# Aplicación de modelos de markov con estados ocultos y una extensión bayesiana en la predicción de etiquetado POS para palabras desconocidas en Hindi y Marathi

Juan Pablo Barrera Avella,
Julieth Andrea López Castiblanco,
Natalia Coy Lozano

Diciembre 9 del 2020

**Resumen**

Los modelos ocultos de Marvok (Hidden Markov models HMMs), actualmente son ampliamente utilizados para el análisis del etiquetado POS o POS tagging. En este trabajo veremos algunos hechos de su historia, exploraremos los conceptos básicos de estos modelos y realizaremos una aplicación de cadenas de markov con estados ocultos en la prediccion de la marcacion POS para palabras desconocidas en una base de datos en idioma Hindi.

## 1. Introdución

El paper en el que fue basado este escrito *Prediction of POS Tagging for Unknown Words for Specific Hindi and Marathi Language* se propone una extension Bayesiana (el algoritmo Naïve Bayes) del enfoque en Modelos de Markov Ocultos, los idiomas considerados en el paper: el Hindi y el marathi, son similares, pues su lengua raíz, el sanscrito, es la misma y además son escritas en devanagari (escritura moderna silábica). Sin embargo y por desgracia, el etiquetado del POS para los idiomas indios no ha sido muy investigado. Las formas de etiquetado actuales usan aprendizaje automático, técnicas estocásticas e información fonética sin tener exito en la obtención de los resultados deseados debido a que las palabras no presentes en el corpus(datos de entrenamiento) o no reconocibles son marcadas como POS DESCONOCIDA. Durante el desarrollo del presente documento buscaremos proporcionar al lector un acercamiento breve alrededor de los temas ya anunciados, enfocandonos en las siguientes preguntas:

- ¿Qué son los modelos de Markov Ocultos?
- ¿Qué es un procedimiento de etiquetado POS y cómo se realiza?
- ¿Cuál es el problema de las palabras desconocidas en Hindi y Marathi? ¿Cuál es una posible solución?

## 2. Marco Teórico

### Historia

La historia de los modelos de estados osculos se divide en dos partes, la historia de los procesos de markov, y la historia de los algoritmos con los que se implementan los modelos. Por un lado,

las cadenas de Markov, fueron desarrolladas por el matemático ruso Andréi Markov alrededor de 1905, sin embargo, los modelos ocultos de Markov fueron descritos por primera vez por Leonard E.Baum en los 60's. Por otro lado, con el desarrollo en las ciencias computacionales, desde los 40's, se empezaron a desarrollar soluciones a varios algoritmos para resolver problemas de la vida real.

Aquí fueron importantes los avances de C.Shannon porque impulsaron la necesidad de integrar la automatización estocástica en dispositivos eléctricos, así como el desarrollo del algoritmo de esperanza-maximización (1977) de A.Dempster, N.Laird, y D.Rubin, que es utilizado para encontrar la estimación por máxima verosimilitud de parámetros en modelos probabilisticos donde éstos dependen de variables latentes, de este se desarrollan otros algoritmos como el Baum-Welch o el Viterbi. Actualmente, los HMMs son ampliamente utilizados en aplicaciones de reconocimiento de voz (60's), procesamiento de lenguaje natural, bioinformática y aplicaciones de secuencias biológicas (80's).

## Modelos de Markov ocultos

Los modelos de marvok ocultos son autómatas estocásticos finitos que pueden aprender y que hoy en día se consideran una forma de las redes bayesianas dinámicas. Un autómata es un sistema secuencial que se puede encontrar en uno de los posibles estados, y donde los valores de las salidas en un momento dado, dependen de los valores de las entradas en dicho momento y del estado anterior o estado interno. Cuando un modelo de Markov oculto tiene estados y observaciones discretas decimos que es una cadena de Markov con estados ocultos.

## Cadenas de Markov con estados ocultos

Consisten en dos procesos estocásticos, el primero es una cadena de markov con sus respectivos estados $S = \{S_1, ..., S_L\}$ ( $S_n$ es el estado de la $0_i$ esima observación) y probabilidades de transicion, sin embargo, los estados de la cadena no se pueden ver, por eso se le llaman ocultos. El segundo proceso produce resultados observables en cada momento y toman valores en el conjunto $O = \{O_1, ..., O_N\}$ basado de un estado dependiente de distribucion de probabilidad. Estos modelos tambien son llamados procesos estocásticos doblemente incrustados.

Sin perdida de generalidad, una cadena de markov con estados ocultos esta completamente definida por 5 elementos: i)Los L estados del modelo, ii)Las M observaciones por estado, iii)La distribución de probabilidad de transición entre estados $A\{a_{ij}\}$ donde $a_{ij} \geq 0$ es la probabilidad de transición:

$$a_{ij} = P(S_{n+1} = j | S_n = i) \qquad 1 \leq i, j \leq N$$

y $\sum_{j=1}^{N} a_{ij} = 1$ para $1 \leq i \leq N$, iv)La distribución de las observaciones en cada estado, $B = b_j(k) \geq 0$ donde $b_j(k)$ es la probabilidad de que la observación sea emitida o producida en el estado $S_j$.

$$b_j(k) = P(v_n = O_k | S_n = j) \qquad 1 \leq k \leq L$$

y $\sum_{k=1}^{M} b_j(k) = 1$ para $1 \leq j \leq N$, tenemos que $v_n$ es el vector de parámetros en el tiempo n, las probabilidades discretas se ambian a funciones de densidad si las observaciones son continuas. v) La distribución inical $\pi = \pi_i$ donde $\pi_i$ es la probabilidad de que el modelo este en el estado $S_i$ en el tiempo n=0

$$\pi_i = PS_n = i \qquad 1 \leq i \leq L$$

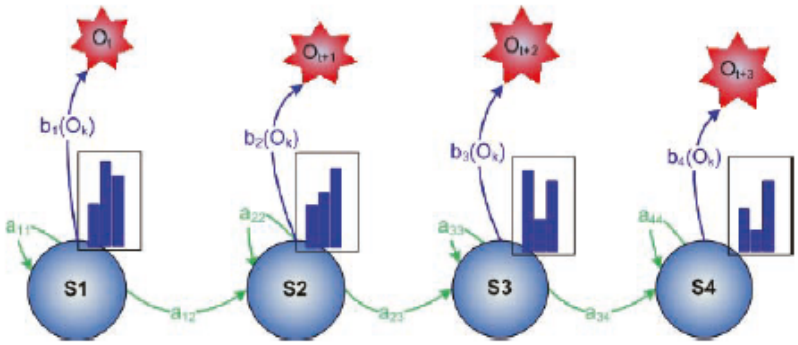


Figura 1: Estructura de un modelo de Markov oculto

Existen 3 problemas básicos que este modelo busca solucionar, el problema de evaluación, el problema de decodificación y el problema de aprendizaje. En este caso el problema de predicción de estiquetado POS es un problema de decodificación, que busca la mejor o más óptima secuencia de estados dada una secuencia de observaciones.

# 3. Una vista al como de la prediccion del etiquetado POS para palabras desconocidas

### Contexto

El etiquetdo POS se usa para examinar un mensaje y extraer características gramaticales de cada palabra teniendo en cuenta la relación de una palabra con las que la acompañan o estan relacionadas en una frase o párrafo. Así, hay ocho formas principales de clasificar una palabra: el sustantivo, el pronombre, adjetivo, verbo, adverbio, preposición, conjunción e interjección, que a su vez podriamos volver a clasificar siendo mas especificos. Ahora bien, el enfoque probabilístico se utiliza para la predicción de la palabra desconocida en el caso de nuestro trabajo, cabe señalar que la precisión del sistema completo de prediccion se degrada con el aumento del número de palabras desconocidas.

### Metodología general

Los datos en bruto se someten a un preprocesamiento. Los datos se dividen en dos partes, a saber, los datos de entrenamiento que se utilizan para entrenar el modelo y los datos de prueba que se utilizan para la evaluación de éste.

# 4. Modelos y Algoritmos inmersos en la prediccion de la enmarcacion POS

### Modelo de Markov Oculto

Los HMMs se utilizan ampliamente para dos propósitos principales: la asignación de etiquetas adecuadas a los datos secuenciales y segundo es la estimación de la probabilidad de una secuencia

o etiqueta de datos.

## Algoritmo de Viterbi

El algoritmo de Viterbi se usa para hallar la secuencia finita más probable de estados ocultos, es decir, la asignación de las partes adecuadas de etiquetas de voz a la frase introducida. Las palabras son observaciónes y las etiquetas son estados.

Este modelo requiere datos clasificados o etiquetados de entrenamiento(corpus). y con ellos calcula frecuencias, probabilidades iniciales, matrices de transición y emisión.Las fórmulas para el cálculo de inicio, transición y emisión se mostraran a continuacion.
-Probabilidades inciales de las etiquetas:

$$\frac{\text{Frecuencia de la etiqueta}}{\text{Numero total de palabras}}$$

-Probabilidad de transicion de la etiqueta $t_1$ a la $t_2$:

$$\frac{\text{Frecuencia de transicion de } t_1 \text{ a la } t_2}{\text{Frecuencia de } t_1}$$

-Probabilidad de emision:

$$\frac{\text{Frecuencia de la palabra w etiquetada con t}}{\text{Frecuencia de t}}$$

Asi, el resultado del algoritmo de Viterbi es la secuencia más probable de etiquetas para la de entrada.

## Algoritmo Bayesiano

La predicción de la etiqueta desconocida (UNK) puede hacerse usando el teorema de Bayes Naïve, por medio de este, la predicción de la etiqueta de palabra desconocida se calcula sobre la base de la probabilidad de transición y de inicio. Para predecir la etiqueta de la palabra desconocida se utiliza:

$$\begin{aligned} b(\text{MAP}) &= \arg\max_b [P(b \mid A)] \\ &= \arg\max_b [(P\ (a \mid b)\ P(b))\ /\ (P(a))] \\ &= \arg\max_b [P\ (A \mid b)\ P(b)] \\ &= \arg\max_b [P\ (a_1, a_2, \ldots a_n | y) P(b)] \\ b &= \arg\max_b P\left(b \prod_{i=1}^{n} P(a_i|b)\right) \end{aligned}$$

donde el vector **A** hace referencia a etiquetas y $b$ es la etiqueta de la palabra desconocida. Además, "MAP" corresponde a la etiqueta más probable. En la siguiente tabla muestra un ejemplo de las etiquetas y sus significados presentes en el NLTK de la India corpus para el marathi y el hindi.

| SR. No. | Tags | Meaning | SR.No. | Tags | Meaning |
|---|---|---|---|---|---|
| 1 | NN | Noun singular | 18 | QFNUM | Quantifier number |
| 2 | JJ | Adjective | 19 | RP | Particle |
| 3 | VFM | Verb finite main | 20 | NEG | Negative word |
| 4 | SYM | Symbol | 21 | QF | Quantifier |
| 5 | NNP | Proper noun | 22 | JVB | Adjective in kriyamula |
| 6 | NNC | Common noun | 23 | NLOC | Noun location |
| 7 | INTF | Intensifier | 24 | VJJ | Verb non-finite adjective |
| 8 | CC | Conjunction | 25 | QW | Question word |
| 9 | PREP | Preposition | 26 | VM | Main verb |
| 10 | PRP | Pronoun | 27 | JJC | Adjective comparative |
| 11 | NVB | Verb past participle | 28 | PSP | Post position |
| 12 | VAUX | Auxiliary verb | 29 | NST | Spatial noun |
| 13 | PUNC | Punctuation | 30 | QC | Cardinal |
| 14 | NNPC | Compound proper noun | 31 | DEM | Demonstrative |
| 15 | VRB | Verb | 32 | WQ | Question word |
| 16 | VNN | Non-finite nominal | 33 | QO | Ordinal |
| 17 | RB | Adverb | 34 | RDP | Reduplication |
| | | | 35 | UNK | Unknown word |

### Aplicación Computacional en Python

Para poder realizar la aplicación anteriormente mencionada, se requiere realizar la estimación de las probabilidades de transición, las probabilidades de estado inicial y de las probabilidades de emisión. Ésto se realiza de la misma manera como se describió en la subsección anterior. Sin embargo, para poder realizar estas estimaciones se requiere un corpus de frases que permitan inferir cómo se comportan las secuencias de POS en el mundo real al usar el lenguaje natural HINDI. Para tal fin se usa el corpus dispuesto por la libreria nltk de python.

No obstante, este corpus está bastante incompleto y es muy pequeño en comparación con otros corpus de la misma libreria. Su ínfimo tamaño (cerca de 9000 frases) genera que, con alta probabilidad, hayan palabras del Hindi que no se encuentran en el corpus y por ellos, el algoritmo de Viterbi las clasificará como desconocidas. En este sentido, usando todo el corpus, se entrena un clasificador del tipo Naive Bayes para así predecir de antemano las palabras desconocidas y, así, no afectar el desempeño del HMM para la tarea de POS Tagging. El funcionamiento del clasificador se muestra en la subsección denominada Algoritmo Bayesiano.

Finalmente, se adjunta una descripción más detallada del código, así como el cuaderno con la codificación en lenguaje Python de la aplicación. El código con el cuál se trabajó se encuentra en el siguiente link de Drive:

`https://drive.google.com/drive/u/0/folders/1G97mnmtKHiyR4SfgWFDkGfoet9TNsw_U`

## 5. Conclusiones

El algoritmo general propuesto, compuesto por el algoritmo Viterbi y el Naive bayes, para el etiquetado POS en el idioma Hindi y Marathi provee una solución para aquellas palabras etiquetadas como desconocidas después de un proceso de clasificacion inicial a partir del corpus, problemas y soluciones similares que se han construido para otros idiomas, contribuyendo así a la solución de un problema que aún es nuevo. Sobre los modelos de markov ocultos HMMs, podemos conluir que son ampliamente utilizados hoy en día, por su comprensibilidad, su intuitivo entendimiento y la adaptabilidad de la naturaleza de su estructura a problemas actuales, además de la facilidad computacional con la que cuenta, que es capaz de competir hoy en día con las redes neuronales. Sin embargo, la evidente desventaja de este modelo es que es que se requiere un corpus de gran tamaño, por eso, el siguiente paso para de desarrollo de este problema es obtener un corpus más rico.

# 6. Referencias


Kouemou G. (2011). History and theoretical basics of hidden markov models. En Hidden markov models, theory and applications. InTech. Croacia. (p. 3-26). Editor: Dymarski Przemyslaw.

Yoon, B. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. Current Genomics, 10, (p. 402-415).

Chiplunkar K., Kharche M., Chaudhari T., Shaligram S., Limkar S. (2021). Prediction of POS tagging for unknown words for specific Hindi and Marathi language. En: Intelligent data engineering and analytics, Frontiers in intelligent computing, volumen 1177. Springer. (p. 133-143). Editores: Satapathy S., Zhang YD., Bhateja V., Majhi R.

Anandika A., Mishra S.P., Das M. (2021) Review on Usage of Hidden Markov Model in Natural Language Processing. In: Intelligent and Cloud Computing. Smart Innovation, Systems and Technologies, vol 194. Springer, Singapore. (p. 435-450). Editores: Mishra D., Buyya R., Mohapatra P., Patnaik S.

Galaviz R. Portafolio. Autómatas. Google sites. Recuperado de https://sites.google.com/site/portafoliorenygalaviz/

Modelo oculto de markov. Numerentur. Recuperado de http://numerentur.org/markov-hmm/