Capítulo 6

Pruebas y resultados

En esta sección se planteará la mecánica de cada prueba y el análisis de los resultados obtenidos.

6.1. Recolección de las muestras

Se requieren dos muestras principales para llevar a cabo las pruebas. En el módulo de entrenamiento es necesario un corpus cuyas características fueron definidas en el capítulo de metodología; en el módulo de pruebas se requiere otro corpus, de la misma manera sus características están descritas en el capítulo de metodología. A continuación se explica la forma de su recolección y la estructura que se les dio.

6.1.1. Córpora de entrenamiento

Se eligieron dos dominios en español: el primero fue **cartas de amor** y el segundo **cartas comerciales**. Como ya se había dicho anteriormente, el criterio utilizado para la elección de los dominios fue la intuición, es decir, nos basamos en la premisa de que intuitivamente estos dominios parecen formulaicos, aunque no se tendrá la certeza de ello hasta después de que se realicen las pruebas. El primer corpus, dominio de las

cartas de amor, se obtuvo buscando páginas en español en la Web, de donde tomamos 34 ejemplares de cartas de amor. Para el segundo corpus, dominio de cartas empresariales, se digitalizaron 27 documentos pertenecientes al material de tareas de un curso de secretariado; se tiene autorización de la autora de los documentos. En el apéndice D está descrito el corpus de entrenamiento. Aunque estos números parecen pequeños, por el tipo de dominio se puede especular que son suficientes; si en efecto el dominio es formulaico no debe afectar el tamaño de la muestra, ya que las fórmulas que realmente lo describan deben tener una aparición importante aunque no sea un gran corpus. Por ejemplo, se puede especular que la fórmula "te amo" es representativa del dominio de las cartas de amor, entonces no importa si son muchas o pocas cartas de amor, la gran mayoría deben tener esta expresión por lo tanto se reconocerá como una fórmula. En realidad en lo único que podría afectar es que no se reconozcan como fórmulas algunas que no son tan representativas para el dominio, o tienen una relevancia media.

Estructura de datos y formato

Los textos fueron guardados en formato .txt y cada dominio se describió en un documento también .txt que se utilizó como directorio de los textos que conformaban cada subcorpus. El corpus de entrenamiento está descrito entonces con un archivo .txt donde se incluye la localización de cada subcorpus que lo conforma.

6.1.2. Corpus de prueba

Fueron recolectados de la misma forma que los textos del corpus de entrenamiento. A diferencia del anterior, en este caso se incluyeron también textos ajenos a los dominios para los que fue entrenado el algoritmo. El corpus está constituido de:

- 13 cartas de amor,

- 14 cartas empresariales y
- 10 documentos de varios dominios desconocidos.

Estructura de datos y formato

Como se había dicho antes el corpus de prueba no está estratificado, es decir, los textos que lo conforman no estarán agrupados por dominios. Existe un archivo .txt dónde se hace referencia a todos los textos sin que se encuentren ordenados o asociados con el dominio al que pertenecen.

6.2. Fase de entrenamiento

6.2.1. Preprocesamiento

Para la generación de fórmulas se experimentó con todas las posibles combinaciones que resultan de utilizar o no los algoritmos de preprocesamiento, lo anterior para ver cómo afectan a la generación de la lista de fórmulas. Como se tienen tres algoritmos y cada uno puede ser usado o no, las combinaciones son de dos en tres y por lo tanto se realizaron ocho experimentos. Al final obtuvimos 8 listas de fórmulas, aunque para la fase de evaluación y la fase de prueba la lista utilizada fue la que se obtuvo como resultado de aplicar el preprocesamiento completo. El resto de las listas de fórmulas simplemente nos sirvió para corroborar que a mayor preprocesamiento mayor calidad de las fórmulas; esta calidad se midió en función del peso promedio de las fórmulas, a mayor peso mayor calidad. Esta medida de calidad es subjetiva o de apreciación. Los resultados obtenidos se pueden apreciar en la figura 6.1.

En la siguiente tabla se puede ver como afecta el grado de preprocesamiento al tamao de la lista de fórmulas generada para cada dominio (Fig. 6.2).

Peso Promedio	Abre	Pal_Vac	Stemming	Abre_Pal_Vac	Abre_Stem	Pal_Vac_Stem	Todos	Ninguno
Dominio1	0.77573531	0.8235294	0.79705884	0.82352944	0.807189566	0.857142883	0.882353	0.781045767
Dominio2	1.29629628	1.2962963	1.33333333	1.296296283	1.333333333	1.333333333	1.3333333	1.296296283

Figura 6.1: Comparativa del peso promedio de las fórmulas obtenidas

No. Palabras	Abre	Pal_Vac	Stemming	Abre_Pal_Vac	Abre_Stem	Pal_Vac_Stem	Todos	Sin
Dominio1	8	7	10	6	9	7	6	9
Dominio2	6	6	6	6	6	6	6	6

Figura 6.2: Comparativa del tamaño de fórmulas con el grado de preprocesamiento

Algoritmo de Porter

En el apéndice E se encuentran las salidas de este algoritmo, en éstas se pueden identificar algunos de los errores que presenta el algoritmo. Estos errores afectaron directamente la calidad de las fórmulas. Uno de los errores que más cometió el algoritmo fue que en algunos casos no reduce lo suficiente; por ejemplo con la palabra "señorita" que es reducida a "señorit" y el resultado deseable es que tenga la misma raíz que "señor" y "señora". Casos como éste afectan el valor del peso de la fórmula, porque la fórmula que tiene la palabra "señorita" no se generaliza en la misma fórmula de "señor" y esto afecta el valor de su representatividad, es decir, disminuye su valor, porque la fórmula general y ésta contarían como dos distintas, en lugar de contar como una y que su valor aumente. Casos más elementales como el verbo "es" que es reducido a nada, esto atenta contra uno de los principios básicos del algoritmo de Porter, donde palabras de tamaño dos no deben ser reducidas. Además en el caso de palabras como "ante" y "cabe" se tuvieron que manejar por fuera porque la implementación del algoritmo fallaba, generaba excepciones.

Palabras vacías

En el apéndice E se encuentran las salidas de este algoritmo. Se puede notar que si se aplica el algoritmo, se afecta el tamaño de la lista de fórmulas; cuando no se aplica, la lista es un poco más larga.

Abreviaturas

En el apéndice E se encuentran las salidas de este algoritmo. Se puede notar que si se aplica el algoritmo, se afecta el tamaño de la lista de fórmulas, porque al sustituirse los valores que representan, éstos pueden reducirse con el algoritmo de Porter y se pueden hacer más generales.

6.2.2. Lista de fórmulas

En el apéndice F se encuentran las listas de fórmulas generadas para cada dominio. Como ya se había dicho, para la fase de evaluación y pruebas se empleó para clasificar la lista que se obtuvo de realizar el preprocesamiento completo así que esta es la que nos interesa.

6.3. Fase de evaluación

6.3.1. *Clustering*

Se obtuvo como resultado un archivo resultado.txt donde se tiene asociado cada texto de los corpora de entrenamiento al dominio al que pertenece y el índice de similitud. En este caso, los resultados del *clustering* de la fase de evaluación se encuentran en el apéndice G.

6.3.2. Análisis

Comparando el archivo obtenido en la etapa de *clustering* y el archivo corpus_ana.txt, donde se tenían previamente clasificados los textos, se realizó el cálculo del **porcentaje de éxito**. Se obtuvo que 85 % de los textos se clasificaron correctamente. Este porcentaje se traduce en que 52 de 61 textos fueron clasificados correctamente, pero si se analizan los dominios por separado se puede ver que todos los textos que son cartas empresariales fueron clasificados correctamente. Por otro lado, los 9 textos clasificados incorrectamente pertenecen al dominio de las cartas de amor. Debido a que no había certeza de que los dominios fueran formulaicos y, al analizar estos resultados, se puede ver que el dominio de las cartas empresariales parece tener un índice de formulaicidad alto; por el contrario, el índice de formulaicidad del dominio de las cartas de amor parece no serlo tanto. Pero no se puede concluir que ese dominio es completamente no formulaico, porque aproximadamente 74 % de estos textos fueron correctamente clasificados. Aunque la última palabra se tiene en la fase de pruebas. Éste es un resultado alentador, ya que pone en evidencia el potencial del método aplicado para la detección del grado de formulaicidad de un dominio.

6.4. Fase de pruebas

6.4.1. *Clustering*

Se obtuvo como resultado un archivo resultado_p.txt donde se tiene asociado cada texto de los córpora de entrenamiento al dominio al que pertenece y el índice de similitud. En este caso, los resultados del *clustering* de la fase de pruebas se encuentran en el apéndice G.

6.4.2. Análisis

Comparando el archivo obtenido en la etapa de clasificación y el archivo corpus_ana_p.txt, donde se tenían previamente clasificados los textos, se realizó el cálculo del **porcentaje de éxito**. Se obtuvo que 81 % de los textos se clasificaron correctamente. Este porcentaje se traduce en que aproximadamente 30 de 37 textos fueron clasificados correctamente.

Al igual que en la fase de evaluación, todos los textos que son cartas empresariales fueron clasificados correctamente. También los textos de dominios desconocidos fueron identificados. Por otro lado, los 7 textos clasificados incorrectamente pertenecen al dominio de las cartas de amor. Éste resultado parece confirmar que el dominio de las cartas empresariales tiene un índice de formulaicidad alto, mientras que el dominio de las cartas de amor tiene un ndice de formulaicidad alto. Las cartas de amor a diferencia de las cartas empresariales tienen un estilo mucho más literario, a pesar de esto se pudieron encontrar fórmulas que pueden clasificar gran parte de los textos que pertenecen a este dominio.

6.5. Dominios

Al final de las pruebas se pudo ver que el primer dominio elegido (cartas de amor) no es tan formulaico, porque aunque es posible identificar una carta de amor, se necesitaría una base de conocimiento mucho más grande; hay muchas variantes en este dominio, y para encontrar fórmulas que sean lo suficientemente representativas se debe tener una muestra mucho más grande. Además, se tendrían mejores resultados si se tuviera un diccionario de sinónimos, porque como el objetivo de estás cartas es ser lo más literario y original, el lenguaje no es tan restringido y, por lo tanto, hay muchas variantes de cada palabra. Aunado a esto el tamaño influye porque, mientras más pequeña la muestra,

pues menos oportunidad hay de que las mismas palabras hayan sido usadas en cada uno de los textos. Creo que el problema es que este dominio podría fácilmente incluir otros, y estos subdominios serían mucho más fáciles de clasificar si la muestra fuera muy grande.

Al final se puede ver que el tamaño de la muestra no afecto negativamente los resultados, se pudieron encontrar fórmulas suficientemente representativas como para que se clasificaran correctamente la mayoría de los textos.

Por otro lado, el segundo dominio elegido presenta un lenguaje más restringido y por lo tanto se obtuvieron mejores resultados en la clasificación. Se puede concluir que hay una dependencia con el dominio, es decir, mientras más restringido el lenguaje, son mejores son los resultados.

6.6. Complejidad del algoritmo

Después de analizar el algoritmo de reconocimiento de patrones para detectar fórmulas se concluyó que su complejidad es n^2 donde n es el número de palabras en una oración. Empíricamente lo anterior se puede ver en la siguiente gráfica, en ella se asoció el tiempo con el número de palabras en una oración:

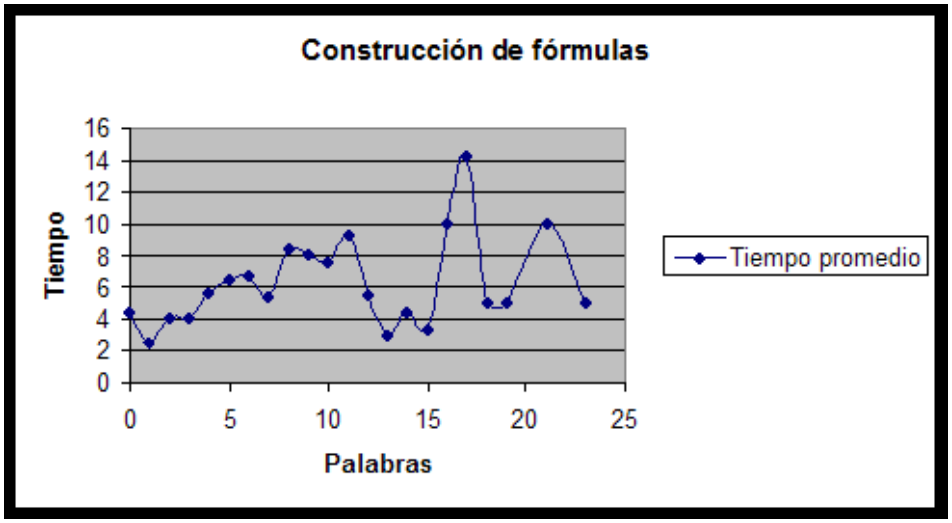

Figura 6.3: Relación tiempo con número de palabras