Métodos de Refinamiento de la Clasificación Translingüe de Documentos

Por

Adelina Escobar Acevedo

Tesis sometida como requisito parcial para obtener el grado de
Maestro en Ciencias del Área de Ciencias Computacionales

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica
INAOE

Supervisada por:

Dr. Manuel Montes y Gómez, INAOE
Dr. Luis Villaseñor Pineda, INAOE

Tonantzintla, Puebla
Septiembre 2009

A mis padres que han sabido formarnos, impulsarnos y motivarnos para ser capaces de alcanzar nuestras metas.

Gracias por tu fortaleza mamá, por tu ecuanimidad papá y por tu ejemplo Marco

Agradecimientos

Agradezco profundamente a mis asesores Dr. Manuel Montes y Gómez y Dr. Luis Villaseñor Pineda por su amable guía y colaboración en la realización de este proyecto.

A mis sinodales: Dr. J. Francisco Martínez Trinidad, Dra. Ma. Pilar Gómez Gil y Dr. Saúl E. Pomares Hernández por sus comentarios, consejos y facilidades prestadas para dar los toques finales al trabajo.

A los profesores de Ciencias Computacionales y al grupo de compañeros de Tecnologías de Lenguaje por siempre estar dispuestos a resolver alguna duda y a proporcionar unos minutos de charla.

A los compañeros de maestría por su amistad, confianza y camaradería. Con especial cariño a Mario.

Al personal de INAOE por las facilidades proporcionadas durante nuestros estudios.

A CONACYT por darnos la oportunidad de continuar estudiando y de otorgarme la beca No. 212424

Resumen

Gracias a los continuos avances tecnológicos, cada día se tiene acceso a documentos escritos en diversas partes del mundo, ameritando el uso de procesos automáticos para su organización. La clasificación automática de textos multilingües se plantea el problema de clasificar documentos escritos en diferentes idiomas bajo las mismas clases. Ante ello, la clasificación translingüe representa una solución viable ya que utiliza herramientas, como los traductores automáticos, para franquear la barrera del lenguaje con el objetivo de aprovechar recursos existentes en uno o varios idiomas. No obstante, la clasificación translingüe ha demostrado ser insuficiente por si sola debido a distorsiones introducidas por el traductor y a las diferencias culturales. En el presente trabajo se proponen dos métodos de refinamiento de la clasificación translingüe inicial usando el conjunto de documentos que se desea clasificar. El primer método aplica un procedimiento posterior a la clasificación translingüe, auxiliándose de las similitudes entre los documentos no clasificados, para hacer un refinamiento sin modificar el clasificador. El segundo método selecciona documentos clasificados de forma confiable, para incorporados al proceso de clasificación translingüe, a fin de adecuar el clasificador al nuevo idioma. Los resultados obtenidos llegan a superar en 17% a la clasificación translingüe tradicional en el mejor caso.

Abstract

Thanks to technological advances, every day we have access to large volumes of multilingual information around the world. That amount of text requires the use of automated processes for its organization. Multilingual Text Classification takes the problem of classifying documents written in different languages under the same classes. In response, Cross-language Text Classification represents a possible solution because, supported by tools like automatic translators, aims to take advantage of existing resources from one language to classify another. However Cross-language Text Classification has proved to be insufficient due to translation issues and cultural differences. This work proposes two refinement methods of cross-language text classification considering information from the target language documents. The first method applies a post-classification procedure, aided by nearest neighbors between classified documents, to make a refinement without changing the classifier. The second method selects reliably classified documents, to be joined to the cross-language classification process, in order to produce an adapted classifier for target language documents. Experimental results are encoraging, showing an improvement in the classification accuracy as high as 17%.

Índice General

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

El lenguaje se originó hace miles de años por nuestros ancestros como consecuencia del desarrollo de la sociedad humana. Las lenguas son instrumentos de que disponen los seres humanos para la interacción y la expresión de ideas, sentimientos, conocimientos, memorias y valores (portal UNESCO, 2009). A través de ellas, el hombre tiene la posibilidad de acumular el conocimiento por medio de la comunicación oral y escrita.

La escritura, base de la existencia de los textos, surge de la voluntad de dar una materialidad perdurable a las experiencias del hombre. En los medios escritos, se encontró una forma de preservar y transmitir la cultura tanto en el espacio como en el tiempo. La creación de medios específicos de divulgación del conocimiento desencadenó una escalada de textos impresos y electrónicos. Dichos documentos conforman hoy repositorios desmedidos de información en continuo crecimiento.

Debido a la imposibilidad humana de manejar enormes cantidades de textos, nace la necesidad de automatizar procesos que manipulen y organicen tales volúmenes de documentos (Galicia Haro y Gelbukh, 2007). A fin de aprovechar la información contenida en los documentos, surgen

distintas líneas de investigación como los sistemas de recuperación de información (*Information Retrieval*), para buscar datos relevantes en grandes repositorios de documentos (Baeza y Ribeiro, 1999); búsqueda de respuestas (*Question Answering*), que permite a los usuarios plantear preguntas en lenguaje natural para la recuperación de información (Vicedo et al., 2003); generación automática de resúmenes (*Text Summarization*), que pretende extraer las líneas importantes de los documentos a fin de agilizar su lectura (Hovy y Lin, 1999); clasificación de textos (*Text Categorization*), para asignar etiquetas a los textos de forma automática (Sebastiani, 2002); entre otros.

1.1 Clasificación Automática de Textos

La clasificación o categorización se define como la tarea de asignar objetos de un universo a dos o más clases predefinidas. Se conocen como clases o categorías a las opciones con las que se puede asignar una etiqueta. Para tomar una decisión de la clase a la que pertenecen los objetos, es necesario conocer las características particulares de cada clase (Sierra Araujo et al., 2006). La clasificación automática de textos, en particular, tiene por objetivo asignar automáticamente la clase apropiada a cada documento y deriva en aplicaciones como filtrar un caudal de noticias para un grupo de interés particular; clasificación de textos de opinión, sentimiento o juicio; atribución de autoría; organizar grandes volúmenes de información de acuerdo a cierta taxonomía, por ejemplo, en bibliotecas, etc.

La clasificación automática de textos en el llamado aprendizaje computacional (o aprendizaje automático) es comúnmente una tarea supervisada, lo que significa que no sólo se conocen previamente las categorías sino que debe contarse con un conjunto de entrenamiento. El

conjunto de entrenamiento es un grupo de documentos previamente etiquetados con las clases a las que pertenecen, donde cada documento es representado por una serie de variables características de la clase. La tarea de clasificación se precede por el aprendizaje de un modelo computacional que usa información del conjunto de entrenamiento, almacenando los valores de las variables junto con la clase a la que pertenecen (Sierra Araujo et al., 2006).

Una parte clave del éxito del clasificador depende del conjunto de entrenamiento, por lo cual es fundamental tener un conjunto representativo de documentos adecuadamente etiquetados. Generalmente, la etiquetación se realiza de forma manual, así es que, un experto en los temas debe ser quien lea los documentos y determine su clase. Por ello, la tarea de etiquetación es costosa, principalmente en tiempo, debido a la cantidad de documentos a etiquetar. Derivado de lo anterior, el clasificador tiene una alta dependencia al idioma del conjunto de entrenamiento. Esa dependencia al idioma debe flexibilizarse ante el actual entorno multilingüe.

1.2 Entorno Multilingüe

Durante años las investigaciones se concentraron en crear recursos para el idioma inglés ya que posee el icono de ser la lengua internacional. El uso del inglés fue impulsado en un acuerdo colectivo implícito de las comunidades políticas, científicas y académicas, tanto en los principales medios de comunicación (revistas, editoriales, libros, Internet) como en reuniones internacionales, en un esfuerzo por difundir el conocimiento y facilitar la colaboración internacional.

Como resultado del desbalanceado desarrollo tecnológico mundial, que está estrechamente relacionado a la organización, infraestructura y financiamiento de cada país, gente de todo el mundo trabaja con idiomas internacionales como inglés o español adicionales a sus lenguajes nativos. No obstante, en el mundo actual coexisten alrededor de 6000 lenguas (portal UNESCO, 2009) y la incursión paulatina de las mismas en medios masivos de comunicación responde a la expansión mundial de la tecnología. Por ejemplo, la Figura 1.1 muestra la presencia de idiomas en internet conforme se incrementa el número de usuarios de diversas nacionalidades.

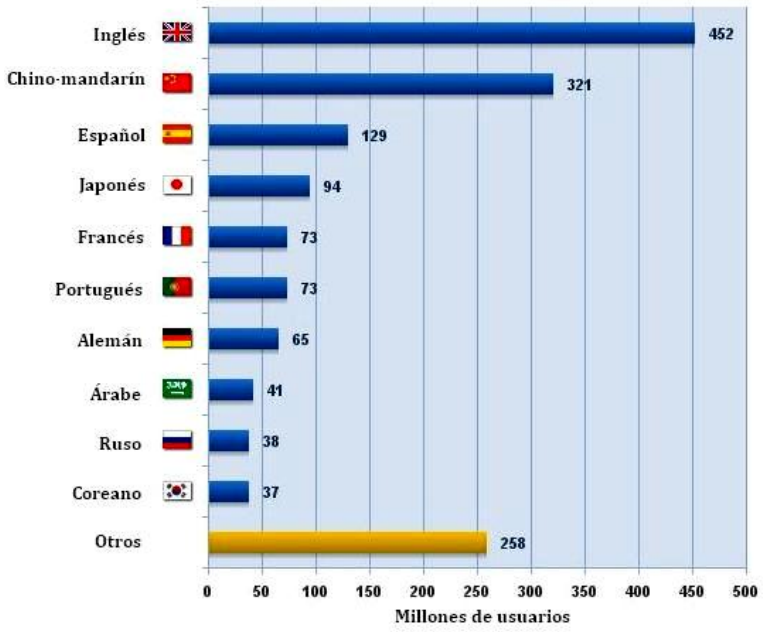


Figura 1.1 Idiomas más usados en internet por millones de usuarios (Nielsen Net Ratings, 2009)

Ante la creciente incursión de más lenguajes en los medios de comunicación, cada día tenemos acceso a documentos publicados en más idiomas, creando con ello grandes volúmenes de información multilingüe. Por ende, los intereses se han concentrado en aprovechar dicha información tratando de franquear la barrera del idioma.

El multilingüismo ha desencadenado investigaciones recientes con el fin flexibilizar los mecanismos desarrollados en procesamiento de texto para permitir la entrada de varios idiomas, esto porque los recursos existentes deben expandirse para satisfacer la demanda mundial. Por ejemplo, muchas compañías e instituciones necesitan buscar y organizar eficientemente repositorios multilingües de documentos. El manejo de estas colecciones de texto heterogéneas incrementa significativamente el costo ya que se requieren expertos de diferentes idiomas que organicen las colecciones.

Entre las soluciones planteadas, se crearon herramientas multilingües automáticas como traductores, diccionarios y ontologías multilingües, que han ido evolucionando hasta convertirse hoy en día en herramientas más confiables. Esto propició la expansión de líneas de investigación como la recuperación de información, búsqueda de respuestas, clustering y clasificación de textos (Mitkov, 2003).

En particular, la mayoría de las aplicaciones de Clasificación de Textos, como librerías digitales, filtrado de noticias, clasificación de páginas web y correos electrónicos, también resultan aplicaciones interesantes del la Clasificación Multilingüe de Textos, donde documentos dados en diferentes idiomas deben clasificarse por tema o criterios similares (de Melo y Siersdorfer, 2007).

1.3 Clasificación Translingüe

Dentro de la clasificación de textos, el entorno multilingüe representa un nuevo reto por la dependencia de la tarea a los conjuntos etiquetados de documentos. Puesto que la creación de corpus de entrenamiento es un trabajo manual efectuado por un experto, el manejo de corpus multilingües

se convierte en un problema debido a que el proceso de etiquetado se multiplica para cada idioma que se desee clasificar. Implícitamente, para construir corpus en varios idiomas es necesario que el experto domine dichos idiomas, o en su defecto, conseguir a un experto por cada idioma a etiquetar. Ante la dificultad de contar con corpus etiquetados para cada idioma, la clasificación translingüe representa una solución prometedora.

La clasificación translingüe se define como la tarea de asignar clases a documentos escritos en un idioma objetivo usando recursos de un idioma fuente (Gliozzo y Strapparava, 2006). Se parte de suponer que existe un buen clasificador para un grupo de categorías en el idioma L_1 y documentos no etiquetados en un idioma diferente L_2. Así surge el cuestionamiento ¿Cómo se pueden clasificar los documentos en las mismas categorías definidas para el idioma L_1 sin realizar un etiquetado manual en L_2?. Usando el paradigma del Aprendizaje Computacional, este problema puede reformularse, ¿Cómo puede entrenarse un clasificador para el idioma L_2 utilizando ejemplos etiquetados para el idioma L_1?. Un algoritmo que sea capaz de desempeñar efectivamente esta tarea reduciría los costos de crear sistemas multilingües ya que el esfuerzo humano se reduciría a proveer el conjunto de entrenamiento en solo un idioma (Rigutini et al. 2005).

1.4 Descripción del Problema

La clasificación translingüe se ha abordado utilizando distintos métodos, uno de los que obtiene mejores resultados es la incorporación de traductores automáticos al proceso. El uso de traductores automáticos tiene la ventaja de que el número de idiomas en los que existen traductores se incrementa y por lo tanto pueden trasladarse recursos existentes de un idioma a cuantos sea necesario, considerando que no es tan limitante en

cuanto a dominio como otras herramientas. El objetivo de la traducción dentro de la clasificación translingüe no es producir un texto traducido con propiedades semánticas fieles al original, sino proveer un texto que asegure una calidad suficiente para la clasificación (Jalam, 2003). Por ello, al considerar la clasificación translingüe utilizando traducción automática, se presentan dos problemas importantes:

Las distorsiones de la traducción: La traducción es suficiente para algunos términos temáticos, pero a pesar de los avances, existen errores que impactarán los resultados de la clasificación. Deben considerarse factores como la traducción con sinónimos, los errores de traducción e incluso la falta de traducción.

La discrepancia cultural: A pesar de suponer una traducción automática perfecta, la traducción no captura toda la información propia del dominio en el lenguaje objetivo. Por ejemplo, si se deseara hacer una clasificación de notas periodísticas donde una clase sea deportes, es muy probable que los deportes en un idioma varíen considerablemente respecto a otro. Recordemos que el idioma es el medio de expresión de un grupo cultural y socialmente homogéneo. De esta manera, noticias sobre deportes de Estados Unidos probablemente mencionen fútbol americano y béisbol mientras que notas de un periódico español puedan contener información sobre pelota vasca y las reseñas francesas sobre rugby.

La hipótesis de este trabajo consiste en que los problemas antes mencionados pueden abordarse utilizando información que se encuentra en el propio conjunto escrito en el idioma objetivo. Utilizando dicha información se puede crear un proceso mediante el cual se encuentren las similitudes existentes entre los documentos no etiquetados, considerando rasgos propios del dominio.

Se plantean dos métodos, en el primero se utilizará un proceso de refinamiento, independiente del proceso de clasificación, mediante el cual se confirma o corrige la clasificación por los documentos no etiquetados. El resultado debe ser una clasificación adaptada aprovechando las características propias de los documentos del idioma objetivo. En el segundo, se realizará un proceso de selección de ejemplos confiables, dichos ejemplos paulatinamente ajustarán el clasificador a los rasgos del conjunto de documentos no etiquetados.

1.5 Objetivo de la Tesis

Objetivo General:

Dado el contexto de la clasificación translingüe basada en traducción automática, con el fin de subsanar los problemas de distorsión cultural aunados a los pormenores típicos de la traducción:

- Definir métodos de refinamiento de la clasificación translingüe que utilicen información extraída del idioma objetivo y sean capaces de adecuar la clasificación con características propias de dicho idioma.

Objetivos Particulares:

- Analizar el impacto de la traducción y discrepancia cultural en la clasificación translingüe.
- Proponer un método de refinamiento de la clasificación usando el conjunto no etiquetado mediante el uso de vecinos más cercanos.

- Proponer un método de ajuste del clasificador usando incorporación de ejemplos confiables provenientes del conjunto no etiquetado.
- Evaluar los métodos propuestos con los dos esquemas de clasificación translingüe existentes: traduciendo el conjunto de entrenamiento y traduciendo el conjunto objetivo.

1.6 Estructura de la Tesis

El presente documento está organizado de la siguiente manera:

En el capítulo 2 se introducen conceptos básicos sobre la clasificación de textos, incluyendo la forma de representación de los documentos, esquemas de pesado, los métodos de clasificación relacionados y las medidas de evaluación para determinar el éxito de la clasificación. Estos conceptos serán utilizados a lo largo del documento.

El capítulo 3 da una descripción de la problemática multilingüe dentro de la clasificación de textos. Cita el trabajo previo desarrollado en el área, incluyendo las primeras propuestas para solucionar el entorno multilingüe dentro de la clasificación de textos y algunas de las herramientas que se han utilizado para facilitar la tarea.

El capítulo 4 hace un análisis de los problemas específicos de la clasificación translingüe con el objetivo de explicar los puntos débiles que es posible corregir mediante los métodos propuestos. Se presentan los dos esquemas conocidos en la clasificación translingüe y se define el corpus de trabajo mostrando resultados iniciales de las mismas.

En el capítulo 5 se definen los métodos propuestos de refinamiento de la clasificación translingüe discutiendo sus ventajas y desventajas. Se presentan los resultados obtenidos incluyendo una comparación entre los métodos considerando ambos esquemas.

El capítulo 6 resume las conclusiones del trabajo de tesis y presenta ideas para trabajo futuro.

Capítulo 2

Marco Teórico

En este capítulo se introducen los conceptos básicos relacionados con el presente trabajo de investigación. Ya que el objetivo es definir un método de clasificación basado en aprendizaje computacional, se inicia con las definiciones de aprendizaje computacional y clasificación. Se enlistan los procesos previos necesarios para realizar la tarea de clasificación automática de textos; posteriormente se mencionan algunos de los métodos de clasificación más relacionados al trabajo y finalmente las medidas de evaluación comúnmente usadas para determinar el éxito de la clasificación.

2.1 Introducción

El aprendizaje computacional es el estudio de algoritmos que mejoran su desempeño en la realización de una tarea a través de la experiencia. Formalmente, se dice que “un programa de computadora aprende de la experiencia *E* con respecto a alguna clase de tarea *T* y desempeño medido *P*, si su desempeño en dicha tarea *T* medido por *P* mejora con la experiencia *E*” (Mitchell, 1997). Así, el objetivo del aprendizaje computacional es desarrollar modelos capaces de aprender a partir de conjuntos de datos

extrayendo información implícita para tomar decisiones y hacer predicciones sobre nuevos datos (Aurajo, 2006). Entre las tareas que se pueden realizar con aprendizaje computacional, se encuentra la de clasificación.

La clasificación consiste en asociar un valor booleano a cada par $(o_j, c_i) \in O \times C$, donde $O = \{o_1, \ldots, o_n\}$ es el conjunto de objetos y $C = \{c_1, \ldots, c_{|c|}\}$ es el conjunto predefinido de clases. El valor V (Verdadero) es asociado a un par (o_j, c_i) si el objeto o_j pertenece a la clase c_i, mientras que el valor F (Falso) es asociado en caso contrario. Se aproxima una función $\breve{\Phi}: O \times C \rightarrow \{V, F\}$ que asocia una o más clases a un objeto o_j de tal forma que la decisión tomada es lo más cercana posible a la que correspondería en la función $\Phi: O \times C \rightarrow \{V, F\}$ llamada clasificador.

Las técnicas de aprendizaje computacional utilizan un conjunto de ejemplos *O* etiquetados, donde el valor de $\Phi: (o_j, c_i)$ es conocido para cada $(o_j, c_i) \in O \times C$ (Sebastiani, 2002). Cada objeto o_j debe ser representado por un conjunto de atributos $T = \{t_1, \ldots, t_m\}$, los cuales son utilizados para discriminar entre clases. Por ello, para poder realizar la clasificación por métodos de aprendizaje computacional se deben realizar una serie de tareas como la selección de los atributos, el pesado de los mismos, la construcción del clasificador, entre otras, antes de finalmente evaluar al clasificador.

El éxito de la clasificación está limitado a tres 'cuellos de botella' potenciales que son la calidad del conjunto de entrenamiento, la capacidad del clasificador para discriminar las clases y la riqueza de representación de los objetos a clasificar (Ifrim et al., 2005). Con respecto a la calidad del conjunto de entrenamiento, generalmente no se tiene control sobre él debido a que es un trabajo costoso de realizar, considerando que el etiquetado suele realizarse de forma manual.

A continuación, se abordan con mayor detalle las tareas relacionadas con la clasificación automática mediante técnicas de aprendizaje computacional, para el caso específico de clasificación de textos.

2.2 Clasificación de Textos

En particular, en la clasificación de textos el conjunto de objetos a clasificar $O = \{o_1, \dots, o_n\}$ es un conjunto de documentos D$= \{d_1, \dots, d_n\}$. El conjunto de documentos D_E, al cual le fue asignado cada valor de $\Phi: (d_j, c_i)$, se conoce como corpus de entrenamiento; usualmente el etiquetado se realiza manualmente por expertos que deben leer los documentos antes de asignar una clase.

El conjunto de prueba D_P contiene datos que no han sido utilizados durante el entrenamiento, es decir, $D_E \cap D_P = \emptyset$. El objetivo de poseer conjuntos de prueba es saber que tan bien trabaja un modelo particular con ejemplos nuevos. Por ello los datos de prueba y de entrenamiento deben ser diferentes para que la prueba sea válida.

Una vez que se cuenta con los dos conjuntos de documentos etiquetados que servirán como entrenamiento y prueba, los documentos deben ser preparados a fin de ser utilizados por el clasificador, para ello es necesario aplicar ciertos pasos previos. A continuación se presentan algunos procesos relacionados a la representación de los documentos.

2.2.1 Pre-procesamiento

Para realizar la clasificación se deben llevar los documentos a una representación con la que pueda trabajar el algoritmo de aprendizaje computacional. Para ello se siguen algunas etapas de pre-procesamiento

que están relacionadas al tipo de clasificación que se desea realizar (temática, no temática):

Limpieza de documentos: Consiste en remover todo aquello que pueda representar ruido, por ejemplo: etiquetas html o similares, encabezados, separadores, tablas, números, caracteres extraños introducidos por el OCR (Optical Character Recognition), etc.

Eliminar palabras vacías: Se les llama palabras vacías a los artículos, pronombres, preposiciones, etc. que aparecen con alta frecuencia en los documentos pero tienen poco impacto para determinar una clase.

Aplicar lematización: La lematización es el proceso mediante el cual se identifica el lema o raíz de la palabra, eliminando declinaciones por conjugación, número o género (caminar, caminamos; doctor, doctora; *walk, walks, walking*). Aunque agrupar varias palabras puede ser benéfico, también es posible que ocurran efectos como la sobre-lematización (*overstemming*), con la cual se confunden palabras que no están relacionadas (torá y torear son llevadas a tor, así como *magnesia, magnesian, magnet, magnetic* son llevadas a *magnes*). Por ejemplo, la palabra *business* será transformada por el lematizador a *busy*, provocando pérdida de información. La lematización está muy relacionada con el idioma; el inglés tiene poca morfología pero otros idiomas son más ricos en inflexiones y derivaciones, y por ello, es necesario un análisis morfológico para realizar el proceso (Manning y Schütze, 2001; Mitkov, 2003).

2.2.2 Representación de Documentos

Entre los modelos de representación basados en análisis estadístico, el más utilizado es el modelo vectorial. Cada documento $d_j \in D$ es representado por un vector m-dimensional de términos indexados o palabras

clave (Sebastiani, 2002). Sea $T = \{t_1, \dots, t_m\}$ el conjunto finito de términos llamado vocabulario, los cuales son extraídos del conjunto de documentos *D*, entonces los documentos quedan representados por un vector $d_j = (w_{1j}, w_{2j} \dots, w_{mj})$ donde cada término indexado corresponde a una palabra en el texto y tiene un peso asociado w que refleja la importancia del término para el documento y/o para la colección completa de documentos.

Existen diferentes métodos de calcular los pesos. Estos métodos se basan en dos observaciones, mientras el término t_k es más frecuente en un documento d_j, más relacionado está con el tema de dicho documento y mientras el término t_k es más frecuente en una colección, menos será utilizado como discriminante entre documentos (véase Anexo A: Ley de Zipf)

2.2.3 Esquemas de Pesado

Sea $\#(t_k, d_j)$ el número de ocurrencias de un término t_k en el documento d_j, $|D|$ el número de documentos del conjunto y $\#D(t_k)$ el número de documentos en los que aparece el término t_k.

Pesado Booleano: Asigna un 1 si aparece la palabra en el documento, y un 0 si no aparece.

$$w_{kj} = \begin{cases} 1 & \text{Si } \#(t_k, d_j) \geq 1 \\ 0 & \text{en caso contrario} \end{cases} \quad \textbf{(2.1)}$$

Frecuencia del Término (TF): Asigna el número de veces que aparece la palabra en el texto y 0 si no aparece.

$$w_{kj} = \#(t_k, d_j) \quad \textbf{(2.2)}$$

TFxIDF: *Term Frequency x Inverse Document Frequency.* Establece una relación entre la frecuencia del término en el documento y la frecuencia de éste en el resto de los documentos de la colección. Se obtiene por la ecuación (2.3).

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|D|}{\#D(t_k)} \quad \textbf{(2.3)}$$

Existen otros esquemas de pesado que consideran factores adicionales, por ejemplo, el tamaño de los documentos, la existencia de clases desbalanceadas con pocos ejemplos de alguna de las clases, entre otros.

2.2.4 Reducción de Dimensionalidad

Un problema central de la clasificación de textos es la alta dimensionalidad de las matrices de representación de atributos, ya que existe una columna por cada palabra encontrada en la colección y el vocabulario puede abarcar desde unos cuantos cientos hasta miles de palabras. Dado que el manejo de matrices grandes es costoso computacionalmente, la selección de atributos tiene como objetivo remover palabras no informativas con el fin de reducir la complejidad computacional y mejorar la eficacia del clasificador (Aas y Eikvil, 1999).

A continuación se presentan algunas de las técnicas de reducción de dimensionalidad más usadas en la clasificación automática de documentos.

Umbral de Frecuencia:

La reducción por umbral de frecuencia es una de las más sencillas, si el número de documentos en los cuales ocurre un término es menor a cierto umbral predefinido, dicho término es removido. Este criterio se basa en la

suposición de que las palabras que rara vez ocurren en una colección no son informativas para la predicción de la clase y por lo tanto no tienen influencia en el desempeño global del clasificador (Aas y Eikvil, 1999) (véase Anexo A: ley de Zipf).

Ganancia de información:

La Ganancia de Información (IG por sus siglas en inglés *Information Gain*) mide la entropía del sistema con la presencia o ausencia de cada palabra en un documento, es decir, mide si el grado de desorden del sistema se incrementa o reduce al conocer el valor de un atributo determinado. Con ello, asigna un 'valor de información' a cada atributo, calculándose como:

$$IG(t_k) = \sum_{i=1}^{|C|} P(c_i) \log P(c_i) + P(t_k) \sum_{i=1}^{|C|} P(c_i|t_k)\, logP(c_i|t_k) + P(\bar{t_k}) \sum_{i=1}^{|C|} P(c_i|\bar{t_k})\, logP(c_i|\bar{t_k}) \quad \textbf{(2.4)}$$

Donde $P(c_i)$ es la probabilidad de la clase c_i y se estima con la cantidad de documentos de la colección total que pertenecen a la clase c_i; $P(t_k)$ es la probabilidad de seleccionar un documento que contenga el término t_k y se estima de la porción de documentos en los cuales ocurre el término; $P(c_i|t_k)$ es la probabilidad condicional de que un documento pertenezca a la clase c_i dado que el documento contiene al término t_k, se obtiene con los documentos de la clase c_i en los que el término ocurre al menos una vez; de la misma forma, $P(c_i|\bar{t_k})$ es la probabilidad condicional de que un documento pertenezca a la clase c_i dado que el documento no contiene el término t_k.

La ganancia de información se calcula para cada palabra del conjunto de entrenamiento y aquellas cuya ganancia es menor a determinado umbral son eliminadas. Típicamente se conservan aquellas palabras cuya IG es positiva.

2.2.5 Medidas de Similitud

Algunos métodos de clasificación requieren determinar que tan semejante es un documento con otro. Las medidas de similitud son capaces de expresar una cantidad numérica entre 0 y 1. Sean d_i y d_j documentos representados de la forma $d_j = (w_{1j}, w_{2j} \ \ldots, w_{mj})$, entonces algunas medidas de similitud o distancia se definen de la siguiente forma:

Manhatan:

La medida Manhatan es la llamada medida de bloques o calles, ya que se suman las distancias horizontales y verticales desde el punto de inicio hasta el punto final (Russell, S. y Norvig, P; 2004).

$$manhatan(d_i, d_j) = \sum |w_{ki} - w_{kj}| \qquad \textbf{(2.5)}$$

Dice:

El coeficiente de Dice determina la similitud entre dos documentos pesados dando importancia a los atributos de la intersección (Salton, G. y McGill, M. J; 1983).

$$dice(d_i, d_j) = \frac{2\sum(w_{ki} \times w_{kj})}{\sum w_{ki} + \sum w_{kj}} \qquad \textbf{(2.6)}$$

Coseno:

La medida cosenoidal es una de las más populares para determinar la similitud de los documentos. El objetivo es determinar el ángulo entre dos vectores, en este caso los vectores de representación de los documentos. Se obtiene aplicando la ecuación (2.7) (Frakes, W. B. y Baeza-Yates, R.; 1992).

$$coseno(d_i, d_j) = \frac{\sum w_{ki} \times w_{kj}}{\sqrt{\sum {w_{ki}}^2} \times \sqrt{\sum {w_{kj}}^2}} \quad \textbf{(2.7)}$$

2.3 Métodos de Clasificación

Dentro del aprendizaje computacional existen múltiples métodos de clasificación. Sin embargo, para el manejo de texto los que han obtenido mejores resultados son Naive Bayes, Máquinas de Vectores de Soporte, Vecinos más Cercanos y Rocchio. En esta sección se describe el funcionamiento de cada uno.

2.3.1 Clasificador Naive Bayes

Este clasificador aplica el teorema de Bayes para calcular la probabilidad de un documento de pertenecer a una clase:

$$P(c_i|d_j) = \frac{P(c_i)P(d_j|c_i)}{P(d_j)} \quad \textbf{(2.8)}$$

Donde $c_i \in C$ representa cada clase en un conjunto finito de clases, $d_j = (w_{1j}, w_{2j} \ldots, w_{mj})$ es el vector de pesos que representa a un documento. A un ejemplo de prueba se le asigna la clase c_i más probable dados los valores de sus atributos:

$$c_{ij} = argmax_{c_i \in C}\left(P\left(c_i \middle| w_{1j}, w_{2j}\ \ldots, w_{mj}\right)\right) \quad \textbf{(2.9)}$$

usando Bayes

$$c_{ij} = argmax_{c_i \in C}\left(\frac{P(w_{1j}, w_{2j}\ \ldots, w_{mj}|c_i)P(c_i)}{P(w_{1j}, w_{2j}\ \ldots, w_{mj})}\right) \quad \textbf{(2.10)}$$

$$c_{ij} = argmax_{c_i \in C}\left(P(w_{1j}, w_{2j}\ \ldots, w_{mj}|c_i)P(c_i)\right) \quad \textbf{(2.11)}$$

$P(c_i)$ se puede estimar con la ecuación (2.12) donde $\#D(c_i)$ es el número de documentos que pertenecen a la clase c_i y $|D|$ es el número total de documentos que posee la colección.

$$P(c_i) = \frac{\#D(c_i)}{|D|} \quad \textbf{(2.12)}$$

Para calcular $P(w_{1j}, w_{2j}\ \ldots, w_{mj}|c_i)$, Naive Bayes está basado en la suposición de independencia de palabras, es decir que los atributos son condicionalmente independientes, dada una clase. Por ello puede expresarse como el producto de probabilidades de cada término que aparece en el documento:

$$P\left(w_{1j}, w_{2j}\ \ldots, w_{mj} \middle| c_i\right) = \prod_k P(w_{kj}|c_i) \quad \textbf{(2.13)}$$

Los valores $P(w_{kj}|c_i)$ se estiman con la frecuencia de los datos observados mediante la ecuación (2.14).

$$P(w_{kj}|c_i) = \frac{1 + \#(t_k, d_j(c_i))}{\#D(c_i) + m} \quad \textbf{(2.14)}$$

Donde $\#(t_k, d_j(c_i))$ es el número de veces que aparece el término t_k aparece en documentos de la clase c_i, y m es el total de términos de la colección.

El clasificador Naive Bayes asigna a cada documento la clase cuya probabilidad sea mayor según la ecuación (2.15)

$$c_{ij} = argmax_{c_i \in C}\left(P(c_i)\prod_k P(w_{kj}\,|c_i)\right) \qquad \textbf{(2.15)}$$

A pesar de que la suposición de independencia generalmente no es cierta en la aparición de palabras en los documentos, el Naive Bayes es un clasificador efectivo para la tarea de clasificación de textos (Mitkov, 2003).

2.3.2 Clasificador Vecinos más Cercanos (KNN)

Sea D_E el conjunto de documentos de entrenamiento, donde el valor de la función $\Phi: (d_j, c_i)$ es conocido para cada $(d_j, c_i) \in D \times C$ y donde cada documento d_j es descrito por el vector $d_j = (w_{1j}, w_{2j} \ldots, w_{mj})$. Entonces existe una función de distancia que permite comparar los documentos (sección 2.2.5). Dado un nuevo documento *d*, el algoritmo del vecino más cercano (KNN *K-nearest neighbour*) busca entre los objetos D_E los *k* documentos más cercanos $d_{nn1}, \ldots, d_{nnk}$ a *d* de acuerdo a la función de distancia como se observa en la Figura 2.1. En un esquema de voto simple, se asigna al documento *d* la clase más frecuente entre las clases de sus vecinos. En el esquema de voto ponderado, la posición del vecino otorga un grado de importancia al voto.

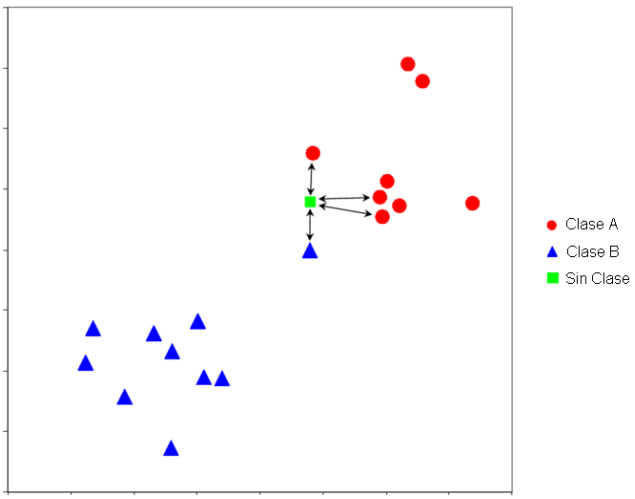


Figura 2.1 Clasificación por vecinos más cercanos (Izhikevich, 2009)

Debido a que la clasificación por vecinos más cercanos es un método que carece de etapas de entrenamiento, se le llama de 'aprendizaje flojo'. La búsqueda de los vecinos más cercanos puede ser costosa computacionalmente si existe un gran número de documentos en el conjunto de entrenamiento, lo que provoca que la clasificación resulte lenta a comparación de otros métodos como demuestra experimentalmente García Adeva (2005).

2.3.3 Clasificador por Prototipos

El clasificador Rocchio es uno de los varios tipos de clasificadores que trabajan con prototipos. La idea general es construir un prototipo a_i de cada clase $c_i \in C$, de forma que, cuando un ejemplo nuevo *d* debe ser clasificado, solamente se compare con los prototipos y se asigne la clase de aquel que sea más similar. No es necesario que el ejemplo prototipo exista, en algunos casos se calcula mediante promedio, suma normalizada o alguna otra medida de sus atributos.

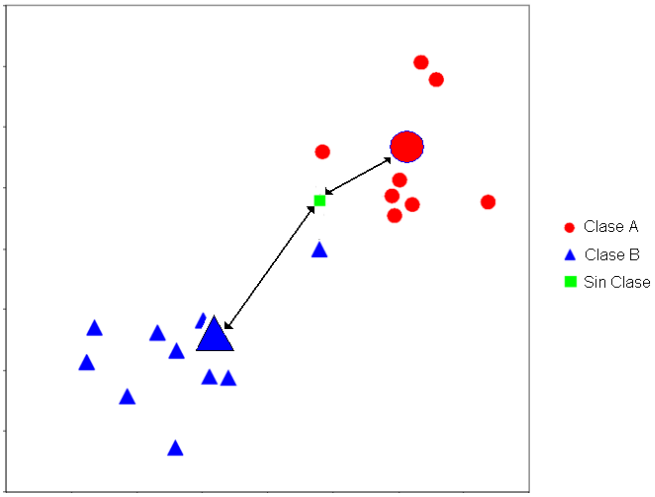


Figura 2.2 Clasificación por prototipos

Específicamente en el clasificador Rocchio, para obtener el vector prototipo, se hace una suma ponderada de los documentos etiquetados con la clase c_i en el conjunto de entrenamiento menos la suma ponderada de los no pertenecientes a la clase c_i :

$$a_i = \beta \sum_{d_j \in c_i} d_j - \gamma \sum_{d_j \notin c_i} d_j \qquad \textbf{(2.16)}$$

Donde β y γ son parámetros para dar peso a la suma. Para más detalles sobre la clasificación basada en prototipos ver Joachims (1997).

2.3.4 Clasificador Máquinas de Vectores de Soporte (SVM)

Mientras la mayoría de los algoritmos de aprendizaje se centran en reducir los errores cometidos por el modelo generado, SVM (Support Vector Machines) no busca reducir el riesgo cometiendo pocos errores, sino que pretende construir modelos confiables (Sierra Araujo et al., 2006). SVM es considerado el primer método kernel porque, para problemas que no son

linealmente separables, SVM usa *funciones de convolución* o *Kernels*. Aplicar kernels consiste en hacer una transformación del problema original mediante una representación de información distinta llevándolo a un espacio de alta dimensionalidad, donde los documentos transformados son linealmente separables (Figura 2.3a). Ya que pueden existir varios hiperplanos capaces de separar las clases (Figura 2.3b), se elige el que está más distanciado de ambas clases porque presenta menos riesgo frente a posibles ruidos en los datos (Figura 2.3c).

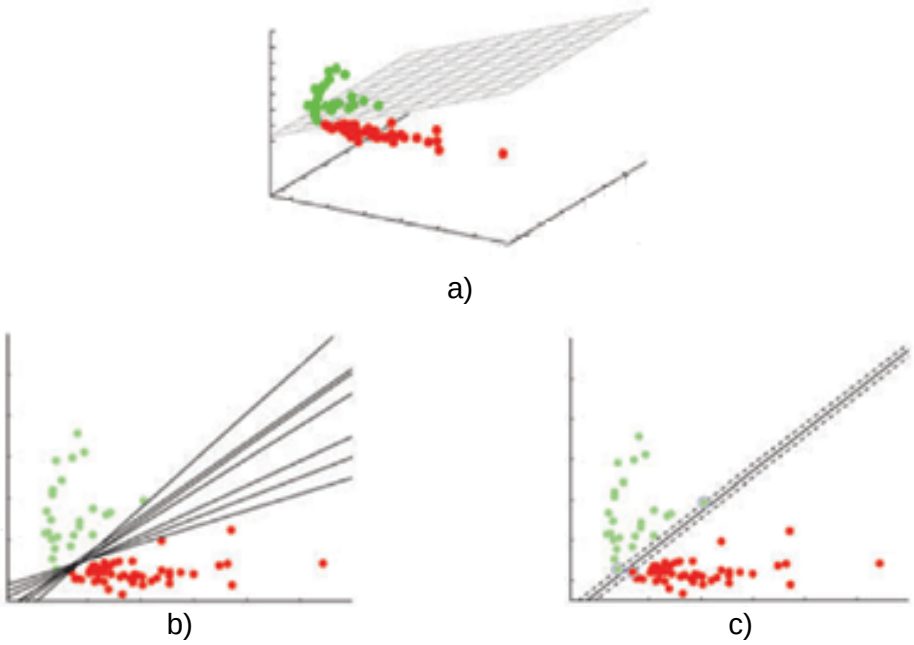


Figura 2.3 Clasificación por máquinas de vectores de soporte (Noble, 2007)

La Figura 2.4 ilustra de forma general la selección del hiperplano. Se muestran dos hiperplanos capaces de separar linealmente las clases. Se visualizan los márgenes de riesgo de error de cada hiperplano. Aquellos ejemplos que representan los limites de decisión para el hiperplano se conocen como ‘vectores de soporte’. El método elegirá el hiperplano con mayor margen.

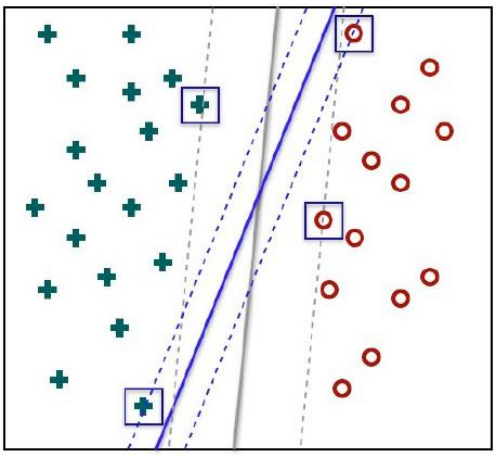


Figura 2.4 Elección del hiperplano por vectores de soporte (Alvarez, 2009)

Este método es aplicable únicamente a problemas de clasificación binaria, en problemas no binarios se utiliza en cascada, donde, de contarse con n clases es necesario construir n-1 clasificadores. Para más detalles ver Hearst et al. (1998).

2.4 Medidas de Evaluación

La evaluación de los clasificadores de textos se conduce de forma experimental usualmente midiendo la exactitud del clasificador, es decir, la capacidad de tomar decisiones correctas.

Un clasificador genera una tabla de confusión como se muestra en la Tabla 2.1, donde TP_i indica el número de *verdaderos positivos*, es decir, cuantos documentos fueron correctamente clasificados bajo la clase c_j. De forma similar, FP_i indica el número de *falsos positivos*, aquellos que fueron clasificados erróneamente como positivos. FN_i corresponde al número de *falsos negativos* y TN_i corresponde al número de *verdaderos negativos*. Con

base en estos valores, se pueden calcular Exactitud (e), Precisión (π) y Recuerdo (ρ), medidas que permiten saber que tan exitoso es el clasificador.

Tabla 2.1 Matriz de confusión de la clase c_i.

Clase c_i		**Decisión del Experto**	
		Si	No
Decisión del Clasificador	Si	TP_i	FP_i
	No	FN_i	TN_i

Exactitud

La exactitud es una medida global ya que se refiere a la capacidad del clasificador para acertar en la clasificación, considerando correctamente clasificados tanto a los ejemplos positivos como negativos. En breve, la exactitud es un valor entre 0 y 1 que representa el porcentaje de documentos correctamente clasificados y se obtiene mediante la sumatoria (2. 17)

$$\hat{A} = \sum_{i=1}^{|C|} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \qquad \textbf{(2.17)}$$

Un inconveniente de esta medida de evaluación es que, si las clases están desbalanceadas y existe tendencia del clasificador a predecir la clase mayoritaria, entonces no reflejará la calidad real del clasificador ya que el asignar la clase más frecuente reflejará buenos resultados con esta medida. (Sebastiani, 2002).

Precisión

La precisión indica la probabilidad de que el documento asignado a cierta clase por el clasificador, efectivamente pertenezca a esa clase. De la matriz de confusión, la precisión se obtiene por la ecuación 2.18.

$$\pi = \frac{TP_i}{TP_i + FP_i} \quad \textbf{(2.18)}$$

Recuerdo

El recuerdo estima la probabilidad de que un documento que pertenezca a cierta clase, sea correctamente asignado durante el proceso de clasificación. De la matriz de confusión, el recuerdo se obtiene por:

$$\rho = \frac{TP_i}{TP_i + FN_i} \quad \textbf{(2.19)}$$

F-measure (F_β)

Comúnmente utilizada para englobar la precisión (π) ec. 2.18 y el Recuerdo (ρ) ec. 2.19, se obtiene con:

$$\boldsymbol{F_\beta} = \frac{(1 + \beta^2)\pi\rho}{\beta^2\pi + \rho} \quad \textbf{(2.20)}$$

Donde β es el parámetro que controla la importancia relativa entre las dos medidas. Usualmente se fija el valor $\beta = 1$ para dar igual importancia a ambos valores.

Las definiciones dadas por (2.18), (2.19) y (2.20) son medidas de evaluación por categoría, para obtener datos globales se calculan los promedios. Existen dos formas de realizar el cálculo global, el micropromedio y el macropromedio.

Micropromedio

El micropromedio calcula los parámetros TP_i , FP_i , FN_i, y TN_i para todas las categorías obteniendo una medida global única, con la cual otorga el mismo peso a cada documento dando a las categorías una importancia proporcional al número de ejemplos positivos que le corresponden.

Macropromedio

Es el promedio obteniendo las medidas de evaluación por categoría y posteriormente obtiene un promedio global. El macropromedio da el mismo peso a cada categoría. La diferencia entre micro y macro promedio debe considerarse si las clases son desbalanceadas. La Tabla 2.2 muestra cómo se obtienen de la matriz de confusión.

Tabla 2.2 Micropromedio y macropromedio a partir de la matriz de confusión

	Micropromedio	**Macropromedio**
Precisión (π)	$\pi = \frac{\sum TP_i}{\sum (TP_i + FP_i)}$	$\pi = \frac{\sum \left(\frac{TP_i}{TP_i + FP_i}\right)}{\lvert C \rvert}$
Recuerdo(ρ)	$\rho = \frac{\sum TP_i}{\sum (TP_i + FN_i)}$	$\rho = \frac{\sum \left(\frac{TP_i}{TP_i + FN_i}\right)}{\lvert C \rvert}$

Capítulo 3

Estado del Arte

El multilingüismo ha sido recientemente abordado en la tarea de clasificación de textos. Hasta ahora, el problema se ha enfrentado construyendo clasificadores monolingües, polilingües y translingües. Dentro del área, el término clasificación translingüe (*cross-language text classification*) se acuñó en el 2003 siendo el enfoque más reciente. En esta sección se explica en qué consiste cada método y se discuten las ventajas y desventajas de los trabajos realizados.

3.1 Clasificación Monolingüe por Idioma

La clasificación monolingüe por idioma es un caso idealizado en el que existen suficientes documentos etiquetados en cada idioma y en cada clase para construir los clasificadores. Cada grupo de documentos en un idioma debe ser enviado a su respectivo clasificador, por lo que se deben construir varios clasificadores, uno por cada idioma que se desee clasificar. La Figura 3.1 ilustra el proceso de clasificación monolingüe para el idioma 1. En ella se aprecia que el conjunto D_p de documentos a clasificar puede contener documentos escritos en diferentes idiomas por lo que se incluye un módulo

de identificación de idioma, que filtra aquellos para los cuales existe un clasificador. En este caso, sólo el subconjunto perteneciente al idioma del clasificador es aceptado. El resultado es un grupo de documentos clasificador por cada idioma.

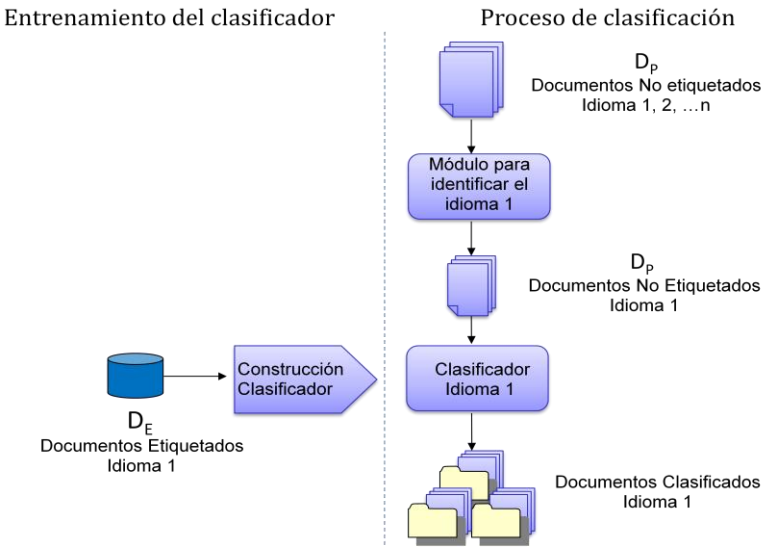


Figura 3.1 Esquema general de la clasificación monolingüe

La construcción de clasificadores monolingües para cada idioma es propuesta por Jalam (2003) como una solución trivial al problema de clasificación multilingüe. García Adeva (2005) también propone en el enfoque al que llama NPNC (n pre-procesamientos n clasificadores) un pre-procesamiento específico por idioma y la construcción de clasificadores monolingües, determinando experimentalmente que esta solución es la que arroja mayor exactitud. Bel (2003), Rigutini (2005), Chih-Ping (2007), entre otros, hacen la clasificación monolingüe para establecer un punto de comparación con las soluciones propuestas en sus trabajos.

3.2 Clasificación por Entrenamiento Polilingüe

Otra solución abordada para la clasificación de textos multilingües, consiste en crear un único clasificador para todos los idiomas. La clasificación por entrenamiento polilingüe se define como aquella en la cual un clasificador es entrenado con documentos etiquetados escritos en diferentes idiomas; considerando incluso la posibilidad de contar con varios idiomas en un mismo documento. El objetivo es que el clasificador reciba los documentos sin especificar de qué idioma se trata y sea capaz de clasificarlos (Bel et al., 2003). La Figura 3.2 es una muestra de la clasificación por entrenamiento polilingüe clásica.

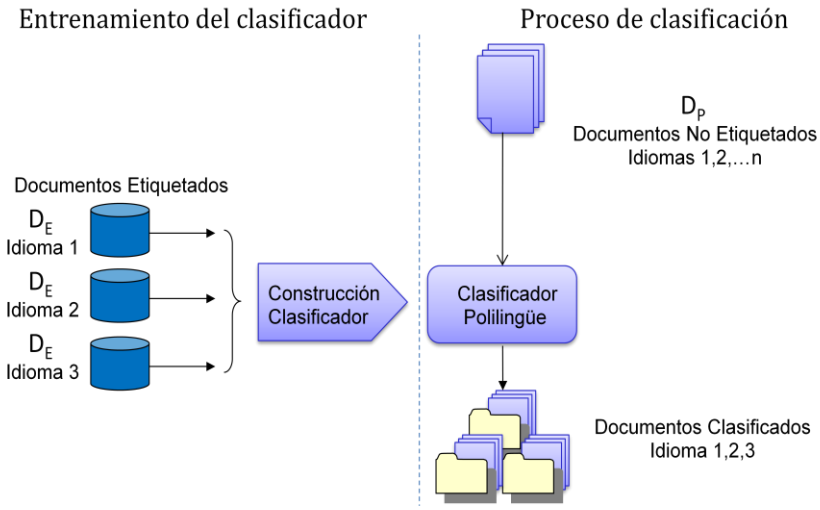


Figura 3.2 Esquema general de la clasificación polilingüe

Se observa que el conjunto D_p de documentos a clasificar puede contener varios idiomas y es recibido por un único clasificador. El resultado es un conjunto polilingüe de documentos clasificados.

Bajo este enfoque, García Vega et al. (2001) propone entrenar una red neuronal LVQ (*Learning Vector Quantization*), la cual utiliza aprendizaje competitivo, con una biblia bilingüe inglés-español utilizando los 66 libros como clases. El objetivo es proporcionar conjuntos de prueba en inglés, en español o en una mezcla de ambos. En su trabajo, compara los resultados de la red neuronal con los resultados de otros clasificadores como el Rocchio. El mejor resultado reportado se obtuvo al clasificar únicamente inglés con 76.87% de exactitud calculada por micropromedio. Con español se obtuvo 73.76%, y con documentos bilingües de 75.11%.

Bel et al. (2003) prueba el entrenamiento polilingüe como una solución a la clasificación multilingüe. El objetivo es obtener tan buenos resultados como al elaborar clasificadores monolingües por idioma, por lo que los resultados monolingües se usan como referencia base. Los experimentos se realizaron con los clasificadores Rocchio (R) y Winnow[1] (W) sobre el corpus ILO[2] en inglés y español con 12 clases. En su trabajo, utiliza palabras como atributos sin aplicar lematización debido a que experimentalmente demuestra que no hay diferencia significativa entre usar lematización y omitir el proceso. Utiliza validación cruzada a 12 dobleces y su promedio alcanza exactitudes de 75%(R) y 81%(W), las que son comparables con su referencia monolingüe en la cual obtiene 80%(R) y 86%(W) para inglés con 76%(R) y 79% (W) para español.

García Adeva (2005) experimenta con tres enfoques para determinar cual funciona mejor clasificando corpus multilingües. El primero, NPNC (n procesamientos, n clasificadores), construye clasificadores monolingües por

[1] Winnow es un clasificador similar al SVM

[2] Es un corpus jerárquico, no público, que contiene documentos relacionados a economía, derechos ciudadanos y laborales, entre otros.

idioma. Los otros dos enfoques son por entrenamiento polilingüe. En el método llamado 1P1C (1 pre-procesamiento, 1 clasificador), un clasificador aprende de un corpus polilingüe de entrenamiento pero la matriz de representación no utiliza tal cual el vocabulario polilingüe. Bajo la suposición de que existen palabras que comparten la misma raíz (lema) independiente del idioma, se utilizan como atributos n-gramas[3] con las primeras letras de cada palabra. La ventaja de este método es que no requiere identificación del idioma ni pre-procesamiento especifico por idioma. Sin embargo este enfoque puede no resultar si los idiomas son totalmente diferentes, en cuyo caso las raíces de las palabras no coincidirán. En el método llamado NP1C (n pre-procesamientos, 1 clasificador), un clasificador aprende de un corpus polilingüe de entrenamiento donde cada idioma recibe un pre-procesamiento diferente previa identificación del idioma. El pre-procesamiento incluye remover palabras vacías y lematización. Se probaron los clasificadores Naive Bayes, Rocchio y Vecinos más cercanos (KNN), sobre un corpus recopilado en español y vasco con 20 clases. En orden, resultó mejor NPNC, NP1C y al final 1P1C; aún cuando sus resultados varían sólo en un par de puntos porcentuales. NPNC oscila entre el 45% y 75% de exactitud.

Chih-Ping (2007) desea aprovechar la ventaja de poseer corpus etiquetados en dos idiomas diferentes con las mismas clases. Para el entrenamiento poililingüe, en primer lugar, construye automáticamente un Tesauro a partir de un corpus paralelo inglés-chino de 2074 noticias, estableciendo experimentalmente los parámetros para su construcción. Utiliza adicionalmente dos corpus construidos en inglés y chino respectivamente, ambos poseen 278 noticias en 8 categorías. El entrenamiento se realiza con una mezcla de ambos seleccionando el 50% de

[3] Un n-grama, en este contexto, es una cadena de n caracteres consecutivos.

los corpus originales, proceso que se realiza 10 veces para evitar selecciones ventajosas. La selección de atributos se realiza por la relevancia de los términos en los conjuntos de entrenamiento y el tesauro. Los conjuntos de prueba están separados por idioma. Se usan los clasificadores Naive Bayes y SVM con los esquemas de pesado booleano, TF y TFxIDF. Se alcanzan exactitudes entre 63% y 72%, aproximadamente 5 puntos porcentuales arriba de su referencia monolingüe establecida entre 58% y 68%.

La ventaja de usar clasificación con entrenamiento polilingüe es que en general los resultados demuestran ser comparables con la clasificación monolingüe por idioma y no necesita módulo de reconocimiento de lenguaje. Sin embargo, retoma la misma suposición que la solución monolingüe, es decir, se debe contar con corpus etiquetados en diferentes idiomas para poder realizarse. Como desventaja adicional, la incorporación de nuevos idiomas requiere una modificación total del clasificador.

3.3 Clasificación Translingüe

La clasificación translingüe se define como la tarea de asignar clases a documentos escritos en un idioma objetivo usando recursos de un idioma fuente (Gliozzo y Strapparava, 2006). Se parte de la suposición de que el conjunto de entrenamiento está disponible únicamente en un idioma y debe ser usado para clasificar documentos en otros idiomas. La Figura 3.3 ilustra el esquema general de la clasificación translingüe. Además del módulo de identificación de idioma, requiere un módulo de relación al idioma fuente. El resultado es un conjunto de documentos clasificados en el idioma fuente.

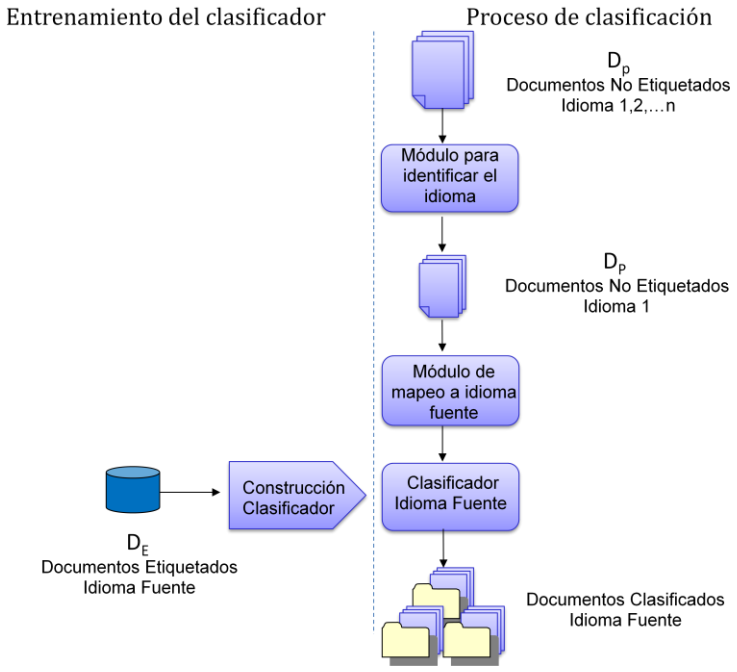


Figura 3.3 Esquema general de la clasificación translilngüe

Variantes del esquema mostrado se presentan en el siguiente capítulo. Entre las formas de relacionar los dos idiomas se han utilizado términos comunes, relaciones conceptuales definidas en tesauros, diccionarios y traducción automática.

3.3.1 Clasificación Translingüe con Términos Comunes

El objetivo de utilizar vocabulario común entre dos o más idiomas es lograr una clasificación evitando las traducciones. Este enfoque puede funcionar si las clases contienen suficientes entidades nombradas y cognados (Ver Anexo A) para ser utilizados como atributos y con ello permitir la discriminación entre clases. Bel et al. (2003) reporta haber realizado un experimento entre inglés y español sin traducciones a pesar de la diferencia

de vocabularios, en el que alcanza 10.75% de exactitud debido únicamente a los términos compartidos entre ambos idiomas.

Gliozzo y Strapparava (2006) establecen que por la diferencia de vocabulario, no es posible utilizar una medida de similitud entre documentos de diferentes idiomas usuando el modelo de representación vectorial, ya que los documentos son pobremente representados al utilizar directamente vocabulario de otro idioma. En su trabajo, explora una solución usando LSA (*Latent Semantic Analisys*) buscando términos comunes entre los idiomas. En sus experimentos utiliza Naive Bayes y SVM sobre un corpus construido con mensajes en inglés e italiano con 4 clases: Calidad de vida, Hecho en Italia, Turismo y Cultura. Como métrica de comparación usa *F-measure* alcanzando 0.55 entrenando con italiano y probado con inglés y 0.65 entrenando con inglés y probado con italiano. La referencia monolingüe es de aproximadamente 0.94 en inglés y 0.92 en italiano. Este enfoque es fácilmente superado cuando se usan otros recursos como tesauros y diccionarios.

3.3.2 Clasificación Translingüe con Diccionarios

Utilizar diccionarios como apoyo en la clasificación translingüe, reduce el esfuerzo de traducción limitándola a traducir sólo ciertas palabras y no documentos completos.

Olsson et al. (2005) utilizan un diccionario probabilístico construido por ellos mismos. El corpus que utilizan es peculiar ya que son transcripciones de conversaciones orales en inglés. El trabajo está inspirado en un problema real ya que no existen datos en idioma checo. Debido a la falta de datos en checo se hace una aproximación traduciendo ejemplos en inglés. El objetivo es realizar el etiquetado automático de las transcripciones por medio de la

clasificación. El clasificador elegido fue KNN, para evaluarse hace una lista de los temas más relevantes, consigue sólo 23% en idioma Checo con una referencia monolingüe en inglés de 73%. No menciona el número de clases.

Gliozzo y Strapparava (2006) se auxilian del diccionario Collins para relacionar dos idiomas por su traducción. Utilizan Naive Bayes y SVM sobre un corpus comparable construido con mensajes en inglés e italiano con 4 clases. Los *F-measure* obtenidos son aproximadamente de 0.88 de inglés a italiano y viceversa. El monolingüe alcanza 0.92 en el mejor caso. Éste fue el mejor de los tres enfoques propuestos: términos comunes, diccionario y tesauro.

3.3.3 Clasificación Transilingüe con Tesauros

Con la idea de relacionar varios idiomas, se ha abordado el uso de tesauros. Un tesauro es una herramienta que a diferencia de un diccionario, tiene definidas las relaciones entre conceptos (Ver Anexo A). El uso de tesauros se basa en la idea de que es posible expresar cada palabra por un concepto y con ello expandir el vocabulario por medio de las relaciones a dicho concepto. El tesauro multilingüe puede ser visto como un ensamble de tesauros monolingües que engloba todos en un mismo sistema de conceptos. Los tesauros tienen conjuntos de sinónimos de palabras llamadas *synsets*, donde cada *synset* tiene un índice único que es el mismo aún si los idiomas son diferentes.

Gliozzo y Strapparava (2006) utilizan el tesauro MultiWordNet[4] (Ver Anexo A) para relacionar los idiomas que manejan, en este caso inglés e italiano, aprovechando que poseen los mismos indices de *synset*. Utilizan

[4] http://multiwordnet.itc.it/english/home.php

Naive Bayes y SVM sobre un corpus construido con mensajes en 4 clases. Los *F-measure* obtenidos son de aproximadamente 0.85 en ambos idiomas contra la referencia monolingüe de 0.92 en el mejor caso.

De Melo y Siersdorfer (2007) no sólo buscan relacionar dos idiomas por medio de un tesauro, sino que aprovechan la estructura del tesauro para adquirir conocimiento adicional. Los experimentos fueron realizados con SVM, sobre el corpus Reuters en inglés y español, usando el tesauro WordNet. Los vectores de documentos usan como atributos palabras con pesado TFxIDF previa eliminación de las palabras vacías. No existe referencia monolingüe ya que el objetivo es apreciar si la clasificación mejora usando tesauros comparada con enfoques más sencillos como el usar sólo palabras. Utiliza exactitud como medida de evaluación calculada por micropromedio, obteniendo 91.88 al usar el tesauro y 80.97 en el esquema más simple usando sólo palabras.

3.3.4 Clasificación Translingüe con Traducción Automática

Consiste en insertar un bloque de traducción automática en el esquema de clasificación. Se han utilizado traductores automáticos tanto para traducir documentos completos o términos elegidos. Dentro de la clasificación translingüe existen dos enfoques, traducir el conjunto objetivo o el de entrenamiento.

Traducción del Conjunto Objetivo (TCP)

El clasificador es entrenado utilizando documentos etiquetados en el idioma fuente, por lo que conjunto de documentos no etiquetados en el idioma objetivo debe traducirse al idioma fuente antes de ser recibidos por el clasificador (Rigutini et al., 2005).

Jalam (2003) usa babelfish[5] como traductor automático. Escoge un sólo idioma para el clasificador y traduce los documentos que se van a clasificar al idioma en el que está construido el clasificador. En el trabajo, selecciona un número fijo de palabras, experimentando con 100 y 200 palabras. Sin embargo, no usa las palabras completas como atributos sino que varía entre 3-gramas, 4-gramas y 5-gramas. Los mejores resultados experimentales se obtienen al seleccionar 200 palabras representadas en 4-gramas. Clasifica con C4.5 y KNN (fijado a 3 vecinos) con un corpus adaptado del CLEF en inglés, francés y alemán con 10 clases. Los resultados del clasificador KNN superan a los de C4.5. Cuando la referencia monolingüe llega a 93% en promedio, se alcanza 90% de precisión y recuerdo en el experimento de francés a inglés y 97% en el experimento de alemán a inglés.

Bel et al. (2003) no realiza traducción de todo el documento sino únicamente de los términos relevantes del corpus objetivo, argumentando que la traducción de todo el documento no asegura mejores resultados. Usa los mismos parámetros y corpus que en la clasificación polilingüe. Los clasificadores son Rocchio(R) y Winnow(W) Como medida de evaluación se utiliza la exactitud por micropromedio. En los resultados se aprecia una pérdida de aproximadamente 10 puntos porcentuales al obtener 58%(R) y 61%(W) de inglés a español comparado contra 76%(R) y 79%(W) en el monolingüe de español y 70%(R) y 79%(W) de español a inglés comparado contra 80%(R) y 86%(W) del monolingüe en inglés.

[5] http://babelfish.altavista.com

Traducción del Conjunto de Entrenamiento (TCE)

Este enfoque requiere la traducción al idioma objetivo del conjunto de documentos etiquetados, dicha traducción se utiliza para entrenar al clasificador. Rigutini et al. (2005) utiliza el esquema TCE argumentando que debido a que el conjunto de documentos no etiquetados puede ser sumamente grande, un enfoque basado en la traducción del conjunto objetivo resultaría muy costosa en tiempo puesto que la traducción es un proceso lento. En cambio, la traducción del conjunto de entrenamiento requiere traducir una sola vez. De forma que al entrenar al clasificador con una traducción, éste puede recibir a los documentos no etiquetados directamente en su idioma original.

Rigutini et al. (2005) construye un clasificador traduciendo el conjunto de entrenamiento al idioma objetivo. Después de asignar clases a los documentos no etiquetados, selecciona un número fijo de palabras con mayor ganancia de información para construir un nuevo clasificador iterativamente, sin volver a utilizar el corpus de entrenamiento. El algoritmo se detiene cuando no existen cambios en la asignación de clases. Cabe mencionar que Rigutini menciona un comportamiento indeseable en el algoritmo en el cual alguna clase se queda con un conjunto vacío de ejemplos. Los experimentos se realizaron con Naive Bayes sobre un corpus creado con mensajes en inglés e italiano en 3 clases: autos, hardware y deportes. Utiliza una versión de prueba del Office Translator Idiomax[6] para realizar la traducción. Los mejores resultados experimentales se obtienen seleccionando 300 palabras en el primer ciclo y 1000 en los posteriores. No menciona el número de iteraciones que se realizaron. Los resultados llegan a

[6] www.idiomax.com

90.6 de recuerdo y 90.64 de precisión con referencia monolingüe en promedio de 94.3 de recuerdo y 93.50 de precisión.

3.4 Resumen del Estado del Arte

La Tabla 3.1 muestra un resumen del estado del arte, la sección de clasificadores monolingües se presenta en la última columna. Los investigadores muestran los resultados de referencia monolingüe a fin de compararlos con los obtenidos por sus métodos. Cada investigador utilizó su propio conjunto de documentos bajo los criterios que consideraron adecuados para su trabajo.

Tabla 3.1 Resumen del estado del arte

	Referencia	Clasificador	Mejor resultado[7]	Referencia Monolingüe
Pollilingüe	García Vega et al. (2001)	RNA	Inglés 76.87% Español 73.76% Bilingüe 75.11%	--
	Bel et al. (2003)	Rocchio (R) Winnow (W)	81% (W) 75% (R) Bilingües	80%(R)-86%(W) inglés 76%(R)-79%(W) español
	García Adeva (2005)	Naive Bayes Rocchio KNN	72%(N) 68%(R) 61%(K) aprox. 1P1C	75.8%(N) 46%(R) 64%(K)
	García Adeva (2005)	Naive Bayes Rocchio KNN	73%(N) 45%(R) 62%(K) aprox. NP1C	75.8%(N) 46%(R) 64%(K)
	Chih-Ping (2007)	Naive Bayes SVM	72.42%(N) inglés 68.87%(S) inglés 72.42% (N) chino 68.87% (S) chino	67.63%(N) inglés 66.14%(S) inglés 65.61% (N) chino 62.33% (S) chino

[7] El porcentaje mostrado corresponde a la exactitud como medida de evaluación calculada con micropromedio, a excepción de aquellos casos en que se indique una medida diferente.

Tabla 3.1 (Cont.) Resumen del estado del arte

		Referencia	Clasificador	Mejor resultado[8]	Referencia Monolingüe
Tanslingüe	Términos comunes	Bell et al. (2003)	Rocchio (R) Winnow (W)	10.75%	76-86%
		Gliozzo y Strapparava (2006)	Naive Bayes SVM	F-measure Italiano-inglés 0.55 Inglés-italiano 0.65	0.94 inglés 0.92 italiano
	Diccionarios	Olsson (2005)	KNN	23% Checo	73% inglés
		Gliozzo y Strapparava (2006)	Naive Bayes SVM	F measure Inglés 0.88 Italiano 0.88	0.94 inglés 0.92 italiano
	Tesauro	Gliozzo y Strapparava (2006)	Naive Bayes SVM	F-measure Inglés 0.85 Italiano 0.84	0.94 inglés 0.92 italiano
		De Melo y S(2007)	SVM	80.97% con palabras 91.88% con tesauro	
	Traducción	Jalam (2003)	KNN C4.5	π=ρ 90% francés a inglés 97% de alemán a inglés	93% promedio
		Bel et al. (2003)	Rocchio (R) Winnow (W)	58%(R)-61%(W) inglés a español 70%(R)-79%(W) español a inglés	80%(R)-86%(W) inglés 76%(R)-79%(W) español
		Rigutini et al (2005)	Naive Bayes	Recuerdo 90.6 Precisión 90.64 inglés e italiano	Recuerdo 94.3 Precisión 93.5

3.5 Discusión

Entre las soluciones propuestas hasta ahora para abordar la clasificación automática de textos en el entorno multilingüe, se identifican la clasificación monolingüe por idioma, la clasificación por entrenamiento polilingüe y la clasificación translingüe.

La construcción de clasificadores monolingües por idioma y la clasificación por entrenamiento polilingüe, son aplicables únicamente ante la

[8] El porcentaje mostrado corresponde a la exactitud como medida de evaluación calculada con micropromedio, a excepción de aquellos casos en que se indique una medida diferente.

existencia de suficientes datos etiquetados en todos los idiomas involucrados. Ya que el proceso de etiquetado es costoso, dichos enfoques resultan poco prácticos. Incluso, si el proceso de etiquetado se realiza de forma manual, se requieren expertos poliglotas o dispersos por el mundo para poder realizarse. Además, cada problema se trata independientemente, dificultando la retroalimentación de cualquier tipo, ignorando la experiencia adquirida para resolver problemas futuros, y desaprovechando recursos existentes en algún idioma para clasificar otros.

Dentro de la clasificación translingüe, se ha considerado el uso de términos comunes, diccionarios, tesauros y traductores automáticos. Los trabajos que utilizan vocabulario común han encontrado que, a pesar de la existencia de palabras en común o cognados[9], el vocabulario entre idiomas es en su mayoría diferente, ocasionando representaciones pobres de los documentos cuando se usa directamente el vocabulario de otro idioma. Aún se busca una representación única para los documentos multilingües, con el objetivo de evitar la traducción.

En la clasificación translingüe, el uso de recursos como diccionarios y tesauros con el fin de efectuar la traducción o ampliar el vocabulario han demostrado ser favorables pero no solucionan por completo el problema, ya que no alcanzan a ser comparables con los resultados monolingües. El uso de diccionarios puede obtener resultados inferiores a otros enfoques debido a que los diccionarios generales no incluyen vocabulario especializado, fracasan en la traducción de palabras compuestas y caen en problemas de ambigüedad (Jalam, 2003). Los tesauros por su parte, tienen la limitación de poseer un vocabulario controlado, aunado a que son complicados de

[9] Palabras con poca variación en los distintos idiomas; ver Anexo A para mayor detalle.

construir, costosos de mantener y difíciles de actualizar. Además, no existen tesauros para tantos idiomas como otros recursos, mencionando por ejemplo diccionarios y traductores, en cuyo caso debe elaborarse primero el tesauro.

La clasificación translingüe con traductores automáticos tiene la ventaja de que el número de idiomas para los cuales existe un traductor se incrementa gradualmente. No son tan restrictivos en cuanto al dominio o cobertura como otras herramientas. Pueden ser usados para trasladar recursos existentes de un idioma a otros idiomas, previa consideración de los efectos de traducción. De los recursos utilizados, el uso de traductores automáticos ha resultado más exitoso.

Hasta ahora la mayoría de los trabajos se limitaban a observar el desempeño general de la clasificación probando distintos clasificadores, preprocesamientos o representaciones para los atributos como n-gramas, palabras completas o al aplicar lematización. Sin embargo la clasificación translingüe simple es insuficiente. Las razones que se mencionan son distorsiones introducidas por las herramientas y la discrepancia cultural entre dominios.

Dado que el enfoque con más ventajas para la clasificación translingüe es el basado en la traducción, el presente trabajo se desarrolla con este enfoque. En el capítulo siguiente se efectúa un análisis de la clasificación translingüe. Dicho análisis tiene como objetivo apreciar los efectos de la traducción y la discrepancia cultural, ya que, si bien los investigadores los mencionan como posibles razones de la insuficiencia de la clasificación translingüe, es conveniente verificarlo. Los experimentos se realizan con los dos esquemas existentes (TCP y TCE), considerando que no se han encontrado trabajos con ambas.

Capítulo 4

Análisis de la Clasificación Translingüe

En el capítulo anterior se mostraron las soluciones actuales al problema del multilingüismo dentro de la clasificación de textos. En el marco de la clasificación translingüe se abordaron las investigaciones que han propuesto alguna forma de franquear la barrera del idioma, como es el uso de términos comunes, diccionarios, tesauros y traductores automáticos. Hasta ahora el uso de traductores automáticos ha demostrado poseer más ventajas sobre los otros enfoques debido a los buenos resultados que se obtienen y a la practicidad en el manejo de la herramienta.

En este capítulo hace un análisis de la clasificación translingüe con el objetivo de apreciar los efectos de la traducción y la discrepancia cultural. Para ello, se definen los dos esquemas conocidos para la clasificación translingüe y se determina el conjunto de documentos con el cual se efectúan los experimentos de referencia de este trabajo.

4.1 Esquemas de la Clasificación Translingüe

Como se mencionó anteriormente, la clasificación translingüe utiliza recursos de un idioma fuente para clasificar documentos de un idioma objetivo. Concretamente se considera el caso en el que se tiene acceso al conjunto de documentos etiquetados necesarios para entrenar a un clasificador en al menos un idioma. Entonces, se desea utilizar dicho clasificador para categorizar documentos escritos en otro idioma. Sea D_E el conjunto de entrenamiento en el idioma fuente, donde el valor de la función $\Phi: (d_j, c_i)$ es conocido para cada $d_j \in D_E$ y $c_i \in C$ donde C es un conjunto finito de clases. Se desea aproximar la función $\breve{\Phi}: (d_k, c_i)$ para el conjunto D_P en el idioma objetivo. C es el mismo conjunto de clases para todos los idiomas y L_1, L_2,…L_n representan los n idiomas que se desean clasificar. Dado que se poseen tanto el conjunto de entrenamiento como el objetivo, se puede aplicar el proceso de traducción a cualquiera de los dos conjuntos, formando con esto dos posibles esquemas para la clasificación translingüe que se explican a continuación.

4.1.1 Traducción del Corpus de Entrenamiento (TCE)

Consiste en aplicar el traductor al conjunto de entrenamiento en el idioma fuente D_E para obtener el mismo conjunto pero ahora en el idioma objetivo. Cabe mencionar que después de la traducción el conjunto de documentos se procesa como si hubiera sido escrito en el idioma objetivo. El clasificador se entrena con atributos extraídos de la traducción del conjunto de entrenamiento D_E como lo muestra la Figura 4.1; el conjunto de documentos a clasificar D_P se representa con los atributos de entrenamiento y el clasificador asigna una clase a cada documento.

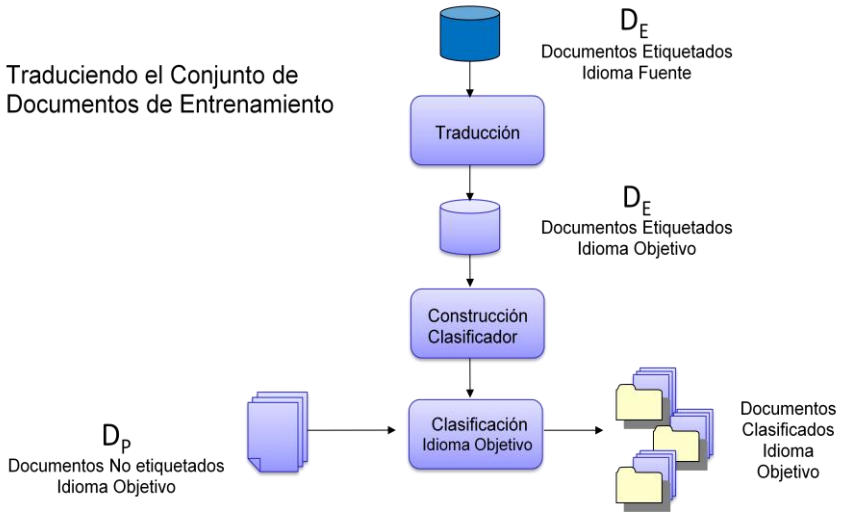


Figura 4.1 Esquema básico traduciendo conjunto de entrenamiento (TCE)

La ventaja del esquema TCE es que no se realiza traducción de los documentos a clasificar, por lo que el proceso de traducción se realiza una sola vez. Una desventaja consiste en que para clasificar n idiomas deben realizarse n clasificadores.

4.1.2 Traducción del Corpus Objetivo (TCP)

Consiste en aplicar el proceso de traducción al conjunto de documentos no etiquetados D_P en el idioma objetivo para obtener el mismo conjunto pero en el idioma fuente. El clasificador se construye con el conjunto de entrenamiento en el idioma fuente como lo muestra la Figura 4.2, donde se aprecia al clasificador recibiendo los documentos objetivo provenientes del módulo de traducción. El conjunto no etiquetado se encuentra representado con las características extraídas del conjunto de entrenamiento en el idioma fuente.

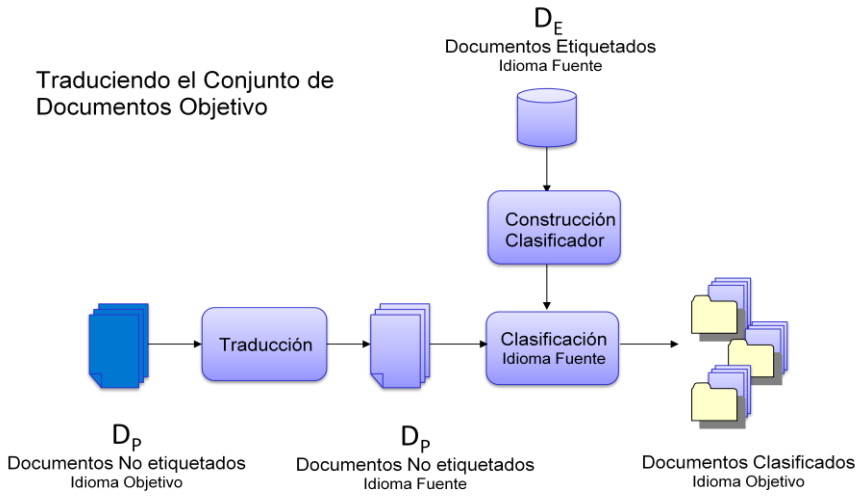


Figura 4.2 Esquema básico traduciendo conjunto objetivo (TCP)

Entre las ventajas se puede mencionar que el clasificador se entrena en el idioma fuente por lo que no existe ruido alguno de traducción. Adicionalmente, debido a que el traductor se utiliza para el conjunto no etiquetado, se usa un único clasificador capaz de recibir los conjuntos correspondientes a cualquier idioma. Sin embargo la cantidad de traducciones podría incrementarse de manera importante debido a que debe traducirse cada conjunto objetivo que se desee clasificar; esto implica que para n conjuntos objetivo, se realizan n traducciones.

Independientemente del conjunto al que se aplique el proceso de traducción, deben considerarse los efectos resultantes de la utilización de un traductor. Por ello se debe identificar los problemas específicos tanto de las traducciones como de realizar clasificación translingüe en general. Antes de abordar la discusión de estos problemas, se describe el corpus de trabajo que se utiliza tanto para el análisis de la clasificación translingüe como para los métodos propuestos descritos en el siguiente capítulo.

4.1.3 Corpus de Trabajo

Para realizar los experimentos de clasificación translingüe se requiere un conjunto multilingüe de documentos etiquetados. Como características particulares, cada idioma debe contar con las mismas clases que los otros idiomas y cada clase debe contener suficientes documentos en cada idioma para formar los conjuntos de entrenamiento y prueba. No existen muchos corpus multilingües disponibles. La mayoría de los investigadores prefiere recopilar noticias o mensajes para cumplir con los requerimientos de su trabajo. Una recopilación conocida es el corpus Reuters.

La colección *Reuters* es actualmente la base de datos más popular en la evaluación de investigación de clasificación de textos (Manning y Schütze, 2001). RCV1 contiene 810,000 noticias en inglés y RCV2 alrededor de 487,000 noticias en 13 idiomas: alemán, francés, italiano, japonés, mandarín, ruso, español, español latinoamericano, portugués, holandés, danés, noruego y sueco[10]. A continuación se presentan algunas características de la colección Reuters.

- El etiquetado se realizó de forma manual por un equipo de trabajo.
- Cada documento puede pertenecer a varias clases según el criterio del etiquetador.
- Existen documentos que no están asignados a clase alguna.
- Las clases no son balanceadas, puede haber más noticias de una clase que de otra e incluso existir clases vacías.
- El corpus está dividido en grupos y subgrupos (Lewis et al., 2004).

[10] http://trec.nist.gov/data/reuters/reuters.html

Para el presente trabajo se seleccionó un subconjunto de noticias de México, E.U y Francia del grupo llamado GCAT. Ya que el corpus Reuters posee documentos asignados a una o varias clases, se filtraron aquellos que pertenecen a una sola clase. Se seleccionaron 4 de los subgrupos: GCRIM, GDIS, GPOL y GSPO a los que llamamos Crímenes, Desastres, Política y Deportes respectivamente; en los que existen suficientes documentos en los 3 idiomas. La Tabla 4.1 presenta una breve descripción del tipo de noticias asignadas a las clases mencionadas.

Tabla 4.1 Descripción de clases seleccionadas del corpus.

Grupo	Subgrupo	Clase	Descripción
GCAT-	GCRIM	Crímenes	Derecho civil y penal, ley y orden, delitos relacionados con drogas, crimen organizado, delincuencia, fraude, asesinato, criminales, mafia, policía.
GCAT-	GDIS	Desastres	Desastres naturales; terremotos, tsunamis, erupciones volcánicas, maremotos, huracanes, tornados. Desastres provocados por el ser humano; desastres aéreos, choques, accidentes carreteros, naufragios.
GCAT-	GPOL	Política	Partidos políticos, parlamento, congreso, constitución
GCAT-	GSPO	Deportes	Historias deportivas, marcadores de encuentros.

En el corpus de trabajo cada clase tiene el mismo número de noticias en todos los idiomas. Se utilizaron 200 noticias en cada idioma para entrenamiento (50 por clase) y 120 como objetivo (30 por clase). Cada noticia cuenta con aproximadamente 400 palabras, aunque las longitudes son variadas. Para el pre-procesamiento de los documentos, se utilizaron listas de palabras vacías en español, inglés y francés, cada lista cuenta con al menos 150 palabras.

El subconjunto seleccionado del corpus Reuters, de la forma descrita anteriormente, es el que se utiliza como corpus de trabajo durante la

investigación. El primer paso es realizar directamente la clasificación translingüe a fin de establecer los datos iniciales.

4.1.4 Resultados de la Clasificación Translingüe

Los investigadores que han realizado estudios previos sobre clasificación translingüe, como Bel et al. (2003), Rigutini (2005) y Chih-Ping (2007) entre otros, incluyen en sus trabajos los resultados que obtuvieron en la clasificación monolingüe. Dichos resultados sirven de referencia para demostrar que tan bien funcionan los métodos aplicados. En este trajo se utiliza el traductor automático Worldlingo[11] y se clasifica con Naive Bayes (sección 2.3.1). La Tabla 4.2 muestra los resultados de la clasificación monolingüe con el corpus definido anteriormente, se aprecia que el promedio de exactitud es de 92.22%.

Tabla 4.2 Resultados de la clasificación monolingüe

Entrenamiento	Objetivo	Exactitud
Inglés	Inglés	91.66%
Español	Español	91.66%
Francés	Francés	93.33%
	Promedio	92.22%

Para evitar sesgos en los resultados de la clasificación translingüe, los experimentos se realizaron con los dos esquemas (TCE y TCP) descritos en la Sección 4.1. No tenemos conocimiento de algún trabajo que experimente con ambos esquemas. La Tabla 4.3 muestra los resultados de la clasificación translingüe traduciendo el conjunto de entrenamiento. Se observa que el promedio es de 79.85% de exactitud. Aunque no se busca en este trabajo

[11] www.worldlingo.com

hacer análisis por idioma, es interesante observar que el entrenamiento proveniente del inglés obtiene los resultados más bajos.

Tabla 4.3 Resultados de la clasificación translingüe TCE

Entrenamiento	Objetivo	Exactitud
Inglés Español	Español	71.66%
Inglés Francés	Francés	75.83%
Español Inglés	Inglés	81.66%
Español Francés	Francés	80.83%
Francés Inglés	Inglés	85.83%
Francés Español	Español	83.33%
	Promedio	79.85%

La Tabla 4.4 muestra los resultados de la clasificación translingüe traduciendo el conjunto objetivo. Se observa que el promedio es de 80.41% de exactitud y de la misma forma que el esquema anterior, el entrenamiento en inglés ofrece resultados más bajos que entrenar con español o francés. Indicando que los resultados entre ambos esquemas son consistentes.

Tabla 4.4 Resultados de la clasificación translingüe TCP

Entrenamiento	Objetivo	Exactitud
Inglés	Español Inglés	75.00%
Inglés	Francés Inglés	76.66%
Español	Inglés Español	85.00%
Español	Francés Español	79.16%
Francés	Inglés Francés	86.66%
Francés	Español Francés	80.00%
	Promedio	80.41%

En general, la caída en exactitud es de hasta 20 puntos porcentuales en comparación de la referencia monolingüe, siendo el promedio 12 puntos en los experimentos realizados. A continuación se describen los problemas específicos de la clasificación translingüe; algunos de ellos ya habían sido

mencionados en investigaciones anteriores como probables inconvenientes relacionados a la clasificación translingüe.

4.2 Problemas Específicos de la Clasificación Translingüe

Cuando se efectúa clasificación translingüe usando traductores automáticos se debe considerar que el objetivo de la clasificación y el objetivo del traductor no están encaminados hacia el mismo fin. Mientras que la clasificación usualmente se realiza buscando similitudes entre documentos por medio de algunos atributos, el objetivo de un traductor automático es producir una versión legible y fiable de los documentos de un idioma a otro (Jalam, 2003). El traductor no necesariamente utilizará el vocabulario con el que se probará el clasificador e inevitablemente inducirá ciertas distorsiones.

4.2.1 Distorsión Introducida por el Traductor:

El uso de un traductor automático, independientemente de cuál sea el elegido, implica el riesgo de introducir ruido durante el proceso de clasificación. Algunos de los efectos son propios del proceso de traducción y se refieren al lenguaje, otros sin embargo, son resultado de errores de la herramienta o del uso inadecuado de la misma. A continuación se enlistan algunos efectos del uso de traductores que pueden tener un impacto directo en los resultados de la clasificación translingüe:

Traducción con Sinónimos: Es el efecto más mencionado en las investigaciones relacionadas y se debe a que el traductor automático no necesariamente utilizará el mismo vocabulario que poseen los conjuntos objetivo. En algunos casos la traducción se vuelve un problema severo cuando la palabra “coche” en español se mapea a “car” en inglés, mientras que el francés “voiture” se traduce a “automobile” en inglés, de forma que el

algoritmo de aprendizaje computacional permanece en desconocimiento de la sinonimia (de Melo y Siersdorfer, 2007). Los efectos de la sinonimia también fueron notados por Bel et al. (2003), quien asegura que muchos de los sinónimos generados en la traducción afectan la distribución de frecuencias y con ello los resultados de la clasificación.

Traducción de Entidades Nombradas: Se les llama entidades nombradas a aquellas palabras que por tratarse de nombres propios como Michael Jackson, tienen poca o ninguna variación entre idiomas (véase Anexo A). Algunas entidades nombradas pueden ser elegidas como atributos para el clasificador, en cuyo caso su permanencia en las traducciones cobra importancia. Las traducciones erróneas no sólo promueven la pérdida de atributos sino que llegan a generar ruido. Por ejemplo, el famoso arquero mexicano Jorge Campos es transformado en Jorge Fields por el traductor, introduciendo vocabulario que puede ser ruidoso en el dominio.

Falta de Traducción: Es posible encontrar palabras, oraciones, párrafos o incluso documentos completos sin traducir. Problemas como éstos deben identificarse y solucionarse rápidamente para evitar una pobre representación de los documentos al carecer de parte del vocabulario. Sucede con más frecuencia en dominios especializados, pero no quedan exentas palabras poco comunes. Por ejemplo: Los padres acongojados…---the ‘acongojados’ parents…

Errores de Traducción: Los errores cometidos por los traductores automáticos son determinados generalmente por fenómenos lingüísticos tales como la polisemia, homonimia, ambigüedad, expresiones coloquiales, colocaciones y falsos cognados que se encuentran presentes de forma

común en el lenguaje natural[12] (Ver Anexo A). Una cara conocida --- An expensive acquaintance[13].

Los efectos antes mencionados se encuentran en cualquier traductor automático y se debe estar consciente de ellos al incluir un proceso de traducción. Más adelante se realiza un análisis del impacto de la traducción en la tarea de clasificación. Ya se mencionó que la traducción afectará la frecuencia de distribución del vocabulario, que propiciará la pérdida de atributos y que es capaz de introducir vocabulario ruidoso. Sin embargo existe otro motivo para que existan diferencias entre el vocabulario del conjunto de entrenamiento y el objetivo, la discrepancia cultural.

4.2.2 Discrepancia Cultural entre Idiomas

Debido a que el lenguaje es una expresión cultural, la diferencia cultural entre idiomas es un factor inherente a la clasificación translingüe. Significa que los términos utilizados comúnmente en un idioma pueden ser escasos o no existir en otro idioma, por ejemplo, es de esperarse que los mexicanos hablen más sobre fútbol soccer que los americanos; y que a su vez, deportes como el rugby, la pelota vasca y el cricket sean más nombrados en países europeos. Considere el ejemplo de la Tabla 4.5; para obtenerla se tomaron 50 noticias de la clase ‘Deportes’ para los idiomas español, inglés y francés, obteniendo la frecuencia (F) de algunas palabras en los tres idiomas.

[12] Llamado así para diferenciarlo de los lenguajes de programación computacional.
[13] Ejemplo tomado de http://paginaspersonales.deusto.es/abaitua/_outside/ikasle/trad_03/alfredo_v.htm

Tabla 4.5 Frecuencia (F) del vocabulario de entrenamiento en la clase deportes

Español	F	Inglés	F	Francés	F
gol	92	hits	161	match	139
fútbol	70	yards	58	rugby	32
portero	27	tennis	41	football	11
cancha	21	goal	10	athletisme	5
tenis	4	baseball	9	tennis	2
beisbol	2	football	6	cricket	1

En el idioma inglés “tennis” es una palabra frecuente con 41 apariciones en el conjunto de noticias, mientras que en español y francés aparece sólo 4 y 2 veces respectivamente. Las noticias en francés incluyen la palabra “rugby” que aparece 32 veces en su conjunto, mientras que en los conjuntos de inglés y español no aparece en el vocabulario, por lo que es de esperarse que los conjuntos español e inglés identifiquen pobremente temas como rugby.

Para determinar el impacto de la traducción y las diferencias culturales en los resultados de la clasificación translingüe, se realizaron varios análisis. Se usaron gráficas de similitud para observar ambos efectos y otros análisis fueron desarrollados para acentuar uno de ellos. En las siguientes secciones se presentan por objetivo los análisis efectuados.

4.3 Impacto de la Traducción

La sección 4.2.1 está dedicada a las distorsiones que pueden introducirse al hacer uso de un traductor automático. En esta sección se realizan dos análisis con el fin de visualizar el impacto de la traducción en los conjuntos de documentos y de determinar su impacto sobre la clasificación.

4.3.1 Análisis por Gráficas de Similitud

Con el objetivo de visualizar los efectos de la traducción se obtienen gráficas de similitud, que comparan cada documento con el resto del conjunto. En ellas se asigna un valor de escala de grises a un rango de similitud. Idealmente los documentos deben parecerse a todos los documentos de su misma clase pero no a documentos de otras clases. Para poder apreciarse los documentos pertenecientes a la misma clase deben ser contiguos, sin ser necesario ordenar los documentos bajo criterio alguno. En este trabajo, la representación vectorial de los documentos se obtiene mediante su vocabulario con pesado booleano y se usa *el coeficiente de Dice* para medir la similitud. Las gráficas de similitud se usan para visualizar los efectos de traducción y discrepancia cultural, se incluyen en el trabajo debido a que la apreciación visual resulta interesante.

La Figura 4.3 es una muestra de dichas gráficas realizada sobre el corpus de trabajo. En ella se aprecia una diagonal tomando el valor más alto de la escala de grises, indicando que para este caso, los documentos se comparan a sí mismos y se obtienen el valor máximo de similitud. Se aprecian 4 clases, las cuales resultan visibles debido a que lo documentos tienen un grado de similitud entre los de su misma clase y diferencias con documentos de otras clases. Algunos documentos también son similares en cierto porcentaje a documentos de otras clases, en consecuencia se observan varios puntos fuera del área de los cuadros. En contraste, la clase inferior izquierda contiene documentos no tan similares entre sí mismos, esto se deduce por la presencia de espacios en blanco, el rango más bajo de la escala de grises.

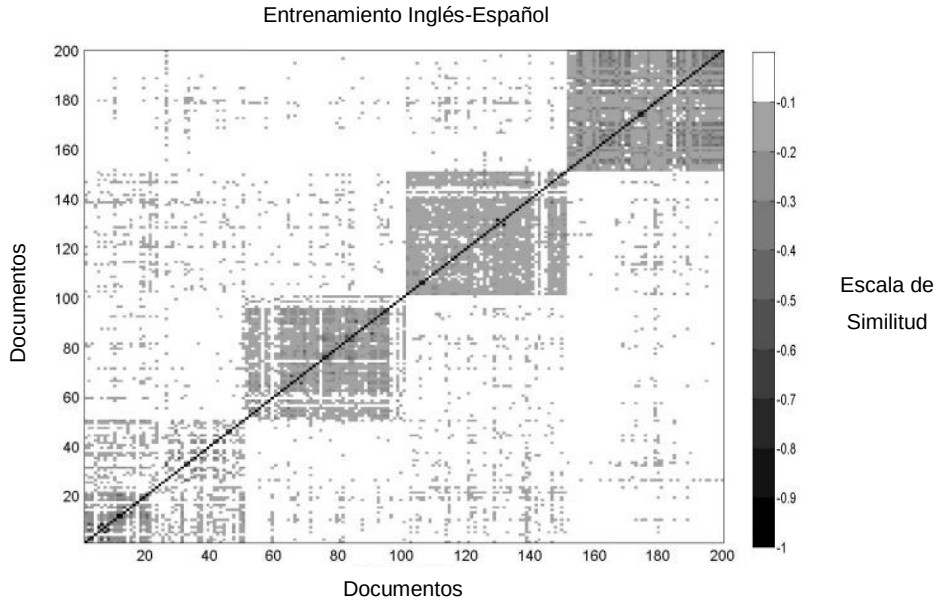
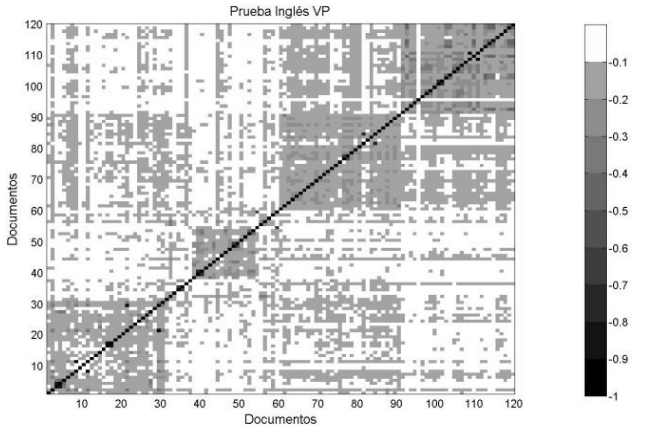


Figura 4.3 Ejemplo de gráfica de similitud de documentos en el corpus de trabajo.

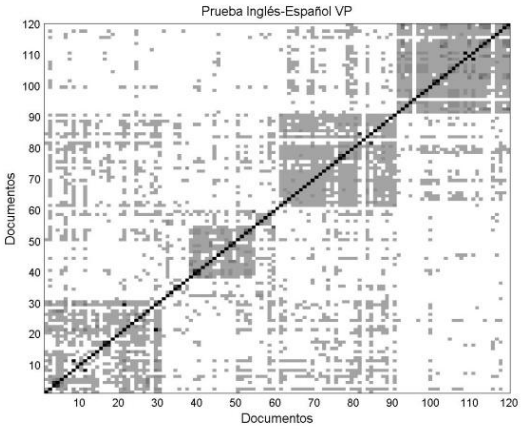
Para observar los efectos de la traducción se obtiene la gráfica de similitud del conjunto de documentos en su idioma original y se compara con la gráfica obtenida del mismo conjunto después de la traducción. Cabe mencionar que el proceso de traducción se efectúa directamente sobre los documentos originales eliminando únicamente etiquetas. Una vez realizada la traducción se ejecutan los procesos para retirar signos de puntuación, caracteres y palabras vacías propias del nuevo idioma. El nuevo vocabulario con el cual se representan los documentos puede favorecer o dificultar la consolidación de las clases.

La Figura 4.4 (a) muestra la gráfica de similitud del corpus objetivo en inglés. Observe que los documentos pertenecientes a la tercera clase también tienen similitud con la cuarta y ligeramente con la primera, mientras que los documentos de la segunda clase no sólo denotan diferencias con documentos de otras clases sino que incluso algunos no resultan similares

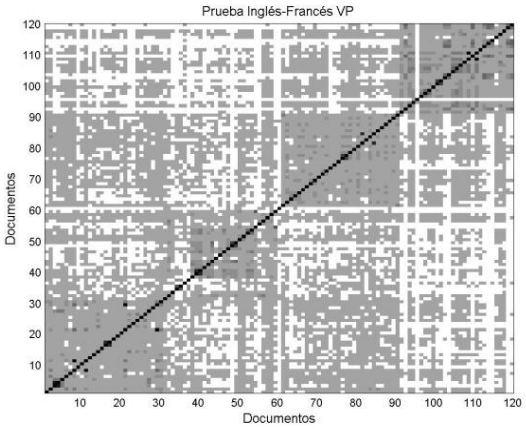
entre sí. Las Figuras 4.4 (b) y (c) presentan las gráficas de similitud para el corpus en inglés traducido al español y francés respectivamente. En este ejemplo, se observa que la traducción al español parece favorable porque degrada algunas similitudes entre documentos de la tercera y cuarta clase, sin embargo, también reduce las similitudes intrínsecas en el conjunto de documentos de la primera clase. En contraste, la traducción al francés propicia las similitudes de los documentos con otras clases. Cabe mencionar que las similitudes modificadas son bajas, la mayoría son menores a 0.2.



a) Conjunto objetivo en inglés



b) Traducción inglés-español



c) Traducción inglés-francés

Figura 4.4 Ejemplo del efecto de la traducción sobre un conjunto objetivo del corpus de trabajo

4.3.2 Análisis del Impacto de la Traducción en la Clasificación

Para medir el impacto directo de la traducción en la tarea de clasificación translingüe, se aplica el proceso de traducción tanto al conjunto de entrenamiento como al objetivo y se efectúa la tarea de clasificación. En este caso, se realizan las traducciones entre los corpus de inglés, español y francés[14]. Los resultados de la clasificación monolingüe sin traducción son comparados con la clasificación monolingüe con traducción en la Tabla 4.6 y la Tabla 4.7

Tabla 4.6 Resultados monolingües sin traducción

Entrenamiento	Objetivo	Exactitud
Inglés	Inglés	91.66%
Español	Español	91.66%
Francés	Francés	93.33%
	Promedio	92.22%

Tabla 4.7 Resultados monolingües con traducción

Entrenamiento	Objetivo	Exactitud
Inglés-Español	Inglés-Español	93.33%
Inglés-Francés	Inglés-Francés	90.83%
Español-Inglés	Español-Inglés	91.66%
Español-Francés	Español-Francés	90.83%
Francés-Inglés	Francés-Inglés	93.33%
Francés-Español	Francés-Español	94.16%
	Promedio	92.35%

A comparación con los resultados monolingües sin traducción, los monolingües con traducción clasifican en general el mismo número de documentos correctos, las variaciones mostradas corresponden a máximo

[14] Jalam (2003) realiza el experimento para alemán-inglés y francés-inglés con otro corpus de trabajo.

dos ejemplos. Las mejoras se deben probablemente debido a que la traducción logra simplificar el vocabulario y las bajas de exactitud probablemente correspondan a los efectos mencionados en la sección 4.2.1. Nuestra comparación muestra que la traducción tiene poco efecto en la clasificación, en consecuencia, los resultados desfavorables de la clasificación translingüe no son directamente atribuibles a la intervención de un traductor automático. A continuación se presentan los análisis realizados para determinar el impacto de la discrepancia cultural.

4.4 Impacto de la Discrepancia Cultural

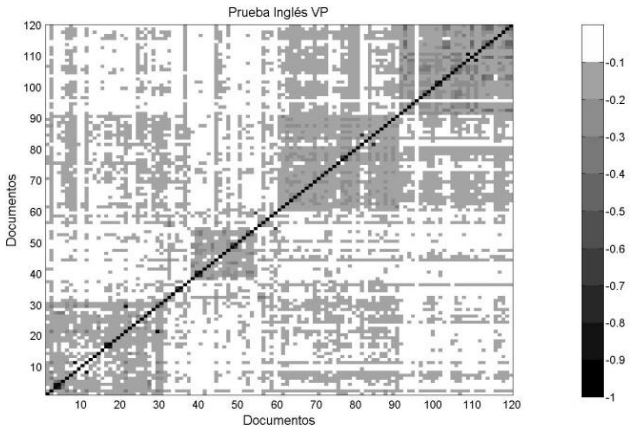
En esta sección se realizan tres análisis para determinar el impacto de la discrepancia cultural. El primero se orienta a la visualización de los conjuntos de documentos objetivo mediante gráficas de similitud. El segundo se enfoca en contabilizar el vocabulario común entre los conjuntos de entrenamiento y objetivo. El tercero aplica reducción de dimensionalidad en los conjuntos de entrenamiento de la clasificación translingüe para acentuar los efectos de la discrepancia cultural.

4.4.1 Análisis por Gráficas de Similitud

Estas gráficas nos permiten visualizar diferencias entre los conjuntos de documentos. Cada conjunto refleja el material con el cual fue construido, de forma que, aún contando con las mismas categorías, es posible encontrar diferencias entre corpus (Manning y Schütze, 2001), esto porque la distribución de vocabulario y por lo tanto los traslapes entre clases varían para cada conjunto de documentos.

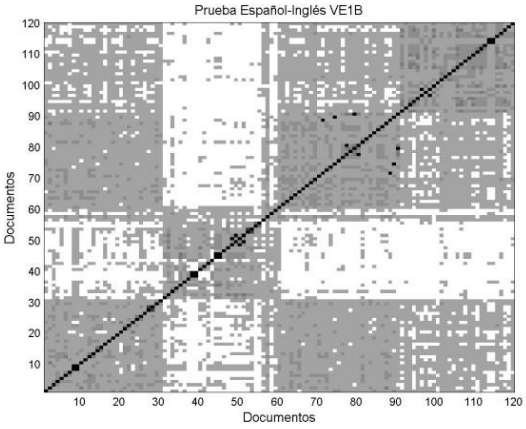
La Figura 4.5(a) muestra la gráfica de similitud del conjunto objetivo en inglés representado por los atributos del conjunto de entrenamiento también

en inglés y muestra similitudes ligeras entre documentos de la primera, tercera y cuarta clase, con disimilitudes intrínsecas entre los documentos de la segunda clase. La Figura 4.5(b) presenta la gráfica de similitud del conjunto en español traducido al inglés representado por el conjunto de entrenamiento en inglés, donde se aprecia que los documentos de la segunda clase presentan similitud entre ellos y diferencias a los de otras clases pero existen similitudes entre documentos de la primera, tercera y cuarta clase. La Figura 4.5(c) corresponde al conjunto objetivo en francés representado por los atributos del conjunto de entrenamiento en inglés. Nótese que la mayoría de los documentos tiene similitud con documentos de otras clases y sólo un grupo reducido de documentos de la cuarta clase sólo son similares a ellos mismos. La totalidad de las gráficas del corpus con traducción y sin ella, se presentan en el Anexo B.

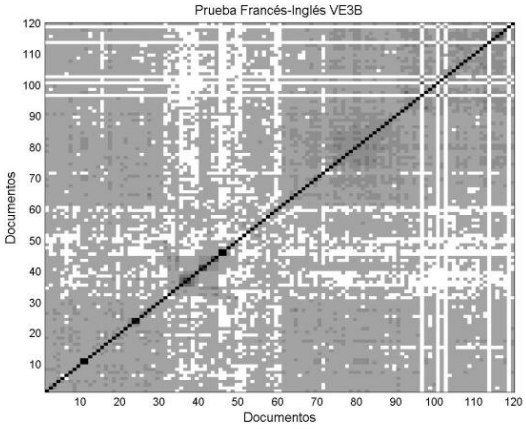


a) Conjunto objetivo en inglés representado por el conjunto de entrenamiento en inglés

Figura 4.5 Ejemplo de discrepancia cultural entre los conjuntos objetivo.



b) Conjunto objetivo en español-inglés representado por el conjunto de entrenamiento en inglés



c) Conjunto objetivo en francés-inglés representado por el conjunto de entrenamiento en inglés

Figura 4.5 Ejemplo de discrepancia cultural entre los conjuntos objetivo.

Los siguientes experimentos están enfocados a determinar el impacto directo de la diferencia cultural en la exactitud de la clasificación translingüe.

4.4.2 Análisis por Vocabulario

El éxito de la clasificación está ligado a la pregunta ¿Qué tan adecuado es el conjunto de entrenamiento para el conjunto que se desea clasificar (objetivo)?. Como menciona Mitchel (1997) “En general aprender es más confiable cuando los ejemplos siguen una distribución similar a los ejemplos de prueba futuros”. Sin embargo, con la clasificación translingüe hemos quebrantado, en cierto grado, dicha suposición en la cual está basada la tarea de clasificación. Incluso, en los conjuntos objetivo no existen muchas de las palabras que contienen los documentos de entrenamiento (Bel et al. 2003).

El clasificador utiliza palabras como atributos. El vocabulario del corpus de entrenamiento debería ser adecuado para representar los

documentos del corpus objetivo. Para comprobarlo se contabilizan los atributos y se hace una comparación. Sea $T_e = \{t_{e1}, \dots, t_{en}\}$ el conjunto de términos extraídos del conjunto de entrenamiento llamados vocabulario de entrenamiento, y sea $T_p = \{t_{p1}, \dots, t_{pm}\}$ el conjunto de términos extraídos del conjunto no etiquetado llamados vocabulario de prueba. Entonces el conjunto que comparten los corpus de entrenamiento y prueba $T_{com} = T_e \cap T_p$ es llamado vocabulario común. Es decir, si el término $t_{ex} \in T_e$ y $t_{ex} \notin T_p$, la columna correspondiente al término t_{ex} posee el valor 0 para todos los documentos del conjunto objetivo. Así mismo, cualquier $t_{py} \in T_p$ y $t_{py} \notin T_e$ no se considera por el clasificador.

La Tabla 4.8 muestra el tamaño de vocabulario en la clasificación monolingüe de referencia. Las dos primeras columnas indican el idioma del conjunto de entrenamiento y del conjunto objetivo respectivamente; la tercera presenta el número total de palabras diferentes con las que cuenta el corpus respectivo de entrenamiento. La cuarta columna indica el número de palabras diferentes del corpus objetivo y finalmente, la última columna indica cuantas de estas palabras se encuentran tanto en el corpus de entrenamiento como en el objetivo. De aquí se determina que, en la clasificación monolingüe de referencia, el vocabulario del conjunto de entrenamiento es de 12,419.66 palabras en promedio, el del conjunto objetivo de 8,322.33 y el vocabulario común llega a 5,544.66 palabras.

Tabla 4.8 Tamaño de vocabulario clasificación monolingüe.

Entrenamiento	Objetivo	T_e	T_p	T_{com}
Inglés	Inglés	10,892	7,658	5,452
Español	Español	12,295	8,051	5,182
Francés	Francés	14,072	9,258	6,000
	Promedio	12,419.66	8,322.33	5,544.66

La Tabla 4.9 muestra el tamaño de vocabulario en la clasificación TCE y la Tabla 4.10 muestra el tamaño del vocabulario en la clasificación TCP. Se observa que, al traducir, existen variaciones en el tamaño del vocabulario utilizado para representar el mismo conjunto de documentos. Por ejemplo, en la primera fila de la Tabla 4.9, la traducción del corpus de entrenamiento de inglés a español 'aumenta' el tamaño del vocabulario en 2,561 palabras, mientras que traducir de español o francés a inglés, indica una reducción del tamaño del vocabulario equivalente; por ejemplo, la quinta fila de la misma tabla indica que en la traducción del conjunto de entrenamiento de francés a inglés se 'decrementan' 2,734 palabras. En realidad, el tamaño del vocabulario no se decrementa o incrementa, es el resultado de representar documentos en otro idioma. Las variaciones de tamaño de vocabulario considerando la traducción de un idioma a otro simplemente denotan la complejidad de los idiomas, su capacidad de expresión con la riqueza de su vocabulario. Por ello, considerando todos los experimentos, el promedio de vocabulario de entrenamiento en la clasificación translingüe resulta similar al de la clasificación monolingüe.

Tabla 4.9 Tamaño de vocabulario clasificación TCE

Entrenamiento	Objetivo	T_e	T_p	T_{com}
Inglés Español	Español	13,453	8,051	3,640
Inglés Francés	Francés	12,426	9,258	4,131
Español Inglés	Inglés	9,012	7,658	3,351
Español Francés	Francés	10,666	9,258	3,793
Francés Inglés	Inglés	11,338	7,658	3,700
Francés Español	Español	14,684	8,051	3,920
	Promedio	11,929.83	8,322.33	3,755.83

Tabla 4.10 Tamaño de vocabulario clasificación TCP

Entrenamiento	Objetivo	T_e	T_p	T_{com}
Inglés	Español Inglés	10,892	6,314	3,295
Inglés	Francés Inglés	10,892	7,731	3,697
Español	Inglés Español	12,295	9,190	3,826
Español	Francés Español	12,295	9,398	3,925
Francés	Inglés Francés	14,071	8,428	4,194
Francés	Español Francés	14,071	7,049	3,749
	Promedio	12,419.33	8,018.33	3,781

Considerando los datos de ambos esquemas, mostrados en las Tablas 4.9 y 4.10, el promedio de vocabulario del conjunto de entrenamiento es de 12,174.58 palabras, el conjunto objetivo llega a ser de 8,170.33 palabras y el vocabulario común de sólo 3,768.41. Esto quiere decir que mientras el número de palabras existentes en los corpus de entrenamiento y el objetivo son similares en ambas clasificaciones, existe una reducción en el porcentaje de vocabulario común. El vocabulario común es aproximadamente 50% del vocabulario de entrenamiento en el caso monolingüe y de aproximadamente 31% en los casos translingües.

La Tabla 4.11 muestra la relación vocabulario común-exactitud en el caso monolingüe. La tercera columna indica el vocabulario común entre los conjuntos de entrenamiento y objetivo, la cuarta columna indica la exactitud reportada en esa clasificación. Las Tablas 4.12 y 4.13 de igual manera, muestran la relación de vocabulario común-exactitud en la clasificación TCE y la TCP respectivamente.

Tabla4.11 Relación vocabulario común-exactitud monolingüe

Entrenamiento	Objetivo	T_{com}	Exactitud
Inglés	Inglés	5,452	91.66%
Español	Español	5,182	91.66%
Francés	Francés	6,000	93.33%

Tabla 4.12 Relación vocabulario común-exactitud TCE

Entrenamiento	Objetivo	T_{com}	Exactitud
Inglés Español	Español	3,640	71.66%
Inglés Francés	Francés	4,131	75.83%
Español Inglés	Inglés	3,351	81.66%
Español Francés	Francés	3,793	80.83%
Francés Inglés	Inglés	3,700	85.83%
Francés Español	Español	3,920	83.33%

Tabla 4.13 Relación vocabulario común-exactitud TCP

Entrenamiento	Objetivo	T_{com}	Exactitud
Inglés	Español Inglés	3,295	75.00%
Inglés	Francés Inglés	3,697	76.66%
Español	Inglés Español	3,826	85.00%
Español	Francés Español	3,925	79.16%
Francés	Inglés Francés	4,194	86.66%
Francés	Español Francés	3,749	80.00%

Se observa la existencia de una relación entre la exactitud y la cantidad de vocabulario común. Sin embargo, esta relación no es directa ya que depende de la complejidad de cada corpus y no se puede establecer una comparación. Por ejemplo, la tercera fila del experimento de la Tabla 4.12, con entrenamiento en español-inglés y objetivo en inglés, tiene 3,351 palabras en común con una exactitud de 81.66%, mientras que en la primera fila el experimento con entrenamiento en inglés-español y objetivo en español cuenta con 3,640 palabras en común, es decir 289 palabras ’más’ y la exactitud es de 71.66%, 10 puntos menor. Podríamos suponer que siendo el español un idioma más complejo, el número de palabras en el experimento resulta insuficiente, mientras que en idioma inglés si resultan suficientes. El argumento anterior es posible, pero no se puede afirmar con los datos mostrados hasta ahora y no es objetivo del trabajo determinar la complejidad de los idiomas. En un intento por determinar la importancia del vocabulario

se hizo un análisis por ganancia de información, el cual se muestra en la siguiente sección.

4.4.3 Análisis por Ganancia de Información

En esta sección realizamos un experimento sencillo aplicando ganancia de información a la clasificación translingüe para acentuar los efectos de la discrepancia cultural. La ganancia de información (IG) como se describe en la sección 2.1.4, otorga un 'valor de importancia' a cada palabra del vocabulario. Usualmente es utilizada para conservar aquellos términos que son relevantes desechando los que no lo son. Utilizando ganancia de información podemos determinar, en promedio, el número de términos relevantes en el vocabulario común (Ver Anexo C).

La teoría indica que si tanto los conjuntos de entrenamiento y objetivo son tomados de la misma distribución, el aplicar un método de reducción de dimensionalidad, que conserve los atributos relevantes para la tarea, no reducirá el desempeño del clasificador. La Tabla 4.14 muestra una comparación de exactitudes al aplicar o no aplicar ganancia de información en el caso monolingüe. En la tercera columna se indica la exactitud de la clasificación monolingüe utilizando todo el vocabulario como atributos, la cuarta columna muestra la exactitud después de aplicar reducción de dimensionalidad con ganancia de información. Se observa que los resultados son cercanos.

Tabla 4.14 Comparación de exactitudes monolingües al reducir y no reducir vocabulario con IG.

Entrena	Objetivo	Exactitud sin reducción	Exactitud con reducción IG
Inglés	Inglés	91.66%	93.33%
Español	Español	91.66%	89.16%
Francés	Francés	93.33%	92.50%
	Promedio	92.22%	91.66%

Las Tablas 4.15 y 4.16 permiten hacer una comparación entre las exactitudes con y sin reducción de dimensionalidad por IG en la clasificación TCE y TCP respectivamente. En esta serie de experimentos, la reducción resulta contraproducente en la mayoría de los casos.

Tabla 4.15 Comparación de exactitudes TCE al reducir y no reducir con IG

Entrena	Objetivo	Exactitud sin reducción	Exactitud con reducción IG
Inglés Español	Español	71.66%	65.83%
Inglés Francés	Francés	75.83%	68.33%
Español Inglés	Inglés	81.66%	80.00%
Español Francés	Francés	80.83%	73.33%
Francés Inglés	Inglés	85.83%	79.16%
Francés Español	Español	83.33%	82.50%
	Promedio	79.86%	74.86%

Tabla 4.16 Comparación de exactitudes TCP al reducir y no reducir con IG

Entrena	Objetivo	Exactitud sin reducción	Exactitud con reducción IG
Inglés	Español Inglés	75.00%	69.16%
Inglés	Francés Inglés	76.66%	70.83%
Español	Inglés Español	85.00%	85.00%
Español	Francés Español	79.16%	80.00%
Francés	Inglés Francés	86.66%	80.83%
Francés	Español Francés	80.00%	76.66%
	Promedio	80.41%	77.08%

En la clasificación translingüe el promedio de exactitud es de 80.14% y con ganancia de información es de 75.97%. Experimentalmente la clasificación translingüe por sí misma produce una caída de 12 puntos porcentuales en promedio; reducir el número de atributos con ganancia de información le hace perder 16.24 puntos porcentuales en promedio. En consecuencia, se infiere que la reducción de dimesionalidad elimina palabras útiles para la clasificación translingüe.

En conclusión, existe una caída en promedio de 12 puntos porcentuales con respecto a la clasificación monolingüe. La clasificación translingüe por sí sola no es suficiente. Una posible solución consiste en incorporar conocimiento del conjunto objetivo a fin de mejorar la clasificación final. En este trabajo se proponen dos métodos con el fin de enfrentar las dificultades propias de la clasificación translingüe.

Capítulo 5

Métodos Propuestos para el Refinamiento de la Clasificación Translingüe

La clasificación translingüe, al utilizar recursos que no fueron creados para clasificar documentos de otro idioma, es susceptible a experimentar los efectos de las diferencias culturales, adicionales a los pormenores del proceso de traducción. El capítulo anterior muestra resultados de los análisis efectuados para visualizar dichos efectos. En general, la exactitud de la clasificación translingüe resulta menor a la obtenida de la clasificación monolingüe; en consecuencia, se proponen dos métodos a fin de mejorar los resultados. En este capítulo se explican los métodos propuestos, presentando los resultados obtenidos y comparándolos con los resultados de referencia.

5.1 Métodos Propuestos

Existen dos observaciones en cuanto a la clasificación translingüe simple. La primera es inherente a la tarea ya que, en la clasificación translingüe, el conjunto de entrenamiento y objetivo no son tomados de la

misma distribución. Aunque se espera que las distribuciones sean similares y compartan características, no se asegura el éxito de la clasificación. Sin embargo, se cuenta con la ventaja de que el propio conjunto objetivo es una muestra de la distribución que nos interesa captar.

La segunda observación es aplicable a la mayoría de los métodos de clasificación. El clasificador simple calcula la probabilidad de cada documento de pertenecer a una clase, utilizando para ello los atributos con los cuales fue representado el documento. Esto quiere decir que en esta forma de realizar la clasificación, ésta se realiza de forma individual sin considerar el resto de los documentos a clasificar y la decisión del clasificador se basa en los atributos del conjunto de entrenamiento, ignorando el resto de las propiedades del documento a clasificar.

Con el objetivo de aprovechar información propia del conjunto objetivo, se desarrollaron dos métodos. En el primero, se busca realizar un proceso de refinamiento, sin afectar al clasificador, a través de una corrección de clasificación mediante vecinos más cercanos del conjunto objetivo únicamente. En el segundo, se incorporan ejemplos confiables del conjunto de documentos no etiquetado al clasificador, con el fin de especializar gradualmente al clasificador a los documentos del idioma objetivo.

5.1.1 Corpus de Trabajo

Se utilizó el corpus explicado en la sección 4.1.3 conformado por un subconjunto de noticias de México, E.U y Francia de las clases Crimen, Desastres, Política y Deportes con 200 noticias en cada idioma para entrenamiento y 120 como objetivo.

5.1.2 Software Utilizado

El software empleado durante el desarrollo de este trabajo incluye:

1. Un traductor automático, llamado Wordlingo. Este programa recibe un texto escrito en un idioma determinado y produce un texto en el idioma especificado. Fue obtenido de www.wordlingo.com.

2. Un módulo de pre procesamiento de documentos. Este software desarrollado en Perl limpia los documentos de caracteres extraños, elimina etiquetas, números, palabras vacías entre otros. La serie de procesos que pueden conformar el pre procesamiento se explican en la sección 2.2.1.

3. Un módulo para representar documentos y asignar pesos. Este software fue programado en Perl. Se utiliza para llevar a una representación vectorial el conjunto de documentos con el que se desea trabajar, dando el formato necesario para ser manejado por el algoritmo de aprendizaje computacional. La sección 2.2.2 y 2.2.3 abordan los conceptos de representación de documentos.

4. El software Weka es una colección de algoritmos de Aprendizaje Computacional desarrollado por la Universidad de Waikato. Se utilizó la versión 3.5.7, programada en java y fue obtenida de www.cs.waikato.ac.nz/ml/weka/.

5. Módulo para obtención de vecinos más cercanos, se encarga de obtener la similitud (sección 2.2.5) del conjunto de documentos enlistando los vecinos más cercanos de cada documento. Este software fue programado en Perl y es utilizado por los métodos descritos en las secciones 5.2 y 5.3.

6. Módulo de refinamiento de la clasificación, se encarga de re etiquetar cada documento con una nueva clase, es utilizado por el método descrito en la sección 5.2, fue desarrollado en Perl.

7. Módulo de selección de confiables, se encarga de identificar documentos confiables, se utiliza en el método descrito en la sección 5.3 y fue programado en Perl.

8. El software para la obtención de las gráficas de similitud de los conjuntos de documentos, sección 4.3.1, fue desarrollado en Matlab R2007a, recibe un archivo de texto con la matriz de similitud resultante del módulo de obtención de vecinos más cercanos y genera con ella las gráficas en escala de grises.

Los tiempos requeridos para efectuar los métodos planteados en el presente trabajo varían de acuerdo al tamaño del corpus. Dado que proceso de traducción no es totalmente automatizado y utiliza una herramienta libre disponible en la web, para el corpus utilizado en este trabajo se requirió aproximadamente un día para cada idioma y el trabajo debe ser corroborado ante la posible falta de traducción. El módulo de representación de documentos puede tomar unos cuantos minutos dependiendo del tamaño del corpus y el resto de los programas consume tan sólo algunos segundos. Sin embargo, algunas etapas requieren más de un programa y existen fases manuales entre ellos. Una vez construido el clasificador, dependiendo del número de iteraciones y considerando la aplicación correcta de la secuencia de programas, se pueden obtener resultados de cualquiera de los dos métodos en aproximadamente 4 horas. Los experimentos fueron realizados en una Pentium M a 1.70 GHz a 0.98Gb de RAM.

5.1.3 Resultados de la Clasificación Translingüe Simple

Como ya se mencionó, los investigadores comparan los resultados de sus métodos con una referencia monolingüe, la cual se obtiene del mismo corpus de trabajo. Adicionalmente, en este trabajo se obtuvo la pérdida relativa de la clasificación translingüe con respecto a la monolingüe, la cual se calcula con la ecuación 5.1.

$$Pérdida\ relativa = 100 - \frac{Exactitud\ Translingüe\ \times 100}{Exactitud\ monolingüe} \quad \textbf{(5.1)}$$

La Tabla 5.1 presenta los resultados de la clasificación monolingüe, mientras que las Tablas 5.2 y 5.3 muestran los resultados de la clasificación translingüe simple traduciendo el corpus de entrenamiento (TCE) y traduciendo el corpus objetivo (TCP) respectivamente. La cuarta columna de dichas tablas muestra el porcentaje de pérdida con respecto a la clasificación monolingüe.

Tabla 5.1 Resultados de la clasificación monolingüe

Entrenamiento	Objetivo	Exactitud
Inglés	Inglés	91.66%
Español	Español	91.66%
Francés	Francés	93.33%

Tabla 5.2 Resultados de la clasificación translingüe TCE

Entrenamiento	Objetivo	Exactitud	Pérdida relativa
Inglés Español	Español	71.66%	21.82%
Inglés Francés	Francés	75.83%	17.27%
Español Inglés	Inglés	81.66%	10.90%
Español Francés	Francés	80.83%	11.81%
Francés Inglés	Inglés	85.83%	08.06%
Francés Español	Español	83.33%	10.71%

Tabla 5.3 Resultados de la clasificación translingüe TCP

Entrenamiento	Objetivo	Exactitud	Pérdida relativa
Inglés	Español Inglés	75.00%	18.17%
Inglés	Francés Inglés	76.66%	16.36%
Español	Inglés Español	85.00%	07.26%
Español	Francés Español	79.16%	13.63%
Francés	Inglés Francés	86.66%	07.14%
Francés	Español Francés	80.00%	14.28%

Recapitulando, el traductor utilizado en todos los experimentos es Worldlingo y se usa Naive Bayes como clasificador. En promedio la clasificación monolingüe alcanza 92.22% de exactitud mientras que la clasificación translingüe simple llega a 80.13% considerando los resultados de ambos esquemas. El porcentaje de pérdida de la clasificación translingüe llega al 21.82% en comparación con la monolingüe, siendo el promedio 13.11%. A continuación se explica el primer método propuesto.

5.2 Primer Método: Refinación de la Clasificación Translingüe mediante Vecinos más Cercanos

El método surge de aprovechar la similitud entre documentos. Se espera que los documentos pertenecientes a la misma clase sean similares entre ellos y no tan similares a documentos de otras clases. Además, se aprovecha la capacidad del clasificador translingüe simple de realizar una clasificación inicial apoyándose en el vocabulario común entre los conjuntos de entrenamiento y objetivo.

La Figura 5.1 muestra el método propuesto para el refinamiento de la clasificación translingüe por medio de vecinos más cercanos. Este método se enfoca en el conjunto objetivo previa clasificación inicial, sin intentar

modificar al clasificador original. En la figura se aprecia el conjunto de documentos no etiquetados en el idioma objetivo D_P, que es utilizado para dos bloques. El primero es la clasificación translingüe respaldada por el conjunto etiquetado en el idioma fuente. Cabe mencionar que puede utilizarse cualquiera de los dos esquemas, incluso, cualquier método de clasificación trasnlingüe de documentos capaz de asignar una clasificación inicial. El segundo proceso, independiente del anterior, consiste en obtener

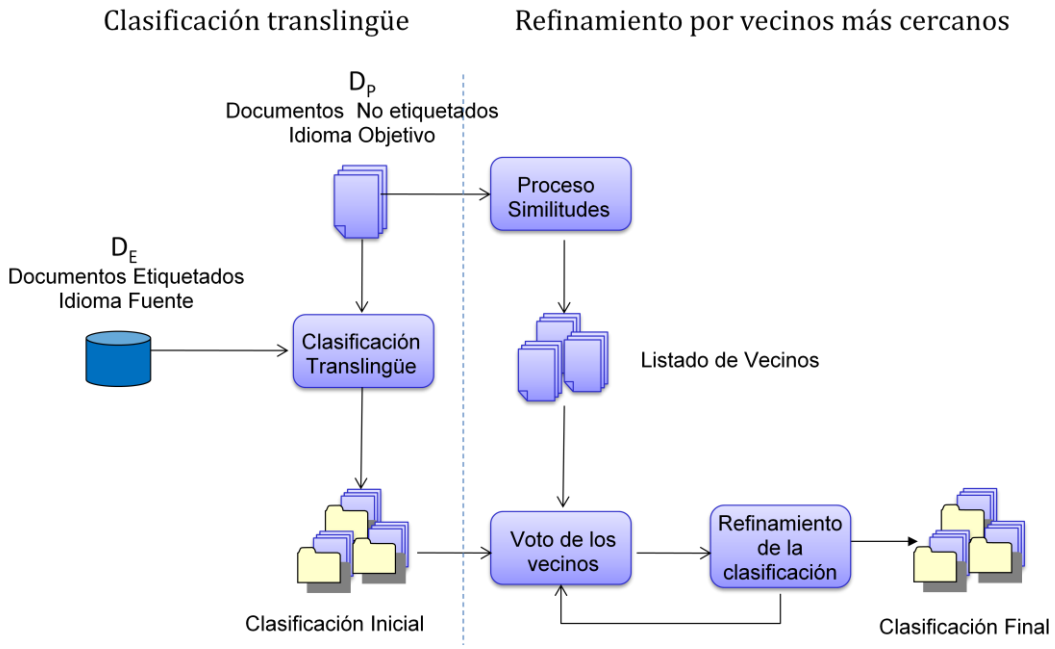


Figura 5.1 Método propuesto de refinamiento de la clasificación mediante vecinos.

las similitudes entre documentos de todo el conjunto objetivo utilizando su propio vocabulario. De aquí se deriva un listado de vecinos más cercanos para cada documento. Con la lista de vecinos más cercanos y las clases asignadas de forma inicial por el clasificador translingüe, se realiza un voto para cada documento, dicho voto es considerado en el siguiente bloque a fin

de cambiar o mantener la clase de cada documento. Las etiquetas se actualizan en el bloque llamado refinamiento de la clasificación para volver al proceso de voto, éstos dos últimos procesos se efectúan iterativamente hasta que no existen cambios en las etiquetas. El resultado es considerado como la clasificación final.

Se observa que el proceso de clasificación translingüe simple se realiza una sola vez, al igual que el proceso de similitudes y listado de vecinos. La parte iterativa del método se centra en la votación de los vecinos y su re-etiquetación. El método se describe en el siguiente algoritmo:

1. Se construye un clasificador C_1 utilizando el corpus de entrenamiento D_E en el idioma fuente L_1.
2. Se clasifica cada documento $d_j \in D_p$ en el idioma objetivo L_2 utilizando cualquiera de los dos esquemas conocidos para la clasificación translingüe a fin de asignar una clasificación inicial, la cual se representa como c_j^0 donde el superíndice indica el número de iteración en la que fue asignada la clase.
3. Se representa al conjunto objetivo D_p con su propio vocabulario $T_p = \{t_{p1}, \dots, t_{pm}\}$ y se enlistan los k vecinos más cercanos $d_{nn1}, \dots, d_{nnk}$ de cada documento.
4. Sea c_j^n la clase asignada por el clasificador al documento d_j en la iteración *n* y $c_{nn1}^n \dots c_{nnk}^n$ las clases de sus k vecinos. Entonces $c_j^{n+1} = c_{nn1}^n$ si $c_{nn1}^n = c_{nn2}^n = \cdots = c_{nnk}^n$, re-clasificando al documento con la clase de sus vecinos y $c_j^{n+1} = c_j^n$ si no se cumple que $c_{nn1}^n = c_{nn2}^n = \cdots = c_{nnk}^n$.
5. Se actualizan las etiquetas de cada documento.

6. El paso 4 y 5 se repiten iterativamente hasta que no existen cambios, es decir $\forall (d_j \in D_p): c_j^{n+1} = c_j^n$ o hasta un número fijo de iteraciones.

A continuación se presentan los resultados experimentales de aplicar el método de refinamiento de la clasificación mediante vecinos más cercanos a los resultados obtenidos de ambos esquemas de clasificación translingüe.

5.2.1 Resultados de aplicar el Primer Método

Los resultados de la clasificación translingüe simple son utilizados como clasificación inicial; se utiliza el *coeficiente de Dice* (ecuación 2.6) como medida de similitud para enlistar los k vecinos más cercanos de cada documento y exactitud (ecuación 2.17) como medida de evaluación. La Tabla 5.4 contiene los resultados experimentales de aplicar el primer método a la clasificación translingüe TCE. Las primeras dos columnas muestran los idiomas de entrenamiento y objetivo; la tercera indica la exactitud de la clasificación inicial, las siguientes muestran los resultados finales de la aplicación del primer método con 3, 4 y 5 vecinos respectivamente.

Tabla 5.4 Refinamiento mediante vecinos más cercanos TCE

Entrenamiento	Objetivo	Exactitud			
		Inicial	3 Vecinos	4 Vecinos	5 Vecinos
Inglés Español	Español	71.66%	72.50%	73.33%	72.50%
Inglés Francés	Francés	75.83%	77.50%	76.66%	76.66%
Español Inglés	Inglés	81.66%	90.00%	90.00%	88.33%
Español Francés	Francés	80.83%	83.33%	81.66%	82.5%
Francés Inglés	Inglés	85.83%	95.83%	92.50%	92.5%
Francés Español	Español	83.33%	84.16%	84.16%	84.16%

En este grupo de experimentos, siempre hubo una ventaja al aplicar el método. El menor de los incrementos es de aproximadamente 1% respecto a

la clasificación inicial, el más alto la supera el 11.65%. En promedio, la clasificación inicial se mejora 4.11% considerando todos los experimentos. Por número de vecinos, al aplicar el método con 3 vecinos el incremento fue de 4.88%, con 4 vecinos de 3.9% y con 5 vecinos de 3.54% respecto a la primera columna. La serie de experimentos con menor crecimiento corresponde a aquellos con entrenamiento Francés-Español y la serie de mayor crecimiento corresponde a los de entrenamiento Francés-Inglés.

La Tabla 5.5 muestra los resultados de aplicar el método a la clasificación translingüe TCP. A diferencia de la tabla anterior, existe un caso con un decremento en la exactitud cercano al 1% con respecto a la clasificación inicial. Con resultados más bajos en general comparados con los anteriores. El incremento más alto fue de 6.73% sobre la clasificación inicial. El promedio de incremento general es de 2.25% considerando que la serie de experimentos con objetivo Español-Inglés no muestran variación. La serie de experimentos con mayor crecimiento son los de objetivo Inglés-Español con 5.54% en promedio.

Tabla 5.5 Refinamiento mediante vecinos más cercanos TCP

Entrenamiento	Objetivo	Exactitud			
		Inicial	3 Vecinos	4 Vecinos	5 Vecinos
Inglés	Español Inglés	75.00%	75.00%	75.00%	75.00%
Inglés	Francés Inglés	76.66%	75.83%	76.66%	76.66%
Español	Inglés Español	85.00%	90.83%	89.16%	89.16%
Español	Francés Español	79.16%	80.83%	80.83%	81.66%
Francés	Inglés Francés	86.66%	92.50%	89.16%	89.16%
Francés	Español Francés	80.00%	81.66%	80.83%	81.66%

Considerando ambos esquemas, se observa que la exactitud se incrementa en promedio 3.18%. En el mejor de los casos se aumenta la clasificación inicial 11.65% (Entrena con Francés-Inglés y objetivo en Inglés a

3 Vecinos) y sólo en un caso se disminuye 1% (Entrena con Inglés, objetivo con Francés-Inglés a 3 Vecinos). En general se obtienen mejores resultados al usar 3 vecinos. Para hacer mejor interpretación de los resultados presentados, en la siguiente sección se considera el número de cambios de etiqueta efectuados por el método y se presentan las gráficas de comportamiento.

5.2.2 Gráficas de Comportamiento para el Primer Método

En esta sección se muestran las gráficas de comportamiento de cada experimento al aplicar el método de refinamiento de la clasificación translingüe mediante vecinos más cercanos. Se contabiliza el número total de cambios de etiqueta efectuados por el método en cada iteración. Para facilitar su visualización se elaboraron gráficas por tipo de esquema y por número de vecinos. El mayor número de cambios se reporta en la Figura 5.2 con 23 cambios en la primera iteración.

La Figura 5.2 muestra el comportamiento para la serie de experimentos de la clasificación TCE con 3 vecinos. Como se puede observar, en la primera iteración todos los experimentos reportaron por lo menos 3 cambios. Resulta interesante que el experimento con mayor número de cambios, entrenamiento Inglés-Español con clasificación inicial de 71.66% de exactitud, alcanzara solamente 72.50%. En cambio, el experimento con entrenamiento Francés-Inglés, el cual también hace un número importante de cambios, es el que obtiene mayor crecimiento con 10 puntos al cambiar de 85.83% a 95.83%. De lo anterior se infiere que no existe relación directa entre el número de cambios efectuados y la exactitud.

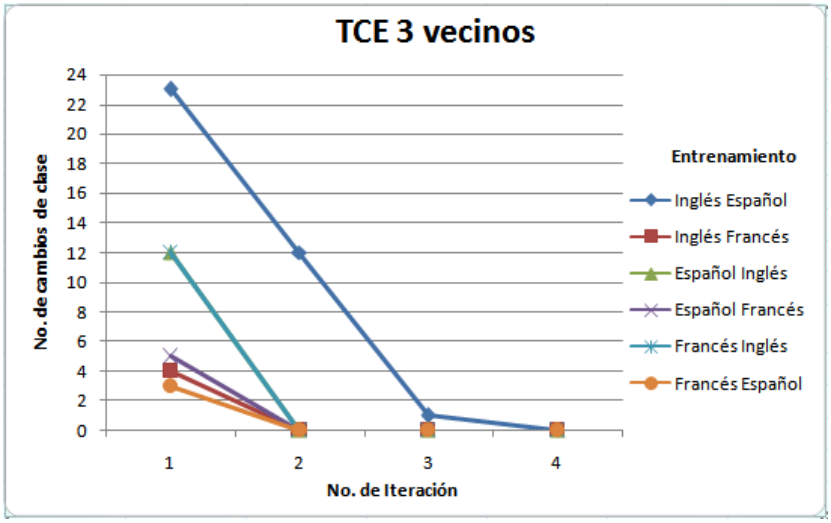


Figura 5.2 Cambios realizados por el primer método TCE 3 vecinos

Las Figuras 5.3 y 5.4 muestran el número de cambios de etiqueta realizados por el método en la clasificación TCE con 4 y 5 vecinos respectivamente. Se observa que a mayor número de vecinos, menor número de cambios efectuados.

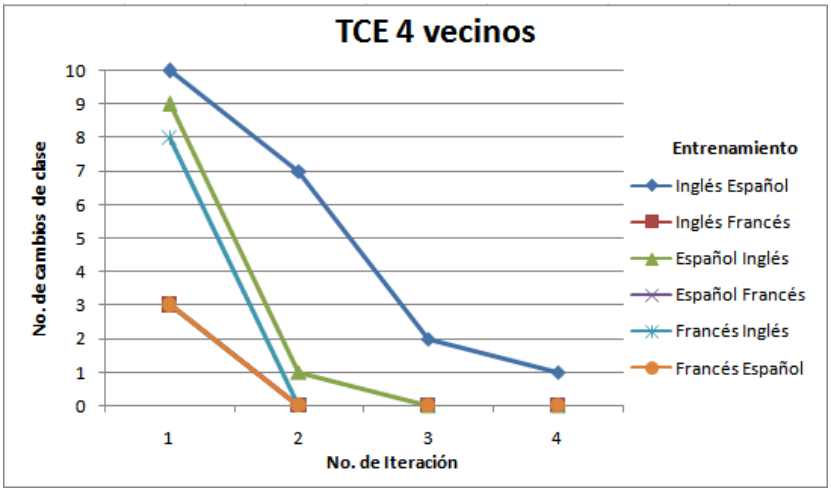


Figura 5.3 Cambios realizados por el primer método TCE 4 vecinos

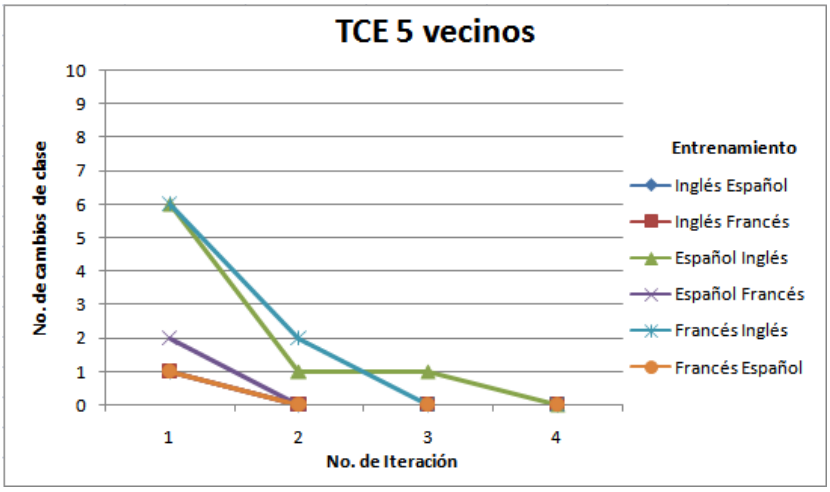


Figura 5.4 Cambios realizados por el primer método TCE 5 vecinos

En general el comportamiento de los experimentos es similar. Del total de 18 experimentos, 5 experimentos decrementan el número de cambios en cada iteración hasta detenerse y 13 realizan cambios en la primera iteración convergiendo en la segunda.

La Figura 5.5 muestra el comportamiento del método en cuanto a número de cambios de etiqueta efectuados en cada iteración para la serie de experimentos de la clasificación TCP con 3 vecinos. El experimento con objetivo Español-Inglés no realizó ningún cambio en las iteraciones, ello explica que la exactitud se mantuviera en 75%.

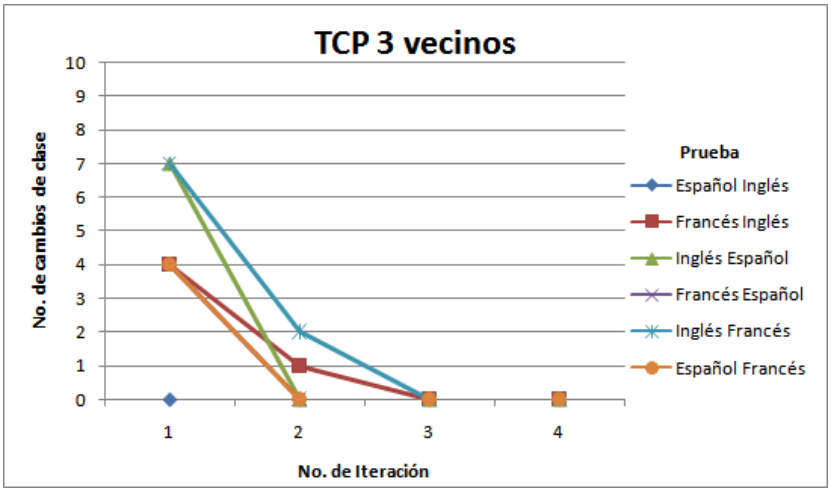


Figura 5.5 Cambios realizados por el primer método TCP 3 vecinos

Las Figuras 5.6 y 5.7 son equivalentes a la anterior para la serie de experimentos de la clasificación TCP con 4 y 5 vecinos respectivamente. Se observan algunas similitudes, por ejemplo, ambas muestran cambios sólo en la primera iteración. Nótese, sin embargo, que en la gráfica de 5 vecinos, son 3 los experimentos sin cambios y en la gráfica de 4 vecinos sólo 2 no reportan cambios. Considere que el enfoque de 5 vecinos es más restrictivo que el de 4, la mayoría de los vecinos que cambian con 5 vecinos también lo harían para 4.

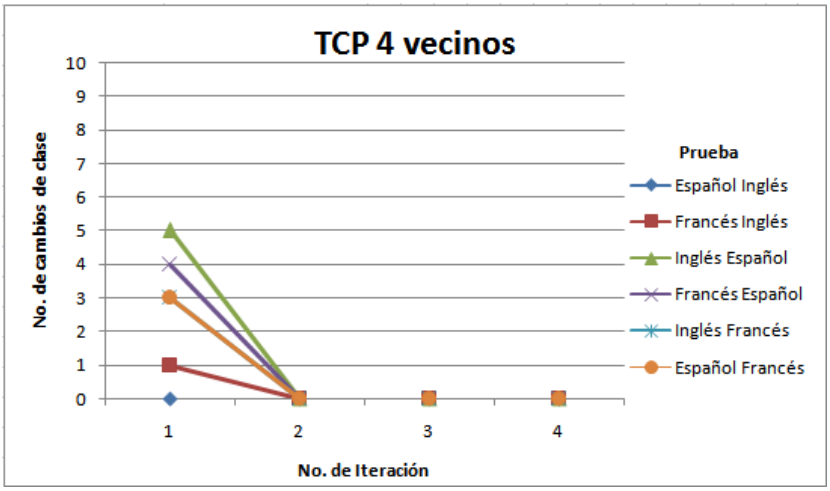


Figura 5.6 Cambios realizados por el primer método TCE 4 vecinos

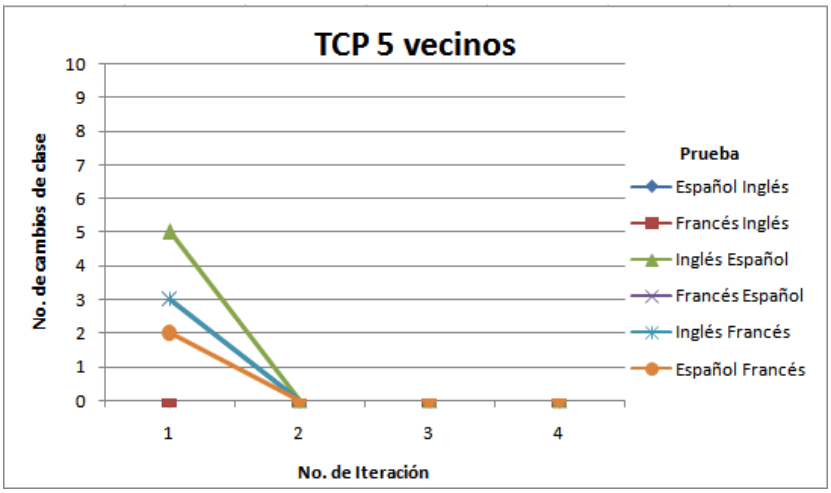


Figura 5.7 Cambios realizados por el primer método TCE 5 vecinos

Sobre el comportamiento observado, se determina que de los 18 experimentos, 2 decrementan el número de cambios en cada iteración, 12 realizan cambios en la primera iteración y convergen en la segunda, y 4 se mantienen sin cambios. En consecuencia, considerando ambos esquemas, alrededor del 70% de los casos converge en la segunda iteración.

5.2.3 Comentarios Finales del Primer Método

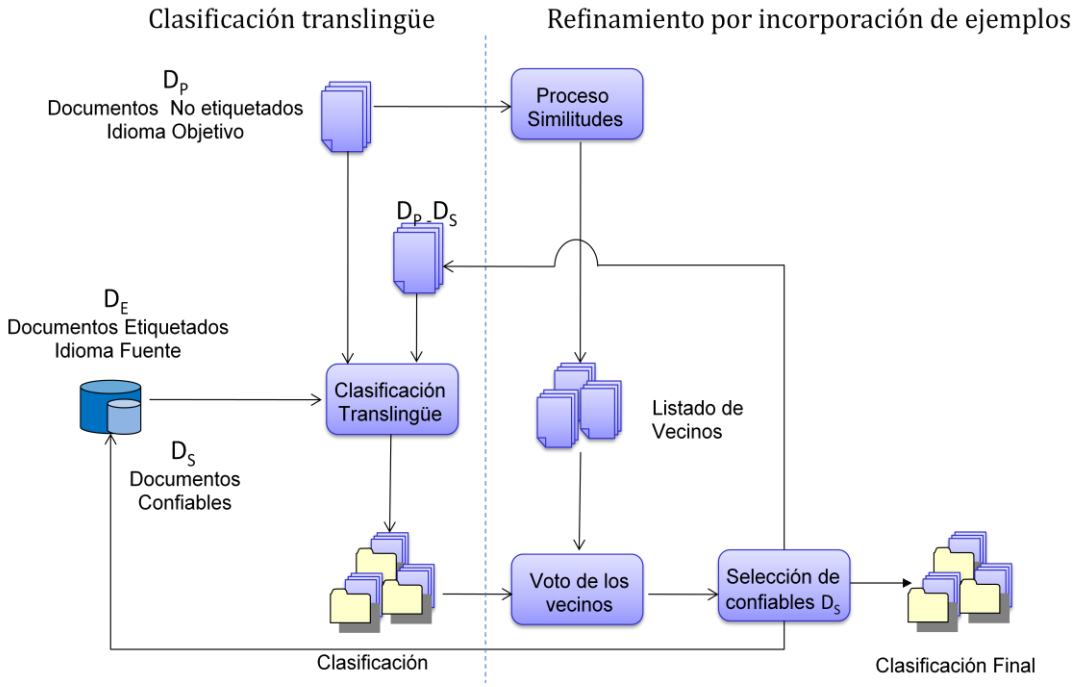
Aplicar el método de refinamiento de la clasificación translingüe mediante vecinos más cercanos demuestra mejoría en el 83% de los experimentos presentados. De ese porcentaje, aproximadamente 34% de los experimentos logran un incremento en la exactitud superior al promedio, el cual se fija en 3.18% respecto a la clasificación inicial para ambos esquemas. Además, para el 91.66% de los experimentos, ya no existen cambios en la primera o segunda iteración.

5.3 Segundo Método: Refinación de la Clasificación Translingüe mediante Incorporación de Ejemplos

Este método tiene como objetivo identificar documentos clasificados de forma confiable para incorporarlos al conjunto de entrenamiento y con ello incorporar vocabulario adecuado para discriminar los documentos del idioma objetivo. Un documento se considera confiable cuando a sus vecinos y a él mismo se les ha asignado una clase idéntica. Por lo tanto, de la misma forma que el método anterior, se utilizan vecinos más cercanos y la clasificación inicial, en ésta ocasión para identificar documentos confiables.

La Figura 5.8 ilustra el método para el refinamiento de la clasificación translingüe mediante incorporación de ejemplos. Este método crea clasificadores iterativamente para adaptarse a los documentos del idioma objetivo. En la figura se aprecia que el conjunto de documentos no etiquetados DP participa en dos procesos; el de similitud proporciona una lista no modificable de vecinos y se realiza una sola vez; y el de clasificación translingüe, usando cualquiera de los dos esquemas, asigna las clases en cada iteración. Utilizando la lista de vecinos y la clase asignada se ejerce el

voto de confianza. Probablemente existan clases con más documentos confiables que otras, en consecuencia, los conjuntos por clase tendrán diferentes cardinalidades. Considerando lo anterior, en el bloque de selección de confiables se eligen aleatoriamente el mismo número de documentos confiables para cada clase con el objetivo de mantener el conjunto de entrenamiento balanceado. Una vez seleccionados los



(a) Etapa iterativa

Figura 5.8 Método propuesto de refinamiento de la clasificación mediante incorporación de ejemplos

documentos confiables se integran al conjunto de entrenamiento con la clase asignada por el clasificador, misma que no se modificará en iteraciones siguientes. Una vez definido el nuevo corpus de entrenamiento, se entrena un nuevo clasificador para asignar clases al resto de los documentos

objetivo. Con esto iterativamente se crean clasificadores que incorporan vocabulario del conjunto objetivo. El criterio de paro se da cuando no existen más ejemplos confiables para incorporar. Considerando las últimas etiquetas como la clasificación final.

A diferencia del método anterior son varios procesos los que se repiten en cada iteración, incluyendo aquellos propios a la construcción del clasificador translingüe. El método se describe con el siguiente algoritmo:

1. Se representa al conjunto objetivo D_p con su propio vocabulario $T_p = \{t_{p1}, \dots, t_{pm}\}$ y se enlistan los k vecinos más cercanos $d_{nn1}, \dots, d_{nnk}$ de cada documento, $\mathrm{D}_S^0 = \emptyset$.
2. Se construye un clasificador C_1 utilizando el corpus de entrenamiento D_P^n en el idioma fuente L_1 donde el superíndice indica el número de iteración en la cual el conjunto es válido.
3. Se clasifica cada documento $d_j \in \mathrm{D}_P^n - \mathrm{D}_S^{n-1}$ en el idioma objetivo L_2 utilizando cualquiera de las dos esquemas conocidos para la clasificación translingüe a fin de asignar una clasificación, la cual se representa como c_j^n.
4. Sea c_j^n la clase asignada por el clasificador al documento d_j y $c_{nn1}^n \dots c_{nnk}^n$ las clases de sus k vecinos. El documento se considera confiable si $c_j^n = c_{nn1}^n = c_{nn2}^n = \cdots = c_{nnk}^n$.
5. Sean $\mathrm{D}_1^n, \mathrm{D}_2^n, \dots \mathrm{D}_m^n$ los conjuntos de las m clases que contienen los documentos seleccionados como confiables de cada clase. Entonces se define *x = min* $\{ |\mathrm{D}_1^n| , |\mathrm{D}_2^n| , \dots |\mathrm{D}_m^n| \}$. Se seleccionan aleatoriamente x documentos de cada conjunto, creando el conjunto homogéneo D_s^n.

6. El conjunto de entrenamiento se redefine como $D_E^{n+1} = D_E^n \cup D_S^n$ incorporando con ello los documentos confiables del corpus objetivo.
7. El algoritmo se repite iterativamente desde el paso 2 hasta que no existe el mínimo de documentos confiables en cada clase, es decir $D_S^n = \emptyset$ o hasta un número fijo de iteraciones.

A continuación se presentan los resultados experimentales de aplicar el método de refinamiento de la clasificación mediante incorporación de ejemplos en ambos esquemas de clasificación translingüe.

5.3.1 Resultados de aplicar el Segundo Método

Los resultados de la clasificación translingüe simple son utilizados como clasificación inicial, en congruencia con el trabajo previo, se utiliza Naive Bayes como clasificador, el *coeficiente de Dice* como medida de similitud para enlistar los k vecinos más cercanos de cada documento y la exactitud como medida de evaluación (ec. 2.17). La Tabla 5.6 muestra los resultados experimentales de aplicar el segundo método en la clasificación TCE. En la tercera columna se indica la exactitud de la clasificación inicial, las siguientes muestran los resultados alcanzados al aplicar el método de incorporación de ejemplos con 3, 4 y 5 vecinos respectivamente.

Tabla 5.6 Refinamiento mediante la incorporación de ejemplos TCE

Entrenamiento	Objetivo	Exactitud			
		Inicial	3 Vecinos	4 Vecinos	5 Vecinos
Inglés Español	Español	71.66%	84.16%	80.83%	70.83%
Inglés Francés	Francés	75.83%	75.83%	75.83%	75.83%
Español Inglés	Inglés	81.66%	91.66%	91.66%	88.33%
Español Francés	Francés	80.83%	80.00%	85.83%	81.66%
Francés Inglés	Inglés	85.83%	92.5%	92.5%	93.33%
Francés Español	Español	83.33%	85.33%	85.33%	85.33%

En este grupo de experimentos, el promedio de incremento es de 5.52%, donde el experimento con entrenamiento Inglés-Español alcanza un incremento de 17.44% respecto a la clasificación inicial. Dos experimentos resultan en un decremento de aproximadamente el 1%. No se observaron cambios en la serie de experimentos con entrenamiento en Inglés-Francés.

La Tabla 5.7 muestra los resultados de aplicar el segundo método en la clasificación translingüe TCP, en donde el mayor incremento es de 13.33% respecto a la clasificación inicial. En los resultados más bajos se consideran 4 experimentos, los cuales no tienen crecimiento por haber permanecido sin variación. El promedio de incremento es de 6.04%.

Tabla 5.7 Refinamiento mediante incorporación de ejemplos TCP

Entrenamiento	Objetivo	Exactitud			
		Inicial	3 Vecinos	4 Vecinos	5 Vecinos
Inglés	Español Inglés	75.00%	85.00%	85.00%	84.16%
Inglés	Francés Inglés	76.66%	76.66%	76.66%	76.66%
Español	Inglés Español	85.00%	90.00%	90.83%	91.66%
Español	Francés Español	79.16%	85.00%	80.00%	79.16%
Francés	Inglés Francés	86.66%	93.33%	93.33%	92.5%
Francés	Español Francés	80.00%	84.16%	85.00%	85.83%

El promedio de incremento general considerando ambos esquemas es de 5.78%, dónde el mejor caso fue de 17.44%. El experimento con entrenamiento en Inglés-Español y objetivo en Español a 5 vecinos y el experimento con entrenamiento Español-Francés y objetivo en Francés a 3 vecinos son los dos casos, ambos de la Tabla 5.6, en los cuales disminuyó la exactitud, la pérdida fue de 0.83 puntos. Nuevamente la columna de 3 vecinos muestra exactitudes más altas que el resto de los experimentos, aumentando 6.52% en promedio la clasificación inicial. A fin de observar el

comportamiento del método, se presentan las gráficas de comportamiento en la siguiente sección.

5.3.2 Gráficas de Comportamiento para el Segundo Método

Las gráficas presentadas a continuación son diferentes a las mostradas en la sección anterior debido a que no reflejan el total de cambios, sino el número de ejemplos incorporados por clase. Es decir, cada valor presentado en la figura se multiplica por 4 para obtener el número total de documentos incorporados porque en este trabajo se están considerando 4 clases. Las gráficas se dividen por esquema y número de vecinos para facilitar su visualización.

Las Figuras 5.9, 5.10 y 5.11 muestran en la clasificación TCE el número de documentos confiables por clase que se incorporan en cada iteración por la aplicación del segundo método con 3, 4 y 5 vecinos respectivamente. Los experimentos con el conjunto de entrenamiento en Inglés-Francés y objetivo Francés, y con el conjunto de entrenamiento en Inglés y objetivo Francés-Inglés, no presentaron variaciones debido a que dichos experimentos no seleccionaron documentos confiables; en consecuencia, se infiere que por lo menos en una clase no existían confiables para incorporar. Como observación los dos casos que disminuyeron su exactitud, conjunto de entrenamiento en Inglés-Español y objetivo Español con 5 vecinos y conjunto de entrenamiento en Español-Francés objetivo Francés con 3 vecinos, fueron los únicos experimentos que integraron un solo documento de cada clase y detuvieron su clasificación.

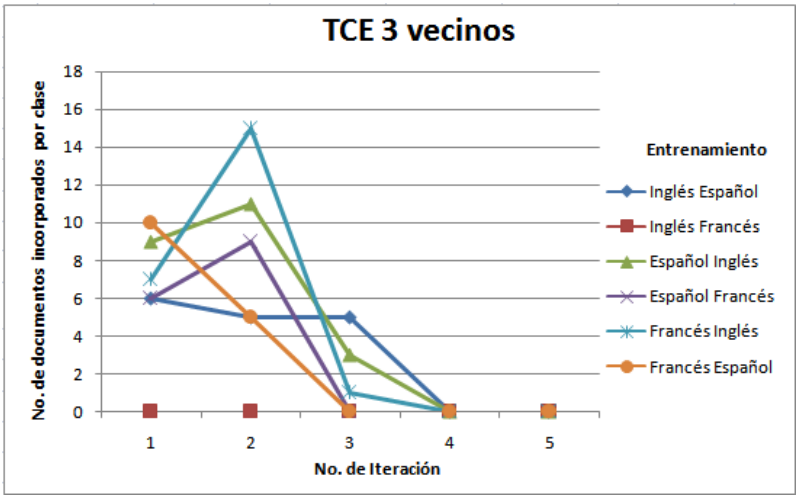


Figura 5.9 Confiables incorporados por clase TCE 3 vecinos

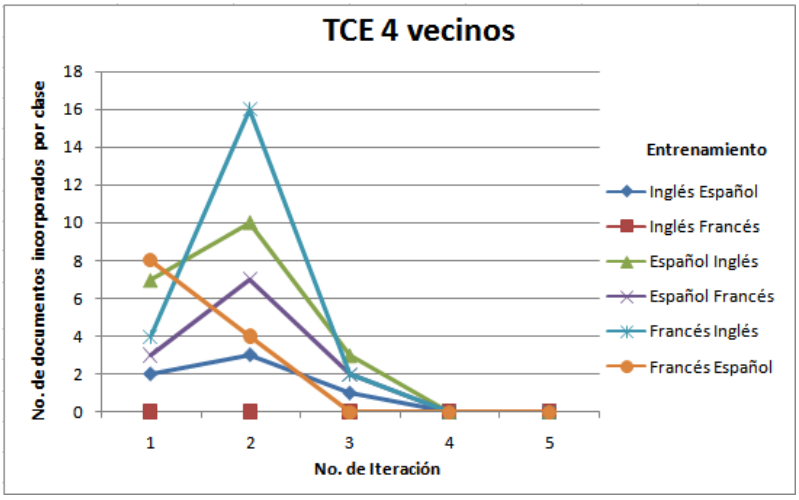


Figura 5.10 Confiables incorporados por clase TCE 4 vecinos

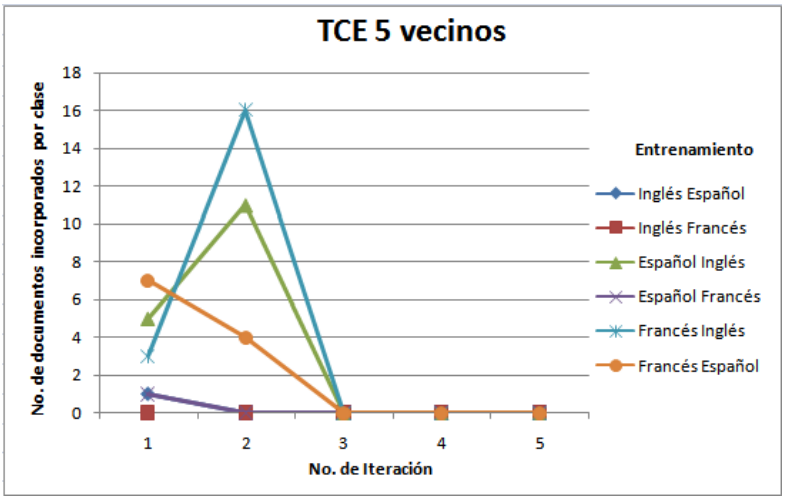


Figura 5.11 Confiables incorporados por clase TCE 5 vecinos

Nótese que el comportamiento del método es diferente al anterior. Se observa una tendencia general a incorporar más ejemplos en la segunda iteración que en la primera. De los 18 experimentos de este esquema, 9 incorporan más ejemplos en la segunda iteración y luego decrecen en iteraciones siguientes hasta detenerse, 4 integran menos documentos en cada iteración, 2 sólo incorporan documentos en la primera iteración deteniéndose en la segunda y finalmente 3 experimentos no reportan cambios.

Las Figuras 5.12, 5.13 y 5.14, equivalentes a las anteriores, muestran el número de documentos por clase que se incorporan al corpus de entrenamiento en cada iteración. Se obtienen de aplicar el segundo método a la clasificación translingüe TCP con 3, 4 y 5 vecinos respectivamente.

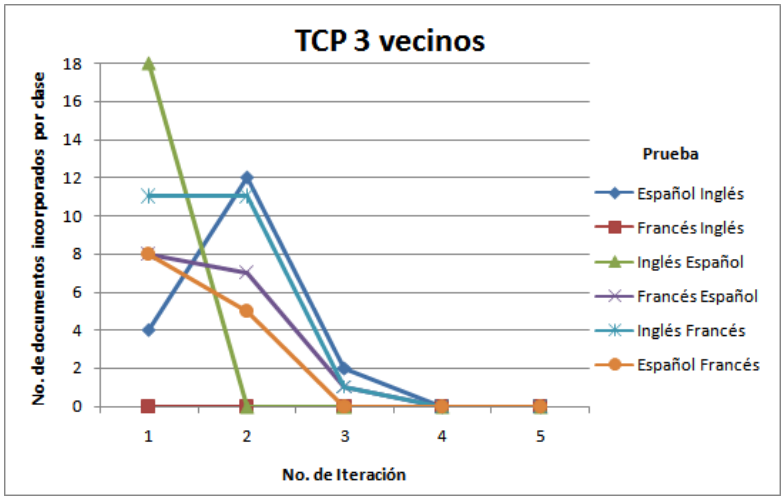


Figura 5.12 Confiables incorporados por clase TCP 3 vecinos

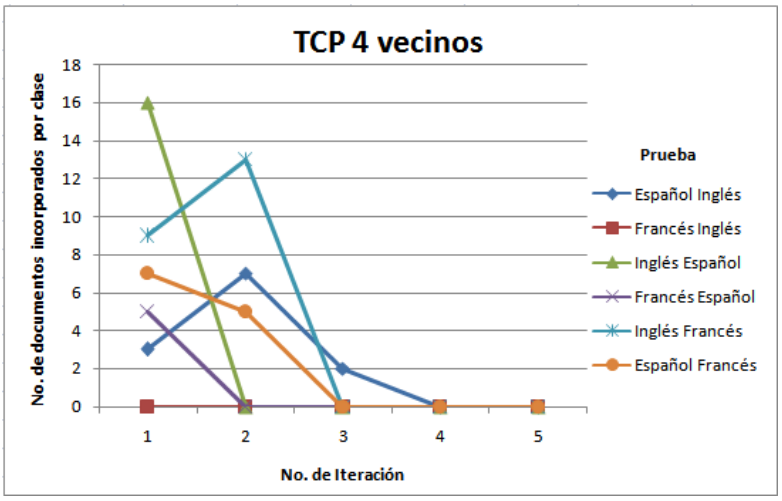


Figura 5.13 Confiables incorporados por clase TCP 4 vecinos

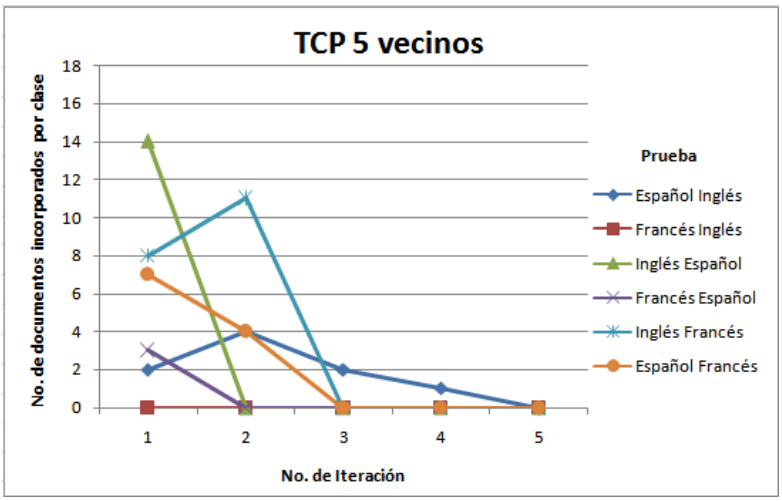


Figura 5.14 Confiables incorporados por clase TCP 5 vecinos

Sobre el comportamiento observado en este esquema, se determina que de los 18 experimentos, 6 aumentan el número de documentos incorporados en la segunda iteración, 4 incrementan menor número de documentos en cada iteración, 5 incrementan sólo en el primer ciclo convergiendo en el segundo y 3 no reportan cambios.

Considerando ambos esquemas, alrededor de 56% de los experimentos se detienen entre la 3 y 4 iteración. La mayor captación de ejemplos reportada fue de 18 documentos por clase, que corresponde al 60% del corpus objetivo. En promedio se agregan 6 documentos por clase en cada ciclo.

5.3.3 Comentarios Finales del Segundo Método

A diferencia del primer método, aplicar el método de refinamiento de la clasificación translingüe mediante incorporación de ejemplos, demuestra mejoría en el 75% de los experimentos presentados. En contraste, de ese porcentaje, aproximadamente 78% de los experimentos logran un incremento

en la exactitud superior al promedio, que se fija en 5.78% respecto a la clasificación inicial para ambos esquemas. En el 67% de los experimentos se deja de incorporar documentos antes de la tercera iteración y el resto de los experimentos lo hace entre la cuarta y la quinta iteración.

5.3.3.1 Comparación con referencias externas

Basados en el método planteado por Rigutini (2005), el cual también incluye un proceso iterativo, se realiza una serie de experimentos adicionales con el corpus utilizado en este trabajo. Los resultados se muestran en la Tabla 5.8, donde las dos primeras columnas muestran los idiomas de los corpus de entrenamiento y objetivo respectivamente. La tercera columna indica la exactitud de la clasificación inicial del esquema TCE. La cuarta columna indica la exactitud final alcanzada por la aplicación del método y la última muestra el número de iteraciones efectuadas antes de cumplirse el criterio de paro.

Tabla 5.8 Resultados Método Rigutini (2005)

Entrenamiento	Objetivo	Inicial	Exactitud	No Iteraciones
Inglés Español	Español	71.66%	78.33%	3
Inglés Francés	Francés	75.83%	76.66%	2
Español Inglés	Inglés	81.66%	87.50%	2
Español Francés	Francés	80.83%	85.00%	2
Francés Inglés	Inglés	85.83%	90.00%	2
Francés Español	Español	83.33%	83.33%	1

Como puede observarse, el método utilizado por Rigutini obtiene resultados aproximados, aunque generalmente inferiores, a los mostrados por el segundo método, por lo que la aplicación de los métodos propuestos en este trabajo resulta prometedora. La siguiente sección presenta la discusión de los resultados.

5.4 Discusión de los Resultados

En esta sección se hace una comparación entre los resultados generados por el método de refinamiento de la clasificación por vecinos más cercanos y el método de refinamiento de la clasificación mediante incorporación de ejemplos. En la Tabla 5.9 se muestran los mejores resultados de ambos métodos aplicados en la clasificación TCE. En las dos primeras columnas se indica el idioma de entrenamiento y objetivo respectivamente. La tercera columna muestra la exactitud de la clasificación inicial asignada por el clasificador translingüe simple. La cuarta columna retoma el mejor resultado del método de refinamiento por medio de vecinos más cercanos indicando en la quinta columna el número de vecinos con los que se obtuvo dicho resultado. La sexta columna muestra el mejor resultado del método de incorporación de ejemplos indicando en la séptima columna el número de vecinos con los que se obtuvo el resultado. Cabe mencionar que cuando la exactitud no varía con el número de vecinos, se considera el menor número de vecinos únicamente por simplicidad. De cualquier forma, para la mayoría de los casos, aplicar los métodos con 3 vecinos resulta benéfico.

Tabla 5.9 Comparación de mejores resultados ambos métodos TCE

Entrenamiento	Objetivo	Exactitud inicial	Mejor exactitud método 1		Mejor exactitud método 2	
Inglés Español	Español	71.66%	73.33%	4 vecinos	84.16%	3 vecinos
Inglés Francés	Francés	75.83%	77.50%	3 vecinos	75.83%	3 vecinos
Español Inglés	Inglés	81.66%	90.00%	3 vecinos	91.66%	3 vecinos
Español Francés	Francés	80.83%	83.33%	3 vecinos	85.83%	4 vecinos
Francés Inglés	Inglés	85.83%	95.83%	3 vecinos	93.33%	5 vecinos
Francés Español	Español	83.33%	84.16%	3 vecinos	85.33%	3 vecinos

Para facilitar la visualización de los mejores resultados, la Tabla 5.10 indica el incremento con respecto a la clasificación inicial de cada método. Si bien algunos de los resultados son mejores en el segundo método, es interesante considerar el experimento con entrenamiento Inglés-Francés, donde el primer método supera al segundo, debido a que éste último no fue capaz de incorporar ejemplos. El caso con entrenamiento en Francés- inglés en el cual a pesar de la incorporación de ejemplos, el primer método también supera al segundo.

Tabla 5.10 Comparación de mejores porcentajes ambos métodos TCE

Entrenamiento	Objetivo	Mejor incremento método 1	Mejor incremento método 2
Inglés Español	Español	2.33%	17.44%
Inglés Francés	Francés	2.20%	0.00%
Español Inglés	Inglés	10.21%	12.25%
Español Francés	Francés	3.09%	6.19%
Francés Inglés	Inglés	11.65%	8.74%
Francés Español	Español	1.00%	2.40%

La Tabla 5.11 muestra una comparación de la clasificación inicial con los mejores resultados de aplicar los dos métodos en la clasificación TCP. Al igual que anteriormente, cuando dos o más vecinos obtienen los mismos resultados, se indica el experimento con el menos número de vecinos. Nuevamente, utilizar 3 vecinos proporciona mayores beneficios. La Tabla 5.12 muestra los incrementos respecto a la clasificación inicial para el mejor resultado de cada método. En este grupo de experimentos aplicar el segundo método es mejor que aplicar el primero. Debe considerarse por ejemplo el experimento con objetivo Español-Inglés que no reporta cambios con el primer método pero con el segundo alcanza un incremento de 13.33% sobre la clasificación inicial. Obsérvese también que el experimento con objetivo Francés-Inglés no tiene cambios con ninguno de los dos métodos.

Tabla 5.11 Comparación de mejores resultados ambos métodos TCP

Entrena	Objetivo	Exactitud inicial	Mejor exactitud método 1		Mejor exactitud método 2	
Inglés	Español Inglés	75.00%	75.00%	3 vecinos	85.00%	3 vecinos
Inglés	Francés Inglés	76.66%	76.66%	4 vecinos	76.66%	3 vecinos
Español	Inglés Español	85.00%	90.83%	3 vecinos	91.66%	5 vecinos
Español	Francés Español	79.16%	80.83%	3 vecinos	85.00%	3 vecinos
Francés	Inglés Francés	86.66%	92.50%	3 vecinos	93.33%	3 vecinos
Francés	Español Francés	80.00%	81.66%	3 vecinos	85.83%	5 vecinos

Tabla 5.12 Comparación de mejores porcentajes ambos métodos en TCP

Entrenamiento	Objetivo	Mejor incremento método 1	Mejor incremento método 2
Inglés	Español Inglés	0.00%	13.33%
Inglés	Francés Inglés	0.00%	0.00%
Español	Inglés Español	6.86%	7.84%
Español	Francés Español	2.11%	7.38%
Francés	Inglés Francés	6.74%	7.70%
Francés	Español Francés	2.07%	7.29%

En resumen, el promedio de la clasificación translingüe simple llega a 80.1% de exactitud en ambos esquemas. El promedio de incremento al aplicar el primer método es de 3.18% sobre la clasificación inicial y 5.78% al aplicar el segundo método. Considerando los experimentos de ambos esquemas, 79% indican incrementos en la clasificación final. El primer método obtuvo experimentalmente 11.65% más que la clasificación inicial en el mejor caso, mientras que el mejor caso del segundo método llega a superar en 17.44% la clasificación inicial. Cabe mencionar que 13 de los 72 experimentos llegan a ser comparables con la referencia monolingüe cuya exactitud promedio es de 92.22%. Ambos métodos cumplen el criterio de paro en el 79% de los casos entre la primera y segunda iteración y 20% entre la cuarta y quinta iteración.

Capítulo 6

Conclusiones

Durante la investigación, se realizaron análisis para determinar los efectos de traducción y de discrepancia cultural en la clasificación translingüe, ya que han sido consideradas como los principales problemas de la clasificación translingüe. Se determinó experimentalmente que el ruido introducido por el traductor no afecta significativamente a la clasificación. La caída en la exactitud en el caso translingüe se debe a la discrepancia cultural, ya que, entre el conjunto de entrenamiento y el conjunto objetivo, por no provenir de la misma distribución, se reduce el vocabulario utilizado para la representación de documentos. En consecuencia, la discrepancia cultural provoca un descenso de exactitudes al aplicar clasificación translingüe.

Existen dos esquemas conocidos para la clasificación translingüe, el primero traduce el conjunto de entrenamiento y el segundo traduce el conjunto objetivo. Debido a que hasta ahora no se han encontrado trabajos que reporten resultados con ambos esquemas, este trabajo desarrolla ambos para evitar sesgos, si bien los primeros experimentos sugieren que ambos esquemas son comparables. Adicionalmente, la mayoría de los trabajos realizados sobre clasificación translingüe, investigan el uso de diferentes clasificadores, representaciones y pesos para evaluar el comportamiento de

la clasificación, centrándose en solucionar la barrera del idioma y no en abordar el problema de discrepancia cultural. Considerando al corpus objetivo como una muestra de la distribución que se desea capturar, en este trabajo se proponen dos métodos que tratan de aprovechar el conjunto no etiquetado para mejorar la clasificación translingüe confirmando la hipótesis inicial: *los problemas de discrepancia cultural, aunados a los pormenores de la traducción, pueden abordarse en la clasificación translingüe utilizando información que se encuentra en el conjunto escrito en el idioma objetivo*.

El primer método propuesto es un proceso iterativo posterior a la clasificación translingüe, el cual utiliza la clase asignada a los vecinos de cada documento para refinar la clasificación sin modificar el clasificador original. Experimentalmente, considerando ambos esquemas, la exactitud se incrementa en promedio 3.18% sobre la clasificación inicial. Llegando a superarla en 11.65% y sólo en un caso se a disminuirla 1%. En general se obtienen mejores resultados al usar 3 vecinos para la votación. Alrededor del 91% de los experimentos cumplen el criterio de paro entre la primera y segunda iteración, sin encontrar relación directa entre el número de cambios efectuados y la exactitud. Por la naturaleza del método no se considera aplicable si las clases tienen un fuerte traslape de vocabulario, ya que se favorecerían los errores y el método es dependiente de la clasificación inicial.

El segundo método crea clasificadores de forma iterativa integrando ejemplos etiquetados de forma confiable del conjunto objetivo con el fin de adaptar el clasificador a los documentos del idioma objetivo. Los documentos candidatos son aquellos considerados confiables porque a sus vecinos y a él mismo se les ha asignado una clase idéntica. Experimentalmente, considerando ambos esquemas, la exactitud se incrementa en promedio 5.78% sobre la clasificación inicial, llegando a superarla en 17.44% y a disminuirla en 1%. En general se obtienen mejores resultados al usar 3

vecinos para la votación. Alrededor del 67% de los experimentos cumple el criterio de paro entre la primera y segunda iteración. La mayor captación de ejemplos fue del 60% del corpus objetivo, agregando en promedio 6 documentos por clase en cada ciclo. En congruencia con el método anterior, por su naturaleza, no se considera aplicable si existe un fuerte traslape de vocabulario entre clases, ya que aumenta la probabilidad de incorporar ejemplos erróneamente etiquetados. El número de ejemplos incorporados debe ser proporcional al tamaño del corpus de entrenamiento para tener impacto en la clasificación.

En resumen, experimentalmente el promedio de la clasificación translingüe simple llega a 80.1% de exactitud en ambos esquemas. Considerando los experimentos, 79% indican incrementos en la clasificación final, de este porcentaje, aproximadamente 23% llegan a ser comparables con la referencia monolingüe cuya exactitud promedio es de 92.22%. En conclusión ambos métodos logran el objetivo de mejorar la clasificación translingüe llegando en algunos casos a hacerla comparable con la monolingüe. Se detienen entre la primera y segunda iteración en el 79% de los casos. Por su naturaleza, no parecen ser aplicables cuando la clasificación translingüe reporta exactitudes bajas debido a la fuerte dependencia de ambos métodos a la clasificación inicial.

6.1 Trabajo futuro

Evaluación de otros idiomas. El español, inglés y francés son considerados idiomas internacionales, en consecuencia, los recursos existentes se encuentran más desarrollados en comparación a los de otros idiomas. Es conveniente considerar que incorporar idiomas diferentes traería

consigo retos adicionales, los cuales pueden derivar en uso de doble traducción, expansión de ejemplos vía internet, clases desbalanceadas, etc.

Crear un método híbrido. Es posible crear un método híbrido que utilice tanto el método de refinamiento de la clasificación mediante incorporación de ejemplos como el método de refinamiento de la clasificación mediante vecinos más cercanos.

Análisis de los esquemas de clasificación translingüe. Durante el trabajo no se demostró que alguno de los dos esquemas funcionara mejor que el otro. Es conveniente realizar un análisis profundo a fin de demostrar, si existen, las ventajas de un esquema sobre el otro sin importar que dichas ventajas existan para casos específicos.

Referencias

Aas, Kjersti, and Eikvil, Line. "Text Categorisation: A Survey." *Technical Report.* Norwegian Computing Center, Junio 1999.

Aceves Pérez, Rita M., Luis Villaseñor Pineda, and Manuel Montes y Gómez. "Using N-gram Models to Combine Query Translations in Cross-Language Question Answering." *International Conference on Intelligent Text Processing and Computational Linguistics CICLing-2007* (Springer) 4394 (2007): 485-493.

Araujo Arredondo, Nadia Patricia. *Clasificación de Textos de Opinión*, Tesis de Maestría en Ciencias Computacionales, INAOE, México, 2006

Baeza, R. y Ribeiro, B. *Modern Information Retrieval*. Addison Wesley, 1999.

Bel, Nuria, Cornelis H.A. Koster, and Marta Villegas. "Cross-Lingual Text Categorization." 7th *European Conference on Digital Libreries, ECDL*. Trondheim Norway, 2003. 126-139.

Bolshakou, Igor, and Alexander Gelbukh. *Computational Linguistics*. México: Fondo de Cultura Económica, 2004.

Chih-Ping, Wei, Shi Huihua, and Christopher C. Yang. "Feature Reinforcement Approach to Poly-lingual Text Categorization". *10th*

International Conference on Asian Digital Libreries, ICADL 2007, Lecture Notes in Computer Sciencie, Springer, 2007. 99-108.

Chung-Hong, Lee, Yang Hsin-Chang, and Ma Sheng-Min. "A Novel Multilingual Text Categorization System using Latent Semantic Indexing." *Proceedings of the First International Conference on Innovative Computing, Information and Control.* IEEE Computer Society, 2006a.

Chung-Hong, Lee, Yang Hsin-Chang, Chen Ting-Chung, and Ma Sheng-Min. "Development of a Multi-Classifier Approach for Multilingual Text Categorization." *Proceedings of the 2006 International Conference on Data Mining, DMIN 06*. 2006b. 73-77.

Cirujeda Jarque, Laura, Claudia Costa Daalmans, Esther Gómez de la Orden, and Rosa Moreno Renart. *Sistemas de Representación y Procesamiento automático del Conocimiento.* 2004. http://personales. upv.es/ ccarrasc/ doc/ 2003-2004 / TesaurosOnto / principal.html #Definiciones (ultimo acceso Abril 17, 2009).

de Melo, Gerard, and Stefan Siersdorfer. "Multilingual Text Classification using Ontologies." *29th European Conference on IR Research, ECIR 2007*, Lecture Notes in Computer Sciencie. Springer, 2007.

Frakes, W. B. y Baeza-Yates, R. *Information Retrieval: Data structures and Algorithms*. Englewod Cliffs, Prentice Hall, 1992.

Galicia Haro, Sofía N., and Alexander Gelbukh. *Investigaciones en Análisis Sintáctico para el Español*. Instituto Politécnico Nacional, México 2007.

García Adeva, Juan José, Rafael A. Calvo, and Diego López de Ipiña. "Multilingual Approaches to Text Categorisation." *The European Journal for the Informatics Professional 6*, no. 3 (2005): 43-51.

García Vega, Manuel, M. Tesera Martín Valdivia, and L. Alfonso Ureña López. "Categorización de Textos Multilingües basada en Redes Neuronales." *Procesamiento del Lenguaje Natural*, no. 27 (2001): 265-271.

Gliozzo, Alfio, and Carlo Strapparava. "Exploiting Corporable Corpora and Biligual Dictionaries for Cross-Language Text Categorization." *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL. Sydney: Association for Computational Linguistics, 2006*. 553-560.

Hearst M., Schölkopf B.,. Dumais S., Osuna E, and J. Platt. *Trends and controversies - Support vector machines*, IEEE Intelligent systems, pp. 18-28, 1998.

Hovy, E. H. y C.Y. Lin 1999. “Automating text summarization in SUMMARIST” *Annual Meeting of the ACL, Proceedings of a workshop* In Mani and Maybury 1999, 81-97.

Ifrim, Georgiana, Martin Theobald, and Gerhard Weikum. "Learning Word-to-Concept Mappings for Automatic Text Classification." *Learning in Web Search, 22nd International Conference on Marchine Learning*. Bonn, Germany, 2005. 18-26.

Izhikevich, Eugene M, Sholarpedia, Septiembre 1, 2009. http://www. scholarpedia.org/article/Image:Knn_sample_plot.png (ultimo acceso Agosto 15, 2009).

Jalam, Radwan. *Apprentissage automatique et catégorisation de textes multilingues*. Lyon: PhD Tesis, Université Lumiere Lyon 2, 2003.

Jezek, Karel, and Michal Toman. "Document Categorization in Multilingual Environment" *Proceedings ELPUB2005 Conference on Electronic Publishing. 2005.*

Joachims, T., A probabilistic analysis of the rocchio algorithm with tfidf for text categorization, in *Proceedings of the 14th International Conference on Machine Learning*, pp. 143–151, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

Lewis, David D., Yiming Yang, Tony G. Rose, and Fan Li. "RCV1: A New Benchmark Collection." *Journal of Machine Learning Research 5* (2004): 361-397.

Loukachevitch, Natalia V. "Knowledge Representation for Multilingual Text Categorization." *AAAI Technical Report* SS-97-05, 1997.

Manning, Christopher D. and Schütze, Hinrich. *Foundations of Statistical Natural Language*, 575-576. The MIT Press, 2001.

Mitchell, Tom M. *Machine Learning*. McGraw-Hill, 1997.

Mitkov, Ruslan. *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2003.

Nielsen//NetRatings, International Telecommunications Union. Internet Word Stats. Diciembre 31, 2008. http://www.internetworldstats.com/ stats7.htm (ultimo acceso Abril 17, 2009).

Noble, William S., Nature Pubishing Group, Nature biotechnology, 2007, http://www.nature.com/nbt/journal/v24/n12/fig_tab/nbt1206-1565_F1.html (ultimo acceso Agosto 15, 2009)

Olsson, J. Scott, Douglas W. Oard, and Jan Hajic. "Cross-Language Text Classification." *SIGIR*. Salvador, Brazil: ACM, 2005. 15-19.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. Portal Unesco, 1995-2009. http://portal.unesco.org/culture/en/ev.php-URL_ID= 34325&URL_DO =DO_TOPIC&URL_SECTION=201.html (ultimo acceso Abril 17 de 2009).

Rigutini, Leonardo, Marco Maggini, and Bing Liu. "An EM based training algorithm for Cross-Language Text Categorization." *Proceedings Web Intelligence Conference.* Compiegne, France: IEEE, 2005.

Russell, S. y Norvig, P. *Inteligencia Artificial, Un Enfoque Moderno*, Segunda ed. Prentice-Hall International, 2004.

Salton, G. y M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill, New York, 1983.

Sebastiani, Fabrizio. "Machine Learning in Automated Text Categorization." *ACM computing Surveys* 34 (March 2002): 1-47.

Sierra Araujo, Basilio. *Aprendizaje Automático: Conceptos básicos y avanzados.* Edited by Miguel Martín Romo. Pearson Educación S.A, 2006.

Tikk, Domonkos, and György Biró. "Text categorization on a multi-lingual corpus." *Proc. of the 4th int. Symp. of Hungarian Researchers on Computational Intelligence*. Budapest, Hungary, 2003. 167-177.

Vicedo J., Rodríguez H., Peñas A. & Massot M. “Los sistemas de Búsqueda de Respuestas desde una perspectiva actual”*. Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, pp. 351-367, 2003.

Villatoro Tello, Esaú, Luis Villaseñor Pineda, and Manuel Montes y Goméz. "Using Word Sequences fot Text Summarization." *LNCS (LNAI)* 4188 (2006): 297.

WebKeySoft, *Simple Text Encodign Converter*, www.webkeysoft.com, 2008 (ultimo acceso Agosto 2008).

Wenyuan, Dai, Xue Gui-Rong, Yang Qiang, and Yu Yong. "Transferring Naive Bayes Classifiers for Text Classification." *Proceedings of the Twenty-Second AAAI conference on Artificial Intelligence.* Vancouver, British Columbia Canada, 2007, 540-545.

Zhu, Lin, Jihong Guan, and Shuigeng Zhou. CWC: A Clustering-Based Feature Weighting Approach for Text Classification. Vol. 4617, in *Modeling Decisions for Artificial Intelligence*, 204-205. Springer, 2007.

Anexo A: Recursos Multilingües

A continuación se enlistan algunos conceptos utilizados en el entorno multilingüe:

Corpus bilingüe o multilingüe: Es un conjunto de textos diferentes escritos en dos o más idiomas que coinciden en la temática.

Corpus Paralelo: Es el conjunto que contiene los mismos textos escritos en diferentes idiomas.

Las **entidades nombradas** son todas aquellas palabras o serie de palabras que no varían: nombres de personas o lugares, expresiones de épocas, cantidades, especies biologícas o substancias, etc.

Los **cognados** son palabras (en ocasiones también símbolos) que como remanentes de su idioma de origen tienen poca variación en los distintos idiomas: *chocolate* (español), *chocolate* (inglés), *chocolat* (francés), *schokolade* (alemán). Existe investigación con el fin de encontrar relaciones entre un idioma y otro. También existen los falsos cognados: *dos* (español: Número arábigo 2, francés: espalda).

Un **lenguaje univoco** es un lenguaje en el cual cada palabra o expresión tiene un solo sentido, una sola interpretación posible y no existe

más que una forma de expresar un concepto dado. El lenguaje natural[15] es **equivoco**: existen varias formas de expresar la misma idea (redundancia), las expresiones poseen comúnmente diversas interpretaciones (ambigüedad) y no todo es explicado en el discurso (función implícita) (Jalam, 2003).

Con el objetivo de abordar las dificultades del lenguaje natural, algunos investigadores utilizan técnicas de **Ingeniería del Conocimiento** para auxiliar la clasificación de textos. Las técnicas de Ingeniería del Conocimiento se basan en la creación de bases, reglas y diccionarios que describan el dominio del conocimiento. Para relacionar palabras con conceptos, se construyen fuentes como lexicones, tesauros y ontologías. (Ifrim et al., 2005). La creación de recursos basados en Ingeniería del Conocimiento requiere una considerable cantidad de trabajo humano y de tiempo de desarrollo. Una vez que el sistema ha sido creado, los cambios en los tipos y formatos de textos, las modificaciones de categorías o la substitución del sistema completo de categorías es trabajo significativo adicional y costoso en tiempo; aunque herramientas como diccionarios y tesauros generados automáticamente han sido creados para reducir el problema (Loukachevitch, 1997).

Un **tesauro** es una herramienta en la cual se manejan conceptos, por ello, la estructura de la terminología de un **tesauro** está basada en las interrelaciones entre los conceptos. Estas interrelaciones pueden ser: jerárquicas, de afinidad y preferenciales. Las relaciones jerárquicas indican términos más amplios o más específicos de cada concepto (Cirujeda et al., 2004). El concepto de **ontología** fue tomado por la Inteligencia Artificial, para definir el vocabulario de un dominio acotado mediante un conjunto de términos básicos y relaciones entre dichos términos. Las ontologías son

[15] Llamado así para hacer diferencia con los lenguajes de programación.

teorías que especifican un vocabulario relativo a un cierto dominio. Este vocabulario define entidades, clases, propiedades, predicados y funciones y, las relaciones entre estos componentes. Las ontologías se diferencian de los tesauros por una descripción formal de los objetos en el mundo, sus propiedades, y las relaciones entre estos objetos. Así, las operaciones que pueden realizarse en un tesauro son comprobar la idoneidad de un término o ver las relaciones existentes con otros términos. Mientras que algunas ontologías ofrecen la habilidad de comprobar la consistencia de la misma y deducir la existencia de los objetos de forma automática.

Existen dos tesauros importantes: **WordNet** (1999), que contiene 150,000 conceptos, cada uno con una pequeña descripción y relaciones semánticas entre los conceptos, y **EuroWordNet**, que abarca gran parte de los idiomas hablados en la unión europea. Al igual que EuroWordNet, **MultiWordNet** es un proyecto que pretende la expansión de WordNet para los principales idiomas europeos. MultiWordNet se ha desarrollado de forma independiente, por ahora solo cuenta con inglés, español e italiano, pero portugués, hebreo y rumano serán integrados en breve. Los dos tesauros difieren en su procedimiento de construcción, en EuroWordNet se construyó un tesauro independiente para cada idioma y posteriormente se buscaron las correspondencias entre los mismos. Al contrario, MultiWordNet se desarrolló manteniendo en medida de lo posible las relaciones establecidas en WordNet. A diferencia del WordNet original de Princeton, la mayoría de los otros WordNets no están disponibles de manera gratuita.

La ley de Zipf: En su libro *Human Behavior and the Principle of Least Efford*, Zipf argumenta haber encontrado un principio unificarte. El principio del Menor esfuerzo, que sustenta que la gente actuará para minimizar su promedio de trabajo, las leyes empiezan cubriendo ciertas distribuciones en el lenguaje. Si se contabiliza la frecuencia de cada palabra en un conjunto de

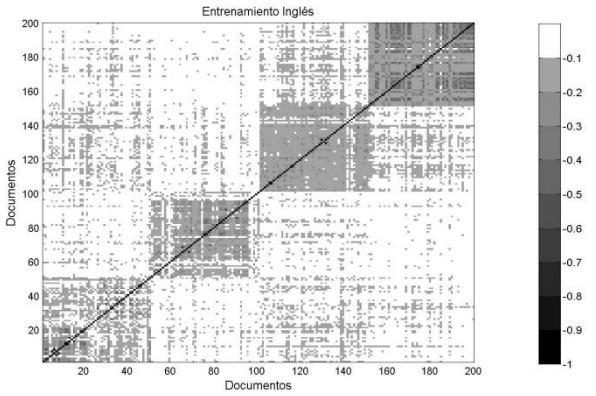
documentos se encuentra una burda descripción de la distribución de las palabras en los lenguajes humanos. Existe una pequeña cantidad de palabras muy comunes, un número medio de palabras medianamente frecuentes y muchas palabras poco frecuentes. De acuerdo a la ley de Zipf, en un discurso, tanto el orador como el escucha están tratando de reducir su esfuerzo. El orador conserva su esfuerzo por medio de poco vocabulario con palabras comunes y el escucha con el uso de un vocabulario amplio incluyendo palabras raras (para que los mensajes sean poco ambiguos). Para la mayoría de las palabras los datos sobre su uso serán dispersos (infrecuentes), sólo para unas cuantas palabras existirán varios ejemplos. (Manning y Schütze, 2001)

Anexo B: Gráficas de Similitud

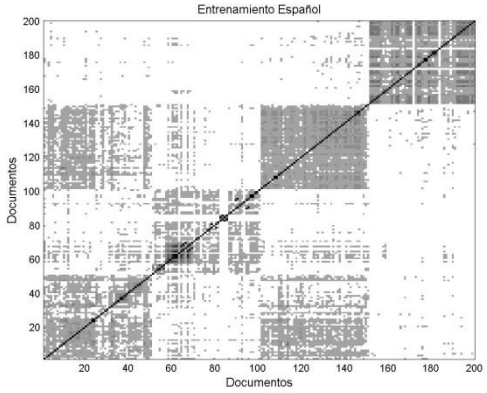
En este anexo se encuentran la totalidad de las gráficas de similitud obtenidas para el presente trabajo. En primer lugar se muestran las del corpus de entrenamiento para los tres idiomas. Posteriormente se presentan por experimento, se muestra el corpus objetivo con ambas traducciones efectuadas colocando siempre el corpus objetivo sin traducción al inicio.

Cuando los corpus están escritos en diferentes idiomas una comparación entre ellos permite apreciar sus diferencias. La gráfica del conjunto de entrenamiento en inglés (a) muestra clases definidas que no se confunden con otras clases. Cada documento perteneciente a una clase se parece a las de su misma clase y es diferente a los de otras clases en cuanto al vocabulario que utiliza. Sin embargo, el conjunto de entrenamiento en español (b) presenta traslapes en la clase 1 y 3, igual que el de francés (c). De esta forma se observa la complejidad de los conjuntos de entrenamiento.

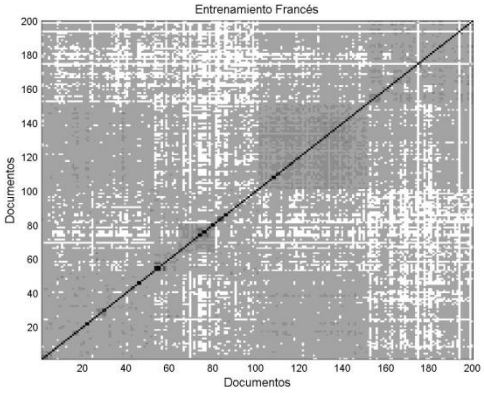
Es interesante que, en estos experimentos, las características esenciales de los corpus de entrenamiento y objetivo se conserven después del proceso de traducción. Por ejemplo, las similitudes entre clases. La totalidad de las gráficas del corpus con traducción y sin ella, se presentan sin más explicación.



a) Conjunto de entrenamiento en inglés

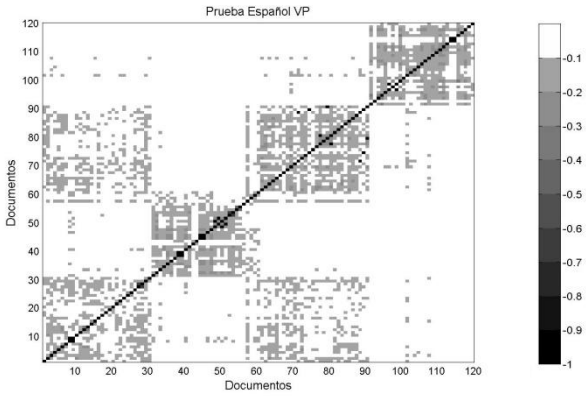


b) Conjunto de entrenamiento en español

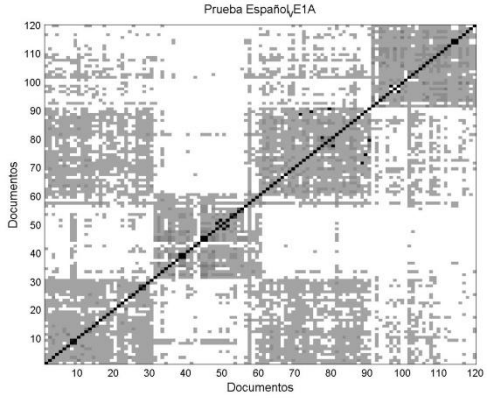


c) Conjunto de entrenamiento en francés

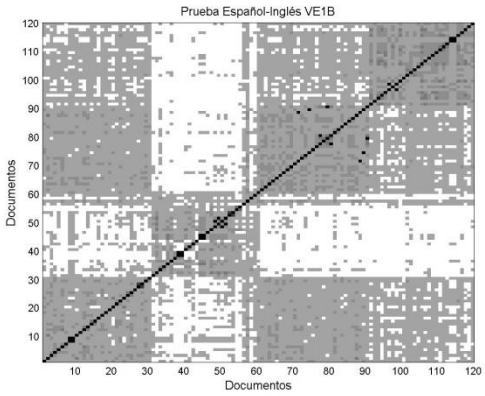
Figura B.6.1 Graficas de similitud de los corpus de entrenamiento en los tres idiomas



Corpus objetivo en Español
Vocabulario con el que se representa: Corpus objetivo en español

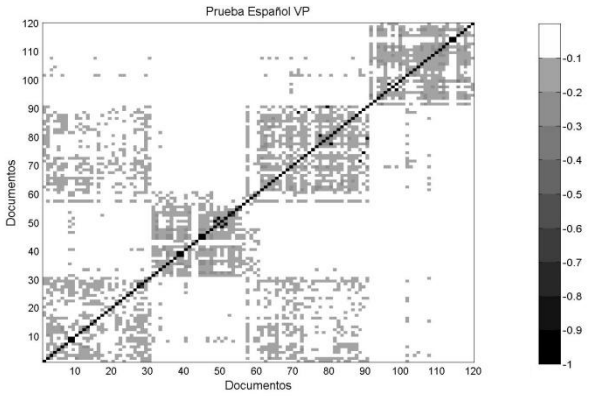


Corpus objetivo en Español
Vocabulario con el que se representa: Corpus de Entrenamiento traducción de inglés a español

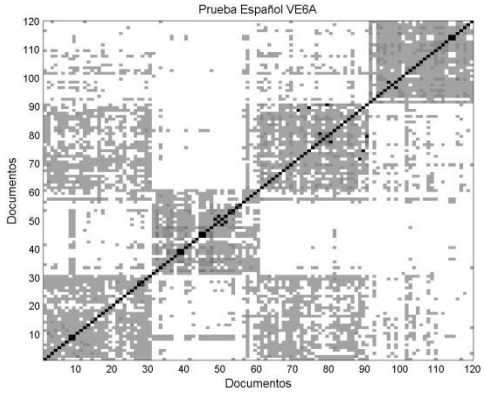


Corpus objetivo en traducción de Español a inglés
Vocabulario con el que se representa: Corpus de Entrenamiento inglés

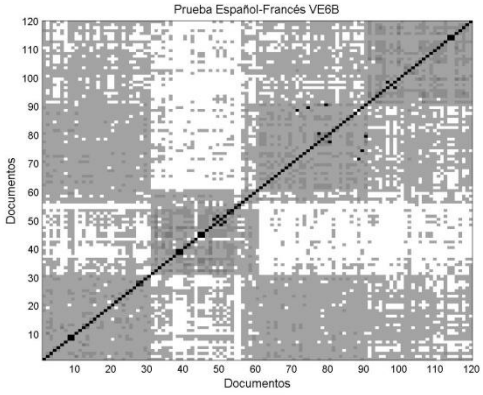
Figura B.6.2 Graficas de similitud con Idioma objetivo: español. Idioma fuente entrenamiento: inglés



Corpus objetivo en Español
Vocabulario con el que se representa: Corpus objetivo en Español

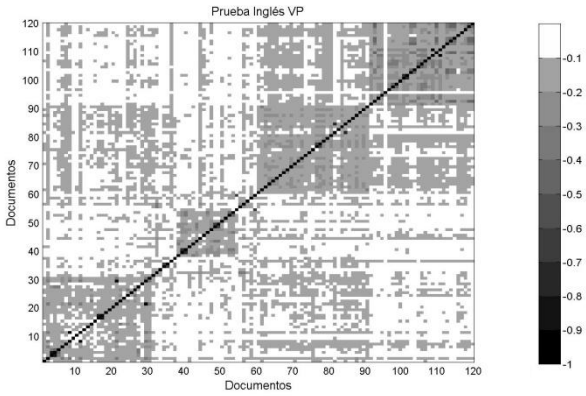


Corpus objetivo en Español
Vocabulario con el que se representa: Corpus de Entrenamiento traducción de Francés a español

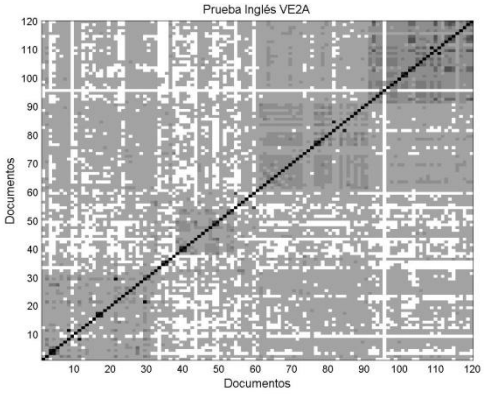


Corpus objetivo en traducción de Español a francés
Vocabulario con el que se representa: Corpus de Entrenamiento Francés

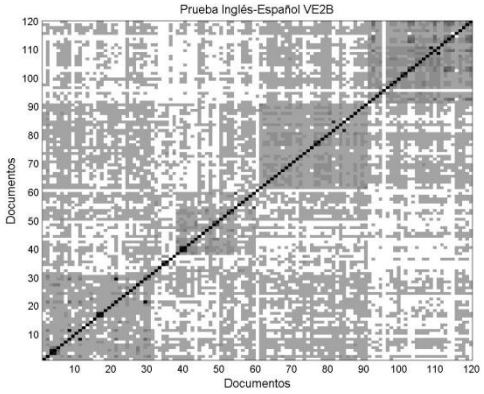
Figura B.6.3 Graficas de similitud con Idioma objetivo: Español Idioma fuente entrenamiento: Francés



Corpus objetivo en Inglés
Vocabulario con el que se representa: Corpus objetivo en Inglés

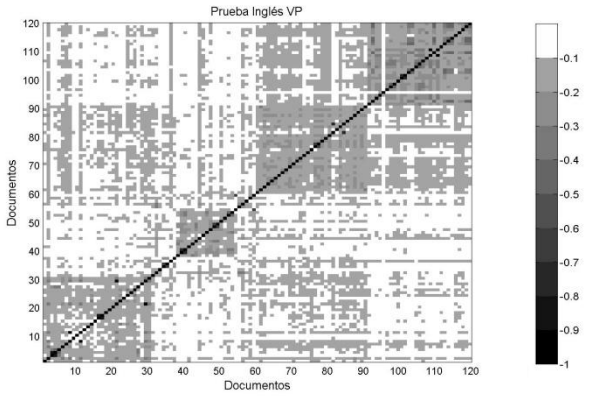


Corpus objetivo en Inglés
Vocabulario con el que se representa: Corpus de Entrenamiento traducción de Inglés a español

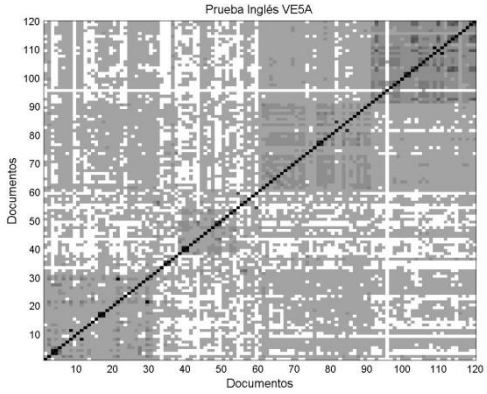


Corpus objetivo en traducción de inglés a español
Vocabulario con el que se representa: Corpus de Entrenamiento Español

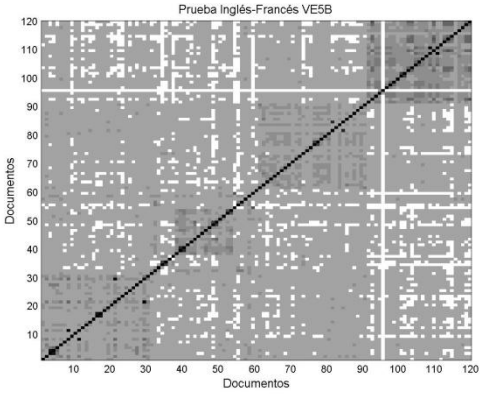
Figura B.6.4 Graficas de similitud con Idioma objetivo: Inglés. Idioma fuente entrenamiento: Español



Gráfica: Corpus objetivo en Inglés
Vocabulario con el que se representa: Corpus objetivo en Inglés

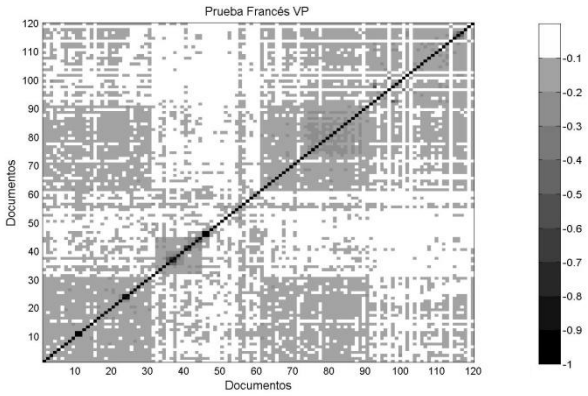


Gráfica: Corpus objetivo en Inglés
Vocabulario con el que se representa: Corpus de Entrenamiento traducción de Inglés a Francés

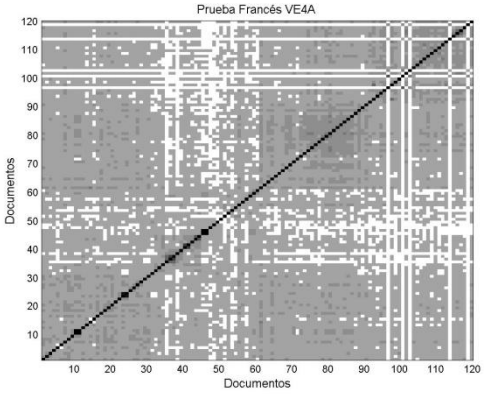


Gráfica: Corpus objetivo en traducción de Inglés a francés
Vocabulario con el que se representa: Corpus de Entrenamiento Francés

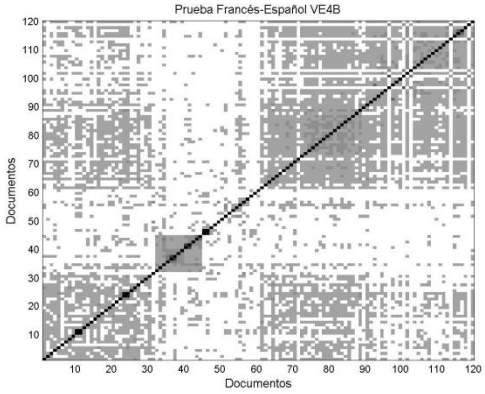
Figura B.6.5 Graficas de similitud con Idioma objetivo: inglés. Idioma fuente entrenamiento: Francés



Gráfica: Corpus objetivo en Francés
Vocabulario con el que se representa: Corpus objetivo en Francés

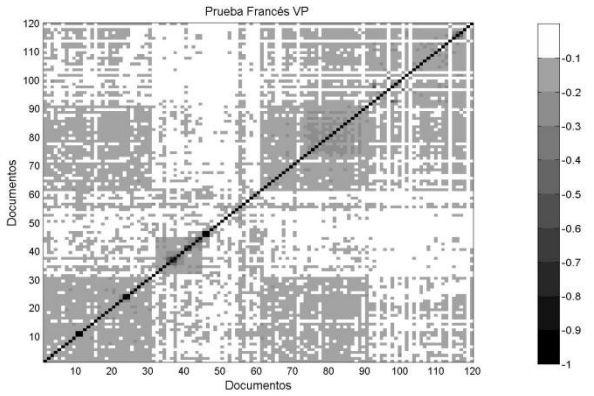


Gráfica: Corpus objetivo en Francés
Vocabulario con el que se representa: Corpus de Entrenamiento traducción de español a francés

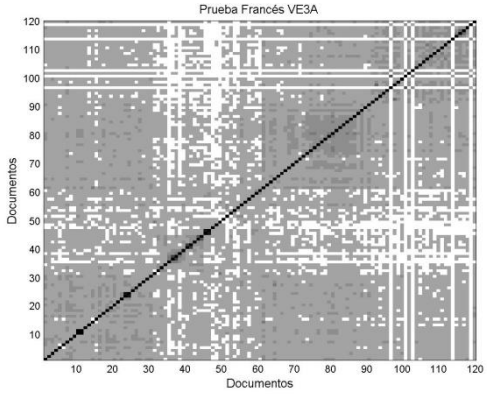


Gráfica: Corpus objetivo en traducción de francés a español
Vocabulario con el que se representa: Corpus de Entrenamiento Español

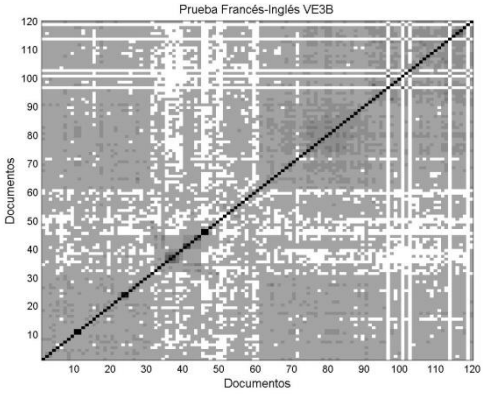
Figura B.6.6 Graficas de similitud con Idioma objetivo: Francés -Idioma fuente entrenamiento: Español



Gráfica: Corpus objetivo en Francés
Vocabulario con el que se representa: Corpus objetivo en Francés



Gráfica: Corpus objetivo en Francés
Vocabulario con el que se representa: Corpus de Entrenamiento traducción de Inglés a Francés



Gráfica: Corpus objetivo en traducción de francés a inglés
Vocabulario con el que se representa: Corpus de Entrenamiento Inglés

Figura B.6.7 Graficas de similitud con Idioma objetivo: Francés. Idioma fuente entrenamiento: Inglés

Anexo C: Experimento Secundario de Análisis por Ganancia de Información.

Como análisis se usa ganancia de información para determinar el número de términos que son relevantes en el vocabulario común, para el corpus de entrenamiento. El objetivo es determinar si el corpus objetivo contiene menor cantidad de palabras relevantes en la clasificación translingüe que en la monolingüe.

La Tabla C.6.1 muestra en la tercera columna el tamaño del vocabulario común entre entrenamiento y objetivo. En la cuarta columna se indica el número de palabras del vocabulario común seleccionadas como relevantes al aplicar ganancia de información (IG) en el corpus de entrenamiento. La quinta columna indica la exactitud de la clasificación.

Tabla C.6.1 Vocabulario común con IG -entrenamiento inglés.

Entrenamiento	Objetivo	Vocabulario Común	Vocabulario con IG según el Entrenamiento	Exactitud
Inglés	Inglés	5,452	1,356	91.66%
Inglés	Español Inglés	3,295	784	75.00%
Inglés	Francés Inglés	3,697	821	76.66%

Para obtener los datos de la cuarta columna, se obtiene la lista de palabras con ganancia de información positiva en el corpus de

entrenamiento, luego se contabilizan las palabras de la lista que aparecen en el vocabulario común. Considerando únicamente el idioma inglés, existen 1,356 palabras con IG en el caso monolingüe y la exactitud es de 91.33%, en cambio, la reducción de vocabulario común en el caso translingüe deja 784 palabras con IG en el caso español-inglés con exactitud de 75% y para francés-inglés 821 palabras con exactitud de 76.66%.

Se observa la existencia de una relación entre el número de palabras relevantes y la exactitud del clasificador. En el caso de la clasificación TCP se cuenta con el mismo corpus de entrenamiento en el caso monolingüe y translingüe. De la misma forma que los experimentos efectuados durante la investigación, debe considerarse que no es posible hacer comparación entre idiomas porque los corpus no contienen las mismas palabras. Por ello la Tabla C.6.2 muestra la comparación para idioma español y la Tabla C.6.3 para el francés.

Tabla C.6.2 Vocabulario común con IG –entrenamiento español

Entrenamiento	Objetivo	Vocabulario Común	Vocabulario con IG según el Entrenamiento	Exactitud
Español	Español	5,182	1,173	91.66%
Español	Inglés Español	3,826	722	85.00%
Español	Francés Español	3,925	716	79.16%

Tabla C.6.3 Vocabulario común con IG –entrenamiento francés

Entrenamiento	Objetivo	Vocabulario Común	Vocabulario con IG según el Entrenamiento	Exactitud
Francés	Francés	6,000	939	93.33%
Francés	Inglés Francés	4,194	595	86.66%
Francés	Español Francés	3,749	580	80.00%

Las Tablas C 6.4, C 6.5 y C 6.6, muestran el equivalente considerando el conjunto objetivo.

Tabla C.6.4 Vocabulario común con IG objetivo– entrenamiento inglés

Entrenamiento	Objetivo	Vocabulario Común	Vocabulario con IG según Objetivo	Exactitud
Inglés	Inglés	5,452	662	91.66%
Inglés	Español Inglés	3,295	366	75.00%
Inglés	Francés Inglés	3,697	323	76.66%

Tabla C.6.5 Vocabulario común con IG objetivo- entrenamiento español

Entrenamiento	Objetivo	Vocabulario Común	Vocabulario con IG según Objetivo	Exactitud
Español	Español	5,182	489	91.66%
Español	Inglés Español	3,826	414	85.00%
Español	Francés Español	3,925	277	79.16%

Tabla C.6.6 Vocabulario común con IG objetivo- entrenamiento francés

Entrenamiento	Objetivo	Vocabulario Común	Vocabulario con IG según Objetivo	Exactitud
Francés	Francés	6,000	421	93.33%
Francés	Inglés Francés	4,194	462	86.66%
Francés	Español Francés	3,749	376	80.00%

No se obtuvieron los datos para los casos monolingües con traducción pero se presenta el resto de los datos extraídos en cuyo caso se observa consistencia.

Tabla C.6.7 Vocabulario común con IG TCE (entrenamiento)

Entrenamiento	Objetivo	Vocabulario Común	Vocabulario con IG según el Entrenamiento	Exactitud
Inglés Español	Español	3,640	770	71.66%
Inglés Francés	Francés	4,131	838	75.83%
Español Inglés	Inglés	3,351	740	81.66%
Español Francés	Francés	3,793	721	80.83%
Francés Inglés	Inglés	3,700	622	85.83%
Francés Español	Español	3,920	590	83.33%

Tabla C.6.8 Vocabulario común con IG TCE (objetivo)

Entrenamiento	Objetivo	Vocabulario Común	Vocabulario con IG según Objetivo	Exactitud
Inglés Español	Español	3,640	353	71.66%
Inglés Francés	Francés	4,131	300	75.83%
Español Inglés	Inglés	3,351	411	81.66%
Español Francés	Francés	3,793	294	80.83%
Francés Inglés	Inglés	3,700	456	85.83%
Francés Español	Español	3,920	353	83.33%

Observando estos resultados, se considera que las diferencias de vocabulario entre el conjunto de entrenamiento y el conjunto objetivo ocasiona la caída de exactitud en la clasificación translingüe.