Original Research / Artículo Original - Tipo 1

Text segmentation by language

Robin Cabeza Ruiz / druizolivero@gmail.com

Universidad de Holguín, Cuba

ABSTRACT There are two approaches for text segmentation by language: first, assuming that language changes happen in the "border" between sentences (never within a sentence); second, assuming that language changes can happen anyplace in the text. This work presents methods for both types of text's segmentation by languages. On the first proposal, the text is initially segmented by sentence, then the language of each sentence is obtained; the second proposal is an adaptation of hidden Markov model to this task. Both cases, according to results obtained in experimental proofs, exceed the state of art.

KEYWORDS Hidden Markov model; text segmentation by language; natural language processing.

Segmentación de textos por idioma

RESUMEN La segmentación de textos por idioma puede ser abordada de dos maneras: la primera, asumiendo que los cambios de idioma solo ocurren en los saltos entre oraciones; y la segunda, asumiendo que el idioma puede cambiar en cualquier lugar del texto. En este trabajo se presentan métodos para segmentar ambos tipos de textos por idiomas. Para el primer caso se segmenta el texto a analizar por oraciones y luego se identifica el idioma de cada oración; la segunda propuesta consiste en la adaptación de los modelos ocultos de Markov a la tarea de segmentación de textos por idiomas. El estado del arte es superado por ambas propuestas, según los resultados obtenidos en la experimentación realizada.

PALABRAS CLAVE Modelos ocultos de Markov; segmentación de textos por idioma; procesamiento del lenguaje natural.

Segmentação de textos por idioma

RESUMO A segmentação de textos por idioma pode ser abordada de duas maneiras: a primeira, assumindo que as alterações da linguagem ocorrem apenas nos saltos entre as frases; e a segunda, partindo do princípio que o idioma pode mudar em qualquer parte do texto. Este trabalho apresenta métodos para segmentar ambos os tipos de textos por idioma. No primeiro caso, o texto é segmentado para analisar frases e, em seguida, identifica-se a língua de cada frase; a segunda proposta consiste na adaptação dos modelos ocultos de Markov à tarefa de segmentação de textos por idioma. O estado da arte é ultrapassado por ambas as propostas, de acordo com os resultados obtidos na experimentação realizada.

PALAVRAS-CHAVE Modelo oculto de Markov; segmentação de textos por idioma; processamento de linguagem natural.

Received / Recepción: Septiembre 5, 2016 - Accepted / Aceptación: Septiembre 27, 2016 *doi 10.18046/syt.v14i38.2289*

I. Introduction

The development currently reached on the web has led to the emergence of many sources of information in text form (for example news sites, blogs, social networks, etc.), so the volume of data available in textual form is growing. The Text Mining [TM] is the research area dedicated to obtain valuable new information from these texts. To apply techniques of Text Mining, documents are usually represented as vectors of terms where the terms are usually nouns, verbal forms or adjectives, according to the task to be solved. For the extraction and selection of terms are used Natural Language Processing [NLP] techniques (Vasquez, Quispe, & Huayana, 2009). Many NLP techniques are language-dependent, so the Language Identification [LI] is an essential phase in the preprocessing of the texts, and aims to identify the language in which it is written a given document.

Most of the work reported in the literature base their operation on the LI in monolingual documents, only one language per document. However, existing documents on the web can be written in more than one language; for example, many scientific articles contain a summary in English and the rest in the main language of the article. Examining such documents, traditional algorithms of LI often fail, give the language most represented in the texts, or abstain.

Among the works reported in the literature in the LI area in multilingual documents, some only identify the languages present in the text –Languages Identification in Multilingual Documents [LIMD]– (Lui, Lau, & Baldwin, 2014), while others also identify text segments for each language –Text Segmentation by Languages [TSL]– (Yamaguchi & Tanaka-Ishii, 2012).

A tool prepared for the LIMD, when analyzing the example string "english string para prueba", should only recognize that English and Spanish languages are present; while one prepared for TSL should return: [("english string", inglés), ("para prueba", español)]. This work focuses on the second variant, because it will allow applying TM tools once the texts are segmented by languages.

TSL can target a sequences labeling task, since it is possible to analyze a text as a sequence of features (words, for example), in which each feature has assigned a tag (that is, each word is written in a language). For example, in the text string above, an algorithm prepared for the task should return [("english", en), ("string", en), ("para", es), ("prueba", es)]. Among the most popular algorithms for sequence

I. Introducción

El desarrollo alcanzado actualmente en la web ha motivado la aparición de muchas fuentes de información en forma de texto (ya sean sitios de noticias, blogs, redes sociales, etcétera), por lo que el volumen de datos disponible en forma textual es cada vez mayor. La Minería de Textos [MT] es el área de investigación que se dedica a obtener información valiosa y novedosa de estos textos. Para poder aplicar técnicas de MT suelen representarse los documentos como vectores de términos, donde los términos generalmente son sustantivos, formas verbales o adjetivos, de acuerdo con la tarea a resolver. Para la extracción y selección de los términos se utilizan técnicas de Procesamiento de Lenguaje Natural [PLN] (Vásquez, Quispe, & Huayana, 2009). Muchas técnicas de PLN son dependientes del idioma, por lo que la Identificación de Idiomas [II] es una fase imprescindible en el preprocesamiento de los textos, y tiene como objetivo identificar el idioma en que está escrito un documento dado.

La mayoría de los trabajos reportados en la literatura basan su funcionamiento en la II en documentos monolingües, solo un idioma por cada documento. No obstante, los documentos existentes en la web pueden estar escritos en más de un idioma; por ejemplo, muchos artículos científicos contienen un resumen en inglés y el resto en el idioma principal del artículo. Al analizar este tipo de documentos, los algoritmos tradicionales de II suelen fallar, dar el idioma más representado en los textos, o bien abstenerse.

Entre los trabajos reportados en la literatura en el área de II en documentos multilingües, algunos solo identifican los idiomas presentes en el texto –Identificación de Idiomas en Documentos Multilingües [IIDM]– (Lui, Lau, & Baldwin, 2014), mientras que otros identifican además los segmentos de texto para cada idioma –Segmentación de Textos por Idiomas, [STI]– (Yamaguchi & Tanaka-Ishii, 2012).

Una herramienta preparada para la IIM, al analizar la cadena de ejemplo "english string para prueba", debería solamente reconocer que están presentes los idiomas inglés y español; mientras que una preparada para la STI debería retornar: [("english string", inglés), ("para prueba", español)]. Este trabajo se enfoca en la segunda variante, pues es la que permitirá aplicar herramientas de MT una vez segmentados los textos por idiomas.

STI puede orientarse a una tarea de etiquetado de secuencias, ya que es posible analizar un texto como una secuencia de rasgos (palabras, por ejemplo), en la que cada rasgo tiene asignada una etiqueta (es decir, cada palabra está escrita en un idioma). Por ejemplo, en la cadena de texto anterior, un algoritmo preparado para la tarea debería retornar [("english", en), ("string", en), ("para", es), ("prueba", es)]. Entre los algoritmos más populares para el etiquetado de secuencias se encuentran los modelos ocultos de Markov, que son los que se propone utilizar en este trabajo, motivado por los buenos resultados obtenidos por esta técnica en otras tareas de PLN.

II. Estado del arte

El principal trabajo reportado en la literatura para la STI es el propuesto por Yamaguchi y Tanaka-Ishii (2012). 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). En su trabajo se denota un texto multilingüe a ser segmentado X como una secuencia de caracteres $x_1,\ldots,x_{|X|}$, donde x_i es el i-ésimo carácter. La segmentación se realiza encontrando el conjunto de bordes $B=B_1,\ldots,B_{|B|}$, donde cada B_i denota el i-ésimo borde como el número de caracteres desde el principio del texto. La lista de segmentos obtenida de B es denotada como $X=[X_0,\ldots,X_{|B|}$, donde la concatenación de estos elementos es igual al texto X. Finalmente, $L=L_1,\ldots,L_{|B|}$ denota la secuencia de los idiomas contenidos en el texto.

Los autores formulan el problema de la siguiente manera (1): dado un texto X, se obtienen los segmentos X para un listado de bordes B en correspondencia con un listado de idiomas L. La longitud de descripción es calculada obteniendo la longitud de descripción de cada segmento X_i para el idioma L_j :

$$(\hat{\mathbb{B}},\hat{\mathbb{L}}) = argmin_{x,L}\sum_{i=0}^{\mathbb{B}} dl_{L_j}(X_i) \quad (1)$$

La función $dl_{Lj}(X_i)$ calcula la longitud de descripción para un texto X_i dado un idioma L_j , ver (2)

$$\hat{dl}_{L_j}(X_i) = argmin_{L_j\in\mathfrak{L}}(-\log_2 P_{L_j}(X_i) + \log_2(|X|) + \log_2(|\mathfrak{L}|) + \gamma) \quad (2)$$

donde el primer término corresponde a la entropía cruzada del texto X_i para el idioma L_j , y el resto de los términos son utilizados para describir los parámetros utilizados. Estos son constantes, por lo que se puede utilizar solamente la entropía cruzada.

Los autores proponen dos métodos para el cálculo de la entropía entre dos segmentos de texto: predicción por coincidencias parciales (Witten & Bell, 1991) y media de estadísticas de coincidencia (Juola, 1997).

III. Materiales y métodos

En esta sección se describen los algoritmos propuestos para la tarea de segmentación de textos por idiomas. Se comenzará con una introducción a los modelos ocultos de Markov, para luego proceder a explicar los algoritmos.

A. Modelos ocultos de Markov

Un modelo oculto de Markov (Blunsom, 2004; Ghahramani, 2001) es un modelo estadístico en el que se asume que el sistema a modelar es un proceso de Markov de parámetros desconocidos. El objetivo es determinar los parámetros desconocidos (u ocultos, de ahí el nombre) de dicha cadena a partir de los parámetros observables.

labeling are hidden Markov models, which are proposed to be used in this work, motivated by the good results obtained by this technique in other NLP tasks.

II. State of the art

Yamaguchi and Tanaka-Ishii (2012) proposed the main work reported in the literature for the TSL. The authors propose a method that uses the concept of Minimum Description Length (Barron, Rissanen, & Yu, 1998) to find the borders of languages (sites where changes occur between languages). In their work, it is denoted a multilingual text to be segmented X as a sequence of characters $x_1,\ldots,x_{|X|}$, where x_i is the i-th character. The segmentation is performed by finding the set of borders $B=[B_1,\ldots,B_{|B|}]$, where each B_i denotes the i-th border as the number of characters from the beginning of the text. The list of segments obtained from B is denoted as $X=[X_0,\ldots,X_{|B|}]$, where the concatenation of these elements is equal to text X. Finally, $L=[L_1,\ldots,L_{|B|}]$ denotes the sequence of the language contained in the text.

The authors formulate the problem as follows (1): given a text X, the segments X for a list of borders B are obtained in correspondence with a list of languages L. The description length is calculated by obtaining the description length of each segment X_i for the language L_j :

$$(\hat{\mathbb{B}},\hat{\mathbb{L}}) = argmin_{x,L}\sum_{i=0}^{\mathbb{B}} dl_{L_j}(X_i) \quad (1)$$

The function $dl_{Lj}(X_i)$ calculates the description length of a text X_i given a language L_j, see (2)

$$\hat{dl}_{L_j}(X_i) = argmin_{L_j\in\mathfrak{L}}(-\log_2 P_{L_j}(X_i) + \log_2(|X|) + \log_2(|\mathfrak{L}|) + \gamma) \quad (2)$$

where the first term corresponds to the cross entropy of the text X_i for the language L_j, the remaining terms are used to describe the parameters used. These are constant, so only cross entropy can be used.

The authors propose two methods for calculating the entropy between two text segments: prediction by partial matching (Witten & Bell, 1991) and mean of matching statistics (Juola, 1997).

III. Materials and methods

In this section, are described the proposed algorithms for the text segmentation task by language. It will begin with an introduction to hidden Markov models, and then proceed to explain the algorithms.

A. Hidden Markov Models (HMM)

A hidden Markov model (Blunsom, 2004; Ghahramani, 2001) is a statistical model that assumes that the system to be modeled is a Markov process of unknown parameters. The objective is to determine the unknown parameters (or hidden, hence the name) of the chain from observable parameters.

In a conventional Markov Model (Rincón, 2008), the state is directly visible to the observer, so the transition probabilities between states are the only parameters. In a hidden Markov model, the state is not directly visible, but the possible output tokens. Consequently, the sequence of tokens generated by an HMM gives some information about the sequence of states, so it is possible to determine the most likely set of states, given a set of observable variables generated by themselves.

- A hidden Markov model is a 5-tuple consisting of:
- set of states;
- set of possible observations;
- initial probabilities $\{\pi_1,\pi_2,\ldots,\pi_N\}$. $p(q_0=S_i)=\pi_i$;
- matrix of transitions between different states A, $a_{ij}=p(q_{(t+1)}=S_j \mid q_t=S_i)$ is the probability of changing from state S_i to state S_j; and
- matrix of observations B (the emission probability of each possible observation, from each of the states): $b_i(k)=P(O_t=k \mid q_t=S_i)$.

B. Text segmentation by language using hidden Markov models

One of the questions answered within the HMM is what is the most likely sequence of states for which transited the model, given a sequence of observations? To solve the tasks of STL using HMM is necessary to define what the states and observations are. In this work the "states" are the languages for which has been trained the model, and the "observations" are the features used (in this case the words).

For clarity, it is shown the following example: for the string "*english string para prueba*", using words like features (observations), the algorithm should return: [("*english", en*), (*"string", en*), (*"para", es*), (*"prueba", es*)]. Note that [] refers to a list, the words in quotation marks (" ") are segments of the text passed as parameter, and the rest of the words (in this case are only *en, es*) are the ISO language code assigned to each segment; being "en" the code for English language and "es" the code for Spanish language. The result is the most likely language way to

En un Modelo de Markov convencional (Rincón, 2008), el estado es visible directamente para el observador, por lo que las probabilidades de transición entre estados son los únicos parámetros. En un modelo oculto de Markov, el estado no es visible directamente, sino que sólo lo son los posibles símbolos de salida. Consecuentemente, la secuencia de símbolos generada por un HMM proporciona cierta información acerca de la secuencia de estados, por lo que es posible determinar el conjunto de estados más probable, dado un conjunto de variables observables generadas por los mismos.

Un modelo oculto de Markov es una 5-tupla que consta de:

- el conjunto de estados;
- el conjunto de posibles observaciones;
- las probabilidades de inicio $\{\pi_1,\pi_2,\ldots,\pi_N\}$. $p(q_0=S_i)=\pi_i$;
- la matriz de transiciones entre los diferentes estados A, $a_{ij}=p(q_{(t+1)}=S_j \mid q_t=S_i)$ es la probabilidad de cambiar del estado S_i al estado S_j; y
- la matriz de observaciones B (la probabilidad de emisión de cada una de las posibles observaciones, a partir de cada uno de los estados): $b_i(k)=P(O_t=k \mid q_t=S_i)$.

B. Segmentación de textos por idiomas utilizando modelos ocultos de Markov

Una de las preguntas a las que se les da respuesta dentro de los HMM es: ¿cuál es la secuencia de estados más probable por la que transitó el modelo, dada una secuencia de observaciones? Para resolver las tareas de STI utilizando HMM es necesario definir cuáles serán los estados y las observaciones. En el presente trabajo los "estados" serán los idiomas para los cuales ha sido entrenado el modelo, y las "observaciones" serán los rasgos utilizados (en este caso las palabras).

Para mayor claridad, se muestra el siguiente ejemplo: para la cadena "*english string para prueba*", utilizando las palabras como rasgos (observaciones), el algoritmo debe retornar: [(*"english", en*), *("string", en*), (*"para", es*), (*"prueba", es*)]. Advertir que [] hace referencia a una lista, las palabras entre comillas (" ") son segmentos del texto pasado como parámetro, y el resto de las palabras (en este caso solo son *en, es*) son el código ISO del idioma asignado a cada segmento (código de dos letras para identificar los principales idiomas del mundo, norma desde 2002). "en" es el código para el idioma inglés y "es" el código para español (). El resultado obtenido es el camino de idiomas más probable para la cadena de prueba.

Los parámetros obtenidos para la segmentación con HMM son: probabilidades de inicio, probabilidades de emisión y probabilidades de transición.

Probabilidades de inicio

Probabilidad de que un texto comience con cada uno de los idiomas disponibles. Se asume que todos los idiomas son equiprobables al empezar un documento (cada autor puede empezar a redactar un texto con el idioma que desee); se

asigna la misma probabilidad 1/N a cada idioma, donde N es la cantidad de idiomas disponibles para el entrenamiento.

Probabilidades de emisión

Matriz que guarda la probabilidad de emitir el rasgo (la palabra) O a partir del estado (idioma) S_i. Estas probabilidades fueron extraídas de los textos del entrenamiento, la probabilidad de emisión del idioma L_i para el rasgo O_j es la cantidad de veces que aparece el rasgo O_j en el idioma L_i, dividido por la cantidad total de rasgos que aparecen en textos escritos en el idioma L_i.

Probabilidades de transición

Probabilidad de pasar de un idioma a otro en un mismo documento. Fueron calculadas de dos maneras (4) y (5):

$$L_{ij}/L_{in} \quad (4)$$

donde, L_{ij} es el número de veces que es vista la transición de L_i a L_j en el corpus, y L_in es la cantidad de saltos existentes en el corpus a partir del idioma L_i a cualquier otro idioma; y

$$p_t(L_i) * p_e(L_j) \quad (5)$$

donde $p_t\ (L_i)$ es la probabilidad de terminar un texto en el idioma L_i, y $p_e\ (L_j)$ es la probabilidad de comenzar la próxima palabra con el idioma L_j. Para esto se almacenan en tablas auxiliares las probabilidades de empezar y terminar en cada idioma con cada una de las palabras.

C. Segmentación de textos por idiomas identificando el idioma de cada oración

Para el funcionamiento de este método (*multilangid*) se tiene una unidad mínima en la cual se asume que el idioma no cambia; dicha unidad mínima será la oración. La segmentación por idiomas se realiza en dos etapas: obteniendo las oraciones del texto e identificando el idioma de cada oración.

Para la segmentación de un documento por oraciones se utilizó la biblioteca NLTK [Natural Language Tool Kit] (Bird, 2006), una plataforma para construir programas en lenguaje de programación Python para el Procesamiento de Lenguaje Natural [PLN] –campo de las ciencias de la computación, la inteligencia artificial y la lingüística, que estudia las interacciones entre las computadoras y el lenguaje humano– . La biblioteca basa su funcionamiento en una lista de símbolos de puntuación que pueden definir bordes entre oraciones ('.', '?', '!'). El algoritmo puede ser entrenado en caso de que se quiera aumentar la cantidad de símbolos que definan cambios de oración en un texto. Sin embargo, para este trabajo no se le realizó ningún entrenamiento.

Langid, la herramienta utilizada para identificar el idioma de cada oración (Lui & Cook, 2012), representa los textos mediante n-gramas de caracteres de diferentes longitudes como rasgos ($1 \leq n \leq 4$); luego aplica un clasificador baye-

string test (Two-letter code to identify the main world languages; it became a standard in 2002).

The parameters obtained for the HMM segmentation are initial probabilities, emission probabilities and transition probabilities.

Initial probabilities

Probability that any text begins with each of the available languages. It is assumed that all languages are equiprobable at the start of a document (each author can start writing a text with the desired language); it is assigned the same probability 1/N to each language, where N is the number of languages available for training.

Emission probabilities

Matrix that keeps the probability of emitting the feature (the word) or from the state (language) S_i. These probabilities were drawn from the texts of training, the emission probability of language L_i for feature O_j is the number of times feature O_j appears in language L_i, divided by the total number of features that appear in written texts in language L_i.

Transition probabilities

Probability of moving from one language to another in the same document. They were calculated in two ways (4) and (5)

$$L_{ij}/L_{in} \quad (4)$$

where, L_ij is the number of times the transition from L_i to L_j is seen in the corpus, and L_in is the number of existing jumps in the corpus from language L_i to any other language; and

$$p_t(L_i) * p_e(L_j) \quad (5)$$

where $p_t\ (L_i)$ is the probability of finishing a text in language L_i, y $p_e\ (L_j)$ is the probability of starting the next word with language L_j. For this purpose, probabilities of starting and finishing in each language with each of the words are stored in auxiliary tables.

C. Text segmentation by language identifying the language of each sentence

For the operation of this method (*multilangid*) it has a minimum unit in which it is assumed that language does not change; said minimum unit will be the sentence. Segmenta-

tion by language is performed in two stages: obtaining the sentences of the text and identifying the language of each sentence.

For the segmentation of a document by sentences, it was used the NLTK [Natural Language Tool Kit] (Bird, 2006) library, a platform for building programs in Python programming language for Natural Language Processing [NLP] –Field of computer science, artificial intelligence and linguistics, which studies interactions between computers and human language –. The library bases its operation on a list of punctuation that can define borders between sentences ('.', '?', '!'). The algorithm can be trained to increase the number of tokens that define changes of sentence in a text. However, this work did not involve any training activities.

Langid, the tool used to identify the language of each sentence (Lui & Cook, 2012), represents texts by n-grams of characters of different lengths as features ($1 \leq n \leq 4$); then applies a Bayesian classifier to identify the most likely language of them. It trains with a set of monolingual documents for each language, although it can also be a single document for each language, ensuring it has the number of features sufficient for performing classification from these.

IV. Experimental framework

This section describes the experiments performed on algorithms, as well as the corpus used for them. In addition, there are the characteristics of the corpus.

A. Corpus Wikipedia-Multi

The *corpus* used (Wikipedia-Multi) was created by Lui, Lau, and Baldwin (2014) using segments sources of mediawiki pages of Wikipedia. Used segments correspond to the period between July and August 2010.

To generate the corpus, documents in the raw of mediawiki were normalized; these contain a paragraph for each line, interspersed with structural elements (tags), so a *corpus* is obtained where language changes occur only in the jumps between paragraphs. Documents were filtered to remove all these structural elements, and kept only the documents that exceeded 25 kb after standardization. Then, of this set were removed languages that contained less than 1000 documents. Remaining 44 languages.

The steps for building documents were:

- select N languages without replacement *($1 \leq N \leq 5$)*;
- select a document (without replacement) for each language;

siano para identificar el idioma más probable de los mismos. Se entrena con un conjunto de documentos monolingües por cada idioma, aunque también puede ser un único documento por cada idioma, procurando que tenga la cantidad de rasgos suficiente para realizar la clasificación a partir de estos.

IV. Marco experimental

En esta sección se exponen los experimentos realizados sobre los algoritmos, así como el corpus utilizado para los mismos. Además se encuentran las características del corpus.

A. Corpus Wikipedia-Multi

El *corpus* utilizado (Wikipedia-Multi) fue creado por Lui, Lau, y Baldwin (2014) utilizando segmentos de las fuentes de mediawiki de las páginas de Wikipedia. Los segmentos usados corresponden al período comprendido entre julio y agosto de 2010.

Para generar el corpus se normalizaron los documentos en bruto de mediawiki; estos contienen un párrafo por cada línea, intercalados con elementos estructurales (etiquetas), por lo que se obtiene un *corpus* en el que los cambios de idioma ocurren solo en los saltos entre párrafos. Se filtraron los documentos para eliminar todos estos elementos estructurales, y se mantuvieron solamente los documentos que excedían los 25 kb luego de la normalización. Luego, de este conjunto se eliminaron los idiomas que contenían menos de 1000 documentos. Quedaron un total de 44 idiomas.

Los pasos a seguir para la construcción de los documentos fueron:

- seleccionar N idiomas sin reemplazo *($1 \leq N \leq 5$)*;
- seleccionar un documento (sin reemplazo) para cada idioma;
- tomar las primeras $^{1}/_{N}$ líneas del inicio de cada documento; y
- unir todas las líneas seleccionadas (sin agregar ningún carácter adicional entre ellas).

Como resultado del algoritmo anterior se obtuvo un *corpus* en el cual ningún documento contará con múltiples segmentos escritos en el mismo idioma. El *corpus* posee 6.000 documentos destinados al trabajo multilingüe: 5.000 para el entrenamiento, que contienen 1.000 para cada valor de N (cantidad de idiomas presentes en cada documento; $1 \leq N \leq 5$); y otros 1.000 para las evaluaciones, esto es 200 por cada valor de N.

B. Evaluaciones y resultados obtenidos

Se evaluaron los métodos propuestos sobre la colección descrita, y se compararon los resultados con el algoritmo propuesto en el estado del arte: *seglang* (Yamaguchi & Tanaka-Ishii, 2012). La probabilidad de inicio para las dos variantes del algoritmo que utiliza HMM fue 1/44 para cada idioma. Los algoritmos *multilangid* y *seglang* están diseñados para obtener como entrenamiendo un documento por cada idioma. Estos fueron construidos a partir de los documentos multilingües existentes en el entrenamiento, concatenando los segmentos de cada idioma en un documento único.

Las medidas de evaluación utilizadas fueron la μF1 y MF1, sujetas a las fórmulas (6) a (12).

$$\mu F_1 = \frac{2.\mu Precisión.\mu Relevancia}{\mu Precisión + \mu Relevancia} \quad (6)$$

$$MF_1 = \frac{\sum_{i=1}^{|C|} F_1(c_i)}{|C|} \quad (7)$$

$$F_1(c_i) = \frac{2.Precisión(c_i).Relevancia(c_i)}{Precisión(c_i) + Relevancia(c_i)} \quad (8)$$

$$\mu Precisión = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FP_i} \quad (9)$$

$$Precisión(c_i) = \frac{TP_i}{TP_i + FP_i} \quad (10)$$

$$\mu Relevancia = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FN_i} \quad (11)$$

$$Relevancia(c_i) = \frac{TP_i}{TP_i + FN_i} \quad (12)$$

donde, |C| representa la cantidad de clases (idiomas); ci es el idioma i; TPi es el número de palabras para las cuales se identificó el idioma ci correctamente; FPi es la cantidad de palabras que fueron clasificadas en el idioma ci sin pertenecer a este realmente, y los FNi las que pertenecen al idioma ci que no fueron clasificadas correctamente.

En la **Figura 1** se observan los resultados de las evaluaciones realizadas a la propuesta que utiliza los modelos ocultos de Markov para la segmentación. Cabe recordar que se tomaron en consideración dos maneras de calcular las probabilidades de transición entre los idiomas Li, Lj:

la primera consiste en la cantidad de transiciones vistas en el entrenamiento desde el idioma Li hacia el idioma Lj dividido por la cantidad total de saltos de idioma vistos desde el idioma Li, hacia cualquier otro; y

la segunda consiste en la probabilidad de terminar un segmento de texto con el idioma Li multiplicado por la probabilidad de comenzar un segmento con el idioma Lj en la próxima palabra.

Como se puede observar, la primera variante obtuvo mejores resultados. Esto se debe a que la segunda variante necesita una colección donde el fenómeno que ella describe aparezca más frecuentemente, para poder tener una mejor estimación de las probabilidades de transición entre idiomas, ya que de ahí se obtienen las probabilidades de terminar y empezar segmentos de texto en los respectivos idiomas. A partir de este momento, a la primera variante se le llamará HMMmultilangid, y será la que se utilice en el resto del trabajo.

En la **Figura 2** se comparan los resultados para las dos variantes de seglang. Se puede observar que se comportó mejor la variante que utiliza Media de Estadísticas de Coincidencias [MMS] para calcular la entropía cruzada entre dos textos. En el artículo donde se propuso seglang (Yamaguchi & Tanaka-Ishii, 2012), los mejores resultados fueron alcanzados utilizando Predicción por Coincidencias Parciales [PCP] para hallar la entropía cruzada. Los autores

- taking the first $1/N$ initial lines of each document; and
- join all the selected lines (without adding any additional character between them).

As a result of the previous algorithm, a *corpus* in which no document will have multiple written segments in the same language was obtained. The *corpus* has 6,000 documents for multilingual work: 5,000 for training, containing 1,000 for each value of N (number of languages present in each document; 1≤N≤5); and 1,000 for evaluations, it means 200 for each value of N.

B. Evaluations and results

The proposed methods for the collection described were evaluated, and the results were compared with the proposed algorithm in the state of the art: *seglang* (Yamaguchi & Tanaka-Ishii, 2012). The initial probability for the two variants of the algorithm that uses HMM was 1/44 for each language. *Multilangid* and *seglang* algorithms are designed to obtain as a training a document for each language. These were built from existing multilingual documents in training, concatenating the segments of each language in a single document.

Evaluation measures used were μF1 and MF1, subject to the formulas (6) to (12).

$$\mu F_1 = \frac{2.\mu Precisión.\mu Relevancia}{\mu Precisión + \mu Relevancia} \quad (6)$$

$$MF_1 = \frac{\sum_{i=1}^{|C|} F_1(c_i)}{|C|} \quad (7)$$

$$F_1(c_i) = \frac{2.Precisión(c_i).Relevancia(c_i)}{Precisión(c_i) + Relevancia(c_i)} \quad (8)$$

$$\mu Precisión = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FP_i} \quad (9)$$

$$Precisión(c_i) = \frac{TP_i}{TP_i + FP_i} \quad (10)$$

$$\mu Relevancia = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FN_i} \quad (11)$$

$$Relevancia(c_i) = \frac{TP_i}{TP_i + FN_i} \quad (12)$$

where, |C| represents the number of classes (languages); c_i is the language i; TP_i is the number of words for which the language c_i was correctly identified; FP_i is the number of words that were classified in the language c_i without belonging to it, and FN_i are the ones which belong to the language c_i that were not correctly classified.

In **Figure 1**, the results of evaluations performed on the proposal that uses hidden Markov models for segmentation

are observed. It should be noted that were considered two ways to calculate the transition probabilities between L_i, L_j languages:

- the first is the number of transitions seen in training from language L_i, to language L_j divided by the total number of jumps language seen from language L_i, to any other; and
- the second is the probability of finishing a segment of text with language L_i multiplied by the probability of starting a segment with language L_j in the next word.

As can be seen, the first variant obtained better results. This is because the second variant needs a collection where the phenomenon it describes appears more frequently in order to have a better estimate of transition probabilities between languages, since from there, probabilities of starting and finishing text segments in the respective languages are obtained. From this moment, the first variant will be called HMMmultilangid, and will be the one used in the rest of the work.

Figure 2 compares the results for the two seglang variants. It can be observed that the variant using Statistical Matches Media [SMM] to calculate the cross entropy between two texts had a better performance. In the article which seglang was proposed (Yamaguchi & Tanaka-Ishii, 2012), the best results to find the cross entropy were achieved using Prediction by Partial Matching [PPM]. The authors suggest that the results for each variant depend on the training. In their experiments, they used an average of 10 kb per language for training, while the corpus used in this work has an average of 653 kb per language.

In **Figure 3**, the results of the best variants of algorithms analyzed are shown and multilangid is included. As can be seen, the latter reached the highest score. Keeping in mind that in the *corpus* used there is no more than one language in the same sentence, it is logical that the best results are obtained for multilangid. However, in texts that do not have this feature, the algorithm will incur in misclassification because it will recognize, in the best case, the most represented language in each sentence, obviating the other languages present. For these cases, seglang or HMMmultilangid can be used, which are designed to detect the language change anywhere in the text. Specifically referred to these two, the one that performed better was the one proposed in this work, so

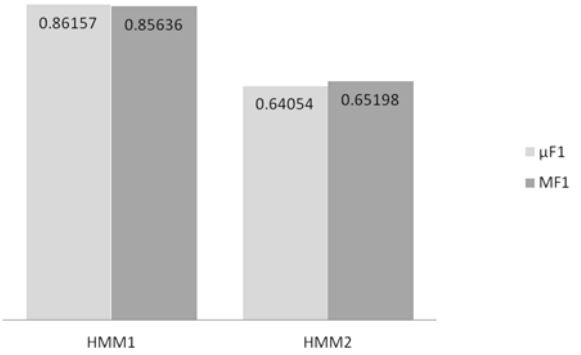

Figure 1. Results achieved by applying HMM for each variant / Resultados alcanzados al aplicar HMM para cada una de las variantes

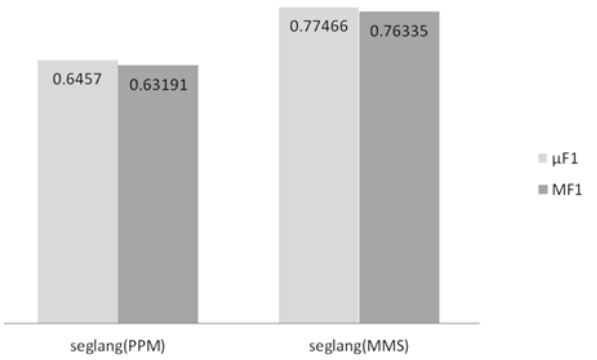

Figure 2. Evaluation results for the two seglang variants / Resultados de las evaluaciones para las dos variantes de seglang

plantean que los resultados para cada variante dependen del entrenamiento. En sus experimentos utilizaron como promedio 10 kb por cada idioma para el entrenamiento, mientras que el corpus utilizado en este trabajo cuenta con 653 kb como promedio por cada idioma.

En la **Figura 3** se muestran los resultados de las mejores variantes de los algoritmos ya analizados y además se incluye multilangid. Como se puede observar, este último alcanzó la mejor puntuación. Teniendo en cuenta que en el *corpus* utilizado no existe más de un idioma en la misma oración, es lógico que los mejores resultados se obtengan para multilangid. Sin embargo, en textos que no posean esta característica, el algoritmo incurrirá en errores de clasificación, pues reconocerá, en el mejor de los casos, el idioma más representado en cada oración, obviando los demás idiomas presentes. Para estos casos se puede utilizar seglang o HMMmultilangid, que están diseñados para detectar el cambio de idioma en cualquier parte del texto. Específicamente de estos dos, se comportó mejor el propuesto en este trabajo, por lo que se recomienda la utilización del mismo para el procesamiento de corpus con documentos en los que el cambio de idioma pueda ocurrir en cualquier parte del texto.

V. Conclusiones

En este trabajo se presentaron los métodos para segmentar textos por idiomas *multilangid* y *HMMmultilangid*. El primero se propuso para el caso en el que el cambio de idioma se de en los saltos entre oraciones, consiste en segmentar el

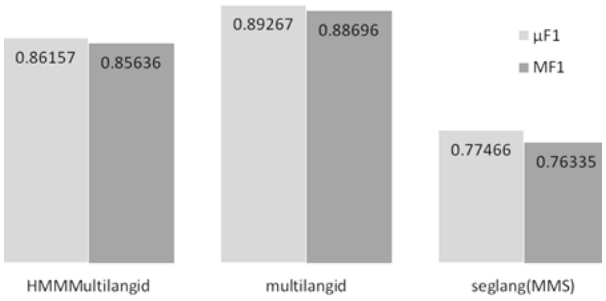

Figure 3. Results of multilangid and the best variants of HMM and seglang / Resultados de multilangid y las mejores variantes de HMM y seglang

texto por oraciones y luego identificar el idioma de cada oración. La segunda variante con¬siste en la adaptación de los modelos ocultos de Markov, considerando los idiomas y las palabras como los "estados" y las "observaciones" del modelo, respectivamente.

Para ambas propuestas se superó el estado del arte, obteniéndose los mejores resultados para *multilangid* debido a las características de la colección, en la que el cambio de idioma solo ocurre entre oraciones. Teniendo en cuenta los resultados obtenidos se considera que los métodos propuestos tienen un impacto favorable en la segmenta¬ción de textos por idiomas.

Para futuras investigaciones se recomienda utilizar una colección donde el cambio entre idiomas ocurra en cualquier parte del texto y no solo en los saltos entre oraciones, y evaluar además la representación de los documentos mediante otros rasgos como los n-gramas de caracteres. S&T

it is recommended the use thereof for corpus processing with documents in which the language change can occur in any part of the text.

V. Conclusions

This paper presented methods for segmenting texts by language *multilangid* and *HMMmultilangid*. The first was proposed for the case in which the change of language is in the jumps between sentences that consists of segmenting the text by sentences and then identifying the language of each sentence. The second variant consists of the adaptation of hidden Markov models, considering the language and words such as "states" and "observations" of the model, respectively.

For both proposals, the state of the art was overcome, obtaining the best results for *multilangid* due to the characteristics of the collection, in which the language change occurs only between sentences. Keeping in mind the results obtained, it is considered that the proposed methods have a favorable impact on text segmentation by language.

For future researches, it is recommended to use a collection where changes between languages occur anywhere in the text and not only in the jumps between sentences, and evaluate the representation of documents by other features such as character n-grams. S&T

References / Referencias

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

Blunsom, P. (2004). *Hidden Markov models*. Retrieved from: http://digital.cs.usu.edu/~cyan/CS7960/hmm-tutorial.pdf

Cabeza, R. (2015). *Segmentación de textos por idiomas: utilizando modelos ocultos de Markov*. Saarbrücken, Germany: EAE.

Ghahramani, Z. (2001). An introduction to hidden Markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, *15*(01),9-42.

Juola, P. (1997). *What can we do with small corpora? Document categorization via cross-entropy*. Edinburgh, UK: University of Edinburgh.

Lui, M. & Cook, P. (2012). langid.py for better language modelling. In: *Proceedings of Australasian Language Technology Association Workshop* (Vol. 10, pp. 107–112). Retrieved from: http://www.alta.asn.au/events/alta2012/proceedings/pdf/U12-1.pdf

Lui, M. (2016). *Langid.py* [app]. Retrieved from: https://github.com/saffsd/langid.py

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.

Rincón, L. (2012). *Introducción a los procesos estocásticos.* México, DF: UNAM. Available at: http://lya.fciencias.unam.mx/lars/Publicaciones/procesos2012.pdf

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.

Witten, I. H. & Bell, T. C. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compresion. *IEEE Transactions on Information Theory*, *37*(40), 1085-1094.

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-Volume 1* (pp. 969-978). Stroudsburg, PA: ACL.

CURRICULUM VITAE

Robin Cabeza Bachelor's degree in Computer Science from Universidad de Oriente (2015) and student of Master in Design Assisted by Computer at the Universidad de Holguín [UHo], Cuba. Currently he is professor of programming and member of CAD/CAM Studies Center at the Faculty of Engineering of UHo, where he researches about biomechanical / Licenciado en Ciencias de la Computación, graduado en la Universidad de Oriente (UO) en 2015. Actualmente profesor en la Facultad de Ingeniería de la Universidad de Holguín (UHo), donde imparte la asignatura de Programación. Pertenece al centro de estudios CAD/CAM de la misma universidad, y forma parte del proyecto que realiza investigaciones en el campo de la biomecánica. Se encuentra cursando la Maestría en Diseño Asistido por Computadoras en la Universidad de Holguín.