Una Representación basada en Atributos Multilingües para el Agrupamiento de Documentos

Por

María Claudia Denicia Carral

M.C. INAOE

Tesis sometida como requisito parcial para obtener el grado de

Doctora en Ciencias Computacionales

en el Instituto Nacional de Astrofísica, Óptica y Electrónica.

Supervisada por

Dr. Manuel Montes y Gómez

Coordinación de Ciencias Computacionales INAOE

Dr. Luis Villaseñor Pineda

Coordinación de Ciencias Computacionales INAOE

A mis grandes amores

Agradecimientos

Agradezco a mis directores de tesis Dr. Manuel Montes y Dr. Luis Villaseñor por su apoyo y guía durante mis estudios, gracias docs. por todos sus consejos, paciencia y buen humor.

A mi comité doctoral conformado por: Dra Pilar Gómez, Dra. Claudia Feregrino, Dr Aurelio López, Dr. Carlos A. Reyes y Dr. David Pinto (BUAP) por su valiosa colaboración en el desarrollo de esta investigación.

A CONACyT por el apoyo asignado para la realización de esta investigación a través de la beca no. 165323

A mi familia, en especial a mi mamá María Elena por el apoyo y amor que me ha brindado siempre y por ayudarme a lograr mis metas; a Miguel y Ana por quererme y darme momentos de alegría.

A mí esposo Valentín por estar siempre a mi lado brindándome su paciencia, amor y cariño. A mi hija Regina Valentina, por soportar mi ausencia y darme siempre una sonrisa de aliento.

Resumen

El agrupamiento multilingüe de documentos (MDC, por sus siglas en inglés) es la tarea de organizar información dentro de grupos de documentos que están relacionados por un tema en común pero que pueden estar escritos en distintos idiomas. El principal reto a vencer en el agrupamiento multilingüe de documentos es la barrera lingüística del idioma que se presenta cuando se manejan distintos idiomas. Para salvar estas barreras y poder comparar documentos, aunque estén escritos en diferentes idiomas, es necesario disponer de herramientas que permitan obtener una representación de cada documento de manera independiente del idioma. En esta tesis se propone un método de representación que no depende de la traducción y que además es independiente de recursos lingüísticos externos. El método representa a los documentos por medio de parejas de palabras que guardan una relación temática. La representación primero fue evaluada en una colección bilingüe de documentos y posteriormente trasladada a una colección multilingüe mediante dos estrategias de agrupamiento. Los resultados obtenidos con la representación propuesta superan a métodos similares, lo que demuestra la pertinencia de ésta en la tarea de agrupamiento multilingüe. Las estrategias de agrupamiento multilingüe propuestas son pioneras en el desarrollo de representaciones independientes de la traducción que pueden manipular más de dos idiomas, porque la mayoría de los trabajos que hacen agrupamiento multilingüe solamente se quedan en un nivel bilingüe.

Abstract

Multilingual document clustering (MDC) involves partitioning documents, written in more than one language, into a set of thematically homogeneous groups. The major challenge facing MDC is achieving cross-lingual interoperability, which builds a bridge among representations of target documents written in different languages. This thesis proposes a translation independent approach especially suited to deal with linguistically related languages. In particular, it proposes representing the documents by pairs of words orthographically or thematically related. The representation was evaluated in a bilingual and multilingual corpus. The experimental evaluation in bilingual collections and using two clustering algorithms demonstrated the appropriateness of the proposed representation, which results are comparable to those from other approaches based on complex linguistic resources such as part-of-speech taggers and named entity recognizers. Proposed multilingual document clustering strategies represent a novel translation independent approach that manages more than two languages.

Contenido

Índice de Tablas

Índice de Figuras

Capítulo 1

Introducción

La era digital en la que vivimos ha propiciado que un mayor número de personas tengan la oportunidad de compartir información, principalmente en formato textual. Este fenómeno ha propiciado que el número de documentos electrónicos tenga un crecimiento continuo, dando como resultado bibliotecas digitales, bases de datos documentales, sitios web, etc. El acelerado crecimiento de la información textual hace que sean necesarios sistemas capaces de buscar y seleccionar la información útil para un usuario o una organización.

En la actualidad, gracias al fenómeno de globalización e internacionalización de los mercados, es inevitable el manejo de información en muchos idiomas, lo que agrega al problema de búsqueda y selección el reto de superar las barreras lingüísticas del idioma. En este contexto, los métodos capaces de manipular información multilingüe se han convertido en una herramienta vital para el manejo de la información.

La investigación en el manejo de datos multilingües ha recibido una atención especial por parte de la comunidad científica debido a que muchas organizaciones multinacionales, compañías, instituciones educativas, partidos políticos y usuarios comunes con conocimientos de otras lenguas, tienen la necesidad de manipular información en muchos idiomas. Algunos ejemplos de la necesidad de sistemas multilingües son los siguientes:

- *Gestión de información multilingüe*. Algunas organizaciones mundiales como la Organización de las Naciones Unidas o la Organización Mundial de la Salud, aunque con campos de actuación distintos,

comparten la necesidad de ofrecer la información que generan en muchos idiomas, así como también necesitan gestionar la información que reciben de distintos idiomas.

- *Análisis de información multilingüe.* Compañías multinacionales, cuentan con un departamento de análisis de noticias. Las motivaciones para contar con este tipo de departamento son muchas, pero el interés principal recae en observar cómo la compañía es visualizada alrededor del mundo. El área de análisis de noticias se encarga de revisar, ordenar y agrupar diariamente aquellas noticias, generadas alrededor del mundo, que hablan sobre la compañía.

- *Búsqueda y recuperación de información multilingüe.* Dado el creciente número de patentes registradas en múltiples países, los usuarios tienen la necesidad o se encuentran interesados en la recuperación de patentes en distintos idiomas. Sin embargo, muchos usuarios tienen dificultades para recuperar las patentes escritas en lenguas extranjeras.

El agrupamiento multilingüe de documentos (MDC, por sus siglas en inglés) es la tarea de organizar la información dentro de grupos de documentos que están relacionados por un tema en común pero que pueden estar escritos en distintos idiomas. El MDC es de gran ayuda en el manejo de información multilingüe porque ayuda a tener una mejor organización, que tiene como consecuencia un manejo más fácil de ésta.

Este documento de tesis se centra en el problema de MDC, el cual es de utilidad en el análisis de información multilingüe pues ayuda a la extracción de información relevante en documentos escritos en distintos idiomas, a la

organización de los resultados de sistemas de búsqueda y a la generación de resúmenes, entre otras aplicaciones.

1.1 Descripción del problema

El agrupamiento de documentos es la tarea de dividir un conjunto de documentos en grupos que contienen documentos de temática similar. En el caso multilingüe, los documentos se encuentra en distintos idiomas, y el objetivo es reunir aquellos documentos que comparten temática sin importar el idioma en el que se encuentren. El principal reto a vencer en el agrupamiento multilingüe de documentos es la barrera lingüística que se presenta cuando se manejan distintos idiomas. Para salvar esta barrera y poder comparar documentos, aunque estén escritos en diferentes idiomas, es necesario disponer de herramientas que permitan obtener una representación de cada documento de manera independiente del idioma.

Los primeros intentos para realizar agrupamiento multilingüe se centraron en el uso de técnicas de traducción (Chen and Lin,2000) (Leftin,2003). La idea general consiste en traducir los documentos o ciertas características distintivas de éstos a un único idioma, posteriormente con los documentos en un único idioma es posible aplicar cualquier técnica de agrupamiento monolingüe. Actualmente, se han explorado nuevas técnicas que permiten realizar agrupamiento multilingüe sin el uso de técnicas de traducción. El objetivo es buscar un espacio común de representación de los documentos sin recurrir a la traducción. Para generar esta representación, la mayoría de los trabajos que realizan agrupamiento multilingüe sin técnicas de traducción parten de la idea de que existen palabras que se conservan sin cambios entre los lenguajes involucrados. Ejemplos de éstas son fechas, números, nombres de personas o lugares, etc. Este tipo de palabras conforma un vocabulario común entre documentos de distintos idiomas. A través de ellas cada uno de los documentos en una colección multilingüe es representado

y mediante un algoritmo de agrupamiento son reunidos documentos de acuerdo a las similitudes observadas en este vocabulario común.

La extracción de estas palabras generalmente se realiza en corpus paralelos, es decir, un conjunto de documentos que contiene los mismos documentos traducidos a distintos idiomas y que se encuentran alineados a nivel frase u oración. En estos corpus la cantidad de atributos comunes es considerable debido a que se comparten los mismos nombres, lugares y fechas. La principal desventaja de estos métodos es que los corpus paralelos son difíciles de crear; trabajos como el de (Michel et al., 1999) muestran los retos a vencer en el alineamiento de corpus.

Para tratar de resolver la desventaja que presenta la obtención de corpus paralelos se utilizan corpus comparables, que contienen textos en distintos idiomas, que sin ser traducciones, comparten similar origen, temática y extensión. Es decir, que los textos no se reúnen de manera arbitraria, sino que se escogen de acuerdo a criterios comunes de selección. Los corpus comparables se acercan más a la realidad de organización de la información; ejemplos de estos corpus son las colecciones de noticias que son ampliamente utilizadas en la investigación de agrupamiento de documentos. Sin embargo, cuando los métodos independientes de la traducción son aplicados a corpus comparables, surgen los siguientes problemas:

- La cantidad de características comunes se torna cada vez menor cuando el número de idiomas aumenta. La carencia de características provoca que los documentos no puedan ser correctamente representados, lo que deriva en malos resultados en el agrupamiento. Este problema ha sido poco abordado, porque los esfuerzos se han enfocado en resolver el problema bilingüe suponiendo que las técnicas que funcionan bien en dos idiomas tendrán un comportamiento similar cuando el número de idiomas aumenta. Sin embargo, esta suposición no es del todo verdadera porque la

aplicación de los métodos bilingües requiere del diseño de estrategias que combinen adecuadamente los métodos bilingües, de tal forma que conserven los resultados obtenidos en el agrupamiento bilingüe.

- Características como entidades nombradas[1] no son útiles en estos corpus porque, aunque en la mayoría de los corpus son abundantes, éstas no se comparten entre distintos idiomas aunque la temática sea similar. Por ejemplo, si se quieren agrupar noticias sobre deportes, es muy probable que los nombres de los jugadores, los lugares y fechas en donde se realizan los juegos sean diferentes entre distintos idiomas. Este hecho hace que las entidades no sean de utilidad ya que no ayudan a establecer cuando dos documentos pertenecen a la misma temática.

A partir de los problemas antes mencionados, surgen las siguientes preguntas de investigación *¿Cómo aumentar el conjunto de características sin recurrir a técnicas de traducción?, ¿Cómo hacer la extracción de características independiente de recursos, tales como corpus comparables?* y *¿Qué estrategias son necesarias para trasladar las soluciones bilingües a un ambiente multilingüe?*

En este trabajo de tesis se propone un método de agrupamiento independiente de la traducción. El principal objetivo de esta tesis es crear una representación de los documentos a través de características obtenidas sin técnicas de traducción. Además, se afronta el problema del escaso número de características comunes proponiendo una estrategia para enriquecer la representación. Por último, este trabajo también propone abordar escenarios multilingües presentando dos estrategias de agrupamiento para dicha situación.

[1] Una entidad nombrada (EN) es una palabra que denota un objeto que puede caer en una de las siguientes categorías generales: persona, organización, lugar, fecha y cantidad.

1.2 Objetivos

El objetivo general que se persigue es:

Proponer una representación de textos orientada a la tarea de agrupamiento de documentos utilizando atributos multilingües obtenidos sin recurrir a recursos externos .

Los objetivos específicos que se plantean son los siguientes:

- Proponer representaciones de documentos a través de características bilingües y/o multilingües.

- Desarrollar métodos independientes de la traducción para la extracción de las características propuestas.

- Evaluar estrategias de agrupamiento multilingüe usando las características bilingües y/o multilingües extraídas.

1.3 Estructura de la Tesis

El documento se organiza de la siguiente manera:

- ***Capítulo 2***. En este capítulo se presentan los conceptos básicos necesarios para introducir al lector en el contexto de esta tesis. En específico se presentan las etapas de un sistema de agrupamiento de documentos, el modelo de representación de documentos, los algoritmos de agrupamiento utilizados y las medidas para evaluar los resultados.

- ***Capítulo 3***. La revisión del estado del arte de los métodos de agrupamiento multilingüe es mostrada en este capítulo. Los métodos se dividen en los basados en traducción y los independientes de la traducción. En este capítulo se revisan las ventajas y desventajas que presentan cada uno de ellos, así como los avances alcanzados.

- ***Capítulo 4***. En este capítulo se presenta la representación propuesta para el agrupamiento bilingüe. Se detallan los métodos de extracción de las características que representarán a los documentos y se muestran los resultados experimentales.

- ***Capítulo 5***. En este capítulo se describe el agrupamiento multilingüe. Se detallan las estrategias aplicadas y las características generadas para lograr el propósito, también se presentan los resultados obtenidos.

- ***Capítulo 6.*** Finalmente, en este capítulo se muestra un resumen de la investigación desarrollada, las aportaciones obtenidas y las direcciones futuras del trabajo de tesis.

Capítulo 2

Agrupamiento de documentos

Este capítulo tiene como objetivo introducir los conceptos básicos para entender el resto del documento. El capítulo se organiza de la siguiente forma: en la sección 2.1 se explica la tarea de agrupamiento de documentos. En la sección 2.2 se describe el modelo de espacio vectorial; dicho modelo es el utilizado en esta tesis para representar los documentos. En la sección 2.3 se describe la medida de similitud para comparar dos documentos; en específico se detalla la medida coseno que es la utilizada en los experimentos desarrollados en este documento. En la sección 2.4 se hace una breve revisión de los tipos de algoritmos de agrupamiento existentes y se explican a detalle los utilizados en este trabajo. Finalmente, en la sección 2.5 se revisan las medidas de evaluación para el agrupamiento y se explican a detalle la medida *F-measure* con la cual se evaluaron los experimentos de esta investigación.

2.1 Agrupamiento de documentos

El agrupamiento de documentos se empezó a investigar dentro de la Recuperación de Información (IR, por sus siglas en inglés) que se encarga de localizar documentos relevantes de acuerdo a una petición de información, llamada consulta. En su forma más simple las consultas están dirigidas a determinar que documentos poseen determinadas palabras en su contenido. Por ejemplo, la consulta *"anatomía humana"*, podría tener como resultado aquellos documentos que contengan a las palabras *"anatomía"* y *"humana"* como parte de su contenido.

En el contexto de la recuperación, Van Rijsbergen (1979), formuló la denominada *"Hipótesis de Agrupamiento"* que sostiene que *"documentos fuertemente asociados tienden a ser relevantes para la misma consulta"*. Basándose en dicha hipótesis, el agrupamiento de documentos tiene en cuenta el contenido de los documentos para agruparlos, ya que los documentos similares contendrán palabras similares.

A partir de esta hipótesis, el agrupamiento de documentos puede verse como la tarea de dividir un conjunto de documentos en grupos de temática similar; dicha temática delimitada por las palabras que contiene cada documento. Es decir, los documentos de un grupo deberán tener el mayor número de palabras en común y el menor número de palabras comunes con otros grupos.

Para definir formalmente el agrupamiento, se han propuesto distintas definiciones, todas ellas generalmente basadas en el término grupo o *cluster*. La definición que utilizaremos en esta tesis es la propuesta por Theodoridis and Koutroumbas (1999) que se presenta a continuación:

Definición 1. *Sea* $D = \{d_1, d_2, \dots, d_N\}$ *un conjunto de documentos. Se define un m-agrupamiento de D como una partición de* D *en m grupos o clases* $c_1, c_2, \dots, c_m$*, de forma que se cumplan las siguientes condiciones:*

- $c_i \neq \emptyset, i = 1, \dots, m$
- $\bigcup_{i=1}^{m} c_i = D$
- $c_i \cap c_j = \emptyset, i \neq j,\ i, j = 1, \dots, m$

Además se busca que los elementos de un grupo c_i sean más similares entre sí y menos similares a los elementos de otros grupos.

De acuerdo a la definición anterior cada documento debe pertenecer a un único grupo. Este tipo de agrupamiento es conocido como "*hard clustering"*.

Los pasos básicos que caracterizan a un proceso de agrupamiento son (Murty, and Flynn, 1999) (Theodoridis and Koutroumbas, 1999): construcción de la representación de los documento, medición de la similitud entre todos los documentos de la colección y la aplicación del algoritmo de agrupamiento. En las siguientes secciones se explican cada uno de estos pasos, poniendo especial énfasis en las técnicas utilizadas en esta investigación.

2.2 Representación de los documentos

El modelo de representación más utilizado es el modelo de espacio vectorial propuesto por Salton et al. (1975) que está basado en la idea de que las palabras de un documento son una representación razonable de su contenido. En este modelo no hay información sobre el orden de las palabras, por lo tanto, esta representación puede verse como una "*bolsa de palabras"*. Es claro que el contenido real de un documento va más allá que un conjunto de palabras; aun así este modelo es ampliamente usado en agrupamiento de documentos, tanto monolingüe como multilingüe, ya que ha demostrado su eficacia en estas y otras tareas.

La idea básica del modelo de espacio vectorial reside en la construcción de una matriz de palabras y documentos, donde las columnas y las filas representan a las palabras y a los documentos respectivamente. Así, las filas de esta matriz son equivalentes a los documentos que se expresan en función de las apariciones de cada término. Formalmente podemos definir la representación de documentos en el Modelo de Espacio Vectorial (VSM) de la siguiente forma (Gliozzo and Strapparava, 2001)*:*

Definición 2*. Sea* $D = \{d_1 d_2 \cdots d_n\}$ *una colección de documentos, y* $W = \{w_1 w_2 \cdots w_k\}$ *su vector de características. En el agrupamiento de documentos, el VSM es un espacio k-dimensional* $\mathbb{R}^k$*, en el cual el documento* d_j *es representado*

por medio de un vector $\overrightarrow{d_j}$ *tal que el* z^{th} *componente de* $\overrightarrow{d_j}$ *es la importancia de la palabra* w_z *en el documento* d_j.

Un conjunto de documentos en el modelo de espacio vectorial puede verse como una matriz *P,* tal como se muestra en la Figura 1. Cada componente $p_{i,j}$ de la matriz representa el peso de la palabra *j* en el documento *i*.

		Palabras				
		w_1	w_2	$\cdots$	w_{k-1}	w_k
	d_1	$p_{1,1}$	$p_{1,2}$	$\cdots$	$p_{1,k-1}$	$p_{1,k}$
	d_2	$p_{2,1}$	$p_{2,2}$	$\cdots$	$p_{2,k-1}$	$p_{2,k}$
Documentos		$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	d_{n-1}	$p_{n-1,1}$	$p_{n-1,2}$	$\cdots$	$p_{n-1,k-1}$	$p_{n-1,k}$
	d_n	$p_{n,1}$	$p_{n,2}$	$\cdots$	$p_{n,k-1}$	$p_{n,k}$

Figura 1. Matriz del modelo de espacio vectorial

Para calcular el peso $p_{i,j}$ de cada término o palabra se utilizan distintas medidas. En el caso más simple, pueden aplicarse exclusivamente valores binarios; de forma que si en el documento d_i aparece la palabra w_j, el valor de $p_{i,j}$ sera 1 y en caso contrario, 0. Como, naturalmente una palabra puede aparecer más de una vez en el mismo documento, y, como además, unas palabras pueden considerarse como más significativas que otras, el valor numérico p_{ij} generalmente se calcula con métodos más sofisticados que tienen en cuenta otros factores, además de la simple ocurrencia o no de un término. En (Salton and Buckley, 1988) se pueden consultar distintos esquemas de peso.

Las medidas más usadas en el agrupamiento de documentos son la frecuencia del término (*tf*) y el pesado *tf-idf* (Salton and Buckley, 1988). En el pesado *tf* a cada palabra en un documento se le asigna un peso proporcional a la cantidad de veces que éste aparece en dicho documento. Generalmente estas frecuencias de aparición son normalizadas con el objetivo de que el peso asociado no dependa de

la frecuencia relativa de esta palabra con otras palabras y para evitar favorecer a documentos más largos que otros. El tipo de normalización más utilizado es la normalización por longitud del documento, que consiste en dividir la frecuencia de la palabra por la longitud del documento.

El esquema de pesado utilizado en esta tesis es el *tf-idf*, que combina la frecuencia de la palabra (*tf*), con la frecuencia de la palabra en la colección de documentos, representado por el factor *idf*. El factor *idf* da un estimado de la importancia de la palabra para describir a un documento en función de cuantos documentos en la colección contienen a dicha palabra; es decir, entre menos documentos contengan a la palabra mayor será la utilidad de ésta para describir a un documento. El peso *tf-idf* se define de la siguiente forma:

$$p_{i,j} = tf(w_j, d_i) \times idf(w_j)\ , idf(w_j) = log\frac{N}{df(w_j)} \qquad (2.1)$$

Donde $tf(w_j, d_i)$ representa la frecuencia de la palabra w_j en el documento d_i, $df(w_j)$ es el número de documentos que contienen a la palabra w_j y N es el número de documentos de la colección.

2.3 Medición de la similitud

Para aplicar un algoritmo de agrupamiento, es necesario determinar el grado de asociación entre todas las parejas de documentos de la colección. Con este propósito, se pueden utilizar distancias, o medidas de similitud y disimilitud. Algunos algoritmos de agrupamiento tienen un requerimiento teórico para el uso de una medida específica, pero lo más común es que el investigador(a) seleccione la medida que utilizará dependiendo del método de agrupamiento seleccionado. La mayoría de los estudios coinciden en que en el agrupamiento de documentos es

adecuado utilizar medidas como Coseno, coeficiente Dice o Jaccard (Baeza-Yates and Ribeiro-Neto, 1999).

La medida de similitud utilizada en esta tesis para comparar dos documentos es la similitud coseno propuesta por Salton (1989). La idea de esta medida es calcular el coseno del ángulo entre el vector del documento d_1 y el vector del documento d_2. Esta medida es ampliamente utilizada en trabajos de agrupamiento de documentos, y su popularidad se debe a que es posible obtener una representación geométrica del modelo vectorial. La fórmula de la similitud coseno es:

$$s_{1,2} = cos(d_1, d_2) = \frac{\sum_{l=1}^{k} p_{l,d_1} \times p_{l,d_2}}{\sqrt{(\sum_{l=1}^{k} p_{l,d1}^2)} \times \sqrt{(\sum_{l=1}^{k} p_{l,d2}^2)}} \qquad (2.2)$$

Donde d_1 y d_2 representan los documentos a comparar; p_{l,d_1} representa el peso de la palabra *l* en el documento d_1; y p_{l,d_2} representa el peso de la palabra *l* en el documento d_2. Los valores obtenidos por la similitud coseno oscilan entre 0 y 1. Un valor cero (ángulo de 90°) indica que los documentos comparados son completamente distintos de acuerdo a las características con las que fueron representados, de manera contraria un valor 1 (ángulo de 0°) indica que los documentos son similares en las características.

Si la colección contiene N documentos, entonces se genera una matriz de similitud S de tamaño $N \times N$ como resultados de la comparación de los documentos. Ésta es una matriz triangular, en la cual $s_{i,j}$ representa la similitud de los documentos d_i y d_j. Los elementos de la diagonal son iguales al máximo valor obtenido por la medida de similitud; en el caso de coseno la matriz tendrá una diagonal llena de 1's. En la Figura 2 se puede observar la matriz de similitud para un conjunto de N documentos; la parte superior derecha se muestra en blanco, porque la matriz es simétrica al igual que la función de similitud, es decir, $(s_{i,j} = s_{j,i})$.

		Documentos				
		d_1	d_2	$\cdots$	d_{n-1}	d_n
	d_1	1				
	d_2	$s_{2,1}$	1			
Documentos	$\vdots$	$\cdots$	$\cdots$	$\cdots$		
	d_{n-1}	$s_{n-1,1}$	$s_{n-1,2}$	$\cdots$	1	
	d_n	$s_{n,1}$	$s_{n,2}$	$\cdots$	$s_{n,n-1}$	1

Figura 2. Matriz de similitud de N documentos

2.4 Algoritmos de agrupamiento

Aunque existen una gran cantidad de algoritmos de agrupamiento, éstos pueden clasificarse en dos clases principales que dependen de la estructura de agrupamiento generada. Esta clasificación está conformada por los algoritmos jerárquicos y los de partición (Manning and Schütze, 2000), (Jain et al., 1999). La Figura 3 muestra esta división.

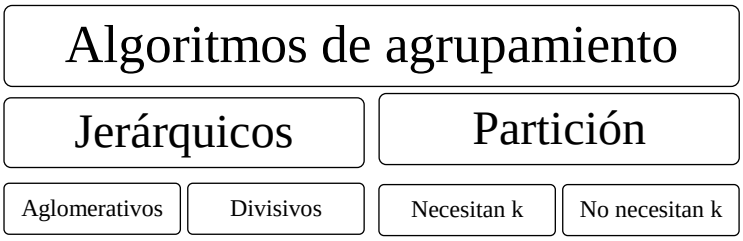


Figura 3. Clasificación de los algoritmos de agrupamiento (Manning and Schütze, 2000), (Jain et al., 1999)

Algoritmos jerárquicos

Los algoritmos jerárquicos se caracterizan por generar una estructura de árbol, llamada dendograma, en la que cada nivel es un agrupamiento posible de los objetos de la colección (Jain et al., 1999). Cada vértice o nodo del árbol es un

grupo de documentos. La raíz del árbol, que es el primer nivel, se compone de un único grupo que contiene todos los documentos. Cada hoja del último nivel del árbol es un grupo compuesto por un único documento, hay tantas hojas como documentos tenga la colección. En los niveles intermedios, cada nodo del nivel *n* es dividido para formar sus hijos del nivel *n* + 1.

La figura 4(b) muestra el dendograma generado para la colección de elementos de la figura 4(a) y el agrupamiento obtenido en cada nivel.

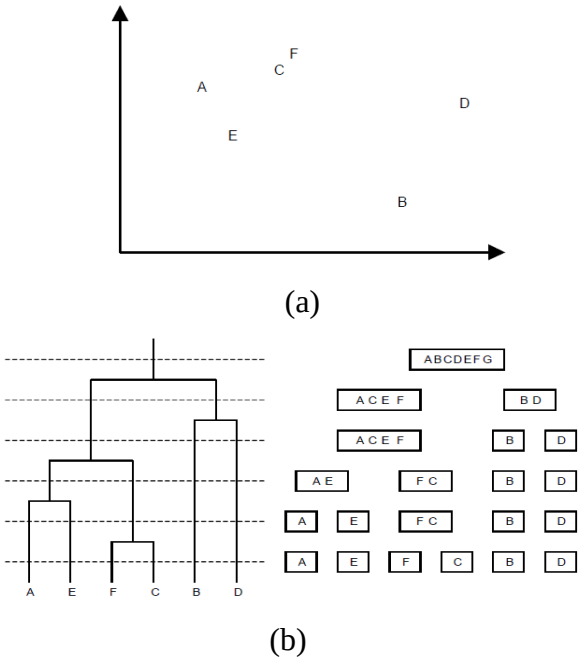


Figura 4. Ejemplo de dendograma (Jain et al., 1999)

De acuerdo a la metodología que se aplica para obtener el dendograma, los algoritmos jerárquicos pueden dividirse en *aglomerativos* y *divisivos*. A continuación se explica de forma general en qué consisten estos dos tipos de algoritmos:

- *Aglomerativos*: Estos algoritmos parten de las hojas del árbol, considerando a cada documento como un grupo. De forma iterativa se van uniendo los grupos más cercanos. Este procedimiento continúa hasta que todos los documentos se encuentran dentro de un grupo o hasta que ciertas condiciones de paro son cumplidas.

- *Divisivos*: El algoritmo de división supone que en un primer paso todos los documentos conforman un único grupo, es decir, inicia desde la raíz del árbol. Este grupo se va dividiendo sucesivamente en grupos más pequeños hasta que cada documento constituye un grupo o hasta que se cumplan ciertos criterios de paro.

Algunos algoritmos que pertenecen a esta categoría son: CURE (Clustering Using Representatives) (Guha et al. 1998), CHAMALEON (Karypis et al., 1999), BIRCH (Balanced Iterative Reducing and Clustering using Hierarchical) (Zhang et al., 1996) y ROCK (RObust Clustering algorithm using linKs) (Guha et al., 2000).

Algoritmos de partición

El agrupamiento de partición, a diferencia del agrupamiento jerárquico, no genera distintos niveles de agrupamiento de los documentos, sino que trabaja en un sólo nivel, en el que se refina el agrupamiento. Los algoritmos particionales se dividen en dos grupos: los que necesitan que se les indique la cantidad de grupos a formar y los que no requieren esta información. Es importante resaltar que los algoritmos del último grupo inducen de manera natural el número de grupos. Los métodos particionales son los más utilizados en el agrupamiento de documentos, ya que la construcción de un dendograma en grandes volúmenes de información resulta costosa.

Algunos algoritmos de clustering que pertenecen a esta clasificación son: K-Means (Huang, 1998), K-prototypes (H., 1995), *Estrella* (Aslam et al., 1999), CLARA (Clustering Large Applications) (Kauffman and Rousseuw, 1990), CLARANS (Clustering Large Applications based on Randomized Search) (Han, 1994). En esta tesis se utilizan dos algoritmos de agrupamiento, *Estrella* y k-means, ambos algoritmos particionales. No es necesario indicarle al algoritmo *Estrella* el número de grupos que se generarán, el *k-means* tiene como requerimiento que se indique el número de grupos.

La elección del algoritmo *Estrella* se debe a que en ambientes documentales ha obtenido buenos resultados. Por otro lado, el algoritmo *k-means* ha sido usado por otros trabajos de agrupamiento multilingüe que no dependen de recursos de traducción, obteniendo buenos resultados. Por esta razón este algoritmo sirve como punto de comparación con otros trabajos de agrupamiento multilingüe.

2.4.1 Agrupamiento *Estrella*

El algoritmo de agrupamiento *Estrella* fue propuesto por Aslam et al. (1999). Éste es un algoritmo particional que se basa en la teoría de grafos. Este algoritmo induce de manera natural el número de grupos y la estructura de los temas dentro del espacio de los documentos. El algoritmo *Estrella* ha sido usado con éxito en trabajos como el de (Pérez-Suarez et al. 2008), . Para entender el funcionamiento del algoritmo *Estrella* es necesario conocer las siguientes definiciones:

***Definición 3**. Sea $D = \{d_1 d_2 \cdots d_n\}$ una colección de documentos y $sim(d_i, d_j)$ la similitud entre dos documentos. Un grafo de similitud G=(V,E,p) es un grafo no dirigido etiquetado en el cual los vértices V representan a los documentos de la colección, cada arista ponderada E representa la $sim(d_i, d_j)$ y p es la función que asigna los pesos de las aristas.*

Definición 4. *Sea σϵR, tal que σϵ[0,1], un umbral de similitud y G un grafo de similitud. Un grafo de σ-similitud se denota por G_σ=(V,Eσ) y es el grafo no dirigido que se obtiene a partir de G si se eliminan todas las aristas eϵE tal que el peso de la arista* $p(e) < \sigma$.

El algoritmo *Estrella* construye un conjunto de grupos *G* a partir del grafo G_σ utilizando sub-grafos en forma de *Estrella*; cada uno de estos sub-grafos determinan un grupo de *G*. Un sub-grafo en forma de *Estrella* es un sub-grafo de *m+1* vértices, en el cual existe un vértice llamado *centro,* denotado por *C,* y *m* vértices denominados satélites, denotados por S_i, en el cual se cumple que: (1) el centro tiene un grado mayor o igual que el resto de los vértices del sub-grafo y (2) existe una arista del centro a cada uno de los satélites. La Figura 5 muestra un ejemplo de un sub-grafo en forma de *Estrella*, con vértice centro C y vértices satélite s1 hasta s7; las aristas son denotadas por líneas sólidas y líneas punteadas. Las líneas sólidas muestran que siempre existen aristas entre el centro y cada uno de los satélites y las líneas punteadas demuestran que además también pueden existir aristas entre los satélites.

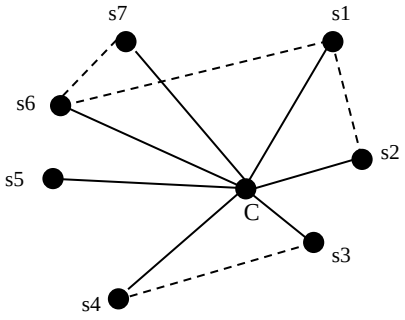


Figura 5. Ejemplo de un grupo con centro C y siete satélites s1 hasta s7

El algoritmo garantiza que la similitud entre el centro y los satélites de la *Estrella* sea al menos σ, pero no garantiza que la similitud entre dos satélites

alcance este valor. Sin embargo, en (Aslam et al., 2004) se demuestran que la similitud entre los satélites es alta y que los sub-grafos *Estrella* conforman grupos con una similitud promedio entre los elementos que componen la *Estrella*. En el trabajo de Aslam et al. (2004), se da evidencia matemática de esta similitud e incluso es posible determinar la similitud esperada entre los satélites.

El algoritmo *Estrella* se muestra en la Figura 6. Este algoritmo no necesita conocer el número de grupos a obtener sino que los descubre a través de las propias características de los documentos de la colección.

Dado un umbral σ:

1. Calcular $G_\sigma=(V,E_\sigma)$, donde $E_\sigma = e \in E: p(e) \geq \sigma$.

2. Poner cada vértice en G_σ inicialmente marcado como "no visitado".

3. Calcular el grado de cada vértice $v \in V$.

4. Tomar el vértice de mayor grado que tenga la etiqueta de "no visitado" como centro de la *Estrella* y construir un grupo con este vértice como centro de la *Estrella* y sus satélites como vértices asociados. Marcar cada nodo de la *Estrella* como "visitado".

5. Repetir el paso 4 hasta que todos los vértices estén marcados como "visitados"

6. Representar cada grupo por medio del documento correspondiente al centro de cada *Estrella*.

Figura 6. Algoritmo *Estrella*

2.4.2 Agrupamiento *k-means*

El agrupamiento por medio de *k-means* propuesto por (Mac-Queen, 1967). Es un algoritmo que sigue una forma fácil y simple para dividir un conjunto de

documentos en *k* grupos. El número de grupos *k* en el cual se divide la colección de documentos debe ser especificado por el usuario.

La Figura 7 muestra el algoritmo *k-means*. El primer paso del algoritmo es la inicialización, que consiste en la selección de los centros para cada uno de los *k* grupos iniciales. La selección de los centros se puede determinar de distintas maneras, siendo la más sencilla la elección aleatoria (Fung, 2001). Cuando los centros han sido seleccionados; se asignan cada uno de los documentos del corpus al grupo que tiene el centro más cercano. Entonces, los centros son actualizados, es decir, se recalcula el centro de cada grupo y todos los documentos se distribuyen de acuerdo a los nuevos centros. Este proceso se repite hasta que se han cumplido ciertas condiciones de parada, por ejemplo, cuando se alcanza un cierto número de iteraciones, cuando los centros ya no han cambiado o cuando el error cuadrático no tenga cambios significativos.

2.5 Evaluación del agrupamiento

Para evaluar la calidad del agrupamiento generado de forma automática se han desarrollado diversas métricas de evaluación; éstas se pueden agrupar en dos categorías: las supervisadas y las no supervisadas. A continuación se detallan cada una de estas categorías.

No supervisadas: Estas medidas permiten evaluar la calidad del agrupamiento cuando no se tiene información de cuál es el grupo al que realmente pertenece cada documento, es decir, no se cuenta con una solución manual generada previamente. Estas medidas calculan la calidad en función de la *cohesión interna y la separación externa de los grupos*, las cuales determinan que tanto se parecen los documentos dentro de un mismo grupo y que tan diferentes son éstos de los documentos de otro grupo. Ejemplos de estas medidas son el error cuadrático medio, la similitud global y el coeficiente Silhouette (Tan, 2006).

Dado un número de grupos k y un conjunto de documentos $D = \{d_1 d_2 \cdots d_n\}$

1. Inicializar k documentos $(x_1, \ldots, x_k)$ como centros, tales que:

$$x_j = d_l, \quad j \in \{1, \ldots, k\}, \quad l \in \{1, \ldots, n\}$$

2. Asociar cada centro x_j a un grupo C_j
3. Repetir
   3.1 Para cada documento d_l, con $l \in \{1, \ldots, n\}$
   Hacer
   Asignar d_l al grupo C_{j*} más cercano y cuyo centro es x_{j*}, es decir, $|d_l - x_{j*}| \leq |d_l - x_j|, j \in \{1, \ldots, k\}$

   3.2 Para cada grupo C_j donde $j \in \{1, \ldots, k\}$
   Hacer
   Recalcular el centro x_j para que sea el centro de todos los documentos actualmente en C_j, de modo que $x_j = \sum_{d_l \in C_j} \frac{d_l}{|C_j|}$

   3.3 Calcular la función de error cuadrático

$$E = \sum_{j=1}^{k} \sum_{i_l \in C_j} |i_l - w_j|^2$$

   Hasta
   Que los centros no cambien o E no tenga cambios significativos.

Figura 7. Algoritmo k-means

Las medidas no supervisadas se basan en la matriz de similitud de documentos, que es generada a partir de la representación de éstos. Estas medidas se enfocan en comparar los resultados obtenidos por distintos algoritmos de agrupamiento y suponen que la representación de los documentos y, por consecuencia la matriz de similitud, no cambia. En esta tesis se realizaron distintas representaciones de los documentos, por lo que la aplicación de esta clase de medidas no es posible debido

a que la matriz de similitud cambia de una representación a otra. Por otro lado, el objetivo primordial de la evaluación presentada en esta tesis es comparar el desempeño de las representaciones y no el funcionamiento de los algoritmos de agrupamiento.

Supervisadas: Estas medidas determinan la calidad del agrupamiento comparando los grupos obtenidos con un conjunto de clases previamente definidas y que generalmente son determinadas manualmente por expertos. Ejemplos de estas medidas son la entropía, pureza, coeficiente de *Jaccar*d, *F-measure* (Tan, 2006).

Este tipo de medidas no dependen de la matriz de similitud sino que se basan en un agrupamiento manual, que es considerado como correcto. Estas medidas pueden ser utilizadas para comparar el desempeño de distintos algoritmos de agrupamiento, así como para comparar distintas representaciones de los documentos. Por esta razón y debido a que los conjuntos de documentos con los que se realizaron los experimentos de esta tesis, cuentan con una solución manual, se eligieron este tipo de medidas para las representaciones desarrolladas. En específico se eligió la *F-measure*, que es ampliamente utilizada en el campo de agrupamiento de documentos. A continuación se explica a detalle esta medida.

2.5.1 *F-measure*

La *F-measure* es una medida de evaluación supervisada propuesta por Van Rijsbergen en 1974. Originalmente fue diseñada para evaluar a los sistemas de recuperación de información (IR, por sus siglas en inglés), y posteriormente adaptada para evaluar los resultados de un algoritmo de agrupamiento. La IR tiene como objetivo encontrar documentos relacionados con una consulta hecha por un usuario. Esta medida se basa en los conceptos de precisión y recuerdo. El recuerdo es el número de documentos relevantes que se recuperan del conjunto de

documentos. La precisión se refiere al número de documentos recuperados que son relevantes a la información que necesita el usuario.

Cuando la *F-measure* es trasladada al agrupamiento se compara una solución ideal generada por un humano con la solución generada por el algoritmo de agrupamiento. Los grupos de la solución manual, generalmente son llamados *clases*, mientras que los grupos generados automáticamente son nombrados *clusters*. La *F-measure* combina las medidas de precisión y recuerdo de la siguiente forma:

$$F(i,j) = \frac{2 \times Recall(i,j) \times Precision(i,j)}{Precision(i,j) + Recall(i,j)} \qquad (2.3)$$

Donde el $Recall(i,j) = \frac{n_{ij}}{n_i}$ y la $Precision(i,j) = \frac{n_{ij}}{n_j}$, n_{ij} es el número de elementos de la clase *i* en el *cluster j*, n_i es el número de elementos de la clase *i* y n_j es el número de elementos del *cluster j*. Para evaluar todos los *clusters* se utiliza la siguiente fórmula:

$$F = \sum_i \frac{n_i}{n} max_j\{F(i,j)\} \qquad (2.4)$$

Si la *F-measure* es más cercana a 1 indica que la estructura del agrupamiento obtenido de forma automática es muy parecida a la estructura del agrupamiento de la solución humana, de manera contraria un *F-measure* cercano a cero indica que el agrupamiento obtenido tiene una menor coincidencia con la solución humana.

Capítulo 3

Agrupamiento multilingüe

En este capítulo se muestra una revisión del estado actual de los sistemas de agrupamiento multilingüe. En la sección 3.1 se describen los enfoques para atacar el problema. El primer enfoque trata sobre los sistemas que utilizan técnicas de traducción para agrupar los documentos, explicado en la sección 3.2. El segundo enfoque comprende a los sistemas que no requieren de técnicas de traducción, este enfoque se detallan en la sección 3.3. Finalmente en la sección 3.4 se presenta una discusión final de los métodos de aplicado en ambientes multilingües.

3.1 Enfoques

Los enfoques para atacar el problema de agrupamiento multilingüe de documentos pueden dividirse en dos grupos: los basados en recursos de traducción y los independientes de estos recursos. El primer enfoque puede dividirse de acuerdo al uso de las técnicas utilizadas para la traducción en los que utilizan traducción automática y los que utilizan recursos multilingües como diccionarios o tesauros multilingües. Este enfoque también puede dividirse tomando en cuenta a aquellos que traducen el documento completo y los que traducen algunos rasgos distintivos de éste.

El segundo enfoque, independiente de los recursos de traducción, se basa en representar los documentos sin utilizar técnicas de traducción. Los métodos desarrollados pueden dividirse de acuerdo a las técnicas utilizadas para seleccionar los atributos de representación. Por un lado, existen los métodos que toman como atributos características comunes y por otro lado los que obtienen dichos atributos

a partir de corpus paralelos y posteriormente los utilizan para representar los documentos de una colección bilingüe.

Un aspecto a tomar en cuenta cuando se realiza agrupamiento multilingüe es la estrategia para realizar el agrupamiento. Estas estrategias pueden dividirse en dos tipos:

- *Un paso*. Consiste en abordar desde un principio la obtención de grupos multilingües a través de una representación multilingüe de los documentos en todos los lenguajes considerados.
- *Dos pasos*. En este caso se parte de un agrupamiento monolingüe por idioma y posteriormente los grupos monolingües son unidos mediante características multilingües, dando lugar a los grupos multilingües.

En las siguientes secciones de este capítulo se describen los sistemas de agrupamiento multilingüe. Primero se abordan los sistemas que utilizan alguna técnica de traducción para representar a los documentos y después aquellos que no utilizan estas técnicas.

3.2 Métodos basados en traducción

El objetivo de los sistemas basados en traducción es agrupar una colección de documentos escritos en distintos idiomas a través de una representación común del conjunto de documentos obtenida por medio de la traducción a un único idioma, llamado *idioma pivote*. Estos métodos pueden dividirse, de acuerdo a la técnica utilizada para realizar la traducción, en aquellos que utilizan traductores automáticos (Chen and Lin, 2000) (Leftin, 2003) y los que utilizan recursos multilingües como diccionarios bilingües (Mathieu et al., 2004), (Rauber et al., 2001) (Saralegi Urizar and Alegria Loinaz, 2007) o tesauros (Steinberger et al., 2002), (Cobo Ortega and Rocha Blanco, 1999), (Pouliquen et al., 2004).

La forma más directa en la que trabajan los enfoques de traducción consiste en traducir el documento completo a un idioma pivote; de esta forma el problema multilingüe se trata con métodos monolingües. Un trabajo que utiliza esta técnica es el desarrollado por Leftin (2003), en el cual se presenta un sistema para generar resúmenes multi-documento obtenidos automáticamente de noticias en inglés y ruso. El agrupamiento multilingüe de documentos es una parte importante en la generación de este tipo de resúmenes, ya que, permite seleccionar los documentos similares de los cuales se obtendrá el resumen. El principal inconveniente cuando se traduce un documento completo es que la traducción se hace literal, es decir término a término. Después de que el documento ha sido traducido se genera una representación que utiliza sólo las palabras más representativas del conjunto de documentos.

La traducción completa generalmente no es conveniente, por tal motivo algunos otros trabajos se dan a la tarea de seleccionar las palabras que son más relevantes a la colección de documentos y posteriormente traducir solamente este conjunto, evitando la traducción del documento completo. Para hacer esta selección de atributos se utilizan distintos métodos. Por ejemplo, en (Rauber et al. 2001) y (Mathieu et al., 2004) se traducen solamente las palabras que sobrepasan un umbral de frecuencia. Otros en cambio, seleccionan como características representativas a algunas categorías gramaticales o a entidades nombradas (Chen and Lin, 2000), (Urizar and Loinaz, 2007).

La traducción de las características seleccionadas, generalmente, se realiza por medio de un diccionario bilingüe. Los diccionarios bilingües reducen la complejidad de traducir el documento completo a la traducción de palabras individuales. Sin embargo, el uso de los diccionarios presenta dos tipos de inconvenientes: la cobertura y la ambigüedad. La cobertura del diccionario puede no ser completa, por lo que algunos términos no pueden ser traducidos. Ejemplos de ellos son los términos técnicos pertenecientes a un dominio en particular, los nombres de personas o de lugares, etc.; este tipo de traducciones no están

contempladas en un diccionario de uso común (Rauber et al., 2001). Para afrontar estos problemas, trabajos como el de (Urizar and Loinaz, 2007), se dan a la tarea de construir sus propios diccionarios de nombres o de entidades nombradas; obviamente este tipo de enfoques son dependientes de la colección de documentos en la que se realiza el agrupamiento.

La ambigüedad se presenta cuando una palabra tiene más de una traducción posible, por ejemplo, la palabra inglesa *"spring"* tiene diversas traducciones en español con significados muy distintos entre sí: *"muelle", "primavera" o "manantial"*. La elección de la traducción más adecuada supone la resolución del problema de ambigüedad, el cual no es una tarea trivial. Para tratar este problema, algunos trabajos como los de (Chen and Lin, 2000) y (Mathieu et al., 2004) se limitan a eliminar las palabras ambiguas; otros tratan de resolverlo eligiendo como sentido correcto aquel que es más frecuente en la colección de documentos (Montalvo et al., 2007)

Otra posibilidad para traducir las características seleccionadas para representar a un documento es recurrir a un tesauro multilingüe. Los sistemas que utilizan tesauros, tienen como base que cada palabra puede expresarse a través de conceptos. En un tesauro multilingüe los conceptos de distintos idiomas se enlazan a través de un índice único que no cambia entre idiomas, es decir, a través del concepto en un idioma es posible encontrar su concepto equivalente en otro idioma, a través del índice. El tesauro multilingüe es EUROVOC de la Comunidad Europea que abarca 22 idiomas oficiales de países miembros de la Unión Europea.

En los trabajos de (Steinberger et al., 2002), (Pouliquen et al., 2004) y (Ortega and Blanco, 1999) se utilizó el tesauro EUROVOC en la tarea de agrupamiento multilingüe. Al igual que en el caso de los diccionarios, los tesauros tienen el problema de cobertura, es decir, no existe un tesauro que abarque todo el vocabulario necesario para representar una colección de documentos. Otra desventaja en el uso de tesauros es la disponibilidad; la cual se refiere a la

posibilidad de encontrar tesauros multilingües que abarquen la totalidad de idiomas de los cuales se compone una colección de documentos.

La Tabla 1 resume los trabajos antes mencionados. En la primera columna se citan los autores de los sistemas; en la segunda columna se específica la tarea en la cual se ha utilizado el agrupamiento multilingüe (en caso de que no sea la tarea primordial del trabajo), esta columna es importante porque muestra la utilidad del agrupamiento multilingüe en distintas tareas. La tercera columna describe el tipo de características utilizadas en la representación; la cuarta columna muestra la estrategia de agrupamiento utilizada para obtenerlas; la quinta columna se refiere a la estrategia de agrupamiento utilizada y finalmente la última columna describe el número de idiomas considerados en el agrupamiento.

En el caso de la estrategia de agrupamiento, se puede notar que la técnica de un paso es la más utilizada en el agrupamiento con traducción, porque al tener los términos o documentos traducidos, el problema se limita a la aplicación del algoritmo de agrupamiento. En el trabajo de (Pouliquen et al., 2004) el agrupamiento se realizó en dos pasos; en el primer paso se hace un agrupamiento monolingüe por idioma, en el segundo paso los grupos de distintos idiomas son unidos por medio de los descriptores de EUROVOC. En (Chen and Lin, 2000) la unión de grupos de distintos idiomas se hace a través de la traducción de verbos, sustantivos y entidades nombradas. El número de idiomas abarcados por estas técnicas, en la mitad de los casos es dos es, decir, sólo se realiza agrupamiento bilingüe. El mayor número de idiomas es el reportado en el trabajo de (Pouliquen et al., 2004), sin embargo, no hay una evaluación de la tarea de agrupamiento, porque forma parte de un sistema de seguimiento de noticias. El tipo de corpus utilizado en la mayoria de los casos es corpus comparable. El único trabajo que utiliza corpus paralelo es el de (Steinberger et al., 2002).

Tabla 1. Métodos basados en traducción

Trabajo	Tarea	Representación	Estrategia	Técnica	Idiomas
(Leftin, 2003)	Resúmenes bilingües automáticos	Documento completo	Un paso	Traducción automática	Inglés y Ruso
(Chen and Lin, 2000)	Resúmenes multi-documentos bilingües	Verbos, sustantivos, entidades nombradas	Dos pasos	Traducción automática	Chino e Inglés
(Cobo Ortega and Rocha Banco, 1999)	Agrupamiento bilingüe	Descriptores de EUROVOC y nombres propios	Un paso	Tesauro multilingüe	Español e Inglés
(Steinberger, et al., 2002)	Calculo de la similitud semántica entre documentos	Descriptores de EUROVOC	Un paso	Tesauro multilingüe	Español e inglés
(Rauber,et al., 2001)	Organización automática de librerías digitales multilingües	Términos más frecuentes	Un paso	Diccionario bilingüe	Ruso, Inglés, Francés y Alemán
(Mathieu, et al., 2004)	Agrupamiento Multilingüe	Términos más frecuentes	Un paso	Diccionario bilingüe	Inglés, Francés y Español
(Saralegi Urizar and Alegria Loinaz, 2007)	Agrupamiento multilingüe	Nombres comunes, entidades y términos multipalabra	Un paso	Diccionario bilingüe	Euskera, Castellano e Inglés
(Pouliquen et al., 2004)	Seguimiento de noticias multilingües	Descriptores de EUROVOC, nombres propios, fechas y números	Dos pasos	Tesauro multilingüe	Español, Inglés, Alemán, Francés e Italiano

3.3 Métodos independientes de la traducción

Los métodos independientes de la traducción se enfocan en tratar de aprovechar las características propias de las colecciones de documentos para realizar el agrupamiento. Algunos trabajos utilizan las características comunes entre idiomas, por ejemplo, fechas, números, entidades nombradas o cognados (definidos en párrafos posteriores) (García Vega et al., 2002), (Montalvo et al., 2006), (Montalvo et al., 2007) otros en cambio recurren al uso de corpus para generar una

representación de los documentos (Wei et al., 2008), (Yogatama and Tanaka-Ishii, 2009).

Los primeros intentos para abordar el agrupamiento multilingüe sin técnicas de traducción consistieron en considerar como características de representación, las palabras comunes que existen entre dos idiomas. En (García Vega et al., 2002), los documentos son representados mediante fechas y nombres de personas, porque no presentan cambios significativos entre idomas. Este tipo de características, en algunos contextos no proporcionan información relevante, por ejemplo, las fechas no son un elemento discriminante para agrupar una colección de documentos médicos.

Con el objetivo de enriquecer este tipo de representaciones, se adicionaron como elementos de representación las entidades nombradas (EN), las cuales incluyen tanto a los nombres de personas y a las fechas, así como nombres de organizaciones y lugares. Ejemplos de entidades son: *"David Beckham", "Pemex", "Europa", "31 de marzo", "3400"*. Algunas entidades nombradas tienden a sufrir pequeños cambios entre idiomas por ejemplo *"Francia"* en inglés se escribe *"France"* y algunas otras se mantienen, por ejemplo, *"John Lennon"*, se escribe igual en inglés y español. En el trabajo de (Montalvo et al., 2006) se utilizan entidades nombradas para representar documentos en español e inglés. Para capturar los cambios que una entidad sufre cuando va de un idioma a otro, ese trabajo se apoya en el concepto de cognado. Los cognados son palabras que perteneciendo a idiomas diferentes, conllevan el mismo significado y exhiben semejanzas ortográficas y/o fonéticas, gracias a que las palabras provienen de un origen lingüístico común. Un ejemplo de cognados es "*silence"* y "*silencio"* en inglés y español respectivamente. Antes de representar los documentos Montalvo y colaboradores (2006) realizan un proceso de extracción de *"entidades nombradas cognadas"*. El primer paso de este proceso es la extracción y clasificación de las entidades. Después se realiza una comparación ortográfica entre las entidades de distintos idiomas y se determina cuáles de ellas tienen más

similitud ortográfica. Las entidades más similares conforman un conjunto de características independientes de la traducción a través de las cuales se puede representar documentos en distintos idiomas. Los resultados obtenidos en ese trabajo demuestran que el uso de este tipo de características es similar a los resultados obtenidos con técnicas de traducción.

El enfoque basado en entidades nombradas es independiente de recursos de traducción como diccionarios o tesauros, sin embargo, depende de reconocedores de entidades nombradas para cada idioma del corpus. El desempeño del agrupamiento depende de la calidad y disponibilidad de estos recursos.

En el trabajo de Montalvo y colaboradores (2007) se presenta un estudio con distintas representaciones de documentos. En ese trabajo los autores combinan distintos tipos de características como entidades nombradas cognadas y cognados de verbos y sustantivos. Los mejores resultados son obtenidos con entidades nombradas y cognados de sustantivos. Este método es aplicable a idiomas de la misma familia lingüística, es decir, idiomas altamente relacionados y a corpus de dominios muy particulares. En temáticas más generales estos métodos no son útiles porque las entidades no son las mismas de un idioma a otro. Por ejemplo, si una colección de noticias se quiere dividir en temas como política, deportes, o espectáculos, seguramente los nombres de lugares y personas de cada tema cambiarán de un lugar a otro, es decir, no serán compartidas entre distintos idiomas.

Otros trabajos han recurrido al uso de corpus paralelos para realizar agrupamiento. El trabajo de (Wei et al., 2008) utiliza un corpus paralelo a partir del cual se aplica análisis semántico latente (LSA, por sus siglas en inglés) (Deerwester et al., 1990) para formar una representación independiente del idioma. En un corpus paralelo cada documento en un idioma tiene un documento traducido, es decir, es equivalente a tener un corpus monolingüe junto con su traducción a otros idiomas. A través de estos corpus es posible conocer con qué frecuencia un término es traducción de otro, así como el contexto en el que se

desenvuelve. La aplicación de LSA a un corpus paralelo dará como resultado un conjunto de relaciones entre términos de distintos idiomas. Este conjunto de términos es utilizado para representar documentos de un corpus bilingüe, distinto al corpus paralelo. La principal desventaja de este método es la creación u obtención de corpus paralelos, ya que generar este tipo de recursos es una tarea costosa, más aún cuando se requiere una colección de documentos en más de dos idiomas, ya que el corpus deberá estar alineado en todos los idiomas de la colección. Las características obtenidas a partir de los corpus paralelos no serán de utilidad en colecciones de dominios específicos, porque las palabras del corpus paralelo pueden no estar presentes en el corpus que se desea agrupar o en caso contrario palabras importantes en el corpus bilingüe no son obtenidas del corpus paralelo.

Yogatama y Tanaka-Ishii (2009) realizan agrupamiento multilingüe de documentos utilizando una técnica de propagación de similitud. Esta técnica requiere el uso de un corpus comparable con la misma temática del conjunto de documentos que será agrupado. Se utiliza la información contenida en el corpus paralelo para mezclar documentos en distintos idiomas. El corpus paralelo debe tener información sobre los grupos a los cuales pertenecen los documentos. Esta información crea enlaces entre dos documentos de distintos idiomas. A partir de esta información la similitud de los documentos es propagada hacia sus vecinos en distintos idiomas, los cuales pertenecen a la colección de documentos que se desea agrupar. Este método depende de información proporcionada por el usuario para agrupar documentos, pero es aplicable incluso a idiomas que no pertenecen a la misma familia lingüística.

La Tabla 2 muestra un resumen de los trabajos independientes de la traducción. En la primera columna se citan los autores; en la segunda columna se específican las características utilizadas para representar los documentos; la tercera columna muestra la estrategia de agrupamiento utilizada en el agrupamiento; la última columna describe el número de idiomas considerados en el agrupamiento.

Todos los trabajos, a excepción de (Montalvo et al., 2006), utilizan agrupamiento de un paso. Sin embargo, los resultados reportados en el trabajo de Montalvo y sus colaboradores (2007) demuestran que la estrategia de un paso obtiene mejores resultados.

Tabla 2. Métodos independientes de la traducción

Referencia	Representación	Estrategia	Idiomas
(García et al., 2002)	Fechas y nombres propios	Un paso	Español e inglés
(Montalvo et al., 2006)	Entidades nombradas	Un paso y dos pasos	Español e inglés
(Montalvo et al., 2007)	Entidades nombradas, Cognados de verbos y sustantivos	Un paso	Español e inglés
(Wei et al., 2008)	LSA	Un paso	Inglés y chino
(Yogatama and Tanaka-Ishii, 2009)	LSA	Un paso	Español, inglés y francés

El agrupamiento en más de dos idiomas solamente se realizó en el trabajo de Yogatama y Tanaka-Ishii, (2009) en el cual se realiza agrupamiento en tres idiomas, pero es necesario proporcionar información sobre los grupos a los que pertenecen los documentos, volviendo al sistema dependiente de esta información.

Es importante observar que todos los enfoques necesitan recursos lingüísticos externos. Los trabajos de (García et al., 2002), (Montalvo et al., 2006) y (Montalvo et al., 2007) requieren de reconocedores de entidades nombradas y de etiquetadores de partes de la oración. Esta característica vuelve al método dependiente del desempeño de estos recursos. Los trabajos de (Yogatama and Tanaka-Ishii, 2009) y de (Wei et al., 2008) necesitan corpus paralelos o comparables como herramientas de entrenamiento, lo que hace que el método dependa de la existencia de estos corpus.

3.4 Discusión

En las secciones anteriores se hizo una revisión de los métodos existentes para el agrupamiento multilingüe de documentos. En esta tesis se desarrolla un método de agrupamiento multilingüe que pertenece a la categoría de métodos independientes de la traducción. El método desarrollado se encuentra en la categoría de los métodos que a partir de características comunes del idioma representan un conjunto de documentos.

El principal problema de los métodos independientes de la traducción que utilizan características comunes es que éstas no son suficientes para representar colecciones de documentos de temáticas generales. Este problema se genera porque las características de representación, generalmente entidades nombradas, no se comparten en dominios generales. El método que se muestra en esta tesis ataca este problema a través de la extracción de un mayor número de características en comparación con los métodos propuestos anteriormente en el estado del arte.

Los métodos que utilizan características comunes utilizan recursos lingüísticos como etiquetadores de entidades nombradas o reconocedores de partes de la oración. Esta característica hace que los métodos dependan de estos recursos y que sus resultados estén basados en el desempeño de dichas herramientas. El método propuesto en este trabajo desarrolla métodos de que no dependen del uso de ningún recurso lingüístico externo, lo que permite al método poder ser aplicado a un número mayor de idiomas y a no depender del desempeño de dichos recursos.

Otro problema que se presenta en los enfoques independientes de la traducción basados en características comunes, es que no han atacado el problema multilingüe, es decir, agrupamiento en más de dos idiomas. La mayoría de los trabajos suponen que el paso del agrupamiento bilingüe al multilingüe es sencillo y que los métodos bilingües funcionan de la misma manera que cuando el número de idiomas aumenta. Esta suposición no es verdadera puesto que en el caso de

características comunes, éstas tienden a disminuir cuando el número de idiomas aumenta, generando representaciones pobres, es decir, los documentos no están representados correctamente y esto deriva en malos resultados en el agrupamiento. En el caso de los métodos que utilizan corpus como entrenamiento, surge el problema de la creación de estos corpus, ya que al aumentar el número de idiomas la alineación entre un mayor número de frases se vuelve una tarea compleja difícil de resolver. En este trabajo proponemos estrategias para trasladar los métodos bilingües a un escenario multilingüe. Dichas estrategias son capaces de mantener el desempeño obtenido con los métodos bilingües.En los siguientes capítulos se muestra la representación desarrollada tanto para el caso bilingüe como para el caso multilingüe.

Capítulo 4

Representación bilingüe

En este capítulo se presenta la propuesta de representación de documentos en dos idiomas para la tarea de agrupamiento multilingüe. La representación presentada en este capítulo es el punto de partida para representar documentos en más de dos idiomas. La representación de documentos está constituida por dos tipos de características, ambas obtenidas con métodos independientes de la traducción y de cualquier recurso lingüístico externo. El capítulo se organiza de la siguiente forma: en la sección 4.1 se describen la representación propuesta, así como los métodos de extracción de las características que las conforman. En la sección 4.2 se presenta la evaluación en agrupamiento multilingüe. Finalmente en la sección 4.3 se muestra un análisis de los resultados obtenidos.

4.1 Representación bilingüe

La mayoría de las investigaciones en agrupamiento multilingüe sin recursos de traducción se basan en las características comunes que existen entre dos idiomas. La ocurrencia de este tipo de palabras disminuye cuando la temática de los corpus es general o cuando los idiomas de la colección aumentan. Esta disminución tiene como consecuencia representaciones dispersas, es decir, las palabras comunes pueden no estar presentes en algunos documentos o ser tan poco comunes que no son de utilidad para dividir la colección de documentos. Para mejorar la representación es necesario aumentar el conjunto de características, es decir, ir más allá de las palabras comunes entre lenguajes. En esta sección se presenta una nueva representación de documentos que afronta el problema de la representación

dispersa a través del descubrimiento de un mayor número de palabras relacionadas en dos idiomas. A continuación se describe detalladamente la propuesta para afrontar el problema de la carencia de atributos comunes y con esto mejorar la representación de documentos en dos idiomas.

Como se mencionó en el Capítulo 2, en el agrupamiento de documentos los documentos son típicamente representados por el modelo de espacio vectorial (VSM); este modelo representa los documentos a través de una matriz de palabras y documentos. En un escenario bilingüe el modelo de espacio vectorial aplicado a una colección de documentos escritos en dos idiomas, puede observarse en la Figura 8. Para ejemplificar la existencia de las palabras en los documentos se ha elegido un peso binario de las palabras, es decir, un 1 en la figura indica la existencia de una palabra en un documento y un cero el caso contrario. Las regiones de la matriz superior derecha e inferior izquierda están llenas de ceros, porque la gran mayoría de las palabras del lenguaje 1 (w_{k1}^{L1}) no están presentes en los documentos del lenguaje 2 (d_{n2}^{L2}) y viceversa. Sin embargo, existen algunas palabras que están presentes en ambos idiomas ($w_m^{L1/L2}$), este conjunto de palabras además de ser *idénticas*, usualmente también comparten *significado*.

Gran cantidad de las palabras idénticas son nombres de personas o lugares que al trasladarse de un idioma a otro no sufren cambios, por ejemplo, la ciudad *"Madrid"* se escribe igual en español e inglés, o el nombre *"George Bush"*, se refiere a la misma persona en español e inglés. Otras palabras de este conjunto son aquellas que al derivarse del mismo origen etimológico no han sufrido cambios ortográficos entre idiomas; por ejemplo, la palabra *"hotel"* en español e inglés significan lo mismo, o la palabra *"triste"* en español y francés mantienen su significado.

		Palabras Lenguaje 1					Palabras Comunes		Palabras Lenguaje 2				
		w_1^{L1}	w_2^{L1}	⋯	w_{k1-1}^{L1}	w_{k1}^{L1}	$w_1^{L1/L2}$	⋯	w_1^{L2}	w_2^{L2}	⋯	w_{k2-1}^{L2}	w_{k2}^{L2}
	d_1^{L1}	0	1	⋯	0	1	0	⋮	0	0	⋯	0	0
Documentos	d_2^{L1}	1	1	⋯	1	0	1	⋮	0	⋱		0	0
Lenguaje 1	⋮	⋯	⋯	⋯	⋯	⋯	⋯	⋮	⋮		0		⋮
	d_{n-1}^{L1}	0	1	⋯	0	0	1	⋮	0	0		⋱	0
	d_{n1}^{L1}	0	1	⋯	0	0	0	⋮	0	0	⋯	0	0
	d_1^{L2}	0	0	⋯	0	0	0	⋮	0	1	⋯	1	1
Documentos	d_2^{L2}	0	⋱		0	0	1	⋮	1	1	⋯	0	1
Lenguaje 2	⋮	⋮		0		⋮	⋮	⋮	⋯	⋯	⋯	⋯	⋯
	d_{n2-1}^{L2}	0	0		⋱	0	0	⋮	0	1	⋯	0	1
	d_{n2}^{L2}	0	0	⋯	0	0	1	⋮	0	1	⋯	1	0

Figura 8. Matriz de características de dos lenguajes

Las palabras idénticas o comunes que pueden encontrarse en dos idiomas han sido la base de la mayoría de los métodos que representan documentos en dos idiomas sin recurrir al uso de la traducción (García et al., 2002), (Montalvo et al., 2006). Sin embargo, el principal inconveniente de esta técnica de representación recae en el número de palabras comunes que pueden encontrarse en dos idiomas. Este número depende del tipo de colección que se utilice y de los idiomas presentes en ella. Como se ha venido mencionando, este inconveniente genera una representación dispersa en la que ciertos documentos no tienen presentes las palabras comunes o la cantidad de palabras comunes es pequeña.

Para generar una representación que pueda proporcionar mejor información sobre la temática del documento es necesario aumentar el número de características, las cuales deben ser capaces de ofrecer información acerca de la temática del documento. De acuerdo a estas exigencias, para incrementar este conjunto de características, partimos de los siguientes supuestos *i)* dos palabras de distintos idiomas que tienen una alta similitud ortográfica tienen un significado

cercano y *ii)* los contextos, es decir, las palabras que co-ocurren en una oración, de uso de estas palabras tienden a guardar una relación de significado.

Tomando en cuenta las suposiciones anteriores, tanto las palabras similares ortográficamente como las palabras de sus contextos, pueden servir para representar colecciones de documentos en distintos idiomas. En esta tesis se desarrollaron dos métodos para extraer este tipo de palabras. Por un lado, se aprovechan las similitudes ortográficas de palabras de distintos idiomas, pero obteniendo dichas palabras a través de un método totalmente independiente de recursos lingüísticos externos, lo que marca la principal diferencia con los métodos propuestos anteriormente (Montalvo et al., 2007), (Wei, Yang, and Lin, 2008). Por otro lado, se aprovechan los contextos de las palabras similares ortográficamente para extraer nuevas características que aportan más información para la división en grupos de la colección de documentos. El método de extracción de dichas características también es independiente de recursos lingüísticos externos y además es una forma novedosa de capturar relaciones temáticas entre palabras de distintos idiomas. En las siguientes secciones se explican a detalle cada uno de los tipos de características y sus métodos de extracción.

Por simplicidad, a las características obtenidas las llamaremos en el resto del documento *parejas bilingües*. A continuación se describe una definición formal de dicho concepto:

Definición 5. *Una pareja bilingüe es un par* (w_i^{L1}, w_j^{L2})*, donde* w_i^{L1} *es una palabra del idioma* $L1$ *y* w_j^{L2} *es una palabra del idioma* $L2$*, cuyos significados se relacionan temáticamente.*

Para representar documentos en dos idiomas se adaptó el modelo de espacio vectorial a un modelo bilingüe basado en parejas bilingües. La Figura 9 muestra gráficamente como se ve la matriz de características. A diferencia de la matriz

mostrada en la Figura 8 donde a cada palabra se le asignaba un valor de existencia, en esta matriz a cada palabra se le asigna un peso denotado por $p_{z,k}$, que representa la importancia de la pareja bilingüe k en el documento *z*.

		Parejas bilingües				
		(w_1^{L1}, w_1^{L2})	(w_2^{L1}, w_2^{L2})	⋯	$(w_{k-1}^{L1}, w_{k-1}^{L2})$	(w_k^{L1}, w_k^{L2})
Documentos Lenguaje 1	d_1^{L1}	$p_{1,1}^{L1}$	⋯	⋯	⋯	$p_{1,k}^{L1}$
	d_2^{L1}	$p_{2,1}^{L1}$		⋯		$p_{2,k}^{L1}$
	⋮	⋮	⋮	⋯	⋮	⋮
	$d_{n_1-1}^{L1}$	$p_{n_1-1,1}^{L1}$		⋯		$p_{n_1-1,k}^{L1}$
	$d_{n_1}^{L1}$	$p_{n_1,1}^{L1}$	⋯	⋯	⋯	$p_{n_1,k}^{L1}$
Documentos Lenguaje 2	d_1^{L2}	$p_{1,1}^{L2}$	⋯	⋯	⋯	$p_{1,k}^{L2}$
	d_2^{L2}	$p_{2,1}^{L2}$		⋯		$p_{2,k}^{L2}$
	⋮	⋮	⋮	⋯	⋮	⋮
	$d_{n_2-1}^{L2}$	$p_{n_1-1,1}^{L2}$		⋯		$p_{n_1-1,k}^{L2}$
	$d_{n_2}^{L2}$	$p_{n_1,1}^{L2}$	⋯	⋯	⋯	$p_{n_1,k}^{L2}$

Figura 9. Matriz de parejas bilingües de dos lenguajes

Para calcular el peso p_{zk}, se modificó la medida *tf-idf*, definida al final de la sección 2.2, de la siguiente forma:

$$p_{zk} = \frac{\#(w_i^{L_1}, d_z) + \#(w_j^{L_2}, d_z)}{|d_z|} \times \log\left(\frac{N}{df\left(w_i^{L_1}, D^{L_1}\right) + df\left(w_j^{L_2}, D^{L_2}\right)}\right) \qquad (4.1)$$

Donde $\#(w_i^{L_1}, d_z)$ y $\#(w_j^{L_2}, d_z)$ indican el número de ocurrencias de las palabras w_i^{L1} y w_j^{L2} en el documento d_z, $df\left(w_i^{L_1}, D^{L_1}\right)$ indica el número de documentos del lenguaje *L* que contienen a la palabra w_k^{Lx} , $|d_z|$ es la longitud en palabras del documento d_z y N el número de documentos de la colección completa.

Con esta representación de los documentos se calcula la matriz de similitud entre documentos utilizando la similitud coseno.

4.1.1 Parejas bilingües con similitud ortográfica

Como se mencionó en la sección anterior, la representación a través de las palabras idénticas o comunes tiene como inconveniente que el número de éstas es reducido. Sin embargo, existen otras palabras que, aunque no son idénticas, tienen cierta similitud ortográfica, es decir, presentan ligeros cambios en su ortografía cuando se trasladan de un idioma a otro. Este tipo de palabras usualmente también tienen un significado común. A este tipo de palabras se les conoce como *cognados*. Por ejemplo, "*presidente"* y "*president"* son cognados entre español e inglés, o *"triste"* en español y francés son cognados. Como se puede apreciar, las palabras idénticas con el mismo significado son un subconjunto de los cognados. La siguiente es una definición formal del concepto cognado.

Definición 6. *Son cognados una pareja de palabras que comparten su significado por tener el mismo origen lingüístico y, en consecuencia, son similares en sus rasgos ortográficos y/o fonéticos (Alcaraz E. and Martinez M., 2004).*

Los cognados surgen como consecuencia de distintos factores, algunos son derivaciones del mismo origen lingüístico, por ejemplo, las palabras "*enciclopedia"* en español y "*encyclopedia"* en inglés; otros son conceptos adoptados de otras lenguas, por ejemplo, la palabra "*jardín"* en español proviene del francés "*jardin"*. Algunos más son términos que se han traducido erróneamente, pero que con el uso continuo de esta traducción se han aceptado como válidos, por ejemplo, *"range"* en inglés que algunas veces se traduce como *"rango"* en español.

Como es de esperarse no todas las parejas que tienen similitudes ortográficas son cognados, algunas parejas como "*dos-dos"*, en español y francés, aunque son idénticas no tienen el mismo significado, en español *dos* se refiere a un número y en francés denota el término espalda. Este tipo de parejas recibe el nombre de

falsos amigos o *cognados falsos*. El número de cognados falsos entre dos idiomas es mínimo comparado con el número de cognados verdaderos que pueden ser encontrados (Alcaraz and Martínez, 2004). Los cognados falsos tienen menor incidencia en corpus de temáticas específicas, es decir, cuando el vocabulario está conformado por términos propios del dominio, la probabilidad de encontrar falsos cognados disminuye. Por ejemplo, en un dominio médico los términos o palabras que se utilizan para nombrar medicamentos, enfermedades o virus no tienen variaciones importantes de ortografía en distintos idiomas y su significado generalmente se mantiene.

Para extraer los cognados se han desarrollado distintos trabajos (Mulloni et al., 2007), en los cuales es necesario saber la categoría de cada palabra, porque los cognados son palabras de la misma categoría, es decir no pueden existir cognados entre sustantivos y verbos. El objetivo de estos métodos es identificar si una pareja de palabras es un cognado o un falso amigo. Generalmente estos métodos no se encargan de extraer los cognados desde documentos sino que tienen las listas de palabras a las cuales hay que asignarles etiqueta de cognado o falso amigo.

Los cognados están limitados a aquellas parejas de palabras que tienen el mismo significado. Por ejemplo, en el estricto sentido de la definición de cognado, parejas como *"president-presidencial"* no empatan con ésta. Sin embargo, aunque estas parejas no tienen el mismo significado, si mantienen cierta relación semántica porque la palabra "*presidencial"* pertenece a la misma familia léxica que "*presidente"*, que es el cognado verdadero de "*president"*. Generalmente, aquellas palabras que provienen de una misma raíz léxica, es decir, que pertenecen a la misma familia léxica, guardan una relación semántica o de significado. Por ejemplo, "*presidente, presidencial, presidencialismo, presidencia"* provienen de la raíz *presidir*. Dadas estas circunstancias, la pareja *"president-presidencial"* cobra sentido para la tarea de agrupamiento, donde el objetivo es formar grupos de documentos relacionados temáticamente.

De acuerdo a esta observación, el método desarrollado se basa en encontrar parejas de palabras que son similares ortográficamente, aunque no cumplan estrictamente con la definición de cognado. El conjunto de parejas bilingües con similitud ortográfica está formado por cognados y algunas parejas que, sin ser cognados en un sentido estricto sí mantienen una relación semántica, porque muchas de estas palabras están formadas con palabras que provienen de las mismas raíces léxicas. A continuación se presenta una definición de las parejas bilingües con similitud ortográfica.

Definición *7.* *Una pareja bilingüe con similitud ortográfica es un par* $(w_i^{L_1}, w_j^{L_2})$, *donde* $w_i^{L_1}$ *es una palabra del idioma* L_1 *ortográficamente similar a una palabra* $w_j^{L_2}$ *del idioma* L_2. *Dada esta similitud ortográfica se supone cierta cercanía semántica entre ambas palabras.*

A continuación se describe el método de extracción de las relaciones bilingües basadas en similitud ortográfica.

Método de extracción de parejas bilingües

El método de extracción de parejas bilingües se basa en el cálculo de la similitud ortográfica entre todas las parejas de palabras que se forman con el vocabulario de dos idiomas. El método no considera la obtención de raíces de las palabras con alguna herramienta, esto es porque fue desarrollado para no depender de ningún recurso lingüístico externo. Esta independencia de recursos le proporciona la posibilidad de ser aplicado a varios idiomas.

A continuación se describe el método de extracción de las parejas bilingües con similitud ortográfica:

Dada una colección de documentos $D = D^{L_1} \cup D^{L_2}$ formada por documentos en los lenguajes L_1 y L_2, la extracción de parejas bilingües con similitud ortográfica se lleva a cabo de la siguiente forma:

1. Determinar el vocabulario de cada lenguaje, es decir, el conjunto de palabras distintas eliminando las palabras vacías[2]. $V^{L_1} = \{w_1^{L_1}, w_2^{L_1} \dots, w_{k_1}^{L_1}\}$ y $V^{L_2} = \{w_1^{L_2}, w_2^{L_2} \dots, w_{k_2}^{L_2}\}$
2. Evaluar la similitud ortográfica para cada pareja de palabras de los vocabularios, es decir, evaluar $sim_{ort}(w_i^{L_1}, w_j^{L_2})$, $i = 1, \dots, k_1$ y $j = 1, \dots, k_2$.
3. Identificar el conjunto de parejas bilingües candidatas ψ^c, es decir, todas las parejas bilingües con similitud ortográfica mayor a un umbral específico definido manualmente.

$$\psi^c = \{(w_i^{L_1}, w_j^{L_2}) : sim_{ort}(w_i^{L_1}, w_j^{L_2}) \geq \beta\} \qquad (4.2)$$

4. Formar el conjunto de parejas bilingües con similitud ortográfica, seleccionando las parejas bilingües candidatas que cumplen la siguiente condición:

$$\psi = \left\{(w_i^{L_1}, w_j^{L_2}) \in \psi^c \mid \max_m \left(sim_{ort}(w_i^{L_1}, w_m^{L_2})\right) \wedge \max_n \left(sim_{ort}(w_n^{L_1}, w_j^{L_2})\right) \rightarrow w_i^{L_1} = w_n^{L_1} \wedge w_j^{L_2} = w_m^{L_2}\right\} \qquad (4.3)$$

[2] Las palabras vacías son palabras frecuentes que no contienen información semántica. Ejemplos de palabras vacías son preposiciones, conjunciones y artículos.

Esta condición permite elegir una sola correspondencia para cada palabra, es decir, la palabra w_i^{L1} debe corresponder solamente con la palabra w_j^{L2}, la cual debe ser la más similar entre todo el conjunto de parejas que puedan formarse con otras palabras

La similitud ortográfica de dos palabras, calculada en el segundo paso de la extracción puede ser calculada con distintas medidas (Inkpen, 2006). Las medidas más utilizadas son la distancia de Levensthein (Levenshtein, 1965) y la Longest Common Subsequences Ratio (LCSR) (Melamed, 1999). En esta tesis se probó el desempeño de ambas, obteniendo resultados ligeramente superiores la LCSR. Por lo tanto, dicha medida es la utilizada para medir la similitud ortográfica requerida. A continuación se describe la forma en la que se obtiene esta medida.

La LCSR de dos palabras w_i y w_j es el cociente de la longitud de la sub-secuencia de caracteres común más larga (LCS) entre w_i y w_j y la longitud en caracteres de la palabra más larga, es decir:

$$LCSR(w_i, w_j) = \frac{length(LCS(w_i, w_j))}{max(length(w_i), length(w_j))} \qquad (4.4)$$

Esta medida está acotada por los valores 0 y 1, un valor igual a 1 indica que ambas palabras son idénticas y un valor cercano a cero indica que el número de letras coincidentes entre las palabras es cada vez menor. Por ejemplo, para las palabras *"australiano*" y *"australien"*, la LCS es *"a·u·s·t·r·a·l·i·n"*, y la palabra con longitud mayor es *"australiano"*, por tanto la LCSR es 9/11=0.81

4.1.2 Parejas bilingües con similitud distribucional

El método descrito en la sección anterior obtiene palabras similares en ortografía; este tipo de características no son suficientes para representar documentos, porque

aunque van más allá de las palabras idénticas entre dos idiomas, no pueden capturar relaciones entre palabras con distinta ortografía. Este tipo de atributos no son suficientes para representar a los documentos de una forma adecuada, es necesario enriquecer la representación con otros tipos de atributos.

Para obtener estos atributos nos basamos en la *"Hipótesis Distribucional"* (Harris, 1970). Esta hipótesis sostiene que las palabras distribuidas de forma similar se relacionan semánticamente. La similitud distribucional permite deducir la similitud de significado de acuerdo al contexto en el que ocurren dichas palabras, es decir, "las palabras con significados similares ocurren en contextos similares" (Rubenstein and Goodenough, 1995), (Schütze and Pedersen, 1995), (Pantel, 2005). Adicionalmente muchos estudios (Carnine et. al, 1984), (Miller and Charles, 1991), (McDonald and Ramscar, 2001) han demostrado que el contexto juega un rol vital en la búsqueda del significado de las palabras.

Para determinar si el contexto de una palabra es similar al de otra se deben mirar a sus vecinos léxicos, es decir, se debe observar con que palabras co-ocurren frecuentemente. Por ejemplo, si las palabras "*voto"* y "*candidato"* co-ocurren frecuentemente con la palabra *"elecciones"* entonces es posible establecer que "*voto"* y "*candidato"* tienen un cierto grado de similitud distribucional. Entonces, el grado de similitud distribucional será mayor si ambas palabras co-ocurren con un mayor número de palabras en común y además este fenómeno se repite constantemente.

La frecuencia de co-ocurrencia de dos palabras w_1 y w_2 es definida como el número de veces que w_2 (llamada *palabra de contexto*) ocurre en una ventana de n palabras alrededor de w_1, en todas las apariciones de w_1 en la colección de documentos. Dado un conjunto *k* de palabras de contexto, cualquier palabra se puede representar con un vector *k*-dimensional de frecuencias de co-ocurrencia. El conjunto de palabras de contexto en una colección de documentos está formado por todo el vocabulario del conjunto de documentos. La matriz de frecuencia de

co-ocurrencia de un conjunto de n palabras es una matriz cuadrada de dimensión $n \times n$; la Figura 10 muestra esta matriz.

.

		Palabras de contexto				
		w_1	w_2	⋯	w_{n-1}	w_n
	w_1	4	0	⋯	0	0
	w_2	1	3	⋯	2	1
Palabras	⋮	⋯	⋯	⋯	⋯	⋯
	w_{n-1}	0	1	⋯	3	1
	w_n	1	4	⋯	0	0

Figura 10. Matriz de co-ocurrencia entre palabras

El tamaño de la ventana de contexto puede extenderse a una frase o incluso a un documento completo. En el campo monolingüe, la gran mayoría de las investigaciones se inclinan por una ventana de tamaño fijo. Por ejemplo (Schütze, 1992) usa una ventana de 1,000 caracteres, argumentando que la co-ocurrencia se mide mejor cuando existe un mayor número de palabras. En el trabajo de (Gale et al., 1994) se usa una ventana de 50 palabras. En ambas investigaciones se concluye que un mayor número de palabras en el contexto es de utilidad cuando los temas del conjunto de documentos son amplios o muy generales, porque la co-ocurrencia muy cercana únicamente ocurre cuando los documentos son de temas muy específicos. La elección del tamaño del contexto tiene una motivación particular, dependiendo de la colección de documentos o de los fines prácticos para los que se utilice, por lo tanto, este parámetro es seleccionado de forma experimental.

Las palabras con distribución similar pueden obtenerse calculando la similitud de los vectores *k*-dimensionales a través de medidas como la similitud coseno. El valor obtenido por la métrica de similitud establece el grado de similitud distribucional que tendrán dichas parejas.

Para encontrar parejas con similitud distribucional, cuando las palabras que conforman dicha pareja pertenecen a distintos idiomas, es necesario establecer un conjunto de palabras de contexto. Este conjunto debe ser capaz de proporcionar una plataforma que permita calcular la similitud semántica entre palabras de distintos idiomas. El método monolingüe calcula la similitud semántica de dos palabras de acuerdo a las palabras co-ocurrentes en común, estas palabras son las mismas porque el vocabulario con el que se describen dichas palabras es el mismo. En el caso multilingüe esto no sucede ya que la mayoría de las palabras de contexto en un idioma son diferentes a las de otro idioma. Para resolver este problema en esta tesis se propone un método que obtiene palabras de contexto comunes en los dos idiomas de las palabras que conforman la pareja.

El método descrito en la sección anterior descubre parejas bilingües con similitud ortográfica, las cuales muestran indicios de cierta relación de significado. Además, hay que recordar las parejas bilingües conforman un vocabulario común para una colección bilingüe de documentos. Las parejas bilingües son utilizadas como el conjunto de palabras de contexto, y a través de ellas es posible comparar y establecer la similitud distribucional de palabras de distintos idiomas. Para ejemplificar la idea, tomemos como palabras de contexto a la pareja bilingüe *"president-presidente"* en inglés y español respectivamente. La Tabla 3 muestra dos párrafos, uno en español y otro en inglés. En la tabla se puede observar que las palabras *"elections y voters"*, en inglés y las palabras "*candidatos y comicios*" en español co-ocurren con la pareja *"president-presidente*". Si esta co-ocurrencia se repite constantemente, entonces es posible establecer que las palabras *"elections y voters"* y "*candidatos y comicios*" son similares en su distribución y por lo tanto están relacionadas temáticamente.

Tabla 3. Ejemplo de similitud distribucional

Español
El *candidato* Fernando Rojas fue derrotado la semana pasada en los *comicios* para elegir *presidente* de Perú …
Inglés
Elections for President and Vice President are indirect *elections* in which *voters* ….

El número de co-ocurrencia de las palabras es capturado en una matriz, llamada matriz de co-ocurrencia bilingüe, la cual puede observarse en la Figura 11. Los componentes de la matriz representan el número de veces que una palabra a co-ocurrido con las palabras de contexto. A diferencia de la matriz del caso monolingüe (Figura 10), esta matriz tiene como palabras de contexto parejas bilingües. Esta representación de la matriz de co-ocurrencia es una forma novedosa para obtener similitud distribucional entre palabras de distintos idiomas. En nuestro método la ventana de contexto es a documento completo, porque el número de características comunes es reducido y por lo tanto la probabilidad de ser encontradas en una ventana de tamaño menor será muy baja.

		Parejas de contexto				
		(w_1^{L1}, w_1^{L2})	(w_2^{L1}, w_2^{L2})	⋯	$(w_{k-1}^{L1}, w_{k-1}^{L2})$	(w_k^{L1}, w_k^{L2})
Palabras Lenguaje 1	w_1^{L1}	0	3	⋯	0	2
	⋮	2	1	⋯	1	0
	w_{n1}^{L1}	⋯	⋯	⋯	⋯	⋯
Palabras Lenguaje 2	w_1^{L2}	0	4	⋯	3	1
	⋮					
	w_{n2}^{L2}	0	1	⋯	0	4

Figura 11. Matriz de frecuencia de co-ocurrencia bilingüe

Para determinar parejas bilingües con similitud distribucional cada vector de las palabras del idioma L_1 es comparado con todos los vectores de las palabras del

idioma L_2, después se eligen aquellos pares de palabras que sobrepasan un umbral de similitud. La definición formal de una pareja bilingüe con similitud distribucional se presenta a continuación.

Definición 8. *Una pareja bilingüe con similitud distribucional es un par* $(w_i^{L_1}, w_j^{L_2})$*, donde* $w_i^{L_1}$ *es una palabra del idioma* L_1 *con distribución similar a una palabra* $w_j^{L_2}$ *del idioma* L_2*, en el contexto de uso de las parejas bilingües con similitud ortográfica.*

La Tabla 4 muestra ejemplos de las parejas bilingües en español e inglés obtenidas con este método. El conjunto de parejas bilingües con similitud distribucional obtenidas con el método desarrollado es variado, ya que pueden encontrarse desde traducciones directas de las palabras (*secuestrados-kidnapped*), así como otro tipo de relaciones (*candidato-elections*). En el caso monolingüe, el trabajo de (Patwardhan and Pedersen, 2006) refiere que cuando la similitud distribucional de las palabras es alta, generalmente se trata de una relación de sinonimia, y mientras más baja la relación temática es más débil. Trasladando esta información al caso bilingüe se puede decir que cuando la similitud distribucional es muy grande probablemente se trata de la traducción de la palabra y cuando es más pequeña se tienen relaciones temáticas, por ejemplo "*hija-born"* o *"condenado-prison"*.

Tabla 4. Ejemplos de parejas bilingües con similitud distribucional

Español	Inglés
candidato	*elections*
aviones	*measures*
secuestrados	*kidnapped*
Hija	*born*
condenado	*prison*

A continuación se describe el proceso de extracción de las parejas bilingües con similitud distribucional.

Método de extracción de parejas bilingües con similitud distribucional

Dada una colección de documentos $D = D^{L_1} \cup D^{L_2}$ formada por documentos en los lenguajes L_1 y L_2, la extracción de parejas bilingües con similitud distribucional se lleva a cabo de la siguiente forma:

1. Determinar el vocabulario de cada lenguaje, es decir, el conjunto de palabras distintas eliminando las palabras vacías: $V^{L_1} = \{w_1^{L_1}, w_2^{L_1} \dots, w_{k_1}^{L_1}\}$ y $V^{L_2} = \{w_1^{L_2}, w_2^{L_2} \dots, w_{k_2}^{L_2}\}$. Sea $V = V^{L_1} \cup V^{L_2}$.
2. Evaluar la similitud ortográfica para cada pareja de palabras de los vocabularios, es decir, evaluar $sim_{ort}(w_i^{L_1}, w_j^{L_2})$, $i = 1, \dots, k_1$ y $j = 1, \dots, k_2$.
3. Identificar el conjunto de parejas bilingües candidatas ψ^c, es decir, todas las parejas bilingües con similitud ortográfica mayor a un umbral específico definido manualmente.

$$\psi^c = \{(w_i^{L_1}, w_j^{L_2}) : sim_{ort}(w_i^{L_1}, w_j^{L_2}) \geq \beta\} \qquad (4.3)$$

4. Formar el conjunto de parejas bilingües con similitud ortográfica, seleccionando las parejas bilingües candidatas que cumplen la siguiente condición:

$$\psi = \left\{\left(w_i^{L_1}, w_j^{L_2}\right) \in \psi^c \mid \max_m \left(sim_{ort}(w_i^{L_1}, w_m^{L_2})\right) \wedge \max_n \left(sim_{ort}\left(w_n^{L_1}, w_j^{L_2}\right)\right) \rightarrow w_i^{L_1} = w_n^{L_1} \wedge w_j^{L_2} = w_m^{L_2}\right\}$$

5. Seleccionar el conjunto de parejas de contexto γ, es decir, las parejas bilingües con similitud ortográfica que sobrepasan un umbral de similitud ortográfica β definido manualmente.

$$\gamma = \{(w_i^{L_1}, w_j^{L_2}) \in \psi : sim_{ort}(w_i^{L_1}, w_j^{L_2}) \geq \beta\} \quad (4.5)$$

6. Representar cada palabra $w_l^{L_x} \in V$ a través del conjunto de parejas de contexto γ, es decir, $w_l^{L_x} = \left(t_{1,w_l^{L_x}}, \ldots t_{k,w_l^{L_x}}\right)$ donde $t_{l,j}$ indica el número de documentos en los cuales co-ocurre la palabra w_l con la pareja de contexto.

7. Calcular la similitud distribucional entre palabras de distintos idiomas, es decir:

$$sim_{dist}(w_i^{L1}, w_j^{L2}) = \frac{\sum_{r=1}^{k} t_{r,w_i^{L1}} \times t_{r,w_i^{L2}}}{\sqrt{(\sum_{r=1}^{k} t_{r,d1}^2)} \times \sqrt{(\sum_{r=1}^{k} t_{r,w_i^{L1}}^2)}} \quad (4.6)$$

8. Formar el conjunto de parejas bilingües con similitud distribucional α definido manualmente, es decir, seleccionar las parejas las parejas de palabras que cumplen la siguiente condición:

$$\varphi = \{(w_i^{L_1}, w_j^{L_2}) : (w_i^{L_1}, w_j^{L_2}) \notin \gamma \wedge sim_{dist}(w_i^{L_1}, w_j^{L_2}) \geq \alpha\} \quad (4.7)$$

El método descrito no depende del uso de ningún recurso lingüístico externo; esta característica permite que pueda ser aplicado a varios idiomas.

4.2 Experimentos 1

En esta sección se describen los experimentos con la representación propuesta. Los experimentos se realizan con una colección bilingüe de documentos descrita en la

sub-sección siguiente. Se muestran experimentos utilizando parejas bilingües con similitud ortográfica para representar los documentos en dos idiomas; esta representación es llamada, *representación de primer orden*. Después, los documentos son representados utilizando parejas bilingües con similitud ortográfica y parejas bilingües con similitud distribucional; esta representación es llamada *representación de segundo orden.*

4.2.1 Configuración experimental

Corpus

Los métodos expuestos se probaron en un corpus comparable, es decir, noticias en diferentes idiomas que hablan sobre el mismo hecho. El corpus fue proporcionado por la Universidad Nacional de Educación a Distancia. Este corpus será llamado en el resto del documento corpus *UNED*. El corpus UNED consiste de una colección de noticias en español e inglés escritas en el mismo periodo de tiempo proporcionadas por la agencia EFE y compiladas por el proyecto HERMES[3]. La colección contiene 192 documentos, 100 en español y 92 en inglés. Los documentos fueron leídos por tres personas y se agruparon en 35 grupos, de los cuales 33 son bilingües y 2 monolingües. La Tabla 5 muestra la distribución del corpus. Como puede apreciarse en la tabla el promedio de palabras de contenido (incluyendo repeticiones) es de aproximadamente 91 y 54 en español e inglés respectivamente, es decir, los documentos son pequeños, ya que sólo contienen un promedio de aproximadamente 4 oraciones.

[3] http://nlp.uned.es/hermes/index.html

Tabla 5. Distribución del corpus UNED

Lenguaje	Documentos	Vocabulario	Promedio de palabras por doc.	Promedio de oraciones por doc.
Español	100	5645	90.39	4.40
Inglés	92	3595	53.37	3.66

Algoritmos de agrupamiento

Para realizar los experimentos se utilizaron dos algoritmos de agrupamiento: el algoritmo *Estrella* y el algoritmo *k-means*. La implementación utilizada para *Estrella* es denominada versión fuera de línea propuesta por (Aslam, et al. 1999). El segundo algoritmo utilizado es el algoritmo *Direct* perteneciente a la biblioteca CLUTO, este algoritmo es una versión del algoritmo *k-means*. Para detalles sobre los algoritmos existentes en la biblioteca CLUTO refiérase a (Karypis, 2002).

En todos los experimentos se utiliza el mismo esquema de pesado, es decir, *tf-idf* definido en la sección 4.1; el coseno como medida de similitud entre documentos, definido en la sección 2.3; y los mismos algoritmos de agrupamiento. Para evaluar los experimentos se utilizó la *F-measure* definida en la sección 2.5.1 en todos los casos.

Baseline

El interés por construir el *baseline* o método de referencia es comparar los métodos desarrollados con enfoques que utilizan una representación independiente de la traducción en la tarea de agrupamiento multilingüe. La representación utilizada como referencia es la propuesta por Montalvo y colaboradores (2006), que utiliza entidades nombradas como características de representación.

Para generar la representación de los documentos es necesario distinguir las entidades nombradas de cada idioma del corpus y después calcular la similitud entre ellas. Para localizar las entidades nombradas se utilizaron los siguientes etiquetadores: FreeLing[4] para el español y Lingpipe[5] para el inglés. En los experimentos sólo se toman en cuenta entidades de tipo persona, organización y lugar. La cantidad de entidades nombradas obtenidas en español e inglés fueron 850 y 636 respectivamente. Para calcular la similitud entre entidades de distintos idiomas se calculó la medida de similitud ortográfica *"Longest Common Subsequence Ratio"* (LCSR), descrita en la sección 4.1.1.

En la Tabla 6 se muestra el número de entidades nombradas similares obtenidas con el método de Montalvo y colaboradores (2006). El umbral β representa la similitud LCSR entre las entidades, también se muestran los documentos que no fue posible representar con el conjunto de características.

Tabla 6. Número de entidades similares y documentos no representados usando el método de Montalvo

β	Número de Características	Documentos no representados
1.0	153	6
0.9	159	6
0.8	176	4
0.7	201	2
0.6	272	0

Los experimentos se realizaron con distintos umbrales de similitud ortográfica β y distintos valores de σ el cual es un valor proporcionado al algoritmo *Estrella* que indica la similitud entre los grupos generados. En específico se realizaron pruebas con β={1.0, 0.9, 0.8, 0.7, 0.6} y σ={0.1, 0.2, 0.3, 0.4, 0.5, 0.6}. En el caso

[4] http://garraf.epsevg.upc.es/freeling/

[5] http://alias-i.com/lingpipe/

del algoritmo *Direct* el número de grupos a generar fue 35, porque coincide con la solución manual.

La Tabla 7 muestra los mejores resultados y la combinación de valores β y σ para obtenerlos. En ambos casos los mejores resultados se obtienen con umbrales de similitud β en los cuales el conjunto de documentos no se encuentra totalmente representado. Esto demuestra que aunque con umbrales menores de similitud los documentos son representados en su totalidad, esta representación no arroja buenos resultados porque existen muchas parejas incorrectas cuando la similitud ortográfica de las entidades es menor.

Tabla 7. Mejores resultados obtenidos con el método de referencia

Algoritmo	*F-measure*	Combinación
Direct	0.84	β=0.7
Estrella	0.83	β=0.8 σ=0.5

4.2.2 Experimentos de primer orden

En esta sección se muestran los resultados con la representación de primer orden, es decir, los documentos son representados por medio de parejas bilingües con similitud ortográfica, obtenidas con el método desarrollado en esta tesis. Los experimentos se realizaron con distintos umbrales de similitud ortográfica β y distintos valores de σ (un valor proporcionado al algoritmo *Estrella* que indica la similitud entre los grupos generados). En específico se realizaron pruebas con β={1.0, 0.9, 0.8, 0.7, 0.6} y σ={0.1, 0.2, 0.3, 0.4, 0.5, 0.6}. En el caso del algoritmo *Direct* el número de grupos a generar fue 35, porque coincide con la solución manual.

La Tabla 8 muestra el número de parejas bilingües con similitud ortográfica obtenidas por umbral. Se utilizaron los mismos umbrales de similitud del caso de las entidades. Como se puede observar en la tabla, al ir disminuyendo el umbral se

obtiene un mayor número de características pero su cambio ortográfico es más notorio, lo cual impacta en su calidad, pues a menor número de letras coincidentes, mayor probabilidad de que dichas parejas sean incorrectas. Otro punto a resaltar en esta tabla es que todos los documentos pueden ser representados incluso con parejas bilingües que tienen similitud β=1, es decir, palabras idénticas en ambos idiomas. Este fenómeno se obtiene porque al manejar la representación a nivel palabra, existe mayor probabilidad de ocurrencia a nivel palabra que a nivel entidades las cuales pueden estar formadas por muchas palabras.

Tabla 8. Número de características en la representación de primer orden

β	# de Características	Documentos no representados
1.0	333	0
0.9	412	0
0.8	866	0
0.7	1415	0
0.6	1950	0

La Tabla 9 muestra los mejores resultados de *F-measure* obtenidos con la representación de primer orden tanto para *Estrella* como para *Direct*. En la primera columna se muestra el algoritmo utilizado; en la columna dos se muestran los resultados de *F-measure*; y en la tercera columna se muestra la combinación de valores con las que se obtuvo el valor de *F-measure* reportado. Los valores obtenidos son superiores a los generados por el método de referencia. Aunque en el caso del algoritmo *Estrella* la mejoría es menos notoria, hay que destacar que el método propuesto tiene como principal ventaja la independencia de recursos lingüísticos externos, tales como reconocedores de Entidades Nombradas necesarios en el método de referencia.

Tabla 9, Mejores resultados obtenidos con la representación de primer orden y la representación de referencia

Representación	Algoritmo	*F-measure*	Parametros
Referencia	***Direct***	0.84	β=0.7
	Estrella	0.83	β=0.8 σ=0.5
Primer Orden	***Direct***	**0.91**	β=0.8
	Estrella	**0.84**	β=0.8 σ=0.3

4.2.3 Experimentos de segundo orden

Los experimentos de primer orden demuestran que es posible obtener una mejor representación de los documentos cuando se tiene un número mayor de características. Para incrementar el número de características, se desarrolló un método que obtiene parejas bilingües con una distribución similar, dicha distribución permite capturar parejas de palabras en distintos idiomas que presentan alguna relación temática. Las parejas bilingües con similitud distribucional no requieren que las palabras que conforman dicha pareja tengan alguna similitud ortográfica. Adicionar parejas bilingües con similitud distribucional a la representación de parejas con similitud ortográfica dará como resultado una representación más rica, que tendrá como consecuencia mejores resultados en el agrupamiento. La representación formada con parejas bilingües con similitud ortográfica y las parejas bilingües con similitud distribucional, es llamada representación de segundo orden. Los resultados con esta representación se muestran a continuación.

Los experimentos con la representación de segundo orden se realizaron combinando distintos umbrales de similitud ortográfica y similitud distribucional. La Tabla 10 muestra solamente el número de las parejas con similitud

distribucional, es decir, no se muestran las combinaciones de los dos tipos de parejas.

Tabla 10. Número de parejas bilingües con similitud distribucional

α	# de Características
1.0	87
0.9	98
0.8	204
0.7	307
0.6	498

Los mejores resultados obtenidos con la representación de segundo orden se muestran en la Tabla 11. Las tablas reflejan que la representación basada en entidades nombradas y la representación de primer orden son superadas por la representación de segundo orden. Con estos resultados se demuestra que una representación más rica de los documentos impacta de forma directa en los resultados del agrupamiento.

Tabla 11. Mejores resultados obtenidos con la representación de Segundo Orden

Representación	Algoritmo	*F-measure*	Combinación
Referencia	***Direct***	0.84	β=0.7
	Estrella	0.83	β=0.8 σ=0.5
Primer Orden	***Direct***	0.91	β=0.8
	Estrella	0.84	β=0.8 σ=0.3
Segundo Orden	***Direct***	**0.93**	β=0.8 α=0.8
	Estrella	**0.89**	β=0.7 σ=0.3; α=0.9

En la siguiente sección se muestra un análisis, que compara los resultados obtenidos con las representaciones desarrolladas y los experimentos de referencia.

4.2.4 Análisis de los resultados

En esta sección se muestra un análisis de los resultados obtenidos. En el caso del algoritmo *Direct*, las representaciones de primer orden supera en un 8.33% a la representación con entidades nombradas y la representación de segundo orden tiene un incremento de 11.9% con respecto a las entidades nombradas. En el trabajo de (Montalvo et al., 2007) se reporta un resultado de *F-measure* de 0.91 cuando se utilizan cognados de entidades nombradas y cognados de sustantivos utilizando el algoritmo *Direct* y la misma colección de documentos que en este trabajo de tesis. Sin embargo, en dicho trabajo las entidades nombradas cognadas son etiquetadas manualmente, lo que elimina los errores derivados del etiquetado manual de entidades. Otro aspecto a destacar en el trabajo de Montalvo, es que las entidades nombradas similares entre dos idiomas solamente pueden obtenerse entre entidades del mismo tipo, es decir, es necesario que se reconozca el tipo de entidad. En el caso de los cognados de sustantivos, su proceso de obtención requiere que el corpus este etiquetado con la categoría gramatical de las palabras.

Debido a que los resultados son sensibles a los parámetros utilizados, en la Tabla 12 se muestran los resultados promedio de *F-measure* de cada una de las representaciones y su correspondiente desviación estándar. Los valores promedio para la representación de primer y segundo orden superan en 8.86% y 13.93%, respectivamente a los resultados obtenidos con entidades, además la desviación estándar es menor en ambos casos lo que refleja que los valores de F-measure, obtenidos con distintos parámetros, son menos variables que los obtenidos con las entidades nombradas.

Tabla 12. Variabilidad de los resultados

Experimento	F-measure	
	Promedio	**Desviación estándar**
Entidades	0.79	0.051
Primer Orden	0.86	0.045
Segundo Orden	0.90	0.026

Los resultados demuestran que el método desarrollado es una buena alternativa cuando las colecciones de documentos están formadas por lenguajes que pertenecen a la misma familia lingüística. La principal ventaja de este método es que no depende de ningún recurso externo, esto le da la posibilidad de ser aplicado a una gran cantidad de idiomas. Los buenos resultados obtenidos con la representación de segundo orden demuestran que las parejas bilingües con similitud distribucional mejoran la representación. La representación de segundo orden puede ser aplicada a dominios más generales, en los cuales las entidades nombradas no son una buena alternativa. A continuación se muestran experimentos en una colección de dominio más general y en otros idiomas

4.3 Experimentos 2

Los experimentos mostrados en esta sección tienen el objetivo de evaluar el desempeño de las representaciones propuestas en otros idiomas y una colección de dominio más general. La colección de documentos utilizada consta de tres idiomas; y los experimentos se realizaron en forma bilingüe, es decir, realizando experimentos con todas las combinaciones de dos idiomas. Los experimentos se realizaron utilizando la representación de primer orden y la representación de segundo orden.

4.3.1 Configuración experimental

Corpus

El corpus está compuesto de noticias provenientes del Reuters Corpus Vol.1 y Vol.2 formado por aproximadamente 487,000 noticias en 13 idiomas distintos al inglés y 810,000 noticias en inglés, que abarcan el mismo periodo de tiempo. La versión en inglés es ampliamente utilizada en la experimentación de algoritmos de clasificación (Pérez et al., 2008) debido a que cada noticia cuenta con una clasificación en distintas categorías. El corpus cuenta con 4 categorías principales (Corporate/Industrial, Economics, Government/Social y Markets), cada categoría tiene sub-categorías, en total existen 99 sub-categorías. Cada noticia está etiquetada por una o más categorías principales y las sub-categorías a las que pertenece.

Para nuestros experimentos seleccionamos un pequeño conjunto de noticias, en el cual nos limitamos a elegir aquellas noticias que pertenecen a la categoría Government/Social y además sólo pertenecen a una de sus sub-categorías. La decisión de elegir noticias que tienen una categoría principal y una sub-categoría es porque en nuestro problema no son tratados aquellos casos en los que una noticia puede pertenecer a más de un grupo. Por otra parte, la elección de la categoría Government/Social fue debido a que los temas de las noticias son más diversos. Finalmente elegimos aleatoriamente un conjunto de 1386 noticias, 491 en español, 431 en inglés y 464 en francés. Las noticias están distribuidas en 16 grupos multilingües, con aproximadamente 30 noticias por clase y por idioma. El número de noticias difiere entre los idiomas porque en algunas de las subcategorías de la categoría Government/Social del Reuters corpus existen menos de 30 noticias. La Tabla 13 muestra algunas estadísticas del corpus seleccionado, el cual a partir de este momento será nombrado como corpus RCV. Como se puede observar los conjuntos de noticias para los tres idiomas tienen

aproximadamente el mismo número de palabras y oraciones por documento, esto indica que las noticias en los tres idiomas tienen tamaños similares. El vocabulario reportado no incluye palabras vacías. Los documentos en la colección son pequeños, cada documento tiene aproximadamente de 3 a 4 oraciones, y el promedio de palabras por documento es menor a 50 palabras.

Tabla 13. Distribución del Corpus RCV

Lenguaje	Documentos	Vocabulario	Promedio de palabras por doc.	Promedio de oraciones por doc.
Español	491	13,437	49.19	3.87
Inglés	431	11,169	41.06	3.03
Francés	464	13,076	47.34	3.67

Los experimentos reportados en esta sección se realizaron bajo las mismas condiciones que los experimentos expuestos en la sección anterior; es decir, se utilizaron los algoritmos *Estrella* y *Direct*; el esquema de pesado *tf-idf*, la medida de similitud coseno; y el *F-measure* para evaluar los resultados del agrupamiento.

Baseline

El *baseline* sigue el método presentado en la sección 4.4. Este método utiliza entidades nombradas para representar a los documentos y fue propuesto por Montalvo y sus colaboradores (2006). Para localizar las entidades nombradas se utilizaron los siguientes etiquetadores: FreeLing[6] para el español, Lingpipe[7] para el inglés y LIA_NE para el francés. La similitud entre entidades de distintos idiomas se realizó midiendo la LCS, definida al final de la sección 4.1.1. La cantidad de entidades nombradas obtenida para cada idioma de los corpus se muestran en la

[6] http://garraf.epsevg.upc.es/freeling/

[7] http://alias-i.com/lingpipe/

Tabla 14. En esta tabla también se muestra el porcentaje de entidades con respecto al vocabulario en cada idioma.

Tabla 14. Entidades nombradas por idioma en el corpus RCV

Idioma	Vocabulario
Español	2199
Inglés	3117
Francés	3638

En la Tabla 15 se muestra el número de entidades nombradas similares encontradas en las distintas parejas de idiomas. Como se puede observar en la tabla no es posible crear una representación en la cual la totalidad de los documentos cuenten con al menos una entidad nombrada. Otro aspecto a notar en la tabla es que aunque en la pareja de idiomas Español-Inglés el número de entidades similares es mayor para todos los umbrales β, la cantidad de documentos no representados también mayor.

Tabla 15. Número de características y documentos no representados en el corpus RCV

	β	Número de entidades similares	Documentos no representados
Español-Inglés	1.0	327	164
	0.9	339	155
	0.8	445	74
	0.7	665	27
	0.6	1154	6
Francés-Español	1.0	202	108
	0.9	216	102
	0.8	302	40
	0.7	477	20
	0.6	918	2
Inglés-Francés	1.0	221	7
	0.9	237	7
	0.8	331	6
	0.7	532	4
	0.6	1091	3

El umbral σ representa el valor proporcionado al algoritmo *Estrella* que indica la similitud mínima entre los elementos de los grupos generados. El número de

grupos indicados al algoritmo *Direct* fue 16, el mismo número de grupos de la solución.

En las Tablas 16 y 17 se muestran los mejores resultados obtenidos por *Estrella* y *Direct* respectivamente, y la combinación de valores para obtenerlos. Los resultados de *F-measure* obtenidos en las tres parejas de idiomas muestran resultados similares; en la mayoría de los casos la similitud β de las entidades nombradas es menor a 0.8, es decir, cuando el número de documentos no representados es menor.

Tabla 16. Mejores resultados de *F-measure* obtenidos con el algoritmo *Estrella* utilizando Entidades Nombradas

Lenguajes	*F-measure*	Combinación
Español-Inglés	0.25	β=0.7 σ=0.3
Francés-Español	0.20	β=0.8 σ=0.2
Inglés-Francés	0.17	β=0.7 σ=0.5

Tabla 17. Mejores resultados de *F-measure* obtenidos con el algoritmo *Direct* utilizando Entidades Nombradas

Lenguajes	*F-measure*	Combinación
Español-Inglés	0.27	β=0.7
Francés-Español	0.21	β=0.7
Inglés-Francés	0.25	β=0.6

4.3.2 Experimentos con las representaciones propuestas

En esta sección se muestran los resultados obtenidos con las representaciones de primer y segundo orden. Los experimentos se realizaron con distintos umbrales de similitud ortográfica β, distintos valores de σ y en el caso de la representación de

segundo orden distintos valores de α. La Tabla 18 muestra el número de parejas con similitud ortográfica y el número de documentos no representados. En los casos de Francés-Español y Francés-Inglés se logra que la totalidad de los documentos quede completamente representada aún con umbrales altos de similitud ortográfica. En el caso de Español-Inglés se logra representar a la totalidad de los documentos cuando el umbral β es menor a 0.8.

Tabla 18. Número de parejas con similitud ortográfica

	β	# de parejas con similitud ortográfica	Documentos no representados
Español-inglés	1.0	790	18
	0.9	948	10
	0.8	2279	5
	0.7	3863	0
	0.6	5166	0
Francés-español	1.0	1073	0
	0.9	1426	0
	0.8	3600	0
	0.7	5656	0
	0.6	6610	0
Inglés-Francés	1.0	1806	0
	0.9	2039	0
	0.8	3909	0
	0.7	5444	0
	0.6	6146	0

Las Tablas 19 y 20 muestran los mejores resultados de *F-measure* obtenidos con el algoritmo *Estrella* y el algoritmo *Direct* para la representación con entidades, la representación de primer orden y la representación de segundo orden. En todos los casos la representación de segundo orden obtuvo mejores resultados que la representación de referencia (entidades) y que la representación de primer orden. Estos resultados demuestran que una mejor representación de los documentos impacta de manera positiva en los resultados del agrupamiento. Además también se puede notar que la adición de parejas bilingües con similitud distribucional genera mejores resultados, debido a que este tipo de parejas están formadas con

palabras en distintos idiomas que están temáticamente relacionadas gracias a su similitud distribucional.

Tabla 19. Resumen de los mejores resultados con *Estrella* en RCV

Lenguajes	Representación	*F-measure*	*Combinación*
Español-Inglés	Entidades	0.25	β=0.7 σ=0.3
	Primer orden	0.28	β=0.7 σ=0.1
	Segundo orden	**0.30**	β=0.7 σ=0.1;α=0.9
Francés-Español	Entidades	0.20	β=0.8 σ=0.2
	Primer orden	0.27	β=0.7 σ=0.1
	Segundo orden	**0.30**	β=0.9 σ=0.1;α=0.9
Inglés-Francés	Entidades	0.17	β=0.7 σ=0.5
	Primer orden	0.18	β=0.8 σ=0.1
	Segundo orden	**0.29**	β=0.8 σ=0.2;α=0.9

Tabla 20. Resumen de los mejores resultados con *Direct* en RCV

Lenguajes	Experimento	*F-measure*	Combinación
Español-Inglés	Entidades	0.27	β=0.7
	Primer orden	0.32	β=0.8
	Segundo orden	**0.37**	β=0.6 α=0.9
Francés-Español	Entidades	0.21	β=0.7
	Primer orden	0.35	β=0.6
	Segundo orden	**0.36**	β=0.8 α=0.8
Inglés-Francés	Entidades	0.25	β=0.6
	Primer orden	0.30	β=0.6
	Segundo orden	**0.35**	β=0.7 α=0.9

Las Tablas 21 y 22 muestran los valores de F-measure promedio para cada tipo de representación en cada uno de los algoritmos. Los mejores valores promedio para todos los casos se obtuvieron con el algoritmo *Direct*. En el caso de Español-Inglés cuando se utiliza el algoritmo Direct se puede observar que la representación de segundo orden tiene una mejora relativa del 25% con respecto a la representación que utiliza entidades nombradas, pasando de un F-measure de 0.24 a uno de 0.30. En el caso de Francés-Español la representación de segundo orden tiene una mejora relativa del 47% con respecto a la representación con entidades. En el caso de Inglés-Francés la mejora obtenida cuando se utiliza la representación de segundo orden es del 40% con respecto a las entidades nombradas. La desviación estándar en todos los casos es muy baja lo que indica que la variación de resultados de F-measure cuando se utilizan distintos parámetros es minima.

Tabla 21. Variabilidad de los resultados con *Estrella*

Lenguajes	Experimento	*F-measure*	
		Promedio	**Desviación estándar**
Español-Inglés	Entidades	**0.20**	0.02
	Primer orden	0.18	0.05
	Segundo orden	0.19	0.05
Francés-Español	Entidades	0.15	0.02
	Primer orden	0.15	0.06
	Segundo orden	**0.16**	0.06
Inglés-Francés	Entidades	0.13	0.02
	Primer orden	**0.17**	0.05
	Segundo orden	**0.17**	0.05

Tabla 22. Variabilidad de los resultados con *Direct*

Lenguajes	Experimento	*F-measure*	
		Promedio	**Desviación estándar**
Español-Inglés	Entidades	0.24	0.02
	Primer orden	0.27	0.04
	Segundo orden	**0.30**	0.06
Francés-Español	Entidades	0.19	0.01
	Primer orden	0.27	0.05
	Segundo orden	**0.28**	0.06
Inglés-Francés	Entidades	0.22	0.02
	Primer orden	0.29	0.01
	Segundo orden	**0.31**	0.03

Los resultados mostrados en esta sección demuestran que las representaciones propuestas son independientes del idioma ya que mantienen los resultados mostrados en la sección anterior, es decir, las representaciones propuestas mejoran los resultados obtenidos con el método de referencia.

Capítulo 5

Representación multilingüe

El agrupamiento multilingüe de documentos, es decir, cuando se desean agrupar documentos en más de dos idiomas, es un problema poco abordado por los enfoques independientes de la traducción. La mayoría del trabajo realizado se ha llevado a cabo en colecciones bilingües, suponiendo que el paso de una representación bilingüe de los documentos a una representación multilingüe es una tarea sencilla. El principal problema es que cuando el número de idiomas aumenta, la aplicación directa de las representaciones bilingües no es viable porque al aumentar el conjunto de idiomas en una colección, el número de características comunes tiende a disminuir.

Para afrontar esta problemática, desarrollamos dos estrategias de agrupamiento que combinan de distintas maneras las características utilizadas en las representaciones bilingües, es decir, las parejas bilingües. La primera estrategia toma directamente a las parejas bilingües como medio de representación, pero calculando la similitud entre documentos de una forma novedosa. La segunda estrategia representa a los documentos mediante atributos multilingües, los cuales están formados por un conjunto de palabras de los distintos idiomas. La obtención de estos atributos se hace a través de parejas bilingües. En las siguientes secciones se explican a detalle ambas estrategias de agrupamiento.

5.1 Estrategia basada en parejas bilingües

Esta estrategia utiliza directamente la representación con parejas bilingües en una colección de documentos en más de dos idiomas. Esta representación es posible

para el agrupamiento porque, como recordamos, un algoritmo de agrupamiento requiere una matriz de similitud, la cual muestra la semejanza entre dos documentos de acuerdo a las características con las que fue representado. En esta matriz la similitud es calculada por pares de documentos. En el caso multilingüe, sin importar el número de idiomas que existan en la colección, pueden existir dos casos en el cálculo de similitud. En el primer caso la similitud se calcula entre documentos del mismo idioma y en el segundo caso entre documentos de distintos idiomas. El cálculo entre documentos del mismo idioma no presenta problemas porque el vocabulario entre ellos es compartido. Para calcular la similitud entre documentos de distintos idiomas es necesario encontrar características comunes en los idiomas involucrados. En el capítulo anterior se demostró que las parejas bilingües son una buena alternativa como medio de representación de documentos en dos idiomas. Entonces estas parejas pueden ser utilizadas para calcular la similitud de documentos en distintos idiomas, porque contienen información en dos idiomas.

Para ejemplificar la idea descrita anteriormente, partimos de una colección de documentos en tres idiomas. La Figura 12 muestra la matriz de características para documentos en 3 idiomas. Esta matriz se ha dividido en zonas denotadas por letras (a,b,c,…,i) y se han colocado pesos binarios para ejemplificar la idea de forma más sencilla, sin embargo, el pesado de las características se puede hacer con otros esquemas. La comparación de los documentos del idioma 1 y del idioma 2 puede realizarse por medio de las regiones *a* y *b* de la matriz, y comparados con los documentos del idioma 3 por medio de las regiones *d* y *f*. De la misma forma, los documentos del idioma 2 y del idioma 3 pueden ser comparados por medio de las regiones *h* e *i*. Para comparar documentos en el mismo idioma se consideran las regiones donde los documentos tienen presencia, por ejemplo, para el idioma 1 la comparación se hace a través de las regiones *a* y *d*. Las regiones *c*, *e* y *g* no son de

utilidad porque las parejas bilingües sólo con capaces de representar a dos idiomas.

		Parejas L1 y L2			**Parejas L1 y L3**			**Parejas L2 y L3**		
		(w_1^{L1}, w_1^{L2})	⋯	(w_1^{L1}, w_1^{L2})	(w_1^{L1}, w_1^{L3})		(w_1^{L1}, w_1^{L3})	(w_1^{L2}, w_1^{L3})		(w_1^{L2}, w_1^{L3})
Documentos L1	d_1^{L1}	0	⋯	1	0	⋯	1	0	⋯	0
	⋮	⋮		⋮ (a)	⋮		⋮ (d)	⋮		⋮ (g)
	d_1^{L1}	1	⋯	0	1	⋯	0	0	⋯	0
Documentos L2	d_1^{L2}	0	⋯	1	0	⋯	0	0	⋯	1
	⋮	⋮		⋮ (b)	⋮		⋮ (e)	⋮		⋮ (h)
	d_1^{L2}	1	⋯	0	0	⋯	0	1	⋯	0
Documentos L3	d_1^{L3}	0	⋯	0	0	⋯	1	0	⋯	1
	⋮	⋮		⋮ (c)	⋮		⋮ (f)	⋮		⋮ (i)
	d_1^{L3}	0	⋯	0	1	⋯	0	1	⋯	0

Figura 12. Matriz de características para 3 idiomas

Dado $L = \{L1, L2, \dots, Lm\}$ un conjunto de idiomas, es necesario generar todas las combinaciones de dos idiomas que pueden formarse con L idiomas, es decir, $\frac{m!}{2(m-2)!}$ combinaciones de idiomas. La generación de las parejas bilingües se realiza con los métodos descritos en las secciones 4.1.1 y 4.1.2. Con los documentos representados por parejas bilingües, la comparación de documentos se hace tomando en cuenta solamente a las parejas bilingües que tienen los idiomas de los documentos comparados.

5.2 Estrategia basada en atributos multilingües

La estrategia basada en atributos multilingües representa a los documentos con un conjunto de palabras que son comunes en todos los idiomas de la colección. Los atributos multilingües son conjuntos de palabras, una por cada idioma del corpus, que presumiblemente tienen una relación temática. Una definición para el término atributo multilingüe se da a continuación:

Definición 9. *En un conjunto de documentos escritos en m idiomas, un atributo multilingüe* $(w_1^{L1}, w_2^{L2}, \ldots, w_k^{Lm})$ *es un conjunto de palabras, una por cada idioma, que tienen una relación temática.*

Los atributos multilingües se construyen a partir de las parejas bilingües que existen en el conjunto de idiomas. Por ejemplo, en una colección de tres idiomas, el atributo multilingüe *(astronauts-astronautas-astronautes)* existe porque pueden encontrarse las parejas bilingües que se forman con las palabras en los tres idiomas, es decir, *(astronauts-astronautas)*, *(austronautas-astronautes)* y *(astronaut-astronautes*) en inglés-español, en español-francés e inglés-francés respectivamente. La existencia de las parejas bilingües da indicios de que existe una relación de transitividad, la cual permite relacionar las parejas bilingües para formar el atributo multilingüe. Sin embargo, es posible que alguna de las parejas no exista en el conjunto de parejas bilingües. Esta situación se presenta por distintos motivos, por ejemplo, porque la pareja bilingüe no existe en el corpus, es decir, alguna de las palabras no está presente en el vocabulario de su respectivo idioma o simplemente porque dicha relación no existe. Aún sin la existencia de una de las parejas bilingües es posible obtener un atributo multilingüe a través de la unión de dos de los atributos. Los atributos obtenidos a través de esta unión serán menos confiables.

Tomando en cuenta estas consideraciones dos clases de atributos son definidos, los atributos multilingües estrictos y los atributos multilingües relajados. Los atributos estrictos son aquellos que se forman a partir de la existencia de todas las parejas bilingües que se forman con los idiomas del corpus. Los atributos relajados son aquellos que se forman cuando una de las parejas bilingües no puede encontrarse. Proponemos la siguiente definición formal para cada uno de estos atributos se presenta a continuación:

Definición 10. U*n atributo multilingüe estricto* $\left(w_i^{L1}, w_j^{L2}, \ldots, w_k^{Lm}\right)$ *con similitud* μ *existe si y solo si se pueden encontrar* $\binom{m}{2}$ *parejas bilingües tales que* $sim\left(w_x^{Lx}, w_y^{Ly}\right) \geq \mu$, $L_x, L_y \in L$.

Definición 11. *Un atributo multilingüe relajado* $\left(w_i^{L1}, w_j^{L2}, \ldots, w_k^{Lm}\right)$ *con similitud* μ *existe si y solo si se pueden encontrar al menos m-1 parejas bilingües tales que* $sim\left(w_x^{L_x}, w_y^{L_y}\right) \geq \mu$, $L_x, L_y \in L$.

La similitud μ de los atributos garantiza que solamente sean unidas parejas con la misma similitud ortográfica, lo que evita que se unan parejas bilingües con mínima relación de significado. En el caso de los atributos bilingües estrictos, es necesario que se cumpla transitividad entre las parejas bilingües. Este hecho hace que los atributos generados sean más confiables, porque existen más restricciones para su creación.

En la Tabla 22 se pueden observar ejemplos de atributos multilingües en tres idiomas, inglés, español y francés, obtenidos automáticamente con el método descrito. Los atributos descubiertos pueden incluir atributos incorrectos. Por ejemplo, *"tube-nube-aube",* en el cual las palabras tienen alta similitud pero los significados no tienen nada común *(tubo,nube,alba)*. Otros atributos tienen dos palabras correctas pero una incorrecta, por ejemplo, *"salvage-salvaje-sauvage"* que significa *rescatar-salvaje-salvaje;* este tipo de relaciones persisten porque la similitud de las parejas bilingües *(salvage-salvaje)* y *(salvaje-sauvage)* con las que se forma el atributo es alta, además la similitud de la pareja bilingüe formada por la transitividad también es alta *(salvage-sauvage).* Otros atributos, aunque incorrectos, son de utilidad "*motorist- motores-motrices"* porque aunque no existe una relación directa como en el caso de otros atributos existe una relación tematica *(motorist y motor).* Otro caso de atributos incorrectos pero que son de utilidad,

son los atributos multilingües formados a partir de parejas bilingües con similitud distribucional, por ejemplo, el atributo multilingüe "*candidato-elections-vote",* aunque incorrecto proporciona información de relaciones distribucionales, es decir, las palabras que conforman el atributo multilingüe tienen similar distribución y por tanto comparten temática.

Tabla 23. Ejemplos de atributos multilingües en tres idiomas

cosmonauts-cosmonautas-cosmonautes
fraud-fraude-fraude
scholars-escolares-scolaires
candidato-elections- vote
salvage-salvaje-sauvage
tube-nube-aube
motorist-motores-motrices

5.3 Experimentos

En esta sección se describen los experimentos con ambas estrategias de agrupamiento. En la primera estrategia los documentos son representados por atributos bilingües y en la segunda con atributos multilingües. Ambas representaciones se evalúan con agrupamiento de un paso y de dos pasos. Recordemos que el agrupamiento de un paso realiza agrupamiento multilingüe desde un principio y el agrupamiento de dos pasos primero hace un agrupamiento monolingüe y después une los grupos monolingües para formar grupos multilingües.

5.3.1 Configuración experimental

Corpus

El corpus utilizado es el descrito en la sección 4.3.1. Para formar una versión multilingüe se combinaron las noticias de los tres idiomas del corpus, es decir, 491 noticias en español, 431 noticias en inglés y 464 noticias en francés. Para una explicación detallada el corpus vea la sección 4.3.1.

Algoritmos de agrupamiento

Los algoritmos de agrupamiento utilizados son *Estrella* y *Direct*, es decir, los mismos algoritmos utilizados en los experimentos bilingües presentados en el capítulo anterior. En el caso de *Estrella* se realizaron experimentos con distintos valores de similitud entre grupos σ, en específico $\sigma=\{0.1,0.2,0.3,0.4,0.5,0.6\}$. En el caso de *Direct*, el número de grupos a generar fue 16, el mismo número de grupos que en la solución manual. El esquema de pesado utilizado es el *tf-idf* , ver sección 4.1 ; la medida utilizada para calcular la similitud entre documentos es la medida coseno, ver sección .3. Para evaluar los resultados del agrupamiento se utilizó la *F-measure* definida en la sección 2.5.1.

Baseline

Los experimentos de referencia se hicieron utilizando un enfoque de traducción, porque en el momento de su realización no se habían reportado trabajos independientes de la traducción que trabajaran con más de dos idiomas. La traducción de los documentos se realizó a documento completo y en cada uno de

los lenguajes existentes, es decir, español, inglés y francés. El traductor utilizado fue el proporcionado por Google[8]. La Tabla 23 muestra el tamaño del vocabulario obtenido para cada idioma; en todos los casos los valores presentados no consideran palabras vacías. Como puede apreciarse, el número de palabras cuando el corpus es traducido aumenta aproximadamente un 100% con respecto al vocabulario de la colección en un solo idioma. Este incremento indica que los vocabularios de cada idioma son muy distintos, o que la traducción es incorrecta, es decir, existen palabras que no son traducidas o bien que son ambiguas y son traducidas con un sentido incorrecto.

Tabla 24. Vocabulario con la traducción en RCV

Idioma Pivote	Vocabulario traducido	Vocabulario por idioma
Español	28,136	13,437
Inglés	20,763	11,169
Francés	27,116	13,076

La Tabla 24 muestra los resultados de *F-measure* obtenidos con *Estrella* y *Direct*. En el caso de *Estrella* los mejores resultados se obtienen, en todos los casos, con σ=0.1. Los valores de F-measure en la traducción no muestran cambios significativos cuando el idioma pivote cambia. Sin embargo, de acuerdo a otros trabajos (Chen and Lin, 2000) (Leftin, 2003) el desempeño del agrupamiento depende en gran medida del traductor utilizado.

Tabla 25. Resultados obtenidos con traducción

Idioma Pivote	*Estrella*	*Direct*
Español	0.24	0.26
Francés	0.25	**0.27**
Inglés	0.25	0.25

[8] www.google.com.mx/language_tools

En la Figura 13 se puede apreciar la variabilidad de los resultados obtenidos con *Estrella* en cada uno de los idiomas, como se ve en la gráfica, los valores de *F-measure* decaen considerablemente cuando el valor σ es incrementado. Esto muestra que la similitud entre los documentos es mínima, aun y cuando los documentos tienen un espacio común de representación. Estos resultados, son atribuidos a la naturaleza del corpus, el cual tiene una estructura difícil de separar en grupos, ver Anexo 1.

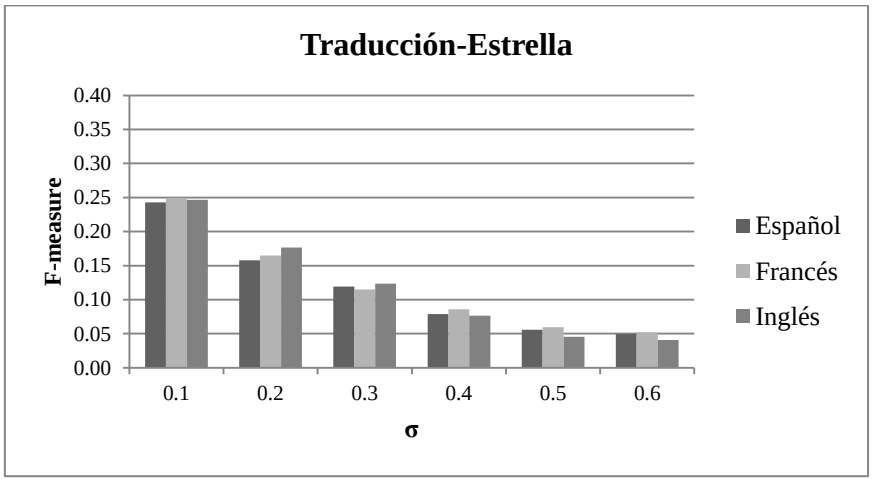


Figura 13. Variación de *F-measure* en el algoritmo *Estrella*

5.3.2 Experimentos con las representaciones propuestas

En esta sección se muestran los experimentos con las representaciones propuestas, es decir, la representación a través de parejas bilingües y la representación a través de atributos multilingües.

Como se mencionó en el capítulo 3, en el agrupamiento multilingüe existen dos tipos de agrupamiento: el de un paso y el de dos pasos. Los experimentos con ambas representaciones se realizaron utilizando los dos tipos de agrupamiento. La estrategia de un paso consiste en abordar el problema multilingüe desde la

representación y la estrategia de dos pasos consiste en generar un agrupamiento monolingüe por idioma y posteriormente unir los grupos monolingües por medio de atributos bilingües. El objetivo de la aplicación de estrategias de dos pasos es enriquecer las representaciones, Cada grupo monolingüe puede verse como un gran documento a partir del cual existe mayor probabilidad de encontrar atributos tanto monolingües como multilingües, pues al aumentar el número de idiomas en los corpus la probabilidad de que un solo documento tenga los atributos de representación disminuye.

Agrupamiento a un paso

La Tabla 17 muestra el número de características obtenidas para cada uno de los tipos de atributos. El número de atributos bilingües es mayor que el número de atributos multilingües, tanto relajados como estrictos. En el caso bilingüe, para todos los umbrales de similitud es posible representar la colección completa de documentos. En el caso de los atributos multilingües existen documentos que no pueden ser representados. Aunque los documentos no representados en el caso relajado y estricto representan un 2.5% y 6.2% de la colección de documentos, repercuten en el agrupamiento de forma negativa, lo cual es demostrado en los resultados presentados en páginas siguientes.

Los resultados del agrupamiento con el algoritmo *Estrella* pueden observarse en la Tabla 26. Como se puede apreciar los resultados obtenidos con *Estrella* no superan al enfoque de traducción. Entre los tres tipos de atributos utilizados puede apreciarse que los atributos multilingües tienen un mejor desempeño. Los atributos estrictos además de superar los resultados con la representación bilingüe tienen la ventaja de generar un espacio de representación más compacto, por ejemplo, para los resultados reportados en la Tabla 26, comparando el número de parejas bilingües (9930, con β=0.8) con los atributos multilingües estrictos (810, con β=0.8) se genera una reducción de 91.9%.

Tabla 26. Número de atributos por umbral

Atributos	Umbral β	# de atributos	No representados
Bilingües	1	3,811	0
	0.9	4,555	0
	0.8	9,930	0
	0.7	15,105	0
	0.6	18,064	0
Multilingüe relajado	1	534	35
	0.9	636	15
	0.8	1,711	1
	0.7	2,547	0
	0.6	2,892	0
Multilingüe estricto	1	361	86
	0.9	427	64
	0.8	810	3
	0.7	952	2
	0.6	982	2

Tabla 27. Mejores resultados de *F-measure* obtenidos con *Estrella* utilizando agrupamiento a un paso

Experimento	*F-measure*	Combinación
Bilingües	0.19	β=0.8; σ=0.1
Multilingüe relajado	0.20	β=0.8; σ=0.2
Multilingüe estricto	0.22	β=0.8; σ=0.2

Los mejores resultados de *F-measure* con el algoritmo *Direct* se muestran en la Tabla 27. En este caso los resultados bilingües superan al enfoque de traducción en un 11.1%. Los resultados con los atributos multilingües no superan a la traducción, sin embargo, están cercanos a los obtenidos con ésta, y tienen la ventaja de que el espacio de representación es menor, además no necesita de ningún recurso lingüístico externo.

Tabla 28. Resultados obtenidos con *Direct* utilizando agrupamiento de un paso

Experimento	*F-measure*	Combinación
Bilingües	0.30	β=0.8
Multilingüe relajado	0.26	β=0.6
Multilingüe estricto	0.24	β=1

Agrupamiento a dos pasos

Los resultados obtenidos en el agrupamiento a un paso no se comportan de la manera esperada, esto se debe a que los atributos utilizados para representar a los documentos no son suficientemente abundantes en el conjunto de documentos utilizado. Según las estadísticas de la Tabla 13 (sección 4.3.1 Corpus), los documentos tienen aproximadamente 50 palabras, es decir son documentos pequeños, en los cuales no se encuentran un gran número de atributos, incluso existen documentos que no tienen ningún atributo. Debido a este problema el agrupamiento de documentos no obtiene resultados superiores porque se ve afectado por el tamaño del documento.

Para resolver este problema, aplicamos una estrategia de agrupamiento de dos pasos. Una estrategia de dos pasos, en el primer paso aplica un agrupamiento monolingüe de documentos; y en el segundo paso genera el agrupamiento multilingüe a partir de la unión de los grupos monolingües obtenidos en el primer paso. El objetivo de realizar agrupamiento monolingüe es considerar a cada grupo monolingüe como un gran documento, en el cual existe mayor probabilidad de encontrar atributos tanto bilingües como multilingües.

A continuación se presentan los resultados utilizando la estrategia de dos pasos. Se realizaron experimentos con los tres tipos de representaciones. En todos los casos, se utilizaron los mismos algoritmos en los dos pasos, sin embargo, se pueden utilizar algoritmos diferentes.

En el caso de *Direct* el agrupamiento del primer paso considera 16 grupos, el mismo número que en la solución manual. En el caso de *Estrella* se realizaron experimentos con σ=0.1, 0.2, 0.3, 0.4, 0.5, 0.6. Los mejores resultados se obtuvieron con los umbrales 0.1 y 0.2, por lo tanto los grupos generados con esta configuración son los utilizados en el segundo paso. Es importante comentar que en el caso monolingüe los mejores resultados alcanzan un *F-measure* de 0.45 en el caso del inglés.

En las Tablas 28 y 29 se muestran los mejores resultados obtenidos en el agrupamiento de dos pasos y la combinación utilizada. En el caso de los resultados de *Estrella*, los mejores valores se obtuvieron con los grupos del primer paso obtenidos con σ=0.1. Los resultados obtenidos tanto por *Direct* como por *Estrella* superan en todos los casos a los resultados obtenidos con un enfoque de traducción.

Una ventaja de la representación con atributos multilingües es que el número de atributos utilizados para representar los documentos en el caso relajado y estricto es de 534 y 361 respectivamente. Estos atributos representan entre un 1% del vocabulario de la colección, lo que demuestra que nuestros atributos son una buena alternativa de representación de documentos en distintos idiomas y que los resultados obtenidos son equivalentes e incluso superan a los resultados obtenidos con métodos de traducción.

Tabla 29. Resultados obtenidos con *Estrella* utilizando agrupamiento de dos pasos

Experimento	*F-measure*	Combinación
Bilingües	0.32	β=0.9; σ=0.1
Multilingüe relajado	0.31	β=1; σ=0.2
Multilingüe estricto	0.31	β=1; σ=0.2

Tabla 30. Resultados obtenidos con *Direct* utilizando agrupamiento de dos pasos

Experimento	*F-measure*	Combinación
Bilingües	0.25	β=0.8
Multilingüe relajado	0.30	β=0.9
Multilingüe estricto	0.30	β=0.7

A diferencia del agrupamiento de un paso en el que existen documentos no representados, en el caso de dos pasos este problema no se presenta, porque al considerar a los grupos como un gran documento, existe mayor probabilidad de

encontrar los atributos tanto bilingües como multilingües. En la siguiente sección se muestra un análisis de los resultados obtenidos en los experimentos.

5.4 Análisis de resultados

Los resultados de los experimentos son variables y vulnerables a los umbrales de los distintos parámetros, por tal razón, en esta sección se muestra un análisis de los resultados promedio obtenidos con nuestras representaciones

Los resultados obtenidos con las distintas representaciones obtienen resultados similares a los métodos de traducción, pero presentan varias ventajas, entre las que destacan la independencia de recursos y el conjunto de representación. La independencia de recursos se refiere a que los métodos desarrollados en este trabajo de investigación no requieren de ningún recurso de traducción ni lingüístico, lo que da al método la posibilidad de ser aplicado a una cantidad mayor de idiomas.

Las Tablas 30 y 31 muestran el *F-measure* promedio para cada tipo de representación en *Estrella* y *Direct*. En las tablas se observa que los experimentos con la estrategia de dos pasos superan a la traducción y a la estrategia a un paso; y que la desviación estándar es menor, lo que indica que hay menor variabilidad de los resultados. En el caso de *Estrella* el mejor resultado obtenido con las representaciones propuestas, es decir, la estrategia de dos pasos con atributos multilingües estrictos tiene una mejora relativa del 125% con respecto a la representación que utiliza traducción (pasando de 0.12 a 0.27). En el algoritmo *Direct* no es posible hacer una comparación promedio, porque solamente se calcula un valor en el caso de la traducción el cual alcanza un *F-measure* máximo de 0.27, el cual tiene una mejora relativa del 7.4% con respecto a la representación multilingüe de dos pasos con atributos multilingües estrictos (pasando de 0.27 a 0.29). La desviación estándar en *Direct* es muy baja lo que indica que los valores de *F-measure* son similares en todos los experimentos.

Tabla 31. Variabilidad de los resultados con Estrella

		F-measure	
	Experimento	**Promedio**	**Desviación estándar**
	Traducción	0.12	0.07
Un paso	Bilingüe	0.13	0.04
	Relajado	0.14	0.03
	Estricto	0.15	0.02
Dos pasos	Bilingüe	0.26	0.02
	Relajado	0.27	0.02
	Estricto	0.27	0.02

Tabla 32. Variabilidad de los resultados con *Direct*

		F-measure	
	Experimento	**Promedio**	**Desviación estándar**
	Traducción	0.27	-
Un paso	Bilingüe	0.25	0.05
	Relajado	0.23	0.02
	Estricto	0.21	0.02
Dos pasos	Bilingüe	0.24	0.005
	Relajado	0.28	0.01
	Estricto	0.29	0.01

Para finalizar se puede concluir que la representación con atributos multilingües combinada con un agrupamiento de dos pasos obtiene los mejores resultados en agrupamiento multilingüe.

Capítulo 6

Conclusiones

En este capítulo se presentan las conclusiones de la tesis. En la sección 6.1 se presenta un sumario de la tesis, en la sección 6.2 las aportaciones generadas y en la sección 6.3 se presentan las direcciones futuras de la investigación. Finalmente, en la sección 6.4 se listan las publicaciones derivadas de este trabajo de tesis.

6.1 Conclusiones

En este trabajo de tesis se desarrolló una representación de documentos en distintos idiomas, independiente de la traducción. La representación desarrollada atacó el problema de carencia de características que se presenta en los enfoques independientes de la traducción. Este problema se abordó desde dos escenarios: el bilingüe y el multilingüe.

En el agrupamiento bilingüe se propuso una representación basada en parejas bilingües. Las parejas bilingües son pares de palabras en dos idiomas cuyos significados se relacionan temáticamente. Dicha relación puede establecerse mediante la similitud ortográfica entre la pareja de palabras o por medio de sus contextos. En el caso multilingüe se desarrollaron dos tipos de estrategias de agrupamiento. La primera estrategia representa a los documentos por medio de parejas bilingües y la segunda describe a los documentos por medio de atributos multilingües construidos a partir de parejas bilingües.

Una característica importante de los métodos desarrollados en esta tesis es que todos son independientes de recursos lingüísticos externos. Esta situación distingue al método como el único enfoque independiente de recursos de traducción, al momento de la publicación de este documento, que cuenta con esta

característica. Además, esta independencia de recursos permite que el método sea aplicado a diferentes idiomas. Una restricción importante acerca del tipo de idiomas que pueden tratarse con el método desarrollado, es que los idiomas deben presentar características morfo-sintácticas muy similares e incluso deben compartir el mismo abecedario. Así el método puede aplicarse a pares de idiomas como español e inglés, español e italiano, etc.; pero no es aplicable a pares de idiomas como chino e inglés o español y árabe.

Las conclusiones obtenidas de este trabajo de tesis se listan a continuación:

- La representación de documentos en dos idiomas con parejas bilingües es una buena opción en el agrupamiento de documentos. Los resultados demostraron que la representación propuesta, es decir, la que combina parejas bilingües con similitud ortográfica y parejas bilingües con similitud distribucional superan a los métodos basados en cognados y a los basados en entidades nombradas propuestos por Montalvo y sus colaboradores (2007). Además la representación propuesta permite obtener resultados equivalentes a los métodos de traducción.
- Un mayor número de atributos en la representación de documentos multilingües impacta de manera favorable en los resultados de agrupamiento. En los experimentos se pudo observar que, cuando los documentos son representados con parejas bilingües con similitud ortográfica y parejas bilingües con similitud distribucional es posible obtener una representación completa del conjunto de documentos, hecho que no sucede con otros tipos de características, como las entidades nombradas.
- La estrategia de dos pasos; es decir, el agrupamiento multilingüe se realiza a partir de un primer paso que genera grupos monolingües por idioma y después en un segundo paso se unen dichos grupos por medio

de un conjunto de atributos multilingües. Esta estrategía obtuvó mejores resultados que la estrategia de un paso y que la traducción porque al considerar a los grupos monolingües como un gran documento, existe mayor probabilidad de que se encuentren tanto parejas bilingües como atributos multilingües, dando pie a una mejor representación de la colección de documentos.

- La representación a través de atributos multilingües combinada con una estrategia de dos pasos, muestra que a través de un número mínimo de atributos se obtienen buenos resultados. Estos resultados demuestran que los atributos propuestos son una buena alternativa para discriminar entre distintas temáticas.

6.2 Aportaciones de la tesis

Las aportaciones que pueden distinguirse en este trabajo de tesis son las siguientes:

- ***Una nueva representación de documentos basada en parejas bilingües orientada al agrupamiento multilingüe***. Esta nueva representación considera nuevas características y nuevos esquemas de pesado, lo que implica que documentos escritos en distintos idiomas son mejor representados. Una buena representación impacta de manera positiva en los resultados obtenidos por el algoritmo de agrupamiento. La representación generada tiene potencial para ser aplicada en otras tareas multilingües como clasificación, generación de resúmenes y recuperación de información multilingüe.
- ***Un nuevo método de extracción de parejas bilingües con similitud distribucional***. Este método genera pares de palabras relacionadas de acuerdo a la similitud semántica. Estas relaciones son de utilidad en distintas tareas, por tanto, generar métodos que obtengan parejas bilingües

de forma automática repercute en el desempeño de sistemas de traducción automática, alineación de corpus y clasificación de textos.

- ***Estrategias de agrupamiento multilingüe***. En esta tesis se abordó el agrupamiento multilingüe mediante dos estrategias de agrupamiento, siendo éstas pioneras en el agrupamiento multilingüe de documentos sin técnicas de traducción. El problema de agrupamiento multilingüe, es decir, cuando la colección de documentos consta de más de dos idiomas a agrupar, es un problema poco abordado por los métodos de agrupamiento que no dependen de recursos de traducción. Esta situación se presenta porque dichos métodos han sido probados en escenarios bilingües. Estos métodos suponen que el paso de las estrategias bilingües a las multilingües es una tarea sencilla, sin embargo, cuando el número de idiomas aumenta esta suposición no se cumple porque las características comunes entre idiomas disminuye cuando aumenta el número de idiomas.

6.3 Trabajo futuro

Como trabajo futuro se plantea la realización de las siguientes tareas:

- **Evaluar el método desarrollado en otros dominios**. La representación de documentos propuesta se basa en parejas bilingües con similitud ortográfica y en parejas bilingües que se obtienen de los contextos de éstas. En dominios más particulares, por ejemplo en el dominio médico, el vocabulario es más particular, por tal motivo, es posible que exista una cantidad considerable de términos que conservan tanto su ortografía como su significado en distintos idiomas. Por ejemplo, los nombres de las enfermedades o de los virus no tiene cambios considerables. Consideramos que la representación de documentos propuesta aplicada a estos dominios obtendrá buenos resultados de agrupamiento.

- **Asignar distintos pesos a las parejas bilingües.** En la representación propuesta, el umbral de similitud, tanto ortográfico como distribucional no es considerado cuando se calcula la importancia de las parejas bilingües. En nuestra representación una pareja bilingüe con alta similitud es considerada con la misma importancia que una pareja con similitud baja. Considerar los pesos de las parejas daría más importancia a aquellas parejas bilingües con mayor similitud, lo que impactaría en los resultados del agrupamiento.

6.4 Publicaciones derivadas de la tesis

Las publicaciones derivadas del trabajo de tesis se nombran a continuación:

- Claudia Denicia-Carral, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda and Rita M. Aceves-Pérez. Bilingual Document Clustering using Translation-Independent Features. International Journal of Computational Linguistics and Applications, Vol. 1, No. 1-2, Enero-Diciembre 2010 (CicLing 2009).

- Claudia Denicia-Carral, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda and David Pinto-Avendaño. Bilingual Document Clustering: Evaluating Cognates as Features. Canadian Journal of Information and Library Science. (Por publicarse en Otoño del 2011).

Bibliografía

Alcaraz, E., and Martínez, M. (2004). *Diccionario de Lingüistica Moderna.* Ariel.

Aslam, J., Pelekhov, K., and Rus, D. (1999). A practical clustering algorithm for static and dynamic information organization. *In Proceedings of the 1999 Symposium on Discrete Algorithms* , 208-217.

Aslam, J., Pelekhov, K., and Rus, D. (2004). The Star Clustering Algorithm for Static and Dynamic Information Organization. *Journal of Graph Algorithms and Applications, Vol. 8, No. 1* , 95-129.

Baeza, R., and Ribeiro, B. (1999). *Modern Information Retrival.* Addison Wesley.

Carnine, D., Kameenui, E., and Coyle, G. (1984). Utilization of contextual information in determining the meaning of unfamiliar words. *Reading Research Quarterly* , 188-204.

Chen, H., and Lin, C. (2000). A Multilingual News Summarizer. *Proceedings of 18 th International Conference on Computational Lingüstics* , 159-165.

Cobo, A., and Rocha, R. (1999). Desarrollo de una aplicación para la gestión, clasificación y agrupamiento de documentos económicos con algoritmos bio-inspirados. *XV Jornadas de ASEPUMA y III Encuentro Internacional* , .

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41 (6)* , 391-407.

Fung, G. (2001). *A Comprehensive Overview of Basic Clustering Algorithms.* University of Winsconsin, Madison, WI, Technical Report.

Gale, W., Church, K., and Yarowsky. (1994). Discrimination decisions for 100,000-dimensional spaces. *Current issues in Computational Linguistics: In honour of Don Walker* , 429–450.

García, M., Martínez, F., Ureña, A., and Martin, M. T. (2002). Generación de un Tesauro de Similitud Multilingüe a partir de un Corpus Comparable Aplicado a CLIR. *Procesamiento del Lenguje Natural , 28*, 55-62.

Gil, R., Badía, J. M., and Pons, A. (2004). Parallel Algorithm for Extended Star Clustering. *Progress in Pattern Recognition, Image Analysis and Applications* , 402-409.

Gliozzo, A., and Strapparava, C. (2001). Cross language Text categorization by acquiring multilingual domains models from comparable corpora. *In Proceedings of the ACL Workshop on Building and Using Parallel Texts* , 123-145.

Gliozzo, A., and Strapparava, C. (2006). Cross language Text categorization by acquiring multilingual domains models from comparable corpora. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* , 553-560.

Guha, S., Rastogi, R., and Kyuseok, S. (1998). Cure: an efficient clustering algorithm for large databases. *In SIGMOD 98: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data* , 73-84.

Guha, S., Rastogi, R., and Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems , 25* (5), 345–366.

Harris, Z. (1970). Distributional structure. *In Papers in structural and transformational Linguistics* , 775–794.

Hsin-Hsi Chen, J.-J. K.-C. (2003). Clustering and Visualization in a Multi-lingual Multi-document Summarization System. *ECIR 2003* , 266-280.

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery* , 283-304.

Inkpen, O. F. (2006). Semi-Supervised Learning of Partial Cognates Using Bilingual Bootstrapping. *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics, COLING-ACL* , 441-448.

Jain, A., Murty, M., and Flynn, P. (1999). Data Clustering: A Review. *ACM Computing Surveys , 31* (3), 264-323.

Karypis, G., Eui-Hong, H., and Kuma, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer, Vol. 32 No. 8* , 68–75.

Kauffman, L., and Rousseuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: John Wiley and Sons.

Leftin, L. J. (2003). Newblaster Russian-English Clustering Performance Analysis. *Columbia computer science technical reports* .

Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR* , 845-848.

MacDonal, S., and Ramscar, S. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. *In Proceedings of the 23rd Annual Conference of the Cognitive Science Society* , 611-616.

Manning, C., and Schütze, H. (2000). *Foundations Of Statistical Natural Language Processing.* Massachusetts, USA: MIT Press Cambridge.

Mathieu, B., Besancon, R., and Fluhr, C. (2004). Multilingual Document Clustering Discovery. *RIAO* , 1-10.

Melamed, D. (1999). Bitext Maps and Alignment via Pattern Recognition. *Computational Linguistics, 25(1)* , 107-130.

Michel, S., George, F., and Isabelle, P. (1993). Using Cognates to Align Sentences in Bilingual Corpora. *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing* , 1071-1082.

Miller, G., and Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes* , *6* (1), 1–28.

Montalvo, S., Martínez, R., Arantza, C., and Fresno, V. (2006). Multilingual News Document Clustering: Two Algorithms Based on Cognate Named Entities. *Text,*

Speech and Dialogue. Lecture Notes in Artificial Intelligence, subseries of Lecture Notes in Computer Science , 165-172.

Montalvo, S., Martínez, R., Casillas, A., and Fresno, V. (2007). Multilingual news clustering: Feature translation vs. identification of cognate named entities. *Pattern Recognition Letters* (28), 2305-2311.

Mulloni, A., Pekar, V., Mitkov, R., and Blagoev, D. (2007). Semantic Evidence for Automatic Identification of Cognates. *Proceedings of the 1st International Workshop on Acquisition and Management of Multilingual Lexicons* , 49-54.

Ng, R., and Hang J. (1994). Very large data bases. *In Proceedings of the 20th International Conference on Very Large Data Bases* , 144–155.

Pantel, P. (2005). Inducing ontological co-occurrence vectors. *In Proceedings of the 43rd Conference of the Association for Computational Linguistics,ACL'05* , 125–132.

Patwardhan, S., and Pedersen, T. (2006). Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. *In: Proceedings of the EACL* , 1-8.

Pouliquen, B., Steinberger, R., Ignat, C., Käsper, E., and Temnikova, I. (2004). Multilingual and Cross-lingual News Topic Tracking. *In: Proceedings of the 20th International Conference on Computational Linguistics , II*, 959-965.

Ralambondrainy, H. (1995). A conceptual version of the k-means algorithm. *Pattern Recognition Letters , 16* (11), 1147–1157.

Rauber, A., Dittenbanch, M., and Merkl, D. (2001). Towars Automatic Content-based Organization of Multilingual Digital Libraries: an English, French and German View of the Russian Information Agency Novosti News. *in Proceedings of the 3rd All-Russian Scientific Conference "Digital Libraries: Advanced Methods And Technologies, Digital Collections"* , 88-95.

Rubenstein, H., and Goodenough, J. (1995). Contextual correlates of synonymy. *Communications of the ACM 8 (10)* , 627–633.

Sahlgren, M. (2008). The distributional hypothesis. *Rivista di Linguistica (Italian Journal of Linguistics) , 20* (1), 33-53.

Salton, G. (1989). *Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer.* Addison-Wesley.

Salton, G., and Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management* , 513-523.

Salton, G., Yang, C., and Wong, A. (1975). A vector space model for automatic indexing. *Communications of the ACM* , 613-620.

Saralegi Urizar, X., and Alegria Loinaz, I. (2007). Similitud entre documentos multilingües de carácter científico-técnico en un entorno Web. *Procesamiento del lenguaje natural* (39), 71-78.

Schütze, H. (1992). Dimensions of meaning. *In Proceedings of the 1992 ACM/IEEE Conference on Supercomputing, Supercomputing'92* , 787-796.

Schütze, H., and Pedersen, J. (1995). Information retrieval based on word senses. *In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval* , 161–175.

Steinberger, R., Pouliquen, B., and Hagman, J. (2002). Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. *Computational Linguistics and Intelligent Text Processing, CICLing* , 415-424.

Tan, P., Steinbach, M., and Kumar, V. (2006). *Cluster Analysis:basic concepts and algorithms in Introduction to Data Mining.* Addison-Wensley.

Theodoridis, S., and Koutroumbas, K. (1999). *Pattern Recognition.* Academic, Press.

Van Rijsbergen, C. (1974). Foundations of evaluation. *Journal of Documentation* , 365-373.

Van Rijsbergen, C. (1979). *Information Retrieval.* London: Butterworths.

Wei, C.-P., Yang, C., and Lin, C.-M. (2008). A Latent Semantic Indexing-based approach to multilingual document clustering. *Decision Support Systems , 45*, 606-620.

Yogatama, D., and Tanaka-Ishii, K. (2009). Multilingual Spectral Clustering using Documente Similarity Propagation. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* , 871-879.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: an efficient data clustering method for very large databases. *In SIGMOD '96: Proceedings of the 1996 ACM SIGMOD international conference on Management of data* , 103-114.

Zhao, Y. (2001). Criterion Functions for Document Clustering. *Technical Report #01-40, University of Minnesota* .

Apéndice 1

Evaluación de los corpus

Los experimentos en los corpus UNED y RCV mostraron diferencias en los resultados de F-measure. En el corpus UNED todos los métodos aplicados obtienen resultados de hasta un 0.93 de F-measure y en el corpus RCV todos los métodos obtienen resultados de hasta el 0.31 de F-measure. La diferencia entre los resultados es derivada de la dureza del corpus ("h*ardness"*). Esta dureza determina que tan sencillo es realizar la separación de los documentos en los grupos propuestos por una solución manual de acuerdo a las características con las que fueron representados.

Para evaluar la dureza de los corpus se utilizaron algunas ideas de los métodos de evaluación de agrupamiento, en específico la evaluación vía correlación (Tan et al., 2006). En particular, dada la matriz de similitud de un corpus así como los grupos a los que cada documento pertenece (de acuerdo a la solución manual), la dureza del corpus se evalúa calculando la correlación entre la matriz de similitud y una matriz de similitud ideal. En una matriz de similitud ideal los documentos que pertenecen al mismo grupo tienen similitud 1, y los documentos de distintos grupos tienen similitud 0. Una alta correlación entre la matriz de similitud y la matriz ideal indica que los documentos bajo el mismo grupo son muy cercanos unos de otros, es decir, el corpus tienen una dureza baja. Una correlación baja entre las matrices indica lo contrario, es decir, el corpus tiene una alta dureza.

La Tabla A1 muestra los valores de correlación obtenidos en los corpus UNED y RCV en los idiomas español-inglés. Las matrices de similitud evaluadas corresponden a las generadas a partir de entidades nombradas y primer orden. Las

matrices de similitud utilizadas en los experimentos corresponden a las matrices de la representación que obtuvo los mejores resultados de F-measure.

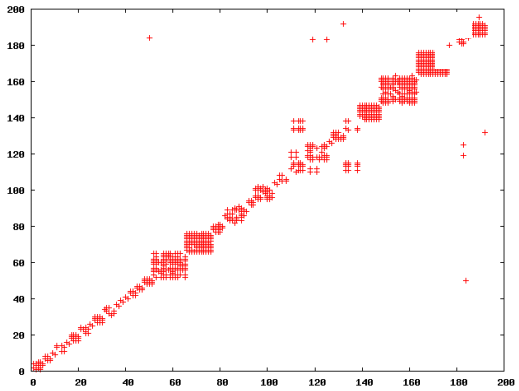
Los resultados de correlación muestran que es más difícil descubrir la estructura del corpus RCV que la del corpus UNED, de acuerdo a las características con las que fueron representados los documentos.

Tabla A1. Correlación en los corpus

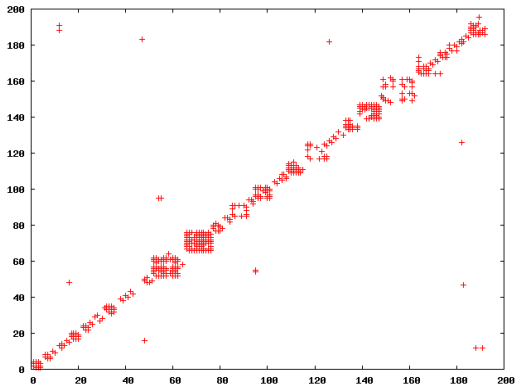
Representación	UNED	RCV
Entidades nombradas	0.868	0.178
Primer orden	0.833	0.237

Una interpretación visual de los resultados mostrados en la Tabla A1 se puede apreciar realizando graficas de las matrices de similitud. La gráfica de un agrupamiento ideal es aquella en la que existen acumulaciones de puntos sobre la línea formada por los puntos $x=y$. En esta situación un documento sólo está relacionado con una clase. En el caso real, esto no sucede ya que los documentos pueden estar relacionados con otros grupos, es decir, existen puntos que están fuera de la región de cúmulos. Las gráficas que se presentan a continuación muestran la distribución de los corpus UNED y RCV de acuerdo a dos tipos de características: entidades nombradas y representación de primer orden.

Las gráficas obtenidas con el corpus UNED se muestran en la Figura A1. Como se puede observar la mayoría de los puntos se encuentran en la línea formada por $x=y$, cuatro casos. Esto demuestra que las características usadas son suficientes para agrupar los documentos de temáticas específicas.



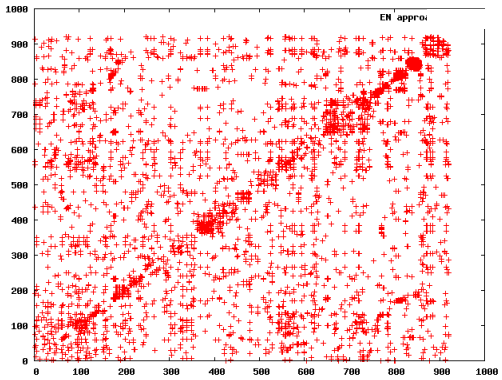
(a)



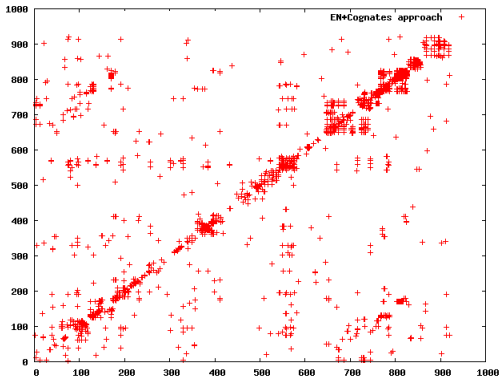
(b)

Figura A1. Análisis del Corpus UNED con (a) Entidades Nombradas y (b) Representación de primer orden

En las gráficas del corpus RCV, mostradas en la Figura A2, se observa que existen una gran cantidad de puntos dispersos, estos puntos representan los documentos que no pueden ser posicionados dentro de algún grupo propuesto en la solución manual.



(a)



(b)

Figura A2. Distribución corpus RCV con (a) Entidades Nombradas, (b) Representación de primer orden