03063
4

VNIVERSIDAD NACIONAL
AVFNOMA DE
MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

UNIDAD ACADÉMICA DE LOS CICLOS PROFESIONALES Y DE POSGRADO DEL COLEGIO DE CIENCIAS Y HUMANIDADES

ALGORITMO REVISADO PARA LA EXTRACCIÓN AUTOMÁTICA DE AGRUPAMIENTOS SEMÁNTICOS

T E S I S

QUE PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

P R E S E N T A

GABRIEL CASTILLO HERNÁNDEZ

DIRECTOR: DR. GERARDO SIERRA MARTÍNEZ

MÉXICO, D. F. **2002**

TESIS CON
FALLA DE ORIGEN

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Autorizo a la Dirección General de Bibliotecas de la UNAM a difundir en formato electrónico e impreso el contenido de mi trabajo recepcional.

NOMBRE: Gabriel Castillo Hernández

FECHA: 6 Oct 2002

FIRMA:

INDICE

Yo que hago dar a luz, ¿no haré nacer? dijo Jehová.
Yo que hago engendrar, ¿impediré el nacimiento? dice tu Dios
Isaías 66:9

Agradecimientos

El trabajo de investigación siempre es producto de las aportaciones de una gran cantidad de personas. Este trabajo no es la excepción, por ello hago un reconocimiento a la participación de aquellas personas que con su esfuerzo diario han permitido esta tesis sea posible.

Paty, que con su apoyo, tiempo y regaños ha logrado impulsar este trabajo hasta su conclusión, ¡siempre juntos! Gracias por tu amor, cariño, tolerancia y por ese par de ángeles disfrazados de pingos: Gabriela Aideé y Pablo David

A la Universidad Nacional Autónoma de México, siempre tan golpeada, siempre tan criticada, pero siempre tan plural, siempre tan tolerante, siempre a la vanguardia de las investigaciones en este país; cuna de investigadores, profesores y egresados de la más alta calidad.

A los compañeros del Instituto de Ingeniería, al grupo de Hidromecánica y al grupo de Ingeniería Lingüística. Gracias por su apoyo.

Muy en especial al Dr. Gerardo Sierra, por su dirección, discusión e ideas vertidas durante el desarrollo de esta tesis; sin mencionar las frecuentes correcciones a mis textos. Gracias.

Sin duda el Dr. Jhon McNaught contribuyó de manera fundamental de esta tesis, pues en su corta estancia en México vertió junto con el Dr. Sierra y un servidor muchas de las ideas presentadas aquí

Los jóvenes becarios que han estado a mi cargo, ciertamente han apoyado y apoyan este trabajo: Martha Cecilia, Miguel, Hilda, Lizeth, Gabriela, Laura y Marlen

Resumen

En 1999 Sierra y McNaught propusieron un algoritmo para la generación automática de agrupamientos semánticos basado en analogías. El algoritmo, denominado en esta tesis algoritmo básico de alineamiento semántico, se aplicó originalmente sobre un diccionario terminológico en el área de metrología en el idioma ingles.

Del análisis de los resultados obtenidos por el algoritmo básico de alineamiento semántico y de un riguroso estudio sobre el algoritmo, en esta tesis se han identificado una serie de opciones que derivan en un conjunto de alternativas para mejorar el número de pares-semánticos reconocidos en el algoritmo. Las observaciones dan pie a 19 líneas de trabajo, divididas en dos grandes grupos: Heurísticas alternativas (16 líneas) y modificaciones a la interfaz hombre-computadora (3 posibilidades) de las cuales en esta tesis se ha trabajado en cuatro líneas: (1) pares semi-iguales y semi-nulos, (2) intercambio de palabras, (3) modificación de costos y (4) rutas múltiples de alineamiento. Como resultado, se desarrollaron 6 algoritmos, los cuales se integraron en lo que denominamos el *algoritmo flexibilizado de alineamiento semántico* y se implantaron en un sistema que puede ser consultado en la página http://iling.iingen.unam.mx

La expectativa respecto al desempeño de las diferentes alternativas desarrolladas, indicaba que el algoritmo de rutas múltiples junto con modificación de costos y pares semi-iguales y semi-nulos ofrecerían los mejores resultados posibles. Con el fin de establecer la certeza de las observaciones cualitativas, se evaluó el desempeño de las diferentes alternativas a través del método *recall* y *precision*. Además, como una medida de comparación entre las diferentes alternativas se propone en esta tesis el *índice de identificación de pares-vinculados* y *el índice de recuperación de pares-semánticos*

Los resultados de la evaluación demostraron que las alternativas de pares semi-nulos, pares semi-iguales y su combinación proporcionan una mayor cantidad de pares-vinculados y de ellos un porcentaje importante corresponde a pares-semánticos. Las alternativas de intercambio de palabras (disyuntivo o no), rutas múltiples y modificación de costos no ofrecieron los resultados esperados, pues aumentan fuertemente el número de pares-vinculados sin incrementar de manera importante el número de pares-semánticos. Sin embargo, estas líneas no deben abandonarse hasta asegurar que no es posible mejorar el desempeño de estas alternativas.

Introducción

El método de programación dinámica es un procedimiento de inferencia que típicamente es utilizado en el diseño de algoritmos para resolver problemas de optimización. La técnica, que fue propuesta en 1962 por Bellman y Dreyfus [BeD1962], se basa en dividir el problema en sub-problemas, donde cada uno es resuelto y eventualmente puede utilizar la solución de sub-problemas resueltos con anterioridad. Desde entonces la técnica ha sido aplicada con éxito a una amplia gama de problemas, entre los cuales podemos mencionar: problemas económico-financieros, problemas de programación de actividades, problemas de comparación de patrones, etc.

En el área de comparación de patrones, este método ha dado origen a las técnicas de alineamiento de cadenas[1]. El problema de alineamiento de cadenas consiste básicamente en determinar el mínimo costo de transformación de una cadena en otra. Este costo es evaluado con base en los cambios que se requieren aplicar para transformar una cadena en otra, utilizando un conjunto de operaciones básicas de transformación: inserción de un símbolo, borrado de un símbolo, sustitución de un símbolo, etc) y donde cada operación tiene asociado un costo. El alineamiento de cadenas es aplicado con éxito en áreas tan diferentes como la biología molecular, la inteligencia artificial (el reconocimiento de voz, traduccion automática, etc), corrección de errores en la transmisión de mensajes a través de un medio, etc.

Específicamente en el área de recuperación de información, las técnicas de alineamiento de cadenas (particularmente el algoritmo propuesto por Wagner y Fisher [WaF1974]) fue empleado como componente de un algoritmo para la identificación y agrupamiento de pares de palabras semánticamente relacionadas. El algoritmo, propuesto por Sierra y McNaught [Sie1999], 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 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* [CED1994] y el *Oxford English Dictionary* [OED1994]).

Los resultados obtenidos indican que el algoritmo propuesto tiene un alto grado de precisión con base en que del número total de pares identificados (32 pares) casi todos (30 pares) son efectivamente pares-semánticos. Sin embargo, el hecho de que sólo se identifican 32 pares de un diccionario con 342 términos, sugiere que, al comparar los resultados con respecto a los pares-semánticos posibles (363 identificables

---

[1] Como se explicará en el capítulo 1, la secuencia de operaciones que se deben aplicar a la cadena inicial para llegar a la cadena final se denomina *alineamiento de cadenas*.

manualmente), la identificación y recuperación obtenida por el algoritmo es pobre. Cabe mencionar que, el trabajo original [Sie1999] no realiza ninguna evaluación del desempeño del algoritmo, por lo que en esta tesis se efectúa una evaluación cuantitativa de los resultados obtenidos tanto por el algoritmo básico de agrupamiento semántico como por las variantes analizadas en este documento.

Siendo básicamente un algoritmo heurístico, el algoritmo de alineamiento semántico puede mejorar la recuperación de los pares-semánticos si se relajan las restricciones sobre las cuales descansa; sin embargo, esa misma relajación puede llevar a una disminución importante en la precisión de los pares identificados. En términos de la aplicabilidad del algoritmo a los problemas de recuperación de información, es necesario buscar un balance entre la recuperación y la precisión. Cualquier variante de las heurísticas subyacentes en el algoritmo debe evaluarse cuantitativamente a fin de tener un criterio para su elección.

El algoritmo de alineamiento semántico ha sido utilizado en un proyecto orientado a la elaboración de un diccionario onomasiológico, y constituye una pieza fundamental de tal proyecto. El proyecto permite la búsqueda de términos a partir de la descripción del concepto en lenguaje natural [SiM2000a]. Como se mencionó arriba, el algoritmo ha mostrado su efectividad, aunque se ha observado que es posible mejorarlo a fin de recuperar mayor número de pares semánticos.

En este marco, la presente tesis busca establecer mejores resultados en términos de un balance entre el número de pares-semánticos correctos y el número de pares-semánticos incorrectamente identificados, buscando siempre un mayor número de pares-semánticos y un menor número de pares identificados incorrectamente.

Objetivo de la tesis

El objetivo general de esta tesis es:

> Mejorar los resultados obtenidos con el algoritmo de alineamiento semántico propuesto por Sierra y McNaught

Las posibilidades de llevar a cabo esto son numerosas, por lo que es necesario delimitar el alcance de este trabajo. En principio, esta tesis se abocará a mejorar el algoritmo, no de proponer uno nuevo, con lo que se hace un aporte significativo –y justificado- al avance del diccionario onomasiológico.

Para ello, se proponen cuatro objetivos específicos:

> Analizar el algoritmo de alineamiento semántico, e identificar aquellas variaciones en la heurística que ofrecen un mejor balance entre el número de pares identificados y la precisión en los pares que efectivamente son pares-semánticos. De entre estas variantes elegir aquellas que, intuitivamente, mejores expectativas ofrezcan y desarrollarlas durante esta tesis.

Establecer una evaluación cuantitativa (no cualitativa) de los resultados ofrecidos por el algoritmo original y por las variantes de las heurísticas subyacentes en el algoritmo para la generación de pares semánticos.

Finalmente, formular un algoritmo con base en las variantes que mejor evaluación obtuvieron.

Resumen de los capítulos

El presente trabajo se encuentra organizado de la siguiente manera:

En el capítulo 1 se introduce el algoritmo de agrupamiento semántico, analizando cada una de sus etapas.

El capítulo 2 establece un conjunto de líneas de trabajo que deben desarrollarse a fin de mejorar los resultados del algoritmo. Los trabajos se dividen en dos tipos:

Heurísticas alternativas. Cuyo objetivo es: (a) identificación de algoritmos existentes o desarrollo de nuevos algoritmos, (b) su incorporación en el algoritmo de alineamiento semántico.

Interfaz hombre-computadora: tiene por fin facilitar la aplicación del algoritmo de alineamiento semántico, permitiendo al usuario la flexibilidad de aplicar diferentes opciones a diferentes diccionarios terminológicos.

El capítulo 3 retoma las Heurísticas alternativas propuestas en el capítulo anterior y aporta al algoritmo básico seis modificaciones con el objetivo de mejorar el desempeño del algoritmo en lo que se refiere a identificación de pares semánticos. Se presentan las consideraciones teórico-prácticas de estas modificaciones.

En el capítulo 4 se presenta la evaluación cuantitativa de la calidad de los pares semánticos generados, tanto por el algoritmo básico como por las modificaciones se presentan. Aunque los resultados son alentadores, indican que es necesario seguir afinando el algoritmo e incorporar nuevos algoritmos que suministren más información sobre la semántica y las partes de la oración, de modo que se afine la identificación de pares semánticos.

Finalmente en el capítulo 5 se presentan las conclusiones del trabajo realizado, divididas en visión general del trabajo, limitaciones, líneas de trabajo analizadas, aportaciones, resultados esperados y obtenidos y trabajos futuros.

1 ALGORITMO BÁSICO DE AGRUPAMIENTO SEMÁNTICO

En el área de recuperación de información, se denomina *agrupamiento semántico* (paradigm en inglés) al un conjunto de palabras semánticamente relacionadas. De acuerdo con Lounsbury (citado por Geckeler [Gec1976]):

> "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"[2]

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

El algoritmo de agrupamiento semántico, presentado en este capítulo, permite agrupar palabras cuyo significado o uso puede considerarse bajo el contexto analizado como sinónimos, aún cuando no guarden una relación sinonímica desde el punto de vista formal. El algoritmo fue desarrollado por Sierra y McNaught [Sie1999] [SiM2000b]. En la presente tesis denominaremos a este algoritmo como *Algoritmo Básico de Agrupamiento Semántico*.

El funcionamiento general 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* [DES1996]
B. ***caída libre***: *descenso de un cuerpo sometido únicamente a la acción de la gravedad* [GDL1996]

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

---

[2] El texto original en inglés es: *"We shall regard as a paradigm any set of linguistic forms wherein: (a) the meaning of every form has a feature in common with the meaning of all other forms of the set, and (b) the meaning of every form differs from that of every other form of the set by one or more additional features"* [Gec1976] (p. 1073)

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 básico 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.

1.1 Alineamiento de dos definiciones

El primer paso del algoritmo de agrupamiento semántico 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 definido por un mismo lexicógrafo.

Para transformar una definición en otra, es necesario establecer un conjunto de operaciones de transformación (inserción, borrado y sustitución de una palabras por otra) que se tiene que aplicar a la definición original para convertirla en la definición objetivo. Una tabla que indica la secuencia de transformaciones que deben aplicarse se denomina *alineamiento*, y por extensión a la acción de obtener un alineamiento se denomina *alineación*.

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.

Así, para las definiciones de *"caída libre"* podemos establecer que una posible manera de alinear estas dos definiciones es:

caída	libre	movimiento	de	un	cuerpo	en	un	campo	gravitatorio	bajo	la	influencia	de	la	gravedad
caída	libre	descenso	de	un	cuerpo	sometido	únicamente	a			la	acción	de	la	gravedad
		Sustitución				Sustitución	Sustitución	Sustitución	Borrado	Borrado		Sustitución			

Tabla 1. Un posible alineamiento de las definiciones del término "caída libre". La primera línea representa la cadena original, la segunda la cadena objetivo y la tercera muestra las operación de transformación.

A partir de esta alineación, se identifican los pares de palabras que son candidatos a formar parte de un mismo grupo semántico. En la siguiente sección se presenta el algoritmo utilizado para obtener el alineamiento de las dos definiciones.

1.1.1 Distancia de edición

Con el fin de alinear un par de definiciones de manera que se obtenga el mínimo costo total de las operaciones aplicadas, se emplea un algoritmo denominado *distancia de edición*. Este algoritmo determina las transformaciones que se deben aplicar a una cadena para convertirla en otra, minimizando el costo total de las operaciones. Aunque originalmente formulado para establecer el grado de similitud de dos cadenas, comparando cada uno de sus caracteres, el algoritmo fácilmente puede ser generalizado para considerar secuencias de palabras en lugar de cadenas de caracteres, de modo que se comparen cada una de las palabras que constituyen la secuencia. Por lo que se considerarán a las palabras como unidades indivisibles.

Se define la *longitud de una secuencia de palabras* A como el número de palabras que constituyen la secuencia y se representa por $n = |A|$. La *secuencia vacía*, ϕ, se define como la secuencia de palabras cuya longitud es cero. La *palabra vacía* (palabra cuyo número de símbolos es cero) la representaremos como ε. La *i-esima palabra* la representamos como a_i, considerando el índice de la primera palabra como 1 y numerando las palabras posteriores ascendentemente. Una *subsecuencia de palabras* se representa por $A_{k..j}$ y comienza con la palabra k y termina hasta la palabra j de la secuencia A. La secuencia $A_{0..0}$ equivale a la secuencia vacía ϕ.

Como ya se mencionó, para transformar la cadena A en la cadena B es necesario aplicar diferentes operaciones; las más comunes son agregar palabras, borrar palabras y sustituir palabras. Si a cada transformación le asignamos un costo, entonces se define a la *distancia de edición* de dos secuencias de palabras, $ed(A,B)$, como el costo mínimo total de las operaciones que deben aplicarse a la secuencia A para transformarla en la secuencia B.

V.I. Levenshtein [Lev1965] y [Lev1966] propuso que las operaciones a considerar fueran:

Insertar la palabra b en la cadena A en la posición i-esima, cuyo costo lo representamos como $w_i(\varepsilon, b)$[3],

Borrar la palabra a de la cadena A en la posición i-esima, el costo de esta operación lo simbolizamos como $w_i(a, \varepsilon)$,

Sustituir la palabra a por la palabra b en la cadena A en la posición i-esima, en donde representamos al costo de esta operación como $w_i(a, b)$;

Además propuso que los costos deberían ser iguales y con valor unitario, a excepción del caso $w_i(a,a)$, para el cual se considera un costo de cero. Bajo estas condiciones, la distancia de edición es igual al número de transformaciones que deben realizarse sobre la secuencia de palabras A. A este valor de edición a menudo se le conoce como "*distancia de edición de Levenshtein*" o simplemente "*distancia de Levenshtein*".

En el algoritmo básico de alineamiento semántico se emplea la técnica propuesta por Wagner y Fisher [WaF1974] para evaluar la distancia de edición de Levenshtein y que se basa en el método de programación dinámica. A continuación se explica éste algoritmo.

Se parte de una matriz C donde las columnas corresponden a cada uno de los símbolos de la secuencia A y los renglones a los símbolos de la secuencia B. Los elementos de la matriz $C_{i,j}$ corresponden a los costos de transformar la secuencia $A_{1..i}$ en la secuencia $B_{1..j}$

El renglón cero representa el costo de las transformaciones que deberían hacerse si la cadena B fuera la secuencia vacía (ϕ). La columna cero corresponde al costo de las transformaciones que deberían hacerse si la secuencia A fuera vacía (ϕ). Debido a que se requiere obtener una secuencia cualquiera (distinta de ϕ) a partir de la secuencia vacía ϕ, y a que las transformaciones son unitarias en costo, es necesario efectuar n inserciones, lo que da por resultado que los costos para esta columna y este renglón estén dados, respectivamente, por: $C_{i,0} = i$ y $C_{0,j} = j$.

La idea básica del algoritmo de programación dinámica nos lleva a una ecuación recurrente, ya que establece que el costo para llegar a la posición $C_{i,j}$ dentro de la matriz de costos se puede calcular con base en el costo de haber llegado a posiciones anteriores, las cuales debieron haber sido calculadas previamente.

Como el objetivo del algoritmo de la distancia de edición es establecer el costo mínimo para transformar una cadena en otra, entonces se propone:

---

[3] A fin de simplificar la notación, si en el texto queda clara la posición para la cual se calcula el valor w_i entonces se omitirá el índice.

$C_{i,0} = i$
$C_{0,j} = j$
$C_{i,j} = min\{ C_{i-1,j}+w(x_i,\varepsilon),\ C_{i,j-1}+w(\varepsilon,y_j),\ C_{i-1,j-1}+w(x_i,y_j) \}$

además

$w(x_i,\varepsilon) = w(\varepsilon,y_j) = 1$
$w(x_i,y_j) = 1$ si $x_i \neq y_j$
$w(x_i,y_j) = 0$ si $x_i = y_j$

Donde la posición $C_{|A|,|B|}$ indica el costo de edición, es decir, $C_{|A|,|B|} = ed(A,B)$. El siguiente algoritmo Wagner y Fisher propusieron es:

Algoritmo: distancia de edición según Wagner y Fisher
Entrada: X, Y *: secuencia de palabras a transformar*
n: Longitud de la secuencia X
m: Longitud de la secuencia Y
Salida: C *: Matriz de costos*
$C_{m,n} = ed(X,Y)$

$C_{0,0} = 0$
for $j = 1$ *to* n
$C_{0,j} = C_{0,j-1} + w(x_j,\varepsilon)$
for $i = 1$ *to* m
$C_{i,0} = C_{i-1,0} + w(y_i,\varepsilon)$
for $j = 1$ *to* n
$C_{i,j} = min\{ C_{i-1,j-1}+w(y_i,x_j),\ C_{i-1,j}+w(y_i,\varepsilon),\ C_{i,j-1}+w(\varepsilon,x_j)\}$

Como puede apreciarse, dado que la evaluación del valor mínimo se efectúa $m \times n$ veces, la complejidad del algoritmo es $O(mn)$ tanto en tiempo como en espacio.

Con este método es posible establecer no sólo el valor de $ed(A,B)$, sino también el conjunto de operaciones requeridas para efectuar la transformación. La secuencia de transformaciones se determina a partir de la matriz completa, Wagner y Fisher sugirieron el siguiente algoritmo para obtener esta secuencia:

Algoritmo: Obtener la secuencia de transformaciones de costo mínimo según Wagner y Fisher
Entrada: *C : Matriz de costos obtenida en el algoritmo anterior*
X , Y : secuencia de palabras a transformar
n: Longitud de la secuencia X
m: Longitud de la secuencia Y
Salida: *La secuencia de transformaciones a efectuar para convertir a la cadena X en la cadena Y*

$i = n$
$j = m$

while (i>0) and (j>0)
if $C_{i,j} = C_{i-1,j} + w(x_i, \varepsilon)$
print(X_i, ε)
$i=i-1$
else if $C_{i,j} = C_{i,j-1} + w(\varepsilon, y_j)$
print(ε, Y_j)
$j = j-1$
else
print(X_i, Y_j)
$i=i-1$
$j=j-1$

El resultado final de la aplicación de este algoritmo es un conjunto secuencial de pares de palabras (incluida la palabra vacía ε) 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.

1.1.2 Aplicación de la distancia de edición de Levenshtein

Como ya se mencionó, se emplea la distancia de edición de Levenshtein dentro del algoritmo de agrupamiento semántico. Aplicando el algoritmo de Wagner y Fisher a las dos definiciones de *"caída libre"* (sección 1.1), obtendremos la siguiente matriz de costos:

	ϕ	caída	libre	movimiento	de	un	cuerpo	en	un	campo	gravitatorio	bajo	la	influencia	de	la	gravedad
ϕ	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
caída	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
libre	2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
descenso	3	2	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14
de	4	3	2	2	1	2	3	4	5	6	7	8	9	10	11	12	13
un	5	4	3	3	2	1	2	3	4	5	6	7	8	9	10	11	12
cuerpo	6	5	4	4	3	2	1	2	3	4	5	6	7	8	9	10	11
sometido	7	6	5	5	4	3	2	2	3	4	5	6	7	8	9	10	11
únicamente	8	7	6	6	5	4	3	3	3	4	5	6	7	8	9	10	11
a	9	8	7	7	6	5	4	4	4	4	5	6	7	8	9	10	11
la	10	9	8	8	7	6	5	5	5	5	5	6	6	7	8	9	10
acción	11	10	9	9	8	7	6	6	6	6	6	6	7	7	8	9	10
de	12	11	10	10	9	8	7	7	7	7	7	7	7	8	7	8	9
la	13	12	11	11	10	9	8	8	8	8	8	8	7	8	8	7	8
gravedad	14	13	12	12	11	10	9	9	9	9	9	9	8	8	9	8	7

Tabla 2. Matriz de costos obtenida a partir del algoritmo de Wagner y Fisher para el cálculo de la distancia de edición

La secuencia de operaciones con costo mínimo es la siguiente:

Def. A	caída	libre	movimiento	de	un	cuerpo	en	un	campo	gravitatorio	bajo	la	influencia	de	la	gravedad
Def. B	caída	libre	descenso	de	un	cuerpo			sometido	únicamente	a	la	acción	de	la	gravedad
Operación *			***S***				***B***	***B***	***S***	***S***	***S***		***S***			
Costos	0	0	1	1	1	1	2	3	4	5	6	6	7	7	7	7

*Tipos de operciones aplicadas **S** = Sustitución, **B** = Borrado, **I** = Inserción*

Tabla 3. Alineamiento del término "caída libre" cuyo costo total de las operaciones es minimo, este alineamiento fue obtenido de acuerdo con el algoritmo de Wagner y Fisher

De acuerdo con el tipo de operaciones que se pueden efectuar, los pares se clasifican en:

a) *Par-igual* (*Equal couple*). 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* (*Matched Couple*). 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.
c) *Par-nulo* (*Null Couple*). Aquella pareja formada por una palabra y la palabra vacía ε, de forma que una palabra debe agregarse (ε, palabra) o borrarse (palabra, ε).

En principio, los pares-nulos carecen de interés semántico, pues indican que debe agregarse o eliminarse una palabra en la definición. En el ejemplo se tienen los siguientes pares nulo: *(en, ε), (un, ε)*.

Los pares-correspondientes, por el contrario, indican que debe sustituirse una palabra por otra para llegar a la cadena destino. En el ejemplo se tienen los siguientes pares correspondientes: (movimiento, descenso), (campo, sometido), (gravitatorio, únicamente), (bajo, a), (influencia, acción).

Estos pares 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.

A fin de establecer lo anterior es necesario determinar el grado de similitud de las palabras que forman el par-correspondiente; dentro del algoritmo de alineamiento semántico se emplean los pares-iguales para determinar el grado de similitud, tal como se presenta en la siguiente sección.

1.2 Cálculo de similitud de pares-correspondientes

Como una medida de comparación entre dos palabras de un par-correspondiente, se propuso, en el algoritmo de agrupamiento semántico básico, 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 derecha y a la izquierda del par-correspondiente, estableciendo cuántos pares-iguales existen a ambos lados antes de encontrar un par-nulo u otro par-correspondiente. El número de pares-iguales más el par-correspondiente es el valor de LCC que se asigna al par analizado.

En la tabla siguiente puede apreciarse el valor de LCC para los pares correspondientes identificados.

Def. A	caída	libre	movimiento	de	un	cuerpo	en	un	campo	gravitatorio	bajo	la	influencia	de	la	gravedad
Def. B	caída	libre	descenso	de	un	cuerpo			sometido	únicamente	a	la	acción	de	la	gravedad
Costos	0	0	1	1	1	1	2	3	4	5	6	6	7	7	7	7
Tipo*	*I*	*I*	*C*	*I*	*I*	*I*	*N*	*N*	*C*	*C*	*C*	*I*	*C*	*I*	*I*	*I*
LCC			*6*						*1*	*1*	*2*		*5*			

* *Tipos de pares identificados **I** = Par igual, **N** = Par nulo, **C** = Par correspondiente*

Tabla 4. Alineamiento semántico de las definiciones del término "Caída Libre", se muestra además los tipos de pares identificados y el valor del coeficiente de similitud LCC calculado para los pares-correspondientes

1.3 Determinación de pares-vinculados

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"*, cuyo alineamiento se mostró en la tabla 4, 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 Sierra y McNaught [Sie1999] determinaron que, para el inglés, un valor de LCC de 5 sugiere un buen grado de similitud. Para el español es necesario hacer una evaluación del valor adecuado de LCC, sin embargo en esta tesis se plantea la evaluación del desempeño del algoritmo para un corpus en inglés, por lo que sin otro elemento la evaluación para el español sale de los objetivos de esta tesis.

Además Sierra y McNaught encontraron que se requiere al menos un par-igual a la derecha y uno a la izquierda del par-correspondiente, denominada ésta *condición de frontera*, para que las palabras de este par-correspondiente sean susceptibles de considerarse intercambiables.

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 6 y 5, respectivamente, además de que ambos cumplen con la condición de frontera, por lo que estos dos pares se consideran pares-vinculados.

1.4 Lista de palabras irrelevantes

Para propósitos de recuperación de información, las palabras en una definición pueden ser palabras clave (o relevantes) o palabras 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 la definición de *"caída libre"*, las palabras *"descenso"*, *"cuerpo"*, *"campo"*, *"gravitatorio"*, *"bajo"*, *"influencia"*, *"gravedad"* pueden considerarse palabras clave mientras que *"de"*, *"un"*, *"en"*, *"a"*, *"la"* son ejemplos de palabras irrelevantes. Por ello, debe darse un tratamiento especial a los pares-vinculados que contengan palabras irrelevantes durante el alineamiento.

Por ejemplo, para las siguientes definiciones:

A. ***Mecánica:*** *parte de la física que trata del equilibrio y del movimiento de los cuerpos sometidos a cualesquiera fuerzas* [RAE1992]
B. ***Mecánica:*** *parte de la física que trata del movimiento y el equilibrio y de las fuerzas que los producen* [MMol1996]

Al aplicar el algoritmo de alineamiento obtenemos los siguientes pares:

mecánica	parte	de	la	física	que	trata	del	**equilibrio**	y	**del**	**movimiento**	ε	de	los	cuerpos	sometidos	a	**cualesquiera**	fuerzas
mecánica	parte	de	la	física	que	trata	del	**movimiento**	y	**el**	**equilibrio**	y	de	las	fuerzas	que	los	**producen**	ε

Tabla 5. Alineamiento obtenido al aplicar el algoritmo de alineamiento semántico a dos definiciones del término: *mecánica*.

Como puede observarse, existen parejas de palabras (del, el) (los, las) (a, los) que poca información relevante suministran para los fines de identificación de alineamientos semánticos, por lo que estos pares deben ser rechazados como pares-vinculados. El par (sometidos, que) también debe rechazarse puesto que contiene una palabra, *"que"*, sin información relevante. Por el contrario, los pares (equilibrio, movimiento) (movimiento, equilibrio) (cuerpos, fuerzas) (cualesquiera, producen) deberán ser analizados a fin de determinar si estos pares son susceptibles de ser considerados pares-vinculados.

En el algoritmo de alineamiento semántico básico se emplea una lista de palabras irrelevantes a fin de rechazar aquellos pares-vinculados 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.

1.5 Sustitución de 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.

Por ejemplo, para *"caída libre"* y utilizando los pares-vinculados (*movimiento, descenso*) y (*influencia, acción*), al realizar el proceso antes descrito, obtenemos:

Def. 1	caída	libre	**~~movimiento~~**	de	un	cuerpo	en	un	campo	gravitatorio	bajo	la	**influencia**	de	la	gravedad
Def. 2	caída	libre	**descenso**	de	un	cuerpo			sometido	únicamente	a	la	**acción**	de	la	gravedad
Costos	0	0	**1**	1	1	1	2	3	4	5	6	6	**7**	7	7	7
LCC	0	0	**6**	0	0	0	0	0	1	1	2	0	**5**	0	0	0
Tipo*	*I*	*I*	***C***	*I*	*I*	*I*	*N*	*N*	*M*	*M*	*M*	*I*	***M***	*I*	*I*	*I*

Def. 1	caída	libre	**descenso**	de	un	cuerpo	en	un	campo	gravitatorio	bajo	la	**acción**	de	la	gravedad
Def. 2	caída	libre	**descenso**	de	un	cuerpo			sometido	únicamente	a	la	**acción**	de	la	gravedad
Costos	0	0	**0**	0	0	0	1	2	3	4	5	5	**5**	5	5	5
LCC	0	0	**0**	0	0	0	0	0	1	1	6	0	**0**	0	0	0
Tipo*	*I*	*I*	***I***	*I*	*I*	*I*	*N*	*N*	*C*	*C*	*C*	*I*	***I***	*I*	*I*	*I*

** Tipos de pares identificados **I** = Par igual, **N** = Par nulo, **C** = Par correspondiente*

Tabla 6. Proceso de sustitución de pares-vinculados. En el primer alineamiento se muestra la sustitución de *movimiento* por *descenso* y *acción* por *influencia* en la definición del término *caída libre*. En el segundo, se ha recalculado el alineamiento con base a estas nuevas definiciones y por tanto ha variado el valor de LCC.

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.

Sierra y McNaught encontraron que al hacer esto el número de pares-vinculados identificados aumentaba. Es importante hacer notar que la sustitución de una palabra por la otra no puede aplicarse de manera indiscriminada; de hecho, también determinaron que para realizar la sustitución es necesario que las dos palabras del par-vinculado deben aparecer 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.

1.6 Generación de agrupamientos semánticos

El algoritmo básico de agrupamiento semántico 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 (sección 1.5). 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 los agrupamientos semánticos antes de comenzar el nuevo ciclo.

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 regal de transitividad de pares-vinculados podemos afirmar que el conjunto $\{a, b, c\}$ forman un agrupamiento semántico. De manera que se forman agrupamientos, creando conjuntos con de todas las palabras que satisfacen la relación de transitividad entre ellas con base en los pares-vinculados identificados.

1.7 Lematización

El proceso computacional que remueve los sufijos de las palabras y transforma las palabras en sus correspondientes raíces se denomina *lematización*. Así por ejemplo a *casa*, *casero*, *casita* le corresponde el lexema *cas*.

Si estas palabras son identificadas durante el proceso de alineamiento (básicamente durante el cálculo de la distancia de Levenshtein), la efectividad en la identificación de pares-vinculados puede incrementarse debido al hecho de que el número total de términos distintos se reducen como consecuencia del proceso de lematización.

En particular se emplea el algoritmo de lematización de Porter [Por1980], debido a que este algoritmo ha demostrado presentar mejores resultados que otros similares [Fra1992].

Se ha encontrado experimentalmente que la utilización de palabras lematizadas (es decir, el uso de los lexemas, sin tomar en cuenta los sufijos) en lugar de las palabras originales durante el proceso de alineamiento produce mejores resultados. Por otro lado, debido a que el objetivo es obtener agrupamientos semánticos, se consideran semánticamente similares una palabra y sus derivaciones y flexiones.

1.8 *Esquema general del algoritmo de agrupamiento semántico*

Integrando todos los elementos que se han expuesto, a continuación presentamos un diagrama que sintetiza todo el algoritmo [SiM2000b]:

TESIS CON
FALLA DE ORIGEN

Como se observa en el diagrama, el proceso comienza con la lectura de parejas de definiciones de un mismo término. Las definiciones se lematizan (sección 1.7) y con base en estas definiciones lematizadas, se calcula la distancia de Levenshtein (sección 1.1.1). Como resultado del cálculo se obtiene un alineamiento de las dos definiciones analizadas (sección 1.1), determinando con base en este alineamiento los valores de LCC para cada alineamiento (sección 1.2). Si el valor de LCC de una pareja en particular es mayor o igual a cinco (sección 1.3) y ninguna de las dos palabras se encuentra dentro de la lista de palabras de irrelevantes (sección 1.4) entonces el par se considera vinculado y se

Algoritmo de agrupamiento semántico [SiM2000b]

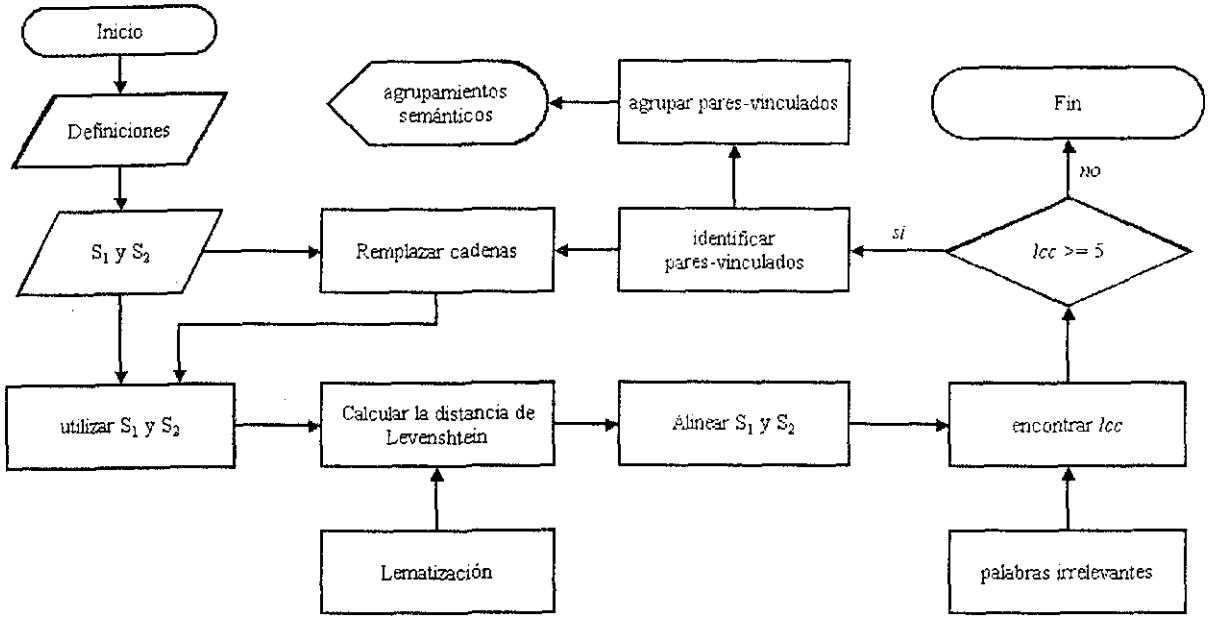


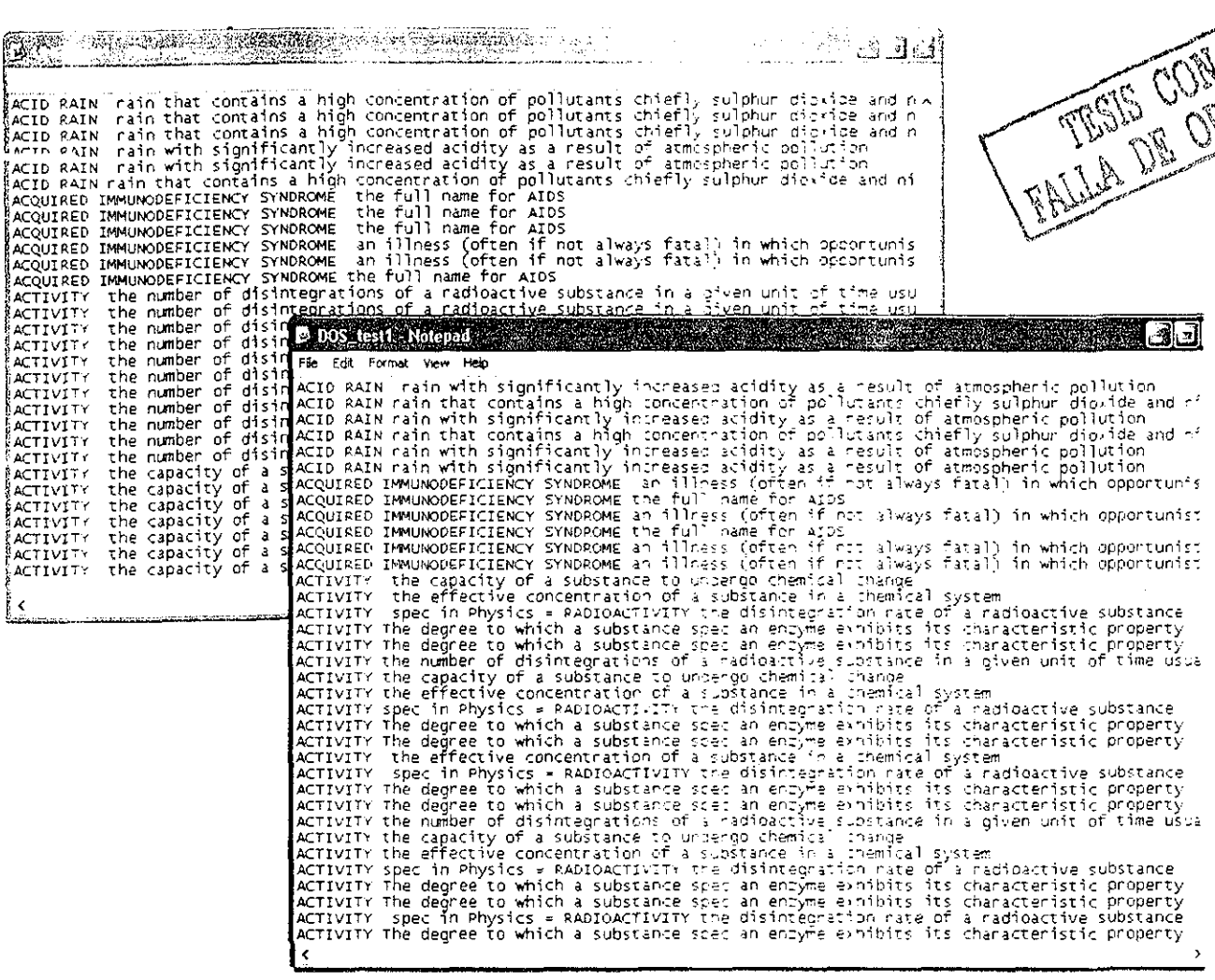
Figura 1. Algoritmo básico de agrupamiento semántico, tal como lo proponen Sierra y McNaught [SiM2000b]

integra a los agrupamientos semánticos (sección 1.6). Además, el par-vinculado se sustituye en las definiciones (sección 1.5). El proceso se repite hasta que no se generen nuevos pares-vinculados.

1.9 Implementación del algoritmo básico de alineamiento semántico

La implementación de este algoritmo se realizó en un ambiente UNIX, en un programa en el lenguaje C. Aunque se han hecho adecuaciones al programa para ejecutarlo en DOS, las adecuaciones limitan el programa resultante porque existe una restricción respecto al tamaño máximo de un arreglo en DOS, motivo por el que es mejor ejecutarlo en el ambiente en el que se diseño el programa.

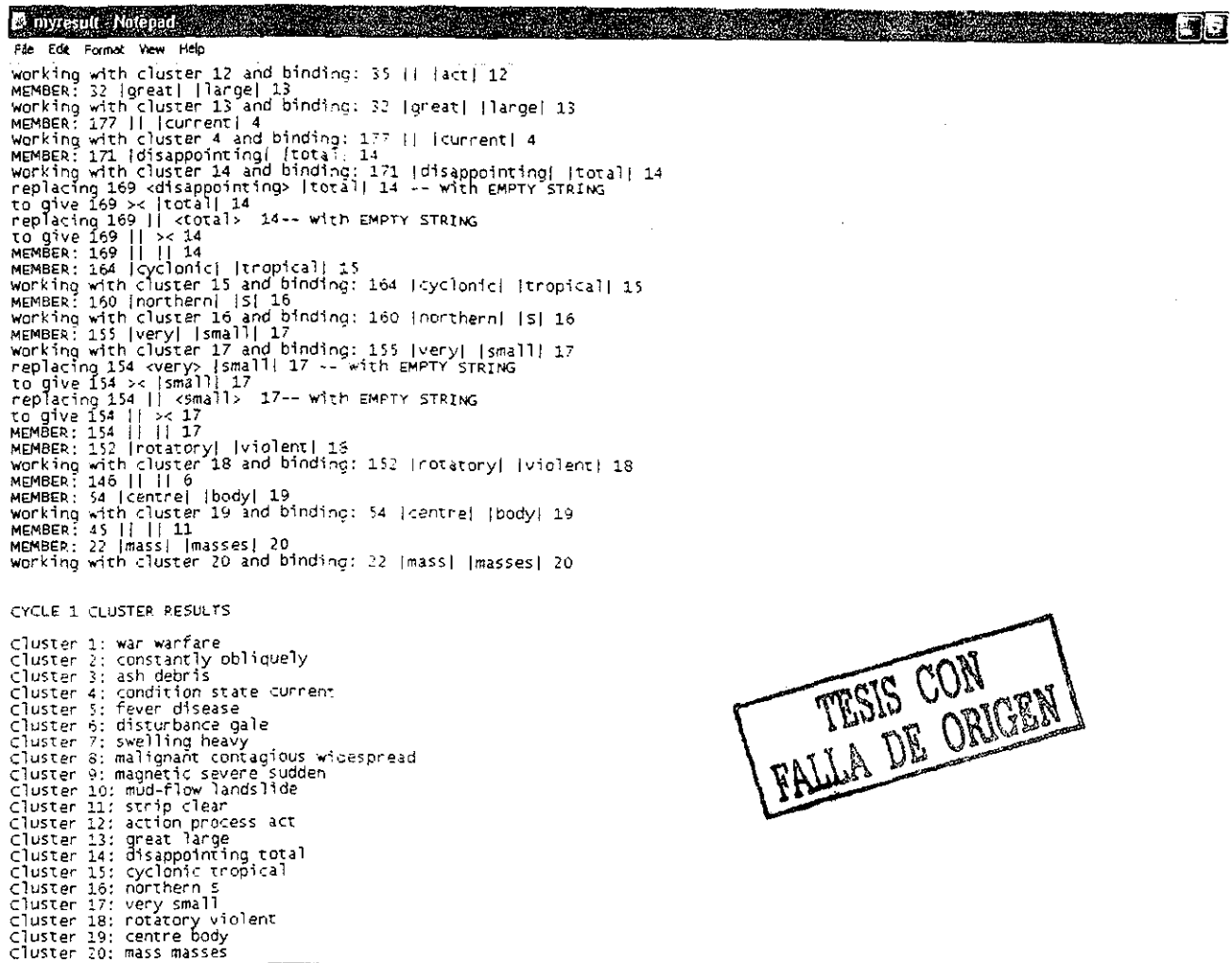
El programa utiliza como entrada dos archivos que suministran cada una de las definiciones que deben aparearse, es decir la primera línea del archivo uno con la primera línea del archivo dos.



TESIS CON
FALLA DE ORIGEN

Figura 2. Archivos que recibe como entrada el programa desarrollado por Sierra y McNaught. Las definiciones que deben alinearse se determinan apareando cada uno de los renglones de los dos archivos. No existe un procedimiento automatizado para realizar este proceso por lo que debe hacerse manualmente.

Los resultados obtenidos es un listado cuya información puede elegirse por medio de parámetros entregados al sistema en el momento en que se ejecuta el programa.



myresult - Notepad

File Edit Format View Help

```
working with cluster 12 and binding: 35 || |act| 12
MEMBER: 32 |great| |large| 13
working with cluster 13 and binding: 32 |great| |large| 13
MEMBER: 177 || |current| 4
working with cluster 4 and binding: 177 || |current| 4
MEMBER: 171 |disappointing| |total| 14
working with cluster 14 and binding: 171 |disappointing| |total| 14
replacing 169 <disappointing> |total| 14 -- with EMPTY STRING
to give 169 >< |total| 14
replacing 169 || <total>  14-- with EMPTY STRING
to give 169 || >< 14
MEMBER: 169 || || 14
MEMBER: 164 |cyclonic| |tropical| 15
working with cluster 15 and binding: 164 |cyclonic| |tropical| 15
MEMBER: 160 |northern| |S| 16
working with cluster 16 and binding: 160 |northern| |S| 16
MEMBER: 155 |very| |small| 17
working with cluster 17 and binding: 155 |very| |small| 17
replacing 154 <very> |small| 17 -- with EMPTY STRING
to give 154 >< |small| 17
replacing 154 || <small>  17-- with EMPTY STRING
to give 154 || >< 17
MEMBER: 154 || || 17
MEMBER: 152 |rotatory| |violent| 18
working with cluster 18 and binding: 152 |rotatory| |violent| 18
MEMBER: 146 || || 6
MEMBER: 54 |centre| |body| 19
working with cluster 19 and binding: 54 |centre| |body| 19
MEMBER: 45 || || 11
MEMBER: 22 |mass| |masses| 20
working with cluster 20 and binding: 22 |mass| |masses| 20


CYCLE 1 CLUSTER RESULTS

Cluster 1: war warfare
Cluster 2: constantly obliquely
Cluster 3: ash debris
Cluster 4: condition state current
Cluster 5: fever disease
Cluster 6: disturbance gale
Cluster 7: swelling heavy
Cluster 8: malignant contagious widespread
Cluster 9: magnetic severe sudden
Cluster 10: mud-flow landslide
Cluster 11: strip clear
Cluster 12: action process act
Cluster 13: great large
Cluster 14: disappointing total
Cluster 15: cyclonic tropical
Cluster 16: northern S
Cluster 17: very small
Cluster 18: rotatory violent
Cluster 19: centre body
Cluster 20: mass masses
```

TESIS CON
FALLA DE ORIGEN

Figura 3. Resultados ofrecidos por el programa desarrollado por Sierra y McNaught. Como puede observarse los resultados se entregan en un archivo tipo texto, la información obtenida, aunque está organizada es difícil de entender a primera vista.

1.10 Resultados

El algoritmo básico fue aplicado a un corpus de metrología en el idioma inglés. Los resultados han sido reportados en [Sie1999] y [SiM2000b]. El algoritmo se aplicó sobre *un diccionario terminológico en el área de metrología en el idioma inglés. El diccionario* está constituido por 342 términos. El algoritmo de básico de alineamiento semántico identificó 32 pares-vinculados, de los cuales 30 son pares-semánticos.

Estos pares generaron 17 agrupamientos semánticos, los resultados se muestran en la tabla siguiente.

1. mass weight	11. hyperbolic radio radiofrequency
2. conditions variations	12. observing tracing
3. swinging turning	13. day sunlight
4. direction inclination	14. apparatus instrument telescope
5. accurate precise	15. amount concentration intensity percentage portion rate salinity strength
6. distances heights	16. celestial heavenly
7. set specific	17. analyse ascertaining determining estimating location measuring recording taking testing
8. method system	
9. field limits	
10. frequency wavelength	

Tabla 7. Agrupamientos semánticos obtenidos por medio del algoritmo de agrupamiento semántico básico al aplicarlo sobre un diccionario de términos en el área de metrología en el idioma inglés

Sierra y McNaught no hacen una evaluación cuantitativa de los resultados, sin embargo, En esta tesis 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

Como puede apreciarse, sólo se recuperaron 30 de los 363 pares posibles, esta baja eficiencia en la recuperación de los pares, originó que se buscará mejorar el desempeño del algoritmo básico de agrupamiento semántico en términos de recuperación de pares semánticos. Cabe hacer notar que, los pares-semánticos son la base para la generación de los agrupamientos, por lo tanto una baja eficiencia en la recuperación de los pares semánticos incide directamente en la generación de los agrupamientos semánticos

1.11 Recapitulación

En este capítulo se presentó el funcionamiento del algoritmo básico de agrupamiento semántico. Se introdujeron los conceptos de alineamiento, par-igual, par-nulo, par-correspondiente, par-vinculado y par-semántico. Se explicó que el algoritmo básico de agrupamiento semántico es un método inferencial que se basa en examinar las definiciones de un término (a partir de determinar los mínimos cambios posibles que es

necesario hacer a una definición para llegar a otra), identifica las palabras que guardan una relación léxica y a partir de esta relación infiere su aplicación a otros contextos.

Tal como se comentó en la sección anterior, la baja eficiencia en la recuperación de pares semánticos, originó el interés de buscar alternativas que mejoraran el desempeño del algoritmo semántico. En capítulo siguiente, se proponen un conjunto de modificaciones con el objetivo de mejorar el desempeño del algoritmo.

2 VARIANTES DEL ALGORITMO BÁSICO

Como ya se mencionó, el algoritmo de agrupamiento semántico básico obtiene agrupamientos semánticos consistentes, sin embrago es posible afinar la técnica.

En cada una de las etapas de las que consta el algoritmo es posible plantear alternativas que mejoren la calidad de los agrupamientos semánticos identificados. Las alternativas propuestas pueden clasificarse en dos grandes rubros:

- *Heurísticas alternativas.* Aquellas heurísticas que llevan al planteamiento de diferentes alternativas dentro de los algoritmos de agrupamiento semántico.
- *Interfaz hombre-computadora.* Aquellas tendientes a mejorar la interfaz hombre-computadora, con el objetivo de que el lingüista computacional pueda aplicar el algoritmo en diferentes campos del conocimiento.

En este capítulo se plantean algunas de estas heurísticas y modificaciones en la interfaz hombre-computadora, esbozándose las posibles alternativas y los resultados esperados. Si bien es cierto que durante el desarrollo de esta tesis se han abordado algunas de estas líneas, no es el objetivo de esta tesis explorar todas las Heurísticas alternativas, pues esto sería materia de una tesis doctoral.

2.1 Selección de la base de datos terminológica

Esta etapa no existe en el sistema original, el objetivo de esta etapa es facilitar la aplicación del algoritmo a diferentes diccionarios terminológicos. Durante esta etapa el usuario del sistema elegirá el diccionario con el cual trabajará durante el proceso de identificación de paradigmas semánticos. Para ello, algunas de las opciones que podrá manejar el usuario son:

- *Base de datos.* Identifica cuál de las bases de datos disponibles se empleará para el algoritmo
- *Área temática.* Comúnmente las bases de datos podrán manejar diferentes áreas temáticas. Por ejemplo en Lingüística se tienen las áreas temáticas de dialectología, fonética, lexicografía, etc.
- *Número de palabras.* Esta opción limita el número de palabras que se considerarán dentro de la definición; las restantes se ignorarán.
- *Idioma.* El Idioma de la base de datos determinará propiedades tales como la lista de palabras irrelevantes que se emplearán, el lematizador, las palabras funcionales, etc.
- *Términos.* Una vez fijada la base y el área temática será posible elegir cuáles de los términos disponibles se emplearán para alimentar al algoritmo.

Considerando lo anterior, la propuesta en este rubro es:

INTERFAZ HOMBRE-COMPUTADORA 1

ALGORITMO BÁSICO

EL ALGORITMO REQUIERE QUE SE CREEN DOS ARCHIVOS, DONDE CADA UNO APORTA UNA DE LAS DOS DEFINICIONES DEL TÉRMINO A ALINEAR. LA ESTRUCTURA DE ESTOS ARCHIVOS, SI BIEN ES CIERTO ES SENCILLA, RESULTA ENGORROSA SU CREACIÓN. ADEMÁS, CUALQUIER MODIFICACIÓN DEBE HACERSE MANUALMENTE Y EN AMBOS ARCHIVOS. (EL CORPUS ORIGINAL SE ENCUENTRA EN INGLÉS Y CONSISTE SOLAMENTE DE TÉRMINOS DEL ÁREA DE METROLOGÍA)

PROPUESTA

DISEÑAR UNA BASE DE DATOS RELACIONAL QUE PERMITA LA CAPTURA Y MANTENIMIENTO DE DIFERENTES DICCIONARIOS Y QUE PERMITA OBTENER LA INFORMACIÓN REQUERIDA.

2.2 Preprocesamiento

La versión original del algoritmo de agrupamiento semántico no realiza ninguna modificación sobre las definiciones originales de los términos. Más aún, se emplea un par de archivos tipo texto como fuentes de entrada. Actualmente se está modificando el algoritmo para tomar las definiciones de la base de datos relacional (desarrollada en MS-SQL).

El preprocesamiento permite preparar las definiciones de modo que se modifiquen las definiciones a fin de mejorar el análisis, como por ejemplo eliminar aquellas partes que se consideren como irrelevantes. A continuación se presentan las diferentes posibilidades de preprocesamiento, todas ellas opcionales y que deberán ser elegidas por el usuario.

2.2.1 Acentos y diéresis para el español

El algoritmo de lematización de Porter [Por1980] fue originalmente planteado para la lematización de palabras en inglés, pero existen adaptaciones para poder aplicarlo a otros idiomas. En particular, en el caso del español, el algoritmo disponible presenta una serie de deficiencias durante la lematización, por lo que es necesario mejorar su desempeño. De hecho, para poder aplicarlo es necesario transformar las palabras acentuadas en dos caracteres al menos, lo mismo ocurre con las *eñes* y las diéresis. Como ejemplo de problemas de lematización considérense las siguientes palabras a lematizar:

Palabra original	Palabra lematizada
demos	dem
demostró	demostr
demostrar	demostr
demuestro	demuestr
demostrarían	demostr
ver	ver
verían	ver
vió	vio
vi	vi
vierón	vieron
público	public
publico	public
publicó	public

Tabla 8. Lematización de algunas palabras del español. Resultados obtenidos utilizando el algoritmo de Porter para el español. Obsérvese la pérdida del significado en el caso de publicar.

Una posible Línea de trabajo es:

LÍNEA DE TRABAJO 1

ALGORITMO BÁSICO

EL ALGORITMO ESTÁ ESCRITO PARA APLICARSE EN UN DICCIONARIO DE TÉRMINOS EN LENGUA INGLESA, EN CONSECUENCIA EL ALGORITMO DE LEMATIZACIÓN UTILIZADO FUE DESARROLLADO PARA EL INGLÉS.

PROPUESTA

ADECUAR EL ALGORITMO AL CONJUNTO DE CARACTERES ISO-LATIN-1 A FIN DE PODER APLICARLO SIN TRANSFORMAR LAS PALABRAS. ADEMÁS, DEDICAR UN TIEMPO RAZONABLE A LA REFINACIÓN DEL ALGORITMO EN ESPAÑOL.

2.2.2 Remoción de puntuación

Es necesario, como parte de la interfaz hombre-máquina, eliminar aquellos símbolos de puntuación no relevantes y que, sin embargo, pueden hacer que el algoritmo se comporte de manera diferente a la esperada. Ejemplo de lo anterior son las comas (,) los apostrofes ('), los puntos (.) etc. También es importante considerar la adecuada interpretación de los paréntesis y guiones.

La Línea de trabajo a desarrollar es:

LÍNEA DE TRABAJO 2

ALGORITMO BÁSICO

EL ALGORITMO NO REALIZA NINGUNA MODIFICACIÓN SOBRE LA PUNTUACIÓN. SIN EMBARGO, DURANTE LAS PRUEBAS REALIZADAS SE ELIMINÓ MANUALMENTE LA PUNTUACIÓN.

PROPUESTA

ESTABLECER LA CORRECTA IDENTIFICACIÓN DEL USO DE LOS SIGNOS DE PUNTUACIÓN Y SU INCIDENCIA DENTRO DEL ALGORITMO DE ALINEAMIENTO SEMÁNTICO, A FIN DE IDENTIFICAR EN QUÉ CASOS ES POSIBLE ELIMINAR LA PUNTUACIÓN Y LAS CONDICIONES BAJO LAS CUALES LA PUNTUACIÓN IMPLICA LA SEPARACIÓN DE IDEAS Y, POR ENDE, LA SEPARACIÓN ADECUADA DE LOS TEXTOS.

2.2.3 Expresiones léxicas

Cuando se dispone de una definición de un término que está compuesta por dos frases entrelazadas por conectivos tales como: *o, y, pero, ni,* etc., es necesario determinar si en realidad se está hablando de una o varias definiciones.

Por ejemplo, al considerar la siguiente definición:

A. ***aceleración**: Es la variación de la velocidad de un cuerpo* **o** *partícula por unidad de tiempo* [GDL1996]

Debido a la existencia de la disyunción *o*, esta definición podría descomponerse en dos:

B. ***aceleración**: Es la variación de la velocidad de un cuerpo por unidad de tiempo*
C. ***aceleración**: Es la variación de la velocidad de una partícula por unidad de tiempo*

LÍNEA DE TRABAJO 3

ALGORITMO BÁSICO

EL ALGORITMO NO REALIZA CONSIDERACIÓN ALGUNA, MANUAL O AUTOMÁTICA, RESPECTO A ESTAS EXPRESIONES.

PROPUESTA

IDENTIFICAR LOS ALGORITMO EXISTENTES Y APLICARLOS AQUÍ PARA UNA CORRECTA IDENTIFICACIÓN DE LAS CONJUNCIONES DISYUNTIVAS Y COPULATIVAS, ASÍ COMO LA ADECUADA SEPARACIÓN DE LAS ORACIONES. EN CASO DE NO EXISTIR ALGORITMOS ADECUADOS DESARROLLARLOS.

2.2.4 Identificación de expresiones léxicas

Existen expresiones léxicas, formadas por colocaciones (por ejemplo, *de un, de acuerdo con, con lo que,* etc.) o términos compuestos (v.gr., *grado superlativo*, *movimiento uniformemente rectilíneo*, etc.) que, aunque formadas por al menos dos palabras, deben ser adecuadamente identificadas para que se consideren como una sola unidad léxica.

LÍNEA DE TRABAJO 4

ALGORITMO BÁSICO

EL ALGORITMO NO REALIZA IDENTIFICACIÓN DE UNIDADES LÉXICAS, POR LO QUE EXPRESIONES TALES COMO: "DE UN", "POR LO TANTO", "POR EJEMPLO" ETC. SON IDENTIFICADAS MANUALMENTE Y SE UNEN ESTAS EXPRESIONES MEDIANTE UN GUIÓN BAJO ("DE_UN", "POR_LO_TANTO", ETC.) DE MODO QUE EL ALGORITMO CONSIDERE ESTAS EXPRESIONES COMO UNA SOLA PALABRA.

PROPUESTA

IDENTIFICAR LOS ALGORITMO EXISTENTES Y APLICARLOS AQUÍ PARA UNA CORRECTA IDENTIFICACIÓN DE LOS TÉRMINOS COMPUESTOS Y COLOCACIONES, PARA SU CORRECTO TRATAMIENTO EN EL ALGORITMO DE ALINEAMIENTO SEMÁNTICO CONSIDERANDO ESTOS TÉRMINOS COMO UNA SOLA UNIDAD. EN CASO DE NO EXISTIR ALGORITMOS ADECUADOS DESARROLLARLOS.

Recientemente, en la Universidad de Manchester se ha comenzado a trabajar en la identificación automática de términos compuestos; la aplicación de un algoritmo de este tipo puede llevar a un mejor desempeño del algoritmo de agrupamiento semántico.

2.2.5 Clasificación gramatical de las palabras

La identificación de las partes de la oración, esto es, la clasificación gramatical de las palabras (verbo, sustantivo, adjetivo, etc.) permitiría limitar al algoritmo a fin de evitar en lo posible agrupar palabras que pertenezcan a diferentes categorías gramaticales.

LÍNEA DE TRABAJO 5

ALGORITMO BÁSICO

NO INCORPORA LA IDENTIFICACIÓN DE LAS PARTES DE LA ORACIÓN, NI AUTOMÁTICA NI MANUALMENTE.

PROPUESTA

ESTABLECER LOS ALGORITMOS NECESARIOS PARA IDENTIFICAR LAS CATEGORÍAS GRAMATICALES EN UNA ORACIÓN. AL INCLUIR LAS CATEGORÍAS

TESIS CON
FALLA DE ORIGEN

GRAMATICALES EN EL ALGORITMO DE ALINEAMIENTOS SEMÁNTICO SE PUEDEN ALINEAR PALABRAS CON LA MISMA CATEGORÍA GRAMATICAL O CON CATEGORÍAS GRAMATICALES AFINES (EN ESTE ÚLTIMO CASO, ES NECESARIO ESTABLECER QUÉ SE ENTENDERÁ POR CATEGORÍA GRAMATICAL AFÍN).

Actualmente existen varios algoritmos, tanto basados en reglas, estadísticos o híbridos, para realizar la identificación de las partes de la oración; sin embargo, como es natural, dependen en gran medida del idioma y es necesario evaluarlos y entrenarlos antes de aplicarlos.

2.3 Modificaciones al algoritmo de distancia de edición

Los cambios sugeridos al algoritmo utilizado para calcular la distancia de edición, y por ende el alineamiento, son:

1. Permitir la modificación de los costos de cada una de las operaciones posibles.
2. Mejorar el tiempo de cálculo y disminuir el espacio requerido para aplicar el algoritmo de Wagner & Fisher.
3. Agregar, como una de las operaciones posibles, el intercambio de dos palabras consecutivas

A continuación se discuten brevemente las posibilidades de estas alternativas.

2.3.1 Modificación de los costos de las operaciones permitidas

Como podrá recordarse, el algoritmo de Levenshtein establece el mínimo número de operaciones de inserción, borrado y sustitución para que, a partir de una cadena, se llegue a otra. Sin embargo, es posible asignar diferentes costos a cada una de las trasformaciones básicas, con lo que se tendrían elementos de evaluación de posibles variantes.

Por ejemplo, al asignar los siguientes costos $w(a,\varepsilon)$ = 2, $w(\varepsilon,b)$ =2, $w(a,b)$ = 1, es de esperarse que se favorezca la sustitución sobre la inserción y el borrado. La modificación a la interfaz permitirá la evaluación de esta u otra propuesta de costos.

LÍNEA DE TRABAJO 6

ALGORITMO BÁSICO

CONSIDERA LOS COSTOS DE LAS OPERACIONES DE SUSTITUCIÓN, INSERCIÓN Y BORRADO IGUALES A UNO. EN EL CASO DE ALINEAMIENTO DE SÍMBOLOS IGUALES, CONSIDERA UN COSTO IGUAL A CERO.

PROPUESTA

ESTABLECER SILOS COSTOS ASIGNADOS A CADA UNA DE LAS OPERACIONES SON LOS ADECUADOS; EN CASO DE NO SERLOS IDENTIFICAR LOS VALORES MÁS CONVENIENTES, TOMANDO EN CUENTA QUE EL OBJETIVO DEL ALGORITMO ES LA ALINEACIÓN DE PALABRAS SEMÁNTICAMENTE RELACIONADAS.

2.3.2 Desempeño del algoritmo de distancia de edición

Otra posibilidad para afinar el algoritmo de alineamiento semántico es evaluar algoritmos alternos para el proceso de alineación.

Como ya se comentó, el algoritmo para el cálculo de la distancia de edición propuesto por Wagner y Fisher tiene un orden de complejidad $O(mn)$ tanto en tiempo como en espacio. Esto hace que para textos grandes el alineamiento sea lento y en situaciones extremas imposible.

LÍNEA DE TRABAJO 7

ALGORITMO BÁSICO

SE EMPLEA EL ALGORITMO DE WAGNER Y FISHER.

PROPUESTA

IDENTIFICAR LOS ALGORITMOS POSIBLES PARA LA EVALUACIÓN DEL PROCESO DE ALINEAMIENTO CONSIDERANDO QUE ESTOS DEBEN SER DE UN ORDEN DE COMPLEJIDAD MENOR A $O(mn)$ YA SEA EN TIEMPO O EN ESPACIO.

2.3.3 Operación de inversión de dos palabras consecutivas

Existe la posibilidad de ampliar las operaciones que pueden realizarse para transformar una secuencia de palabras en otra, tal como incluir el intercambio de palabras consecutivas. De este modo, se tendrían como operaciones posibles para el proceso de alineamiento la inserción de una palabra, el borrado de una palabra, la sustitución de una palabra y el intercambio de dos palabras consecutivas. Con la inclusión de esta operación, es posible alinear palabras que sean intercambiables sin modificar el sentido de la frase.

En español es común el intercambio del sustantivo y adjetivo cuando estos se expresan en conjunto, así por ejemplo "velocidad alta" y "alta velocidad" hacen referencia al mismo concepto y con las mismas características.

Por ejemplo, si alineamos las siguientes definiciones:

A. ***Estroboscopio**: Aparato de alta velocidad que permite congelar para la vista un movimiento periódico* [EDE1986]
B. ***Estroboscopio**: Instrumento de velocidad alta que permite visualizar movimientos periódicos* [DCO1983]

El algoritmo de agrupamiento semántico actual nos ofrece el siguiente resultado

Def. 1	estroboscopio	aparto	de	alta	velocidad	que	permite
Def. 2	estroboscopio	instrumento	de	velocidad	alta	que	permite
LCC	0	3	0	2	3	0	0
Tipo*	*I*	*C*	*I*	*C*	*C*	*I*	*I*

Def. 1	...	congelar	para	la	vista	un	movimiento	periódico
Def. 2	...					visualizar	movimientos	periódicos
LCC	...	0	0	0	0	3	0	0
Tipo*	...	*N*	*N*	*N*	*N*	*C*	*I*	*I*

* *Tipos de pares identificados I = Par igual, N = Par nulo, C = Par correspondiente*

Tabla 9. Intercambio de palabras durante el proceso de alineamiento. Si se inluye como operación el intercambio de dos palabras, se considerarían *alta velocidad* y *velocidad alta* como una misma frase, con lo que *aparato* e *instrumento* serían identificados como un par-vinculado con un LCC de 7.

Como puede observarse a simple vista el segundo par (*aparato, instrumento*) tienen un valor de LCC = *3*, y se observa además que *alta velocidad* y *velocidad alta* se alinearon en las parejas (*alta, velocidad*) y (*velocidad, alta*), que fueron identificadas como pares-correspondientes. Si se dispusiera el algoritmo de intercambio de palabras el resultado debería identificar a las expresiones *alta velocidad* y *velocidad alta* como la misma expresión y elevar de esta manera el valor de LCC de la pareja (*aparato, instrumento*) a 7. Con ello, podría asignársele a la pareja (*aparato, instrumento*) el estado de par-vinculado y por ende identificar las palabras como sinónimos.

LÍNEA DE TRABAJO 8

ALGORITMO BÁSICO

NO SE CONSIDERA LA INVERSIÓN DE DOS PALABRAS CONSECUTIVAS COMO UNA POSIBLE OPERACIÓN DE EDICIÓN.

PROPUESTA

ESTABLECER EL VALOR DEL INCREMENTO DE LCC (1 O 2) ASIGNADO A LA PAREJA IDENTIFICADA COMO INTERCAMBIO. ADEMÁS DE EVALUAR CUANTITATIVAMENTE EL RESULTADO DE AGREGAR LA OPERACIÓN DE INTERCAMBIO. DEBERÁ ESTUDIARSE EL EFECTO QUE LA INCLUSIÓN DE ESTA OPERACIÓN TIENE EN LA APLICACIÓN DEL ALGORITMO DE AGRUPAMIENTO SEMÁNTICO EN DICCIONARIOS DE OTROS IDIOMAS.

2.4 Alineamiento

En el área de alineamiento existen algunas alternativas que deberán ser evaluadas para el mejoramiento del algoritmo.

Uno de los puntos no evaluados en el algoritmo básico es la identificación de rutas alternas, las cuales ofrecen el mismo costo final pero a través de diferentes conjuntos de transformaciones.

LÍNEA DE TRABAJO 9

ALGORITMO BÁSICO

SÓLO SE REALIZA LA EVALUACIÓN DE UNA RUTA POSIBLE. EN CASO DE TENERSE *DIFERENTES ALTERNATIVAS CON EL MISMO COSTO, LA RUTA ELEGIDA SIEMPRE* ELIGE PRIMERO LA SUSTITUCIÓN DE PALABRAS.

PROPUESTA

EVALUAR LA CONVENIENCIA DE BUSCAR TODAS LAS POSIBLES COMBINACIONES COMPARANDO CONTRA EL BENEFICIO QUE PUEDE OBTENERSE CARACTERIZADO ESTE COMO EL NÚMERO DE PARES-VINCULADOS ADICIONALES ENCONTRADOS. ADEMÁS, BUSCAR ALGORITMOS QUE PERMITAN RECORTAR EL NÚMERO DE COMBINACIONES A LOCALIZAR SUPUESTO QUE NO TODAS LAS COMBINACIONES SON ÚTILES PARA EL ALINEAMIENTO SEMÁNTICO.

2.5 Cálculo de LCC

El cálculo del valor de LCC para los pares relacionados es particularmente importante, pues de su valor depende que el par se convierta en un par-vinculado y por tanto forme parte de los agrupamientos semánticos resultantes. En las secciones siguientes se muestran algunas propuestas que tienen como objetivo afinar el cálculo del LCC y calificar de alguna manera el par-vinculado resultante.

2.5.1 Grado de relación en las palabras de un par-vinculado

El algoritmo básico no realiza evaluación alguna respecto a la *"calidad"*[4] en la relación de las palabras que forman una pareja. Si se dispusiera de un calificador de esta relación, podría evaluarse con mayor precisión si un par correspondiente debe convertirse en un par-vinculado, además de establecer qué tan común es que se utilice ese par dentro del *área de conocimiento estudiada.*

---

[4] Utilizamos el término *calidad* para hacer referencia al grado de relación semántica que guardan los pares. Así por ejemplo: *transporte* y triciclo tienen una menor calidad en comparación con *automóvil* y *camión*. Es común utilizar WordNet © para realizar una evaluación cuantitativa a este respecto.

Una posibilidad es utilizar la categoría gramatical a la que pertenece un par, el número de veces que ese par fue utilizado dentro de la base disponible, el valor del LCC (puede usarse el máximo o el promedio), la evaluación del par con un sistema de referencia estándar (por ejemplo, con WordNet o un diccionario de sinónimos), etc.

LÍNEA DE TRABAJO 10

ALGORITMO BÁSICO

EL VALOR DE LCC ES EMPLEADO PARA IDENTIFICAR LOS PARES SEMÁNTICOS, SIN EMBARGO NO SE REALIZA EVALUACIÓN ALGUNA DE LA CALIDAD DEL PAR SEMÁNTICO GENERADO.

PROPUESTA

ESTABLECER UNA FÓRMULA QUE PROPORCIONE UN PESO AL GRADO DE RELACIÓN QUE GUARDAN DOS PALABRAS DE UN PAR-VINCULADO.

2.5.2 Eliminación de palabras funcionales en la cadena

Otra variante posible en el algoritmo es la eliminación de todas aquellas palabras funcionales, a fin de únicamente procesar las palabras que pueden ser consideradas como "relevantes" dentro del texto. Los resultados mediante esta variante deben ser comparados con los obtenidos originalmente a fin de establecer si la disminución del LCC, debida a la eliminación de palabras funcionales, impacta en la identificación de los agrupamientos semánticos.

LÍNEA DE TRABAJO 11

ALGORITMO BÁSICO

NO SE ELIMINAN NI MANUAL NI AUTOMÁTICAMENTE LAS PALABRAS FUNCIONALES.

PROPUESTA

EVALUAR EL IMPACTO DE LA ELIMINACIÓN DE PALABRAS FUNCIONALES DENTRO DE LAS DEFINICIONES, ESTABLECIENDO ALTERNATIVAS PARA LA EVALUACIÓN DEL LCC Y LA PROMOCIÓN DE *PARES CORRESPONDIENTES* A *PARES-VINCULADOS*.

2.5.3 Pares semi-nulos

Frecuentemente, podemos encontrar que, en el momento de alinear dos definiciones los pares-correspondientes no son promovidos a pares-vinculados debido a que existe un par-nulo en al menos uno de los extremos de la alineación. Sin embargo, es posible que este par-nulo pueda ser considerado como un par igual bajo ciertas circunstancias.

Por ejemplo: Supóngase que se desea alinear las siguientes definiciones de dinámica:

A. ***Dinámica***: *parte de la mecánica que trata de las leyes del movimiento en relación con las fuerzas que lo producen* [RAE1992]
B. ***Dinámica***: *parte de la mecánica que estudia las leyes del movimiento en relación con las fuerzas que lo producen* [MMol1996].

Def. 1	dinámica	parte	de	la	mecánica	que	trata	de	las	leyes
Def. 2	dinámica	parte	de	la	mecánica	que	estudia	ε	las	leyes
LCC	0	0	0	0	0	0	6	0	0	0
Tipo	*I*	*I*	*I*	*I*	*I*	*I*	*C*	*N*	*I*	*I*

Def. 1	...	del	movimiento	en	relación	con	las	fuerzas	que	lo	producen
Def. 2	...	del	movimiento	en	relación	con	las	fuerzas	que	lo	producen
LCC	...	0	0	0	0	0	0	0	0	0	0
Tipo	...	*I*	*I*	*I*	*I*	*I*	*I*	*I*	*I*	*I*	*I*

Tipos de pares identificados I = Par igual, N = Par nulo, C = Par correspondiente

Tabla 10. Aplicación de los pares-seminulos. En la definición de *dinámica* el par (*trata*, *estudia*) no cumple con las condiciones de frontera, sin embargo si eliminamos de la primera definición la preposición *de* la condición de frontera se cumple y el par (*trata*, *estudia*) es identificado como un par-correspondiente.

Obsérvese que a la pareja (*trata* y *estudia*) le corresponde un LCC = 6, pero no cumple con las condiciones de frontera. Sin embargo, intuitivamente, estas palabras deberían considerarse como un par-correspondiente, pues la diferencia entre estas dos definiciones sólo está dada por la preposición: *de*. De hecho, si la segunda definición tuviera esta preposición en la posición adecuada, el proceso de la alineación indicaría que a la pareja (*trata* y *estudia*) le correspondería un LCC = 20 y cumpliría además con la condición de frontera, indicando de esta manera que existe una fuerte relación sinonímica entre *cantidad* y *proporción*. Denominamos al par (*de*, ε) *par semi-nulo*, con el objetivo de indicar que "bajo ciertas circunstancias" (como en el caso de las definiciones anteriores) este par puede considerarse un *par-igual*.

LÍNEA DE TRABAJO 12

ALGORITMO BÁSICO

SÓLO SE CONSIDERAN PARES NULOS, SIN EFECTUARSE NINGÚN ANÁLISIS DE ESTOS PARES.

PROPUESTA

IDENTIFICAR AQUELLAS CONDICIONES BAJO LAS CUALES UN PAR NULO PUEDE CONSIDERARSE UN PAR SEMI-NULO, Y LAS CONDICIONES BAJO LAS CUALES UN PAR SEMI-NULO PUEDE CONSIDERARSE UN PAR IGUAL. ADEMÁS, DEBERÁN EVALUARSE LOS RESULTADOS OBTENIDOS Y COMPARARLOS CUALITATIVA Y CUANTITATIVAMENTE CON EL ALGORITMO BÁSICO.

2.6 Lista de palabras irrelevantes

En la etapa de determinación de pares-vinculados se hace uso de una lista de palabras irrelevantes (stop list) a fin de rechazarlos como pares correspondientes, ya que semánticamente no tienen sentido alguno. Debido a que el algoritmo básico se desarrolló para el caso de una terminología en inglés, la lista de palabras funcionales también era en este idioma. Además de las palabras funcionales, es posible añadir, según el área temática, palabras irrelevantes a la lista.

INTERFAZ HOMBRE-COMPUTADORA 2

ALGORITMO BÁSICO

LA LISTA DE PALABRAS IRRELEVANTES SE ENCUENTRA ALMACENADA EN UN ARCHIVO.

PROPUESTA

DEBE PERMITIRSE QUE EL USUARIO TENGA LA POSIBILIDAD DE CONSIDERAR O NO LA PERTINENCIA DE LA APLICACIÓN DE CADA UNA DE LAS PALABRAS IRRELEVANTES DISPONIBLES. INCLUIR LA POSIBILIDAD DE EMPLEAR UNA BASE DE DATOS DE PALABRAS IRRELEVANTES ORDENADAS POR IDIOMA, CAMPO DEL CONOCIMIENTO Y ÁREA TEMÁTICA, DISPONIBLE PARA EMPLEARSE EN CUALQUIERA DE LOS DICCIONARIOS. ADEMÁS, DE MANERA AUTOMÁTICA, DEBE PROPONERSE UNA LISTA CON BASE EN EL IDIOMA DEL DICCIONARIO ANALIZADO.

2.7 Lematización

Actualmente se emplea el algoritmo de Porter para lematizar; sin embargo, éste presenta deficiencias en sus resultados y es necesario afinarlo para el caso de idiomas tales como el español, donde la representación de los acentos, tildes y diéresis implica además diferencia de significados (tabla 8, sección 2.2.1)

LÍNEA DE TRABAJO 13

ALGORITMO BÁSICO

SE EMPLEA EL ALGORITMO DE LEMATIZACIÓN DE PORTER PARA EL INGLÉS.

PROPUESTA

DESARROLLAR UN LEMATIZADOR CUYA CONCEPCIÓN ORIGINAL SEA PARA EL IDIOMA ESPAÑOL. TAMBIÉN DEBE SER CONSIDERADA LA POSIBILIDAD DE EMPLEAR LEMATIZADORES YA EXISTENTES Y SU INCORPORACIÓN AL ALGORITMO DE ALINEAMIENTO SEMÁNTICO.

2.8 Sustitución de pares-vinculados

Con base en el agrupamiento semántico al que pertenece un par-vinculado es posible elegir una palabra representativa del agrupamiento (denominada centroide) y utilizarla durante el proceso de sustitución en cada par de definiciones donde se localice un par-vinculado asociado al agrupamiento semántico.

LÍNEA DE TRABAJO 14

ALGORITMO BÁSICO

NO SE ESTABLECE NINGUNA CONSIDERACIÓN RESPECTO A UN CENTROIDE.

PROPUESTA

ESTABLECER LOS CRITERIOS PARA LA IDENTIFICACIÓN DE LOS CENTROIDES EN LOS AGRUPAMIENTOS SEMÁNTICOS Y EVALUAR LOS RESULTADOS OBTENIDOS AL SUSTITUIR ÉSTOS EN LAS DEFINICIONES. TAMBIÉN DEBE CONSIDERARSE LA POSIBILIDAD DE UTILIZAR EN LUGAR DE UN CENTROIDE UN SÍMBOLO QUE IDENTIFIQUE AL AGRUPAMIENTO SEMÁNTICO ASÍ COMO ANALIZAR LOS RESULTADOS OBTENIDOS CON RESPECTO A LOS ORIGINALES Y LOS GENERADOS CON UN CENTROIDE.

2.9 Generación de agrupamientos semánticos

Tal como está formulado actualmente el algoritmo, la generación de los agrupamientos semánticos no permite el traslape de agrupamientos.

Actualmente, cuando un agrupamiento contiene al menos una palabra en común con otro, estos agrupamientos se reordenan generando un sólo agrupamiento, resultado de la fusión de los dos agrupamientos anteriores. Esta fusión no siempre es adecuada, por ejemplo considérense los agrupamientos formados por: {*posee*, *tiene*} y {*tiene*, *contiene*} como ambos contienen la palabra *tiene* se fusionan dando lugar al agrupamiento semántico {*posee*, *tiene*, *contiene*} lo cual es correcto. Sin embargo los agrupamientos {*grupo*, *sistema*} y {*marco*, *sistema*} dan origen al agrupamiento {*grupo*, *marco*, *sistema*} lo que produce un agrupamiento semántico cuestionable. Lo ideal, en este caso es mantener los agrupamientos semánticos separados, con una o varias palabras en común pero sin forzar su unión.

LÍNEA DE TRABAJO 15

ALGORITMO BÁSICO

EL ALGORITMO BÁSICO FORMA AGRUPAMIENTOS SEMÁNTICOS NO TRASLAPADOS.

PROPUESTA

BUSCAR Y REVISAR TÉCNICAS QUE PERMITAN GENERAR AGRUPAMIENTOS QUE PRESENTEN TRASLAPES. ESTO REDUNDARÁ EN UNA MEJOR REPRESENTACIÓN DEL CONCEPTO ASOCIADO A CADA UNO DE LOS AGRUPAMIENTOS SEMÁNTICOS. EN CUALQUIER CASO ES NECESARIO EFECTUAR LA EVALUACIÓN CORRESPONDIENTE.

2.10 Procedimiento iterativo

Como ya se mencionó anteriormente, el algoritmo de agrupamiento semántico tiene naturaleza cíclica. Originalmente parte de un conjunto de definiciones y un conjunto vacío de agrupamientos semánticos; conforme evoluciona se generan los agrupamientos que ciclo a ciclo van refinándose y mezclándose hasta que no se generen nuevos pares-vinculados. Una segunda opción de partida es comenzar con un conjunto de agrupamientos reconocidos para el idioma y como base para un campo de conocimiento y un área temática en particular. Esto puede llevar a una mejor identificación de los agrupamientos subyacentes en las definiciones a analizar.

INTERFAZ HOMBRE-COMPUTADORA 3

ALGORITMO BÁSICO

LOS AGRUPAMIENTOS SEMÁNTICOS GENERADOS PARTEN DE UN CONJUNTO VACÍO DE PALABRAS.

PROPUESTA

PERMITIR LA ELECCIÓN DEL ESTILO DE ITERACIÓN, ES DECIR, PARTIR DE UN CONJUNTO VACÍO DE AGRUPAMIENTOS SEMÁNTICOS O ELEGIR LOS AGRUPAMIENTOS INICIALES DISPONIBLES.

2.11 Cambios en la estructura del algoritmo

La estructura del algoritmo puede ser modificada a fin de explorar nuevas alternativas para la generación de los agrupamientos semánticos; por ejemplo, una posibilidad es generar primero todos los pares-vinculados e iterar hasta que ya no se tengan más pares-vinculados. Una vez que se dispongan de todos los pares, establecer los agrupamientos semánticos. Otra posibilidad es la de incorporar los elementos de los agrupamientos semánticos obtenidos para la transformación de las definiciones y para la consecuente generación de nuevos agrupamientos semánticos.

LÍNEA DE TRABAJO 16

ALGORITMO BÁSICO

CON CADA CICLO SE AGRUPAN LOS PARES-VINCULADOS FORMANDO NUEVOS AGRUPAMIENTOS SEMÁNTICOS. ADEMÁS, ESTOS AGRUPAMIENTOS NO SON CONSIDERADOS PARA LA PRODUCCIÓN DE NUEVOS PARES-VINCULADOS.

PROPUESTA

EVALUAR LA POSIBILIDAD DE GENERAR PRIMERO TODOS LOS PARES-VINCULADOS POSIBLES Y DESPUÉS ESTABLECER LOS AGRUPAMIENTOS SEMÁNTICOS. OTRA ALTERNATIVA ES INCLUIR LOS AGRUPAMIENTOS SEMÁNTICOS EN LA GENERACIÓN DE PARES-VINCULADOS.

2.12 Recapitulación

En este capítulo se revisó el algoritmo básico y se identificaron 19 líneas de trabajo posibles a explorar, separadas en 16 alternativas y 3 modificaciones a la interfaz hombre-computadora. Estas líneas son el resultado de un profundo análisis que se realizó, junto con los autores del algoritmo básico. Cada una de las líneas en si constituyen tema suficiente para futuros trabajos a realizarse.

En la elección y desarrollo de un conjunto de líneas de trabajo no es posible, al menos en este momento, presentar argumentos cuantitativos más que intuitivos. De acuerdo con la experiencia de los autores del algoritmo básico de agrupamiento semántico se eligieron en para esta tesis cuatro líneas: pares semi-iguales y semi-nulos (Línea de trabajo 12), intercambio de palabras (Línea de trabajo 8), modificación de costos (Línea de trabajo 6), y rutas múltiples de alineamiento (Línea de trabajo 9), las cuales se describen a detalle en el siguiente capítulo.

3 Algoritmo de alineamiento semántico flexibilizado

Una vez que en el capítulo anterior se presentaron diversas líneas de trabajo alternativas con las cuales es posible mejorar el algoritmo básico de agrupamiento semántico, en este capítulo se desarrollan algunas de ellas y se presentan los algoritmos resultantes de este trabajo. El planteamiento teórico se presenta en este capítulo y en el siguiente capítulo se evalúan los resultados de manera formal.

Con base en la experiencia de los autores y a falta de elementos cuantitativos que normen la elección de las líneas de trabajo, se escogieron cuatro variantes a desarrollar. Se espera que estas alternativas mejoren sustancialmente el número de pares semánticos identificados. Las líneas son: modificación de costos (Línea de trabajo 6), intercambio de palabras (Línea de trabajo 8), rutas múltiples de alineamiento (Línea de trabajo 9) y pares semi-nulos (Línea de trabajo 12). De estas cuatro líneas se derivan seis propuestas específicas:

- Inversión de palabras, que permite incluir la posibilidad de analizar el intercambio de palabras como una operación válida en la determinación del alineamiento, y contempla la posibilidad del intercambio de dos palabras contiguas (sección 3.1.1) o de dos palabras separadas por una conjunción (sección 3.1.2);

- Rutas múltiples de alineamiento, que evalúa la conveniencia de utilizar rutas múltiples de costo mínimo para identificar pares-vinculados (sección 3.2.1).

- En paralelo al análisis de la línea de rutas múltiples, se desarrolla una alternativa de la línea de modificación de costos, que evalúa la posibilidad de utilizar diferentes costos para cada una de las operaciones de edición (sección 3.2.4);

- Pares semi-nulos (sección 3.3), que considera distinto peso a dichos pares, con el fin de considerarlos dentro de la evaluación del LCC, y de esta línea se desprende también el análisis de pares semi-iguales (sección 3.4).

3.1 Inversión de dos palabras

Como se mencionó en la sección 2.3.3, considerar la inversión de palabras como una operación válida puede redundar en un mejoramiento de la identificación de pares semánticos.

3.1.1 Inversión de dos palabras consecutivas

Se han propuesto diferentes algoritmos para identificar cadenas con intercambio, entre ellos Lowrence & Wagner [LoW1975], Amir, Auman, Landau, Lewenstein & Lewenstein [AAL1997], Lee, Kim, Park & Cho [LKP1997], han atacado el problema. Debido a la flexibilidad del método propuesto por Lowrence y Wagner para recordar los costos y las operaciones efectuadas durante la determinación del alineamiento, así como por

constituir la base para los posteriores algoritmos, a continuación se describe este método, para después abordar su aplicación al problema de alineamiento semántico.

Lowrance y Wagner desarrollaron un algoritmo de complejidad $O(mn)$ para extender el algoritmo de Wagner y Fisher (sección 1.1.1), en el que se contempla el caso de transposición. El algoritmo propuesto utiliza el método de programación dinámica para el alineamiento de dos cadenas, modificando la manera en que se calculan los costos de las operaciones dentro del método de programación dinámica, a fin de considerar la inversión de dos palabras consecutivas.

La idea básica del algoritmo es calcular el costo de intercambiar dos palabras como la suma del costo que corresponde a llegar desde el inicio de la cadena hasta antes de las palabras en cuestión más el costo de la operación de intercambio.

Al costo por intercambio de dos símbolos consecutivos $(a_i b_{j-1},\ a_{i-1} b_j)$ lo representamos como $w_s(a_i b_{j-1},\ a_{i-1} b_j)$, y se calcula como:

$$w_S(a_i b_{j-1}, a_{i-1} b_j) = \begin{cases} 1 & si\ (a_i = b_{j-1}),\quad (a_{i-1} = b_j) \quad y \quad (a_i \neq b_j) \\ \infty & c.o.c. \end{cases}$$

De esta manera la fórmula propuesta por Lowrence & Wagner es:

$$C_{i,j} = C_{i-2,j-2} + w_s(a_i b_{j-1},\ a_{i-1} b_j)$$

La ecuación de transposición debe ser agregada en la minimización de la relación de recurrencia $C_{i,j}$ de manera que:

$$C_{i,j} = \min \begin{cases} C_{i-1,j} + w(a_i, \varepsilon) \\ C_{i,j-1} + w(\varepsilon, b_j) \\ C_{i-1,j-1} + w(a_i, b_j) \\ C_{i-2,j-2} + w_S(a_i b_{j-1}, a_{i-1} b_j) \end{cases}$$

Debido a que ahora el costo $C_{i-2,j-2}$ está involucrado en los cálculos, es necesario hacer consideraciones especiales para cuando $i < 2$ o $j < 2$, pues en estos casos los costos involucrados en la ecuación de recurrencia son $C_{-1,j}$, $C_{-2,j}$, $C_{i,-1}$ o $C_{i,-2}$ lo cual no tiene sentido. Valores de frontera adicionales se requieren a fin de evitar estas inconsistencias, cuando $i < 2$ o $j < 2$: se debe asignar a $C_{i-2,j} = \infty$ y $C_{i,j-2} = \infty$, con lo cual se previene que la operación de transposición se elija para el primer símbolo, de modo que se evitan transposiciones que pueden involucrar símbolos previos al inicio de la secuencia de palabras. Con fines computacionales, para representar infinito es suficiente con elegir w_s mayor que la longitud máxima de las cadenas a alinear.

3.1.2 Inversión conjuntiva

Si bien es cierto que en la literatura se han considerado intercambios de palabras, ya sea consecutivas o no consecutivas, no se ha reportado el intercambio de palabras que considere dos palabras relacionadas a través de una conjunción[5] (y). A continuación proponemos, con base en el propuesto por Lowrence & Wagner, un algoritmo que considera como una operación válida, para el proceso de alineamiento, el intercambio de palabras conectadas por una conjunción.

Definiremos una *inversión conjuntiva* como la operación resultante de considerar el intercambio de dos palabras conectadas por una conjunción (y). Si dos palabras (a,b) se encuentran separadas por una conjunción y forman parte de una pareja de definiciones, las cuales se desean alinear, entonces estas palabras cumplen con la condición de que $a_i = b_{j-2} = a$ y $a_{i-2} = b_j = b$ y $a_{i-1} = b_{j-1} \in \{y\}$. Por tanto, el costo del intercambio de estas dos palabras, representado como w_{SC}, puede calcularse como:

$$w_{SC}(a_i b_{j-2}, a_{i-2} b_j) = \begin{cases} 1 & si\ (a_i = b_{j-2}),\ (a_{i-2} = b_j) \quad y \quad (a_i \neq b_j) \\ \infty & c.o.c. \end{cases}$$

Además, a fin de garantizar que se está considerando una conjunción debemos incluir el término:

$$w_C(a_i, b_j) = \begin{cases} w_{SC} & si\ a_{i-1} = b_{j-1} \in \{y\} \\ \infty & c.o.c \end{cases}$$

Incorporando esta modificación a la fórmula propuesta por Lowrence & Wagner obtenemos:

$$C_{i,j} = C_{i-3,j-3} + w_C(a_i, b_j)$$

La ecuación de recurrencia para calcular el costo queda:

$$C_{i,j} = \min \begin{cases} C_{i-1,j} + w(a_i, \varepsilon) \\ C_{i,j-1} + w(\varepsilon, b_j) \\ C_{i-1,j-1} + w(a_i, b_j) \\ C_{i-3,j-3} + w_C(a_i, b_j) \end{cases}$$

Como en el caso del intercambio de palabras consecutivas, ahora el término $C_{i-3,j-3}$ está involucrado en los cálculos, por lo que valores de frontera adicionales se requieren

---

[5] La conjunción analizada es la copulativa (y) debido básicamente a que la conjunción disyuntiva implica la posibilidad de dividir la definición en dos (vease la sección 2.2.3)

cuando $i < 3$ o $j < 3$; al asignar a $C_{i-3,j} = \infty$ y $C_{i,j-3} = \infty$ se previenen transposiciones que pueden involucrar símbolos previos al inicio de la secuencia de palabras.

3.1.3 Ecuación de recurrencia

Integrando las ecuaciones de recurrencia para el caso de intercambio de dos palabras consecutivas y para el caso de intercambio conjuntivo, se establece una nueva ecuación, que a continuación se presenta:

$$C_{i,j} = \min \begin{cases} C_{i-1,j} + w(a_i, \varepsilon) \\ C_{i,j-1} + w(\varepsilon, b_j) \\ C_{i-1,j-1} + w(a_i, b_j) \\ C_{i-2,j-2} + w_S(a_i b_{j-1}, a_{i-1} b_j) \\ C_{i-3,j-3} + w_C(a_i, b_j) \end{cases}$$

3.2 Rutas múltiples

En la sección 2.4 se formuló como una Línea de trabajo la búsqueda de rutas múltiples con costo mínimo para el alineamiento de dos definiciones. Sin embargo, el alineamiento de dos definiciones que minimiza la distancia de edición, por lo general, no es único. Por ejemplo, considérense las siguientes dos definiciones del término *ammeter*:

A. ***ammeter***: An instrument for estimating the force of electric currents [CED1994]
B. ***ammeter***: An instrument for measuring an electric current in amperes [OED1994]

Si aplicamos el algoritmo de alineamiento, obtenemos la siguiente tabla de costos (aplicando previamente un proceso de lematización y considerando únicamente las operaciones de sustitución, inserción y borrado).

	ϕ	*ammeter*	*An*	*instrument*	*for*	*estimating*	*the*	*force*	*of*	*electric*	*currents*
ϕ	0	1	2	3	4	5	6	7	8	9	10
ammeter	1	0	1	2	3	4	5	6	7	8	9
An	2	1	0	1	2	3	4	5	6	7	8
instrument	3	2	1	0	1	2	3	4	5	6	7
for	4	3	2	1	0	1	2	3	4	5	6
measuring	5	4	3	2	1	1	2	3	4	5	6
an	6	5	4	3	2	2	2	3	4	5	6
electric	7	6	5	4	3	3	3	3	4	4	5
current	8	7	6	5	4	4	4	4	4	5	4
in	9	8	7	6	5	5	5	5	5	5	5
amperes	10	9	8	7	6	6	6	6	6	6	6

Tabla 11. Matriz de costos para la definición del término *ammeter* de acuerdo con el algoritmo de Wagner y Fisher.

El alineamiento que se obtiene a partir del algoritmo de Wagner & Fisher es:

Def. 1	ammeter	an	instrument	for	estimating	the	force	of	electric	currents
Def. 2	ammeter	an	instrument	for	measuring	an	electric	current	in	amperes
Costos	0	0	0	0	1	2	3	4	5	6

Tabla 11. Alineamiento del término *ammeter* según el algoritmo de Wagner y Fisher

Sin embargo, los siguientes también son alineamientos cuyo costo de edición es mínimo e igual a 6.

Def. 1	ammeter	an	instrument	for	estimating	the	force	of	electric	currents	ε	ε
Def. 2	ammeter	an	instrument	for	ε	ε	measuring	an	electric	current	in	amperes
Costos	0	0	0	0	1	2	3	4	4	4	5	6

Def. 1	ammeter	an	instrument	for	estimating	the	force	of	electric	currents	ε	ε
Def. 2	ammeter	an	instrument	for	ε	measuring	ε	an	electric	current	in	amperes
Costos	0	0	0	0	1	2	3	4	4	4	5	6

Def. 1	ammeter	an	instrument	for	estimating	the	force	of	electric	currents	ε	ε
Def. 2	ammeter	an	instrument	for	measuring	ε	ε	an	electric	current	in	amperes
Costos	0	0	0	0	1	2	3	4	4	4	5	6

Def. 1	ammeter	an	instrument	for	estimating	the	force	of	electric	currents	ε	ε
Def. 2	ammeter	an	instrument	for	ε	measuring	an	ε	electric	current	in	amperes
Costos	0	0	0	0	1	2	3	4	4	4	5	6

Def. 1	ammeter	an	instrument	for	estimating	the	force	of	electric	currents	ε	ε
Def. 2	ammeter	an	instrument	for	measuring	ε	an	ε	electric	current	in	amperes
Costos	0	0	0	0	1	2	3	4	4	4	5	6

Def. 1	ammeter	an	instrument	for	estimating	the	force	of	electric	currents	ε	ε
Def. 2	ammeter	an	instrument	for	measuring	an	ε	ε	electric	current	in	amperes
Costos	0	0	0	0	1	2	3	4	4	4	5	6

Tabla 12. Seis alineamientos de costo mínimo alternativos del término *ammeter*.

Con el fin de incluir en el alineamiento semántico de dos definiciones las posibles alternativas de costo mínimo, a continuación se presenta el algoritmo para la determinación del conjunto de todos los alineamientos de dos definiciones que cumplen con la condición de minimizar la distancia de edición. Este algoritmo fue encontrado en la página Web de Christian Charras y Thierry Lecroq del LIR (Laboratoire d'Informatique de

Rouen) y del ABISS (Atelier Biologie Informatique Statistique Socio-linguistique)[6] y se adecuó a las necesidades de nuestro problema.

En la discusión que sigue sólo se consideran como operaciones posibles la sustitución, la inserción y el borrado, aunque es posible generalizar a fin de considerar otras operaciones (intercambio e intercambio conjuntivo, por ejemplo).

3.2.1 Algoritmo de determinación de rutas múltiples

La solución al problema de determinar el conjunto de alineamientos de dos definiciones que minimiza la distancia de edición puede ser establecida a través de un algoritmo recursivo. Este algoritmo debe partir de la posición final (m,n) y evolucionar hasta encontrar la posición inicial $(0,0)$ de la matriz de costos. El avance a través de la matriz se realiza examinando cada posible operación que pudo haberse aplicado en un paso anterior (inserción, borrado y sustitución) y que llevó a la posición actualmente examinada. Si existe más de una operación que pudo llevar al resultado actual entonces el algoritmo se bifurca y efectúa una llamada recursiva por cada uno de los posibles trayectos. Si la posición a la que se llega es la inicial $(0,0)$ entonces la llamada termina. Así pues, el algoritmo es:

Algoritmo: Generación de todos los alineamientos posibles

Entrada: *A, B :* *Secuencia de palabras a alinear*
C *Matriz de costos calculada según Wagner & Fisher*
r, c *celda a analizar dentro de la matriz de costos (r = renglón y c = columna)*
Z : *Pares de palabras alineadas hasta esta llamada*

Salida: *Z :* *Cada vez que se llama a* ***reporta alineamiento*** *(Z) se presenta un alineamiento posible de las cadenas A y B tal que su costo es mínimo*

```
Alinear(A,r,B,c,Z,C)
if r ≠ 0
        if c ≠ 0
                if C_r,c = C_r-1,c-1 − w(a_r,b_c)
                        Alinear(A,r-1,B,c-1,(a_r,b_c)·Z,C)
                if C_r,c = C_r-1,c +w(a_r,ε)
                        Alinear(A,r-1,B,c,(a_r, ε)·Z,C)
                if C_r,c = C_r,c-1 +w(ε,b_c)
                        Alinear(A,r,B,c-1,( ε, b_c)·Z,C)
        else
                Alinear(A,r-1,B,0,(a_r, ε)·Z,C)
else
        if c≠ 0
                Alinear(A,0,B,c-1,( ε,b_c)·Z,C)
        else
                reporta_alineamiento(Z)
```

---

[6] http://www-igm.univ-mlv.fr/~lecroq/seqcomp/index.html

Este es el algoritmo utilizado para generar los siete alineamientos de *ammeter* presentados con anterioridad.

3.2.2 El problema de alineamiento en términos de un autómata finito

Ukkonen [Ukk1985] formuló el problema de alineamiento en términos de una gráfica dirigida con vértices etiquetados, denominada ***gráfica de edición*** G_{AB} y permite comparar secuencias de palabras. Los vértices de G_{AB} son los pares (i,j). Los vértices se organizan en un arreglo matricial $(n+1)x(m+1)$. El vértice inicial lo representamos por Θ y corresponde a la posición $(0,0)$ dentro de la malla. El vértice Φ representa el vértice final o destino y corresponde a la posición (n,m). Las reglas para determinar los vértices, y sólo estos, que pertenecen a $G_{A,B}$ son:

a) si $i \in [1,m]$ y $j \in [0,n]$, entonces existe un ***vértice de borrado*** $(i-1,j) \rightarrow (i,j)$ etiquetado como $\begin{bmatrix} a_i \\ \varepsilon \end{bmatrix}$

b) si $i \in [0,m]$ y $j \in [1,n]$, entonces existe un ***vértice de inserción*** $(i,j-1) \rightarrow (i,j)$ etiquetado como $\begin{bmatrix} \varepsilon \\ b_j \end{bmatrix}$

c) si $i \in [1,m]$ y $j \in [1,n]$, entonces existe un ***vértice de substitución***[7] $(i-1,j-1) \rightarrow (i,j)$ etiquetado como $\begin{bmatrix} a_i \\ b_j \end{bmatrix}$

La figura siguiente presenta G_{AB} para A = "*una bata blanca*" y B = "*una cabra blanca saltó*"

Esta perspectiva presenta entonces a un alineamiento como una ruta en G_{AB} en donde se han concatenado las etiquetas de los vértices que constituyen dicha ruta. Así, un alineamiento entre A_i y B_j puede ser:

a) un alineamiento entre A_{i-1} y B_j concatenado con $\begin{bmatrix} a_i \\ \varepsilon \end{bmatrix}$

b) un alineamiento entre A_i y B_{j-1} concatenado con $\begin{bmatrix} \varepsilon \\ b_j \end{bmatrix}$

c) un alineamiento entre A_{i-1} y B_{j-1} concatenado con $\begin{bmatrix} a_i \\ b_j \end{bmatrix}$

---

[7] Se asume en esta sección que sustitución contempla dos posibilidades: sustituir a_i *por* b_j *si* $a_i \neq b_j$ y no efectuar operación alguna en caso contrario

TESIS CON
FALLA DE ORIGEN

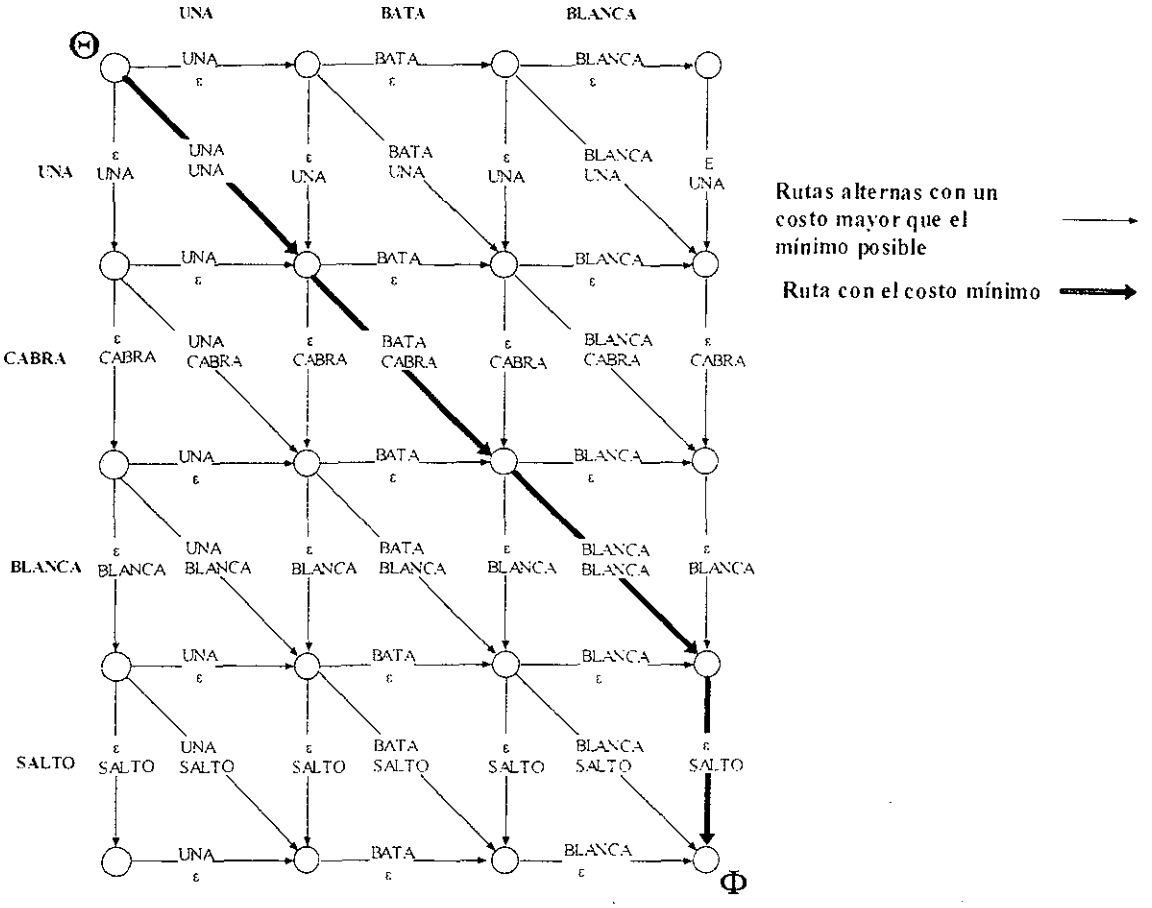


Figura 4. Gráfica de edición de las frases: *"Una bata blanca"* y *"una cabra blanca saltó"*

De lo anterior es posible afirmar que G_{AB} constituye un autómata finito que acepta el conjunto de todos los alineamientos posibles entre A y B, bajo las operaciones de edición antes definidas.

Bajo este punto de vista, el algoritmo de la distancia de edición tiene como objetivo establecer la ruta más corta entre el nodo inicial y el final, asignando costos a cada uno de los vértices y los cuales corresponden a los de la operación asociada al vértice, es decir $w(a_i,\varepsilon)$, $w(\varepsilon,b_j)$, $w(a_i,b_j)$ para las operaciones de borrado, inserción y sustitución, respectivamente. En nuestro ejemplo, la ruta más corta se ha marcado con un trazo más grueso.

3.2.3 La matriz de costos como una gráfica dirigida

Si en la matriz de costos consideramos cada elemento $C_{i,j}$ como un nodo de una malla dirigida, los vértices de cada nodo indican la operación -inserción, borrado o sustitución- elegida con base en la fórmula:

$$C_{i,j} = \min \begin{cases} C_{i-1,j} + w(a_i, \varepsilon) \\ C_{i,j-1} + w(\varepsilon, b_j) \\ C_{i-1,j-1} + w(a_i, b_j) \end{cases}$$

donde la relación entre los nodos $C_{i,j}$, $C_{i-1,j}$, $C_{i,j-1}$ y $C_{i-1,j-1}$ puede visualizarse como:

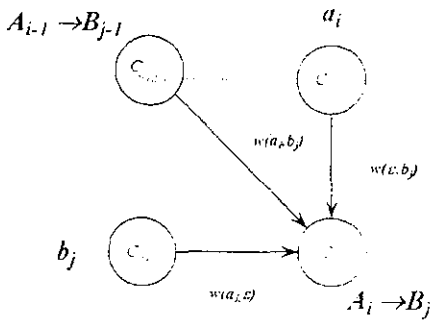


Figura 5. Operaciones de inserción, borrado y sustitución esquematizada como vértices de una malla dirigida.

La gráfica resultante de esta representación es una malla orientada generada durante la búsqueda del alineamiento de costo mínimo, donde los vértices corresponden a la operación u operaciones elegidas para llegar al nodo $C_{i,j}$ y se etiquetan de acuerdo con la suma de los costos de las operaciones elegidas para llegar desde el nodo inicial $C_{0,0}$ hasta el nodo $C_{i,j}$. Para el ejemplo presentado, la gráfica correspondiente es:

	ε	UNA	BATA	BLANCA
ε	0	1	2	3
UNA	1	0	1	2
CABRA	2	1	1	2
BLANCA	3	2	2	1
SALTO	4	3	3	2

Figura 6. Malla orientada que muestra las posibles rutas de alineamiento de costo mínimo de las frases : *"Una bata blanca"* y *"una cabra blanca saltó"*

El ejemplo inicialmente analizado es realmente sencillo, y de hecho sólo existe una ruta que genera un alineamiento de costo mínimo. Sin embargo, al alinear las definiciones de *ammeter*, la malla resultante es más compleja y se muestra en la gráfica siguiente.

TESIS CON
FALLA DE ORIGEN

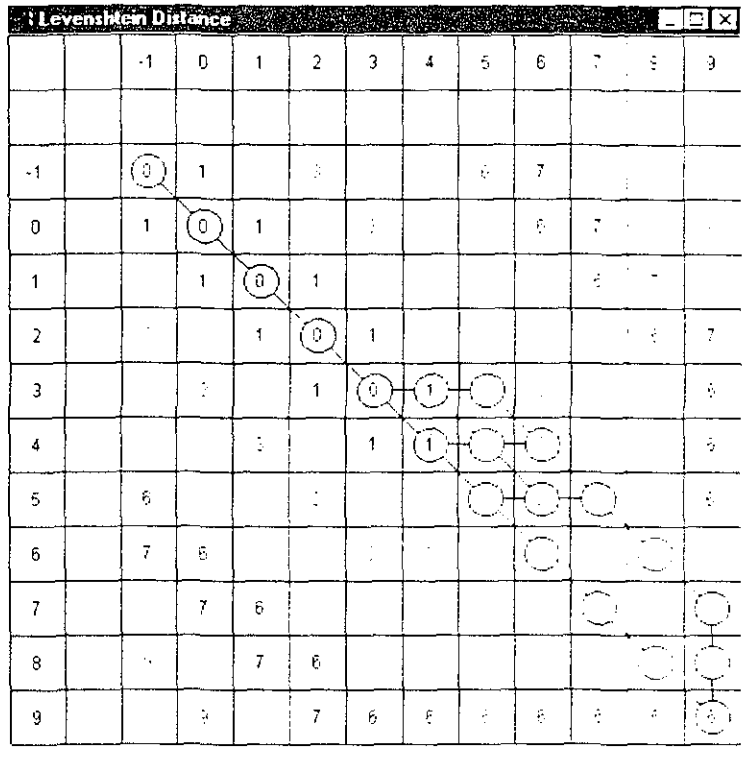


Figura 7. Malla orientada que muestra las posibles rutas de alineamiento de costo mínimo del término *ammeter*

Si la diferencia entre las longitudes de las cadenas analizadas es grande, la cantidad de alineamientos posibles crece rápidamente. Por ejemplo, considérense las siguientes dos definiciones del término *abney level:*

A. ***Abney_level***: A small hand instrument used by surveyors for measuring slopes and angles above the horizon [OED1994] (16 palabras)
B. ***Abney_level***: A surveying instrument consisting of a spirit level and a sighting tube used to measure the angle of inclination of a line from the observer to another point (29 palabras) [CED1994]

El algoritmo de alineamiento semántico genera la siguiente tabla de costos para la forma lematizada de las definiciones; en este caso, la primera columna y el primer renglón representan, a través de un número índice (comenzando en 0), cada una de las palabras de las definiciones, y sólo se han representado en una malla las rutas que generan un alineamiento de costo mínimo:

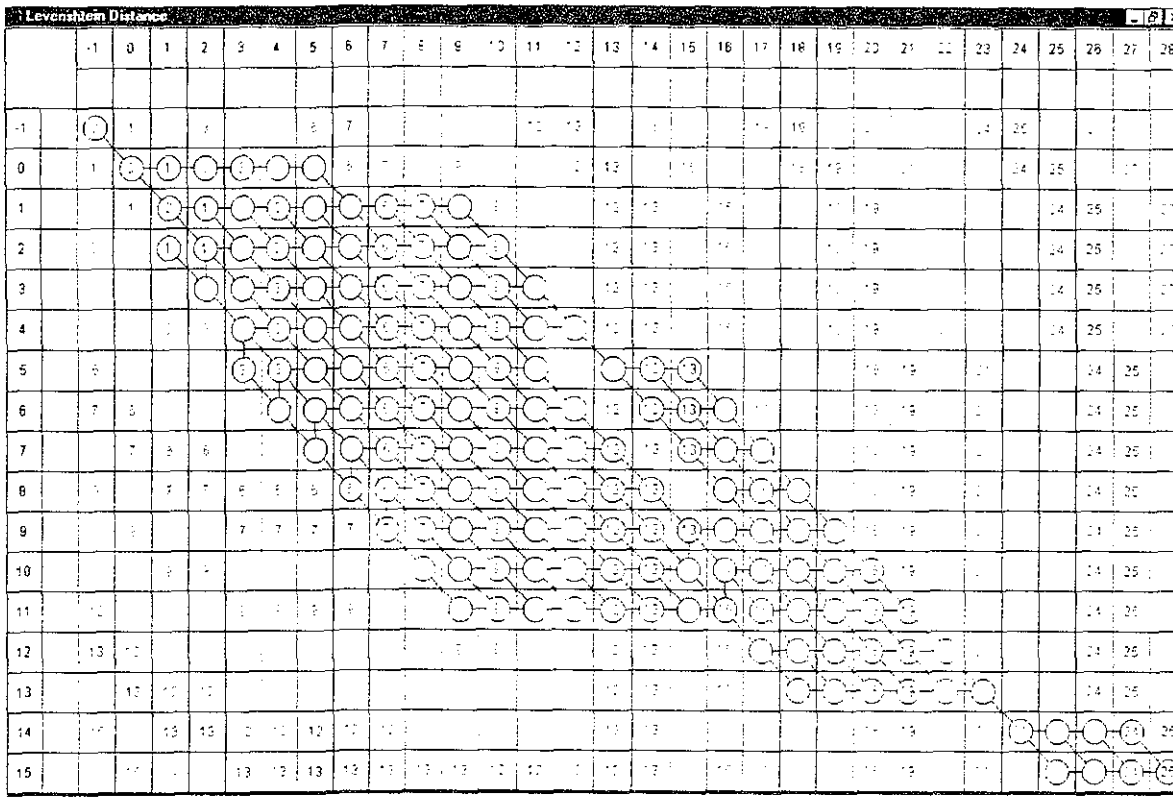


Figura 8. Malla orientada que muestra las posibles rutas de alineamiento de costo mínimo del término *abney level*

El número de rutas posibles que ofrecen un costo mínimo para alinear las dos definiciones es de **970,364**: este valor fue calculado utilizando el algoritmo para la generación de todos los posibles alineamientos, presentado al principio de esta sección, pero únicamente contando los posibles alineamientos.

Se aprecia que las rutas posibles en este caso han crecido drásticamente, con lo que es fundamental establecer un algoritmo que permita dirigir la búsqueda a fin de obtener un número razonable de opciones. En esta tesis sólo se ha restringido la generación de los alineamientos a los primeros k alineamientos encontrados, con lo que el algoritmo empleado es:

TESIS CON
FALLA DE ORIGEN

Algoritmo: Generación de los primeros K_{max} alineamientos posibles

K_{max} : *Valor global y representa el máximo número de alineamientos a generar (permite limitar la búsqueda)*

k: *Valor global y representa el número de alineamientos generados hasta esta corrida. Su valor inicial antes de efectuar la primera llamada recursiva es 0*

Entrada: *A, B :* *Secuencia de palabras a alinear*

C : *Matriz de costos calculada según Wagner & Fisher*

r, c *celda a analizar dentro de la matriz de costos (r = renglón y c = columna)*

Z : *Pares de palabras alineadas hasta esta llamada*

Salida: *Z :* *Cada vez que se llama a* ***reporta alineamiento****(Z) se presenta un alineamiento posible de las cadenas A y B tal que su costo es mínimo*

Alinear*(A,r,B,c,Z,C)*
if $k<K_{max}$
  if $r \neq 0$
    if $c \neq 0$
      if $C_{r,c} = C_{r-1,c-1} + w(a_r, b_c)$
        Alinear*(A,r-1,B,c-1,*(a_r, b_c)*·Z,C)*
      if $C_{r,c} = C_{r-1,c} + w(a_r, \varepsilon)$
        Alinear*(A,r-1,B,c,*(a_r, ε)*·Z,C)*
      if $C_{r,c} = C_{r,c-1} + w(\varepsilon, b_c)$
        Alinear*(A,r,B,c-1,*(ε, b_c)*·Z,C)*
    else
      Alinear*(A,r-1,B,0,*(a_r, ε)*·Z,C)*
  else
    if $c \neq 0$
      Alinear*(A,0,B,c-1,*(ε, b_c)*·Z,C)*
    else
      reporta_alineamiento*(Z)*
      $k = k + 1$

TESIS CON
FALLA DE ORIGEN

3.2.4 Una alternativa de elección de las rutas múltiples

En los párrafos anteriores se ha considerado la aplicación del algoritmo de Levenshtein para la generación de los alineamientos semánticos. Al considerar iguales a 1 los costos de las tres transformaciones fundamentales, este algoritmo no favorece la aplicación de alguna de las tres operaciones con el fin de dirigir la búsqueda del alineamiento óptimo.

Una alternativa para favorecer la alineación de palabras iguales durante la generación de los alineamientos es modificar los costos del algoritmo de Levenshtein a fin de "premiar" la alineación de palabras iguales y "castigar" la aplicación de alguna otra transformación.

Considérese nuevamente la definición *abney level* presentada con anterioridad. Si empleamos el algoritmo de Levenshtein para alinear estas definiciones obtenemos la siguiente matriz de costos:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
2	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
3	2	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
4	3	2	2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
5	4	3	3	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
6	5	4	4	3	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
7	6	5	5	4	4	4	5	6	7	8	9	10	11	12	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
8	7	6	6	5	5	5	5	6	7	8	9	10	11	12	13	13	14	15	16	17	18	19	20	21	22	23	24	25	26
9	8	7	7	6	6	6	6	6	7	8	9	10	11	12	13	14	14	15	16	17	18	19	20	21	22	23	24	25	26
10	9	8	8	7	7	7	7	7	7	8	9	10	11	12	13	13	14	15	16	17	18	19	20	21	22	23	24	25	26
11	10	9	9	8	8	8	8	8	8	8	9	10	11	12	13	14	14	15	16	17	18	19	20	21	22	23	24	25	26
12	11	10	10	9	9	9	9	9	9	8	9	10	11	12	13	14	15	15	16	17	18	19	20	21	22	23	24	25	26
13	12	11	11	10	10	10	10	10	10	9	9	10	11	12	13	14	15	15	16	17	18	19	20	21	22	23	24	25	26
14	13	12	12	11	11	11	11	11	11	10	10	10	11	12	13	14	15	16	16	17	18	19	20	21	22	23	24	25	26
15	14	13	13	12	12	12	12	12	12	11	11	11	11	12	13	14	14	15	16	17	18	19	20	21	21	22	23	24	25
16	15	14	14	13	13	13	13	13	13	12	12	12	12	12	13	14	15	15	16	17	18	19	20	21	22	22	23	24	25

Tabla 13. Matriz de costos del término *abney level* obtenida con el algoritmo de Wagner y Fisher, considerando los costos propuestos por Levenshtein

Los primeros cinco alineamientos obtenidos se muestran a continuación.

Alineamiento 1

Def. 1	abney_level						a				small	hand	instrument	used			by	surveyors	for	measuring	slopes	and	angles	above	the				horizon
Def. 2	abney_level	a	surveying	instrument	consisting	of	a	spirit	level	and	a	sighting	tube	used	to	measure	the	angle	of	inclination	of	a	line	from	the	observer	to	another	point
Costos	0	1	2	3	4	5	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	21	22	23	24	25
Tipo	I	N	N	N	N	N	I	N	N	N	C	C	C	I	N	N	C	C	C	C	C	C	C	C	I	N	N	N	C

Alineamiento 2

Def. 1	abney_level	a									small	hand	instrument	used			by	surveyors	for	measuring	slopes	and	angles	above	the				horizon
Def. 2	abney_level	a	surveying	instrument	consisting	of	a	spirit	level	and	a	sighting	tube	used	to	measure	the	angle	of	inclination	of	a	line	from	the	observer	to	another	point
Costos	0	0	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	21	22	23	24	25
Tipo	I	I	N	N	N	N	N	N	N	N	C	C	C	I	N	N	C	C	C	C	C	C	C	C	I	N	N	N	C

Alineamiento 3

Def. 1	abney_level						a			small		hand	instrument	used			by	surveyors	for	measuring	slopes	and	angles	above	the				horizon
Def. 2	abney_level	a	surveying	instrument	consisting	of	a	spirit	level	and	a	sighting	tube	used	to	measure	the	angle	of	inclination	of	a	line	from	the	observer	to	another	point
Costos	0	1	2	3	4	5	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	21	22	23	24	25
Tipo	I	N	N	N	N	N	I	N	N	C	N	C	C	I	N	N	C	C	C	C	C	C	C	C	I	N	N	N	C

Alineamiento 4

Def. 1	abney_level	a								small		hand	instrument	used			by	surveyors	for	measuring	slopes	and	angles	above	the				horizon
Def. 2	abney_level	a	surveying	instrument	consisting	of	a	spirit	level	and	a	sighting	tube	used	to	measure	the	angle	of	inclination	of	a	line	from	the	observer	to	another	point
Costos	0	0	1	2	3	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	21	22	23	24	25
Tipo	I	I	N	N	N	N	N	N	N	C	N	C	C	I	N	N	C	C	C	C	C	C	C	C	I	N	N	N	C

Alineamiento 5

Def. 1	abney_level						a		small			hand	instrument	used			by	surveyors	for	measuring	slopes	and	angles	above	the				horizon
Def. 2	abney_level	a	surveying	instrument	consisting	of	a	spirit	level	and	a	sighting	tube	used	to	measure	the	angle	of	inclination	of	a	line	from	the	observer	to	another	point
Costos	0	1	2	3	4	5	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	21	22	23	24	25
Tipo	I	N	N	N	N	N	I	N	C	N	N	C	C	I	N	N	C	C	C	C	C	C	C	C	I	N	N	N	C

TESIS CON
FALLA DE ORIGEN

Si ahora en lugar de considerar los costos propuestos por el algoritmo de la distancia de edición utilizamos los siguientes costos: inserción $w(\varepsilon,b) = 1$, borrado $w(a,\varepsilon) = 1$, sustitución $w(a,b) = 1$, e igualdad $w(a,a) = -1$ (con lo cual se premia el alineamiento de palabras iguales y se castiga cualquier otra operación), obtenemos entonces la siguiente tabla de costos:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
2	0	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
3	1	-1	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
4	2	0	0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
5	3	1	1	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
6	4	2	2	0	0	1	2	3	4	5	6	7	8	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
7	5	3	3	1	1	1	2	3	4	5	6	7	8	8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
8	6	4	4	2	2	2	2	3	4	5	6	7	8	9	9	9	10	11	12	13	14	15	16	17	18	19	20	21	22
9	7	5	5	3	3	3	3	3	4	5	6	7	8	9	10	10	10	11	12	13	14	15	16	17	18	19	20	21	22
10	8	6	6	4	4	4	4	4	4	5	6	7	8	9	10	9	10	11	12	13	14	15	16	17	18	19	20	21	22
11	9	7	7	5	5	5	5	5	5	5	6	7	8	9	10	10	10	11	12	13	14	15	16	17	18	19	20	21	22
12	10	8	8	6	6	6	6	6	6	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	18	19	20	21	22
13	11	9	9	7	7	7	7	7	7	5	5	6	7	8	9	10	11	10	11	12	13	14	15	16	17	18	19	20	21
14	12	10	10	8	8	8	8	8	8	6	5	6	7	8	9	10	11	11	11	12	13	14	15	16	17	18	19	20	21
15	13	11	11	9	9	9	9	9	9	7	7	7	7	8	9	10	9	10	11	12	13	14	15	16	15	16	17	18	19
16	14	12	12	10	10	10	10	10	10	8	8	8	8	8	9	10	10	10	11	12	13	14	15	16	16	16	17	18	19

Tabla 14. Matriza de costos obtenida con el algoritmo de Wagner y Fisher. considerando como costos de las operaciones: inserción $w(\varepsilon,b) = 1$, borrado $w(a,\varepsilon) = 1$, sustitución $w(a,b) = 1$, e igualdad $w(a,a) = -1$

Los primeros cinco alineamientos se presentan en las tablas siguientes.

Alineamiento 1

Def. 1	abney_level	a	small	hand	instrument											used			by	surveyors	for	measuring	slopes	and	angles	above	the				horizon
Def. 2	abney_level	a		surveying	instrument	consisting	of	a	spirit	level	and	a	sighting	tube	used	to	measure	the	angle	of	inclination	of	a	line	from	the	observer	to	another	point
Costos	-1	-2	-1	0	-1	0	1	2	3	4	5	6	7	8	7	8	9	10	11	12	13	14	15	16	17	16	17	18	19	20
Tipo	I	I	N	C	I	N	N	N	N	N	N	N	N	N	I	N	N	C	C	C	C	C	C	C	C	I	N	N	N	C

Alineamiento 2

Def. 1	abney_level	a	small	hand	instrument											used			by	surveyors	for	measuring	slopes	and	angles	above	the				horizon
Def. 2	abney_level	a	surveying		instrument	consisting	of	a	spirit	level	and	a	sighting	tube	used	to	measure	the	angle	of	inclination	of	a	line	from	the	observer	to	another	point
Costos	-1	-2	-1	0	-1	0	1	2	3	4	5	6	7	8	7	8	9	10	11	12	13	14	15	16	17	16	17	18	19	20
Tipo	I	I	C	N	I	N	N	N	N	N	N	N	N	N	I	N	N	C	C	C	C	C	C	C	C	I	N	N	N	C

Alineamiento 3

Def. 1	abney_level	a	small	hand	instrument											used		by		surveyors	for	measuring	slopes	and	angles	above	the				horizon
Def. 2	abney_level	a		surveying	instrument	consisting	of	a	spirit	level	and	a	sighting	tube	used	to	measure	the	angle	of	inclination	of	a	line	from	the	observer	to	another	point
Costos	-1	-2	-1	0	-1	0	1	2	3	4	5	6	7	8	7	8	9	10	11	12	13	14	15	16	17	16	17	18	19	20
Tipo	I	I	N	C	I	N	N	N	N	N	N	N	N	N	I	N	C	N	C	C	C	C	C	C	C	I	N	N	N	C

Alineamiento 4

Def. 1	abney_level	a	small	hand	instrument											used		by		surveyors	for	measuring	slopes	and	angles	above	the				horizon
Def. 2	abney_level	a	surveying		instrument	consisting	of	a	spirit	level	and	a	sighting	tube	used	to	measure	the	angle	of	inclination	of	a	line	from	the	observer	to	another	point
Costos	-1	-2	-1	0	-1	0	1	2	3	4	5	6	7	8	7	8	9	10	11	12	13	14	15	16	17	16	17	18	19	20
Tipo	I	I	C	N	I	N	N	N	N	N	N	N	N	N	I	N	C	N	C	C	C	C	C	C	C	I	N	N	N	C

Alineamiento 5

Def. 1	abney_level	a	small	hand	instrument											used	by			surveyors	for	measuring	slopes	and	angles	above	the				horizon
Def. 2	abney_level	a		surveying	instrument	consisting	of	a	spirit	level	and	a	sighting	tube	used	to	measure	the	angle	of	inclination	of	a	line	from	the	observer	to	another	point
Costos	-1	-2	-1	0	-1	0	1	2	3	4	5	6	7	8	7	8	9	10	11	12	13	14	15	16	17	16	17	18	19	20
Tipo	I	I	N	C	I	N	N	N	N	N	N	N	N	N	I	C	N	N	C	C	C	C	C	C	C	I	N	N	N	C

TESIS CON
FALLA DE ORIGEN

Como puede observarse, los alineamientos obtenidos favorecen el alineamiento de palabras iguales. Los primeros cinco alineamientos generados tienen cinco parejas de palabras iguales cada uno, mientras que el algoritmo básico generó alineamientos con cuatro parejas iguales cada uno.

Cabe hacer notar que el alineamiento 1 obtenido en este último caso se localizó en la posición 187 de los alineamientos generados con los costos originales.

El número de alineamientos calculados para *abney level* con los costos propuestos en esta sección son **32,256**, esto es, se redujo un 96.67% de los alineamientos posibles considerados en el algoritmo básico.

3.3 Pares semi-nulos

En la sección 2.5.3 se mencionó la posibilidad de considerar algunos pares nulos como si fueran pares iguales, con lo que algunos pares-vinculados que inicialmente no habrían sido considerados como pares-vinculados, bajo estas nuevas condiciones si lo serían. A continuación se propone una posible condición bajo la cual un par nulo puede considerarse un par *"semi-nulo"*.

Denominamos *par semi-nulo* a aquel par-nulo que contiene una palabra irrelevante, perteneciente a la lista de palabras irrelevantes propuesta por el usuario (*stop list*).

Las palabras contenidas en la lista de palabras irrelevantes no aportan información relevante durante el proceso de identificación de pares-vinculados, por lo tanto, los pares nulos que agrupen palabras de esta lista, y sólo de esta lista, pueden ser considerados pares iguales, bajo la óptica de que alinear las palabras irrelevantes con la cadena vacía (ε) puede ser equivalente, en este contexto, a insertar la palabra involucrada en lugar de ε.

Por ejemplo, considérense las siguientes definiciones de *decelerometer*:

A. ***decelerometer*** an instrument for ascertaining the deceleration of a moving body [OED1994]
B. ***decelerometer*** an instrument for measuring deceleration [CED1994]

Al alinear estas definiciones obtenemos:

Def. 1	decelerometer	an	instrument	for	ascertaining	the	deceleration	of	a	moving	body
Def. 2	decelerometer	an	instrument	for	measuring	ε	deceleration	ε	ε	ε	ε
Costo	0	0	0	0	1	2	2	3	4	5	6
LCC	0	0	0	0	5	0	0	0	0	0	0
Tipo	*I*	*I*	*I*	*I*	*C*	*N*	*C*	*N*	*N*	*N*	*N*

** Tipos de pares identificados I = Par igual, N = Par nulo, C = Par correspondiente*

Tabla 15. Alineamiento del término *decelerometer*

De este ejemplo, se observa que, bajo la alternativa propuesta, los pares nulos (*the*. ε), (*of*, ε), (*a*, ε), son ahora semi-nulos por estar formados por una palabra funcional y la cadena vacía. El primer par (*the*, ε), puede ser omitido de la cadena sin alterar el sentido de la misma, debido a que después de este existe un par igual. De igual manera, es posible agregar la palabra funcional *the* en la segunda definición y conservar así el sentido. En otras palabras, son intercambiables los elementos de estos pares.

En este alineamiento el par-correspondiente (*ascertaining measuring*) tiene un valor de LCC=5, pero no cumplen con los valores de frontera. Ahora, con los pares semi-nulos, aumenta el valor de LCC en cuatro unidades y además cumple con las condiciones de frontera.

Ahora bien, si se observa el alineamiento, al final de las definiciones aparecen los pares (*of*, ε) (*a*, ε) (*moving*, ε) (*body*, ε) donde las palabra *of* y *a* son palabras funcionales (se encuentran dentro de la lista de palabras irrelevantes proporcionada al sistema). Puede observarse que el segmento *of a moving body* es información adicional que contiene la primera definición y que bien puede ser incluida en la segunda. De hecho, con los pares semi-nulos se ofrece la posibilidad de contemplar la complementariedad de las definiciones, esto es, se puede determinar la información que contiene una y que no contiene la otra.

3.4 Pares semi-iguales

Bajo la misma idea de los pares semi-nulos se puede plantear el *par semi igual*.

Definimos como *par semi-igual* a 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 del LCC de otros pares correspondientes.

Por ejemplo, considérese las siguientes definiciones de *acidimeter*:

A. ***acidimeter*** an instrument for measuring the strength of acids [OED1994]
B. ***acidimeter*** any instrument for determining the amount of acid in a solution [CED1994]

Estas definiciones pueden alinearse de la siguiente manera:

Def. 1	acidimeter	an	instrument	for	**measuring**	the	**strength**	of	acids	ε	ε	ε
Def. 2	acidimeter	any	instrument	for	**determining**	the	**amount**	of	acid	in	a	solution
Costos	0	1	1	1	2	2	3	3	3	4	5	6
LCC	0	4	0	0	4	0	3	0	0	0	0	0
Tipo	I	C	I	I	C	I	C	I	I	N	N	N

Tipos de pares identificados I = Par igual, N = Par nulo, C = Par correspondiente

Tabla 16. Alineamiento del término *acidimeter*

Independientemente de la consideración de los pares semi-nulos, el par (*measuring. determining*) no ha sido identificado como un par-vinculado porque su valor de LCC es de 4. Sin embargo, tres lugares a la izquierda se encuentra el par correspondiente (*an. any*) constituido por palabras funcionales; si consideramos a estas palabras como iguales, bajo la premisa de que dichas palabras son *"equivalentes"* debido a que ambas son funcionales, entonces el valor de LCC para (*measuring, determining*) se verá incrementado en dos unidades y se identificará entonces como un par-vinculado.

3.5 Algoritmo flexibilizado de alineamiento semántico

Las variantes al *"algoritmo de alineamiento semántico básico"*, presentadas en lo que va de este capítulo, se integran en lo que lamamos el *"algoritmo flexibilizado de alineamiento semántico"*, y que se muestra en la siguiente figura.

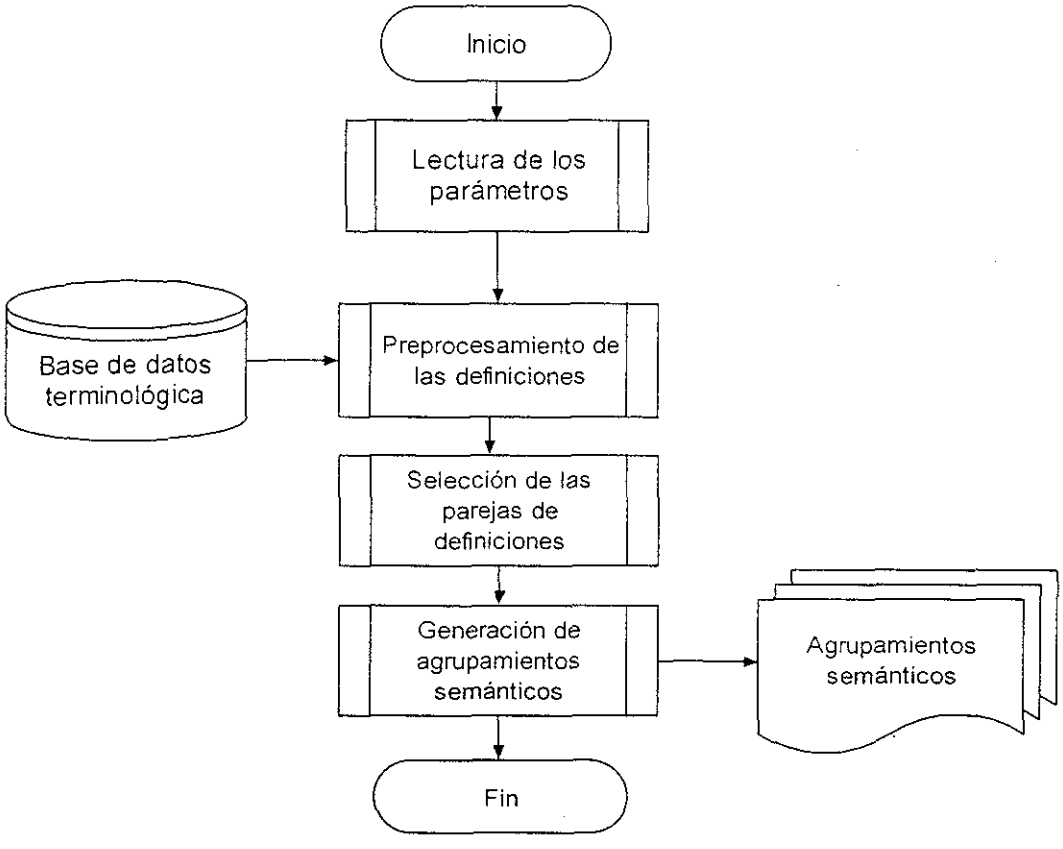


Figura 8. Diagrama general del algoritmo flexibilizado de alineamiento semántico

En ella se distinguen cuatro etapas fundamentales:

- En la proceso de lectura de parámetros se lee información que permite elegir las diferentes variantes del algoritmo. Durante el desarrollo de esta tesis, las siguientes variantes han sido desarrolladas:

    - Las operaciones que se utilizarán además de las operaciones de edición básicas.
    - Los costos de estas operaciones.
    - El idioma en que se encuentran las definiciones terminológicas.
    - El uso del lematizador.
    - La generación de un solo alineamiento o de alineamientos múltiples.
    - La consideración de pares semi-nulos y/o pares semi-iguales.
    - Diferenciar mayúsculas y minúsculas.

- Durante el preprocesamiento de las definiciones, se eliminan los signos de puntuación y, en caso de solicitarse, se lematizan.
- La generación de las parejas de definiciones se realiza conforme al algoritmo básico, es decir, se eligen pares de una misma definición cuya acepción ha sido dada por diferentes fuentes.
- Finalmente, se efectúa el proceso de generación de los agrupamientos semánticos. Este proceso constituye el corazón del algoritmo flexibilizado de alineamiento semántico.

El proceso de generación de agrupamientos semánticos se presenta en el diagrama de flujo de la figura 9. Como lo muestra el diagrama, el proceso general es el siguiente:

Los pares de definiciones se procesan cada uno generando la matriz de costos de acuerdo con Wagner y Fisher y con base en los costos de las operaciones establecidas. Al principio las definiciones empleadas son las originales, sin embargo con cada nuevo ciclo se generan nuevas definiciones que deben ser alineadas nuevamente.

Para cada matriz de costos se establecen los alineamientos posibles y se identifican los pares-vinculados generados.

Si se generaron nuevos pares-vinculados entonces se unen estos nuevos pares para formar nuevos agrupamientos semánticos o para actualizar los ya existentes. En caso contrario el algoritmo termina.

Una vez generados nuevos agrupamientos semánticos, los pares-vinculados se sustituyen en las definiciones, generándose así un nuevo conjunto de definiciones, las cuales son alimentadas nuevamente al algoritmo.

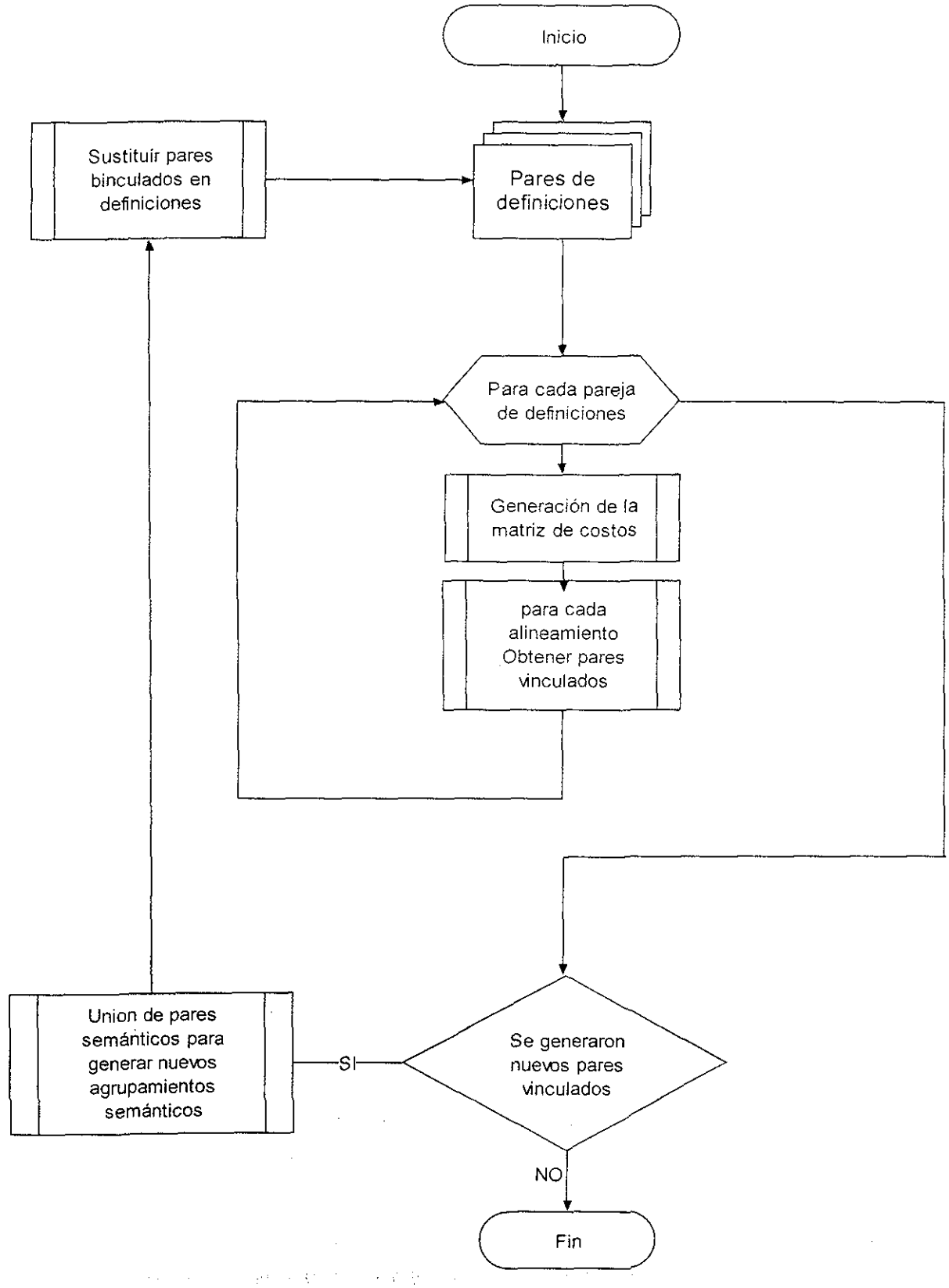


Figura 9. Diagrama de flujo para la generación de agrupamientos semánticos

El algoritmo flexibilizado de alineamiento semántico se detalla a continuación.

El algoritmo comienza con la generación de la matriz de costos para cada par de definiciones con base en las operaciones de edición a usar (inserción, borrado. sustitución, intercambio, intercambio conjuntivo). Se aplica entonces la ecuación de recurrencia presentada en la sección 3.1.3:

$$C_{i,j} = \min \begin{cases} C_{i-1,j} + w(a_i, \varepsilon) \\ C_{i,j-1} + w(\varepsilon, b_j) \\ C_{i-1,j-1} + w(a_i, b_j) \\ C_{i-2,j-2} + w_S(a_i b_{j-1}, a_{i-1} b_j) \\ C_{i-3,j-3} + w_C(a_i, b_j) \end{cases}$$

El algoritmo a aplicar es:

Algoritmo: Generación de la matriz de costos con intercambio de palabras consecutivas e intercambio conjuntivo

Entrada: A , B : *secuencia de palabras a transformar*
n: Longitud de la secuencia A
m: Longitud de la secuencia B

Salida: C : *Matriz de costos*
$C_{m,n} = ed(A,B)$

$C_{0,0} = 0$
for $j = 1$ *to* n
$C_{0,j} = C_{0,j-1} + w(\varepsilon, b_j)$
for $i = 1$ *to* m
$C_{i,0} = C_{i-1,0} + w(a_i, \varepsilon)$
for $j = 1$ *to* n
Cálculo de $C_{i,j}$:

$$C_{i,j} = \min \begin{cases} C_{i-1,j} + w(a_i, \varepsilon) \\ C_{i,j-1} + w(\varepsilon, b_j) \\ C_{i-1,j-1} + w(a_i, b_j) \\ C_{i-2,j-2} + w_S(a_i b_{j-1}, a_{i-1} b_j) \\ C_{i-3,j-3} + w_C(a_i, b_j) \end{cases}$$

Una vez calculada la matriz de costos se procede a determinar el alineamiento o los alineamientos correspondientes para cada par de definiciones, para ello se aplica el algoritmo propuesto en la sección 3.2, considerando las propuestas presentadas en esta

tesis a fin de tomar en cuenta las operaciones de intercambio de dos palabras consecutivas y de intercambio conjuntivo. A continuación se presenta el algoritmo para la generación de los alineamientos semánticos

Algoritmo: Generación de todos los alineamientos posibles, incluyendo las operaciones de inserción, borrado, sustitución, intercambio de palabras e intercambio conjuntivo de palabras

Entrada: *A, B :* *Secuencia de palabras a alinear*
C : *Matriz de costos calculada según el algoritmo de distancia de edición con intercambio de palabras consecutivas e intercambio conjuntivo*
r, c *celda a analizar dentro de la matriz de costos (r = renglón y c = columna)*
Z : *Pares de palabras alineadas hasta esta llamada*

Salida: *Z :* *Cada vez que se llama a* ***reporta alineamiento*** *(Z) se presenta un alineamiento posible de las cadenas A y B tal que su costo es mínimo*

Alinear(A,r,B,c,Z,C)
if $r \neq 0$
  if $c \neq 0$
    if $C_{r,c} = C_{r-1,c-1} + w(a_r,b_c)$
      Alinear$(A,r-1,B,c-1,(a_r,b_c)\cdot Z,C)$
    if $C_{r,c} = C_{r-1,c} + w(a_r,\varepsilon)$
      Alinear$(A,r-1,B,c,(a_r,\varepsilon)\cdot Z,C)$
    if $C_{r,c} = C_{r,c-1} + w(\varepsilon,b_c)$
      Alinear$(A,r,B,c-1,(\varepsilon,b_c)\cdot Z,C)$
    if $((r >= 2)$ *and* $(c>=2))$ *and* $(C_{r,c} = C_{r-2,c-2} + w_s(a_r b_{c-1}, a_{r-1} b_c))$
      Alinear$(A,r-2,B,c-2,(a_{r-1},b_{c-1})\ (a_r,b_c)\cdot Z,C)$
    if $((r >= 3)$ *and* $(c>=3))$ *and* $(C_{r,c} = C_{r-3,c-3} + w_c(a_r,b_c))$
      Alinear$(A,r-3,B,c-3,\ (a_{r-2},b_{c-2})\ (a_{r-1},b_{c-1})\ (a_r,b_c)\cdot Z,C)$
  else
    Alinear$(A,r-1,B,0,(a_r,\varepsilon)\cdot Z,C)$
else
  if $c \neq 0$
    Alinear$(A,0,B,c-1,(\varepsilon,b_c)\cdot Z,C)$
  else
    reporta_alineamiento(Z)

Una vez establecidos el o los alineamientos se identifica los pares-vinculados. Para ello se toma en cuenta o no a los pares semi-nulos y a los pares semi-iguales tal como se describe en las secciones 3.3 y 3.4.

Los pares-vinculados que contengan palabras irrelevantes son rechazados y no se consideran en el procesamiento posterior. Este proceso se describió en la sección 1.4.

Tal como se expone en la sección 1.5, los pares-vinculados se sustituyen en las parejas de definiciones, generando de esta manera nuevas definiciones y a partir de estas puede realizarse de nuevo el análisis.

Finalmente se mezclan los pares-vinculados obtenidos de acuerdo con lo propuesto por el algoritmo básico, cuya descripción se encuentra en la sección 1.6.

El algoritmo se repite hasta que no se generen nuevos pares-vinculados.

La evaluación de los resultados de esta propuesta se muestra en el siguiente capítulo.

3.6 Recapitulación

Con el fin de mejorar sustancialmente el número de pares-vinculados e identificados correctamente (pares-semánticos), en este capítulo se escogieron cuatro Heurísticas alternativas principales a desarrollar. Se presentaron 6 algoritmos, los cuales se integraron en lo que denominamos el algoritmo flexibilizado de alineamiento semántico y se implantaron en un sistema desarrollado durante la presente tesis, que puede ser consultado en la página http://iling.iingen.unam.mx.

Como resultado del desarrollo de este capítulo se puede observar que:

> Si bien la inversión de dos palabras consecutivas ha sido considerada por otros autores, en esta tesis se presentó el algoritmo de intercambio conjuntivo de palabras, algoritmo que fácilmente puede ser generalizado a otro tipo de conectores.
>
> La inclusión de rutas múltiples plantea el problema de que los posibles alineamientos crecen rápidamente en función de la diferencia de las longitudes de las definiciones. Sin embargo, al incluir el costo negativo en el alineamiento de pares de palabras iguales se limita significativamente este crecimiento, obteniéndose intuitivamente alineamientos de mejor calidad. Cabe considerar que los costos negativos no han sido documentados en la literatura.
>
> La introducción del concepto de pares semi-nulos y pares semi-iguales mejora intuitivamente la identificación de pares-vinculados.

Con el sistema desarrollado se realizaron corridas para varios conjuntos de definiciones, correspondientes a diferentes áreas temáticas. En un análisis preliminar se observó que el número de pares-vinculados aumenta no sólo en cantidad, sino que también en calidad. En el siguiente capítulo se hace una evaluación cuantitativa de los resultados obtenidos.

4 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 con respecto al algoritmo básico.

Es importante señalar que en la discusión siguiente se ha tenido especial cuidado en utilizado 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.

Debido a que en esta tesis todas las modificaciones propuestas alteran el proceso de identificación de pares semánticos (y no la generación de agrupamientos semánticos), la evaluación de resultados, tanto del algoritmo básico como del algoritmo flexibilizado, se realiza con base en los pares semánticos.

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 manualmente obtenidos 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.

4.1 Método de evaluación

Cuando se considera el problema de evaluar la eficiencia en los algoritmos de recuperación de información es necesario considerar primero la naturaleza del proceso de recuperación que se debe llevar a cabo. En este sentido, se pueden identificar dos grandes grupos:

- Procesos autónomos: Es un proceso, conocido como "batch", donde el usuario hace una solicitud y la computadora simplemente responde.
- Procesos interactivos: Implican una sesión de trabajo donde el usuario y la computadora interactúan para llegar a un resultado final.

La evaluación de procesos como el que se presenta en esta tesis, sigue la línea planteada por la evaluación de los procesos autónomos. Una de las técnicas más utilizadas para la evaluación de los procesos de recuperación de información se

conocen en inglés como: ***recall and precision***[8]. A continuación se explican brevemente estos dos conceptos.

4.1.1 La técnica de "Recall and Precision"

En la discusión que sigue, I representa la variante del algoritmo de alineamiento que se desea evaluar, A es el conjunto de pares-vinculados obtenidos mediante la variante I; R es el conjunto de pares semánticos identificados manualmente. $|X|$ representa el número de elementos del conjunto X.

El conjunto de los pares-vinculados que se recuperaron de acuerdo con la variante I y que están también en el conjunto de pares semánticos identificados manualmente se puede calcular como la intersección del conjunto pares-vinculados obtenidos (A) y el conjunto de pares semánticos identificados manualmente (R), representamos a este conjunto de pares-vinculados identificados por el algoritmo como Ra, de modo que:

$$Ra = R \cap A$$

La tabla 17 resume la nomenclatura utilizada presentada en esta sección y utilizada a lo largo de este capítulo, se han incluido, por claridad en la consulta de la tabla, la nomenclatura presentada en la sección 4.1.2

Las medidas *Recall* y *Precision* se definen como:

> **Recall** (*Re*) es la relación que existe entre de pares-semánticos identificados y los pares semánticos obtenidos manualmente.
>
> $$\mathrm{Re} = \frac{|Ra|}{|R|}$$
>
> Un valor de *Re* igual a 1 indicará que la variante I ha recuperado todos los pares semánticos que se han identificado manualmente. Por el contrario, un valor de *Re* igual a cero indicará que la variante I no ha recuperado ninguno de los pares semánticos identificados manualmente. Como se puede apreciar, entre más cercano de 1 se encuentre *Re* indicará que se han recuperado un mayor número de pares semánticos.
>
> **Precision** (*Pr*) Es la relación que existe entre los pares-semánticos identificados por la variante I y el total de pares-vinculados generados por la variante I.

---

[8] Los término de *recall* y *precision* se han traducido de distintas maneras en la literatura. Sin embargo, las propuestas, sobretodo para *recall*, resultan ambiguas. Por ello, en esta tesis se optó por conservar los términos en inglés.

$$\Pr = \frac{|Ra}{|A}$$

Un valor igual a 1 indicará que todos los pares-vinculados recuperados son correctos (es decir todos son pares-semánticos). Un valor igual a cero indicará que ninguno de los pares-vinculados recuperados es correcto. Como se puede apreciar entre más cercano a uno esté el valor de *precision* mayor será la precisión de los pares-semánticos identificados.

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.

I	Representa la variante del algoritmo de alineamiento que se desea evaluar
$\|X\|$	Representa el número de elementos del conjunto X
A	El conjunto de pares-vinculados obtenidos mediante la variante I.
R	Es el conjunto de pares-semánticos identificados manualmente.
Ra	Conjunto de pares-semánticos identificados por la variante I.
$\|Ra_0\|$	Número de pares-semánticos identificados por el algoritmo de comparación base.
$\|A_0\|$	Número de pares-vinculados identificados por el algoritmo de comparación base
Re	(*Recall*) Relación que existe entre de pares-semánticos identificados y los pares-semánticos obtenidos manualmente.
Pr	(*Precision*) Relación que existe entre los pares-semánticos identificados por la variante I y el total de pares-vinculados generados por la variante I.
ΔA	(Índice de identificación de pares-vinculados) indica el porcentaje en que se incrementó la identificación de pares-vinculados al incluir las variantes del algoritmo de comparación alternativo.
ΔRa	(Índice de recuperación de pares-semánticos) indica el porcentaje en que incrementó la eficiencia del algoritmo de comparación base al incluir las alternativas contempladas en el algoritmo de comparación alternativo.

Tabla 17. Nomenclatura empleada para la evaluación de las alternativas.

4.1.2 Comparación entre alternativas

La técnica *recall* y *precision* permite evaluar el desempeño de un algoritmo respecto a resultado ideal esperado, en este caso, respecto a los pares-semánticos identificados manualmente. Sin embargo, no ofrece ninguna evaluación del incremento o decremento de los resultados obtenidos por una alternativa de un algoritmo respecto a los resultados que ofrece otra alternativa.

La evaluación relativa del desempeño de las variantes del algoritmo básico de alineamiento semántico entre sí es necesaria, pues en esta tesis se requiere establecer qué combinación de alternativas ofrece los mejores resultados.

El algoritmo respecto a la cual se evaluarán los resultados ofrecidos al agregar una o más alternativas se denomina *algoritmo de comparación base*, por extensión, los resultados obtenidos por ese algoritmo se denominan *resultados de comparación base*.

Un *algoritmo de comparación alternativo* será el que se obtiene al incluir en el algoritmo de comparación base una o más alternativas y cuyo resultado se desea evaluar (*resultados de comparación alternativos*).

En esta tesis se proponen los siguientes factores de evaluación:

Índice de identificación de pares-vinculados: indica el porcentaje en que se incrementó la identificación de pares-vinculados al incluir las variantes analizadas:

$$\Delta\mathbf{A} = (|A|-|A_0|)/|A_0|$$

En este cociente, $|A_0|$ corresponde al número de pares-vinculados identificados por el algoritmo de comparación base. La interpretación de la variación en la identificación de pares vinculados es la siguiente:

Un valor de $\Delta\mathbf{A}=0$ indica que las variantes incluidas en el algoritmo de comparación alternativo no ofrecen una variación en el número de pares-vinculados identificados.

Un valor positivo de $\Delta\mathbf{A}$ indica un aumento en el número de pares-vinculados respecto a los identificados por el algoritmo de comparación base (un valor de 1 indica un aumento del 100%).

Por el contrario, un valor negativo de $\Delta\mathbf{A}$ indica una disminución en el número de pares-vinculados respecto a los identificados por el algoritmo de comparación base (un valor de -1 indica una disminución del 100%).

Índice de recuperación de pares-semánticos. Esta medida indica el porcentaje en que incrementó la eficiencia del algoritmo de comparación base respecto al incluir las alternativas contempladas en el algoritmo de comparación alternativo.

$$\Delta\mathbf{Ra} = (|Ra|-|Ra_0|)/|Ra_0|$$

El término $|Ra_0|$ corresponde al número de pares-semánticos identificados por el algoritmo de comparación base. El índice de recuperación de pares vinculados se interpreta de la siguiente manera:

Un valor de Δ**Ra** =0 indica que las variantes incluidas en el algoritmo de comparación alternativo no ofrecen una variación en el número de pares-semánticos.

Un valor positivo de Δ**Ra** indica un aumento en el número de pares-semánticos correctos respecto a los identificados por el algoritmo de comparación base (un valor de 1 indica un aumento del 100%).

Por el contrario, un valor negativo de Δ**Ra** indica una disminución en el número de pares-semánticos respecto a los identificados por el algoritmo de comparación base (un valor de -1 indica una disminución del 100%).

Como es de esperarse, al confrontar el algoritmo de comparación base consigo mismo tanto el índice de identificación de pares-vinculados como el índice de recuperación de pares-semánticos es cero.

Los índices *recall* y *precision*, como también los índices de identificación de pares vinculados y de recuperación de pares semánticos, fueron utilizados para evaluar las distintas combinaciones propuestas en esta tesis.

4.2 Corpus de metrología

El algoritmo básico se aplicó a un corpus de metrología, cuyas características se resumen a continuación.

El corpus de metrología 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 definicion para el mismo diccionario.

En esta tesis se analizaron las definiciones de ese diccionario y con ayuda de un traductor español-inglés certificado se identificaron un total de 363 pares-semánticos: 285 pares-semánticos simples y 78 pares semánticos compuestos[9].

4.3 Pruebas realizadas

Para la evaluación fue necesario considerar las diferentes variaciones del algoritmo básico de agrupamiento semántico que se exploraron en esta tesis

Las pruebas efectuadas se dividieron en dos grandes grupos:

---

[9] Como se comentó en la sección 1.10, un par-semántico simple es aquel en el que cada elemento del par está constituido por una sola palabra, mientras que un par-semántico compuesto cuenta con más de una palabra en alguno de los elementos que forman el par.

Pruebas con un solo alineamiento: aquellas que emplearon el alineamiento original propuesto por Wagner y Fisher (pruebas 1 a 9)

Pruebas considerando alineamientos múltiples: aquellas que emplearon alineamientos múltiples, limitando, a 20 posibles alineamientos[10] (pruebas 10 a 18).

Dentro de cada uno de estos dos grupos se evaluaron los resultados al considerar una o más de las siguientes variantes del algoritmo:

a) pares semi-iguales.
b) pares semi-nulos.
c) modificación de costos (sólo se evalúa $w_i(a,a) = -1$).
d) intercambio de palabras consecutivas e intercambio conjuntivo de palabras.

La tabla 18 resume las pruebas realizadas.

No. de prueba	*Alineamientos múltiples*	*Semi iguales*	*Semi nulos*	*Modificación de costos*	*Intercambio de palabras consecutivas*	*Intercambio conjuntivo de palabras*
1	no	no	no	no	no	no
2	no	**si**	no	no	no	no
3	no	no	**si**	no	no	no
4	no	no	no	**si**	no	no
5	no	no	no	no	**si**	**si**
6	no	**si**	**si**	no	no	no
7	no	**si**	**si**	**si**	no	no
8	no	**si**	**si**	no	**si**	**si**
9	no	**si**	**si**	**si**	**si**	**si**
10	**si**	no	no	no	no	no
11	**si**	**si**	no	no	no	no
12	**si**	no	**si**	no	no	no
13	**si**	no	no	**si**	no	no
14	**si**	no	no	no	**si**	**si**
15	**si**	**si**	**si**	no	no	no
16	**si**	**si**	**si**	**si**	no	no
17	**si**	**si**	**si**	no	**si**	**si**
18	**si**	**si**	**si**	**si**	**si**	**si**

Tabla 18. Pruebas realizadas para la evaluación cuantitativa del algoritmo flexibilizado de alineamiento semántico.

[10] Como ya se comentó en la sección 3.2, el número de alineamientos posibles crece rápidamente con la diferencia de longitudes de las cadenas a alinear. Por lo que en esta tesis, por razones de tiempo de procesamiento y espacio, se ha optado por evaluar sólo las primeras 20 alternativas. Será tema de trabajos futuros explorar la ventaja o no de incrementar o decrementar este parámetro.

Como puede apreciarse la prueba 1 representa los resultados del algoritmo básico, mientras que las modificaciones propuestas en esta tesis están representadas por las pruebas 2 a 18.

Las pruebas se realizaron sobre el corpus en el área de metrología en el idioma inglés.

El apéndice 1 muestra los pares semánticos-identificados manualmente para el corpus de metrología.

El apéndice 2 presenta los pares-semánticos obtenidos por las diferentes variantes del algoritmo para el corpus analizado sin utilizar alineamientos múltiples, mientras que el apéndice 3 muestra los resultados obtenidos al usar alineamientos múltiples limitando la profundidad a 20 alineamientos posibles.

4.4 Resultados

La evaluación de los resultados obtenidos de la aplicación del algoritmo de alineamiento semántico en el corpus de inglés se resume en las tablas 19 y 20 y en la gráfica 10. Las pruebas se han ordenado con base en los índices de *recall* y *precision*.

Una sola ruta para el alineamiento de definiciones

Prueba	|R|	|Ra|	|A|	Recall	Precision	Semi igual	Semi nulo	Modificación de costos	Intercambios
8	363	90	185	0.2479	0.4865	**SI**	**SI**	NO	**SI**
6	363	90	185	0.2479	0.4865	**SI**	**SI**	NO	NO
9	363	89	188	0.2452	0.4734	**SI**	**SI**	**SI**	**SI**
7	363	89	188	0.2452	0.4734	**SI**	**SI**	**SI**	NO
2	363	62	97	0.1708	0.6392	**SI**	NO	NO	NO
3	363	45	71	0.1240	0.6338	NO	**SI**	NO	NO
5	363	30	32	0.0826	0.9375	NO	NO	NO	**SI**
1	363	30	32	0.0826	0.9375	NO	NO	NO	NO
4	363	30	33	0.0826	0.9091	NO	NO	**SI**	NO

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

Tabla 19. Resultados de las pruebas realizadas considerando un solo alineamiento. Se muestra la evaluación de los índices de *recall* y *precision*; las pruebas se han ordenado en orden ascendente de acuerdo con el índice de *recall*.

Análisis de 20 rutas posibles para el alineamiento de definiciones

Prueba	| R |	| Ra |	| A |	*Recall*	*Precision*	Semi igual	Semi nulo	Modificación de costos	Intercambios
18	363	110	394	0.3030	0.2792	SI	SI	SI	SI
16	363	110	394	0.3030	0.2792	SI	SI	SI	NO
17	363	103	409	0.2837	0.2518	SI	SI	NO	SI
15	363	103	414	0.2837	0.2488	SI	SI	NO	NO
12	363	74	211	0.2039	0.3507	NO	SI	NO	NO
11	363	68	113	0.1873	0.6018	SI	NO	NO	NO
13	363	30	34	0.0826	0.8824	NO	NO	SI	NO
10	363	30	34	0.0826	0.8824	NO	NO	NO	NO
14	363	30	35	0.0826	0.8571	NO	NO	NO	SI

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

Tabla 20. Resultados de las pruebas realizadas considerando 20 alineamientos posibles. Se muestra la evaluación de los índices de *recall* y *precision*; las pruebas se han ordenado en orden ascendente de acuerdo con el índice de *recall*

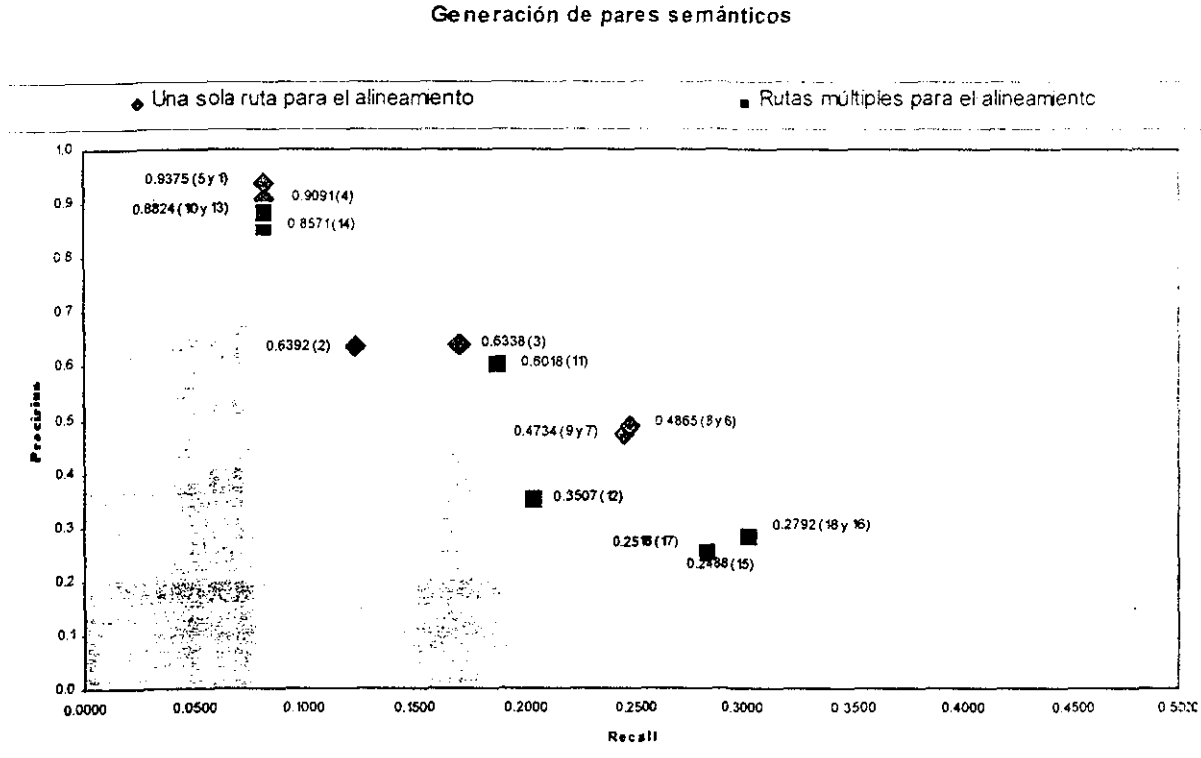


Figura 10. Grafica *Recall* vs *Precision* para las 18 pruebas realizadas. Nota: El número de prueba correspondiente a cada punto se indica entre paréntesis

TESIS CON
FALLA DE ORIGEN

4.5 Análisis de los resultados

El análisis de las pruebas incluye una evaluación de los índices *recall* y *precision* así como índice de identificación de pares-vinculados y del índice de recuperación de pares-semánticos.

Para el análisis de las pruebas realizadas, se han elegido tres variantes del algoritmo de agrupamiento semántico con respecto a las cuales se evaluarán el índice de identificación de pares-vinculados y el índice de recuperación de pares-semánticos. Cada una de estas alternativas las hemos denominado *ejes de comparación*:

a) Eje de comparación: Algoritmo básico de alineamiento semántico (prueba número 1).
b) Eje de comparación: Algoritmo de alineamiento semántico considerando rutas múltiples (prueba 10)
c) Eje de comparación: Algoritmo de alineamiento semántico considerando pares semi-iguales y pares semi-nulos simultáneamente. En este caso particular se han efectuado análisis adicionales para determinar el comportamiento de esta variante al considerar las opciones de un solo alineamiento y alineamientos múltiples. A estas alternativas de análisis las hemos denominado sub-ejes de comparación.

Las secciones 4.5.1 a 4.5.3 presentan los análisis y las observaciones para cada uno de los ejes de comparación. En cada una de las secciones se presenta una tabla que resume el análisis establecido. En caso de existir sub-ejes de comparación, se presentan dos tablas, una para cada sub-eje analizado. El primer renglón de cada tabla representa el eje o sub-eje y en los renglones subsecuentes se muestran los resultados obtenidos al agregar las distintas alternativas analizadas.

Las observaciones al análisis se dividen en dos: aquellas correspondientes al eje o sub-eje de comparación respecto a los resultados ideales y aquellas correspondientes a los resultados obtenidos al agregar cada variante y comparar estos resultados contra el eje.

4.5.1 Eje de comparación: Algoritmo básico

La tabla 21 resume los resultados obtenidos al cotejar, contra el eje de comparación correspondiente al algoritmo de alineamiento semántico básico, las posibles variantes.

Alineamiento Múltiple	Semi iguales	Semi nulos	Modificación de costos	Intercambios	|Ra|	| A |	*Recall*	*Precision*	ΔRa	ΔA
NO	NO	NO	NO	NO	30	32	0.0826	0.9375	0.00%	0.00%
SI	NO	NO	NO	NO	30	34	0.0826	0.8824	0.00%	6.25%
NO	SI	NO	NO	NO	62	97	0.1708	0.6392	106.67%	203.13%
NO	NO	SI	NO	NO	45	71	0.1240	0.6338	50.00%	121.88%
NO	NO	NO	SI	NO	30	33	0.0826	0.9091	0.00%	3.13%
NO	NO	NO	NO	SI	30	32	0.0826	0.9375	0.00%	0.00%

Nota: Se consideró $|Ra_0| = 30$ $|A_0| = 32$

Tabla 21. Análisis de resultados obtenidos al considerar el eje de comparación *Algoritmo básico*.

4.5.1.1 Análisis del eje de comparación algoritmo básico

Respecto al eje de comparación podemos observar que el algoritmo básico genera 32 pares-vinculados, de los cuales 30 son pares-semánticos, con lo que los índices *recall* y *precision* son 0.0826 y 0.9375, respectivamente. El bajo valor de *recall* indica que se han recuperado muy pocos pares-semánticos (8.26% 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.

4.5.1.2 Análisis de las alternativas respecto al eje de comparación algoritmo básico

Al evaluar las diferentes alternativas, se puede observar en la tabla 21, que al considerar alineamientos múltiples, modificación de costos o intercambios no se obtiene ninguna mejora respecto a los resultados obtenidos por el algoritmo básico de alineamiento semántico; incluso en las dos primeras alternativas, si bien se mantuvo constante el índice *recall*, disminuyó el valor de *precision* (de un 93% a un 88% y 90% respectivamente).

Sin embargo, 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 0.1708 y 0.1240, 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.

4.5.1.3 Conclusiones sobre el eje de comparación: algoritmo básico

De las observaciones anteriores, puede establecerse que el algoritmo de alineamiento semántico básico 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 inclusión de alguna otra variante (alineamientos múltiples, modificación de costos o intercambios) no aporta mejora alguna al algoritmo básico.

4.5.2 Eje de comparación: Rutas múltiples

La tabla 22 resume los resultados obtenidos al comparar las posibles variantes contra el eje de comparación correspondiente a la alternativa de rutas múltiples.

Alineamiento Múltiple	Semi iguales	Semi nulos	Modificación de costos	Intercambios	|Ra|	| A |	*Recall*	*Precision*	|ΔRa	ΔA
SI	NO	NO	NO	NO	30	34	0.0826	0.8824	0.00%	0.00%
SI	SI	NO	NO	NO	68	113	0.1873	0.6018	126.67%	232.35%
SI	NO	SI	NO	NO	74	211	0.2039	0.3507	146.67%	520.59%
SI	NO	NO	SI	NO	30	34	0.0826	0.8824	0.00%	0.00%
SI	NO	NO	NO	SI	30	35	0.0826	0.8571	0.00%	2.94%

Nota: Se consideró $|Ra_0| = 30$ $|A_0| = 34$

Tabla 22. Análisis de resultados obtenidos al considerar las posibles alternativas del algoritmo y confrontarlas con el eje de comparación *Rutas múltiples*

4.5.2.1 Análisis del eje de comparación rutas múltiples

De la tabla 22 se observa que los valores de *recall* y *precision* obtenidos por un algoritmo que incluya la evaluación de 20 rutas múltiples, indican que el número de pares-semánticos identificados respectos al universo de pares-semánticos es muy pequeño, del orden de un 8%. Por otra parte, de los pares-vinculados identificados un 88% de ellos sí son pares-semánticos, lo que indica una alta precisión en este rubro.

4.5.2.2 Análisis de las alternativas respecto al eje de comparación rutas múltiples

En la tabla 22 se observa que, al incluir la alternativa de par semi-igual o par semi-nulo, el desempeño del algoritmo mejora notablemente, ofreciendo un incremento en la identificación de pares semánticos del 126% y 146%, respectivamente. Lo anterior se ve reflejado en el índice *recall*, que mejoró respecto al obtenido por el eje de comparación (0.0826), colocándose en un 0.1873 y 0.2039. Sin embargo, para llegar a estos índices *recall*, las modificaciones contempladas en este párrafo disminuyeron su índice *precision*, pasando de un 0.8824 a un 0.6018 y 0.3507, respectivamente, lo cual indica que fue necesario identificar muchos más pares-vinculados: de hecho, el incremento en la identificación de pares-vinculados respecto a los obtenidos en la alternativa eje fue de un 232% y 520%. Lo anterior indica que es mejor elegir alineamientos múltiples y pares semi-iguales que alineamientos múltiples y pares semi-nulos.

Por otro lado, la aplicación de la alternativa correspondiente a modificación de costos o a intercambios, no ofrece mejoras respecto al eje de comparación: de hecho, al incluir intercambios, el valor de *precision* disminuye de un 0.8824 a un 0.8571, manteniéndose el valor de *recall* en 0.0826, lo que indica que en este caso se incrementó el número de pares-vinculados y no se mejoró el número de pares-semánticos.

4.5.2.3 Conclusiones sobre el eje de comparación: rutas múltiples

Nuevamente, las alternativas que mejores resultados ofrecen sólo son pares semi-iguales y pares semi-nulos. En particular, la mejor alternativa que se puede elegir respecto al eje de comparación correspondiente a rutas múltiples es pares semi-iguales. Se obtiene un incremento del 126% en pares semánticos y un incremento del 232% en pares-vinculados.

4.5.3 Eje de comparación: Par semi-igual y par semi-nulo

Como resultado de los análisis del eje de comparación algoritmo básico de alineamiento semántico y del eje de comparación rutas múltiples, se observó que las alternativas que mejores resultados ofrecen son par semi-igual y par semi-nulo. Por este motivo se buscó analizar el efecto que estas dos alternativas combinadas tienen sobre el algoritmo básico de alineamiento semántico y sobre la alternativa de rutas múltiples, por lo que se propone analizar el efecto combinado de estas alternativas mediante el eje de comparación par semi-igual y par semi-nulo.

Para el análisis de este eje se han elegido dos semi-ejes de comparación: la utilización de una sola ruta de alineamiento (algoritmo básico) y un conjunto de 20 rutas posibles (como lo propone la alternativa de alineamientos múltiples). Cada uno de estos dos semi-ejes es analizado en las secciones 4.5.3.1 y 4.5.3.2, para finalmente en la sección 4.5.3.3 ofrecer las conclusiones respecto a este eje de comparación.

4.5.3.1 Semi-eje: Una ruta de alineamiento

Las tablas 23 y 24 resumen los resultados obtenidos al comparar las posibles variantes contra el semi-eje de comparación correspondiente a una sola ruta de alineamiento. La primera de ellas muestra la comparación del semi-eje elegido contra el algoritmo básico de alineamiento semántico, en tanto la segunda muestra los resultados obtenidos al incorporar las distintas alternativas al semi-eje de comparación.

Semi iguales	Semi nulos	Alineamiento Multiple	Modificación de costos	Intercambios	|Ra|	| A |	*Recall*	*Precision*	ΔRa |	ΔA
NO	NO	NO	NO	NO	30	32	0.0826	0.9375	0.00%	0.00%
NO	SI	NO	NO	NO	45	71	0.1240	0.6338	50.00%	121.88%
SI	NO	NO	NO	NO	62	97	0.1708	0.6392	106.67%	203.13%
SI	SI	NO	NO	NO	90	185	0.2479	0.4865	200.00%	478.13%

Nota: Se consideró $|Ra_0| = 30$ $|A_0| = 32$

Tabla 23. Análisis de resultados obtenidos al comparar las alternativas de pares semi-iguales y semi-nulos con respecto al algoritmo básico de alineamiento semántico

Semi iguales	Semi nulos	Alineamiento Multiple	Modificación de costos	Intercambios	|Ra|	| A |	*Recall*	*Precision*	ΔRa	ΔA
SI	SI	NO	NO	NO	90	185	0.2479	0.4865	0.00%	0.00%
SI	SI	NO	SI	NO	89	188	0.2452	0.4734	-1.11%	1.62%
SI	SI	NO	NO	SI	90	185	0.2479	0.4865	0.00%	0.00%
SI	SI	NO	SI	SI	89	188	0.2452	0.4734	-1.11%	1.62%

Nota: Se consideró $|Ra_0| = 90$ $|A_0| = 185$

Tabla 24. Análisis de resultados obtenidos al incorporar las distintas alternativas al semi-eje de comparación *par semi-igual y par semi-nulo* considerando un sólo alineamiento

De la observación de estas tablas se desprenden los siguientes análisis.

4.5.3.1.1 Análisis del semi-eje de comparación: un solo alineamiento

De la tabla 23 se observa que la aplicación simultánea de la alternativa de par semi-igual y par semi-nulo trae como consecuencia un aumento en el índice *recall* (0.2479); este aumento supera el obtenido por la aplicación individual de par-semi-igual (0.1708) o par semi-nulo (0.1240). 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%.

4.5.3.1.2 Análisis de las alternativas respecto al semi-eje de comparación de un solo alineamiento

La tabla 23 muestra que no se obtiene mejora alguna en el desempeño del algoritmo al agregar las variantes de modificación de costos e intercambio de palabras. Incluso se percibe una ligera disminución tanto en el índice *recall* como en el índice *precision*.

4.5.3.2 Semi-eje: Rutas múltiples

Las tablas siguientes resumen los resultados obtenidos al comparar las posibles variantes contra el semi-eje de comparación correspondiente a un conjunto de 20 rutas de alineamiento. La primera de ellas muestra la comparación del semi-eje elegido contra el algoritmo básico de alineamiento semántico, mientras la segunda muestra los resultados obtenidos al incorporar las distintas alternativas al semi-eje de comparación.

Semi iguales	Semi nulos	Alineamiento Múltiple	Modificación de costos	Intercambios	|Ra|	| A |	*Recall*	*recision*	ΔRa	ΔA
NO	NO	SI	NO	NO	30	34	0.0826	0.8824		
SI	NO	SI	NO	NO	68	113	0.1873	0.6018	126.67%	232.35%
NO	SI	SI	NO	NO	74	211	0.2039	0.3507	146.67%	520.59%
SI	SI	SI	NO	NO	103	414	0.2837	0.2488	243.33%	1117.65%

Nota: Se consideró $|Ra_0| = 30$ $|A_0| = 34$

Tabla 26. Análisis de resultados obtenidos al considerar el semi-eje de comparación *par semi-igual y par semi-nulo* y *20 rutas de alineamiento.*

De la observación de estas tablas se desprenden los siguientes análisis.

4.5.3.2.1 Análisis del semi-eje de comparación: Rutas múltiples

La alternativa combinada de par semi-igual y par semi-nulo incrementa el desempeño del algoritmo que contempla la alternativa de rutas múltiples, incrementando el índice *recall* de 0.0826 a 0.2837. Sin embargo, este desempeño se ve contrarrestado totalmente por la notable baja, de 0.8824 a 0.2488, del índice *precision*. Puede observarse, además, que el eje de comparación requirió aumentar un 1117% el número de pares-vinculados para aumentar los pares semánticos identificados en un 243%, lo cual es un indicador de la baja eficiencia de este eje.

4.5.3.2.2 Análisis de las alternativas respecto al semi-eje de comparación rutas múltiples

La inclusión de las alternativas en el eje de comparación muestra un incremento ligero en el índice *recall*, pasando de 0.28 a 0.30 en el mejor de los casos. Por su parte, se observa que el índice *precision* aumentó de 0.24 a 0.27. El número de pares-vinculados identificados disminuyó un 4.8% respecto a los identificados por el eje de comparación.

4.5.3.3 Conclusión del eje de comparación par semi-igual y par semi-nulo

La aplicación combinada de las alternativas de par semi-igual y par semi-nulo al *algoritmo básico (sin considerar rutas múltiples) ofrece los mejores resultados posibles* para este eje de comparación. Por el contrario, al aplicar rutas múltiples, estas dos alternativas juntas disminuyen notablemente su desempeño. Por otro lado, la aplicación de las alternativas de rutas múltiples, modificación de costos e intercambio de palabras ofrecen ligeras mejoras respecto al semi-eje de comparación de rutas múltiples. Debido al pobre e incluso mal desempeño del semi-eje de rutas múltiples, esta variación debe ser en principio desechada.

4.6 Recapitulación

En este capítulo se analizaron los resultados obtenidos de la aplicación de cinco alternativas: rutas múltiples (considerando 20 rutas alternativas), par semi-igual, par semi-nulo, modificación de costos e intercambio de palabras. Las alternativas se

TESIS CON
FALLA DE ORIGEN

analizaron tanto individualmente como en combinación. Para ello se realizaron un total de 14 *pruebas, divididas en dos grandes grupos: con una sola ruta y rutas múltiples (20* rutas posibles). Las pruebas se realizaron sobre un corpus en área de metrología en el idioma inglés, cuyos resultados detallados se muestran en los apéndices.

La conclusión principal de estos análisis es que la aplicación de rutas múltiples, modificación de costos e intercambio de palabras no ofrecen beneficios notables en el desempeño del algoritmo.

Por el contrario, 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 0.12 al 0.28, 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.*

En el capítulo siguiente se presentan las conclusiones de esta tesis y los posibles trabajos futuros para afinar el algoritmo

5 CONCLUSIONES Y TRABAJOS FUTUROS

Como resultado del desarrollo de esta tesis se han generado una serie de productos:

- ✓ Un sistema en Web que permite la experimentación con diferentes corpus
- ✓ Un conjunto de 16 heurísticas alternativas y 3 mejoras a la interfaz hombre-computadora
- ✓ El desarrollo y evaluación cuantitativa de cuatro de las 16 líneas propuestas
- ✓ La elaboración de un algoritmo de alineamiento semántico que mejora el desempeño del algoritmo básico

A continuación se presenta un resumen de los problemas y logros obtenidos.

5.1 Visión general del trabajo

En esta tesis se presentó el funcionamiento del algoritmo básico de agrupamiento semántico. En él se introducen los conceptos de par igual, par nulo, par correspondiente, par-vinculado y par-semántico. El algoritmo parte en 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 (simplemente conjuntos de palabras que pueden relacionarse sinonímicamente).

Del análisis de los resultados obtenidos por el algoritmo básico y de un riguroso estudio sobre éste, se han desprendido una serie de observaciones que derivan en un conjunto de alternativas para mejorar el número de pares-semánticos identificados en el algoritmo. Las observaciones dan pie a 19 líneas de trabajo, divididas en dos grandes grupos: Heurísticas alternativas (16 líneas) y modificaciones a la interfaz hombre-computadora (3 posibilidades).

En esta tesis, y con el fin de mejorar el número de pares-vinculados e identificados correctamente, se ha trabajado en cuatro líneas: pares semi-iguales y semi-nulos, intercambio de palabras, modificación de costos y rutas múltiples de alineamiento. Se desarrollaron 6 algoritmos, los cuales se integraron en lo que denominamos el algoritmo flexibilizado de alineamiento semántico y se implantaron en un sistema que puede ser consultado en la página http://iling.iingen.unam.mx.

5.2 Limitaciones

El algoritmo de alineamiento semántico es un método de comparación de dos definiciones que se basa en la comparación 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 esta tesis, 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 original y de 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*.

Sin embargo 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.

5.3 Líneas de trabajo analizadas

Respecto a las Heurísticas alternativas analizadas se puede comentar que:

A. Si bien la inversión de dos palabras consecutivas ha sido considerada por otros autores, en esta tesis se propuso el algoritmo de intercambio conjuntivo de palabras, algoritmo que fácilmente puede ser generalizado a otro tipo de conectores.

B. La inclusión de rutas múltiples plantea el problema de que los posibles alineamientos crecen rápidamente en función de la diferencia de las longitudes de las definiciones. Sin embargo, al incluir el costo negativo en el alineamiento de pares de palabras iguales se limita significativamente este crecimiento, obteniéndose alineamientos de mejor calidad. Cabe mencionar que los costos negativos como valores asignados a las diferentes operaciones de transformación no han sido documentados en la literatura.

C. La introducción del concepto de pares semi-nulos y pares semi-iguales mejora la identificación de pares-semánticos.

Con el sistema desarrollado se realizaron corridas para varios conjuntos de definiciones, correspondientes a diferentes áreas temáticas. En un análisis preliminar se observó que el número de pares-vinculados aumenta no sólo en cantidad, sino que también en calidad. Sin embargo, y con el fin de valorar las modificaciones propuestas, se efectuó una evaluación de estas alternativas. El proceso de evaluación de los resultados buscó obtener elementos cuantitativos más que cualitativos para establecer las bondades del algoritmo flexibilizado de alineamiento semántico con respecto al algoritmo básico.

La evaluación de resultados, tanto del algoritmo básico como del algoritmo flexibilizado, se realizó con base en los pares-vinculados y los pares-semánticos identificados. Las pruebas se realizaron sobre un corpus en área de metrología en el idioma inglés, debido a que este fue el idioma en que se desarrolló el algoritmo básico de alineamiento semántico. Los resultados detallados se muestran en los apéndices.

5.4 Aportaciones

Una de las primeras aportaciones de esta tesis consiste en la sistematización de los diccionarios terminológicos; al incluir la tecnología de bases de datos, se facilita de manera importante la evaluación de las diferentes heurísticas además de el algoritmo de alineamiento semántico puede aplicarse a los diccionarios ya disponibles o a nuevos diccionarios sin la necesidad de efectuar cambios en el sistema desarrollado.

Además, en el sistema se emplearon las técnicas de Internet, de modo que ya se encuentra disponible para su consulta en la página del Grupo de Ingeniería Lingüística del Instituto de Ingeniería: *http://iling.iingen.unam.mx* . Los corpus terminológicos disponibles a la fecha son: metrología (inglés), lingüística (español), mecánica clásica (español), desastres (inglés y español).

En esta tesis se introdujo el concepto de par semi nulo y par semi igual. La premisa básica detrás de estos conceptos es que las palabras contenidas en la lista de palabras irrelevantes (stop list) no aportan información relevante durante el proceso de identificación de pares semánticos; entonces, los pares nulos que agrupen palabras de esta lista, y sólo de esta lista, pueden ser considerados pares iguales, ya que alinear las palabras irrelevantes con la cadena vacía (ε) puede ser equivalente, en este contexto, a insertar la palabra involucrada en lugar de ε. Por su parte, los pares correspondientes que agrupen palabras de esta lista, y sólo de esta lista, pueden ser considerados pares iguales, bajo la premisa de que alinear dos palabras irrelevantes es equivalente, en este contexto, a sustituir una de las palabras por la otra, sin que el significado de las definiciones involucradas se alteren sensiblemente.

En la literatura se han introducido el intercambio de palabras como una operación válida, ya sean palabras consecutivas o no consecutivas; sin embargo, no se ha reportado el intercambio que considere dos palabras relacionadas a través de una conjunción (*y, o*). Con base en el algoritmo propuesto por Lowrence & Wagner, en esta tesis se desarrolló un algoritmo que considera como una operación válida, para el proceso de alineamiento, el intercambio de palabras conectadas por una conjunción.

El alineamiento de definiciones cuyo costo total de las operaciones involucradas sea mínimo, no es único. En esta tesis se introdujo la posibilidad de considerar los posibles alineamientos para un mismo par de definiciones. El algoritmo empleado se desprende de la teoría relacionada con la programación dinámica. La consideración de rutas múltiples de alineamiento no modifica la manera en que se evalúan los pares-vinculados, más bien aporta un mayor número de alineamientos. El algoritmo básico sólo evalúa una sola posibilidad sin haber establecido una justificación cuantitativa de este hecho.

Al considerar rutas múltiples, se observó que el número de alineamientos crecía rápidamente dependiendo de dos factores: la diferencia en el número de palabras utilizadas en cada definición y el número de palabras iguales utilizadas en ambas definiciones. Además, un alineamiento que tenga mayor número de pares iguales es

mejor para la identificación de pares semánticos. Con el fin de reducir el número de alternativas y favorecer el alineamiento de palabras iguales se buscó beneficiar la operación de igualdad respecto a las demás. Para ello se encontró que una posibilidad era asignar un costo negativo a la operación de igualdad: de esta manera, los costos totales se ven disminuidos con cada asignación por igualdad. Como el algoritmo de alineamiento busca minimizar los costos, de manera natural, con este cambio en los costos, el algoritmo elige rutas de alineamiento que utilicen el mayor numero de asignaciones iguales, disminuyendo notablemente las posibles rutas.

Con todas las alternativas, se elaboró lo que hemos denominado el algoritmo *flexibilizado de alineamiento semántico*. Este algoritmo tiene la característica de que permite al usuario elegir de entre las posibles alternativas cuáles quiere considerar en un momento dado. La evaluación de los resultados obtenidos por este algoritmo y por el algoritmo básico permitió contrastar los resultados y determinar cuál o cuáles alternativas eran las que mejores beneficios ofrecían.

5.5 Resultados esperados y obtenidos

La expectativa respecto al desempeño de las diferentes alternativas, indicaba que el algoritmo de rutas múltiples junto con modificación de costos y pares semi-iguales y semi-nulos ofrecerían los mejores resultados posibles. Con el fin de establecer la certeza de las observaciones cualitativas, se evaluó el desempeño de las diferentes alternativas.

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 semi-nulos, pares semi-iguales y su combinación proporcionan la mayor cantidad de pares semánticos, con una proporción muy alta de pares-vinculados.

Las alternativas de intercambio de palabras (disyuntivo o no), rutas múltiples y modificación de costos no ofrecieron los resultados esperados, pues aumentan fuertemente el número de pares-vinculados sin incrementar de manera importante el número de pares semánticos. Sin embargo, no deben abandonarse estas líneas hasta asegurar que no es posible mejorar el desempeño de estas alternativas.

En el caso de rutas múltiples, la alternativa sólo ofrece diferentes alineamientos, pero en ningún caso modifica la manera en que a partir de éstos se identifican lo pares-vinculados. Es de esperarse que si mejora la manera en que se identifican los pares-vinculados, los resultados obtenidos por la alternativa de rutas múltiple mejore también.

TESIS CON
FALLA DE ORIGEN

ESTA TESIS NO SALE
DE LA BIBLIOTECA

5.6 Trabajos futuros

En esta tesis 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 esta tesis 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.

Finalmente, cabe mencionar que las 12 Heurísticas restantes que se plantearon en esta tesis abren, por sí solas, la posibilidad de crecimiento y desarrollo en diversas áreas, tales como lingüística computacional, procesamiento de lenguaje natural, semántica y lenguajes formales, etc., a la vez que permitirá afinar el método de alineamiento semántico. Intuitivamente, y sujeto a desarrollarse y evaluarse, la identificación de la categoría gramatical de cada palabra promete una mejora sustancial en cuanto a la evaluación de los pares-vinculados, para las alternativas propuestas de pares semi-nulos, pares semi-iguales, intercambio de palabras y rutas múltiples.

REFERENCIAS

[AAL1997] Amir A., Auman Y., Landau G., Lewenstein M., & Lewenstein N., (1997). "Pattern matching with swaps". In Proc. FOCS'97. pp. 144-153.

[BeD1962] Bellman R., Dreyfus S. (1962) Applied Dynamic Programming. Princeton University Press, Princeton NJ.

[BaR1999] Baeza-Yates R., Ribeiro-Neto B. (1999) Modern Information Retrieval. ACM press, Addison Wesley.

[CED1994] Collins English dictionary. (1994). Glasgow: Harper Collins Publishers.

[EDE1986] Diccionario del estudiante, Salvat Multimedia, (1996). Salvat Editores.

[DCO1983] Diccionario Conciso Océano, (1983), Editorial Océano.

[DES1996] Diccionario enciclopédico Salvat Multimedia, (1996). Salvat Editores.

[Fra1992] Frakes W. B., (1992). "Stemming algorithms". In Information retrieval: Data *Structures & Algorithms*. W.B. *Frakes and R.Baeza-Yates (eds)*. New Jersey. Prentice Hall.

[GDL1996] Gran diccionario de la Lengua Española, Larousse, edición electrónica. (1996) Editorial Larousse Planeta.

[Gec1976] Geckeler, H (1976). Semántica estructural, Madrid. Gredos.

[Hat1973] Haton J. P. (1973). Contribution à l'Analyse, Paramétrisation et la Reconnaissance Automatique de la Parole, Thèse de doctorat d'état, Université de Nancy, France.

[Hir1975] Hirschberg, D. S. (1975). A linear Space algorithm for computing maximal common subsequences. Communication of the ACM. Vol 18(6). pp. 341-343.

[LKP1997] Lee J., Kim D., Park K., Cho Y. (1997). Efficient algorithms for approximate string matching with swaps. In Proc. CPM'97. LNCS 1264, Springer-Verlag. pp 28-39.

[Lev1965] Levenshtein, V.I. (1965). Binary codes capable of correcting spurious insertions and deletions of ones. Problems of information Transmission, (1). pp. 8-17.

[Lev1966] Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics - Doklady , 10 (8). pp 707-710.

[LoW1975] Lowrence R., Wagner R. A., (1975). An extension of the string-to-string correction problem. Journal of ACM, Vol. 22 (2). pp 177-183.

[MMol1996] María Moliner, (1996) Diccionario de uso del español, edición electrónica, versión 1.0., Editorial Gredos. S.A.

[NeW1970] Needleman S. B., Wunsch C. D. (1970). A general method applicable to search for similarities in amino-acid sequence of two proteins. Journal of Molecular Biology, Vol. 48. pp 443-453.

[OED1994] Oxford English dictionary. (1994). Oxford: Oxford University Press and Rotterdam: Software B.V.

[Por1980] Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14(3). pp 130-137.

[RAE1992] Diccionario de la Real Academia de la Lengua Española. 1992

[RCW1973] Reitchert T. A., Cohen D. N., Wong A. K. C. (1973). An application of information theory to genetic mutations and matching of polypeptide sequences. Journal of Theoretical Biology, Vol 42. pp. 245-261.

[SaC1970] Sakoe H., Chiba S. (1970). A similarity evaluation of speech patterns by dynamic programming. (Japanese) Institute of electronic Communications Engineering of Japan, July 1970. pp 136.

[SaC1971] Sakoe H., Chiba S. (1971). A dynamic programming approach to continuos speech recognition. Proceedings of the international Congress of Acoustics, Budapest, Hungary, paper 20 C 13.

[San1972] Sankoff D. (1972) Matching sequence under deletion-insertion constraints. Proceedings of National Academy of Sciences of the USA, Vol. 69. pp 4-6

[SaK1983] Sankroff D., Kruskall J. B. (eds.) (1983) Time Wraps, String Edits, and Macromolecules: the Theory and Practice of sequence Comparations, Addison Wesley, Reading, MA.

[Sal1968] Salton, G. 1968. Automatic information organization and retrieval. New York: McGraw Hill.

[Sie1999] Sierra G. (1999). Design of a concept-oriented tool for terminology. PhD Thesis University of Manchester, Institute of Science and Technology.

[SiM2000a] Sierra G. & McNaught J., (2000) "Design of an onomasiological search system: A concept-oriented tool for terminology". Terminology. Vol. 6 (1).

TESIS CON
FALLA DE ORIGEN

[SiM2000b] Sierra G. & McNaught J., (2000), "Extracting semantic clusters from MRD for an onomasiological search dictionary". International Journal of Lexicography. Vol. 13 (4)

[Ukk1985] Ukkonen, E. (1985). Algorithms for approximate string matching. Information and control, Vol. 64. pp 100-118.

[VeZ1970] Velichko V. M., Zagoruyko N. G. (1970). Automatic Recognition of 200 words . International Journal of Man-Machine Studies, Vol. 2. pp 223-234.

[Vin1968] Vintsyuk T.K. (1968) Speech discrimination by dynamic programming. Cybernetics, Vol 4 (1). pp. 52-57

[WaF1974] Wagner R. A., Fisher M. J. (1974). The string-to-string correction problem Journal of the ACM, Vol. 21(1). Pp. 168-173

TESIS CON
FALLA DE ORIGEN

APÉNDICE 1. IDENTIFICACIÓN MANUAL DE PARES SEMÁNTICOS

A continuación se presentan los pares semánticos identificados manualmente a partir del diccionario de términos de metrología en el idioma inglés. Esta identificación la realizó un traductor especializado, para ello tomo las definiciones por pares (cuidando que fueran siempre de un diferente autor –y así evitar considerar acepciones de un mismo término–).

Palabra 1	Palabra 2
8 quarts	peck measure
a person	one who
accurate	precise
acetic acid	vinegar
adjustable	regulated
air capacity	breathing power
aircraft	moving body
aircraft	rocket
altitude	elevations
amount	quantity
amount	value
amount	strengths
amount of salt	salinity
angle	inclination
angle with the horizontal	inclination
apparatus	instrument
apparatus	machine
apparatus	spectrometer
arc	quarter circle
arms	legs
ascertaining	recording
assist	guide
atmospheric conditions	weather
automatically	continuous
axis	spindle
bearing	direction
binocular	two eyes
breathing	respiration
celestial	heavenly
celestial body	heavenly body
circular scale	graduated circle
clicking sound	beats
concentration	amount
concentration	intensity
concentration	strength
condition	phenomena
conditions	variations
consisting	composed
constricted	narrow

Palabra 1	Palabra 2
container	vessel
counting	detecting
counting	records
crash	accident
crystal structure	crystallography
crystals	mineral
current	stream
current	electricity
curvature	sphericality
chambers	vessel
changes	fluctuations
changes	variations
changes in dimension	dilatation
changes in dimension	expansion
charged	electrified
charged bodies	electrified body
checks	monitoring
day	sunlight
declination	variation
decreases	falls
deflection	torque
degrees	amounts
density	weight
detect	indicates
detect	recording
detecting	ascertaining
detecting	keeping count
detecting	measuring
detecting	recording
detecting	displaying
determination	measuring
determine	gauging
determine	indicating
determine	judging
determine	measuring
determines	registering
determining	ascertaining
determining	measuring
device	apparatus

Palabra 1	Palabra 2
device	contrivance
device	indicator
device	instrument
device	machine
device	meter
device	piece
device	spark counter
device	system
device	gauge
device	photocell
devices	instrument
dilatation	expansion
direction	angles
drawing	describing
drawing	forming
electric	galvanic
electric current	electricity
electric current	galvanic electricity
electric charge	electricity
electrical activity of the heart	electrocardiograms
elevation	inclination
employing	operating
energy	heat
equipment	installation
equipment	system
escapes	rushes out
esp	usually
especially	usually
estimating	finding
exhaled	breathed out
exhibiting	measuring
exit pipes	narrow aperture
field	limits
finding	determining
fitted	affixed
fluctuating	variations
fluctuation	variation
fluid	water
fluid	wind
fluid	substance
fluid level	depth of liquid
follow	track
force	strength
forming	production
gas	vapour
gasholder	vessel for holding gas
giving	determining

Palabra 1	Palabra 2
graduated scale	rule
having	consisting
heat radiation	radiant heat
height	altitude
how long	the time
Incas	ancient Peruvians
inclination	direction
inclination	slope
inclination	slopes
indicate	demonstrate
indicate	measuring
indicate	registering
indicate	show
indicates	recording
indicates	mark
indicates	marking
indicates	measure
indicates	registering
indicates	registers
indicating	appoints
indicating	fixes
indicating	measuring
indicating	registering
indicating	showing
indicator	gauge
informal	colloquial
inhaled	breathed in
inserted	pressed
inserted into	pressed into
instrument	contrivance
instrument	indicator
instrument	machine
instrument	one
instrument	particle counter
instrument	astrolabe
instrument	toy
instrument	hydrometer
intensity	amount
intensity	force
intervals	times
instrument	counter
investigating	testing
joined	connected
keeps a continuous record	monitoring
keyboard	typewriting
lead plumb	ball of lead
lead plumb	piece of lead
lead plumb	weight

Palabra 1	Palabra 2
legs hinged together	bowed legs
length	distance
length	line
level	depth
light	lightweight
line	string
linked	connected
liquid	liquor
magnetic field	magnetism
magnetized	magnetic
magnitude	quantity
manometer	barometer
mass	weight
mass	density
mass	weight
measure	estimating
measure	take
measurement	determining
measurement	location
measurement	measure
measures	determining
measures	indicates
measures	indicating
measures	sounds
measuring	ascertaining
measuring	asserting
measuring	counting
measuring	estimating
measuring	indicates
measuring	making measurements
measuring	reads
measuring	registering
measuring	surveying
measuring	taking
measuring	testing
measuring	determines
measuring angles	making angle measurements
mechanism	device
meter	contrivance
meter	instrument
method	procedure
method	system
minerals	rocks
moisture	water
monitor	recording
monitoring	controlling

Palabra 1	Palabra 2
mounted	fixed
mounted	flown
neck	channel
observed	sighted
observing	detecting
observing	viewing
observing	investigating
obtaining	viewing
one	single
one quarter	fourth part
operate	activating
optical density	photographic density
order	sequence
pair	two
particles	events
percentage	amount
percentage	proportion
person	one who
photographic	visual
photons	events
pick up	detecting
piece of equipment	device
piece of equipment	instrument
pivot	movable joint
pivoted	mounted
polariscope	instrument
power	energy
predetermined	set
predicts	foretells
pressure	barometric
pressure	force
pressure	tension
producing	displaying
quality	intensity
radar beacons	ground stations
radiant energy	heat
radiation	rays emitted
radiofrequency	radio
rate	force
rate	amount
readings	measuring
readings	recording
recording	indicating
recording	analyze
recording	measure
recording	measuring
recording	registering
recording	registers

TESIS CON
FALLA DE ORIGEN

Palabra 1	Palabra 2
recording	taking
recording	tracing
recording	counting
records	measuring
records	registering
records	collecting
records	keeping count
reflect	reverberates
registers	measuring
relative density	specific gravity
removing	draw out
removing	extracting
resonator	soundboard
revolutions	circular arcs
right angles	perpendiculars
rocket	moving body
room	confined space
rotate	moving
scale	graduated attachment
sea	water
self recording	records automatically
sensitive	sensitiveness
sensitivity	sensitiveness
sensitivity	intensity
several	a number of
short range radar	secondary radar
showing	detecting
shows	illustrating
signals	echoes
signals	waves
sky	atmosphere
slope	angle
small	minute
small	short
soluble	dissolved
sound	tones
space	sky
specific	set
spectrometer	instrument
spectroscope	instrument
speed	force
speed	rate of motion
speed	motion
speed	velocity
spherical	closed
spirit level	levelling instrument
star	celestial body
stars	celestial bodies

Palabra 1	Palabra 2
station	establishment
stick	rigid
storing	records
strength	amount
string	cord
string	line
supersaturated vapour	water vapour
surveying	surveyors
surveying instrument	astrolabe
suspended	poised
swinging	turning
system	navigator
tank	vessel
tape	line
taxi	cab
telescope	astronomical instrument
telescope	instrument
tempo	time
tests	proves
textile	cloth
theodolite	graduated instrument
theodolite	instrument
thermometer	temperature measuring device
thermometer	device
thread	filament
tilt	obliquity
timepiece	instrument for measuring time
tracing	observing
trademark	proprietary name
transit	passage
traveled	traversed
trickle	runs in
upstream	in opposition to the fluid flow
vane	plate
vanes	arms
variation	fluctuations
vehicle	automobile
vehicle	device
vehicle	instrument
vertical	depth
vinegar	acid
volumes	quantities
walking	on foot
water	liquid

Palabra 1	Palabra 2
water	moist
wavelength	frequency

Palabra 1	Palabra 2
weight	bob
weight	plummet

TESIS CON
FALLA DE ORIGEN

APÉNDICE 2. PRUEBAS UTILIZANDO UNA SOLA RUTA DE ALINEACIÓN

A continuación se presentan los resultados obtenidos al aplicar el algoritmo modificado de alineamiento semántico para las pruebas 1 a 9 descritas en el capítulo cinco de esta tesis. Las primeras dos columnas corresponden al par semántico generado por el algoritmo mientras que la tercera indica si este par forma parte de aquellos identificados manualmente por el traductor.

Prueba 1. Algoritmo básico

Palabra 1	Palabra 2	Identificado manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
apparatus	instrument	si
ascertaining	determining	si
ascertaining	measuring	si
day	sunlight	si
determining	measuring	si
direction	inclination	si
field	limits	si
heavenly	celestial	si
indicating	measuring	si
location	measurement	si
measuring	recording	si
measuring	testing	si
observing	tracing	si

Palabra 1	Palabra 2	Identificado manualmente
precise	accurate	si
proportion	percentage	si
radiofrequency	radio	si
recording	analyze	si
set	specific	si
strength	amount	si
strength	concentration	si
system	method	si
taking	measuring	si
telescope	instrument	si
turning	swinging	si
variations	conditions	si
wavelength	frequency	si
weight	mass	si
hyperbolic	radio	no
salinity	amount	no

Prueba 2. Una sola ruta y considerando pares semi iguales

Palabra 1	Palabra 2	Identificado manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
amount	strengths	si
apparatus	device	si
apparatus	instrument	si
ascertaining	detecting	si
ascertaining	*determining*	*si*
ascertaining	measuring	si
closed	spherical	si
counting	detecting	si
counting	records	si
day	sunlight	si
determining	measuring	si

Palabra 1	Palabra 2	Identificado manualmente
device	instrument	si
direction	angles	si
direction	inclination	si
employing	operating	si
estimating	measuring	si
field	limits	si
fluid	substance	si
force	*rate*	*si*
force	speed	si
heavenly	celestial	si
inclination	angle	si
inclination	slope	si
indicates	measuring	si
instrument	meter	si

Palabra 1	Palabra 2	Identificado manualmente
line	string	si
location	measurement	si
measuring	detecting	si
measuring	determines	si
measuring	recording	si
measuring	testing	si
minute	small	si
mounted	pivoted	si
observing	tracing	si
precise	accurate	si
proportion	percentage	si
quantity	magnitude	si
radiofrequency	radio	si
readings	measuring	si
recording	analyze	si
recording	counting	si
records	collecting	si
registering	recording	si
sensitivity	intensity	si
set	specific	si
slope	angle	si
spindle	axis	si
strength	amount	si
strength	concentration	si
system	method	si
taking	measuring	si
telescope	instrument	si
tension	pressure	si
turning	swinging	si
usually	esp	si
variations	conditions	si
vessel	container	si
wavelength	frequency	si
weight	mass	si
air	materials	no

Palabra 1	Palabra 2	Identificado manualmente
astronomical	measuring	no
calculated	sound	no
carpenter	device	no
correcting	carried	no
degrees	sensitivity	no
depth	water	no
detecting	producing	no
distances	ascertaining	no
distances	depends	no
distances	readings	no
field	light	no
form	device	no
gage	indicates	no
hyperbolic	radio	no
instruments	surveying	no
means	manometer	no
measuring	producing	no
objects	velocity	no
passage	telescope	no
point	pressure	no
presence	position	no
quality	amount	no
quality	ascertaining	no
readings	ascertaining	no
readings	instrument	no
records	making	no
rushes	mounted	no
salinity	amount	no
square	testing	no
substances	source	no
time	reticle	no
triggers	distance	no
tube	reduce	no
value	nitrogen	no

Prueba 3. Una sola ruta y considerando pares semi-nulos

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
amount	strengths	si
apparatus	instrument	si
ascertaining	determining	si

Palabra 1	Palabra 2	Identificado Manualmente
ascertaining	measuring	si
day	sunlight	si
determining	finding	si
determining	measuring	si
direction	inclination	si
elevation	inclination	si

Palabra 1	Palabra 2	Identificado Manualmente
energy	power	si
exhibiting	measuring	si
field	limits	si
guide	assist	si
heat	energy	si
heavenly	celestial	si
indicating	measuring	si
location	measurement	si
measuring	recording	si
measuring	testing	si
observing	investigating	si
observing	tracing	si
obtaining	viewing	si
phenomena	condition	si
precise	accurate	si
proportion	percentage	si
radiofrequency	radio	si
recording	analyze	si
set	specific	si
showing	detecting	si
strength	amount	si
strength	concentration	si
system	method	si
take	measure	si
taking	measuring	si
telescope	instrument	si
turning	swinging	si
variation	fluctuations	si
variations	conditions	si
vertical	depth	si

Palabra 1	Palabra 2	Identificado Manualmente
visual	photographic	si
wavelength	frequency	si
weight	mass	si
air	atmosphere	no
automatically	making	no
buildings	aerofoils	no
compass	motór	no
consumed	passed	no
drop	carburettor	no
form	position	no
horizon	observer	no
hyperbolic	radio	no
instrument	informal	no
keeping	detecting	no
kinds	instruments	no
measurement	escape	no
mech	nautical	no
motion	variation	no
name	number	no
need	faces	no
phenomena	features	no
quantity	presence	no
salinity	amount	no
sixth	sighting	no
speed	number	no
time	water	no
vapour	alcohol	no
various	variously	no
water	salt	no

Prueba 4. Una sola ruta y modificación de costos

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
amount	strengths	si
apparatus	instrument	si
ascertaining	determining	si
ascertaining	measuring	si
day	sunlight	si
determining	finding	si
determining	measuring	si

Palabra 1	Palabra 2	Identificado Manualmente
direction	inclination	si
elevation	inclination	si
energy	power	si
exhibiting	measuring	si
field	limits	si
guide	assist	si
heat	energy	si
heavenly	celestial	si
indicating	measuring	si
location	measurement	si

TESIS CON
FALLA DE ORIGEN

Palabra 1	Palabra 2	Identificado Manualmente
measuring	recording	si
measuring	testing	si
observing	investigating	si
observing	tracing	si
obtaining	viewing	si
phenomena	condition	si
precise	accurate	si
proportion	percentage	si
radiofrequency	radio	si
recording	analyze	si
set	specific	si
showing	detecting	si
strength	amount	si
strength	concentration	si
system	method	si
take	measure	si
taking	measuring	si
telescope	instrument	si
turning	swinging	si
variation	fluctuations	si
variations	conditions	si
vertical	depth	si
visual	photographic	si
wavelength	frequency	si
weight	mass	si
air	atmosphere	no

Palabra 1	Palabra 2	Identificado Manualmente
automatically	making	no
buildings	aerofoils	no
compass	motor	no
consumed	passed	no
drop	carburettor	no
form	position	no
horizon	observer	no
hyperbolic	radio	no
instrument	informal	no
keeping	detecting	no
kinds	instruments	no
measurement	escape	no
mech	nautical	no
motion	variation	no
name	number	no
need	faces	no
phenomena	features	no
quantity	presence	no
salinity	amount	no
sixth	sighting	no
speed	number	no
time	water	no
vapour	alcohol	no
various	variously	no
water	salt	no

Prueba 5 Una sola ruta e intercambio de palabras

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
apparatus	instrument	si
ascertaining	determining	si
ascertaining	measuring	si
day	sunlight	si
determining	measuring	si
direction	inclination	si
field	limits	si
heavenly	celestial	si
indicating	measuring	si
location	measurement	si
measuring	recording	si
measuring	testing	si

Palabra 1	Palabra 2	Identificado Manualmente
observing	tracing	si
precise	accurate	si
proportion	percentage	si
radiofrequency	radio	si
recording	analyze	si
set	specific	si
strength	amount	si
strength	concentration	si
system	method	si
taking	measuring	si
telescope	instrument	si
turning	swinging	si
variations	conditions	si
wavelength	frequency	si
weight	mass	si

TESIS CON
FALLA DE ORIGEN

TESIS CON
FALLA DE ORIGEN

Palabra 1	Palabra 2	Identificado Manualmente
hyperbolic	radio	no
relative	photometric	no
salinity	amount	no

Prueba 6. Una sola ruta, considerando pares semi iguales y semi nulos

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
amount	strengths	si
apparatus	device	si
apparatus	instrument	si
arms	vanes	si
ascertaining	detecting	si
ascertaining	determining	si
ascertaining	measuring	si
ascertaining	recording	si
cab	taxi	si
closed	spherical	si
contrivance	device	si
counting	detecting	si
counting	records	si
day	sunlight	si
degrees	amounts	si
detecting	displaying	si
determining	finding	si
determining	measuring	si
device	gauge	si
device	instrument	si
direction	angles	si
direction	inclination	si
elevation	inclination	si
employing	operating	si
energy	power	si
estimating	measuring	si
events	particles	si
exhibiting	measuring	si
field	limits	si
fluid	substance	si
force	rate	si
force	speed	si
guide	assist	si
heat	energy	si
heavenly	celestial	si

Palabra 1	Palabra 2	Identificado Manualmente
inclination	angle	si
inclination	slope	si
indicates	measuring	si
instrument	meter	si
line	string	si
location	measurement	si
measuring	determines	si
measuring	recording	si
measuring	registers	si
measuring	testing	si
minute	small	si
monitoring	controlling	si
mounted	pivoted	si
observing	investigating	si
observing	tracing	si
obtaining	viewing	si
phenomena	condition	si
precise	accurate	si
proportion	percentage	si
quantity	magnitude	si
radiofrequency	radio	si
readings	measuring	si
readings	recording	si
recording	analyze	si
recording	counting	si
records	collecting	si
registering	indicating	si
registering	recording	si
registers	indicates	si
sensitivity	intensity	si
set	specific	si
show	indicate	si
showing	detecting	si
slope	angle	si
spindle	axis	si
strength	amount	si
strength	concentration	si
system	method	si

Palabra 1	Palabra 2	Identificado Manualmente
system	navigator	si
take	measure	si
taking	measuring	si
telescope	instrument	si
tension	pressure	si
turning	swinging	si
usually	esp	si
variation	fluctuations	si
variations	conditions	si
vertical	depth	si
vessel	container	si
visual	photographic	si
wavelength	frequency	si
weight	mass	si
air	atmosphere	no
air	materials	no
applying	connected	no
ascertaining	reach	no
assembly	device	no
astronomical	measuring	no
automatically	making	no
buildings	aerofoils	no
calculated	sound	no
carpenter	device	no
centre	determine	no
circle	sextant	no
commonest	study	no
compass	motor	no
consumed	passed	no
correcting	carried	no
current	measured	no
declination	stars	no
degrees	sensitivity	no
depth	water	no
detecting	current	no
device	pair	no
displaying	current	no
distances	readings	no
drop	carburettor	no
drop	measure	no
electrical	sensitive	no
field	light	no
form	device	no
form	position	no
form	variation	no
gage	indicates	no
gage	presence	no

Palabra 1	Palabra 2	Identificado Manualmente
gauge	pair	no
horizon	observer	no
hyperbolic	radio	no
indicates	intensity	no
installation	aircraft	no
instrument	depends	no
instrument	esp	no
instrument	informal	no
instrument	kinds	no
instrument	self	no
instruments	surveying	no
intensity	detecting	no
intensity	measuring	no
kinds	instruments	no
means	manometer	no
measure	presence	no
measurement	escape	no
measuring	producing	no
mech	nautical	no
motion	form	no
motion	variation	no
name	number	no
need	faces	no
objects	velocity	no
open	small	no
passage	telescope	no
phenomena	features	no
point	pressure	no
presence	position	no
process	direction	no
property	formerly	no
provides	recording	no
quality	amount	no
quality	ascertaining	no
readings	ascertaining	no
readings	depends	no
recording	instrument	no
recording	kinds	no
records	making	no
rushes	mounted	no
salinity	amount	no
sea	depth	no
sixth	sighting	no
sounding	measures	no
spec	esp	no
speed	number	no
square	testing	no

TESIS CON
FALLA DE ORIGEN

Palabra 1	Palabra 2	Identificado Manualmente
strength	current	no
substances	source	no
take	finding	no
time	long	no
time	reticle	no
time	water	no
triggers	distance	no
tube	reduce	no

Palabra 1	Palabra 2	Identificado Manualmente
value	nitrogen	no
vapour	alcohol	no
variation	earth	no
various	variously	no
voltage	loop	no
water	salt	no
wind	direction	no

Prueba 7. Una sola ruta considerando pares semi nulos y semi iguales y modificando los costos de las operaciones.

Palabra 1	Palabra 2	Identificado Manualmente
accident	crash	si
amount	concentration	si
amount	intensity	si
amount	rate	si
amount	strengths	si
apparatus	device	si
apparatus	instrument	si
arms	vanes	si
ascertaining	detecting	si
ascertaining	determining	si
ascertaining	measuring	si
cab	taxi	si
closed	spherical	si
counting	detecting	si
day	sunlight	si
degrees	amounts	si
detecting	displaying	si
determining	finding	si
determining	measuring	si
device	gauge	si
device	instrument	si
device	photocell	si
direction	angles	si
direction	inclination	si
echoes	signals	si
elevation	inclination	si
employing	operating	si
energy	power	si
estimating	measuring	si
events	particles	si
exhibiting	measuring	si

Palabra 1	Palabra 2	Identificado Manualmente
field	limits	si
fixes	indicating	si
fluid	substance	si
force	rate	si
force	speed	si
guide	assist	si
heat	energy	si
heavenly	celestial	si
inclination	angle	si
indicates	measuring	si
instrument	meter	si
line	string	si
location	measurement	si
measuring	determines	si
measuring	recording	si
measuring	registers	si
measuring	testing	si
minute	small	si
monitoring	controlling	si
mounted	pivoted	si
observing	investigating	si
observing	tracing	si
obtaining	viewing	si
phenomena	condition	si
precise	accurate	si
pressed	inserted	si
proportion	percentage	si
radiofrequency	radio	si
readings	measuring	si
recording	analyze	si
recording	counting	si

TESIS CON
FALLA DE ORIGEN

Palabra 1	Palabra 2	Identificado Manualmente
records	collecting	si
registering	indicating	si
registering	recording	si
registers	indicates	si
sensitivity	intensity	si
set	specific	si
showing	detecting	si
slope	angle	si
slope	inclination	si
spindle	axis	si
strength	amount	si
strength	concentration	si
system	method	si
system	navigator	si
take	measure	si
taking	measuring	si
telescope	instrument	si
turning	swinging	si
usually	esp	si
variation	fluctuations	si
variations	conditions	si
vehicle	device	si
vertical	depth	si
vessel	container	si
visual	photographic	si
wavelength	frequency	si
weight	mass	si
air	atmosphere	no
air	materials	no
applying	connected	no
ascertaining	reach	no
assembly	device	no
astronomical	*measuring*	*no*
automatically	making	no
aviation	based	no
buildings	aerofoils	no
calculated	sound	no
carpenter	device	no
centre	determine	no
commonest	study	no
compass	motor	no
consumed	passed	no
correcting	carried	no
current	measured	no
declination	stars	no
detecting	current	no
device	pair	no

Palabra 1	Palabra 2	Identificado Manualmente
direction	local	no
direction	room	no
displaying	current	no
distance	number	no
distance	position	no
drop	carburettor	no
drop	measure	no
eg	intervals	no
electrical	sensitive	no
event	cause	no
form	device	no
form	position	no
form	variation	no
gage	indicates	no
gauge	pair	no
horizon	observer	no
hyperbolic	radio	no
installation	aircraft	no
instrument	depends	no
instrument	esp	no
instrument	informal	no
instrument	kinds	no
instrument	self	no
instruments	surveying	no
kinds	instruments	no
means	manometer	no
measurement	escape	no
measuring	producing	no
mech	nautical	no
motion	form	no
motion	variation	no
name	number	no
need	*faces*	*no*
open	small	no
passage	plane	no
phenomena	features	no
point	pressure	no
presence	position	no
process	direction	no
properly	formerly	no
provides	recording	no
quality	amount	no
quantity	electrical	no
readings	ascertaining	no
recording	depends	no
recording	instrument	no
recording	kinds	no

Palabra 1	Palabra 2	Identificado Manualmente
relative	photometric	no
return	presence	no
rushes	mounted	no
salinity	amount	no
sea	depth	no
show	fixes	no
sixth	sighting	no
sounding	measures	no
spec	esp	no
spectrum	radiation	no
speed	number	no
strength	current	no
substances	source	no
succession	revolution	no
surface	bolometer	no
take	finding	no
taking	producing	no

Palabra 1	Palabra 2	Identificado Manualmente
time	long	no
time	punching	no
time	reticle	no
time	stamping	no
time	water	no
tube	reduce	no
usually	mounted	no
vapour	alcohol	no
variation	earth	no
various	variously	no
vinegar	acetic	no
voltage	loop	no
water	salt	no
wavelengths	detecting	no
wind	direction	no

Prueba 8. *Una sola ruta considerando pares semi iguales y semi nulos e intercambio de* palabras

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
amount	*strengths*	*si*
apparatus	device	si
apparatus	instrument	si
arms	vanes	si
ascertaining	detecting	si
ascertaining	determining	si
ascertaining	measuring	si
ascertaining	recording	si
cab	taxi	si
closed	spherical	si
contrivance	device	si
counting	detecting	si
counting	records	si
day	sunlight	si
degrees	amounts	si
detecting	displaying	si
determining	finding	si
determining	measuring	si
device	gauge	si
device	instrument	si

Palabra 1	Palabra 2	Identificado Manualmente
direction	angles	si
direction	inclination	si
elevation	inclination	si
employing	*operating*	*si*
energy	power	si
estimating	measuring	si
events	particles	si
exhibiting	measuring	si
field	limits	si
fluid	substance	si
force	rate	si
force	speed	si
guide	assist	si
heat	energy	si
heavenly	celestial	si
inclination	angle	si
inclination	slope	si
indicates	measuring	si
instrument	meter	si
line	string	si
location	measurement	si
measuring	determines	si
measuring	recording	si

Palabra 1	Palabra 2	Identificado Manualmente
measuring	registers	si
measuring	testing	si
minute	small	si
monitoring	controlling	si
mounted	pivoted	si
observing	investigating	si
observing	tracing	si
obtaining	viewing	si
phenomena	condition	si
precise	accurate	si
proportion	percentage	si
quantity	magnitude	si
radiofrequency	radio	si
readings	measuring	si
readings	recording	si
recording	analyze	si
recording	counting	si
records	collecting	si
registering	indicating	si
registering	recording	si
registers	indicates	si
sensitivity	intensity	si
set	specific	si
show	indicate	si
showing	detecting	si
slope	angle	si
spindle	axis	si
strength	amount	si
strength	concentration	si
system	method	si
system	navigator	si
take	measure	si
taking	measuring	si
telescope	instrument	si
tension	pressure	si
turning	swinging	si
usually	esp	si
variation	fluctuations	si
variations	conditions	si
vertical	depth	si
vessel	container	si
visual	photographic	si
wavelength	frequency	si
weight	mass	si
accurately	instrument	no
air	atmosphere	no
air	materials	no

Palabra 1	Palabra 2	Identificado Manualmente
applying	connected	no
ascertaining	reach	no
assembly	device	no
astronomical	measuring	no
automatically	making	no
buildings	aerofoils	no
calculated	sound	no
carpenter	device	no
centre	determine	no
circle	sextant	no
commonest	study	no
compass	motor	no
consumed	passed	no
correcting	carried	no
current	measured	no
declination	stars	no
degrees	sensitivity	no
depth	water	no
detecting	current	no
device	pair	no
displaying	current	no
distances	readings	no
drop	carburettor	no
drop	measure	no
electrical	sensitive	no
field	light	no
form	device	no
form	position	no
form	variation	no
gage	indicates	no
gage	presence	no
gauge	pair	no
horizon	observer	no
hyperbolic	radio	no
indicates	intensity	no
installation	aircraft	no
instrument	depends	no
instrument	esp	no
instrument	informal	no
instrument	kinds	no
instrument	self	no
instruments	surveying	no
intensity	detecting	no
intensity	measuring	no
kinds	instruments	no
means	manometer	no
measure	presence	no

Palabra 1	Palabra 2	Identificado Manualmente
measurement	escape	no
measuring	producing	no
mech	nautical	no
motion	form	no
motion	variation	no
name	number	no
need	faces	no
objects	velocity	no
open	small	no
passage	telescope	no
phenomena	features	no
presence	position	no
process	direction	no
properly	formerly	no
provides	recording	no
quality	amount	no
quality	ascertaining	no
readings	ascertaining	no
readings	depends	no
recording	instrument	no
recording	kinds	no
records	making	no
rushes	mounted	no

Palabra 1	Palabra 2	Identificado Manualmente
salinity	amount	no
sea	depth	no
sixth	sighting	no
sounding	measures	no
spec	esp	no
speed	number	no
square	testing	no
strength	current	no
substances	source	no
take	finding	no
time	long	no
time	reticle	no
time	water	no
triggers	distance	no
tube	reduce	no
value	nitrogen	no
vapour	alcohol	no
variation	earth	no
various	variously	no
voltage	loop	no
water	salt	no
wind	direction	no

Prueba 9. Una sola ruta considerando pares semi nulos y semi iguales, modificación de costos de las operaciones de edición e intercambio de palabras

Palabra 1	Palabra 2	Identificado Manualmente
accident	*crash*	*si*
amount	concentration	si
amount	intensity	si
amount	rate	si
amount	strengths	si
apparatus	device	si
apparatus	instrument	si
arms	*vanes*	*si*
ascertaining	detecting	si
ascertaining	determining	si
ascertaining	measuring	si
cab	taxi	si
closed	spherical	si
counting	detecting	si
day	*sunlight*	*si*
degrees	amounts	si

Palabra 1	Palabra 2	Identificado Manualmente
detecting	*displaying*	*si*
determining	finding	si
determining	measuring	si
device	gauge	si
device	instrument	si
device	photocell	si
direction	angles	si
direction	*inclination*	*si*
echoes	signals	si
elevation	inclination	si
employing	operating	si
energy	power	si
estimating	measuring	si
events	particles	si
exhibiting	*measuring*	*si*
field	limits	si

Palabra 1	Palabra 2	Identificado Manualmente
fixes	indicating	si
fluid	substance	si
force	rate	si
force	speed	si
guide	assist	si
heat	energy	si
heavenly	celestial	si
inclination	angle	si
indicates	measuring	si
instrument	meter	si
line	string	si
location	measurement	si
measuring	determines	si
measuring	recording	si
measuring	registers	si
measuring	testing	si
minute	small	si
monitoring	controlling	si
mounted	pivoted	si
observing	investigating	si
observing	tracing	si
obtaining	viewing	si
phenomena	condition	si
precise	accurate	si
pressed	inserted	si
proportion	percentage	si
radiofrequency	radio	si
readings	measuring	si
recording	analyze	si
recording	counting	si
records	collecting	si
registering	indicating	si
registering	recording	si
registers	indicates	si
sensitivity	intensity	si
set	specific	si
showing	detecting	si
slope	angle	si
slope	inclination	si
spindle	axis	si
strength	amount	si
strength	concentration	si
system	method	si
system	navigator	si
take	measure	si
taking	measuring	si
telescope	instrument	si

Palabra 1	Palabra 2	Identificado Manualmente
turning	swinging	si
usually	esp	si
variation	fluctuations	si
variations	conditions	si
vehicle	device	si
vertical	depth	si
vessel	container	si
visual	photographic	si
wavelength	frequency	si
weight	mass	si
air	atmosphere	no
air	materials	no
applying	connected	no
ascertaining	reach	no
assembly	device	no
astronomical	measuring	no
automatically	making	no
aviation	based	no
buildings	aerofoils	no
calculated	sound	no
carpenter	device	no
centre	determine	no
commonest	study	no
compass	motor	no
consumed	passed	no
correcting	carried	no
current	measured	no
declination	stars	no
detecting	current	no
device	pair	no
direction	local	no
direction	room	no
displaying	current	no
distance	number	no
distance	position	no
drop	carburettor	no
drop	measure	no
eg	intervals	no
electrical	sensitive	no
event	cause	no
form	device	no
form	position	no
form	variation	no
gage	indicates	no
gauge	pair	no
horizon	observer	no
hyperbolic	radio	no

Palabra 1	Palabra 2	Identificado Manualmente
installation	aircraft	no
instrument	depends	no
instrument	esp	no
instrument	informal	no
instrument	kinds	no
instrument	self	no
instruments	surveying	no
kinds	instruments	no
means	manometer	no
measurement	escape	no
measuring	producing	no
mech	nautical	no
motion	form	no
motion	variation	no
name	number	no
need	faces	no
open	small	no
passage	plane	no
phenomena	features	no
point	pressure	no
presence	position	no
process	direction	no
properly	formerly	no
provides	recording	no
quality	amount	no
quantity	electrical	no
readings	ascertaining	no
recording	depends	no
recording	instrument	no
recording	kinds	no
relative	photometric	no
return	presence	no

Palabra 1	Palabra 2	Identificado Manualmente
rushes	mounted	no
salinity	amount	no
sea	depth	no
show	fixes	no
sixth	sighting	no
sounding	measures	no
spec	esp	no
spectrum	radiation	no
speed	number	no
strength	current	no
substances	source	no
succession	revolution	no
surface	bolometer	no
take	finding	no
taking	producing	no
time	long	no
time	punching	no
time	reticle	no
time	stamping	no
time	water	no
tube	reduce	no
usually	mounted	no
vapour	alcohol	no
variation	earth	no
various	variously	no
vinegar	acetic	no
voltage	loop	no
water	salt	no
wavelengths	detecting	no
wind	direction	no

APÉNDICE 3. PRUEBAS CONSIDERANDO HASTA 20 POSIBLES RUTAS DE ALINEACIÓN

A continuación se presentan los resultados obtenidos al aplicar el algoritmo modificado de alineamiento semántico para las pruebas 10 a 18 descritas en el capítulo cinco de esta tesis. Las primeras dos columnas corresponden al par semántico generado por el algoritmo mientras que la tercera indica si este par forma parte de aquellos identificados manualmente por el traductor.

Prueba 10. Rutas múltiples

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
apparatus	instrument	si
ascertaining	determining	si
ascertaining	measuring	si
day	sunlight	si
determining	measuring	si
direction	inclination	si
field	limits	si
heavenly	celestial	si
indicating	measuring	si
location	measurement	si
measuring	recording	si
measuring	testing	si
observing	tracing	si
precise	accurate	si

Palabra 1	Palabra 2	Identificado Manualmente
proportion	percentage	si
radiofrequency	radio	si
recording	analyze	si
set	specific	si
strength	amount	si
strength	concentration	si
system	method	si
taking	measuring	si
telescope	instrument	si
turning	swinging	si
variations	conditions	si
wavelength	frequency	si
weight	mass	si
buildings	aerofoils	no
hyperbolic	radio	no
salinity	amount	no
time	hours	no

Prueba 11. Rutas múltiples y pares semi iguales

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
amount	strengths	si
apparatus	device	si
apparatus	instrument	si
apparatus	spectrometer	si
ascertaining	detecting	si
ascertaining	determining	si
ascertaining	measuring	si
counting	detecting	si
counting	records	si
day	sunlight	si

Palabra 1	Palabra 2	Identificado Manualmente
determining	measures	si
determining	measuring	si
device	instrument	si
direction	inclination	si
estimating	measuring	si
field	limits	si
fluid	substance	si
force	rate	si
force	speed	si
heavenly	celestial	si
inclination	angle	si
inclination	slope	si
indicates	measuring	si

Palabra 1	Palabra 2	Identificado Manualmente
installation	equipment	si
instrument	hydrometer	si
instrument	meter	si
line	string	si
location	measurement	si
measuring	detecting	si
measuring	determine	si
measuring	determines	si
measuring	recording	si
measuring	testing	si
minute	small	si
monitoring	controlling	si
mounted	pivoted	si
observing	tracing	si
polariscope	instrument	si
precise	accurate	si
proportion	percentage	si
radiofrequency	radio	si
readings	measuring	si
recording	analyze	si
recording	counting	si
records	collecting	si
registering	recording	si
registers	indicates	si
registers	measuring	si
registers	recording	si
sensitivity	intensity	si
set	specific	si
slope	angle	si
sounds	measures	si
strength	amount	si
strength	concentration	si
system	method	si
taking	measuring	si
telescope	instrument	si
tension	pressure	si
turning	swinging	si
variations	conditions	si
vessel	container	si
visual	photographic	si
wavelength	frequency	si
weight	mass	si
activating	mechanism	no
aid	examination	no

Palabra 1	Palabra 2	Identificado Manualmente
airfield	airports	no
astronomical	measuring	no
buildings	aerofoils	no
carpenter	device	no
correcting	carried	no
degrees	sensitivity	no
detecting	instrument	no
detecting	irradiation	no
detecting	producing	no
dip	determine	no
distances	ascertaining	no
earth	studying	no
electrical	sensitive	no
employed	fringes	no
form	device	no
gage	indicates	no
hyperbolic	radio	no
instrument	operate	no
instruments	surveying	no
kept	light	no
means	manometer	no
measuring	producing	no
mechanically	various	no
objects	velocity	no
observation	axes	no
presence	position	no
process	causes	no
quality	amount	no
readings	ascertaining	no
readings	distance	no
records	making	no
salinity	amount	no
sounding	anglers	no
speed	number	no
square	testing	no
substances	source	no
time	hours	no
time	long	no
time	reticle	no
transformation	measurement	no
triggers	measured	no
value	nitrogen	no
wind	direction	no

Prueba 12. Rutas múltiples y pares semi nulos

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
amount	strengths	si
apparatus	instrument	si
ascertaining	detecting	si
ascertaining	determining	si
ascertaining	measuring	si
automatically	continuous	si
contrivance	device	si
crystals	mineral	si
day	sunlight	si
determine	measuring	si
determining	finding	si
determining	measures	si
determining	measuring	si
device	gauge	si
device	instrument	si
device	photocell	si
direction	angles	si
direction	inclination	si
elevation	inclination	si
employing	operating	si
energy	power	si
estimating	measuring	si
exhibiting	measuring	si
field	limits	si
fluid	substance	si
force	speed	si
heavenly	celestial	si
height	altitude	si
instrument	hydrometer	si
intensity	quality	si
line	length	si
line	string	si
location	measurement	si
measuring	recording	si
measuring	testing	si
monitoring	controlling	si
mounted	pivoted	si
observing	investigating	si
observing	tracing	si
obtaining	viewing	si
plate	vane	si
polariscope	instrument	si

Palabra 1	Palabra 2	Identificado Manualmente
precise	accurate	si
proportion	percentage	si
radiofrequency	radio	si
recording	analyze	si
registering	recording	si
registers	measuring	si
registers	recording	si
sensitivity	intensity	si
set	specific	si
short	small	si
showing	detecting	si
showing	indicating	si
sounds	measures	si
strength	amount	si
strength	concentration	si
system	method	si
take	measure	si
taking	measuring	si
telescope	instrument	si
tension	pressure	si
testing	investigating	si
turning	swinging	si
usually	esp	si
variation	fluctuations	si
variations	conditions	si
viewing	observing	si
visual	photographic	si
wavelength	frequency	si
weight	mass	si
11th	employed	no
alternating	electric	no
amount	pressure	no
apparatus	reference	no
apparatus	signal	no
ascertaining	action	no
astr	instrument	no
astronomical	optical	no
atmospheric	pressure	no
attached	small	no
automatically	making	no
automatically	registers	no
brass	like	no
buildings	aerofoils	no
bushel	measuring	no
cloth	truncated	no

Palabra 1	Palabra 2	Identificado Manualmente
compass	motor	no
consumed	passed	no
crystals	solutions	no
chest	respiratory	no
degrees	sensitivity	no
density	atmospheric	no
density	kind	no
density	measuring	no
determining	drawing	no
determining	studying	no
device	pair	no
device	reference	no
elevation	vane	no
engine	informal	no
establishment	radio	no
estimating	ascertaining	no
estimation	determining	no
exhibiting	pressure	no
extracting	generally	no
extracting	measuring	no
fitted	aircraft	no
flowing	passed	no
force	arterial	no
foreign	angles	no
form	device	no
form	position	no
form	variation	no
fourth	measuring	no
gas	pressure	no
gauge	graduated	no
gauge	pair	no
generally	measuring	no
graduated	telescope	no
great	increase	no
ground	instrument	no
handle	like	no
hence	pressure	no
horizon	figure	no
hyperbolic	radio	no
index	upper	no
indicating	pressure	no
indices	index	no
indices	refractive	no
instrument	astronomy	no
instrument	depends	no
instrument	informal	no
instrument	physics	no

Palabra 1	Palabra 2	Identificado Manualmente
instruments	type	no
judging	usually	no
keeping	detecting	no
kind	atmospheric	no
line	determine	no
line	plumb	no
measures	keeps	no
measuring	comparing	no
measuring	solution	no
measuring	studying	no
measuring	tape	no
mech	nautical	no
mechanical	storage	no
mechanically	various	no
motion	device	no
motion	form	no
motion	variation	no
movements	pneumogram	no
movements	sphygmogram	no
name	number	no
notes	thing	no
number	esp	no
observing	measuring	no
orig	room	no
passage	indicating	no
passage	plane	no
phenomena	features	no
photographic	material	no
photographic	specimen	no
phototube	sensitive	no
plate	scale	no
poised	centre	no
position	usually	no
presence	deflection	no
presence	position	no
pressure	measuring	no
projection	graduated	no
properly	formerly	no
purpose	small	no
quality	amount	no
quality	magnitude	no
radio	consisting	no
radio	determining	no
reaction	designed	no
registering	magnetic	no
rising	altitudes	no
rule	scale	no

Palabra 1	Palabra 2	Identificado Manualmente
salinity	amount	no
shadow	gnomon	no
soil	surface	no
sonograph	components	no
sonograph	frequencies	no
sonograph	sound	no
sonograph	way	no
soundboard	employed	no
sounding	determine	no
spec	esp	no
specific	earth	no
speed	sun	no
sphere	graduated	no
stamp	measuring	no

Palabra 1	Palabra 2	Identificado Manualmente
tension	moisture	no
tension	pitch	no
time	long	no
time	reticle	no
transformation	determining	no
transformation	measurement	no
tube	height	no
vapour	alcohol	no
vessel	various	no
water	pressure	no
wind	force	no
wood	shaped	no
yarn	content	no

Prueba 13. Rutas múltiples y modificación de costos

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
apparatus	instrument	si
ascertaining	determining	si
ascertaining	measuring	si
day	sunlight	si
determining	measuring	si
direction	inclination	si
field	limits	si
heavenly	celestial	si
indicating	measuring	si
location	measurement	si
measuring	recording	si
measuring	testing	si
observing	tracing	si
precise	accurate	si
proportion	percentage	si

Palabra 1	Palabra 2	Identificado Manualmente
radiofrequency	radio	si
recording	analyze	si
set	specific	si
strength	amount	si
strength	concentration	si
system	method	si
taking	measuring	si
telescope	instrument	si
turning	swinging	si
variations	conditions	si
wavelength	frequency	si
weight	mass	si
buildings	aerofoils	no
hyperbolic	radio	no
relative	photometric	no
salinity	amount	no

Prueba 14. Rutas múltiples e intercambio de palabras

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si

Palabra 1	Palabra 2	Identificado Manualmente
apparatus	instrument	si
ascertaining	determining	si
ascertaining	measuring	si

Palabra 1	Palabra 2	Identificado Manualmente
day	sunlight	si
determining	measuring	si
direction	inclination	si
field	limits	si
heavenly	celestial	si
indicating	measuring	si
location	measurement	si
measuring	recording	si
measuring	testing	si
observing	tracing	si
precise	accurate	si
proportion	percentage	si
radiofrequency	radio	si
recording	analyze	si

Palabra 1	Palabra 2	Identificado Manualmente
set	specific	si
strength	amount	si
strength	concentration	si
system	method	si
taking	measuring	si
telescope	instrument	si
turning	swinging	si
variations	conditions	si
wavelength	frequency	si
weight	mass	si
buildings	aerofoils	no
hyperbolic	radio	no
salinity	amount	no
time	hours	no

Prueba 15. Rutas múltiples y pares semi iguales y semi nulos

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
amount	strengths	si
apparatus	device	si
apparatus	instrument	si
apparatus	spectrometer	si
appoints	indicating	si
ascertaining	detecting	si
ascertaining	determining	si
ascertaining	measuring	si
ascertaining	recording	si
automatically	continuous	si
cab	taxi	si
contrivance	device	si
counting	detecting	si
crystals	mineral	si
day	sunlight	si
degrees	amounts	si
detecting	displaying	si
determine	measuring	si
determining	finding	si
determining	measures	si
determining	measuring	si
device	gauge	si
device	instrument	si
device	photocell	si

Palabra 1	Palabra 2	Identificado Manualmente
direction	angles	si
direction	inclination	si
displaying	producing	si
elevation	inclination	si
employing	operating	si
energy	power	si
estimating	measure	si
estimating	measuring	si
events	particles	si
exhibiting	measuring	si
field	limits	si
fluid	substance	si
force	rate	si
force	speed	si
heavenly	celestial	si
height	altitude	si
inclination	angle	si
inclination	slope	si
indicates	measuring	si
indicating	recording	si
installation	equipment	si
instrument	hydrometer	si
instrument	meter	si
intensity	quality	si
line	length	si
line	string	si
location	measurement	si

Palabra 1	Palabra 2	Identificado Manualmente
measuring	determines	si
measuring	recording	si
measuring	registers	si
measuring	testing	si
minute	small	si
monitoring	controlling	si
mounted	pivoted	si
observing	investigating	si
observing	tracing	si
obtaining	viewing	si
plate	vane	si
polariscope	instrument	si
precise	accurate	si
proportion	percentage	si
radiofrequency	radio	si
readings	measuring	si
readings	recording	si
recording	analyze	si
recording	counting	si
recording	measure	si
recording	registers	si
records	collecting	si
registering	recording	si
registers	indicates	si
sensitivity	intensity	si
set	specific	si
short	small	si
showing	detecting	si
showing	indicating	si
slope	angle	si
sounds	measures	si
strength	amount	si
strength	concentration	si
system	method	si
take	measure	si
taking	measuring	si
telescope	instrument	si
tension	pressure	si
testing	investigating	si
turning	swinging	si
usually	esp	si
variation	fluctuations	si
variations	conditions	si
vehicle	device	si
vessel	container	si
viewing	observing	si
visual	photographic	si

Palabra 1	Palabra 2	Identificado Manualmente
wavelength	frequency	si
weight	mass	si
11th	employed	no
activating	appoints	no
activating	automatically	no
activating	fixes	no
activating	indicates	no
adapted	detect	no
aid	examination	no
air	materials	no
airfield	free	no
alternating	electric	no
alternating	quantity	no
amount	pressure	no
apparatus	reference	no
apparatus	signal	no
appoints	device	no
appoints	instrument	no
appoints	operate	no
ascertaining	action	no
ascertaining	measure	no
ascertaining	reach	no
assembly	device	no
astr	instrument	no
astronomical	measuring	no
astronomical	optical	no
atmospheric	pressure	no
attached	small	no
automatically	making	no
automatically	operate	no
automatically	registers	no
automatically	time	no
beam	instrument	no
beam	measuring	no
beam	reference	no
blood	variations	no
brass	like	no
buildings	aerofoils	no
bushel	measuring	no
business	thing	no
carangeot	faces	no
carpenter	device	no
centre	hinged	no
circular	number	no
cloth	truncated	no
column	pressure	no
compass	motor	no

Palabra 1	Palabra 2	Identificado Manualmente
consisting	usually	no
consumed	passed	no
correct	indicates	no
correcting	carried	no
count	number	no
counting	keeping	no
crystals	solutions	no
current	relative	no
change	calculated	no
chest	respiratory	no
degrees	sensitivity	no
density	atmospheric	no
density	kind	no
density	measuring	no
depth	fluid	no
depth	water	no
detecting	instrument	no
detecting	producing	no
determining	depth	no
determining	drawing	no
device	pair	no
dial	surveyors	no
dip	determine	no
distances	recording	no
earth	inclination	no
earth	intensity	no
earth	studying	no
effect	amount	no
electrical	resistance	no
electrical	sensitive	no
elevation	mounted	no
elevation	vane	no
employed	fringes	no
engine	informal	no
equal	sighting	no
esp	determination	no
esp	measuring	no
establishment	radio	no
estimating	ascertaining	no
estimation	determining	no
expiration	inhaled	no
extracting	form	no
extracting	generally	no
extracting	measuring	no
extracting	surgery	no
eyeball	tension	no
fitted	aircraft	no

Palabra 1	Palabra 2	Identificado Manualmente
fitted	records	no
fixes	appoints	no
fixes	operate	no
flowing	passed	no
force	arterial	no
force	beam	no
force	escape	no
force	measuring	no
force	recording	no
foreign	angles	no
form	device	no
form	measuring	no
form	position	no
form	variation	no
fourth	measuring	no
furnished	consisting	no
gage	indicates	no
gage	presence	no
gas	detecting	no
gas	maintained	no
gas	pressure	no
gauge	graduated	no
gauge	pair	no
gauging	density	no
generally	form	no
generally	measuring	no
graduated	piece	no
graduated	telescope	no
great	increase	no
ground	instrument	no
hand	sighting	no
handle	like	no
hence	pressure	no
horizon	attached	no
horizon	figure	no
horizon	level	no
horizontality	heights	no
hyperbolic	radio	no
illustrating	detection	no
inclination	line	no
index	upper	no
indicates	intensity	no
indicating	aircraft	no
indicating	force	no
indicating	instrument	no
indicating	operate	no
indicating	person	no

Palabra 1	Palabra 2	Identificado Manualmente
indicating	thing	no
indices	index	no
indices	refractive	no
inspiration	air	no
instrument	astronomy	no
instrument	depends	no
instrument	esp	no
instrument	informal	no
instrument	measuring	no
instrument	mechanical	no
instrument	operate	no
instrument	physics	no
instrument	produces	no
instrument	reference	no
instrument	self	no
instrument	tube	no
instruments	surveying	no
instruments	type	no
intensity	gas	no
intensity	maintained	no
intensity	measuring	no
interference	acoustic	no
ionizing	occurrences	no
judging	usually	no
keeping	detecting	no
keeping	records	no
kept	light	no
kind	atmospheric	no
like	gas	no
line	depth	no
line	determine	no
line	plumb	no
lungs	breathing	no
made	indicating	no
main	compressibility	no
marking	music	no
means	esp	no
means	manometer	no
means	tempo	no
measure	presence	no
measures	keeps	no
measuring	comparing	no
measuring	solution	no
measuring	studying	no
measuring	tape	no
mech	naut	no
mech	nautical	no

Palabra 1	Palabra 2	Identificado Manualmente
mechanically	various	no
motion	device	no
motion	form	no
motion	variation	no
movements	pneumogram	no
movements	sphygmogram	no
music	indicating	no
music	instrument	no
name	number	no
naut	nautical	no
number	copies	no
objects	velocity	no
observation	axes	no
observed	esp	no
observing	magnetic	no
observing	measuring	no
orig	room	no
paper	anemometer	no
parallel	telescope	no
passage	indicating	no
passage	keeps	no
passage	plane	no
passage	thing	no
passenger	fare	no
phenomena	features	no
photographic	material	no
photographic	specimen	no
photographic	transmission	no
phototube	sensitive	no
plane	spirit	no
plate	scale	no
point	water	no
poised	centre	no
position	relative	no
presence	deflection	no
presence	position	no
pressure	measuring	no
pressure	movement	no
process	causes	no
process	pulse	no
projection	graduated	no
properly	formerly	no
provides	recording	no
purpose	small	no
quadrant	navigation	no
quality	amount	no
quality	magnitude	no

Palabra 1	Palabra 2	Identificado Manualmente
quantity	electrical	no
radio	determining	no
reaction	designed	no
readings	ascertaining	no
readings	distance	no
recording	copies	no
recording	instrument	no
recording	keeping	no
recording	number	no
records	making	no
registering	spec	no
registering	storage	no
rising	altitudes	no
rule	scale	no
salinity	amount	no
screen	blip	no
sea	depth	no
shadow	gnomon	no
showing	instrument	no
shown	makes	no
sights	scale	no
soil	surface	no
sonograph	components	no
sonograph	frequencies	no
sonograph	sound	no
sonograph	way	no
soundboard	employed	no
sounding	anglers	no
sounding	determine	no
spec	esp	no
specific	earth	no
speed	number	no
speed	sun	no
speed	vehicle	no
sphere	graduated	no
square	testing	no
stamp	measuring	no
substances	source	no
sugars	rotates	no
surgery	form	no

Palabra 1	Palabra 2	Identificado Manualmente
surgery	generally	no
surgery	measuring	no
take	direction	no
take	drawing	no
take	finding	no
taking	producing	no
temperature	automatically	no
temperature	called	no
temperature	coil	no
tension	moisture	no
tension	pitch	no
tests	person	no
thermometric	thermometer	no
time	exact	no
time	fixes	no
time	long	no
time	operate	no
time	passage	no
time	reticle	no
transformation	consisting	no
transformation	force	no
transformation	measurement	no
triggers	measured	no
tube	height	no
turning	arms	no
value	nitrogen	no
vapour	alcohol	no
velocity	etc	no
volume	voltage	no
water	bed	no
water	gas	no
water	pressure	no
weighing	generally	no
wheel	mechanically	no
wind	direction	no
wind	indicating	no
wood	shaped	no
yarn	content	no

Prueba 16. Rutas múltiples con pares semi iguales y semi nulos y modificación de costos.

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
amount	strengths	si
apparatus	device	si
apparatus	instrument	si
apparatus	spectrometer	si
ascertaining	detecting	si
ascertaining	determining	si
ascertaining	measuring	si
ascertaining	recording	si
automatically	continuous	si
cab	taxi	si
contrivance	device	si
counting	detecting	si
crystals	mineral	si
day	sunlight	si
degrees	amounts	si
detecting	displaying	si
determine	measuring	si
determining	finding	si
determining	measures	si
determining	measuring	si
device	gauge	si
device	instrument	si
device	photocell	si
direction	angles	si
direction	inclination	si
displaying	producing	si
echoes	signals	si
elevation	inclination	si
employing	operating	si
energy	power	si
estimating	measure	si
estimating	measuring	si
events	particles	si
exhibiting	measuring	si
field	limits	si
fixes	indicating	si
fluid	substance	si
force	rate	si
force	speed	si
forming	drawing	si
guide	assist	si

Palabra 1	Palabra 2	Identificado Manualmente
heat	energy	si
heavenly	celestial	si
height	altitude	si
inclination	angle	si
inclination	slope	si
indicates	measuring	si
indicating	recording	si
installation	equipment	si
instrument	hydrometer	si
instrument	meter	si
intensity	quality	si
line	length	si
line	string	si
location	measurement	si
measuring	determines	si
measuring	recording	si
measuring	registers	si
measuring	testing	si
minute	small	si
monitoring	controlling	si
mounted	pivoted	si
observing	investigating	si
observing	tracing	si
obtaining	viewing	si
plate	vane	si
polariscope	instrument	si
precise	accurate	si
pressed	inserted	si
proportion	percentage	si
radiofrequency	radio	si
readings	measuring	si
readings	recording	si
recording	analyze	si
recording	measure	si
records	collecting	si
registering	recording	si
registers	indicates	si
registers	recording	si
sensitivity	intensity	si
set	specific	si
short	small	si
show	indicate	si
showing	detecting	si
showing	indicating	si

Palabra 1	Palabra 2	Identificado Manualmente
sky	space	si
slope	angle	si
sounds	measures	si
strength	amount	si
strength	concentration	si
system	method	si
take	measure	si
taking	measuring	si
telescope	instrument	si
testing	investigating	si
turning	swinging	si
usually	esp	si
variation	fluctuations	si
variations	conditions	si
vehicle	device	si
vehicle	instrument	si
vertical	depth	si
vessel	container	si
viewing	observing	si
visual	photographic	si
wavelength	frequency	si
weight	mass	si
11th	employed	no
action	power	no
activating	causes	no
actuated	escape	no
adapted	detect	no
aid	examination	no
airfield	airports	no
alternating	electric	no
alternating	quantity	no
alternations	tilt	no
apparatus	reference	no
apparatus	signal	no
appoints	switch	no
ascertaining	action	no
ascertaining	measure	no
ascertaining	reach	no
assembly	device	no
astronomical	measuring	no
astronomical	optical	no
astronomical	time	no
attached	small	no
automatically	making	no
automatically	registers	no
axis	mounted	no
blood	variations	no

Palabra 1	Palabra 2	Identificado Manualmente
brass	like	no
buildings	aerofoils	no
bushel	measuring	no
business	thing	no
carangeot	faces	no
carpenter	device	no
circular	distance	no
cloth	truncated	no
column	pressure	no
compass	motor	no
consisting	arm	no
consisting	usually	no
consumed	passed	no
containing	tracks	no
correct	indicates	no
correcting	carried	no
count	number	no
counting	keeping	no
current	velocity	no
chest	respiratory	no
degrees	sensitivity	no
density	measuring	no
density	pressure	no
depth	fluid	no
designed	producing	no
designed	spectrum	no
detecting	instrument	no
detecting	producing	no
device	pair	no
dial	surveyors	no
dip	determine	no
direction	room	no
distance	number	no
distance	sent	no
distances	readings	no
distances	recording	no
earth	days	no
earth	slope	no
echoes	target	no
echoes	viewed	no
effect	amount	no
effect	shows	no
electrical	sensitive	no
elevation	axis	no
elevation	mounted	no
elevation	vane	no
employed	fringes	no

Palabra 1	Palabra 2	Identificado Manualmente
engine	informal	no
engine	power	no
esp	determination	no
esp	measuring	no
establishment	radio	no
estimating	ascertaining	no
estimation	determining	no
extracting	form	no
extracting	measuring	no
extracting	surgery	no
fitted	aircraft	no
fixes	regulator	no
flow	upstream	no
flowing	passed	no
force	arterial	no
force	escape	no
force	measuring	no
force	recording	no
foreign	angles	no
form	based	no
form	device	no
form	measuring	no
form	position	no
form	removing	no
form	variation	no
fourth	measuring	no
gage	indicates	no
gas	maintained	no
gas	pressure	no
gauge	graduated	no
gauge	pair	no
gauge	various	no
gauging	density	no
generally	measuring	no
graduated	piece	no
ground	instrument	no
hand	sighting	no
handle	like	no
heat	increase	no
heat	resistance	no
hence	pressure	no
horizon	figure	no
horizon	level	no
hyperbolic	radio	no
illustrating	detection	no
inclination	earth	no
index	upper	no

Palabra 1	Palabra 2	Identificado Manualmente
indicates	intensity	no
indicating	force	no
indicating	instrument	no
indicating	thing	no
instrument	acoustic	no
instrument	astronomy	no
instrument	depends	no
instrument	esp	no
instrument	informal	no
instrument	interference	no
instrument	measuring	no
instrument	mechanical	no
instrument	reference	no
instrument	self	no
instrument	tube	no
instruments	surveying	no
instruments	type	no
intensity	measuring	no
interference	acoustic	no
interference	optical	no
ionizing	occurrences	no
judging	usually	no
keeping	detecting	no
keeping	number	no
keeping	records	no
kept	light	no
kind	density	no
like	gas	no
line	determine	no
line	plumb	no
lungs	breathing	no
made	indicating	no
main	compressibility	no
marking	music	no
means	manometer	no
means	tempo	no
measurement	escape	no
measures	keeps	no
measuring	comparing	no
measuring	producing	no
measuring	solution	no
measuring	studying	no
measuring	tape	no
mech	naut	no
mech	nautical	no
mechanically	various	no
motion	device	no

Palabra 1	Palabra 2	Identificado Manualmente
motion	form	no
motion	variation	no
motor	power	no
movements	pneumogram	no
movements	sphygmogram	no
music	indicating	no
music	instrument	no
name	number	no
naut	nautical	no
night	axis	no
number	producing	no
observation	axes	no
observed	esp	no
observing	magnetic	no
observing	measuring	no
open	small	no
opposition	fluid	no
opposition	open	no
orig	direction	no
orig	room	no
paper	anemometer	no
passage	indicating	no
passage	keeps	no
passage	plane	no
passenger	fare	no
phenomena	features	no
photographic	material	no
photographic	specimen	no
photographic	transmission	no
phototube	sensitive	no
plane	spirit	no
plate	scale	no
point	water	no
pointing	placed	no
poised	centre	no
presence	deflection	no
presence	position	no
pressure	measuring	no
pressure	movement	no
process	pulse	no
projection	graduated	no
properly	formerly	no
provides	recording	no
purpose	small	no
quality	amount	no
quality	magnitude	no
quantity	electrical	no

Palabra 1	Palabra 2	Identificado Manualmente
radio	determining	no
reaction	designed	no
readings	ascertaining	no
recording	instrument	no
registering	spec	no
registering	storage	no
relative	photometric	no
rising	altitudes	no
rule	scale	no
salinity	amount	no
sea	depth	no
sensitive	temperature	no
setting	bullet	no
shadow	gnomon	no
sights	scale	no
slope	line	no
soil	surface	no
sonograph	components	no
sonograph	frequencies	no
sonograph	sound	no
sonograph	way	no
soundboard	employed	no
sounding	anglers	no
sounding	determine	no
spec	esp	no
specific	earth	no
speed	number	no
speed	sun	no
speed	vehicle	no
sphere	graduated	no
square	testing	no
substances	source	no
succession	revolution	no
sugars	rotates	no
surgery	form	no
surgery	measuring	no
take	direction	no
take	finding	no
temperature	coil	no
tension	moisture	no
tension	pitch	no
tests	person	no
threads	arrangement	no
time	exact	no
time	long	no
time	punching	no
time	reticle	no

Palabra 1	Palabra 2	Identificado Manualmente
time	stamping	no
transformation	force	no
transformation	measurement	no
traveled	bicycle	no
triggers	measured	no
tube	height	no
value	nitrogen	no
vapour	alcohol	no
variation	specific	no
various	variously	no
velocity	etc	no
velocity	relative	no
vessel	check	no

Palabra 1	Palabra 2	Identificado Manualmente
volume	voltage	no
water	bed	no
water	gas	no
water	pressure	no
weighing	generally	no
wheel	mechanically	no
wind	direction	no
wind	indicating	no
wood	shaped	no
work	power	no
yarn	content	no

Prueba 17. Rutas múltiples con pares semi iguales y semi nulos e intercambio de palabras

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
amount	strengths	si
apparatus	device	si
apparatus	instrument	si
apparatus	spectrometer	si
appoints	indicating	si
ascertaining	detecting	si
ascertaining	determining	si
ascertaining	measuring	si
ascertaining	recording	si
automatically	continuous	si
cab	taxi	si
contrivance	device	si
counting	detecting	si
crystals	mineral	si
day	sunlight	si
degrees	amounts	si
detecting	displaying	si
determine	measuring	si
determining	finding	si
determining	measures	si
determining	measuring	si
device	gauge	si
device	instrument	si
device	photocell	si

Palabra 1	Palabra 2	Identificado Manualmente
direction	angles	si
direction	inclination	si
displaying	producing	si
elevation	inclination	si
employing	operating	si
energy	power	si
estimating	measure	si
estimating	measuring	si
events	particles	si
exhibiting	measuring	si
field	limits	si
fluid	substance	si
force	rate	si
force	speed	si
heavenly	celestial	si
height	altitude	si
inclination	angle	si
inclination	slope	si
indicates	measuring	si
indicating	recording	si
installation	equipment	si
instrument	hydrometer	si
instrument	meter	si
intensity	quality	si
line	length	si
line	string	si
location	measurement	si

Palabra 1	Palabra 2	Identificado Manualmente
measuring	determines	si
measuring	recording	si
measuring	registers	si
measuring	testing	si
minute	small	si
monitoring	controlling	si
mounted	pivoted	si
observing	investigating	si
observing	tracing	si
obtaining	viewing	si
plate	vane	si
polariscope	instrument	si
precise	accurate	si
proportion	percentage	si
radiofrequency	radio	si
readings	measuring	si
readings	recording	si
recording	analyze	si
recording	counting	si
recording	measure	si
recording	registers	si
records	collecting	si
registering	recording	si
registers	indicates	si
sensitivity	intensity	si
set	specific	si
short	small	si
showing	detecting	si
showing	indicating	si
slope	angle	si
sounds	measures	si
strength	amount	si
strength	concentration	si
system	method	si
take	measure	si
taking	measuring	si
telescope	instrument	si
tension	pressure	si
testing	investigating	si
turning	swinging	si
usually	esp	si
variation	fluctuations	si
variations	conditions	si
vehicle	device	si
vessel	container	si
viewing	observing	si
visual	photographic	si

Palabra 1	Palabra 2	Identificado Manualmente
wavelength	frequency	si
weight	mass	si
11th	employed	no
activating	appoints	no
activating	automatically	no
activating	fixes	no
activating	indicates	no
adapted	detect	no
aid	examination	no
airfield	free	no
alternating	electric	no
alternating	quantity	no
amount	pressure	no
apparatus	reference	no
apparatus	signal	no
appoints	device	no
appoints	instrument	no
appoints	operate	no
ascertaining	action	no
ascertaining	measure	no
ascertaining	reach	no
assembly	device	no
astr	instrument	no
astronomical	measuring	no
astronomical	optical	no
atmospheric	pressure	no
attached	small	no
automatically	making	no
automatically	operate	no
automatically	registers	no
automatically	time	no
beam	instrument	no
beam	measuring	no
beam	reference	no
blood	variations	no
brass	like	no
buildings	aerofoils	no
bushel	measuring	no
business	thing	no
carangeot	faces	no
carpenter	device	no
centre	hinged	no
circular	number	no
cloth	truncated	no
column	pressure	no
compass	motor	no
consisting	usually	no

Palabra 1	Palabra 2	Identificado Manualmente
consumed	passed	no
correct	indicates	no
correcting	carried	no
count	number	no
counting	keeping	no
crystals	solutions	no
current	relative	no
change	calculated	no
chest	respiratory	no
degrees	sensitivity	no
density	atmospheric	no
density	kind	no
density	measuring	no
depth	fluid	no
detecting	instrument	no
detecting	producing	no
determining	depth	no
determining	drawing	no
device	pair	no
dial	surveyors	no
dip	determine	no
distances	recording	no
earth	inclination	no
earth	intensity	no
earth	studying	no
effect	amount	no
electrical	resistance	no
electrical	sensitive	no
elevation	mounted	no
elevation	vane	no
employed	fringes	no
engine	informal	no
equal	sighting	no
esp	determination	no
esp	measuring	no
establishment	radio	no
estimating	ascertaining	no
estimation	determining	no
expiration	inhaled	no
extracting	form	no
extracting	generally	no
extracting	measuring	no
extracting	surgery	no
eyeball	tension	no
fitted	aircraft	no
fitted	records	no
fixes	appoints	no

Palabra 1	Palabra 2	Identificado Manualmente
fixes	operate	no
flowing	passed	no
force	arterial	no
force	beam	no
force	escape	no
force	measuring	no
force	recording	no
foreign	angles	no
form	device	no
form	measuring	no
form	position	no
form	variation	no
fourth	measuring	no
furnished	consisting	no
gage	indicates	no
gage	presence	no
gas	maintained	no
gauge	graduated	no
gauge	pair	no
gauging	density	no
generally	form	no
generally	measuring	no
graduated	piece	no
graduated	telescope	no
great	increase	no
ground	instrument	no
hand	sighting	no
handle	like	no
hence	pressure	no
horizon	attached	no
horizon	figure	no
horizon	level	no
horizontality	heights	no
hyperbolic	radio	no
illustrating	detection	no
inclination	line	no
index	upper	no
indicates	intensity	no
indicating	aircraft	no
indicating	force	no
indicating	instrument	no
indicating	operate	no
indicating	person	no
indicating	thing	no
indices	index	no
indices	refractive	no
inspiration	air	no

Palabra 1	Palabra 2	Identificado Manualmente
instrument	astronomy	no
instrument	depends	no
instrument	esp	no
instrument	informal	no
instrument	measuring	no
instrument	mechanical	no
instrument	operate	no
instrument	physics	no
instrument	produces	no
instrument	reference	no
instrument	self	no
instrument	tube	no
instruments	surveying	no
instruments	type	no
intensity	measuring	no
intensity	pressure	no
interference	acoustic	no
ionizing	occurrences	no
judging	usually	no
keeping	detecting	no
keeping	records	no
kept	light	no
kind	atmospheric	no
like	gas	no
line	depth	no
line	determine	no
line	plumb	no
lungs	breathing	no
made	indicating	no
main	compressibility	no
marking	music	no
means	esp	no
means	manometer	no
means	tempo	no
measure	presence	no
measures	keeps	no
measuring	comparing	no
measuring	solution	no
measuring	studying	no
measuring	tape	no
mech	naut	no
mech	nautical	no
mechanically	various	no
motion	device	no
motion	form	no
motion	variation	no
movements	pneumogram	no
movements	sphygmogram	no
music	indicating	no
music	instrument	no
name	number	no
naut	nautical	no
number	copies	no
objects	velocity	no
observation	axes	no
observed	esp	no
observing	magnetic	no
observing	measuring	no
orig	room	no
paper	anemometer	no
parallel	telescope	no
passage	indicating	no
passage	keeps	no
passage	plane	no
passage	thing	no
passenger	fare	no
phenomena	features	no
photographic	material	no
photographic	specimen	no
photographic	transmission	no
phototube	sensitive	no
plane	spirit	no
plate	scale	no
point	water	no
poised	centre	no
position	relative	no
presence	deflection	no
presence	position	no
pressure	detecting	no
pressure	measuring	no
pressure	movement	no
process	causes	no
process	pulse	no
projection	graduated	no
properly	formerly	no
provides	recording	no
purpose	small	no
quadrant	navigation	no
quality	amount	no
quality	magnitude	no
quantity	electrical	no
radio	determining	no
reaction	designed	no
readings	ascertaining	no

Palabra 1	Palabra 2	Identificado Manualmente
readings	distance	no
recording	copies	no
recording	instrument	no
recording	keeping	no
recording	number	no
records	making	no
registering	spec	no
registering	storage	no
rising	altitudes	no
rule	scale	no
salinity	amount	no
screen	blip	no
sea	depth	no
shadow	gnomon	no
showing	instrument	no
shown	makes	no
sights	scale	no
soil	surface	no
sonograph	components	no
sonograph	frequencies	no
sonograph	sound	no
sonograph	way	no
soundboard	employed	no
sounding	anglers	no
sounding	determine	no
spec	esp	no
specific	earth	no
speed	number	no
speed	sun	no
speed	vehicle	no
sphere	graduated	no
square	testing	no
stamp	measuring	no
substances	source	no
sugars	rotates	no
surgery	form	no
surgery	generally	no
surgery	measuring	no
take	direction	no
take	drawing	no
take	finding	no
taking	producing	no
temperature	automatically	no
temperature	called	no
temperature	coil	no
tension	moisture	no
tension	pitch	no

Palabra 1	Palabra 2	Identificado Manualmente
tests	person	no
thermometric	thermometer	no
time	exact	no
time	fixes	no
time	long	no
time	operate	no
time	passage	no
time	reticle	no
transformation	consisting	no
transformation	force	no
transformation	measurement	no
triggers	measured	no
tube	height	no
turning	arms	no
value	nitrogen	no
vapour	alcohol	no
velocity	etc	no
volume	voltage	no
water	bed	no
water	pressure	no
weighing	generally	no
wheel	mechanically	no
wind	direction	no
wind	indicating	no
wood	shaped	no
yarn	content	no

Prueba 18. Rutas múltiples, pares semi nulos y semi iguales, modificación de costos e intercambio de palabras

Palabra 1	Palabra 2	Identificado Manualmente
amount	concentration	si
amount	intensity	si
amount	rate	si
amount	strengths	si
apparatus	device	si
apparatus	instrument	si
apparatus	spectrometer	si
ascertaining	detecting	si
ascertaining	determining	si
ascertaining	measuring	si
ascertaining	recording	si
automatically	continuous	si
cab	taxi	si
contrivance	device	si
counting	detecting	si
crystals	mineral	si
day	sunlight	si
degrees	amounts	si
detecting	displaying	si
determine	measuring	si
determining	finding	si
determining	measures	si
determining	measuring	si
device	gauge	si
device	instrument	si
device	photocell	si
direction	angles	si
direction	inclination	si
displaying	producing	si
echoes	signals	si
elevation	inclination	si
employing	operating	si
energy	power	si
estimating	measure	si
estimating	measuring	si
events	particles	si
exhibiting	measuring	si
field	limits	si
fixes	indicating	si
fluid	substance	si
force	rate	si
force	speed	si
forming	drawing	si

Palabra 1	Palabra 2	Identificado Manualmente
guide	assist	si
heat	energy	si
heavenly	celestial	si
height	altitude	si
inclination	angle	si
inclination	slope	si
indicates	measuring	si
indicating	recording	si
installation	equipment	si
instrument	hydrometer	si
instrument	meter	si
intensity	quality	si
line	length	si
line	string	si
location	measurement	si
measuring	determines	si
measuring	recording	si
measuring	registers	si
measuring	testing	si
minute	small	si
monitoring	controlling	si
mounted	pivoted	si
observing	investigating	si
observing	tracing	si
obtaining	viewing	si
plate	vane	si
polariscope	instrument	si
precise	accurate	si
pressed	inserted	si
proportion	percentage	si
radiofrequency	radio	si
readings	measuring	si
readings	recording	si
recording	analyze	si
recording	measure	si
records	collecting	si
registering	recording	si
registers	indicates	si
registers	recording	si
sensitivity	intensity	si
set	specific	si
short	small	si
show	indicate	si

Palabra 1	Palabra 2	Identificado Manualmente
showing	detecting	si
showing	indicating	si
sky	space	si
slope	angle	si
sounds	measures	si
strength	amount	si
strength	concentration	si
system	method	si
take	measure	si
taking	measuring	si
telescope	instrument	si
testing	investigating	si
turning	swinging	si
usually	esp	si
variation	fluctuations	si
variations	conditions	si
vehicle	device	si
vehicle	instrument	si
vertical	depth	si
vessel	container	si
viewing	observing	si
visual	photographic	si
wavelength	frequency	si
weight	mass	si
11th	employed	no
action	power	no
activating	causes	no
actuated	escape	no
adapted	detect	no
aid	examination	no
airfield	airports	no
alternating	electric	no
alternating	quantity	no
alternations	tilt	no
apparatus	reference	no
apparatus	signal	no
appoints	switch	no
ascertaining	action	no
ascertaining	measure	no
ascertaining	reach	no
assembly	device	no
astronomical	measuring	no
astronomical	optical	no
astronomical	time	no
attached	small	no
automatically	making	no
automatically	registers	no

Palabra 1	Palabra 2	Identificado Manualmente
axis	mounted	no
blood	variations	no
brass	like	no
buildings	aerofoils	no
bushel	measuring	no
business	thing	no
carangeot	faces	no
carpenter	device	no
circular	distance	no
cloth	truncated	no
column	pressure	no
compass	motor	no
consisting	arm	no
consisting	usually	no
consumed	passed	no
containing	tracks	no
correct	indicates	no
correcting	carried	no
count	number	no
counting	keeping	no
current	velocity	no
chest	respiratory	no
degrees	sensitivity	no
density	measuring	no
density	pressure	no
depth	fluid	no
designed	producing	no
designed	spectrum	no
detecting	instrument	no
detecting	producing	no
device	pair	no
dial	surveyors	no
dip	determine	no
direction	room	no
distance	number	no
distance	sent	no
distances	readings	no
distances	recording	no
earth	days	no
earth	slope	no
echoes	target	no
echoes	viewed	no
effect	amount	no
effect	shows	no
electrical	sensitive	no
elevation	axis	no
elevation	mounted	no

Palabra 1	Palabra 2	Identificado Manualmente
elevation	vane	no
employed	fringes	no
engine	informal	no
engine	power	no
esp	determination	no
esp	measuring	no
establishment	radio	no
estimating	ascertaining	no
estimation	determining	no
extracting	form	no
extracting	measuring	no
extracting	surgery	no
fitted	aircraft	no
fixes	regulator	no
flow	upstream	no
flowing	passed	no
force	arterial	no
force	escape	no
force	measuring	no
force	recording	no
foreign	angles	no
form	based	no
form	device	no
form	measuring	no
form	position	no
form	removing	no
form	variation	no
fourth	measuring	no
gage	indicates	no
gas	maintained	no
gas	pressure	no
gauge	graduated	no
gauge	pair	no
gauge	various	no
gauging	density	no
generally	measuring	no
graduated	piece	no
ground	instrument	no
hand	sighting	no
handle	like	no
heat	increase	no
heat	resistance	no
hence	pressure	no
horizon	figure	no
horizon	level	no
hyperbolic	radio	no
illustrating	detection	no

Palabra 1	Palabra 2	Identificado Manualmente
inclination	earth	no
index	upper	no
indicates	intensity	no
indicating	force	no
indicating	instrument	no
indicating	thing	no
instrument	acoustic	no
instrument	astronomy	no
instrument	depends	no
instrument	esp	no
instrument	informal	no
instrument	interference	no
instrument	measuring	no
instrument	mechanical	no
instrument	reference	no
instrument	self	no
instrument	tube	no
instruments	surveying	no
instruments	type	no
intensity	measuring	no
interference	acoustic	no
interference	optical	no
ionizing	occurrences	no
judging	usually	no
keeping	detecting	no
keeping	number	no
keeping	records	no
kept	light	no
kind	density	no
like	gas	no
line	determine	no
line	plumb	no
lungs	breathing	no
made	indicating	no
main	compressibility	no
marking	music	no
means	manometer	no
means	tempo	no
measurement	escape	no
measures	keeps	no
measuring	comparing	no
measuring	producing	no
measuring	solution	no
measuring	studying	no
measuring	tape	no
mech	naut	no
mech	nautical	no

Palabra 1	Palabra 2	Identificado Manualmente
mechanically	various	no
motion	device	no
motion	form	no
motion	variation	no
motor	power	no
movements	pneumogram	no
movements	sphygmogram	no
music	indicating	no
music	instrument	no
name	number	no
naut	nautical	no
night	axis	no
number	producing	no
observation	axes	no
observed	esp	no
observing	magnetic	no
observing	measuring	no
open	small	no
opposition	fluid	no
opposition	open	no
orig	direction	no
orig	room	no
paper	anemometer	no
passage	indicating	no
passage	keeps	no
passage	plane	no
passenger	fare	no
phenomena	features	no
photographic	material	no
photographic	specimen	no
photographic	transmission	no
phototube	sensitive	no
plane	spirit	no
plate	scale	no
point	water	no
pointing	placed	no
poised	centre	no
presence	deflection	no
presence	position	no
pressure	measuring	no
pressure	movement	no
process	pulse	no
projection	graduated	no
properly	formerly	no
provides	recording	no
purpose	small	no
quality	amount	no

Palabra 1	Palabra 2	Identificado Manualmente
quality	magnitude	no
quantity	electrical	no
radio	determining	no
reaction	designed	no
readings	ascertaining	no
recording	instrument	no
registering	spec	no
registering	storage	no
relative	photometric	no
rising	altitudes	no
rule	scale	no
salinity	amount	no
sea	depth	no
sensitive	temperature	no
setting	bullet	no
shadow	gnomon	no
sights	scale	no
slope	line	no
soil	surface	no
sonograph	components	no
sonograph	frequencies	no
sonograph	sound	no
sonograph	way	no
soundboard	employed	no
sounding	anglers	no
sounding	determine	no
spec	esp	no
specific	earth	no
speed	number	no
speed	sun	no
speed	vehicle	no
sphere	graduated	no
square	testing	no
substances	source	no
succession	revolution	no
sugars	rotates	no
surgery	form	no
surgery	measuring	no
take	direction	no
take	finding	no
temperature	coil	no
tension	moisture	no
tension	pitch	no
tests	person	no
threads	arrangement	no
time	exact	no
time	long	no

Palabra 1	Palabra 2	Identificado Manualmente
time	punching	no
time	reticle	no
time	stamping	no
transformation	force	no
transformation	measurement	no
traveled	bicycle	no
triggers	measured	no
tube	height	no
value	nitrogen	no
vapour	alcohol	no
variation	specific	no
various	variously	no
velocity	etc	no

Palabra 1	Palabra 2	Identificado Manualmente
velocity	relative	no
vessel	check	no
volume	voltage	no
water	bed	no
water	gas	no
water	pressure	no
weighing	generally	no
wheel	mechanically	no
wind	direction	no
wind	indicating	no
wood	shaped	no
work	power	no
yarn	content	no