Discussion paper / Artículo de reflexión / Artigo de reflexão - Tipo 2

Conditional random fields in text segmentation by language

Robin Cabeza Ruiz / robbinc91@uho.edu.cu

Universidad de Holguín, Cuba.

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. Research let conclude that conditional random fields are a powerful tool for segmentation of multilingual text.

KEYWORDS Text segmentation by language; conditional random fields.

Los campos aleatorios condicionales en la segmentación de textos por idioma

RESUMEN En este trabajo se propone la utilización de los campos aleatorios condicionales para la resolución de la tarea de segmentación de textos por idioma, considerándola como una tarea de etiquetado de secuencias. La metodología considera que el cambio entre un idioma y otro en los documentos ocurrirá en cualquier parte del texto y asume que las observaciones en el sistema estarán dadas por las palabras en el texto y que los estados serán los diferentes idiomas. De la investigación se concluye que los campos aleatorios condicionales son una herramienta muy poderosa para la segmentación de textos multilingües.

PALABRAS CLAVE Segmentación de textos por idiomas; campos aleatorios condicionales.

Campos aleatórios condicionais na segmentação de textos por idioma

RESUMO Neste artigo, é proposto o uso de campos condicionais aleatórios para a resolução da tarefa de segmentação de textos por idioma, considerando-a como uma tarefa de marcação de sequências. A metodologia considera que a mudança entre um idioma e outro nos documentos ocorrerá em qualquer parte do texto e pressupõe que as observações no sistema serão dadas pelas palavras no texto e que os estados serão os diferentes idiomas. Conforme os resultados da pesquisa, conclui-se que os campos aleatórios condicionais são uma ferramenta muito poderosa para a segmentação de textos multilíngues.

PALAVRAS-CHAVE Segmentação de textos por idiomas; campos aleatórios condicionais.

Received / Recepción: Diciembre 1, 2017 - Accepted / Aceptación: Diciembre 30, 2017 *doi: 10.18046/syt.v15i43.2712*

I. Introduction

Since the emergence of the Internet and databases, the number of sources of information in the form of text available on the Web has grown in a dramatic way, mainly in news sites, blogs, social networks, etc., which causes, a constant growth of data in text format. This accumulation of data is impossible to qualify or analyze by people in an efficient manner. So, computer tools capable of doing it in an automated way have been created.

Text mining is the research field dedicated to obtaining new and valuable information from these documents. It includes some tasks such as the grouping of documents according to their themes or their classification in predefined themes.

Text mining techniques usually represent documents as vectors of terms, where generally these terms are nouns, verb forms, etc., according to the task to be solved. This selection of terms is enhanced through the *Natural Language Processing* [NLP] (Vásquez, Quispe, & Huayana, 2009), a field of computer science, artificial intelligence, and linguistics that studies the interactions between computers and human language.

Many of the NLP techniques are based on the language, that is, it is necessary to know beforehand in what languages the documents are written in order to be able to process them. Therefore, applying previously language identification techniques is an essential phase.

The identification of languages aims to identify the language (or languages) available in a document. Taking into account that an existing document on the web can contain segments written in several languages, it is necessary to identify those languages in multilingual documents.

The identification of the languages available in a document may not be sufficient, it is more useful to have tools capable of obtaining the segments of each of the languages include in it. This task is called text segmentation by language.

This investigation is focused on the second variant and aims to provide a new method to identify in an efficient manner what languages are included on it and what portions of the text are written in each one of them, during the analysis of a document.

The first of these variants has been addressed in several studies, including those by Lui, Lau, and Baldwin (2014); Singh and Gorla (2007); and Baldwin and Lu

I. Introducción

Desde el surgimiento de la 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, en sitios de noticias, blogs, redes sociales, etc., lo que provoca, como es evidente, un constante crecimiento de datos en formato de texto. Este cúmulo de datos es imposible 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 es el área de investigación dedicada a la obtención de información novedosa y valiosa de estos documentos. Dentro de ella existen tareas como el agrupamiento de documentos de acuerdo con su temática o su clasificación en temáticas predefinidas.

Las técnicas de minería de textos 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 realza utilizando el Procesamiento de Lenguaje Natural [PLN] (Vásquez, Quispe, & Huayana, 2009), 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 poder procesarlos, por lo que aplicar previamente técnicas de identificación de idiomas resulta una fase imprescindible.

La identificación de idiomas tiene como objetivo la identificación del idioma (o idiomas) presente en un documento. Teniendo en cuenta que un documento existente en la web puede contener segmentos escritos en varios idiomas, es necesario realizar la identificación de esos idiomas en documentos multilingües.

Identificar los idiomas presentes en un documento puede no ser suficiente, resulta más útil contar con herramientas capaces de obtener los segmentos de cada uno de los idiomas presentes en él. Esta tarea se denomina segmentación de textos por idiomas.

La presente investigación se centra en la segunda variante y tiene como objetivo de aportar un nuevo método para, al analizar un documento, identificar, de manera eficiente qué idiomas están presentes en él mismo y qué porciones del texto están escritas en cada uno de ellos.

La primera de estas variantes ha sido tratada en variados estudios, entre ellos los de Lui, Lau, y Baldwin (2014); Singh y Gorla (2007); y Baldwin y Lu (2010). La segunda, en cambio, se trata en muy pocas investigaciones, entre ellas, las de Yamaguchi y Tanaka-Ishii (2012); y Cabeza (2016).

Ruiz utilizó los modelos ocultos de Markov [*Hidden Markov Models*, 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) los idiomas en los que ha sido redactado.

Para el etiquetado de secuencias, los campos aleatorios condicionales [*Conditional Random Fields,* CRF] han resultado más poderosos que los HMM gracias a las posibilidades que brindan de caracterización de sistemas más caóticos y con solapamiento de clases, como es el caso de un documento escrito en varios idiomas, más aún considerando que existen palabras que pertenecen a más de un idioma: por ejemplo, "hospital" que puede estar escrita en 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 que existe 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.

II. Estado del arte

Yamaguchi y Tanaka-Ishii (2012) proponen un método que utiliza el concepto de Descripción de Longitud Mínima [*Minimal Description Length,* MDL]] (Barron, Rissanen, & Yu, 1998) para encontrar los "bordes" de los idiomas, es decir los 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.

Cabeza (2016) utiliza HMM 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. El autor tomó: como observaciones, a las palabras existentes dentro de los textos; y como etiquetas (o estados), al conjunto de idiomas disponibles para la identificación y segmentación. Además, el autor propone 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 del *Natural Language Toolkit* [NLTK] (Bird, 2006), para segmentar los documentos por oraciones, y el identificador de idiomas monolingüe LangID (Cook & Lui, 2012) para obtener el idioma de cada oración.

III. Materiales y métodos

En esta sección está descrito el funcionamiento del algoritmo propuesto para la segmentación de textos por idiomas. Se comienza con la introducción a los CRF.

A. Los campos aleatorios condicionales

Un CRF es un modelo utilizado habitualmente para etiquetar secuencias de datos o extraer información de documentos (Lafferty, McCallum, & Pereira, 2001). Los CRF han sido utilizado exitosamente en áreas de procesamiento de textos (Peng & McCallum, 2004; Settles, 2005; Sha & (2010). The second, on the other hand, is treated in very few researches, among them, those of Yamaguchi and Tanaka-Ishii (2012); and Cabeza (2016).

Ruiz used the Hidden Markov Models [HMM] to solve the task, considering the documents as a set of observations in which a set of labels is assigned in the classification process. In this case, the observations would be the words included in the text and the labels (or stages) would be the languages in which it was written.

For the labeling of sequences, the Conditional Random Fields [CRF] have been more powerful than the HMM thanks to characterization possibilities of more chaotic systems and with overlapping of classes, such as a written document in several languages, even more considering that there are words that belong to more than one language: for example, "hospital" that can be written in Spanish or English.

Considering the improvements that the use of CRF can offer to the labeling of sequences, the present research has as an objective to propose these models for its application to the task of texts segmentation by languages, taking into account the compromise that exists between the system efficiency and the amount of RAM available on the computer that is intended for execution. The formulation of CRF allows affirming, from the point of view of the author, that this is a model that will improve the results obtained until the date of the task resolution.

II. State of the art

Yamaguchi and Tanaka-Ishii (2012) propose a method that uses the concept of Minimum Length Description [CDM] (Barron, Rissanen, and Yu, 1998) to find the "edges" of languages, that is, the sites where changes occur between languages within a document. The authors formulate the problem in the following way: Given a text X, the segments are obtained for a list of edges B in correspondence with a list of languages L.

Cabeza (2016) uses HMM to treat a document as a sequence of observations to which a sequence of labels can be assigned, which is the most likely. The author took: as observations, the existing words within the texts; and as labels (or stages), the set of languages available for identification and segmentation. In addition, the author proposes a way to obtain the languages in documents in which the change of language can occur only in the jumps between the sentences or the paragraphs within the documents, through the library of the Natural Language Toolkit [NLTK] (Bird, 2006). The above in order to segment the documents by sentences, and

the monolingual language identifier LangID (Cook & Lui, 2012) with the aim of obtaining the language of each sentence.

III. Materials and methods

This section describes the functioning of the algorithm proposed for the texts segmentation by languages. It begins with the introduction to the CRF.

A. Conditional random fields

A CRF is a model commonly used to label data sequences or extracts information from documents (Lafferty, McCallum, & Pereira, 2001). CRFs have been used successfully in areas of word processing (Peng & McCallum, 2004, Settles, 2005, Sha & Pereira, 2003), bioinformatics (Liu, Carbonell, Weigele, & Gopalakrishnan, 2006), and computer vision (He, Zemel, & Carreira-Peripiñián, 2004).

The main difference between the HMMs and the CRFs is that the HMM calculate the conditional probability of the labels given the input variables, $p(z|x)$, while the CRFs calculate the joint probability of both, $p(z,x)$. In addition, CRFs have access to more observations than HMMs. It can be represented as a non-directed graph $G = (V,E)$, which defines a linear distribution of a set of labels for a given sequence of observations, in which each vertex represents a random variable whose probability distribution must be deduced, and each edge indicates a dependence between the variables of the vertices it connects. The graph satisfies the Markov property extended to graphs (see Equation 1).

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

Where ~ means that S_i, S_j are connected by an edge. Regarding O_i observations, it is most often a vector, instead of a scalar value, having multidimensional observations.

The CRF models the probability of a sequence of z labels given a sequence of observations x in the way described in Equation 2.

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

Where N is the number of observations, and F the number of characteristic functions defined for the operation of the model. Also, Z is called the normalization factor or partition function, which is calculated by equation 3.

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

Pereira, 2003), bioinformática (Liu, Carbonell, Weigele, & Gopalakrishnan, 2006), y visión por computadora (He, Zemel, & Carreira-Peripiñián, 2004)

La principal diferencia entre los HMM y los CRF es que los HMM calculan la probabilidad condicional de las etiquetas dadas las variables de entrada, $p(z|x)$, mientras que los CRF calculan la probabilidad conjunta de ambas, $p(z,x)$. Además, los CRF tienen acceso a más observaciones que los 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 (ver Ecuación 1).

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

Donde ~ significa que Si, Sj están conectados por una arista. En cuanto a las observaciones Oi, lo más frecuente es que sea un vector, en vez de un valor escalar, teniendo observaciones multidimensionales.

Los CRF modelan la probabilidad de una secuencia de etiquetas z dada una secuencia de observaciones x de la manera expresada en la Ecuación 2.

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

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, que se calcula mediante la ecuación 3.

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

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 la ecuación 4, en la que "ES" representa el código utilizado para el idioma español, y "texto" representa la palabra "texto" del idioma español.

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

La utilización de este rasgo depende de su correspondiente peso λ_i. Si $\lambda_i > 0$, cuando esté activa f_i(es decir, cuando en 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: los CRF preferirían la etiqueta ES para la palabra "texto". Por otro lado, si $\lambda_i < 0$, los CRF evitarán dicha etiqueta para la palabra en cuestión.

B. Entrenamiento de un campo aleatorio condicional

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

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

IV. Segmentación de textos por idiomas utilizando campos aleatorios condicionales

Un 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.

A. Funciones características

Estas funciones son 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: de transición, de observación y de valores de la etiqueta.

Funciones de transición (Ecuación 6)

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

Donde, $t(z_{n-1},z_n)$ 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, $p(z_n|z_{n-1})$ (Cabeza, 2016).

Funciones de observación (Ecuación 7)

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

Donde, words[zn] es el conjunto de todas las palabras observadas a partir del idioma zn, en el conjunto de textos utilizados para el entrenamiento; y x[n] es la palabra que se analiza "actualmente" en el texto de prueba.

The characteristic functions are the main component of a CRF. In general, a characteristic function has the form $f_i(z_{n-1},z_n,x_{1:N},n)$, and searches in a pair of adjacent stages z_{n-1},z_n, the complete sequence $z_{1:N}$, and in which part of the sequence the model is found at that same instant of time n. These arbitrary functions produce real values.

An example of a characteristic function for texts segmentation by languages is shown in equation 4, in which "ES" represents the code used for the Spanish language, and "texto" represents the word "text" of the Spanish language.

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

The use of this feature depends on its corresponding weight λ_i. If $\lambda_i > 0$, when it is active f_i (that is when the word "texto" is being analyzed and the ES label is assigned) it increases the probability of the label sequence $z_{1:N}$. In other words: CRFs would prefer the ES label for the word "texto". On the other hand, if λ_i <0, CRFs will avoid that label for the word in question.

B. Training a conditional random field

The training of a CRF consists in finding the parameters that maximize the probability of the sequences of training labels for the observations, which is obtained through equation 5. It is necessary to have sequences of observations previously labeled $\{(x^{(1)},z^{(1)}),\ldots,(x^m,z^m)\}$, where $x^{(1)} = x^{(1)}_{1:N}$.

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

IV. Texts segmenting by language through conditional random fields

A CRF needs a set of sequences of observations for its training (in this study, the observations are the words within the texts), obtained from the corpus training. Additionally, it is necessary to define the characteristic functions of the system. The creation of these functions is explained below.

A. Characteristic functions

These functions are the most important point of defining a CRF. For the task of texts segmenting by language, the characteristic functions can be generated from the union of three sets of auxiliary functions: transition, observation and label values.

Transition functions (Equation 6)

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

Where, $t(z_{n-1}, z_n)$ is 1 if in the training appears the transition from language z_{n-1} to language z_n. These functions favor the appearance of the language z_{n-1} followed by the language z_n in any text. The word that is in the current position of the text string is not taken into account.

This set of functions is similar to the transition matrix A used by HMM, $p(z_n|z_{n-1})$ (Cabeza, 2016).

Observation functions (Equation 7)

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

Where, words [zn] is the set of all the words observed from the zn language, in the set of texts used for the training; and x [n] is the word that is "currently" analyzed in the test text.

These functions are created with the aim of encouraging the punctuation of a word from the language in which they were generated within the training documents.

Label values functions (Equation 8)

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

In equation 8 the input sequence is favored when it comes to the word xp and the zp language is being analyzed, both considered as extra parameters. These functions, together with the observation functions, play a role equivalent to the observation matrix used by the HMMs, which keep the probability of emitting the word xp from the zp language.

The described functions allow creating a CRF model for the task of texts segmentation by languages. Therefore, you can train a CRF using a series of documents with their respective metadata (the language of each word in the texts).

It should be noted that, although CRFs are models that can provide more benefits than other systems, such as HMMs, it is necessary to consider the consumption of memory (and time) that their use may represent. The characteristic functions increase the perfection degree of the model, but a boundary must be found in this way, since many of these functions are synonymous of the system delay, due to the model must evaluate more possibilities for the labeling of the input sequence.

The model was successfully tested at the laboratory level. However, due to limitations in the computing infrastructure, only small-scale tests of the algorithm could be executed. To test its performance, the algorithm was trained with a very small subset of the Wikipedia-multi corpus used in Lui et al., (2014) and Cabeza (2016). The algorithm showed to be efficient in the training used.

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 (Ecuación 8)

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

En la ecuación 8 se favorece la secuencia de entrada cuando se trata de la palabra xp y se está analizando el idioma zp, ambas pasadas como parámetros extra. Estas funciones, junto con las funciones de observación, juegan un papel equivalente a la matriz de observaciones utilizada por los HMM, que guardan la probabilidad de emitir la palabra xp a partir del idioma zp.

Las funciones descritas 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 los CRF constituyen 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 del sistema, ya que del modelo debe evaluar más posibilidades para el etiquetado de la secuencia de entrada.

El modelo fue probado con éxito a nivel de laboratorio. Sin embargo, debido a limitaciones en la infraestructura de cómputo, solo se pudieron ejecutar pruebas a pequeña escala del algoritmo. Para probar su funcionamiento se entrenó el algoritmo con un subconjunto ínfimo del corpus Wikipedia-multi utilizado en Lui et al., (2014) y Cabeza (2016). El algoritmo mostró ser eficiente en el entrenamiento utilizado.

V. Conclusiones y trabajo futuro

Los campos aleatorios condicionales son una herramienta muy poderosa para el etiquetado de secuencias. Como se dijo, 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 cual, al igual que cuando se utilizan 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 Campos Aleatorios de Markov [*Markov Random Fields*], en aras de hallar vías más eficientes para la resolución de la tarea. S&T

V. Conclusions and future works

Conditional random fields are a very powerful tool for labeling sequences. As mentioned, they have been widely used in text processing, bioinformatics, and computer vision fields.

The formulation of CRFs allows us to affirm that they can be a useful tool for a multilingual texts segmentation by languages, taking this task as a problem of labeling sequences, in which, in the same way as using the Hidden Markov Models, the stages represent the languages in which a document is (or may be) written, and the words in the text constitute the observations of the system.

For future research, we intend to test the algorithm on a larger scale, to make real comparisons with other methods used to segment texts by language. It is also thought to explore the so-called Markov Random Fields, in order to find more efficient ways to solve the task. S&T

References / Referencias

Baldwin, T. & Lu, M. (2010). Multilingual language identification: ALTW 2010 shared task dataset. In *Proceedings of the Australasian Language Technology Association Workshop* (pp. 4-7).

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory, 44*(6), 2743-2760.

Bird, S. (2006). NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions,* (pp. 69-72). Stroudsburg, PA: Association for Computational Linguistics.

Cabeza, R. (2016). Text segmentation by language. *Sistemas & Telemática,14*(38), 65-74. doi 10.18046/syt.v14i38.2289

Cook, P. & Lui, M. (2012). langid.py for better language modelling. In *Proceedings of the Australasian Language Technology Association Workshop,* (pp. 107-112).

He, X., Zemel, R. S., & Carreira-Perpiñán, M. Á. (2004). Multiscale conditional random fields for image labeling. In *Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on* (Vol. 2, pp. II-II). IEEE.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic model for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning,* (pp. 282-289).

Liu, Y., Carbonell, J., Weigele, P., & Gopalakrishnan, V. (2006). Protein fold recognition using segmentation conditional random fields (SCRFs). *Journal of Computational Biology, 13*(2), 394-406.

Lui, M., Lau, J. H., & Baldwin, T. (2014). Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics, 2,* 27-40.

Peng, F. & McCallum, A. (2004). Accurate information extraction from research papers using conditional random fields. In *Human Language Texhnology Conference and North American Chapter of the Association for Computational Linguistics.* Retrieved from: https://people.cs.umass.edu/~mccallum/papers/hlt2004.pdf

Settles, B. (2005). Abner: An open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics, 21*(14), 3191-3192.

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. In *NAACL '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology,* (Vol. 1, pp. 134-141). Stroudsburg, PA: Association for Computational Linguistics.

Singh, A. K., & Gorla, J. (2007). Identification of languages and encodings in a multilingual document. In *Building and Exploring Web Corpora (WAC3-2007): Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval* (Vol. 4, p. 95). Louvain, Belgium: Louvain Université.

Vásquez, A. C., Quispe, J. P., & Huayana, A. M. (2009). Procesamiento de lenguaje natural. *Revista de Investigación de Sistemas e Informática, 6*(2), 45-54.

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

CURRICULUM VITAE

Robin Cabeza Ruiz. Master in Design Assisted by Computer from the Universidad de Holguín (Cuba, 2015) with a bachelor's degree in Computer Science from Universidad de Oriente (Cuba, 2017). Currently he is professor of informatics II and member of CAD/CAM Studies Center at the Faculty of Engineering at the Universidad de Holguín. His main areas of interest in research are biomechanical and text segmentation by computer / Licenciado en Ciencias de la Computación, graduado en la Universidad de Oriente (Cuba, 2015), y Máster en Diseño Asistido por Computadoras de la Universidad de Holguín (Cuba, 2017). Actualmente es profesor en la Facultad de Ingeniería de la Universidad de Holguín, donde imparte la asignatura de Informática II. Pertenece al centro de estudios CAD/CAM de la misma universidad. Sus intereses en investigación están focalizados en biomecánica y segmentación de textos por computador.