Algoritmo flexibilizado de agrupamiento semántico

Gabriel Castillo
Gerardo Sierra
Instituto de Ingeniería, UNAM

In this paper, we propose a flexible algorithm for semantic clustering. We introduce equal couple, semi-equal couple, null couple, semi-null couple, match couple, binding and semantic couple. The algorithm matches the words of two strings by a set of operations (insertion, deletion and substitution). We determine the minimum cost for each operation required to change one string into another, using the alignment algorithm of Wagner & Fisher. Following these transformations, we align pairs of words for two definitions, to obtain bindings (the strongest pairs of words candidates for semantic couples), and the semantic clusters (semantically related words in a given context). The flexible algorithm for semantic clustering was implemented into a system accessible on-line at: http://iling.torreingenieria.unam.mx.

En este trabajo se presenta el funcionamiento del algoritmo flexibilizado de agrupamiento semántico. En él se introducen los conceptos de par igual, par semi-igual, par nulo, par semi-nulo, par correspondiente, par-vinculado y par-semántico. El algoritmo parte de un conjunto de operaciones sobre dos cadenas (inserción, borrado y sustitución), a partir de las cuales se determinar el mínimo número de cambios necesarios sobre una definición para llegar a otra, empleando el algoritmo de alineamiento propuesto por Wagner y Fisher. A partir de estas transformaciones se establece lo que se ha denominado alineamiento semántico y, con base en él, se identifican los denominados pares-vinculados. El producto final del algoritmo son: los pares-vinculados (fuertes candidatos a ser pares semánticos) y los agrupamientos semánticos (conjuntos de palabras que pueden relacionarse semánticamente). El algoritmo flexibilizado de agrupamiento semántico se implantó en un sistema que puede ser consultado en la página http://iling.torreingenieria.unam.mx.

Palabras clave: *paradigmas semánticos, alineamiento, extracción de información, semántica léxica, lingüística computacional.*
Fecha de recepción del manuscrito: abril del 2004

Gabriel Castillo y Gerardo Sierra.
Grupo de Ingeniería Lingüística, Instituto de Ingeniería
UNAM, Torre de Ingeniería, 2º Piso, Circuito Interior
04510 México D. F.
correos electrónicos: gch@pumas.iingen.unam.mx., gsm@pumas.iingen.unam.mx.

Introducción

Un diccionario onomasiológico electrónico permite que un usuario introduzca un conjunto de palabras (palabras clave) que él considera describen adecuadamente un término cuyo nombre escapa a su memoria o su conocimiento. El diccionario deberá determinar, con base en ese conjunto de palabras, cuál de los términos disponibles es el más adecuado.

Una primera técnica para realizar esta tarea consiste en buscar la existencia de las palabras clave en el texto de la definición del término; sin embargo, los resultados de esta técnica son pobres, pues implica que el usuario introduzca como palabras clave aquellas palabras que efectivamente se encuentran en la definición. Por ejemplo, considérese que en la siguiente definición, obtenida de un diccionario terminológico,

***Caída libre**: descenso de un cuerpo sometido únicamente a la acción de la gravedad* [GDL1996] el usuario introdujo las palabras *descenso* y *cuerpo* como palabras clave, el resultado sería que *caída libre* podría ser uno de los términos buscados por el usuario. Sin embargo si el usuario introdujo *desplome* y *objeto* como palabras clave del término buscado, la técnica indicaría que no se localizó ningún término.

Una técnica alternativa para mejorar los resultados consiste en expandir la búsqueda. Este proceso consiste en asociar a un término varios conjuntos de palabras; cada conjunto agrupa palabras que están relacionadas semánticamente; por ejemplo, los conjuntos {*caída, descenso, bajada, desplome*} y {*cuerpo, objeto, cosa, entidad*} pudieran asociarse al término *Caída libre*. El proceso de búsqueda de un término consiste ahora en buscar cada palabra clave en los conjuntos de palabras asociadas a los términos, de modo que si el usuario introduce *desplome* y *objeto* como palabras clave, un resultado podría ser *Caída libre*, puesto que ambas palabras se encuentra en los conjuntos asociados al término.

El principal problema que enfrenta esta alternativa es la determinación de los elementos de cada conjunto. Cada conjunto es un *agrupamiento semántico* de acuerdo con la siguiente definición.

En el área de recuperación de información, se denomina *agrupamiento semántico* (*cluster* en inglés) a un conjunto de palabras semánticamente relacionadas. De acuerdo con Lounsbury (citado por Geckeler 1976):

"Consideramos como un *agrupamiento semántico* cualquier conjunto de formas lingüísticas en donde: (a) el significado de cada forma tiene una característica en común con el significado de todas las demás formas del conjunto, y (b) el significado de cada forma difiere de todas las demás formas del conjunto por uno o más sentidos del significado de la forma"

Por extensión, definimos un *par-semántico* como una pareja de palabras que guardan una relación semántica en el sentido propuesto por Lounsbury.

El algoritmo aquí propuesto se basa en uno desarrollado por Sierra y McNaught (1999, 2000b) (al cual denominaremos algoritmo básico de agrupamiento semántico o simplemente algoritmo básico), que es un método heurístico y, en esencia, se basa en analogías. Utiliza como entrada un conjunto de términos y sus definiciones (provenientes

de diferentes fuentes), compara estas definiciones e identifica pares de palabras con relaciones semánticas (pares-semánticos), integrándolos después en conjuntos de palabras con una relación semántica en común. El algoritmo permite agrupar palabras cuyo significado o uso pueden considerarse bajo el contexto analizado como sinónimos, aún cuando no guarden una relación sinonímica desde el punto de vista formal

El algoritmo básico fue aplicado a un diccionario de términos en el área de metrología en el idioma inglés. El diccionario contiene 342 términos, cuyas definiciones se obtuvieron de dos diccionarios (el *Collins English Dictionary* (1994) y el *Oxford English Dictionary* (1994) y los resultados obtenidos también se muestran en este artículo.

Algoritmo

El funcionamiento general del algoritmo es el siguiente: con base en un conjunto de términos y sus definiciones (todos los términos dentro de un área del conocimiento), se toman pares de definiciones de un mismo término provenientes de diferentes fuentes (diccionarios, expertos, etc.) y a partir de estos pares se establecen parejas de palabras que pueden sustituirse unas por otras y cuyo cambio en el significado de las definiciones resulta irrelevante. Este tipo de parejas de palabras forman lo que se ha denominado como par-semántico.

Por ejemplo, considérense las definiciones:

> ***A. caída libre****: movimiento de un cuerpo en un campo gravitatorio bajo la influencia de la gravedad* (DES 1996)
> ***B. caída libre****: descenso de un cuerpo sometido únicamente a la acción de la gravedad* (GDL 1996)

El algoritmo identifica que la pareja de palabras *movimiento* y *descenso* guardan una relación sinonímica. Esto significa, básicamente, que al sustituir *movimiento* por *descenso,* en la definición *A,* la variación del significado es mínima y, por tanto, las palabras de la pareja *movimiento y descenso*, bajo el contexto de esta definición, pueden ser sustituidas una por la otra:

> **C. caída libre**: descenso de un cuerpo en un campo gravitatorio bajo la influencia de la gravedad
> **D. caída libre**: descenso de un cuerpo sometido únicamente a la acción de la gravedad

La búsqueda de pares-semánticos se realiza sobre todas las definiciones del diccionario terminológico. Una vez establecidos todos los pares de palabras, se sustituye la primera palabra por la segunda en todos aquellos pares de definiciones en donde aparecen ambos términos en su texto.

Terminada la sustitución, el proceso de búsqueda de pares se repite. El algoritmo termina hasta que ya no se identifican nuevos pares. Al final de cada ciclo, los pares de palabras se combinan para formar conjuntos más grandes de palabras, todas ellas relacionadas semánticamente (agrupamiento semántico).

El algoritmo de agrupamiento semántico es un método inferencial que se basa en examinar las definiciones de un término, identifica las palabras que guardan una relación semántica y a partir de esta relación infiere su aplicación a otros contextos. A continuación se presentan cada una de las etapas del método, presentando los algoritmos sobre los que se basa la etapa examinada.

Funcionamiento

El primer paso del algoritmo consiste en analizar solamente las definiciones de un término agrupadas en pares, donde cada par de definiciones proviene de una fuente distinta. Esto último con el fin de no analizar acepciones diferentes del mismo término.

Para cada pareja de definiciones se busca determinar que operaciones de transformación (inserción de una palabra, borrado de una palabra y sustitución de una palabras por otra) se tiene que aplicar a la primera definición para convertirla en la segunda definición. Una tabla que indica la secuencia de transformaciones que deben aplicarse se denomina *alineamiento*.

Normalmente cada operación tiene un costo asociado y los algoritmos de alineamiento buscan, generalmente, minimizar el costo total del alineamiento, es decir minimizar la suma de los costos asociados a cada una de las operaciones aplicadas.

Con el fin de alinear un par de definiciones de manera que se obtenga el mínimo costo total de las operaciones aplicadas, aquí se emplea un algoritmo denominado *distancia de edición*. Wagner y Fisher [WaF1974] propusieron una técnica para evaluar la distancia de edición y se basa en el método de programación dinámica. El resultado final de la aplicación de este algoritmo es un conjunto secuencial de pares de palabras (incluida la palabra vacía e) que representan el mínimo número de operaciones necesarias para que, a partir de la definición ***A***, se llegue a la definición ***B***. Esta secuencia representa un posible alineamiento de las dos cadenas.

Para propósitos de recuperación de información, las palabras en una definición pueden ser palabras clave (o relevantes) o palabras irrelevantes. Por ejemplo, en la definición de *"caída libre",* las palabras *"movimiento"* , *"cuerpo"*, *"campo"*, *"gravitatorio"*, *"bajo"*, *"influencia"*, *"gravedad"* pueden considerarse palabras clave mientras que *"de", "un", "en", "a", "la"* son ejemplos de palabras poco significativas o irrelevantes.

El término *palabra clave* se utiliza para designar cualquier palabra que pueda ser considerada importante dentro de una definición, desde el punto de vista de las propiedades del concepto descrito. El término *palabra irrelevante*, en oposición a las palabras clave, se utiliza para designar a aquellas palabras que no son significativas para propósitos de recuperación de información, aunque estas palabras son importantes para conectar las palabras clave y hacer, de esta manera, comprensible el concepto.

En el algoritmo básico de alineamiento semántico se emplea una lista de palabras irrelevantes o *stop list* a fin de rechazar aquellas parejas de palabras cuyo significado es poco útil dentro del proceso de agrupamiento semántico. En esencia, esto es equivalente a determinar la categoría gramatical de cada una de las palabras y rechazar aquel par-vinculado que asocia pares de palabras con categoría gramatical diferente, por ejemplo sustantivos con artículos.

De acuerdo con el tipo de operaciones que se pueden efectuar y con las palabras que integran cada pareja, los pares de palabras se clasifican en:

a) *Par-igual*. Aquella pareja de palabras (palabra$_1$, palabra$_2$) cuyos elementos son idénticos, lo cual indica que no se debe efectuar transformación alguna en esa palabra.

b) *Par-correspondiente* Aquella pareja de palabras (palabra$_1$, palabra$_2$) cuyos elementos son diferentes, que indican que una de ellas (palabra$_1$) debe sustituirse por la otra (palabra$_2$) durante el proceso de transformación, siempre y cuando ninguna de las dos palabras formen parte de la lista de palabras irrelevantes.

c) *Par semi-igual.* Aquel par-correspondiente que está formado únicamente por palabras irrelevantes, consideradas dentro de la stop list. Aquí se considera como par igual al par semi-igual para efectos de la evaluación de un coeficiente de similitud denominado LCC y que se presenta en los siguientes párrafos.

d) *Par-nulo*. Aquella pareja formada por una palabra y la palabra vacía e, de forma que una palabra debe agregarse (e, palabra) o borrarse (palabra, e), tal que la palabra no forma parte de la lista de palabras irrelevantes.

e) *Par semi-nulo.* Aquel par-nulo que contiene una palabra irrelevante, perteneciente a la lista de palabras irrelevantes o stop list propuesta por el usuario.

En principio, los pares-nulos carecen de interés semántico, pues indican que debe agregarse o eliminarse una palabra en la definición.

Los pares-correspondientes, por el contrario, indican que debe sustituirse una palabra por otra para llegar a la cadena destino. Por ejemplo los pares correspondientes: (*movimiento, descenso*), (*campo, sometido*), (*gravitatorio, únicamente*), *(bajo, a*), (*influencia, acción*), lo que en esencia es una de las propiedades de un par-semántico, deben ser analizados en su contexto para determinar si existe alguna relación semántica entre ellos. Es decir, se debe establecer si es posible que un miembro del par pueda ser sustituido por el otro sin modificar apreciablemente el significado de la definición.

Como una medida de correlación entre dos palabras de un par-correspondiente, se propuso, en el algoritmo básico de agrupamiento semántico, el uso de un coeficiente de similitud denominado LCC (por sus siglas en inglés de *longest collocation couple*).

El *coeficiente de similitud LCC* examina cada par-correspondiente y las parejas a la antes y después del par-correspondiente, estableciendo cuántos pares-iguales, pares semi-iguales o pares semi-nulos existen antes y después de cada par-correspondiente hasta antes de encontrar un par-nulo u otro par-correspondiente. El número total de pares-iguales, pares semi-iguales y pares semi-nulos más el par-correspondiente es el valor de LCC que se asigna al par analizado.

Así, para las definiciones de *"caída libre"* podemos establecer que una posible manera de alinear estas dos definiciones es la presentada en la Tabla 1.

Tabla 1. Posible alineamiento de las definiciones de término "caída libre"

Cadena Fuente	**Cadena Destino**	**Operación de transformación**	**Tipo de par**	**LCC**	
caída	caída		*igual*		
libre	libre		*igual*		
movimiento	**descenso**	**Sustitución**	***correspondiente***	***8***	
de	de		*igual*		LCC=8
un	un *igual*				
cuerpo	cuerpo *igual*				
en		Borrado	*semi-Nulo*		
un		Borrado	*semi-Nulo*		
campo	sometido	Sustitución	*correspondiiente*	*1*	
gravitatorio	únicamente	Sustitución	*correspondiente*	*1*	
bajo	a	Sustitución	*correspondiente*	*2*	
la	la		*igual*		
influencia	**acción**	**Sustitución**	***correspondiente***	***5***	
de	de		*igual*		LCC=5
la	la		*igual*		
gravedad	gravedad		*igual*		

La primera columna representa la cadena original; la segunda, la cadena objetivo; y la tercera muestra las operaciones de transformación. La columna cuarta indica el tipo de par, mientras que la quinta proporciona el coeficiente de similitud para los pares correspondientes.

Entre más alto sea el valor de LCC mayor es la similitud del par-correspondiente, y es más probable que puedan intercambiarse las palabras del par en cualquiera de las definiciones del término sin que el texto resultante sufra alteraciones en su significado. Por ejemplo, en las definiciones del término *"Caída libre"*, al sustituir *acción* por *influencia* en la primera definición, obtenemos el texto *"caída libre movimiento de un cuerpo en un campo gravitatorio bajo la ~~influencia~~ acción de la gravedad"*.

Experimentalmente se determinó que, para el inglés y el español, un valor de LCC de 5 sugiere un buen grado de similitud. Además se encontró que se requiere al menos un par-igual antes y uno después del par-correspondiente para que las palabras sean susceptibles de considerarse intercambiables: denominamos a esta restricción *condición de frontera*.

Un par-correspondiente que cumple con que el valor de LCC sea mayor o igual a cinco y satisface la condición de frontera se denomina *par-vinculado* (*binding*).

En nuestro ejemplo, sólo los pares-correspondiente (*movimiento, descenso*) e (*influencia, acción)* tienen un LCC igual a 8 y 5, respectivamente, además de que ambos cumplen con la condición de frontera, por lo que estos dos pares se consideran pares-vinculados.

En principio, los pares-vinculados representan pares de palabras que pueden ser utilizadas con el mismo significado dentro de un contexto en particular. Si tomamos la pareja de definiciones para la cual un conjunto de pares-vinculados fueron extraídos, y remplazamos, por ejemplo, en la primera definición a la primera palabra del par-vinculado con la segunda palabra del par-vinculado, observamos que esta definición no ha variado significativamente su sentido.

Al recalcular la distancia de Levenshtein sobre las definiciones modificadas, encontramos dos efectos interesantes:

a) El costo de edición calculado se reduce como consecuencia de que ahora hay más palabras coincidentes. Lo cual indica una mayor similitud entre ambas definiciones.
b) Los pares-correspondientes que no han sido considerados como pares-vinculados pueden aumentar su valor de LCC, por lo que probablemente serán tomados en cuenta si aplicamos nuevamente el algoritmo.

Al hacer lo anterior, el número de pares-vinculados identificados aumenta. Es importante hacer notar que la sustitución de una palabra por la otra no puede aplicarse de manera indiscriminada; de hecho, para realizar la sustitución es necesario que las dos palabras del par-vinculado aparezcan en los textos de las dos definiciones donde se desea realizar la sustitución; en caso contrario, no debe realizase la sustitución del par-vinculado.

La sustitución, sujeta a la restricción anterior, debe ser aplicada para cada par-vinculado en todos los pares de definiciones empleados. Con las nuevas definiciones, se comenzará un nuevo ciclo del algoritmo.

El algoritmo examina todas las parejas de definiciones disponibles y establece todos los pares-vinculados, eliminando aquellos pares-vinculados que contengan palabras irrelevantes.

Una vez establecidos los primeros pares-vinculados, el proceso se repite utilizando ahora las definiciones resultantes del proceso de sustitución de pares-vinculados. El proceso se repite iterativamente hasta que no se generen nuevos pares-vinculados

En cada ciclo, una vez identificados los pares-vinculados, se procede a establecer conjuntos de palabras denominados *"agrupamientos semánticos"*. Los agrupamientos se generan a través de la siguiente *regla de transitividad entre pares-vinculados*:

Sean (*a*,*b*) y (*b*,*c*) dos pares-vinculados formados por las palabras *a*, *b* y *c* ; además, dado que *a* mantiene una relación semántica con *b*, y a su vez *b* mantiene una relación semántica con *c,* entonces se puede afirmar que *a* mantiene una relación semántica con *c*.

Con base en la regla de transitividad de pares-vinculados podemos afirmar que el conjunto {*a*, *b*, *c*} forma un agrupamiento semántico. De esta manera, los agrupamientos semánticos son conjuntos de todas las palabras que satisfacen la relación de transitividad entre ellas con base en los pares-vinculados identificados.

Por ejemplo, considérense los pares vinculados (*ascertaining, measuring*), (*amount, concentration*), (*measuring, estimating*) y (*determining, ascertaining*) dan lugar, bajo la regla de transitividad, a dos agrupamientos semánticos: {*ascertaining, measuring, determining, estimating*} y {*amount, concentration*}.

Evaluación de resultados

El proceso de evaluación de los resultados debe buscar obtener elementos cuantitativos más que cualitativos para establecer las bondades del algoritmo flexibilizado de alineamiento semántico.

Es importante señalar que en la discusión siguiente se ha tenido especial cuidado en utilizar el término de pares-vinculados para referirse al conjunto de todos los pares que han sido identificados por el sistema, mientras que el término de pares-semánticos se emplea para referirse al conjunto de pares-vinculados que efectivamente guardan una relación semántica. El término *pares-semánticos identificados manualmente* se emplea para designar a aquellos pares semánticos que un traductor especializado español-inglés identificó.

El proceso de evaluación consiste en establecer manualmente, con base en el propio conocimiento del idioma, los pares-semánticos contenidos en el diccionario terminológico analizado. Mediante la comparación de los pares-semánticos identificados manualmente con respecto a los resultados ofrecidos por el algoritmo de agrupamiento semántico propuesto, se establece el grado de aceptación del algoritmo; la comparación necesariamente debe arrojar un resultado cuantitativo (no subjetivo) y permitir una correcta evaluación de la propuesta.

La técnica de *Recall* and *Precision* se aplica aquí directamente al problema de evaluación de pares-semánticos. En [BaR1999] puede encontrarse una descripción más general de esta técnica.

El algoritmo básico se aplicó a un corpus de metrología, que recoge las definiciones de 342 términos presentadas en dos diccionarios (el *Collins English Dictionary* [CED1994] y el *Oxford English Dictionary* [OED1994]). Cada acepción se separó en un registro distinto, por lo que a un término puede corresponderle más de una definición

para el mismo diccionario. Se analizaron las definiciones de ese diccionario y con ayuda de un traductor certificado se identificaron dos tipos de pares semánticos:

> *pares-semánticos simples*: son pares-semánticos en los que cada elemento del par está constituido por una sola palabra
> *pares-semánticos compuestos*: es decir, pares-semánticos con más de una palabra en alguno de los elementos que forman el par.

En el análisis se obtuvieron 285 pares-semánticos simples y 78 pares semánticos compuestos, para un total de 363 pares semánticos. Debido a la naturaleza del algoritmo, el análisis presentado se realiza con respecto a los 285 pares-semánticos simples.

Pruebas realizadas

Se evaluaron los resultados al considerar una o más de las siguientes variantes del algoritmo:

1. No considerar la existencia de pares semi-iguales y de pares pares semi-nulos
2. Considerar la existencia de pares semi-iguales
3. Considerar la existencia de pares pares semi-nulos
4. Considerar la existencia de pares semi-iguales y de pares pares semi-nulos

La evaluación de los resultados obtenidos de la aplicación del algoritmo de alineamiento semántico en el corpus de inglés se resume en la tabla 2 y en la gráfica 1. Las pruebas se han ordenado con base en los índices de *recall* y *precision*.

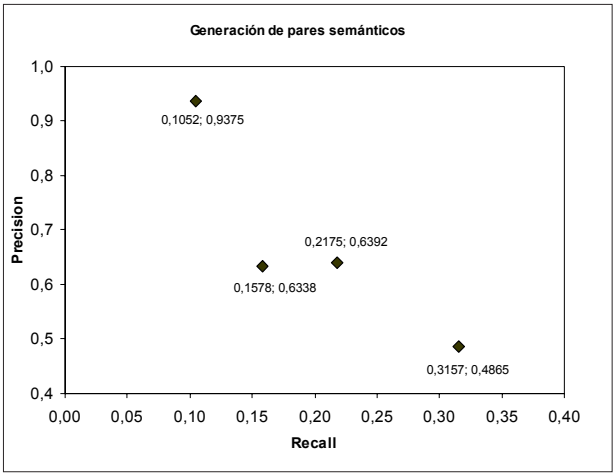
Tabla 2. Resultados de las pruebas realizadas considerando un solo alineamiento

Prueba	| R |	| Ra |	| A |	Recall	Precision	Semi igual	Semi nulo
4	285	90	185	0.3157	0.4865	**SI**	**SI**
2	285	62	97	0.2175	0.6392	**SI**	NO
3	285	45	71	0.1578	0.6338	NO	**SI**
1	285	30	32	0.1052	0.9375	NO	NO

|R| : Número de pares-semánticos obtenidos manualmente |Ra| : Número de pares-vinculados obtenidos
|A| : Número de pares-vinculados generados por la variante evaluada

Se muestra aquí la evaluación de los índices de *recall* y *precision*; las pruebas se han ordenado ascendentemente de acuerdo con el índice de *recall*.

Gráfica 1. Gráfica *Recall* vs *Precision* para las pruebas realizadas



Análisis de los resultados

El algoritmo básico genera 32 pares-vinculados, de los cuales 30 son pares-semánticos, con lo que los índices *recall* y *precision* son 0.1026 y 0.9375, respectivamente. El bajo valor de *recall* indica que se han recuperado muy pocos pares-semánticos (10.52% del universo posible), mientras que el valor alto de *precision* indica que, del total de pares-vinculados, el 93.7% de los pares son pares-semánticos.

Las alternativas de pares semi-iguales o semi-nulos mejoraron el desempeño del algoritmo al incrementar la identificación de pares-semánticos un 106 % y 50%, respectivamente, y en consecuencia los valores de *recall* se incrementaron a 0.2175 y 0.1578, respectivamente. Por otra parte, la generación de pares-vinculados se incrementó un 203% y un 121%, respectivamente, para obtener valores de *precision* de 0.6392 y 0.6338, lo que indica que para obtener más pares-semánticos fue necesario incrementar la identificación de pares-vinculados: de hecho, del total de pares-vinculados sólo el 63% eran pares-semánticos.

De las observaciones anteriores, puede establecerse que el algoritmo mejora notablemente al incluir la variante de par semi-igual o la variante de par semi-nulo; en particular, la primera ofrece mejores resultados que la segunda.

La aplicación simultánea de la alternativa de par semi-igual y par semi-nulo trae como consecuencia un aumento en el índice *recall* (0.3157); este aumento supera el obtenido por la aplicación individual de par-semi-igual (0.2175) o par semi-nulo (0.1578). Sin embargo, este aumento se ve contrarrestado por la disminución en el índice *precision* obtenido en la aplicación conjunta (0.4865) respecto a los valores obtenidos por la aplicación de las alternativas individuales (0.6338 y 0.6392).

Al considerar los porcentajes de incremento obtenidos se observa que la aplicación de la alternativa par semi-nulo implica un aumento del 50% en la identificación de pares- semánticos, contra un aumento del 121% en la identificación de pares-vinculados. En el caso de pares semi-nulos, estos porcentajes corresponden a 106% y 203%, respectivamente, mientras que en el caso de la combinación de las alternativas los incrementos obtenidos son del 200% y 478%.

La aplicación de las alternativas correspondientes a par semi-igual, par semi-nulo y su combinación, mejoran los resultados obtenidos por el algoritmo de alineamiento semántico en valores en el índice *recall* que van del 0.15 al 0.31, contra valores en índice *precision* que van del 0.63 al 0.24. Particularmente, la alternativa que ofrece la mejor relación costo-beneficio corresponde a la consideración de pares semi-iguales.

Limitaciones

El algoritmo de alineamiento semántico es un método de comparación de dos definiciones que se basa en el alineamiento de la secuencia de las palabras que las constituyen. Las palabras son analizadas desde un punto de vista tal que su semántica no es incluida en el análisis. Esta pérdida de información conduce a que eventualmente se agrupan palabras sin ninguna relación semántica.

Las distintas alternativas planteadas en este trabajo permiten relajar las restricciones del algoritmo original, incrementando el índice de *recall* pero disminuyendo en consecuencia el índice de *precision*. La evaluación cuantitativa del algoritmo y sus las alternativas propuestas demostró que mientras no se incorpore información semántica en las definiciones, un incremento del índice *recall* tendrá por consecuencia una disminución del índice *precision*. Mientras se siga visualizando los textos como una secuencia de símbolos sin información adicional, los resultados no mejoraran en cuanto a los índices de evaluación.

Resultados esperados y obtenidos

La evaluación sistemática de los resultados a través del método *recall* y *precision* ayudó a evitar consideraciones cualitativas que eventualmente podrían sesgar los juicios respecto a las bondades de las variantes del algoritmo propuesto.

Entre los resultados obtenidos, se estableció que las alternativas de pares seminulos, pares semi-iguales y su combinación proporcionan la mayor cantidad de pares semánticos, con una proporción muy alta de pares-vinculados.

Trabajos futuros

En este trabajo no se evaluaron los resultados que los algoritmos ofrecen cuando se aplican a un corpus en español. Como parte de los trabajos futuros deberán analizarse las modificaciones y adecuaciones necesarias para el idioma español.

A fin de contrastar los resultados no sólo contra los ideales (los resultados que un algoritmo ideal debería obtener) sino entre las diferentes alternativas, es necesario establecer una medida cuantitativa a través de un análisis costo beneficio. En principio, podría evaluar el costo que se tiene al generar un par semántico en función del costo asociado al número de pares-vinculados identificados. En la literatura del área de recuperación de información no se ha encontrado referencia a un indicador como éste.

Si bien en este trabajo se desarrolló y evaluó el algoritmo flexibilizado con distintas alternativas, éstas todavía son susceptibles de revisarse para tener mejores resultados. Por ejemplo, un análisis semántico de cada una de las definiciones, así como la inclusión de un etiquetador de las partes de la oración posiblemente mejore los resultados obtenidos.

Referencias

BAEZA-YATES R., Ribeiro-Neto B. (1999) *Modern Information Retrieval*. Boston, Mass.: Addison-Wesley:

Collins English dictionary (CED) (1994) Glasgow: Harper Collins Publishers.

Diccionario enciclopédico Salvat Multimedia (DES) (1996). Barcelona: Salvat Editores.

Gran Diccionario de la Lengua Española, edición electrónica (GDL) (1996). Madrid: Editorial Larousse Planeta.

GECKELER, H (1976) *Semántica estructural*. Madrid: Gredos.

Oxford English dictionary (OED) (1994). Oxford: Oxford University Press and Rotterdam: Software B.V.

SIERRA G. (1999) *Design of a concept-oriented tool for terminology*. PhD Thesis, Manchester: University of Manchester, Institute of Science and Technology.

SIERRA G. & McNaught J., (2000), «Extracting semantic clusters from MRD for an onomasiological search dictionary». *International Journal of Lexicography*. Vol. 13 (4): 264-286.

WAGNER R. A., Fisher M. J. (1974) "The string-to-string correction problem". *Journal of the ACM*, Vol. 21(1): 168-173.

Agradecimientos

Agradecemos al CONACyT (R37712) y a la DGAPA-UNAM (IN402900) por su apoyo para el desarrollo de este proyecto