Clasificación Translingüe de Documentos usando Similitudes del Conjunto Objetivo[1]

Escobar-Acevedo, Adelina[1,2], Montes-y-Gómez, Manuel[2],
Villaseñor-Pineda, Luis[2], Guzmán-Cabrera Rafael[3]

[1]Universidad Tecnológica del Valle de Toluca.
[2]Instituto Nacional de Astrofísica, Óptica y Electrónica
{aescobar, mmontesg, villasen}@inaoep.mx
[3]Universidad de Guanajuato, DICIS
guzmanc@ugto.mx

Resumen— La Clasificación Translingüe de Documentos (CTD) pretende aprovechar recursos existentes en un idioma para clasificar documentos en otro idioma. Adicionales a los efectos de la traducción, la CTD también enfrenta la discrepancia cultural entre idiomas que implica que documentos pertenecientes a la misma clase traten tópicos diferentes. Para determinar el impacto de la discrepancia cultural, se realiza un análisis de vocabulario y un análisis basado en gráficas de similitud en los conjuntos de entrenamiento en tres idiomas. Dado que en la CTD el conjunto de entrenamiento y el conjunto objetivo provienen de distribuciones diferentes, se aplica un método para aprovechar las similitudes existentes entre documentos del conjunto objetivo a fin de mejorar la exactitud de la clasificación.

1 Introducción

Actualmente se genera alrededor del mundo gran cantidad de información. Debido a la imposibilidad humana de manejar enormes cantidades de textos, es necesario automatizar procesos que manipulen y organicen tales volúmenes de documentos [1]. Existen distintas líneas de investigación como los sistemas de recuperación de información (Information Retrieval) [2], búsqueda de respuestas (Question Answering) [3]; generación automática de resúmenes (Text Summarization) [4], y clasificación de textos (Text Categorization) [5]; entre otros.

La clasificación automática de textos se define como la tarea de asignar documentos a clases predefinidas [5]. La clasificación automática de textos es una tarea supervisada, lo que significa que no sólo se conocen previamente las categorías sino que debe contarse con un conjunto de entrenamiento. La generación de los conjuntos de entrenamiento es un proceso costoso ya que usualmente es una tarea manual efectuada por un experto. Bajo el entorno multilingüe, la clasificación automática de textos multilingües plantea el problema de clasificar documentos escritos en diferentes idiomas bajo las mismas clases.

[1] El presente trabajo se realizó gracias al apoyo recibido por el PROMEP/103.5/11/4403

Algunas compañías e instituciones necesitan buscar y organizar eficientemente repositorios multilingües de documentos. El manejo de colecciones multilingües se convierte en un problema debido a que el proceso de etiquetado se multiplica para cada idioma que se desee clasificar. Ante ello, la CTD representa una solución viable ya que se define como la tarea de asignar clases a documentos escritos en un idioma objetivo usando recursos de un idioma fuente [6] y promete reducir el esfuerzo humano a proveer el conjunto de entrenamiento en solo un idioma [7].

La CTD presenta dos características particulares que la convierten en una tarea muy compleja. Por una parte la distorsión inevitable introducida por el traductor, y por otra, la discrepancia cultural entre idiomas. Recordemos que el idioma es el medio de expresión de un grupo cultural y socialmente homogéneo. 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. Términos comunes en un grupo pueden ser escasos o inexistentes en otro; es de esperarse que los periódicos mexicanos aborden más sobre fútbol soccer que los americanos y a su vez deportes como el rugby, la pelota vasca y el cricket sean más nombrados en países europeos.

El presente trabajo está organizado de la siguiente manera. La sección 2 presenta el estado del arte mencionando los principales trabajos relacionados. En la sección 3 se describe el corpus de trabajo y hace un análisis relacionado a la solidez de los conjuntos de entrenamiento y prueba. La sección 4 muestra el método de refinamiento. En la sección 5 se presentan los resultados obtenidos; por último, la sección 6 presenta las conclusiones e ideas futuras.

2 Trabajo Relacionado

Existen trabajos que proponen realizar la clasificación multilingüe mediante la creación de clasificadores monolingües [8], [9], [10]. Bajo este enfoque, cada conjunto de prueba debe ser enviado a su respectivo clasificador, por lo que se construye un clasificador por cada idioma que se desee incluir. Otra propuesta consiste en realizar la clasificación por entrenamiento polilingüe [8], [10], [11], la cual consiste en crear un único clasificador para todos los idiomas con el objetivo de recibir documentos sin especificar de qué idioma se trata; cada idioma agregado de forma posterior implica la actualización del clasificador. Tanto las soluciones que proponen crear clasificadores monolingües como las soluciones por entrenamiento polilingüe, suponen la existencia de conjuntos etiquetados en todos los idiomas requeridos.

La CTD asume que el corpus etiquetado está disponible en un único idioma y debe ser usado para clasificar documentos de otros idiomas. Para franquear la barrera del idioma se han usado recursos como términos comunes [8], diccionarios [6], [12], tesauros [6], [13], y traducción automática [7], [8], [10]. El uso de traductores automáticos presenta la ventaja de estar disponibles para varios idiomas. A diferencia de los métodos que se centran principalmente en el problema de la traducción, en este artículo se propone abordar directamente la diferencia cultural.

La mayoría de los métodos calcula la probabilidad del documento de pertenecer a una clase utilizando únicamente los atributos del conjunto de entrenamiento, con ello,

el resto de las propiedades del conjunto objetivo se ignoran. En contraste, el método de refinamiento incluye una segunda fase en la que cada documento es representado con atributos propios del conjunto objetivo y considera el voto de sus documentos similares.

3 Análisis del Corpus de Trabajo

Con el fin de observar los efectos de la discrepancia cultural en la CTD, se realizaron dos análisis sobre el corpus de trabajo. El primero consiste en las gráficas de similitud que permiten observar la solidez de las clases en cada conjunto de documentos y comparar los traslapes entre clases por idioma. El segundo análisis consiste en identificar el vocabulario común entre conjuntos de entrenamiento y prueba, ya que será éste quien determine la exactitud de la clasificación inicial.

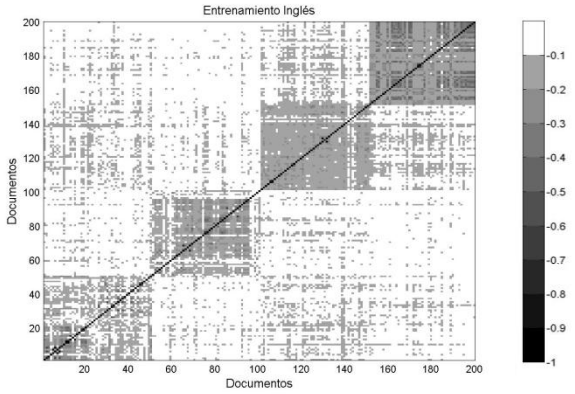
Se utilizó un subconjunto balanceado del corpus Reuters RCV-1 [14], considerando 3 idiomas: inglés, español y francés, con noticias de cuatro clases diferentes correspondientes a '*Policía', 'Desastres', 'Política'* y *'Deportes'*. Para cada idioma se utilizaron 200 noticias para entrenamiento y 120 como prueba, 50 y 30 por clase respectivamente. Para el pre-procesamiento de los documentos, se utilizaron listas de palabras vacías en los idiomas correspondientes. Finalmente, se utilizó el traductor automático Wordlingo para traducir los conjuntos de entrenamiento del idioma fuente al idioma objetivo.

3.1 Análisis por Gráficas de Similitud

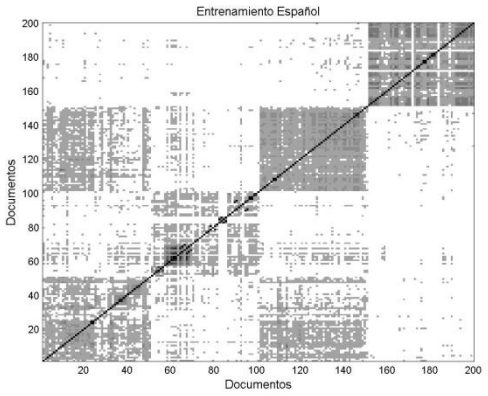
Dada la cantidad de documentos, es complicado para un humano hacer un análisis global del conjunto. Por ello se realizan gráficas de similitud que permiten ver qué tan bien están representadas las clases por sus vocabularios identificando traslapes y disimilitudes. Las gráficas se obtienen al representar vectorialmente cada documento mediante su vocabulario con pesado booleano, y compararlo con el resto de la colección mediante una medida de similitud como el Coeficiente de Dice.

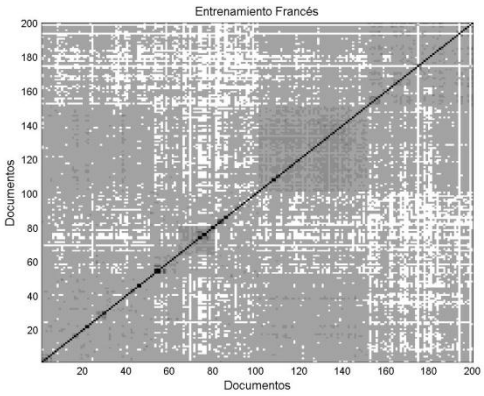
En la Fig. 1, la escala de grises determina el grado de similitud de los documentos. Se muestra que la gráfica del conjunto en inglés (a) tiene clases definidas que no se confunden con otras clases. Cada documento perteneciente a una clase se parece a las de su misma clase y es diferente a los de otras clases en cuanto al vocabulario que utiliza. Particularmente la primera clase resulta problemática ya que es poco sólida, contiene varios documentos que no son similares entre sí o cuyo grado de similitud es bajo. El conjunto de español (b) presenta traslapes en la clase 1 y 3 pero la clase 4 es sólida. El conjunto de francés (c) muestra vocabulario similar en todas sus clases aun cuando las similitudes son bajas, entre 0.2 y 0.3, en este conjunto, la clase 3 es la más marcada por sus grados de similitud mayores.

Los traslapes de clases indican que existe vocabulario utilizado comúnmente entre ellas y el clasificador puede confundirlas. Las clases poco sólidas al contrario, contienen documentos cuya representación es diferente al resto de los documentos de su clase [15]. Esta representación suele aportar escasa información al clasificador para asignar clases y a su vez al método propuesto por similitudes.



(a)





(b) (c)

Figura 1. Gráficas de similitud de los conjuntos de entrenamiento (a) inglés, (b) español y (c) francés

3.2 Análisis por Vocabulario

Como menciona [16] "En general aprender es más confiable cuando los ejemplos siguen una distribución similar a los ejemplos de prueba futuros". Aunque se espera que las distribuciones sean similares y compartan características, el conjunto de entrenamiento y prueba no son tomados de la misma distribución en la CTD. Incluso, en los conjuntos de prueba no existen muchas de las palabras que contienen los documentos de entrenamiento [8]. Para determinar el porcentaje de términos comunes se realizó una comparación simple.

Sea $T_e = \{t_{e1},...,t_{en}\}$ el conjunto de términos llamados vocabulario de entrenamiento, obtenidos de la traducción del conjunto de entrenamiento del idioma fuente al objetivo y sea $T_p = \{t_{p1},...,t_{pm}\}$ el conjunto de términos llamados vocabulario de prueba en su idioma original; entonces el conjunto $T_{com} = t_e \cap t_p$ es el vocabulario común.

Tabla 1: Tamaño de Vocabulario para la CTD

Fuente	**Objetivo**	$\mathbf{T_e}$	$\mathbf{T_p}$	$\mathbf{T_{com}}$
Inglés	Inglés	10,892	7,658	5,452
Español	Español	12,295	8,051	5,182
Francés	Francés	14,072	9,258	6,000
Inglés	Español	13453	8051	3640
Inglés	Francés	12426	9258	4131
Español	Inglés	9012	7658	3351
Español	Francés	10666	9258	3793
Francés	Inglés	11338	7658	3700
Francés	Español	14684	8051	3920

De la Tabla 1, se obtiene que el vocabulario común es aproximadamente 50% del vocabulario de entrenamiento en el caso monolingüe y de aproximadamente 31% en el caso translingüe. Esta caída en el vocabulario común afecta la exactitud de la CTD.

4 Método de Refinamiento de Clases

Bajo la premisa de que el conjunto de prueba es una muestra de la distribución que nos interesa captar, se espera que los documentos pertenecientes a la misma clase sean similares entre sí. Así, se utiliza un método de refinamiento de la clasificación [17], en el cual, adicional a la clasificación asignada se recibe la votación de los documentos con mayor similitud, ver Fig. 2.
El método se describe en el siguiente algoritmo:

1. Se construye un clasificador C_1 utilizando el conjunto 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 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 $T_p = \{t_{p1},...,t_{pm}\}$ y se enlistan los k vecinos más cercanos $d_{nn1},...,d_{nnk}$ de cada documento.

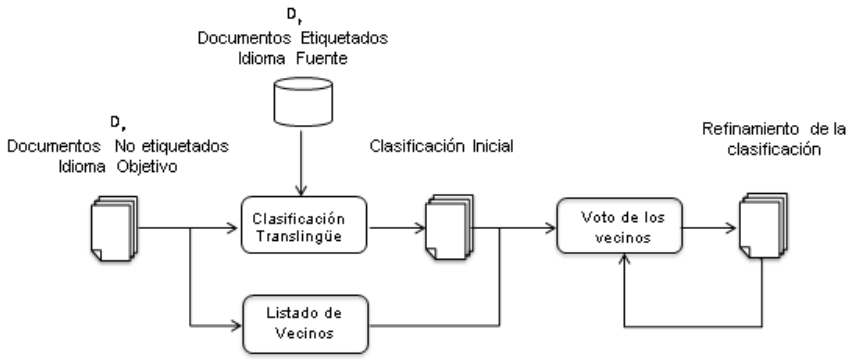


Figura 2. Método propuesto de refinamiento de CTD.

4. Sea c_j^n la clase asignada por el clasificador al documento d_j en la iteración n y $c_{nn1}^n ... c_{nnk}^n$ las clases de sus k vecinos. Entonces $c_j^{n+1} = c_{nn1}^n$ sí $c_{nn1}^n = c_{nn2}^n = ... = c_{nnk}^n$, re-clasificando al documento con la clase de sus vecinos y $c_j^{n+1} = c_j^n$ si no se cumple que $c_{nn1}^n = c_{nn2}^n = ... = c_{nnk}^n$.
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.

5 Resultados

La clasificación se realiza utilizando como método de aprendizaje Näive Bayes y como medida de evaluación la exactitud, que indica el porcentaje de documentos que fueron asignados correctamente por el clasificador. Dado que se considera un consenso entre vecinos pare refinar la clasificación, el método considera un mínimo de 3 vecinos. La Tabla 2 muestra los resultados de referencia monolingües. La Tabla 3 proporciona los resultados de la CTD sin intervención del método y al aplicarlo considerando el voto de sus 3, 4 y 5 documentos más similares.

Tabla 2. Resultados de Referencia Monolingües

Fuente	Objetivo	Exactitud
Inglés	Inglés	0.916
Español	Español	0.916
Francés	Francés	0.933

Tabla 3. Resultados del Método de Similitud

Fuente	Objetivo	Exactitud			
		Inicial	**3**	**4**	**5**
Inglés	Español	0.716	0.725	**0.733**	0.725
Inglés	Francés	0.758	**0.775**	0.766	0.766
Español	Inglés	0.816	**0.900**	0.900	0.883
Español	Francés	0.808	**0.833**	0.816	0.825
Francés	Inglés	0.858	**0.958**	0.925	0.925
Francés	Español	0.833	**0.841**	0.841	0.841

La comparación entre la exactitud de la clasificación monolingüe simple mostrada en la Tabla 2 y la exactitud de la clasificación translingüe simple mostrada en la tercera columna de la Tabla 3 hace evidente que existe una caída de exactitud al realizar CTD. La aplicación del método de refinamiento por similitudes reporta siempre un incremento en la exactitud entre el 1 y el 11.65% respecto a la clasificación inicial, llegando a ser comparables incluso con el caso monolingüe. Cabe mencionar que el número de iteraciones promedio se encuentra entre 3 y 4 antes de la convergencia del método.

6 Conclusiones

Los análisis realizados a los corpus nos permiten apreciar la dificultad y problemática de la CTD. Las gráficas de similitud permiten anticipar el desempeño que tendrá un sistema de clasificación automática de documentos en función de la separación de clases y traslapes que se observan. Esta información puede ser útil al momento de seleccionar los atributos que serán utilizados como entrenamiento para llevar a cabo la tarea. Particularmente los efectos de la discrepancia cultural tienen efectos relevantes en la clasificación debido a la reducción del vocabulario común con el que se representan los conjuntos de prueba. Sin embargo, aprovechar la propia distribución del conjunto objetivo es benéfico para la clasificación final, ya que provee información adicional para apoyar la decisión del clasificador facilitando la correcta asignación de clases. En general, aplicar el método invariablemente incrementó la exactitud inicial. Los mejores resultados llegan a ser comparables con la clasificación monolingüe. En un trabajo futuro se probarán métodos con incorporación de ejemplos con el fin de hacer una adaptación paulatina del clasificador al idioma objetivo.

Referencias

[1] Galicia Haro S. N. y Gelbukh A. Investigaciones en análisis sintáctico para el español. Instituto Politécnico Nacional, México 2007.

[2] Bolshakov I. and Gelbukh A. Computational Linguistics: models, resources, applications. México: Fondo de Cultura Económica, 2004.

[3] Aceves Pérez R. M., Villaseñor-Pineda L., and Montes-y-Gómez M. "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.

[4] Villatoro-Tello E., Villaseñor-Pineda L., and Montes-y-Goméz M. "Using Word Sequences for Text Summarization". LNCS (LNAI) 4188 (2006): 297.

[5] Sebastiani F. "Machine Learning in Automated Text Categorization". ACM computing Surveys 34 (March 2002): 1-47.

[6] Gliozzo A. and Strapparava C. "Exploiting Comparable 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.

[7] Rigutini L., Maggini M., and Bing L. "An EM based training algorithm for Cross-Language Text Categorization". Proceedings Web Intelligence Conference. Compiegne, France: IEEE, 2005.

[8] Bel N., Koster C., and Villegas M. "Cross-Lingual Text Categorization". 7th European Conference on Digital Libreries, ECDL. Trondheim Norway, 2003. 126-139.

[9] García Adeva J.J., Calvo R.A., and López de Ipiña D. "Multilingual Approaches to Text Categorisation". The European Journal for the Informatics Professional 6, no. 3 (2005): 43-51.

[10] Jalam R. Apprentissage automatique et catégorisation de textes multilingues. Lyon: PhD Tesis, Université Lumiere Lyon 2, 2003.

[11] Wei C.P, Shi H., and Yang C.C. "Feature Reinforcement Approach to Polylingual Text Categorization". 10th International Conference on Asian Digital Libreries, ICADL 2007, Lecture Notes in Computer Sciencie, Springer, 2007. 99-108.

[12] Olsson J.S.,. Oard D.W, and Hajic J. "Cross-Language Text Classification". SIGIR. Salvador, Brazil: ACM, 2005. 15-19.

[13] De Melo G. and Siersdorfer S. "Multilingual Text Classification using Ontologies". 29th European Conference on IR Research, ECIR 2007, Lecture Notes in Computer Sciencie. Springer, 2007.

[14] Lewis D.D., Yang Y., Rose T.G., and Li F. "RCV1: A New Benchmark Collection". Journal of Machine Learning Research 5 (2004): 361-397.

[15] Manning C.D. and Schütze H. Foundations of Statistical Natural Language. The MIT Press, 2001.

[16] Mitchell T.M. Machine Learning. McGraw-Hill, 1997.

[17] Escobar-Acevedo A., Montes-y-Gómez M., and Villaseñor-Pineda L., “Using Nearest Neighbor Information to Improve Cross-Language Text Classification”, MICAI 2009. Guanajuato, Mexico. Lecture Notes in Artificial Intelligence 5845, Springer.