronales (p. ej., RNNs y *Transformers*) para aprender representaciones contextuales.

1.2. Etiquetas UPOS en español

- **ADJ** (*Adjective*) — Adjetivo: describe propiedades o cualidades de un sustantivo. *Ej.:* grande, pequeño, internacional, feliz.
- **ADP** (*Adposition*) — Adposición (en español casi siempre preposición). *Ej.:* de, en, a, con, por, para, sobre.

- **ADV** (*Adverb*) — Adverbio: modifica verbo, adjetivo u otro adverbio. *Ej.:* rápidamente, muy, ayer, aquí.
- **AUX** (*Auxiliary*) — Verbo auxiliar: forma tiempos compuestos, pasivas y perífrasis. *Ej.:* haber, estar ("está comiendo"), ser ("es estudiado").
- **CCONJ** (*Coordinating conjunction*) — Conjunción coordinante. *Ej.:* y, o, pero, sino, ni.
- **DET** (*Determiner*) — Determinante (artículos y otros). *Ej.:* el, la, los, las, un, una, este, aquella, algunos.
- **INTJ** (*Interjection*) — Interjección. *Ej.:* ąay!, ąeh!, hola, adiós, gracias (como exclamación).
- **NOUN** (*Noun*) — Sustantivo común. *Ej.:* casa, perro, precio, inflación, libro.
- **NUM** (*Numeral*) — Numeral (cardinales, ordinales, etc.). *Ej.:* uno, dos, primero, segundo, 3, 2024.
- **PART** (*Particle*) — Partícula (palabras gramaticales difíciles de clasificar como otra cosa). *Ej.:* no, ya, pues (según el análisis), que (en ciertas construcciones).
- **PRON** (*Pronoun*) — Pronombre. *Ej.:* yo, tú, él, ella, nosotros, esto, eso, alguien.
- **PROPN** (*Proper noun*) — Nombre propio. *Ej.:* México, Juan, Madrid, América Latina, Pemex.
- **PUNCT** (*Punctuation*) — Signo de puntuación. *Ej.:* . , ; : ? ! ( ) ...
- **SCONJ** (*Subordinating conjunction*) — Conjunción subordinante. *Ej.:* que, porque, aunque, cuando, si (condicional).
- **SYM** (*Symbol*) — Símbolo. *Ej.:* $, %, , +, =, @.
- **VERB** (*Verb*) — Verbo léxico ("verbo de contenido"). *Ej.:* comer, subir, caer, pensar, estudiar.
- **X** (*Other*) — Otro / desconocido: elementos que no encajan bien en ninguna otra categoría. *Ej.:* trozos de código, errores de OCR, palabras extranjeras raras, etc.

2. Aplicaciones

2.1. Ambigüedad léxica: el POS cambia según el significado

En español existen **homógrafos** (misma forma escrita, distinto significado) que pueden pertenecer a **categorías gramaticales diferentes**. Por eso, *la misma oración* puede admitir más de un etiquetado POS dependiendo de la interpretación.

Ejemplo: *"El vino tinto vino de Chile."*
Este tipo de ambigüedad ilustra por qué un etiquetador POS necesita **contexto** (y un modelo estadístico) para elegir la secuencia de etiquetas más plausible y poder diferenciar enre vino y vino (NOUN vs VERB).

3. Otras aplicaciones del etiquetado de secuencias

- **Análisis de sentimiento:** determina el tono identificando adjetivos como "feliz" o "triste".
- **Modelado de temas:** ayuda a categorizar documentos etiquetando sustantivos y verbos clave.
- **Extracción de palabras clave:** mejora la precisión de las búsquedas identificando palabras clave importantes en las consultas.
- **Revisión gramatical:** identifica si las palabras se usan correctamente.
- **Autocompletado:** sugiere la siguiente palabra basándose en las palabras escritas previamente.
- **Comprensión de comandos:** ayuda a asistentes como Siri o Alexa a entender y ejecutar comandos de voz.
- **Reconocimiento de entidades:** identifica nombres de personas, lugares y organizaciones en el texto.
- **Traducción precisa:** garantiza la traducción correcta de las palabras según la estructura de la oración.
- **Comprensión de la entrada del usuario:** mejora las respuestas de los chatbots al entender la estructura de las oraciones.
- **Análisis de tendencias:** analiza publicaciones para identificar tendencias y temas populares.

3.1. Modelo sustituto / *Surrogate model*

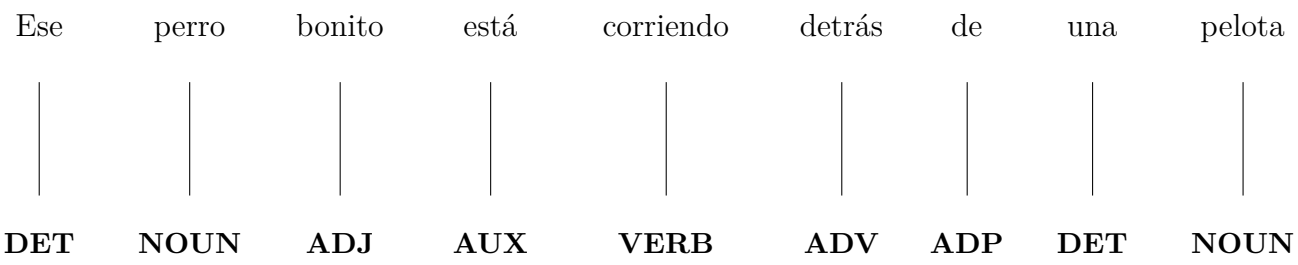
Librerías recientes como *UDPipe* utilizan redes neuronales profundas para generar sus modelos de lenguaje. Estos algoritmos suelen considerarse como una *"caja negra"* (*black-box*).

Una alternativa es aproximar su comportamiento mediante *Hidden Markov Models* (HMMs), lo cual puede ofrecer:

- Menos parámetros y, en consecuencia, cómputo más rápido.
- Interpretabilidad mediante el uso de estados ocultos.

4. POS Tagging

Ese perro bonito está corriendo detrás de una pelota.



Interpretación: **DET** determinante, **NOUN** sustantivo, **ADJ** adjetivo, **AUX** auxiliar, **VERB** verbo, **ADV** adverbio, **ADP** preposición.

4.1. Etiquetado POS usando Modelos Ocultos de Markov (HMM)

La tarea de **etiquetado morfosintáctico** (POS tagging) consiste en: dada una oración (secuencia de palabras), asignar una etiqueta gramatical a cada palabra (por ejemplo, `NOUN`, `VERB`, `ADJ`, etc.). Desde la perspectiva de los **Modelos Ocultos de Markov (HMM)**, esta tarea puede formularse como el **segundo problema del HMM**: dada una secuencia de **observaciones** (las palabras), inferir la secuencia de **estados ocultos** (las etiquetas) que la generó.

Correspondencia HMM $\leftrightarrow$ POS tagging. En este marco, los componentes del HMM se interpretan así:

- **Estados (eventos):** etiquetas POS (p. ej., `NOUN`, `ADJ`, `VERB`).
- **Observaciones:** palabras de la oración (p. ej., *"El", "pastel", "es", "delicioso"*).
- **Probabilidades iniciales:** probabilidad de iniciar con la etiqueta T_i,

$$\pi_i \;=\; P(T_1 = T_i).$$

- **Probabilidades de transición:** probabilidad de que una etiqueta siga a otra,

$$a_{ij} \;=\; P(T_t = T_i \mid T_{t-1} = T_j).$$

- **Probabilidades de emisión:** probabilidad de observar una palabra dada una etiqueta,

$$b_i(w) \;=\; P(W_t = w \mid T_t = T_i).$$

Objetivo (decodificación). Dada una oración $w_{1:T}$ (observaciones), se busca la secuencia de etiquetas más probable:

$$\hat{t}_{1:T} \;=\; \arg\max_{t_{1:T}} P(t_{1:T} \mid w_{1:T}),$$

lo cual típicamente se resuelve con el algoritmo de **Viterbi**.

tikz

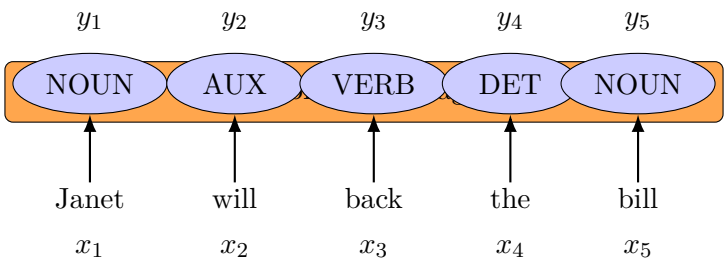


Figura 1: Tarea de etiquetado POS: mapeo de palabras de entrada $x_1, \ldots, x_n$ a etiquetas $y_1, \ldots, y_n$.

4.2. Etiquetado POS: mapeo de palabras a etiquetas

El etiquetado de categorías gramaticales (POS tagging) puede verse como una función que toma una secuencia de palabras de entrada $x_{1:n}$ y produce una secuencia de etiquetas $y_{1:n}$, una por token.

4.3. Entrenamiento del modelo y obtención de etiquetas POS

Entrenar el modelo:

- Calcular las matrices de **probabilidades de transición** y **emisión**.

Encontrar etiquetas POS:

- Usar el algoritmo de **Viterbi**.
- Consta de un **paso hacia adelante** y un **paso hacia atrás**.
- **Paso hacia adelante:** encontrar el mejor camino posible; es decir, el camino hacia cada nodo con la menor **probabilidad logarítmica negativa** (o costo) acumulada.
- **Paso hacia atrás:** reconstruir el camino óptimo mediante **retrotrazado** (*backtrace*).

Referencias

- `Jurafsky, D., Martin, J. H. (2025). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, with language models (3rd ed., online manuscript released August 24, 2025). https://web.star`
- Zucchini, W., MacDonald, I. L., & Langrock, R. (2017). *Hidden Markov Models for Time Series: An Introduction Using R* (2nd ed.). Chapman and Hall/CRC, Boca Raton.
- `nlp-programming-en-04-hmm.pdf`. Phontron (slides). https://phontron.com/slides/nlp-programming-en-04-hmm.pdf
- *An introduction to part-of-speech tagging and the Hidden Markov Model* (freeCodeCamp). https://www.freecodecamp.org/news/an-introduction-to-part-of-speech-tagging-and-the-hidden-markov-model-953d45338f24/
- UPOS (Universal POS) en español (Universal Dependencies). https://universaldependencies.org/u/pos/