LIC. J. R. BARRERAS MILANÉS
SAIN SALAZAR MARTINEZ
LIC. MILDREY CARBONELL CASTRO
RICARDO A. SOTOLONGO TRUJILLO
Universidad Central de las Villas. Santa Clara. Cuba. barreras@uclv.edu.cu

Caracterización de Alfa Relaciones en Lenguajes Naturales

Introducción

La aplicación de las modernas técnicas de cómputo en los más diversos ámbitos de la actividad humana ha originado un considerable desarrollo en las materias que fundamentan dichos sistemas. Particularmente, los aspectos referidos a la comunicación hombre-máquina, han merecido especial atención en los últimos años.

Si consideramos que la forma predominante de comunicación humana es la realizada a través del lenguaje, nos veremos en la obligación de establecer esquemas lingüísticos que sean cada vez más factibles en el proceso de entendimiento entre el hombre y la máquina. La formalización de un lenguaje eficiente para resolver todos los posibles eventos en cierto campo deviene enormes esfuerzos por parte de los especialistas. Digamos concretamente, que solo han podido establecerse convenios parciales aplicables en un reducido marco dentro del universo de trabajo. En la actualidad, un gran numero de investigaciones incursionan en la invención de modelos y herramientas teóricas capaces de lograr un procesamiento efectivo del lenguaje común (o natural, lenguajes usados por el hombre al comunicarse con sus semejantes).

En este sentido, un importante desempeño tienen los métodos dedicados a la determinación de la pertenencia de una palabra a cierto lenguaje. Como los lenguajes naturales no cuentan con estructuras regulares susceptibles de ser modeladas de manera sencilla y uniforme, la decisión sobre la pertenencia o no de la palabra al lenguaje ha de realizarse mediante el empleo de un diccionario. La búsqueda exhaustiva en un diccionario resulta un mecanismo sumamente costoso, impráctico en la mayoría de los casos. Consideremos entonces una solución alternativa, basada en la búsqueda de ciertas relaciones en las estructuras de las palabras (relaciones internas). Estas relaciones nos permitirán elaborar una *medida* cuantitativa del *grado de pertenencia al lenguaje* de cada palabra. Como podemos ver, en estos casos disponemos de una medida difusa (por ejemplo valores entre 0 y 1), a diferencia de los acostumbrados valores *verdadero* y *falso* que caracterizan la solución usando diccionarios.

Desarrollo Teórico.

Comencemos aclarando algunos conceptos fundamentales empleados en la descripción formal de nuestro trabajo.

Definición de Lenguaje (Diccionario)

Un lenguaje es un conjunto de palabras y métodos para combinar palabras que es usado y entendido por un extenso grupo de personas. [1] *.*

En la anterior definición de lenguaje se involucra el concepto de palabra. Aclararemos que la palabra no es más que un arreglo o cadena de longitud n, formada a partir de un conjunto de símbolos denominado alfabeto. Si A es el alfabeto, entonces denotamos la palabra p

$$p = x_1x_2x_3 \ldots x_n\ ,\ \ x_i \in A\ ,\ i = 1,2,\ldots,n \qquad \text{(n pertenece a los naturales)}$$

y a L(A) como el conjunto de palabras pertenecientes al lenguaje.

Características de los Lenguajes

Usualmente en los lenguajes existe un número reducido de palabras primitivas sujetas a ciertas leyes o métodos de combinación a partir de las cuales es posible obtener el resto.

Tal es el caso del ejemplo que a continuación mostramos:

Sea el alfabeto A={a,b} y

L(A)={ a, ab, abb, abbb, . . . , aa, aab, aabb, aabbb, . . . }

el lenguaje formado a partir de A.

Las palabras de L(A) poseen una estructura especifica que podemos generalizar de la siguiente manera

$$L(A) = \{\ a^n b^m\ ,\ n= 1,2, \ldots ,\ m= 1,2, \ldots\}$$

Sin lugar a duda, cada palabra de L(A) es susceptible de ser formada por la aplicación de simples reglas que primariamente pueden expresarse como:

1) Colocar sobre el arreglo vacío (denotado como $\wedge$) la letra *a* n veces

2) Colocar la letra *b* m veces.

De esta forma, cada palabra de L(A) resulta de la aplicación de reglas sumamente sencillas.

Examinemos otro ejemplo. Sea

L(B) = { casa, perro, limar mata, zapato, risa, ... ,
casero, perrera, limar, zapatero, risible ... }

formado por las palabras del habla española común y sus derivaciones[i].

En este caso resulta evidente la carencia de una estructura general en las palabras. Al observar el conjunto, nos percataremos de ciertas regularidades presentes en las primeras letras de diversas palabras; no obstante, dichas regularidades no son suficientes para formular un conjunto de reglas como en el ejemplo inicial. Su solución se convierte en el planteamiento de una gran cantidad de complejas reglas de combinación que en la mayoría de los casos solo reproducen un conjunto parcial del lenguaje.

Empleando estos dos ejemplos podemos hablar de manera informal de dos grupos de lenguajes: los lenguajes con estructuras precisas de formación de sus palabras y aquellos en los que resulta complejo la determinación de reglas para sus estructuras.

A pesar de no tener por el momento una definición formal del concepto de estructura, no es ocioso notar su vínculo con la naturaleza del conjunto de reglas de combinación de las palabras.

Los lenguajes que como L(A) pertenecen al primer grupo se denominan **Lenguajes Formales.**

Los lenguajes que usan las personas para comunicarse se denominan **Lenguajes Naturales**(L(B), ingles, etc). Estos pertenecen al segundo grupo.

Lenguajes Formales. Gramáticas.

Definamos un lenguaje formal como sigue:

Con A conjunto finito. Un lenguaje formal L(A) definido sobre A es un subconjunto de A* , el conjunto de todas los arreglos (o cadenas) con los elementos de A; definido por medio de una lista de reglas que el lenguaje ha de obedecer.

Usualmente en la definición de un lenguaje formal manejamos el concepto de gramática (o sistema de estructuración de frases).

Una gramática no es más que un alfabeto finito T (o conjunto de símbolos terminales), un conjunto finito N de símbolos no terminales o de apoyo, un conjunto finito P de $[(N \cup T^*) - T^*] x (N \cup T)^*$ llamado conjunto de composiciones (reglas) y un símbolo $s_0 \in N$.

La gramática especifica en P un serie de reglas expresadas con símbolos de N y T empleados en la formación de todas las palabras generadas por ella a partir del símbolo inicial s_0.

Resulta que toda gramática G genera un lenguaje formal G(L) formado por palabras cuyas estructuras fundamentales se forman a partir de reglas definidas en el conjunto de composición P de dicha gramática. Mostremos una gramática para $L(A)=\{ (ab)^n , n=1,2,...\}$.

$G =(N,T,P,s_0)$, donde $N=\{S_0,S\}$. $T=\{a,b\}$ $P=\{ (s_0 \rightarrow aS), (S \rightarrow bS_0), (S \rightarrow b)\}$, S_0 símbolo inicial.

Cuando manejamos un lenguaje formal podemos encontrar una gramática que modele su estructura y la *Máquina de Turing* [1] capaz de reconocer sus palabras.

De este modo la cuestión de reconocer la pertenencia o no de una palabra al lenguaje constituye un problema resuelto.

Estructura en un lenguaje natural

Cada palabra de un lenguaje puede fragmentarse en varias secuencias de símbolos (que a su vez pueden pertenecer al lenguaje) de distintas longitudes, iguales o no entre sí. En los lenguajes naturales existen secuencias de símbolos que tienen una mayor incidencia en la formación de las palabras. Es decir, la frecuencia de aparición de estas secuencias en cadenas típicas del lenguaje es mayor que el resto de las posibles secuencias. A dichas secuencias las denominaremos ***elementos constructivos fundamentales (ECF)***. Por ejemplo, en el español, podemos formar las palabras a partir de la concatenación de las sílabas, que constituyen a su vez los ECF de dicho lenguaje. Al ser una característica fundamental de los lenguajes naturales su constante enriquecimiento, no hay certeza de la forma en que puedan ir evolucionando dichas ECF (mediante la aparición de nuevas ECF, caducidad de otras y cambio de unas terceras). Es por ello que su estudio merece gran importancia.

Ahora bien, estas ECF no tienen una longitud fija y normalmente no logramos identificarlas a todas, ni la importancia relativa entre ellas en un lenguaje natural desconocido.

No existirá dificultad alguna si intentamos determinar la pertenencia o no de una palabra a un lenguaje formal. La gramática que genera dicho lenguaje y la *máquina* a ella asociada, permiten clasificar dicha palabra como miembro o

no del lenguaje. En un lenguaje natural no tenemos esta facilidad; y la determinación de la pertenencia de la palabra entra en los límites de la apreciación del observador o a la revisión del conjunto total de palabras de dicho lenguaje. En la mayoría de los casos, como ya hemos mencionado, el lenguaje en cuestión se desarrolla, por lo que este último método resulta ineficiente. Por otro lado, al decir que esta clasificación depende del criterio del observador estamos entrando en el campo de la lógica difusa o borrosa. En esta lógica no se trabaja con los acostumbrados conceptos *verdadero* o *falso*, sino que se obtiene una medida de cuán cerca del *verdadero* nos encontramos [3].

Medida de Pertenencia.

Desarrollemos al efecto una función que nos permita determinar el grado de pertenencia de cierta palabra a un lenguaje natural cualquiera. Definamos dicha función como G: $A^* \rightarrow \nabla^+$; y denotemos G(w) medida del grado de pertenencia de la palabra w $\in A^*$ al lenguaje Υ (A). Cada lenguaje natural tendrá asociado una función G ajustada a sus características estructurales.

Sea Υ(A) un lenguaje natural, si suponemos que todos sus ECF son de longitud p, una medida apropiada del grado de pertenencia de w $\in A^*$ a Υ (A) es

$G_p(w) = F_p(w)$, donde $$F_p = \prod_{i=1}^{n-p+1} f(w_{i,p})$$, y n es la longitud de w

$w_{i,,p}$ denota la subcadena contenida en w a partir de la posición i tomando p símbolos a la derecha.

$f(w_{i,p})$ denota la frecuencia de aparición de la secuencia $w_{i,p}$ en las palabras del lenguaje Υ (A).

$F_p \rightarrow 0$ únicamente cuando w contenga al menos una secuencia de longitud p de frecuencia cero.

Es válido aclarar la forma de F_p mediante el empleo del siguiente ejemplo:

Con w = cancion. No hay duda que $F_1 = f(c)^2 f(a)f(n)^2 f(i) f(o)$. Pero w puede dividirse de diversas formas en elementos de 2 letras ca-nc-io-n , ca-an-nc-ci-io-on. Debemos decidir entonces el convenio a tomar. Si consideramos la fragmentación de la cadena sin el uso de solapamiento tendremos que $F_2 = f(ca)f(nc)f(io)$. Para F_3 , F_4, . . . ocurre de forma análoga.

A partir de $G_p(w)$ podemos deducir una generalización G(w) para lenguajes naturales en los cuales no conozcamos de antemano la longitud de sus ECF. En este caso al no conocer los elementos estructurales decisivos en la formación de las palabras, no podremos decidir certeramente qué secuencias son las que brindan la medida más real del grado de pertenencia y podríamos emplear un criterio errado. G(w) es la suma de todas las posibles frecuencias F_j donde $m \geq j \geq 1$ (m es la longitud de la mayor cadena del lenguaje natural en cuestión) pesadas con el parámetro α_j (medida del peso que tienen los elementos estructurales de longitud j en la formación de las palabras).

$$G(w) = \alpha_1 F_1(w) + \alpha_2 F_2(w) + \alpha_3 F_3(w) + \ldots + \alpha_m F_m(w)$$

Usualmente en los lenguajes naturales las palabras poseen una longitud promedio de cadenas relativamente pequeña. Evidentemente el conjunto de pesos ha se ser normalizados, es decir:

$$\sum_{i=1}^{m} \alpha_i = 1$$

La mayor dificultad estriba entonces en la búsqueda de los valores para los parámetros α_1, α_2, α_3, . . . , α_m que expresen apropiadamente las relaciones estructurales en un lenguaje natural determinado.

Determinación de los Parámetros α_i de un Lenguaje

Estamos en presencia de un problema de búsqueda de valores óptimos α_i para el cual no hemos desarrollado una metodología determinada. Apliquemos, sin embargo, un método que explota la búsqueda aleatoria 'guiada' que ha ganado popularidad en los últimos años. Nos referimos a los algoritmos evolutivos. [2]

Un algoritmo general que describe las cuestiones esenciales de la familia de dichos algoritmos es el siguiente:

MIENTRAS NO *condición de parada* HACER

{ *Generar(Población)*

SI *Población* ES PEOR QUE *Población_Anterior*

Población = Población_Anterior

Mejores_Individuos = Seleccionar_Mejores(Poblacion)

Aplicar_Operadores(Mejores_Individuos) }

En todos los casos, una población estará compuesta por individuos y cada uno de ellos constituye una solución al problema de optimización que se quiere resolver. Todos los algoritmos mantienen una población de individuos para

guiar la búsqueda de modo paralelo en todo el espacio. En nuestro caso cada individuo (o cromosoma) no es mas que un conjunto de parámetros $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_m$ (llamados genes).

El proceso *Generar,* como su nombre lo indica, es el encargado de generar una población con una cantidad fija de individuos de forma aleatoria, es decir generar conjuntos de α parámetros (números entre 0 y 1).

El proceso *Seleccionar_Mejores,* realiza la selección de algunos individuos una vez evaluada la calidad de estos como solución al problema planteado. La evaluación se realiza siguiendo diversos criterios. En nuestro caso tomaremos a priori un conjunto de palabras pertenecientes al lenguaje natural, con que se calcularon las frecuencias $f(w_{i,p})$, e igual cantidad de palabras que no pertenecen. En la medida en que la función G asociada (obtenida con los parámetros $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_m$) logre clusterizar estos dos conjuntos, estaremos en capacidad para incluir a ($\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_m$) en el conjunto de los mejores individuos.

El proceso *Aplicar_Operadores* se le aplica a los mejores individuos de la población y consiste en la mutación o recombinación de estos. El operador de recombinación intenta tomar información de varios individuos para combinarla en uno nuevo que presumiblemente sea mejor, pues incluye buenas características de los combinados. Generalmente se realiza con la participación de dos individuos llamados padres; que intercambian entre si varios de sus parámetros. La mutación ocurre de manera probabilística con la participación de un individuo seleccionado aleatoriamente; al cual se le inducen cambios en su naturaleza o valor informativo (genes). En nuestro caso se realiza mediante el intercambio de un bit cualquiera en un parámetro escogido aleatoriamente.

El algoritmo detiene su ejecución una vez que la población obtenida supere cierto límite de individuos que brinden una solución satisfactoria al problema tratado o debido al establecimiento de un límite de tiempo de cómputo que puede obedecer a restricciones de recursos, etc.

Los mejores individuos de la población, una vez detenido el algoritmo, pueden brindarse como solución definitiva al problema tratado. En nuestro caso, realizamos la selección de un individuo promedio entre el conjunto de candidatos para la formación de la función G asociada al lenguaje natural investigado.

Estudios practicados

Para nuestro estudio seleccionamos como base un texto del idioma español, a partir del cual formamos un diccionario con una cantidad total de 8000 palabras. Las frecuencias $f(w_{i,p})$ fueron calculadas siguiendo las secuencias de letras de diferentes longitudes presentes en cada una de las palabras contenidas en el diccionario. Dadas las características particulares del lenguaje español, seleccionamos solamente las secuencias de 1,2 y 3 letras de longitud. Esta decisión se encuentra a tono con la longitud promedio de las sílabas del lenguaje español. La frecuencia de cada secuencia de longitud 1,2 o 3 se determina por la cantidad de veces que se repite dicha secuencia en las palabras del diccionario entre la cantidad de total de secuencias de dicha longitud. La frecuencia de aquellas secuencias del lenguaje que no aparezcan en el diccionario empleado se considerarán un valor límite ε próximo a cero.

Según el texto analizado, entre las secuencias más frecuentes se encuentran:

Longitud 1		Longitud 2		Longitud3	
Letra	Frecuencia	Sec.	Frecuencia	Sec.	Frecuencia
a	1.2146×10^{-1}	en	4.5387×10^{-2}	ent	2.7963×10^{-2}
e	1.1346×10^{-1}	es	4.1470×10^{-2}	ion	2.4002×10^{-2}
o	8.5026×10^{-2}	ar	3.8092×10^{-2}	con	1.8992×10^{-2}
i	8.3433×10^{-2}	re	3.5253×10^{-2}	cio	$1,9866 \times 10^{-2}$
r	8.0633×10^{-2}	on	3.5037×10^{-2}	ado	2.4468×10^{-2}

Con la utilización del algoritmo genético los valores α_i fueron mejorados progresivamente hasta lograr una población donde todos sus individuos lograron clusterizar las palabras de prueba con un 98 % de confiabilidad (recordemos que se eligen palabras que pertenecen al lenguaje y palabras que no pertenecen para conformar ambos cluster). Al seleccionar el mejor individuo de dicha población, los valores de los parámetros arrojaron los siguientes resultados:

$\alpha_1 = 0.0182506$

$\alpha_2, = 0.3758103$

$\alpha_3, = 0.7058397$

Intuitivamente los valores α en el español deberían cumplir $\alpha_1 < \alpha_2 < \alpha_3$. hasta el momento los valores han reafirmado esto.

Conclusiones

La aplicación del método descrito a lenguajes naturales como el español y el inglés, ha arrojado resultados alentadores. En primer lugar, la convergencia del algoritmo genético hacia valores concretos nos muestra evidencia acerca de la existencia de lenguajes naturales con relaciones internas determinables. En segundo lugar, la posibilidad de contar con un sencillo y eficiente criterio de decisión para la pertenencia al lenguaje.

Este método desarrolla una función *medida* que, en todos los casos, brinda un grado de *pertenencia de la palabra* (criterio de cuán semejante en su estructura, resulta la palabra respecto a las palabras del lenguaje) y no, si la palabra pertenece definitivamente al lenguaje. Por ejemplo: *rampimpona, insorter, regonencia.* Palabras muy semejantes a las palabras comunes del español, y aunque nuestra *medida* no brinda un resultado definitivo para ellas; es capaz de distinguirlas de otras como kthjzolki, jezyhonl, llhutnenl, cuyo aspecto es desconocido al lenguaje. Es importante notar que este hecho determina la naturaleza de los problemas que pueden ser resueltos por medio del uso de las *relaciones internas.*

Bibliografía.

1. *Matemáticas Discretas*. Richrad Johnsonbaugh. DePaul University. Chicago, 1986.
2. *Genetic Algorithms + Data Structures = Evolution Programs*. Zbigniew Michalewicz. Springer Series. New York, 1995.
3. *The Fuzzy Systems Handbook.* Earl Cox. London,1994

---

[i] Por comodidad usamos el alfabeto inglés, aunque las palabras pertenezcan al español.