

## Los Campos Aleatorios Condicionales en la Segmentación de Textos por Idiomas Title (without the word *title*).

### Conditional Random Fields in Text Segmentation by Language

Robin Cabeza Ruiz<sup>1</sup>

<sup>1</sup>Centro de Estudios CAD/CAM, Universidad de Holguín. Cuba. Robbinc91@gmail.com

#### RESUMEN

En este trabajo se propone la utilización de los Campos Aleatorios Condicionales para la resolución de la tarea de Segmentación de Textos por Idiomas, considerándola como una tarea de etiquetado de secuencias. Se considera que el cambio entre un idioma y otro en los documentos ocurrirá en cualquier parte del texto, se asume que las observaciones en el sistema estarán dadas por las palabras en el texto, y los estados serán los diferentes idiomas.

Palabras clave: Segmentación de Textos por Idiomas; Campos Aleatorios Condicionales; observaciones; estados.

#### ABSTRACT

This work presents using Conditional Random Fields for solving the task of Text Segmentation by Language, considering it as a sequence tagging task. Language changes are considered to occur in every part of the text, observations are assumed to be the words in the text, and the states are the different languages.

Keywords: Text Segmentation by Language; Conditional Random Fields; observations; states.

#### 1. INTRODUCCIÓN

Desde el surgimiento de Internet y las bases de datos, el número de fuentes de información en forma de texto disponibles en la web ha crecido de una manera vertiginosa (ya sea en sitios de noticias, blogs, redes sociales, etcétera). Esto provoca que cada vez sea mayor el volumen de datos en formato textual. Estos datos, son imposibles de calificar o analizar por las personas de una manera eficiente, por lo que se han creado herramientas computacionales capaces de hacerlo de manera automatizada. La Minería de Textos (MT) es el área de investigación dedicada a la obtención de información novedosa y valiosa de estos documentos. Dentro de la Minería de Textos existen tareas como el agrupamiento de documentos de acuerdo a su temática, o su clasificación en temáticas predefinidas, entre otras.

La utilización de técnicas de MT se suelen representar los documentos como vectores de términos, donde generalmente estos términos son sustantivos, formas verbales, etcétera, de acuerdo con la tarea a resolver. Esta selección de términos se realiza utilizando el Procesamiento de Lenguaje Natural (PLN) (Vásquez, Quispe, & Huayana, 2009). El PLN es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano.

Muchas de las técnicas de PLN dependen del idioma (es decir, es necesario conocer antes en qué idiomas están escritos los documentos para procesarlos), por lo que resulta una fase imprescindible el aplicar Identificación de Idiomas (II) antes. Esta tarea tiene como objetivo la identificación del idioma (o los idiomas) en que está escrito un documento. Teniendo en cuenta que un documento existente en la web puede

contener segmentos escritos en varios idiomas, es necesario realizar una Identificación de Idiomas en documentos multilingües. Sin embargo, con solo identificar los idiomas presentes en el documento puede no ser suficiente: resulta más útil contar con herramientas capaces de obtener los segmentos de cada uno de los idiomas presentes en el documento. Esta tarea se denomina Segmentación de Textos por Idiomas. Esta investigación se centra en la segunda variante, con el objetivo de aportar un nuevo método para obtener, de manera eficiente, al analizar un documento, qué idiomas están presentes en el mismo, y qué porción está escrita en cada uno de estos idiomas.

La primera de estas variantes ha sido tratada en variados estudios, entre ellos (Lui, Lau, & Baldwin, 2014), (Singh & Gorla, 2007) y (VRL, 2010). La segunda, por otro lado, se trata en muy pocas investigaciones, contándose entre la bibliografía revisada con (Yamaguchi & Tanaka-Ishii, 2012) y (Ruiz, 2016).

El último de estos investigadores, utilizó los Modelos Ocultos de Markov(HMM) para resolver la tarea, tratando a los documentos como un conjunto de observaciones a las que se le asigna, en el proceso de clasificación, un conjunto de etiquetas. En este caso las observaciones serían las palabras presentes en el texto, y las etiquetas (o estados) son los idiomas en los que ha sido redactado.

Para el etiquetado de secuencias, los Campos Aleatorios Condicionales (CRF del inglés Conditional Random Fields) han resultado más poderosos que los HMM gracias a la posibilidad que brindan de caracterizar sistemas más caóticos y el solapamiento de clases, como es el caso de un documento escrito en varios idiomas, sabiendo que existen palabras que pertenecen a más de un idioma (por ejemplo, la palabra “hospital”, puede estar escrita en idioma español o inglés).

Teniendo en cuenta las mejoras que puede brindar el uso de CRF al etiquetado de secuencias, la presente investigación se plantea como objetivo proponer estos modelos para su aplicación a la tarea de segmentación de textos por idiomas, teniendo en cuenta el compromiso existente entre eficiencia del sistema y la cantidad de memoria RAM disponible en el ordenador que sea destinado a su ejecución. La formulación de CRF permite afirmar, desde el punto de vista del autor, que este es un modelo que permitirá mejorar los resultados obtenidos hasta la fecha en la resolución de la tarea.

## **Estado del arte**

El primer trabajo de la bibliografía consultada cuyo objetivo es segmentar los documentos por idiomas, es el propuesto en (Yamaguchi & Tanaka-Ishii, 2012). En este los autores proponen un método que utiliza el concepto de Descripción de Longitud Mínima (Barron, Rissanen, & Yu, 1998) para encontrar los bordes de los idiomas (sitios en los que ocurren los cambios entre idiomas dentro de un documento). Los autores formulan el problema de la siguiente manera: Dado un texto X, se obtienen los segmentos para un listado de bordes B en correspondencia con un listado de idiomas L.

El segundo trabajo orientado a la segmentación de textos por idiomas es (Ruiz, 2016), en el que se utilizan los Modelos Ocultos de Markov para tratar a un documento como una secuencia de observaciones, a las que se le asigna una secuencia de etiquetas, que es la que tiene mayor probabilidad. Los autores tomaron como observaciones las palabras existentes dentro de los textos, y como etiquetas (o estados), el conjunto de idiomas disponibles para la identificación y segmentación. Además proponen una manera de obtener los idiomas en documentos en los que el cambio de idioma pueda ocurrir solamente en los saltos entre oraciones o párrafos dentro de los documentos, utilizando la biblioteca NLTK (Bird, 2006) para segmentar los documentos por oraciones, y el identificador de idiomas monolingüe langid (Lui & Cook, 2012) para obtener el idioma de cada oración.

## **2. METODOLOGÍA**

Un CRF (del inglés Conditional Random Fields) es un modelo utilizado habitualmente para etiquetar secuencias de datos o extraer información de documentos (Lafferty, McCallum, & Pereira, 2001). Han sido utilizados exitosamente en áreas de procesamiento de texto (F & McCallum, 2004) (Settles, 2005) (Sha & Pereira, 2003), bioinformáticas (Liu, Carbonell, & GopalaKrishnan, 2006), y visión por computadora (He, Zemel, & Carreira-Peripíñán, 2004).

La principal diferencia entre HMM (Modelos Ocultos de Markov) y CRF es que HMM calcula la probabilidad condicional de las etiquetas dadas las variables de entrada,  $p(z|x)$ , mientras que CRF calcula la probabilidad conjunta de ambas,  $p(z, x)$ , además de que CRF tiene acceso a más observaciones que HMM. Se puede representar como un grafo no dirigido  $G = (V, E)$  que define una distribución lineal de un conjunto de etiquetas para una secuencia dada de observaciones, en el que cada vértice representa una variable aleatoria cuya distribución de probabilidad debe ser deducida, y cada arista indica una dependencia entre las variables de los vértices que conecta. El grafo cumple la propiedad de Markov extendida a grafos (1).

$$P(S_i | O, S_j; i \neq j) = P(S_i | O, S_j; i \sim j) \quad (1)$$

Donde  $\sim$  significa que  $S_i, S_j$  están conectados por una arista. En cuanto a las observaciones  $O_i$ , lo más frecuente es que sea un vector, en vez de un valor escalar, teniendo observaciones multidimensionales.

CRF modela la probabilidad de una secuencia de etiquetas  $z$  dada una secuencia de observaciones  $x$  de la manera expresada en (2):

$$P(z_{1:N} | x_{1:N}) = \frac{1}{Z} \exp(\sum_{n=1}^N \sum_{i=1}^F \lambda_i f_i(z_{n-1}, z_n, x_{1:N}, n)) \quad (2)$$

Donde  $N$  es la cantidad de observaciones, y  $F$  la cantidad de funciones características definidas para el funcionamiento del modelo. Asimismo,  $Z$  es llamado factor de normalización, o función de partición, y se calcula mediante la fórmula representada en (3):

$$Z = \sum_{z_{1:N}} \exp(\sum_{n=1}^N \sum_{i=1}^F \lambda_i f_i(z_{n-1}, z_n, x_{1:N}, n)) \quad (3)$$

Las funciones características son el componente principal de un CRF. En general, una función característica tiene la forma  $f_i(z_{n-1}, z_n, x_{1:N}, n)$ , y busca en un par de estados adyacentes  $z_{n-1}, z_n$ , la secuencia completa  $x_{1:N}$ , y en qué parte de la secuencia se encuentra el modelo en ese mismo instante de tiempo  $n$ . Estas funciones arbitrarias producen valores reales.

Un ejemplo de una función característica para la segmentación de textos por idiomas se muestra en (4), en la que "ES" representa el código utilizado para el idioma español, y "texto" representa la palabra "texto" del idioma español.

$$f_1(z_{n-1}, z_n, x_{1:N}, n) = \begin{cases} 1 & \text{si } z_n = ES \text{ y } x_n = \text{texto} \\ 0 & \text{en caso contrario} \end{cases} \quad (4)$$

La utilización de este rasgo depende de su correspondiente peso  $\lambda_i$ . Si  $\lambda_i > 0$ , cuando esté activa  $f_1$  (es decir, cuando se esté analizando la palabra "texto" y se le asigne la etiqueta ES), incrementa la probabilidad de la secuencia de etiquetas  $z_{1:N}$ . En otras palabras: el modelo CRF preferiría la etiqueta ES para la palabra "texto". Por otro lado, si  $\lambda_i < 0$ , CRF evitará dicha etiqueta para la palabra en cuestión.

### 3. RESULTADOS Y DISCUSIÓN

#### Entrenamiento de un Campo Aleatorio Condicional

El entrenamiento de un CRF consiste en encontrar los parámetros  $\lambda_i$  que maximicen la probabilidad de las secuencias de etiquetas del entrenamiento para las observaciones, lo cual se logra con (5). Es necesario contar con secuencias de observaciones previamente etiquetadas  $\{(x^{(1)}, z^{(1)}), \dots, (x^m, z^m)\}$ , donde  $x^{(1)} = x_{1:N}^{(1)}$ .

$$p(z|x) = \sum_{i=1}^m \log p(z^{(j)} | x^{(j)}) \quad (5)$$

CRF necesita para su entrenamiento un conjunto de secuencias de observaciones (en este trabajo las observaciones son las palabras dentro de los textos), obtenidas del corpus de entrenamiento. Además es necesario definir las funciones características del sistema. A continuación se expone la creación de estas funciones.

#### Funciones características

Estas funciones son, se mencionó anteriormente, el punto más importante de la definición de un CRF. Para la tarea de segmentar textos por idiomas, las funciones características pueden ser conformadas a partir de la unión de tres conjuntos de funciones auxiliares: las funciones de transición (6), funciones de observación (7), y funciones de valor de etiqueta (8).

Funciones de transición:

$$f_i(z_{n-1}, z_n, x_{1:N}, n) = \begin{cases} 1 & \text{si } t(z_{n-1}, z_n) \\ 0 & \text{en caso contrario} \end{cases} \quad (6)$$

$f_i(z_{n-1}, z_n, x_{1:N}, n)$  es 1 si en el entrenamiento aparece la transición del idioma  $z_{n-1}$  al idioma  $z_n$ . Estas funciones favorecen la aparición del idioma  $z_{n-1}$  seguido del idioma  $z_n$  en un texto cualquiera. No se tiene en cuenta la palabra que está en la posición actual de la cadena de texto. Este conjunto de funciones es semejante a la matriz de transición  $A$  utilizada por HMM,  $p(z_n | z_{n-1})$  (Ruiz, 2016).

Funciones de observación:

$$f_i(z_{n-1}, z_n, x_{1:N}, n) = \begin{cases} 1 & \text{si } x[n] \in \text{words}[z_n] \\ 0 & \text{en caso contrario} \end{cases} \quad (7)$$

$\text{words}[z_n]$ : conjunto de todas las palabras observadas a partir del idioma  $z_n$ , en el conjunto de textos utilizados para el entrenamiento.

$x[n]$ : palabra que se analiza actualmente en el texto de prueba.

Estas funciones son creadas con el objetivo de fomentar la puntuación de una palabra a partir del idioma en el que fueron emitidas dentro de los documentos del entrenamiento.

Funciones de valores de etiqueta:

$$f_i(z_{n-1}, z_n, x_{1:N}, n) = \begin{cases} 1 & \text{si } z_n = z_p \text{ y } x[n] = x_p \\ 0 & \text{en caso contrario} \end{cases} \quad (8)$$

Favorece la secuencia de entrada cuando se trata de la palabra  $x_p$  y se está analizando el idioma  $z_p$ , ambas pasadas como parámetros extra. Estas funciones, junto con las funciones de observación, juegan un papel equivalente a la matriz de observaciones utilizada por HMM, que guarda la probabilidad de emitir la palabra  $x_p$  a partir del idioma  $z_p$ .

Las funciones descritas anteriormente permiten crear un modelo CRF para la tarea de segmentación de textos por idiomas. A partir de este punto se puede entrenar un CRF utilizando una serie de documentos con sus respectivos metadatos (idioma de cada palabra en los textos).

Cabe resaltar que, si bien CRF es un modelo que puede brindar más bondades que otros sistemas, como los HMM, es necesario tener en cuenta el consumo de memoria (y tiempo) que puede representar su utilización. Las funciones características aumentan el grado de perfección del modelo, pero debe encontrarse un límite en este sentido, pues muchas de estas funciones son sinónimo de tardanza por parte del sistema, ya que del modelo debe evaluar más posibilidades para el etiquetado de la secuencia de entrada.

Debido a limitaciones en la infraestructura de cómputo, solo se pudieron ejecutar pruebas a pequeña escala al algoritmo. Para probar su funcionamiento se entrenó el algoritmo con un subconjunto ínfimo del corpus Wikipedia-multi utilizado en (Lui, Lau, & Baldwin, 2014) y (Ruiz, 2016). El algoritmo mostró ser eficiente en el entrenamiento utilizado.

### 3. CONCLUSIONES

Los Campos Aleatorios Condicionales resultan una herramienta muy poderosa para el etiquetado de secuencias. Han tenido gran utilización en áreas de procesamiento de texto, bioinformáticas y visión por computadora. La formulación de los CRF permite afirmar que pueden ser una herramienta útil para la segmentación de textos multilingües por idiomas, tomando esta tarea como un problema de etiquetado de secuencias, en la que, al igual que utilizando los Modelos Ocultos de Markov, los estados representan los idiomas en los que está (o puede estar) escrito un documento, y las palabras del texto constituyen las observaciones del sistema.

Para futuras investigaciones, se pretende probar el algoritmo a mayor escala, para realizar comparaciones reales con otros métodos utilizados para la segmentación de textos por idiomas. Se piensa también explorar los llamados Markov Random Fields, en aras de hallar vías más eficientes para la resolución de la tarea.

### 4. REFERENCIAS BIBLIOGRÁFICAS

- Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *Information Theory*, 6.
- Bird, S. (2006). NLTK: The natural language toolkit. *Proceedings of the COLING/ACL on Interactive presentation sessions*, (págs. 69-72).
- F, P., & McCallum, A. (2004). Accurate information extraction from research papers using conditional random fields. *Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics*.
- He, X., Zemel, R. S., & Carreira-Peripinán, M. A. (2004). Multiscale conditional random fields for image labelling., (pág. Conference on Computer Vision and Pattern Recognition).
- Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional Random Fields: Probabilistic Model for Segmenting and Labeling Sequence Data. *Proceedings of the 18th International Conference on Machine Learning*, (págs. 282-289).
- Liu, Y., Carbonell, J., & GopalaKrishnan, V. (2006). Protein fold recognition using conditional random fields (SCRFs). *Journal of Computational Biology*, 394-406.
- Lui, M., & Cook, P. (2012). langid.py for better language modelling. *Australasian Language Technology Association Workshop*, (pág. 107).
- Lui, M., Lau, J. H., & Baldwin, T. (2014). Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, (págs. 27-40).
- Ruiz, R. C. (2016). Text segmentation by language. *Sistemas & Telemática*, 61-70.
- Settles, B. (2005). Abner: An open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 3191-3192.
- Sha, F., & Pereira, F. (2003). Shallow Parsing with conditional random fields. *Conference of Human Language Technology and North American Association for Computational Linguistics*, (págs. 213-220).
- Singh, A. K., & Gorla, J. (2007). Identification of languages and encodings in a multilingual document. *Building and Exploring Web Corpora (WACe-2007): Proceedings of the 3rd Web as Corpus Workshop*, (pág. 95).
- Vásquez, A. C., Quispe, J. P., & Huayana, A. M. (2009). Procesamiento de Lenguaje Natural. *Revista de investigación de Sistemas e Informática*, 45-54.
- VRL, N. (2010). Multilingual Language Identification: ALTW 2010 Shared Task Dataset. *Australasian Language Technology Association Workshop*, (pág. 4).

Yamaguchi, H., & Tanaka-Ishii, K. (2012). Text Segmentation by Language using minimum description length. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume1*, (págs. 969-978).

#### **SOBRE LOS AUTORES**

Robin Cabeza Ruiz: Licenciado en Ciencias de la Computación, 2015. Máster en Diseño y Fabricación Asistidos por Computadora, 2017. Profesor de Informática II en la Facultad de Ingeniería, Universidad de Holguín, Cuba. Investigador y miembro del Centro de Estudios CAD/CAM en la misma facultad.