Capítulo 3

Metodología

Este proyecto consiste en intentar la categorización de textos basados en su pertenencia a dominios formulaicos, según las expresiones contenidas en los textos, las cuales se hayan identificado como las características de cada dominio. En este capítulo se describirán las cuatro grandes fases en las que está dividido este trabajo. Estas fases son:

1. **entrenamiento**: para detectar las expresiones formulaicas dentro de textos de cada dominio formulaico;
2. **evaluación**: para determinar cuándo el entrenamiento ha sido suficiente (según ciertos criterios de terminación);
3. **refinamiento de fórmulas**: para obtener mejores resultados;
4. y **pruebas**: para medir la eficiencia del método desarrollado.

Las fases de evaluación y pruebas, a su vez, se subdividen en clasificación y análisis. En adelante cada fase será descrita en tres secciones: entradas, salidas y experimentos.

Las expectativas del proyecto son lograr la mayor cantidad de documentos clasificados correctamente, es decir maximizar el valor del **porcentaje de xito**; el clculo de esta variable se explica ms adelante.

Debido a que, como ya se dijo en el capítulo 2, al inicio de este trabajo no se había desarrollado aún en lingüística una metodología formal para comprobar si un dominio es formulaico o para obtener qué tan formulaico es dicho dominio (índice de formulaicidad), intuitivamente se podía decir que exista un alto riesgo de que se hiciera una mala elección de los dominios para hacer las pruebas. La razón es obvia ya que ésta se basaba puramente en la intuición y por lo tanto es subjetiva. Entonces al realizar las pruebas existirá la incertidumbre porque no se sabe que esperar, esto cambiará gracias a este trabajo.

Al finalizar este trabajo seremos capaces de afirmar a ciencia cierta si un dominio es formulaico o no y si lo es, en qu grado.

Adems el tamao de la muestra ser un factor determinante en los casos en que el ndice de formulaicidad sea bajo, porque existir el riesgo de que la muestra no sea lo suficientemente representativa.

Existen además otros detalles a considerar, que influenciarán los criterios de evaluación del rendimiento de está aplicación, es decir se deben definir aquellas variables que van a afectar directamente la calidad de fórmulas que se encuentren y también aquellas que afecten el desempeño del algoritmo. Para empezar, ya se dijo que el tamaño de muestra influirá directamente en la calidad de las fórmulas; por otro lado, la extensión de los textos (longitud) afectará el desempeño del algoritmo (tiempo de ejecución), como se explicará más adelante (Véase sección 3.1.3).

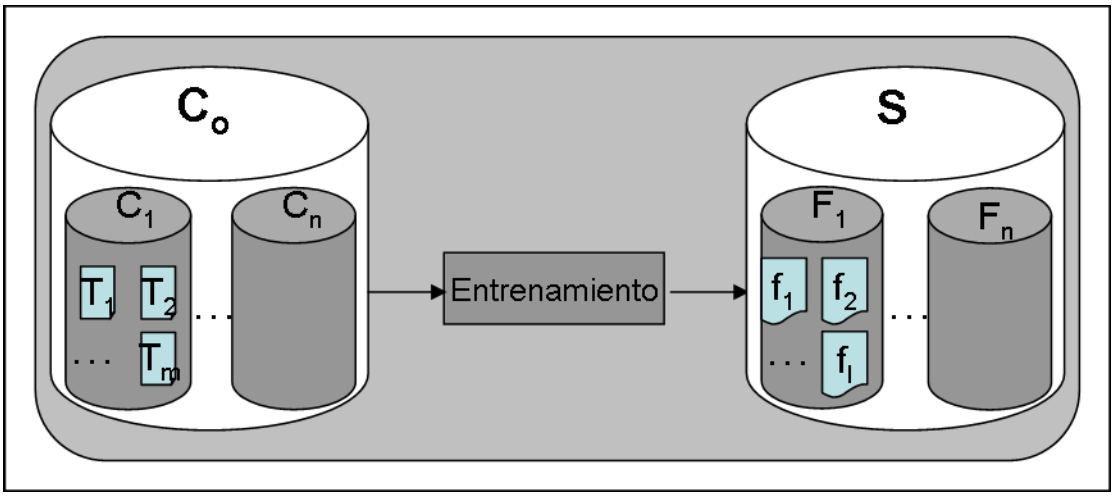

Figura 3.1: Fase de entrenamiento

3.1. Fase de entrenamiento

3.1.1. Entradas

Tiene como entrada un corpus estratificado C_o (Ecuación 3.1), que está conformado por varios subcorpora o estratos C_i con textos que pertenecen a distintas categorías temáticas (Ec. 3.2). Cada categoría se referirá a un dominio formulaico D_i. La recolección de textos se llevará a cabo de acuerdo a la disponibilidad de documentos de dominio público en Internet.

$$C_o = \{C_1, C_2, C_3, C_4, ..., C_n\}, \tag{3.1}$$

donde n es el número de subcorpora o dominios formulaicos en el corpus de entrada C_o.

$$C_i = \{T_1, T_2, T_3, T_4, ..., T_m\}, \tag{3.2}$$

donde m_{C_i} es el número de textos T que conforman cada subcorpus C_i (Fig. 3.1).

Cada subcorpus C_i tiene las siguientes propiedades:

- representa a un dominio formulaico D_i;
- tiene un m_{C_i} que es el tamaño de la muestra de dicho dominio, es decir, el número de textos que se recolecten para formar cada subcorpus muestra C_i;
- tiene un idioma en el que estarán escritos todos sus textos (para fines de nuestra aplicación estos serán inglés o español). Se manejará sólo un idioma por corpus.

Cada uno de los subcorpora C_i será una colección de textos digitales que pertenecen a un solo dominio formulaico. Para hacer la recolección de la muestra, primero se escogerán algunos dominios que intuitivamente parecen formulaicos; como se explicó en el capítulo anterior, no hay manera formal de determinar si un dominio es formulaico, así que la intuición es uno de los criterios más usados para la detección de formulaicidad. Aunque con los resultados obtenidos en nuestra etapa de entrenamiento se podrá determinar si en efecto el dominio resultó formulaico e incluso calcular el índice de formulaicidad, ésta es una utilidad ajena a los objetivos de este proyecto y por lo tanto no entraremos a discusión en este detalle. Posteriormente, se tomará una muestra de textos que representen a cada uno de estos dominios, cada una de estas muestras formará un subcorpus de la entrada.

Para obtener los textos que formarán el corpus que usaremos para cada fase, debemos definir la población de donde se va a tomar la muestra, entendiéndose por muestra "un subgrupo de la población" [24] y por población "un conjunto de todos los casos que concuerdan con determinadas especificaciones" [24]. La población será descrita formalmente más adelante.

La selección de cada muestra se llevará a cabo de una manera arbitraria, no probabilística y por criterios, es decir, se seleccionarán los textos aleatoriamente pertenecientes a dominios que, basados en nuestro criterio, son formulaicos.

El tamaño de la muestra, como ya se dijo, será una variable dentro de nuestros experimentos, ya que *a priori* no se conoce el tamaño de la población total y, debido a que este estudio se llevará a cabo desde un enfoque cualitativo según Sampieri [24], no se necesita que la muestra sea representativa.

Debido a lo anterior, se optará por una **muestra no probabilística**. En este tipo de muestra el procedimiento de selección es informal, la elección de los elementos no depende de la probabilidad sino de las características de la investigación o del criterio de los investigadores. Y así, a partir de ellas, se hacen inferencias sobre toda la población.

Recolectar la muestra de esta forma tiene la utilidad de que para el enfoque cualitativo no requiere la representatividad de los elementos en una población, sino una cuidadosa y controlada elección de sujetos con ciertas características especificadas por el investigador.

Anteriormente cuando se habló de características de la investigación se hacía referencia a los alcances de ésta. Una vez que se ha realizado la revisión de literatura se ha determinado que en esta investigación se realizarán estudios **exploratorios** y **descriptivos** [24]. Estos influirán también en la elección de los sujetos de la muestra.

3.1.2. Salidas

El proceso de entrenamiento es sencillo: se procesará cada subcorpus o estrato C_i por separado (Fig. 3.1), arrojando como resultado para cada uno un conjunto finito de fórmulas F_i (lista de fórmulas) (Ec. 3.3). Cada una de estas listas caracterizará a un dominio específico. Con el fin de obtener la lista final para cada dominio, se llevará a cabo este proceso varias veces experimentando con los valores de las distintas variables, y al final se evaluará cuál de todas las listas concernientes a un dominio resulta ser la

mejor.

$$S = \{F_1, F_2, F_3, F_4, ..., F_n\}, \tag{3.3}$$

donde S es la salida, F_i es la lista de fórmulas que describe el dominio i, y n es el tamaño de corpus de entrada C_o (Fig. 3.1).

$$F_i = \{f_1, f_2, f_3, f_4, ..., f_l\}, \tag{3.4}$$

donde l es el número de fórmulas que conforman cada lista de fórmulas F_i.

Definimos a una sola fórmula de esa lista de la siguiente manera:

$$f_k = ll[v|ll]^+, \tag{3.5}$$

donde v se refiere a una **palabra vacía** y ll se refiere a una **palabra llena**, estos conceptos fueron definidos en el capítulo anterior.

3.1.3. Los experimentos

Los experimentos deben ser consistentes, es decir, si en alguno de los subcorpora se utiliza un método, en el resto de los subcorpora también tiene que usarse. Aunque en cada experimento se puede jugar con los valores de las variables para evaluar qué combinación de valores dan mejores resultados para cada dominio.

El idioma será una de las variables, ésta influirá en la elección de los algoritmos que emplearemos para la limpieza del texto. Primero, podríamos preprocesar o no los textos (Fig. 3.2). Si decidiéramos hacerlo, podríamos usar el algoritmo de Porter y/o un algoritmo de eliminación de palabras vacías. El segundo tiene sentido cuando se trata de textos en español, debido a que el español se caracteriza por el uso generalizado de preposiciones y artículos, a diferencia del inglés que utiliza sustantivos compuestos. Además se puede o no eliminar las abreviaturas sustituyéndolas por sus valores. Después de tratarlos, cada texto T_i será separado en oraciones; cada oración será analizada

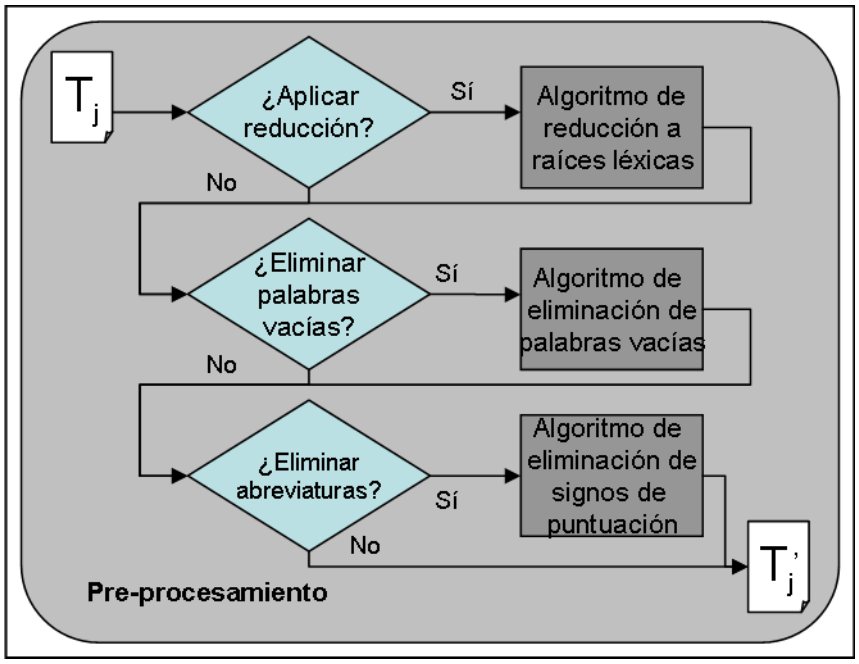

Figura 3.2: Preprocesamiento

y separada en relaciones y todas éstas estarán representadas en un grafo G (Fig. 3.3). Cada relación será un subgrafo dentro de G, y representará la relación entre dos palabras, además sólo se incluirá en el grafo una aparición de cada relación. Entonces cada subgrafo será dirigido, cada nodo representará una palabra y cada vértice la relación entre dos palabras, dada por la distancia que existe entre ellas; la distancia entre dos palabras cualesquiera estará dada por el número de palabras que las separan.

Para definir el tamaño del grafo es importante definir primero a T_i.

$$T_i \ = \ p_1, p_2, p_3, p_g, \tag{3.6}$$

donde cada p_j es una palabra dentro del texto T_i y donde g_i es el número de palabras que contiene el texto T_i (sin repeticiones).

Se puede decir que en el peor de los casos el número de relaciones posibles es la combinación de g_i en $g_i - 1$, es decir que cada palabra está relacionada con todas las palabras excepto consigo misma. Por lo tanto, el orden del grafo es de ${g_i}^2$ y como

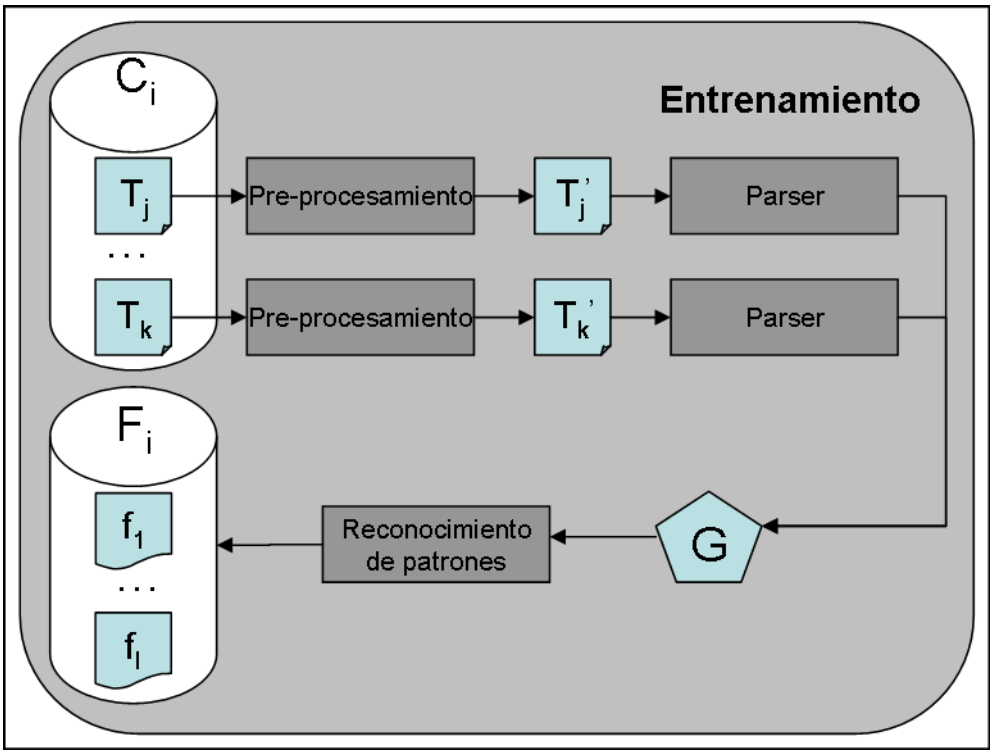

Figura 3.3: Fase de entrenamiento: detalle.

podemos ver está dado por el número de palabras que contienen los textos. Lo anterior nos explica la razón por la que la extensión de los textos influye el desempeño de los algoritmos, mientras mayor sea su extensión (mayor número de palabras) más tiempo tomará construir y navegar el grafo.

La salida del reconocimiento de patrones formará la lista de fórmulas F_i para cada dominio D_i (Fig. 3.3). Posteriormente, en la siguiente fase, se evaluarán todas estas listas de fórmulas generadas por cada uno de los experimentos realizados, y así se podrá determinar qué combinación de los valores de las variables en los experimentos tuvo mejores resultados.

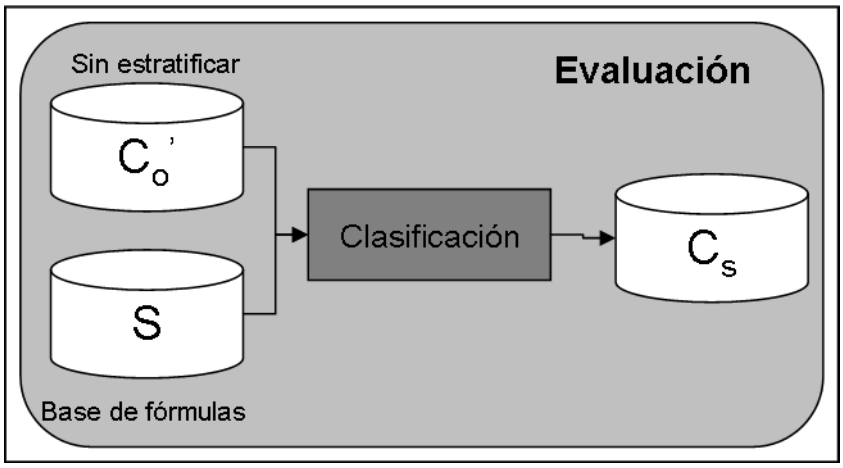

Figura 3.4: Etapa de clasificación

3.2. Fase de evaluación

Esta fase está dividida en dos partes: la primera es la etapa de clasificación y la segunda es la de análisis de resultados. Para la etapa de clasificación se tiene:

3.2.1. Entradas de clasificación

En esta etapa se utilizan los mismos textos de la entrada en la fase de entrenamiento, con la diferencia de que en este caso la entrada estará constituida por un solo corpus que no estará estratificado, es decir, los textos no se encontrarán agrupados *a priori* por dominios. Debido a que el contenido y el tamaño de este corpus son los mismos del C_o y lo único que varía es la agrupación de sus textos, denominaremos C_o' a esta entrada de la etapa de clasificación. En calidad de segunda entrada se tendrá la salida S obtenida como resultado de la fase de entrenamiento, que consiste en el conjunto de las listas de fórmulas generadas para cada dominio o subcorpus (Fig. 3.4).

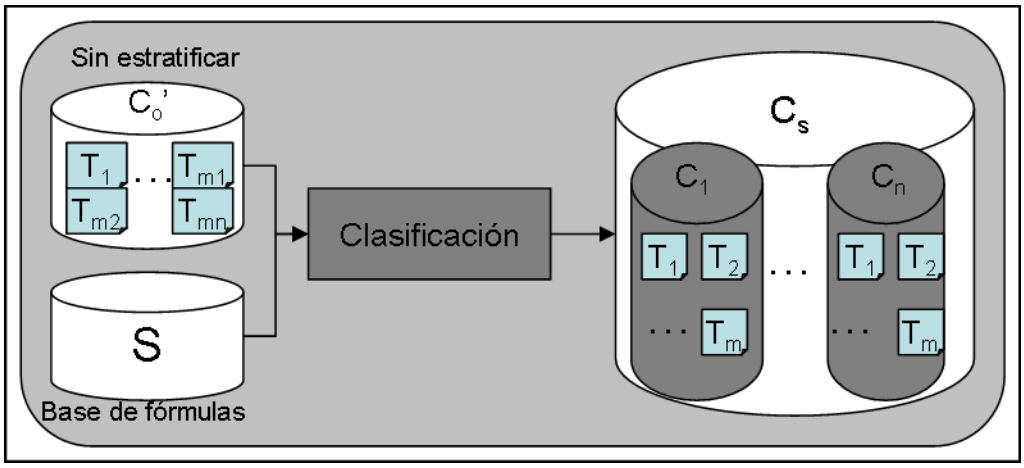

Figura 3.5: Detalle de la etapa de clasificación

3.2.2. Salidas de clasificación

La salida es un corpus estratificado C_s, es decir, a cada texto del $C_o{}'$ se le asociará un dominio y una vez que cada texto tenga asociado un dominio, se les agrupará en base a esto (Fig. 3.5). Idealmente, C_s será idéntico al corpus C_o que nos había servido como entrada a la fase de entrenamiento. De ser ese el caso, la fase de evaluación habrá tenido un éxito del 100 %. El cálculo de porcentaje de éxito se detalla en la sección 3.2.6.

3.2.3. Experimentos de clasificación

Cada texto va a tener un preprocesamiento, como se había mencionado antes: tenemos que asegurarnos que se lleve a cabo el mismo preprocesamiento (Porter, eliminación de palabras vacías, etc.) con el que obtuvimos las listas de fórmulas con las que vamos a comparar cada texto. Una vez procesado, se parsea a grafo y utilizamos el algoritmo de clasificación comparando cada texto con cada una de las listas de fórmulas, para obtener un índice de similitud de ese texto con cada dominio (Fig. 3.6); el cálculo de este indicador se detallará más adelante. Luego se eligen los dominios que estén por arriba del 50 % de similitud y éstos se ordenan en forma ascendente y se elige al mayor

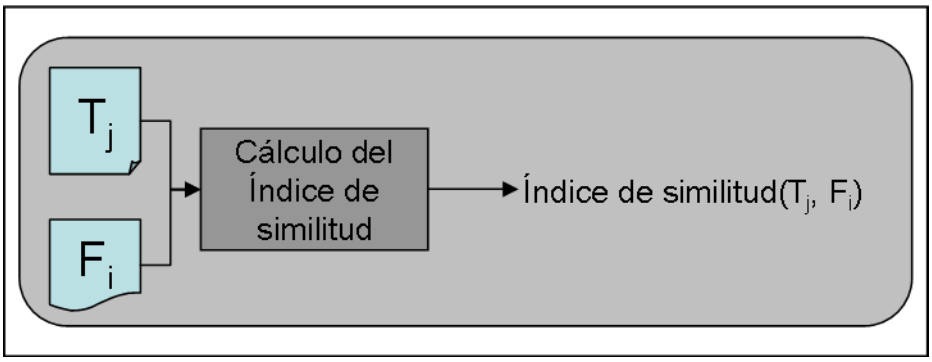

Figura 3.6: Cálculo del índice de similitud

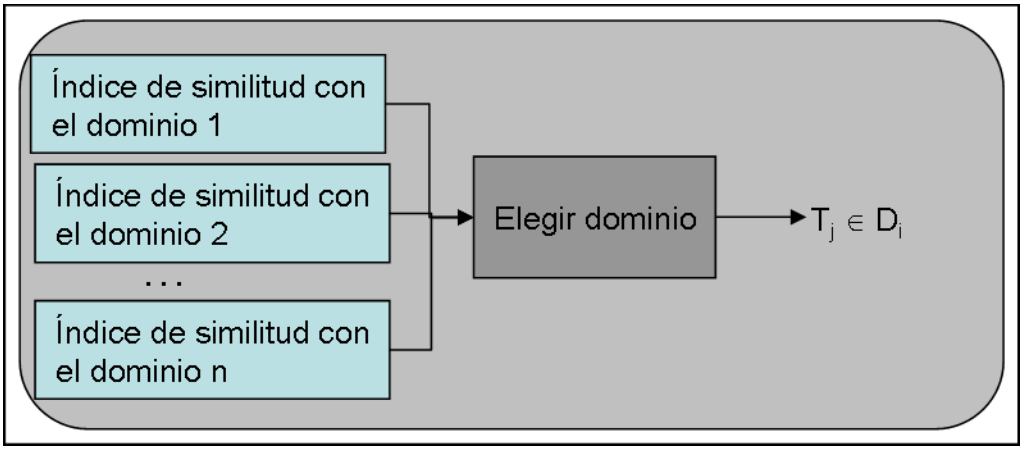

Figura 3.7: Determinación del dominio

(Fig. 3.7). Cada fórmula tiene asociado un peso o relevancia, la cual se calcula como el producto entre la carga semántica y la representatividad de la fórmula; estos conceptos se definen a continuación.

$$Peso(C_i, f_k) = Carga\ semántica(f_k) * Representatividad(C_i, f_k) * Longitud(f_k), \tag{3.7}$$

donde la longitud de f_k es el número total de términos en la fórmula. La carga semántica es un concepto que permite restar relevancia a fórmulas que están compuestas por palabras vacías, que, como definimos en el capítulo anterior, carecen de significado

semántico. Para calcular la carga semántica tenemos que:

$$Carga\ semántica(f_k) \ = \ \frac{No.\ de\ términos\ llenos}{Longitud(f_k)}, \tag{3.8}$$

Está ecuación les da prioridad a las fórmulas que tienen mayor cantidad de palabras con un significado semántico. A su vez, la representatividad de la fórmula denota qué tan característica es ésta dentro de un dominio dado, y se establecerá en la fase de refinamiento de fórmulas.

$$Representatividad(C_i, f_k) \ = \ \left(\frac{Partición(C_i, f_k)}{Tamaño\ de\ C_i}\right) \tag{3.9}$$

donde partición se entiende como el número de textos de C_i en los que aparece la fórmula f_k.

$$Frecuencia(C_i, f_k) \ = \ \left(\frac{Total\ de\ ocurrencia(C_i, f_k)}{Tamaño\ de\ C_i}\right), \tag{3.10}$$

donde

$$\left(\frac{Total\ de\ ocurrencia(C_i, f_k)}{Tamaño\ de\ C_i}\right) \ = \ \left(\frac{\sum Total\ de\ ocurrencia(T_j, f_k)}{Tamaño\ de\ C_i}\right), \tag{3.11}$$

donde $T_j \in C_i$. Se decidió usar la representatividad en lugar de la frecuencia, debido a que nos interesa conocer la probabilidad de que la fórmula ocurra dentro de un texto en un dominio (subcorpus). Entre más textos la contengan, más representativa del dominio es la fórmula. El que una fórmula aparezca más de una vez dentro de un texto no es indicador de que vaya a aparecer en los demás textos del mismo dominio. Lo anterior confirma la aseveración de que la frecuencia no es indicador de la formulaicidad pero sí un factor para su determinación. Debido a esto, normalizamos a 1 todas las incidencias de una fórmula en un solo texto para descartar sus múltiples apariciones dentro del mismo, generando de esta manera la noción de representatividad. Así, la representatividad es en efecto la frecuencia normalizada de la fórmula en un dominio

dado, y con su uso pretendemos considerar la frecuencia de fórmulas ligada al contexto en el que ocurren.

El peso de las fórmulas es uno de los factores para el cálculo del índice de similitud entre un texto T_j y una lista de fórmulas F_i que caracteriza el dominio D_i representado por la muestra del subcorpus C_i. Dicho índice determina la probabilidad de pertenencia de T_j a D_i y por lo tanto a C_i; basados en él se toma la decisión de agrupar los textos en subcorpora.

$$\textit{Índice de similitud}(T_j, F_i) \;=\; \left(\frac{\sum Peso(F_i, f_{oc})}{\sum Representatividad(F_i, f_{oc})} \right), \tag{3.12}$$

donde $f_{oc} \in F_i \wedge f_{oc} \in T_j$.

Por otro lado, para la etapa de análisis de resultados se tiene lo siguiente:

3.2.4. Entradas de análisis

Se tienen dos elementos como entrada: el primero es el corpus estratificado obtenido de la etapa de clasificación C_s, el segundo es el corpus de entrada de la fase de entrenamiento que también se encuentra estratificado C_o (Fig. 3.8).

3.2.5. Salidas de análisis

Se calcula el porcentaje de éxito (Fig. 3.8), obtenido como resultado de comparar los corpus de entrada C_s y C_o.

Este porcentaje de éxito se define como el grado de similitud entre C_s y C_o y será usado como **criterio de terminación**, para dar por terminada ésta fase (Fig. 3.8).

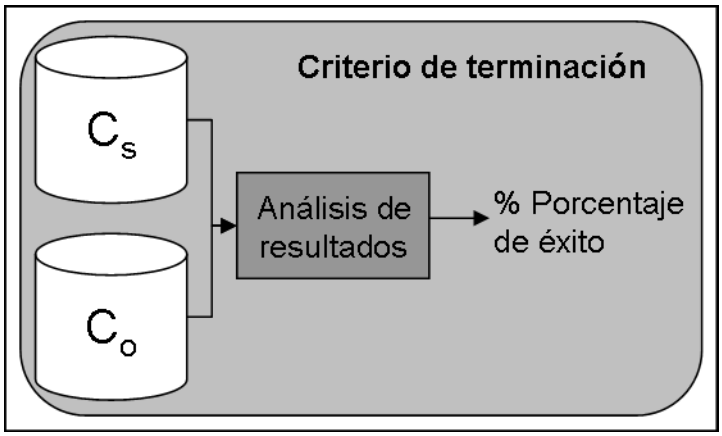

Figura 3.8: Fase de evaluación: análisis de éxito

$$Porcentaje\ de\ éxito(C_s, C_o) = 100 * \left(\frac{No.\ de\ textos\ clasificados\ correctamente\ en\ C_s}{Tamaño\ de\ C_o} \right) \%, \tag{3.13}$$

donde el tamaño de C_o es el total de textos a clasificar y un mismo texto deberá pertenecer a la misma clase en C_s que en C_o para ser considerado como clasificado correctamente. Es decir, se debe cumplir la siguiente condición para el máximo posible de los textos (Ec. 3.14).

$$T_i \in C_s \wedge T_i \in C_o \rightarrow T_i \in C_i \wedge C_i \in C_s \wedge C_i \in C_o \tag{3.14}$$

3.2.6. Experimentos de análisis

El resultado esperado de esta fase es la maximización del porcentaje de éxito en la clasificación de los textos en sus respectivos dominios. En caso de que el porcentaje de éxito sea menor al 100 %, se repetirá varias veces el proceso de entrenamiento, hasta que el porcentaje de éxito se estabilice y ya no mejore más, es decir, hasta que hayamos alcanzado un máximo local; dicho máximo local constituirá nuestro criterio

de terminación para dar paso a la siguiente fase del algoritmo.

3.3. Refinamiento de fórmulas

Es una fase opcional por lo tanto se le puede brincar y pasar directamente a la fase de pruebas. La razón por la que esta fase puede ser omitida es que no es un paso esencial para que el algoritmo de clasificación funcione.

3.3.1. Entradas

Como entrada tenemos el conjunto de fórmulas obtenidas en la etapa de entrenamiento S: en esa lista las fórmulas tienen tamaño definido y términos fijos, es decir, las fórmulas están compuestas por una o dos palabras como máximo y que no son variables, en otras palabras, que siempre están presentes y con el mismo valor.

3.3.2. Salidas

Como salida obtendremos una nueva lista de fórmulas S' la cual estará conformada de fórmulas de tamaño variable, es decir, las fórmulas pueden estar compuestas de cualquier número de términos y además podrán contener términos variables (var). Las fórmulas entonces tendrán la siguiente estructura:

$$f_k = const[var|const]^*, \tag{3.15}$$

donde $const$ se refiere a un término fijo y

$$var = [n|o], \tag{3.16}$$

donde n se refiere a que una variable se encuentra presente en la fórmula necesariamente y o es una variable opcional.

$$const = [v|ll], \tag{3.17}$$

$$n = [v|ll], \tag{3.18}$$

$$o = [v|ll], \tag{3.19}$$

donde v se refiere a una **palabra vacía** y ll se refiere a una **palabra llena**.

3.3.3. Experimentos

En realidad, el objetivo de ésta fase es obtener fórmulas mucho más generales, introduciendo el concepto de variable. Como la definición de la fórmula cambió para esta fase (ec. 3.15), será necesario redefinir la carga semántica de la fórmula. La forma de calcular la carga semántica cambia debido a que se introduce un nuevo criterio en su determinación, necesitamos tomar en cuenta que los términos pueden ser constantes y variables. Es por ello que podemos considerar dos formas alternativas de calcularla:

1.
$$\left(\frac{No.\ de\ términos\ fijos\ consecutivos\ (const)}{Tamaño\ de\ la\ fórmula}\right), \tag{3.20}$$

2.
$$\left(\frac{Total\ de\ términos\ fijos\ llenos}{Tamaño\ de\ la\ fórmula}\right), \tag{3.21}$$

La utilidad de la ecuación 3.20 será influenciada por el idioma en el que estén escritas las fórmulas, por ejemplo en inglés existen sustantivos compuestos, lo cual se traduce en un mayor número de **palabras llenas** consecutivas a diferencia del español donde los sustantivos se combinan con **palabras vacías**, por ello es más difícil que haya términos no vacíos consecutivos.

La ecuación 3.21 sólo se conforma con que la fórmula tenga constantes llenas, sin importar si son consecutivas, por lo cual es más adecuada para idiomas con gramáticas

similares a la del español, mientras que la ecuación 3.20 provee resultados más exactos para idiomas cuya gramática sea similar a la del inglés.

Las ecuaciones anteriores son opciones: dependiendo del caso se puede elegir una sobre la otra y ver cuál funciona mejor dentro de nuestros experimentos. Estos cálculos influirán directamente en el orden de relevancia o peso de las fórmulas asociadas a un dominio.

Las fórmulas de salida son generadas concatenando las fórmulas de entrada y sustituyendo algunos de sus valores por variables. El proceso de sustitución de valores por variables fue desarrollado por Ostrovskaya [21]. La manera de concatenar las fórmulas se explica a continuación.

Generalización

Una vez que tengamos ordenadas las fórmulas por relevancia, podemos generalizar las fórmulas extendiéndolas, es decir, aplicar transitividad y concatenar varias fórmulas. Lo anterior, se logrará definiendo un lenguaje lógico formal usando operadores lógicos ($\wedge$,$\vee$,$\rightarrow$). Así nuestra lista de fórmulas tendrá la siguiente estructura:

$$a \rightarrow b, b \rightarrow c, a \rightarrow c, \tag{3.22}$$

donde a, b y c son palabras y, aplicando transitividad, podemos concluir que:

$$a \rightarrow b \rightarrow c, \tag{3.23}$$

por lo tanto nuestra nueva fórmula tendrá ahora tres términos fijos(const). La anterior representación no es la definitiva. Se tiene que tomar en cuenta el valor de la relación, es decir, la distancia con la que las palabras están relacionadas. Ésta influirá en la estructura de la generalización, esto debido a que la distancia entre dos palabras puede

variar. En realidad tendremos una lista de fórmulas como la que sigue:

$$a \overset{x1}{\rightarrow} b, a \overset{y1}{\rightarrow} b, b \overset{x2}{\rightarrow} c, b \overset{y2}{\rightarrow} c, a \overset{x3}{\rightarrow} c, \tag{3.24}$$

donde cada implicación es una relación, y $x1$, $x2$, $y1$, $y2$ y $x3$ son las distancias con las que se relacionan dos palabras. Se desea reducir el número de formas en que se relacionan dos términos, para ello será necesario elegir la distancia que englobe o incluya al mayor número de casos y después agregar el número de variables que sean necesarias para hacer mucho más generales las fórmulas. Entonces, para elegir la distancia se pueden hacer dos cosas:

1. tomar el valor de la distancia más común en los textos, es decir la relación que tenga mayor frecuencia; pero esto podría dejar de lado varios casos, que de esta forma parecerían casos particulares, pero que en realidad son un caso aparte,

2. tomar la relación cuya distancia sea la mayor aunque es probable que también se pierdan algunas fórmulas.

La estructura de las relaciones de transitividad debe ser la siguiente para poder llevar a cabo la generalización:

$$\bar{ab} \;=\; s1,\; \bar{bc} = s2 \; y \; \bar{ac} = s3, \tag{3.25}$$

donde

$$s3 \geq (s1 + s2), \tag{3.26}$$

con las características anteriores se obtiene:

$$a \overset{s1}{\rightarrow} b \overset{\geq(s1+s2)}{\rightarrow} c \tag{3.27}$$

donde $s1$, $s2$ y $s3$ son los valores de las relaciones entre palabras. Una vez que se tiene generalizada esta fórmula, se hará un análisis entre las relaciones de estas palabras

y aquellas que se relacionen con una distancia mayor a uno ameritará, la inserción de términos variables necesarios. Como se había mencionado anteriormente, la elección de la distancia ideal puede tener como consecuencia que algunas de las variaciones de las fórmulas sean omitidas, es por ello que una vez que se ha terminado este proceso, se seguirá repitiendo hasta que ya no sea necesario hacer más generalizaciones, es decir hasta que ya no hay condiciones para realizar más cambios.

Con el proceso anterior se obtienen fórmulas mucho más generales, qué están en lenguaje natural y que desde el punto de vista de usuario describen mucho mejor al dominio porque tienen un formato mucho más natural.

3.4. Fase de pruebas

Esta fase, al igual que la fase de evaluación, se puede dividir en dos etapas: clasificación y análisis de resultados. La única diferencia es que cambian los valores de entrada y de salida, mientras que los experimentos permanecen los mismos.

3.4.1. Entradas de clasificación

El procedimiento de la primera etapa será el mismo que el de la fase de evaluación, por lo cual el corpus de entrada tendrá las mismas características, a saber: heterogéneo y no estratificado. La diferencia radica en que el corpus de prueba C_p no podrá contener los mismos textos que los córpora de entrenamiento y evaluación (C_o y C_o'), debido a que el objetivo de esta etapa es probar la eficacia del algoritmo en situaciones nuevas, es decir, con textos nunca antes vistos por el mismo. Los tamaños de los textos pueden variar, incluso puede haber textos que no pertenezcan a ninguno de los dominios para los que fue entrenado el algoritmo. En calidad de segunda entrada se tendrá la salida S obtenida como resultado de la fase de entrenamiento, que consiste en el conjunto de

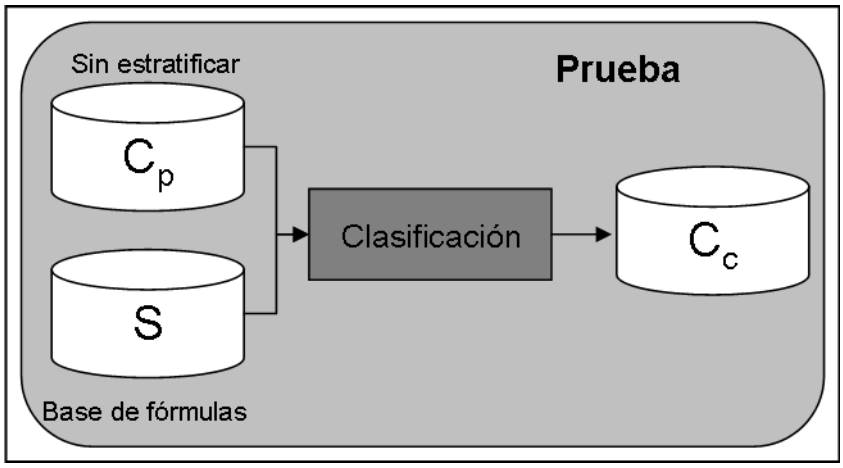

Figura 3.9: Fase de prueba: etapa de clasificación

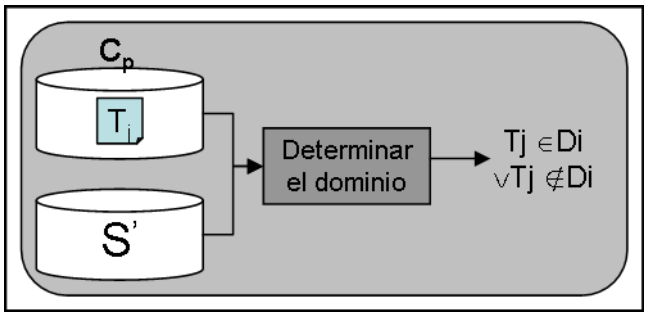

Figura 3.10: Fase de prueba: detalle etapa de clasificación

las listas de fórmulas generadas para cada dominio o subcorpus. Si es que se utiliza la fase de refinamiento de fórmulas, la segunda entrada para las pruebas será la salida S' de aquella fase opcional (Fig. 3.9).

3.4.2. Salidas de clasificación

El resultado de la primera subfase es la clasificación de cada uno de los textos C_c, esperando que sea de una manera eficaz. En el caso de los textos que pertenecen a dominios para los que no fue entrenado el algoritmo, el sistema debe aislarlos e indicar que no pertenecen a un dominio conocido (Fig. 3.10)

3.4.3. Experimentos de clasificación

Al igual que en la fase de evaluación, cada texto va a tener un preprocesamiento. Es importante asegurarse de que prevalezca la consistencia, es decir, el corpus de prueba debe ser tratado de la misma manera que el de entrenamiento. Después del preprocesamiento se sigue el mismo proceso que en la fase de evaluación (Véase sección 3.2.3: se parsea a grafos el corpus y se utiliza el algoritmo de clasificación, obteniendo el índice de similitud con cada dominio para cada uno de sus textos (Fig. 3.6). Comparando los índices de similitud de un texto a cada uno de los dominios, se toma la decisión de clasificar dicho texto como perteneciente al dominio con el cual presenta mayor índice de similitud. Para desambiguar entre dos índices con el mismo valor, se recalcularán dichos índices basándose en la frecuencia en lugar de la representatividad, y se optará por el mayor. Lo anterior no nos dará la certeza de una clasificación inequívoca para el 100 % de los casos ambiguos pero nos proporcionará un criterio para tomar la decisión.

3.4.4. Entradas de análisis

En la segunda subfase se probará la eficacia del método de clasificación mediante un análisis de los resultados que éste arrojó. Para esta subfase se tendrá de entrada el corpus $C_p^{'}$ formado por los mismos textos de C_p pero que están clasificados por un humano, es decir, cada texto tiene asociado el dominio al que pertenece. En calidad de segunda entrada se tendrá el corpus C_c que es el corpus de salida de la primera subfase.

3.4.5. Salida de análisis

Esta subfase arrojará como salida el porcentaje de éxito, el cual medirá el porcentaje de textos que fueron clasificados correctamente.

3.4.6. Experimentos

Una vez clasificados los textos, podremos evaluar el comportamiento del sistema, es decir, determinar su eficacia para la categorización de los textos. Ésta será calculada como el porcentaje de éxito (Ec. 3.22).

$$Porcentaje\,de\,éxito(C_c, C_p^{'}) \;=\; 100 * \left(\frac{N\acute{u}mero\,de\,textos\,clasificados\,correctamente\,en\,C_c}{Tamaño\,de\,C_p} \right), \tag{3.28}$$

donde el tamaño de C_p es el total de textos a clasificar.

Al inicio de este capítulo, se señalo como expectativa u objetivo de este proyecto lograr la maximización del **porcentaje de éxito** en la clasificación de documentos, concepto que ya ha sido definido propiamente. Por lo anterior se puede concluir que el porcentaje de éxito de esta fase debe ser por lo menos igual al porcentaje de éxito obtenido en la etapa de evaluación. Lo anterior es nuestra premisa para la evaluación de la eficacia del proceso de clasificación de textos. La determinación de la correcta clasificación de textos consiste en la comparación de estos dos porcentajes que será llevada a cabo por el experto humano, es decir, determinará con su criterio si estos dos porcentajes son aceptables y con ello concluir qué tan bueno fue el comportamiento del sistema.