Minería de textos para la asociación de contextos semánticos

Ramiro Aguilar, Vivian López, Luis Alonso, María Moreno, Angélica González

Departamento de Informática y Automática
Universidad de Salamanca
Plaza de la Merced S/N, 37008 Salamanca, Spain
ramiro@tejo.usal.es, {vivian, lalonso, mmg, angelica}@usal.es

Abstract. En este trabajo se presenta un método alternativo para el reconocimiento semántico. Se procesan datos en formato de texto para los cuales se realiza minería de textos encontrando asociaciones y patrones secuenciales los cuales son aprendidos por una red neuronal de Hopfield que encuentra asociaciones entre patrones con el mismo contexto. Así, para contextos nuevos o contextos incompletos, la red permite discernir a qué concepto pertenecen.

1 Introducción

La minería de textos es el proceso de extraer conocimiento válido y disperso en un documento y utilizarlo para organizar mejor la información. El análisis automatizado de textos en lenguaje natural es una tarea muy importante pues actualmente la mayor parte de la información se encuentra en forma de textos no estructurados. Los libros, los artículos de investigación, los manuales de productos, los correos electrónicos y todo tipo de documentación en la web, contienen información textual. Analizar enormes cantidades de información textual implica a menudo tomar decisiones importantes y correctas. Solucionar la tarea del análisis de tal información, puede proporcionar los medios para procesar de forma inteligente y automatizada el océano de textos que requieren la atención humana cada día. Tareas como el análisis preliminar de varios artículos, libros, planes de negocio, casos de ley, etcétera, podrían ser realizadas por las máquinas, siendo el primer paso clasificar esa información. En este sentido, en [9] se propone la combinación de dos tipos de modelos del conocimiento, los simbólicos, que utilizan estructuras de datos discretas, presentes en los diccionarios, gramáticas, etcétera y los conexionistas, donde el conocimiento emerge del contexto existente.

En dicha aproximación considerando conjuntos contextuales y textos básicos sobre algún universo de discurso, se propone un esquema para realizar un proceso de minería de textos para obtener un conocimiento sintáctico-semántico coherente respecto del texto base. Con esto se deseaba mejorar el procesamiento del lenguaje natural plasmado en datos textuales presentes en alguna fuente de datos, pretendiéndose así no solo automatizar el descubrimiento del conocimiento,

sino también, la generación de los datos para la minería. Con el objetivo de estructurar y obtener conocimiento semántico sobre algún texto escrito en lenguaje natural, se debían realizar las siguientes tareas:

1. Definir contextos dentro de algún corpus linguístico.
2. Seleccionar un texto como base para adquirir el conocimiento y codificar su léxico asociado (codificar su contenido).
3. Descubrir los patrones secuenciales en el texto base codificado.
4. Definir la gramática del texto base a partir de los patrones secuenciales antes obtenidos.
5. Etiquetar las reglas de producción de la gramática en base a diferentes contextos claves.
6. Entrenar una red neuronal con las reglas de producción etiquetadas para lograr el conocimiento semántico.

En el presente trabajo pretendemos mostrar que la codificación del texto base puede ser automática utilizando técnicas tradicionales de comparación de cadenas, de descubrimiento de asociaciones y de patrones secuenciales. Estos patrones son aprendidos por una red neuronal de Hopfield, como un modelo sencillo para explicar cómo ocurren las asociaciones entre conceptos. Usaremos este tipo de red con el fin de desarrollar memorias asociativas por contenidos: los elementos a memorizar no estarían ordenados por índices numéricos, sino según parte de su contenido, pudiéndose recuperar datos completos a partir de datos parciales utilizando la memoria asociativa de la red. A diferencia de los modelos tradicionales de redes neuronales que solo reciben datos numéricos, se pretende utilizar una red neuronal simbólica que permita expresar las relaciones entre palabras, categorizar las mismas, agruparlas e incluso construir redes semánticas.

2 Codificación automática del texto base

La codificación automática de los datos textuales necesita de un vocabulario C definido como la unión de conjuntos léxicos cl_i, $\forall i = 1..ncl$ (donde ncl es el número de conjuntos léxicos). El vocabulario puede ser generado con el siguiente algoritmo:

```
Algoritmo de generación de conjuntos léxicos
cl=0; // al principio no hay ningun conjunto léxico
Mientras hayan palabras en el corpus hacer:
    token = tomar(corpus); //se toman las palabras una tras otra
    ncl = busca(token); //A qué conjunto léxico pertenece el token?
    si (ncl != NULL) entonces
        codigo_token=code(ncl); //código del conjunto léxico
    en caso contrario
        cl=cl+1;
        codigo_token=code(cl); //crea un nuevo conjunto léxico
    fin si
fin mientras
fin del algoritmo
```

La generación del léxico creará tantos conjuntos como palabras existan en el léxico. Si la cantidad de conjuntos es alta se realiza un proceso de descubrimiento de asociaciones y de patrones secuenciales para sintetizar y disminuir el número de conjuntos léxicos.

2.1 Descubrimiento de asociaciones y patrones secuenciales

El descubrimiento de asociaciones consiste de los siguientes pasos:

1. Asociar con un identificador a cada unidad codificada
2. Ordenar secuencialmente las unidades según su identificador
3. Cuantificar las ocurrencias de las unidades presentes en las subsecuencias creando un vector en el que se coloca la cuenta de cada una. Aquellas unidades en las que la cuenta está por debajo de un "umbral" se eliminan.
4. Combinar en una matriz las transacciones atributo-valor y realizar el conteo de ocurrencias eliminando aquellos que no superen el valor del umbral
5. Tras repetir sucesivamente los pasos 3 y 4 se encuentran las asociaciones entre atributos.

En el paso 3, la cuantificación de las ocurrencias de las unidades o palabras usa la técnica de comparación por semejanza típica en el método de alineamiento de pares de secuencias en el tema de genómica la cual, no realiza comparaciones fijas en las ocurrencias, sino encuentra ocurrencias "semejantes" [13].

El descubrimiento de patrones secuenciales tiene el siguiente funcionamiento:

1. Identificar el atributo relacionado con el tiempo, es decir, cada subsecuencia.
2. En función del periodo de tiempo para el que se quiere descubrir los patrones secuenciales, crear una estructura ordenada por el identificador de la subsecuencia
3. Crear otra estructura de datos concatenando los subsecuencias
4. Según el porcentaje de soporte, inferir los patrones secuenciales (eliminar los que no superen la medida del umbral)

Definamos 5 contextos los cuales tienen una cantidad variada de términos clave [14]:

1. Deportivo: club de fútbol, presidente, ex-presidente, dinero, economía, socios, sociedad, goles, competición, jugadores, copa, fútbol.
2. Administrativo: certificado, matrícula, solicitud, renovación, documentación.
3. Científico: enfermedad, doctor, cirujano, síndrome, intestino irritable, tumor, cirugía, operación, traumatismo, cólico nefrítico, riñón, posición de los pies, prótesis, cadera, gota, mancha en la piel, soplo.
4. Educativo: cuerpo de revolución, cilindro, cono, esfera, triángulo, círculo, área, superficie, cuerpo geométrico, matemático.
5. Político: guerra, Golfo Pérsico, ataque, petróleo, liberación, invasión, representatividad, democracia, protesta, interés económico.

Cada contexto tiene asociado un corpus de donde se pueden adquirir patrones. En primer lugar se codifica cada corpus como se describió anteriormente. Sobre la codificación se descubren patrones secuenciales para formar los ejemplos o patrones del conjunto de entrenamiento. Cada patrón estará etiquetado respecto de su contexto.

Como en [9] se puede definir a los códigos compuestos de subcadenas de 3 caracteres del tipo 1**, el carácter 1 denota que se trata de un conjunto léxico conocido y los caracteres ** el identificador del número correspondiente al conjunto. Por ejemplo, si la palabra "amistad" está definida dentro del conjunto léxico 2, entonces la subcadena correspondiente a la palabra será 102. Podemos disponer de hasta 99 conjuntos léxicos. Si la subcadena no corresponde a un conjunto definido se codifica como 888 (ver figuras 1 y 2). Cada subcadena del código del texto base es un elemento de algún conjunto (esto es análogo a los procesos en genómica, el texto base será el cromosoma, la codificación sería el genoma y las subcadenas, los genes; secuenciar el genoma sería conocer la sintaxis y comprender el genoma, comprender la semántica. Ver [1] y [3]). En las subcadenas del código del texto base, la aparición de código 888 expresa desconocimiento. Para conocer la expresión de tales subcadenas se hace una aproximación a la subcadena más cercana conocida con la técnica de comparación por semejanza global. Por ejemplo, si tenemos la aparición de la subcadena 104 888 108 y se tienen subcadenas 104 102 108 lo más seguro es que se sustituya el código 888 por 102 a través de la expresión de la segunda subcadena.

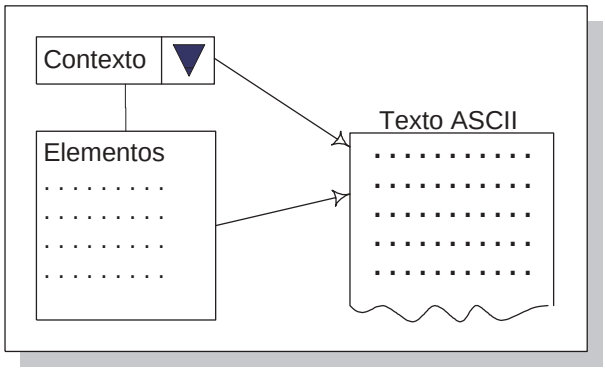

Fig. 1. Interrelación entre los elementos contextuales y el texto a analizar.

3 Red neuronal artificial de Hopfield

La red de Hopfield es un tipo de red neuronal recurrente de una sola capa que consiste en un conjunto de N neuronas interconectadas (figura 3) que actualizan

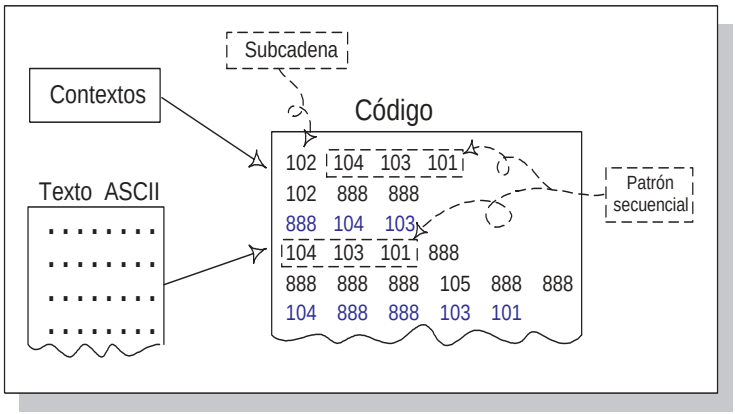

Fig. 2. Codificación del Cromosoma del texto original, se identifican algunos patrones secuenciales.

en paralelo sus valores de activación en cada ciclo temporal [7]. Son redes de adaptación probabilística, funcionalmente estarían en la categoría de memorias asociativas, que aprenden a reconstruir los patrones de entrada que memorizan durante el entrenamiento, reproducen conceptos previamente elaborados, recuerdan impresiones pretéritas y evocan ideas desde conceptos parecidos o situaciones semejantes. Por su modo de operación, esta memoria se puede describir como un sistema asociativo de procesamiento de información como:

- Memoria asociativa: sistema que permite emular las funciones de la memoria humana
- Memorias direccionables por el contenido.

En definitiva recuperan la información deseada a partir de una información parcial, "deteriorada" o con ruido.

3.1 Arquitectura de la red

La Arquitectura de la red se muestra en el gráfico de la figura 3.

Siendo: w_{ij} el peso asociado a la conexión entre la neurona i y la neurona j, que coincide con el valor w_{ji}.

e_i^k: Valor de la componente i-ésima del vector correspondiente a la información k-ésima que debe aprender la red.

N: Número de neuronas de la red, y por tanto, tamaño de los vectores de aprendizaje.

3.2 Funcionamiento de la red

La red Hopfield tiene un mecanismo de aprendizaje "off line". Por tanto existe una etapa de aprendizaje y otra de funcionamiento. En la etapa de aprendizaje se

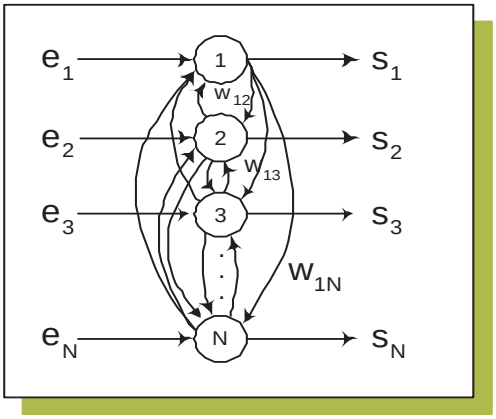

Fig. 3. Arquitectura de la red neuronal de Hopfield.

fijan los valores de los pesos en función de las informaciones que se pretende que memorice o almacene la red. Una vez establecidos, la red entra en funcionamiento tal y como se describió en el apartado anterior.

Esta red utiliza un aprendizaje no supervisado de tipo hebbiano, de tal forma que el peso de una conexión entre una neurona i y otra j se obtiene mediante el producto de los componentes i-ésimo y j-ésimo del vector que representa la información o patrón que debe almacenar. Si el número de patrones a aprender es M, entonces el valor definitivo de cada uno de los pesos se obtiene mediante la suma de los M productos obtenidos por el procedimiento anterior, un producto por información a almacenar.

3.3 Etapa de aprendizaje

El aprendizaje de la red consiste en seleccionar los vectores de pesos de las neuronas de manera que los posibles estados de equilibrio de la red representen los prototipos o contextos que se pretenden almacenar. Con este criterio, la entrada a la red es el conjunto de patrones secuenciales que se obtienen en la codificación del texto base. Así, durante la etapa de aprendizaje, los diferentes patrones de entrada se almacenan en la red, como si de una memoria se tratase. Posteriormente, en la fase de reconocimiento, se presenta una oración de entrada de algún contexto y luego la red evoluciona hasta estabilizarse en algún patrón o contexto antes memorizado. Si, por el contrario, la información de entrada no coincide con ninguna de las almacenadas, por estar distorsionada o incompleta, la red evoluciona generando como salida la más parecida. De esta forma se pueden reconocer contextos basados en contenido.

Los pasos a seguir en la fase de aprendizaje son:

1. Formar el conjunto de entrenamiento de M patrones secuenciales $\{E_1, E_2, ..., E_M\}$ donde:

$E_1 = [e_1^{(1)}, e_2^{(1)}, ..., e_N^{(1)}], E_2 = [e_1^{(2)}, e_2^{(2)}, ..., e_N^{(2)}], ..., E_M = [e_1^{(M)}, e_2^{(M)}, ..., e_N^{(M)}]$

2. Utilizando esta notación, el aprendizaje consistira en la creación de la matriz de pesos W a partir de los M vectores o informaciones de entrada $\{E_1, ..., E_M\}$ que se enseñan a la red. Es decir:
   $W = \sum_{k=1}^{M} (E_k^T \cdot E_k - I)$

Donde la matriz E_k^T es la traspuesta de la matriz E_k, I es la matriz identidad de dimensiones NxN que anula los pesos de las conexiones autorrecurrentes (w_{ii}).

3.4 Etapa de reconocimiento

La etapa de reconocimiento puede verse como la evolución temporal descrita similar a un sistema de ecuaciones no lineales en diferencias de primer orden: $x(t+1) = F(x(t))$.

En nuestro caso, x es el patrón E, F es una función *escalón*, por ejemplo la función *signo* de forma que el reconocimiento viene expresado por $S = sgn(W \cdot E)$.

Cuando el proceso realiza un gran número de iteraciones se tiene un *comportamiento asintótico del problema* y, más aun cuando los ejemplos de entrenamiento no son ortogonales entre sí. Si por el contrario, el número de iteraciones es finito se alcanza un *punto fijo* (punto de equilibrio) donde $F(x) = x$.

Todo lo anterior se puede justificar con teoremas conocidos del modelo de Hopfield.

Teorema 1: Si la matriz W formada con los vectores de pesos de la red es una matriz simétrica y definida positiva, entonces la red de Hopfield es convergente. Exigir que la matriz sea definida positiva, requiere que todos los elementos de su diagonal principal sean positivos. Esto significa que todas las neuronas tengan una autoconexión exitatoria, lo cual no está previsto en el esquema de la red de Hopfield. Si se impone que los elementos de la diagonal principal sean nulos, la red puede no ser globalmente estable, aunque se mantenga la simetría de la matriz.

Teorema 2: Si la matriz de pesos W es simétrica las trayectorias de la red de Hopfield convergen a un punto fijo o a un ciclo de período 2.

Teorema 3: Si la matriz de pesos W es simétrica y los elementos de la diagonal principal son no negativos, la red es globalmente convergente si se emplea una *dinámica asíncrona*.

4 Síntesis de la red de Hopfield para la codificación de contextos

El problema de sintetizar mediante una red de Hopfield una memoria asociativa que almacene una serie de vectores predeterminados es complejo. Sin embargo, se construye una memoria asociativa a partir de la información que se desea almacenar. Es necesario generar la matriz de pesos W de forma que los puntos

fijos de la red se correspondan con los patrones asociados a los contextos. En este caso, las entradas a la red serán las subcadenas que empieza con el caracter '1'.

4.1 Mejora de los ejemplos de aprendizaje

Para mejorar el aprendizaje de la red de Hopfield, los ejemplos de entrenamiento de cada contexto se han ortogonalizado. Para lograr un reconocimiento adecuado, el número máximo de neuronas de la red se fijó en cien. Con esto se pueden reconocer, como mínimo, un número diferente de catorce patrones o contextos, lo que hace viable al método ya que se definieron 5 contextos.

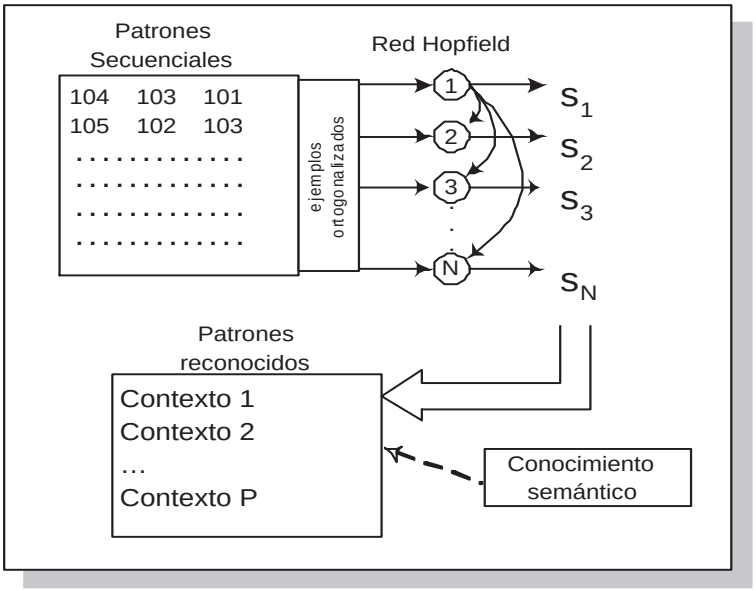

Fig. 4. Síntesis de la red de Hopfield para la asociación de contextos.

El problema de ortogonalización puede ser encarado, además, mediante determinados procedimientos que garanticen una diferencia suficiente entre la información que debe aprender la red. En definitiva, lo que se pretende siempre es que cada par de patrones de entrada difieran en, al menos $\frac{N}{2}$ componentes, siendo N el número total de componentes por cada patrón. Esta condición puede expresarse como:

$\sum_{i=1}^{N} e_i^{(k)} e_i^{(l)} \leq 0, \forall k \neq l$

donde $e_i^{(k)}$ y $e_i^{(l)}$ son valores binarios (-1 o +1) de los componentes i-ésimos de dos vectores (patrones) diferentes (k y l) a almacenar en la red. Esta condición de

ortogonalidad que establece que el número de componentes diferentes (también llamado distancia "Hamming") de dos patrones, sea al menos la mitad del total ($0.5N$) puede ser relajada, estableciendo una distancia mínima del 30% del total para que sean "casi" ortogonales, garantizándose todavía un funcionamiento aceptable. En este caso considerando valores binarios de -1 y +1, la expresión será:

$\sum_{i=1}^{N} e_i^{(k)} e_i^{(l)} \leq 0.7N - 0.3N = 0.4N$

ya que se permite un 70% de componentes diferentes y un 30% de componentes iguales.

4.2 Tipo de Asociación

Existen dos formas de realizar esta asociación entre entrada/salida que se corresponde con la naturaleza de la información almacenada. Una primera es la denominada heteroasociación, que se refiere al caso en que la red aprende parejas de datos $[(A_1, B_1), (A_2, B_2), ..., (A_N, B_N)]$, de tal forma que cuando se presente cierta información de entrada A_i, deberá responder generando la correspondiente salida B_i.

La segunda forma de asociación se conoce como *autoasociación*, donde la red aprende ciertas informaciones $A_1, A_2, ..., A_N$, de tal forma que cuando se le presente una información de entrada realizará una autocorrelación, respondiendo con uno de los datos almacenados, el más parecido al de la entrada. Este tipo de asociación es la que utilizaremos para reconocer los contextos.

4.3 Algunos resultados

Se han utilizado 5 corpus de diferentes contextos como definimos en la sección 2. Estos contextos pueden ser codificados como se observa en la tabla 1.

Contexto	*Patrón*
Deportivo	+1 +1 -1 -1 -1
Administrativo	-1 -1 -1 +1 +1
Científico	+1 -1 -1 -1 +1
Educativo	-1 +1 +1 +1 -1
Político	+1 +1 +1 +1 +1

Table 1. Patrones de entrenamiento para la red de Hopfield.

En cada contexto, luego de codificar las unidades lexicales, se han seleccionado 100 ejemplos o patrones secuenciales de entrenamiento a los cuales se ha asociado un concepto para su aprendizaje por la red de Hopfield. La implementación se realizó en MatLab 6.1. El siguiente trozo de código, muestra el entrenamiento de cinco patrones y el reconocimiento de un patrón de prueba.

```
% Etapa de entrenamiento
% de la Red Hopfield
E = [1 1 -1 -1 -1;
    -1 -1 -1 1 1;
    -1 1 1 1 -1;
    1 -1 -1 1 1;
    1 1 1 1 1]';
net = newhop(E); % crea la red neuronal de Hopfield
[Y, Pf, Af] = sim(net, 5, [], E); % entrenamiento
Y % patrones aprendidos
. . .
% Fase de reconocimiento autoasociación
% o autocorrelación en 10 intentos
X = {[1.00; 1.00; 0.70; 0.98; 0.90]}; % pat. prueba
[Y,Pf,Af] = sim(net,{1 10},{},X);
Y{1} % salida reconocida
```

Luego del aprendizaje de la red se pueden introducir entradas no necesariamente conocidas para que la red las asocie con el contexto más parecido. La tabla 2 muestra la asociación de las 8 oraciones descritas abajo; debe notarse que algunas oraciones no aparecen tal cual en el corpus de su respectivo contexto, sin embargo, la red las aproxima y las autocorrelaciona.

1. Ha caducado el plazo para la matrícula
2. El fútbol maneja mucho dinero
3. El hombre fue capaz de producir cuerpos y vasijas circulares más perfectos cuando se inventó la rueda del alfarero
4. Quién puso tan mal al Real Madrid
5. La opinión pública rechaza la guerra del Golfo Pérsico
6. Es necesaria una fotografía en la documentación
7. El doctor Antonio Sierra es cirujano
8. Los tumores del intestino suelen necesitar intervención médica

Oración	*Patrón codificado*	*Patrón asociado*	*Contexto reconocido*
1	0.99 -1.00 -0.90 -0.89 1.00	1.00 -1.00 -0.21 -0.19 1.00	Administrativo
2	1 1 -1 -1 0.90	1.00 1.00 -0.22 -0.22 1.00	Deportivo
3	-1.00 1.00 1.00 1.00 -1.00	-1.00 1.00 0.22 0.22 0.63	Educativo
4	1 0.8 -0.6 -1 0.80	1.00 0.93 0.05 -0.41 1.00	Deportivo
5	1.00 1.00 0.70 0.98 0.90	1.00 1.00 0.02 0.35 1.00	Político
6	0.89 -0.78 -0.80 -1.00 1.00	1.00 -0.91 -0.08 -0.32 1.00	Administrativo
7	1.00 -1.00 -1.00 -1.00 1.00	1.00 -1.00 -0.22 -0.22 1.00	Científico
8	1.00 -0.90 -0.99 -1.00 0.59	1.00 -1.00 -0.22 -0.23 0.99	Científico

Table 2. Asociación de oraciones a los contextos aprendidos por la red de Hopfield.

5 Conclusiones

Las redes de Hopfield tienen una naturaleza regenerativa: una vez que aprenden patrones, pueden reconstruir otro patrón asociándolo con el más parecido que haya memorizado.

Para reconocer la semántica de los textos, se obtienen automáticamente patrones secuenciales que describen los contextos y nos permiten descubrir el significado de los mismos [9]. Sin embargo, algunos contextos pueden aparecer incompletos por lo que la red de Hopfield nos va a permitir reconstruirlos.

Debe notarse que una limitación del modelo de Hopfield plantea que el número de patrones a reconocer es aproximadamente el 14% del número total de neuronas con lo que la fase de ortogonalización de los ejemplos de entrada es fundamental.

References

1. R. Aguilar (2002). *Análisis de técnicas de minería de datos, comparación de resultados en diferentes dominios de aplicación.* Memoria para el Grado de Salamanca de la Universidad de Salamanca. España.
2. R. Aguilar (2002). "Pautas Para la Simbiosis: Minería de Datos y Lógica Borrosa". *XI Congreso Español sobre Tecnologías y Lógica Fuzzy.* Universidad de León. ESTYLF 2002.
3. R. Aguilar (2003). *Minería de datos: fundamentos, técnicas y aplicaciones.* Universidad de Salamanca.
4. M. Berry, G. Linoff (1999). *Mastering Data Mining: The Art and Science of Customer Relationship Management.* John Wiley & Sons.
5. U. Fayyad, et.al. (eds)(1996). *Knowledge Discovery in Databases.* MIT Press.
6. J. Han, M. Kamber (2000). *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers.
7. S. Haykin (1999). *Neural network. a comprehensive foundation.* Prentice Hall.
8. International Human Genome Sequencing Consortium. (2001). "Initial Sequencing and Analysis of the Human Genome". *Nature* 409.
9. V. López, R. Aguilar (2002). "Minería de textos y aprendizaje Automático en el procesamiento del lenguaje natural", WorkShop de Minería de Datos y Aprendizaje Automático, IBERAMIA 2002, Universidad de Sevilla.
10. V. López (1996). *Desambiaguación Semántica Basada en Métodos Conexionistas para un Problema de Traducción Automática Alemán-Español.* Tesis Doctoral de la Universidad de Valladolid. España.
11. V. López, L. Alonso, M. Moreno (2000). "Mapas Organizados para la Minería de Datos en Procesamiento del Lenguaje Natural". *Actas del II Taller Iberomericano sobre Aplicaciones e Implementaciones de Redes Neuronales en Reconocimiento de Patrones.* Universidad de Salamanca. España.
12. S. Wallis, G. Nelson (2001). "Knowledge Discovery in Grammatically Analysed Corpora". *Data Mining and Knowledge Discovery.* Kluwer Academic Publisher.
13. Universidad Autónoma de Madrid, Materiales del curso de doctorado en Bioinformática. En: http://www.pdg.cnb.uam.es/cursos/BioInfo2004 /pages/Aliseq/Teoria/img12.htm
14. Yahoo-Geocities Server, Página personal de Ramiro Aguilar. En: http://www. geocities.com/ramirohp/corpus.html