risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 30/05/2020
Aceitação/Acceptance: 21/08/2020

Detección de paráfrasis basada en la energía, entropía y temperatura textual

Alberto Iturbe-Herrera[1], Armando Rojas-Valdez[1],
Noé Alejandro Castro-Sánchez[1], Gerardo Sierra[2]

Iturbe@cenidet.edu.mx; armando.rojas17ce@cenidet.edu.mx; noe.cs@cenidet.tecnm.mx; gsierram@iingen.unam.mx

[1] Tecnológico Nacional de México - Cenidet, Interior internado Palmira S/N, Col. Palmira, 62490, Morelos, México.

[2] Instituto de Ingeniería, Universidad Nacional Autónoma de México, Av. Universidad, 3000, Coyoacán, 04510, Ciudad de México, México.

DOI: 10.17013/risti.39.35-51

Resumen: La paráfrasis es la reformulación de un texto utilizando un vocabulario distinto para plasmar la idea original con nuestras propias palabras. En esta investigación se presenta un método para la detección de paráfrasis incorporando los conceptos de Entropía y Temperatura Textual a un modelo previo que centró su contribución en la implementación de las redes neuronales recurrentes de Hopfield para generar una medida de distancia llamada Energía Textual. Utilizando la Entropía y la Temperatura se generó un Contexto de Afinidad Libre, basándose en el Model Ising, lo que permitió medir la distribución semántica entre pares de oraciones. Este modelo fue evaluado utilizando el recurso *Microsoft Research Paraphrase Corpus*, permitiendo superar los resultados del modelo anterior y logrando identificar más de la mitad de la paráfrasis de la muestra analizada.

Palabras-clave: detección de paráfrasis; energía textual; entropía; temperatura textual; red de Hopfield.

Paraphrase detection based on Energy, Entropy and Textual Temperature

Abstract: Paraphrases are the reformulation of a text using different vocabulary to capture the original idea in our own words. In this research, a method for paraphrase detection is presented. We incorporate two variables, Entropy and Textual Temperature, into a previous model which implemented a Hopfield Network to generate a distance measure called Textual Energy. A Context of Free Affinity was generated using Entropy and Temperature based on the Ising Model, which allowed us to measure the semantic distribution between pairs of sentences. Our model was evaluated using Microsoft's Research Paraphrase Corpus improving the results of the previous model and was able to identify more than half of the paraphrases presented in the analyzed sample.

Keywords: Paraphrase detection; Textual Energy; Entropy; Textual Energy; Hopfield Network.

1. Introducción

Las características que definen la 4ta. Revolución Industrial se basan en desarrollos tecnológicos que incorporan principalmente Inteligencia Artificial, y en la creciente cantidad de datos que se generan y los algoritmos que se requieren para procesarlos. En noviembre del 2018, la empresa Seagate (líder global en soluciones de almacenamiento) y la consultora IDC (principal proveedor mundial de inteligencia de mercado, servicios de asesoramiento y eventos para las TIC), publicaron un informe titulado *The Digitalization of the World From Edge to Core* (Reinsel, Gantz & Rydning, 2018) donde concluyen que en el año 2025 se habrán generado 5 veces más datos que en el 2018, un aproximado de 175ZB, de los cuales, el 79% se encontrarán en formato de texto.

Bajo este escenario, las soluciones que ofrece el área de Procesamiento de Lenguaje Natural cobran vital importancia. Entre las diferentes tareas que aborda se encuentra la detección de paráfrasis. Se entiende como tal a la expresión de un enunciado utilizando palabras distintas, pero conservando la misma idea; por lo tanto, cuando se habla de paráfrasis se sabe que existe un enunciado original y al menos una versión distinta de él que mantiene el mismo significado.

La detección de paráfrasis es útil en tareas como la generación automática de resúmenes, la traducción automática, la búsqueda de respuestas e incluso para resolver problemas de índole legal, como la detección de plagio, delito que se ha incrementado en los últimos años debido a la gran cantidad de datos existentes en formato de texto y su fácil acceso a través de Internet.

En este trabajo se propone un modelo para la detección de paráfrasis empleando un método de aprendizaje no supervisado con el que se extiende el umbral de recuperación de información de una Red Neuronal Recurrente de Hopfield, el modelo emplea además la Entropía y la Temperatura textual para la extracción de características. El resultado es un modelo de búsqueda de similitud semántica entre pares de enunciados que llega a encontrar más de la mitad de la paráfrasis que puede existir en un corpus lingüístico.

El artículo se divide en las siguientes secciones: en el capítulo 2 se mencionan los trabajos relacionados a la detección automática de la paráfrasis, el capítulo 3 introduce la definición de los conceptos clave que permiten tener una mejor contextualización teórica del trabajo, en el capítulo 4 se explica el modelo de solución propuesto para la identificación de la paráfrasis, el capítulo 5 abarca los experimentos y resultados y finalmente en el capítulo 6 se señalan las conclusiones obtenidas en este trabajo.

2. Trabajos relacionados

Para detectar paráfrasis en textos se puede recurrir a la intervención manual, y cuando la intensión es resolver un uso ilegal de ella, a los peritos correspondientes. Atenderse de esta manera implica una inversión de tiempo y dinero, además de que, como en todo trabajo manual, no está exenta de incurrir en errores. A lo largo de los años

han surgido diferentes propuestas para resolver esta tarea de forma automática. Los trabajos publicados tradicionalmente se han dividido en dos grandes categorías según las técnicas que emplean: por una parte, se encuentran trabajos que se basan en el uso de funciones de similitud para resolver si en un par de oraciones existe paráfrasis, y, por otro lado, se tienen los trabajos que utilizan aprendizaje automático, donde a partir de un corpus de entrenamiento extraen características y generan un modelo que predice la clase correspondiente de un par de oraciones.

En (Mohamed & Wael, 2019), se hace un estudio que abarca trabajos en una ventana de tiempo de 10 años, considerando métodos más recientes y los clasifican en enfoques no supervisados y enfoques supervisados, en este último incluyen enfoques clásicos de aprendizaje automático y enfoques de aprendizaje profundo.

Socher, Huang, Pennington, Ng & Manning (2011) presentan un método para la detección de paráfrasis basado en *autoencoders* recursivos (RAE por sus siglas en inglés). Los RAE no supervisados aprenden vectores de características para frases en árboles sintácticos. Estas características se utilizaron para medir la similitud entre palabras y frases entre dos oraciones. Además, los autores introducen una nueva capa de agrupación dinámica que calcula una representación de tamaño fijo a partir de las matrices de tamaño variable, la cual, se utiliza como entrada para un clasificador.

Dentro de los trabajos relacionados en la detección de paráfrasis, Milajevs, Kartsaklis, Sadrzadeh & Purver (2014) realizan un estudio comparativo entre representaciones neuronales de palabras y espacios vectoriales tradicionales basados en recuentos de concurrencia, en una serie de tareas de composición. En éste se utilizan tres espacios semánticos diferentes y se implementan siete modelos de composición basados en tensores, evaluados en tareas que involucran desambiguación de verbos y similitud de oraciones.

Por otro lado, Qiu, Kan & Chua (2006) proponen un modelo que no se centra en la detección de la paráfrasis, sino en identificar las diferencias entre las oraciones, su motivación se basa en que oraciones similares comparten una cantidad sustancial de información, y si existe información adicional en las oraciones, su eliminación no sería significativa. Su modelo se implementa en dos fases, en la primera usan un detector de similitud, donde buscan información similar entre oraciones, y en la segunda aplican un clasificador de diferencias, donde valoran si la información adicional es significativa.

En Mihalcea, Corley & Strapparava (2006) se presenta un método para medir la similitud semántica de los textos, utilizando medidas de similitud basadas en el corpus y el conocimiento. El trabajo previo sobre este problema se ha centrado principalmente en documentos grandes (clasificación de texto, recuperación de información) o palabras individuales (pruebas de sinonimia).

Enfoques más actuales que implementan soluciones basadas en redes neuronales se pueden observar en Lan & Xu (2018) quienes analizan cinco diferentes implementaciones de redes neuronales para el modelado de pares de sentencias, lo cual es fundamental para resolver diversas tareas de Procesamiento de Lenguaje Natural, como la similitud semántica, la identificación de paráfrasis, la inferencia de lenguaje natural, preguntas-

respuestas y la comprensión automática. Sus experimentos los realizaron en 8 conjuntos de datos: dos conjuntos NLI, tres PI, un conjunto de datos STS y dos sobre QA.

3. Marco conceptual

Los fundamentos teóricos en los que se basa este trabajo abarcan diversos conceptos. El primero de ellos es la energía textual, que surge a partir del análisis de las redes neuronales recurrentes de Hopfield, las cuales basan su comportamiento en el modelo termodinámico de Ising. Se propone el uso de dos variables, la Temperatura textual y la entropía, que sirven para reforzar la capacidad de la Red Neuronal Recurrente de Hopfield. Este planteamiento permite medir la distribución semántica entre pares de oraciones. A continuación, todos estos conceptos se definirán en el orden en que se mencionaron.

3.1. Energía Textual

La energía textual es un concepto acuñado por el grupo de investigadores del Laboratorio de Informática Aplicada de la Universidad de Aviñón (Fernández, SanJuan & Torres-Moreno, 2007), quienes identificaron que el comportamiento de las Red Neuronal Recurrente de Hopfield puede ser utilizado para distintas tareas de procesamiento de lenguaje, tales como agrupamiento semántico (Torres-Moreno, Molina y Sierra, 2010). La base que toman es el modelo vectorial, lo que permite interpretar los elementos de una matriz documento-EL (Entidades Léxicas) como los espines o neuronas de una Red Neuronal de Hopfield, definiendo a los documentos como secuencias de neuronas, activando o desactivando la neurona si la aparece o no en la frase respectivamente.

Los autores hacen énfasis en la importancia de la interacción de las unidades consigo mismas desde el punto del procesamiento del lenguaje, a diferencia del modelo de Hopfield donde dicha interacción no es considerada.

3.2. Red Neuronal Recurrente de Hopfield

El concepto de Red neuronal artificial (RNA) se define, según Fausett (1994), como "*un sistema de procesamiento de información que posee ciertas características de rendimiento en común con las redes neuronales biológicas*". Se aplica para resolver una gran variedad de problemas, como almacenar y recuperar datos o patrones, clasificar patrones y en la agrupación, además de la búsqueda de soluciones a problemas de optimización restringidos (Copeland, 2016).

Goodfellow, Bengio & Courville (2016) señalan que una de las RNA de mayor popularidad son las Redes Neuronales Recurrentes (RNR, utilizadas comúnmente para procesamiento y aprendizaje de datos secuenciales). Las RNR son máquinas de aprendizaje que tienen como objetivo utilizar información secuencial. Se utilizan en diversos contextos donde la dependencia temporal de los datos es una característica implícita de alto impacto. Calculan de forma recursiva nuevos estados implementando funciones de transferencia a estados y entradas previas.

En la Figura 1 se muestra la arquitectura de una RNR simple.

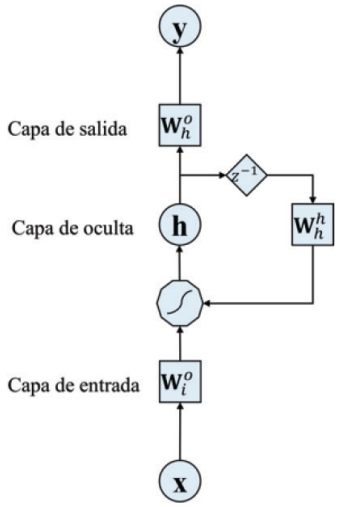

Figura 1 – Arquitectura de una RNR simple

Los datos de entrada se representan con , la variable representa los nodos de salida y corresponde a los nodos ocultos. , y son matrices que representan los pesos de entrada, ocultos y salida, respectivamente. El polígono representa la transformación no lineal realizado por las neuronas. Finalmente, corresponde al operador de demora de la unidad.

Hopfield (1982) propuso una RNR denominada Red de memoria asociativa. Es eficiente para resolver problemas de optimización, memorias asociativas, procesamiento de lenguaje, partición de gráficos, visión estéreo, entre otras. Asocia dos valores de activación binario (0, 1) o bien bipolar (1,-1), que se determinan si las unidades superan o no un determinado umbral. Por lo tanto, la definición posible para la unidad de activación es:

$$a_i \leftarrow \begin{cases} 1 & \text{si } \sum_j w_{ij} s_j > \theta_i \\ 0 & \text{en caso contrario.} \end{cases}$$

Donde: w_{ij} representa la fuerza del peso de la conexión de la unidad j a la unidad i, denominado peso de conexión; y, s_j y θ_i corresponden al estado y al umbral de la unidad i, respectivamente.

La Red Recurrente de Hopfield posee un valor escalar asociado a cada estado de la red conocido como Energía (E) de la red, que se define de la siguiente forma:

$$E = -\frac{1}{2}\sum_{i,j} w_{ij} s_i s_j + \sum_i \theta_i s_i$$

Este valor recibe el nombre de Energía, debido a que la definición indica que, si las unidades son elegidas al azar para actualizar los valores de activación, la red convergerá a estados que son mínimos locales de la función de energía.

3.3. Modelo de Ising

Ernst Ising propone un modelo físico para estudiar el comportamiento de materiales ferromagnéticos. Consta de variables discretas que representan momentos dipolares magnéticos de espines atómicos que pueden presentar uno de dos estados (+1 o −1). Los giros se organizan en un gráfico, generalmente una red (donde la estructura local se repite periódicamente en todas las direcciones), permitiendo que cada giro interactúe con sus vecinos. Los giros vecinos que coinciden tienen una energía menor que los que no coinciden; el sistema tiende a la energía más baja, pero el calor perturba esta tendencia, creando así la posibilidad de diferentes fases estructurales (Wolf, 2000).

El modelo permite la identificación de transiciones de fase, como un modelo simplificado de la realidad. El modelo Ising bidimensional de celosía cuadrada es uno de los modelos estadísticos más simples para mostrar una transición de fase.

3.4. Temperatura Textual

La Temperatura Textual es una medida de distancia entre un par de enunciados de acuerdo con su estructura sintáctica (Stepanov, 2012). Tiene su origen en el Modelo de Ising el cual es la base del diseño del modelo computacional propuesto en esta investigación. Para llevar a cabo el cálculo de la temperatura textual se decidió emplear el Coseno Suave, dado que esta variable está estrechamente relacionada con la entropía que tiene como objetivo reducir el ruido (desorden) que es generado en una fuente de información de acuerdo con la Temperatura de la misma. Dicho de otra manera, la Entropía es constante a su Temperatura.

3.5. Entropía

Una de las acepciones de la palabra entropía, según la Real Academia de la Lengua Española, es *Medida del desorden de un sistema* (Real Academia Española, 2020). Considerándola desde el punto de vista de datos, se define como la medida del desorden que existe dentro de una fuente de información, la cual indica el nivel de entropía en un par de oraciones (Grzymala-Busse, 1993). De esta manera, y considerando el orden alfabético, la cadena "GDCBAEF" cuenta con un nivel de entropía alto, pero la cadena "ABCDEF" tiene un nivel de entropía bajo.

3.6. Distribución semántica

También conocida como semántica distribucional (Firth, 1968) es un área de investigación que tiene como objetivo desarrollar y estudiar teorías y métodos para cuantificar y clasificar las similitudes semánticas entre elementos lingüísticos según sus propiedades distribucionales en grandes muestras de datos. La idea básica de la distribución semántica se resume en la hipótesis distribucional (Sahlgren, 2008), la cual tiene origen en la teoría semántica del uso lingüístico, es decir, palabras que se usan y aparecen en los mismos contextos tienden a transmitir significados similares.

Los métodos más empleados para generar una Distribución Semántica son TF-IDF, *K-means*, *Naive-Bayes*, Shannon ID3, *Word Embeddings*, entre otros.

4. Método de solución

En este trabajo se propone incorporar las variables de Entropía y Temperatura textual para reforzar la capacidad de la Red Neuronal Recurrente (RNR) de Hopfield que se utiliza en el modelo de la Energía textual. La salida generada por esta red es un vector que contiene una cantidad indefinida de ruido, por lo que se considera que las variables que se proponen harían posible reducirlo. Esto finalmente permitiría aumentar el umbral de recuperación de información de la red neuronal.

El modelo que se propone (Figura 2) recibe pares de oraciones, una considerada original y otra como posible paráfrasis. Para fines de la experimentación se utilizó el recurso *Microsoft Research Paraphrase Corpus*.

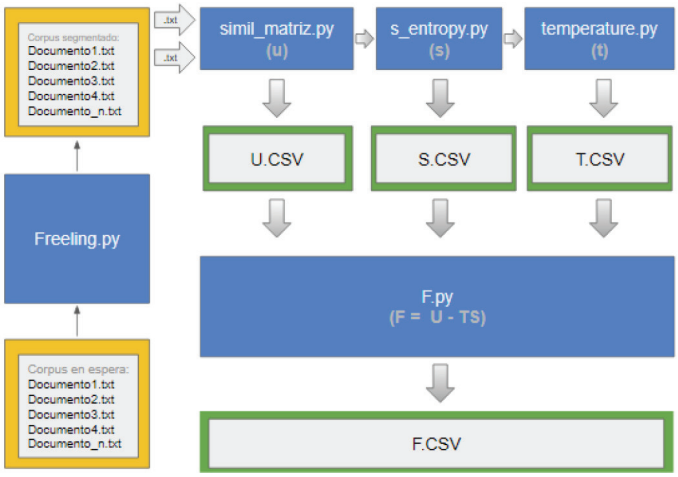

Figura 2 – Método de solución

Ambas oraciones son procesadas con el etiquetador morfosintáctico Freeling, a través del cual se extraen los lemas y se eliminan *stopwords*.

Con esta información se emplea la Red Neuronal Recurrente de Hopfield, de la cual se obtiene una matriz diagonal que permite comparar dos oraciones para identificar los *términos (palabras)* comunes.

Considérense como ejemplo los siguientes dos enunciados:

- Enunciado 1: "*Bienvenidos a la Casa de la Primavera, donde tienen el mejor clima y a las mejores personalidades.*"
- Y enunciado 2: "*Bienvenidos a la Ciudad de la Eterna Primavera, donde tienen el mejor clima y a las mejores personalidades.*"

En la matriz diagonal (Figura 3) se coloca el índice "1" a los términos repetidos en ambas oraciones y "0" si no coinciden.

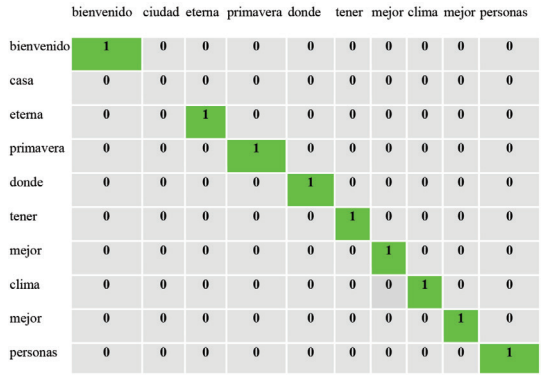

Figura 3 – Matriz de espines (Energía Textual)

Se obtiene el tamaño de la diagonal (en este caso 10 elementos) y la frecuencia de ocurrencia de los índices con valores positivos (en este caso 9 elementos). Dicho tamaño de la diagonal representa el nivel de semejanza de un par de enunciados. Por lo tanto, la similitud semántica entre un par de enunciados es proporcional a la cantidad de valores que existen dentro de la diagonal principal respecto a su cantidad de índices con valores positivos.

La Energía Textual se genera mediante el conteo de los valores de activación positivos dentro de la diagonal principal respecto a la totalidad de elementos existentes dentro de la misma. El resultado es el mínimo local que converge la diagonal principal de la matriz bidimensional que emplea la Red de Hopfield, basándose en el comportamiento de los valores de activación del Modelo de Ising. Retomando el ejemplo, el cálculo del valor de la Energía Textual (**U**) devuelve un resultado de ***0.90%.***

Posteriormente se calcula la Entropía Textual (**S**) entre un par de cadenas de texto según la frecuencia de ocurrencia de un token y su respectivo lema. Esta información permite generar un *árbol de decisión* que se usa para indicar la semejanza de acuerdo con la Entropía de Shannon, como se muestra en la Tabla 1. Las columnas Enunciado 1 y Enunciado 2 muestran los términos de dichos enunciados, la columna Lemas, el término lematizado y **m** y **n** indican con el valor 1 que el término aparece en ambas oraciones y con el valor 0, lo contrario.

Enunciado 1	Lema	m	Enunciado 2	Lema	n
Bienvenidos	bienvenido	1	Bienvenidos	bienvenido	1
casa	casa	0	ciudad	ciudad	0
eterna	eterna	1	eterna	eterna	1
⋮	⋮	⋮	⋮	⋮	⋮
clima	clima	1	clima	clima	1
mejores	mejor	1	mejores	mejor	1
personas	persona	1	personas	persona	1

Tabla 1 – Entropía de información

De tal manera que la entropía entre el par de enunciados debe ser una medida de distancia que se genera haciendo uso de un árbol de decisión basado en la teoría de la información de Shannon (Figura 4).

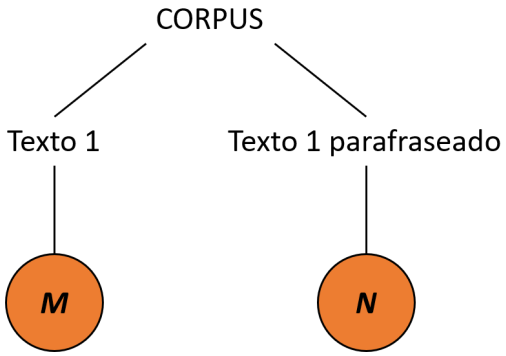

Figura 4 – Árbol de decisión ID3

En este caso, el valor de **S** es igual a **0%**, ya que para medir la distancia se identifica cuál de los dos valores (**m** ó **n**) es mayor y es ocupado como el minuendo para realizar una sustracción, siendo el valor de n el sustraendo.

La Temperatura Textual (**T**) se calcula utilizando el coseno suave definido de la siguiente fórmula:

$$\text{soft\_ cosine}_1(a,b) = \frac{\sum_{i,j}^{N} s_{ij} a_i b_j}{\sqrt{\sum_{i,j}^{N} s_{ij} a_i a_j} \sqrt{\sum_{i,j}^{N} s_{ij} b_i b_j}}$$

En este caso, el Coseno Suave un valor de **0.90**. Al invertir el sentido de su valor se obtiene una medida de similitud.

$$T = Coseno\_Suave\ (Enunciado\ 1,\ Enunciado\ 2)$$

$$T = 1 - T'$$

$$T = 1 - 0.90$$

$$T = \mathbf{0.10}$$

Estos valores pertenecen al Contexto de Afinidad Libre (basado análogamente en la Energía Libre del Modelo de Ising), y que en términos de asociación se definen como:

$$\boldsymbol{F = U - (T \cdot S)}$$

Y finalmente, en términos de disociación como:

$$\boldsymbol{F = 1 - F'}$$

Tal contexto proporciona dos variables para generar una Distribución Semántica, la primera es la Asociación, la cual es un índice de similitud textual y la segunda es la Disociación, medida que sirve para estimar la distancia en un par de piezas de texto. A continuación, se muestra un ejemplo en la Tabla 2.

Entropía (S)	Energía (U)	Temperatura (T)	F (asociación)	$F = U - (T \cdot S)$
0	0.9	0	1	0

Tabla 2 – Contexto de Afinidad Libre

Por lo tanto, el resultado en términos de Asociación es F = 0.90 y en términos de Disociación, 1 – F = 0.10. Este método computacional genera finalmente un Contexto de Afinidad Libre que es útil para detectar paráfrasis.

5. Experimentación y resultados

Para la evaluación del modelo se usó el *Microsoft Research Paraphrase Corpus*. Este corpus fue originalmente evaluado por dos jueces humanos que determinaban si las oraciones eran lo suficientemente cercanas en el sentido semántico como para ser consideradas paráfrasis. Un tercer juez intervenía para resolver desacuerdos existentes. Los pares considerados paráfrasis se denotan con el valor 1, y en caso contrario con el valor 0.

En este trabajo se tomaron sólo 2,000 oraciones que se dividieron en los grupos I y J, cada uno con 1000 oraciones, debido al costo computacional que implica el cálculo de las variables de Energía, Entropía y Temperatura textual.

Una vez obtenidos los grupos de evaluación, se procedió con el cálculo respectivo de las variables de la ecuación $F = U - (T \cdot S)$. *Donde* $\boldsymbol{U}$ es la Energía Textual, $\boldsymbol{T}$ representa la Temperatura Textual y $\boldsymbol{S}$ corresponde a la Entropía Textual. En la Tabla 3 se muestran resultados de U, T y S con oraciones comparadas contra sí mismas y contra otras dos oraciones distintas.

#	Oración original	Oración candidata	U	T	S
1.1	*PCCWs chief operating officer Mike Butcher and Alex Arena the chief financial officer will report directly to Mr So.*	*PCCWs chief operating officer Mike Butcher and Alex Arena the chief financial officer will report directly to Mr So.*	1	1	0
1.2	*PCCWs chief operating officer Mike Butcher and Alex Arena the chief financial officer will report directly to Mr So.*	*Rich media doubled its share increasing from 3% in Q2 2002 to 6% in Q2 2003.*	0	0.05	0.235
1.3	*PCCWs chief operating officer Mike Butcher and Alex Arena the chief financial officer will report directly to Mr So.*	*Thursday afternoon the Standard & Poors 500-stock index was trading up just 148 points or 01 percent to 104959.*	0.075	0.103	0.101

#	Oración original	Oración candidata	U	T	S
2.1	*Ms Laffertys lawyer Thomas Ezzell told a Kentucky newspaper: My understanding of this is that there is a lower percentage of successful impregnations with frozen.*	*Ms Laffertys lawyer Thomas Ezzell told a Kentucky newspaper: My understanding of this is that there is a lower percentage of successful impregnations with frozen.*	1	1	0
2.2	*Ms Laffertys lawyer Thomas Ezzell told a Kentucky newspaper: My understanding of this is that there is a lower percentage of successful impregnations with frozen.*	*Last year Congress passed similar though less expensive buyout legislation for peanut farmers ending that Depression-era program.*	0	0.044	0.16
2.3	*Ms Laffertys lawyer Thomas Ezzell told a Kentucky newspaper: My understanding of this is that there is a lower percentage of successful impregnations with frozen.*	*Oracle on Friday launched a $51 billion hostile takeover bid for PeopleSoft.*	0	0.107	0.083

Tabla 3 – Ejemplos de los valores de Energía (U), Temperatura (T) y Entropía (S)

En la Tabla 4 se muestra el valor resultante F para cada par de oraciones. Dado que representa la asociación, también fue necesario calcular el nivel de disociación definido como *1 - F*, el cual es una medida basada en términos de diferencia o distancia vectorial.

#	Entropía	Energía	Temperatura	F (asociación)	1-F (disociación)
1.1	0	1	1	1	0
1.2	0.235	0	0.05	-0.01175	1.01175
1.3	0.101	0.0756	0.103	0.065197	0.934803
2.1	0	1	1	1	0
2.2	0.16	0	0.044	-0.00704	1.00704
2.3	0.083	0	0.107	-0.008881	1.008881

Tabla 4 – Valor de F para cada par de oraciones

Por lo tanto, los pares de oraciones 1.1, 2.1 y 3.1 con valores de asociación y disociación de 1 y 0, respectivamente, determinan que la similitud (asociación) es bastante alta y la disociación es nula. Es decir, a valores de *F* más cercano a 1, los pares de oraciones tienden a ser más parecidos, caso contrario tienden a no serlo.

Los resultados se evaluaron utilizando las métricas de precisión y cobertura:

Precisión: esta métrica se calcula a partir del número de oraciones clasificadas correctamente por el sistema a partir del total de oraciones dentro del corpus. Es la fracción de instancias relevantes entre el total de las instancias recuperadas.

$$precisión = \frac{|\{oraciones\ relevantes\}| \cap |\{oraciones\ recuperadas\}|}{[\{oraciones\ recuperadas\}]}$$

Cobertura: también conocida como exhaustividad, es la fracción de la cantidad total de instancias relevantes que realmente se recuperaron, es decir, es la cantidad de oraciones clasificadas correctamente a partir del total de oraciones clasificadas.

$$cobertura = \frac{|\{oraciones\ relevantes\}| \cap |\{oraciones\ recuperadas\}|}{[\{oraciones\ relevantes\}]}$$

Los grupos I y J fueron sometidos a evaluación utilizando como referencia los resultados de la Energía Textual del modelo propuesto originalmente por Fernández, SanJuan & Torres-Moreno (2007), para comparar los resultados implementando la Energía Textual con la propuesta del uso de la Entropía y Temperatura Textual.

En la Tabla 5 se muestran los valores de precisión (P) y cobertura (C) para ambas implementaciones (modelo original y nuestra propuesta) para los grupos I y J, con diferentes valores de corte (Corte).

	Grupo I				**Grupo J**			
	Original		**Propuesto**		**Original**		**Propuesto**	
Corte	**P**	**C**	**P**	**C**	**P**	**C**	**P**	**C**
1	61	17	60	35	61	17	60	35
2	61	17	61	36	61	17	61	36
3	62	18	61	38	62	18	61	38
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
7	64	20	66	40	64	20	66	40
8	**69**	**22**	**72**	**41**	**69**	**22**	**72**	**41**
9	69	21	61	40	69	21	61	40

Tabla 5 – Umbral con mejores resultados

Se observa que los resultados de precisión y cobertura son más altos para un valor de corte igual a 8.

En la Tabla 6 se muestran resultados de experimentos, pero considerando valores de corte en el rango de 0.80 a 0.89. Se observa que los valores 0.80, 0.81, 0.87 y 0.88 fueron los umbrales con mayor recuperación de información útil para ambos grupos.

	Grupo I				Grupo J			
	Original		Propuesto		Original		Propuesto	
Corte	**P**	**C**	**P**	**C**	**P**	**C**	**P**	**C**
0.80	69	20	**72**	**41**	**0.80**	69	20	**72**
0.81	68	20	**72**	**41**	**0.81**	68	20	**72**
0.82	68	20	71	40	0.82	68	20	71
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.86	69	22	59	40	0.86	69	22	59
0.87	**70**	**22**	55	40	**0.87**	**70**	**22**	55
0.88	**70**	**22**	54	40	**0.88**	**70**	**22**	54

Tabla 6 – Umbral con mejores resultados

En conclusión, se determinó que implementar entropía y temperatura textual a una Red Neuronal Recurrente de Hopfield mejora significativamente el umbral de recuperación de información, tal y como se muestra en la Tabla 7.

Energía Textual (mínimo local)			
	Precisión	**Cobertura**	**Umbral**
Grupo I - Grupo J	70	22	46%
Grupo I - Grupo J	72	41	56.6%

Tabla 7 – Umbral original de la RNR de Hopfield

$$Threshold = \frac{(precision * recall)}{2}$$

Tomando como base la información de la tabla anterior es posible determinar que el umbral de ganancia de información (Gain' Umbral F-E) incrementó un 10.5% como se muestra en la Tabla 8. El valor de umbral se encuentra dado por:

Modelo	**Umbral de detección de paráfrasis**
Modelo prévio (Energía Textual (E))	46%
Modelo propuesto (F = U - (T · S))	56.5%

Tabla 8 – Incremento del umbral

La experimentación con el enfoque propuesto representa una mejora sobre el modelo original, que debe contrastarse con trabajos relacionados. En la Tabla 9 se muestra una comparativa entre el modelo propuesto y trabajos que hicieron uso del mismo corpus. Las medidas que se usaron para realizar esta comparación son

Exactitud (*accuracy*) y medida F1, las cuales se definen de la siguiente manera:

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Modelo	***Exactitud***	**F1**
(Socher et al., 2011)	76.8%	83.6%
(Milajevs et al., 2014)	73.0%	82.0%
(Mihalcea et al., 2006)	65.4%	75.3%
(Qiu et al., 2006)	72%	81.6%
Nuestro trabajo	**48%**	**53.98%**

Tabla 9 – Comparación de resultados

La comparación muestra que aún es necesario realizar mayor experimentación bajo el enfoque propuesto para obtener resultados competitivos. El incremento en la ganancia de información respecto al modelo anterior resulta alentador pues define una posible línea de investigación que puede continuar explorándose.

6. Conclusiones

El método propuesto permitió desarrollar un Contexto de Afinidad Libre (CAL), capaz de detectar paráfrasis y clasificar una colección de documentos de acuerdo con su género contextual. Se proponen dos nuevos patrones de comportamiento, entropía y temperatura textual respectivamente, que mejoran el proceso de detección de paráfrasis mediante una Red Neuronal Recurrente de Hopfield.

El modelo genera un Contexto de Afinidad Libre a partir del mínimo local que converge a través de una Red de Hopfield, que sirve para indicar y cuantificar la distribución semántica que pueda existir en un conjunto de documentos de texto. Se logró detectar el 56.5% de la paráfrasis existente en un corpus de pares de oraciones, generando un umbral de precisión de 72% y una cobertura del 41%. Lo que refiere, en suma, un aumento de 10.5% con respecto a la versión que utiliza únicamente Energía Textual.

La utilidad teórica de este trabajo está estrechamente relacionada con la Búsqueda y Recuperación de Información. Respecto a su utilidad práctica, el empleo del modelo desarrollado para encontrar paráfrasis se recomienda únicamente para textos cortos (como títulos o snippets), pues calcular la Energía, Temperatura y Energía Textual tiene un costo computacional elevado. Tareas de Procesamiento de Lenguaje Natural implementadas en esta investigación (por ejemplo, etiquetado PoS, lematización, detección de entidades nombradas, entre otras) tienen una complejidad cuadrática (O(n^2)). Por otro lado, la complejidad del coseno suave (entropía) es cuadrática para las operaciones entre vectores dispersos dado que solo toma en cuenta las dimensiones distintas a cero. Finalmente, el cálculo de la configuración de energía en el modelo de Ising se considera NP-Hard. Respecto a la complejidad de espacio, se demandan 1.3 GB y 804.82 MB de memoria para las ejecuciones del algoritmo completo, omitiendo las tareas de PLN.

La experimentación del modelo con las 2000 oraciones demoró 17.3 horas en un equipo con procesador AMD Ryzen 5 y 16 GB de RAM sin GPU dedicada.

Agradecimientos

Parte de esta investigación fue patrocinada por CONACYT (becas de posgrado y SNI), proyecto A1-S-27780, PRODEP, TecNM.

Referencias

Copeland, M. (2016). The Difference Between AI, Machine Learning, and Deep Learning?: NVIDIA Blog. Recuperado el 8 de Octubre de 2020, de https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Fausett, L. (1994). Fundamentals of Neural Networks: Architectures, Algorithms, and Applications. Prentice-Hall, Inc.

Fernández, S., Sanjuan, E., & Torres-Moreno, J.-M. (2007). Énergie textuelle de mémoires associatives. In *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles* (pp. 25–34). Toulouse: Association pour le Traitement Automatique des Langues. https://hal.archives-ouvertes.fr/hal-01320370

Firth, J. R. (1957). A Synopsis of Linguistic Theory 1930-55. In *F. R. Palmer (ed.). 1968. Selected papers of JR. Firth* 1952-59, (pp. 1-32). https://www.semanticscholar.org/paper/A-Synopsis-of-Linguistic-Theory-1930-1955%22-in-in-Firth/32f140fbb9514fd3ead5177025c467b50896db30

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT press. https://www.deeplearningbook.org/

Grzymala-Busse, J. (1993). Selected Algorithms of Machine Learning from Examples. *Fundam. Informaticae, 18*, 193-207.

Hopfield, J. (5 de 1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. In *Proceedings of the National Academy of Sciences of the United States of America (vol.79*, pp. 2554-8). https://doi.org/10.1073/pnas.79.8.2554

Lan, W., & Xu, W. (2018). Neural Network Models for Paraphrase Identification, Semantic Textual Similarity, Natural Language Inference, and Question Answering. In *Proceedings of the 27th International Conference on Computational Linguistics (pp*. 3890–3902). Association for Computational Linguistics. https://www.aclweb.org/anthology/C18-1328

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence (vol. 1,* pp. 775–780). AAAI Press.

Milajevs, D., Kartsaklis, D., Sadrzadeh, M., & Purver, M. (2014). Evaluating Neural Word Representations in Tensor-Based Compositional Settings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 708–719). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1079

Mohamed, I., & Gomaa, W. (2019). Exploring the Recent Trends of Paraphrase Detection. *International Journal of Computer Applications, 182*, 1-5. https://doi.org/10.5120/ijca2019918317

Qiu, L., Kan, M.-Y., & Chua, T.-S. (2006). Paraphrase Recognition via Dissimilarity Significance Classification. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 18–26). Association for Computational Linguistics.

Real Academia Española (2020). Entropía. En Diccionario de la lengua española (23.ª ed.). Recuperado el 8 de octubre de 2020, de https://dle.rae.es/entrop%C3%ADa

Reinsel, D., Gantz, J., & Rydning, J. (2018). The digitization of the world from edge to core. https://www.seagate.com/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

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

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., & Manning, C. D. (2011). Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In *Procedings of the 24th International Conference on Neural Information Processing Systems* (pp. 801–809). Curran Associates Inc.

Stepanov, I. (2012). Exact solutions of the one-dimensional, two-dimensional, and three-dimensional Ising models. Nano Science and Nano Technology: *An Indian Journal, 6*, 118-122.

Torres-Moreno, J.-M., Molina, A., & Sierra, G. (2010). La energía textual como medida de distancia en agrupamiento de definiciones. *Proceedings of 10th International Conference of Statistical Analysis of Textual Data* (pp. 215-226).

Wolf, W. P. (2000). The Ising model and real magnetic materials. *Brazilian Journal of Physics, 30*(4), 794–810. https://doi.org/10.1590/s0103-97332000000400030