UNIVERSIDAD CATÓLICA DE LA SANTÍSIMA CONCEPCIÓN

Facultad de Ingeniería

Ingeniería Civil Informática

COMBINACIÓN DE MÉTRICAS Y RASGOS LÉXICO-SEMÁNTICOS PARA EL ANÁLISIS DE SIMILITUD TEXTUAL ENTRE DOS FRASES

SEBASTIÁN NICOLÁS OLIVA ARENAS

INFORME DE PROYECTO DE TÍTULO PARA OPTAR AL TÍTULO DE

INGENIERO CIVIL INFORMÁTICO

Profesor Guía

Jose Abreu Salas

Concepción, Agosto 2017

Resumen

A partir del procesamiento del lenguaje natural, ha surgido una gama de problemas a resolver durante los años, y uno de ellos ha sido la similitud semántica textual.

La similitud semántica textual, problema que tiene aplicaciones en variados tópicos, como por ejemplo en textos de resumen, traducción automática, la mejora de la eficacia de los motores de búsqueda semánticos, educación como revisión de respuestas breves. Resolver y optimizar las aplicaciones de las áreas en general tiene mucho interés en la comunidad científica. Lo que hacen los algoritmos hoy en día es dar una puntuación de similitud a las frases que se comparan a través de ciertas métricas. Si bien se han hecho conferencias para resolver este tipo de problema, ya hay variados enfoques que dan una puntuación a las frases similares, aún no se ha logrado dar con un enfoque exacto para resolver este problema. Lo que propuso esta investigación para abordar el problema, fueron cuatro enfoques combinando métricas tanto semánticas y léxicas, desambiguando las frases de dos maneras distintas y entrenando los datos con algoritmos de aprendizajes automáticos. Por ende, una hipótesis que siguió esta investigación fue al combinar métricas tanto léxicas como semánticas se puede obtener mejores resultados.

Los experimentos realizados con el modelo propuesto en esta investigación, permitieron ver que el enfoque A da mejores resultados, pero con la prueba de Wilcoxon se concluyó que el enfoque A no tiene mayor relevancia que el enfoque B en los modelos utilizados (Random Forest, Dagging, Linear Regression, SMOreg).

Abstract

This research covered textual semantic similarity, a problem that has applications in various areas, such as summary texts, automatic translation, improving the effectiveness of semantic search engines, education as a review of short answers. To address the problem, there were four proposed approaches, combining both semantic and lexical metrics, disambiguating sentences in two different ways and training data with automatic learning algorithms.

The experiments performed show the results of the four proposed approaches to the problem, trained in four different algorithms. Although the results were not optimal, there are results showing which combination is best and which model is the best of the four analyzed.

Índice

Índice de figuras

Índice de tablas

Capítulo 1

1. Introducción

El problema de la similitud semántica textual ha sido abordado por diversos enfoques, como también competiciones, una de ellas es SemEval, que es un Workshop, competición internacional, donde invita a participar proponiendo diferentes temas de investigación. En general, alrededor de 30 grupos participantes promedio por tema obtiene SemEval cada año, donde cada grupo aborda con diferentes enfoques los temas propuestos. Saber si dos frases tienen el mismo significado o no, es vital para una buena comunicación. Existen técnicas y recursos para abordar este problema, como WordNet, Wikipedia y ontologías como SUMO. Además, se ha abordado el problema a través del modelo del espacio vectorial de recuperación de información, en el que cada texto se modela como una bolsa de palabras y se representa usando un vector. Otro enfoque con el cual se ha abordado es a través de la suposición de que si dos frases de textos cortos son semánticamente equivalentes, se debe ser capaz de alinear sus palabras o expresiones. La alineación sirve como una medida de similitud.

Para esta investigación, se modela el problema como 2 frases en un hiperespacio, las cuales se necesita obtener la distancia de aquellas frases. Para ello se seleccionan un conjunto de métricas, cada una con formulas diferentes que entregan distinta información, por ende, entrega un valor representando la distancia de aquellas frases.

Como análisis general, dada dos frases, la similitud semántica textual se mide a través de la asignación de un valor en la escala de 0 a 5 donde:

- 0: Las dos frases son completamente diferentes.
- 1: Las dos frases no son similares, pero están en el mismo tema.
- 2: Las dos frases no son similares, pero comparten algunos detalles.
- 3: Las dos frases son medianamente similares.

- 4: Las dos frases son altamente similares, pero algunos detalles difieren.
- 5: Las dos frases son completamente similares.

Por lo que es un problema de regresión lineal. Este problema aborda áreas tales como la traducción automática, motores de búsqueda semántica, lo cual permite poder mejorar tales áreas en un futuro próximo.

Se bien se han propuesto variados enfoques para modelar la similitud textual. La combinación de diferentes métricas léxicos-semántica, ha brindado buenos resultados, ya que estas intentan capturar similitudes entre las frases a diferentes niveles (léxicos y semánticos). En el trabajo se propone estudiar el efecto de enriquecer un modelo basado en la combinación de métricas mediante la inclusión de nuevas métricas, siendo la hipótesis que al combinar métricas es bueno para determinar la similitud semántica.

1.1. Objetivo general

Proponer un enfoque basado en la combinación de métricas y rasgos léxicos-semánticos para medir el grado de similitud textual entre dos frases y estudiar el efecto que tiene variar el modo en que se desambigua.

1.2. Objetivos específicos

- Revisar bibliografía sobre similitud semántica.
- Definir un esquema para combinar las métricas de similitud léxica y semántica.
- Validar experimentalmente la propuesta.

1.3. Delimitación del problema

Análisis de similitud semántica textual entre 2 frases cortas.

Idioma solo inglés.

Corpus SemEval[1] de los años 2012 - 2013 - 2014 - 2015 - 2016.

Métricas:

- Leacock & Chodorow similarity (Leacock and Chodorow, 1998).
- Wu and Palmer (Wu and Palmer, 1994).
- Resnik (Resnik, 1995).
- Lin (Lin, 1998).
- Jian & Conrath (Jiang and Conrath, 1997).
- Path Length (Pedersen et al., 2004).
- Similitud de palabras (Chavez et al., 2014).

1.4. Justificación del problema

Existen enfoques previos con resultados alentadores basados en la combinación de métricas, por lo que, como elementos que justifican tratar este problema se encuentran:

- La oportunidad de mejorar resultados y potenciales aplicaciones.
- Los resultados de hoy en día no son 100 % exactos.
- Diversas áreas donde se aplica la similitud semántica textual.

1.5. Metodología

1.5.1. Revisión bibliográfica sobre similitud semántica.

- Búsqueda de artículos relacionados con la similitud semántica.

[1] Workshop y competición internacional para abordar problemas del lenguaje natural

- Análisis crítico de los materiales recopilados.

1.5.2. Definir un esquema para combinar las métricas de similitud léxicas y semánticas.

- Identificar las métricas que se incluirán en el estudio.
- Definir un modelo de integración de las métricas.
- Implementar el modelo.

1.5.3. Validación experimental de la propuesta.

- Ajustar el modelo.
- Preparación de los corpus.
- Organizar experimentos de prueba.
- Analizar resultados de experimentos.

Capítulo 2

2. Marco teórico

Este proyecto de investigación tiene como objetivo proponer un enfoque a través de la combinación de métricas y rasgos léxico-semánticos para el análisis de la similitud semántica textual.

En resumen, se realizó un pre-procesamiento de los textos que consiste en tokenizar, taggear y eliminar stopwords, para luego realizar una desambiguación de las frases y extraer el sentido dependiendo del contexto en el cuál ocurre. Se siguió con el proceso de medir las distancias con métricas tanto léxicas como semánticas a nivel de frase, donde cada métrica genera una matriz en función de costo para luego con el algoritmo húngaro reducir la matriz a un solo valor.

Se generó un vector con todos los valores de las métricas para cada frase con el objetivo de ser entrenados con un algoritmo de aprendizaje automático. Existen dos tipos de aprendizaje, supervisado y no supervisado. El aprendizaje supervisado se refiere a los problemas de clasificación. Básicamente necesitan un conjunto de datos de entrenamiento para la supervisión del aprendizaje, para luego entregar en su salida una predicción de su variable dependiente. El aprendizaje no supervisado se refiere a los problemas de agrupación. Básicamente los datos no están clasificados por clases, de esta forma el aprendizaje no se supervisa. Para esta investigación se ocuparon algoritmos de aprendizaje automáticos supervisados.

Para este proceso, se debe conocer los conceptos y métricas que se necesitan para comprender la investigación.

2.1. *Stopwords*

Los stopwords son palabras que se producen con frecuencia en un documento, pero que no tienen sentido en términos de recuperación de información. Generalmente suelen ser preposiciones y artículos (Perkins, 2014).

2.2. Tokenizar

Es el proceso para dividir un fragmento de texto en muchas partes, ya sea por oraciones o por palabras. Esta unidad de división que queda después del proceso de tokenizar se conoce como “token”(Perkins, 2014).

2.3. *Part Of Speech*

Es una etiqueta para identificar la palabra si es un sustantivo, adjetivo, verbo entre otras. El proceso de etiquetado forma una tupla (palabra, etiqueta) por cada palabra, donde cada palabra lleva su etiqueta (Perkins, 2014).

2.4. N-gramas

Un n-grama es una subsecuencia de n-caracteres de una palabra. Si la subsecuencia es un caracter, se denomina uni-grama, si la subsecuencia es un par de caracteres, se denomina bi-gramas y así sucesivamente (Cavnar et al., 1994) .

2.5. *WordNet*

WordNet es una base de datos de léxicos de inglés. Se puede describir como un diccionario de inglés. Agrupa a los sustantivos, verbos, adjetivos y adverbios en conjuntos de

sinónimos (*synsets*), cada uno de ellos expresando un concepto distinto. Los *synsets* se organizan en sentidos, dando así los sinónimos de cada palabra, y también en relaciones como hipónimo / hiperónimo (es decir, ES-UN), y meronimia / holónimo (es decir, PARTE-DE). *WordNet* es una red semántica de interconexión y grupos de palabras por medio de relaciones léxicas y conceptuales representados por etiquetados de dominio (Fellbaum, 2005).

2.6. *Synsets*

Los *synsets* son un grupo de sinónimos de una palabra, cada uno de ellos expresando un concepto distinto. Una palabra puede tener muchos *synsets* como también solo uno (Perkins, 2014).

2.7. Algoritmo Húngaro

El algoritmo Húngaro (Kuhn, 1955) es un método de optimización para problemas de asignación de costos. El algoritmo modela un problema de asignación como una matriz de costes $n \times m$.

2.8. Máquina vectores de soporte (SVM) y *Dagging*

La máquina de soporte de vectores (SVM), es un algoritmo para la clasificación de datos lineales y no lineales. El SVM utiliza una proyección no lineal para transformar los datos de entrenamiento en una dimensión superior, en la cual busca el hiperplano de separación óptimo lineal. El SVM encuentra este hiperplano usando vectores de soporte (tuplas de entrenamiento) y márgenes definidos por los vectores de soporte (Han et al., 2011).

Dagging es un modelo el cuál combina varios modelos usando el mismo algoritmo de aprendizaje para los modelos base (Ting and Witten, 1997). Crea un número de particiones

disjuntos fuera de los datos y alimenta cada fragmento de datos a una copia del clasificador base suministrado.

2.9. Regresión Lineal

El modelo de regresión lineal es un algoritmo que busca la mejor línea para ajustar dos atributos, donde uno es usado para predecir el otro. La regresión lineal múltiple implica dos o más atributos, donde los datos se ajustan a una dimensión lineal (Han et al., 2011).

2.10. *Random Forest*

Random Forest es un modelo que contiene un conjunto de modelos, donde cada modelo es un árbol de decisión. Cada árbol de decisión se genera usando una selección aleatoria de atributos en cada nodo para determinar la división (Han et al., 2011).

2.11. *Cross-Validation*

La validación cruzada corresponde a la técnica de dividir aleatoriamente k-particiones los datos en un subconjunto de particiones para un entrenamiento. Su función es tomar una partición para medir la precisión de la predicción (prueba) y las demás particiones se utilizan para entrenar (Han et al., 2011).

2.12. Prueba no paramétrica Wilcoxon

La prueba no paramétrica Wilcoxon se aplica cuando no se puede realizar la prueba t. Es la alternativa no paramétrica a la comparación de dos promedios independientes a través de la t-student. Se utiliza cuando se quiere realizar la comparación de dos grupos en quienes se les ha medido una variable cuantitativa continua que no tiene una distribución normal

o cuando la variable es de tipo cuantitativa discreta (Gómez-Gómez et al., 2003). Tiene 3 hipótesis:

1. La variable independiente es dicotómica y la escala de medición de la variable dependiente es al menos ordinal.
2. Los datos son de muestras aleatorias de observaciones independientes de dos grupos independientes, por lo que no hay observaciones repetidas.
3. La distribución de la población de la variable dependiente para los dos grupos independientes comparte una forma similar no especificada, aunque con una posible diferencia en las medidas de tendencia central.

2.13. Análisis semántico latente

Una métrica para calcular el grado de similitud semántica entre palabras es el análisis semántico latente (LSA), la cuál extrae y ocupa información derivada de un gran corpus de texto (Landauer et al., 1998). Para aquello, en primera instancia, se representa el texto como una matriz, donde cada fila es una palabra y cada columna es un contexto. El valor de cada celda es la frecuencia con la que se da la palabra en el contexto. Luego, el LSA aplica la descomposición de valor singular (SVD). Consiste en que la matriz se descompone en el producto de otras tres matrices. En general, este enfoque sigue la hipótesis que las palabras que se producen en el mismo contexto tienden a tener significados similares.

Una de las variaciones de LSA es la llamada hiperespacio analógico para el lenguaje (HAL) (Burgess et al., 1998). HAL es un modelo que adquiere representaciones de significado al capitalizar la información de co-ocurrencia a gran escala. HAL se basa en la co-ocurrencia dentro de un contexto común. Esto consiste en contar el número en que aparecen dos palabras en la distancia n, llamada ventanas. Las ventanas son el número de palabras intermedias entre dos palabras (Chavez et al., 2014). Para calcular el grado de similitud de HAL entre dos palabras se utilizó la medida del coseno.

2.14. Similitud de coseno

Una métrica muy utilizada en el procesamiento del lenguaje es la similitud del coseno, que calcula el ángulo entre dos vectores. Comúnmente se utiliza la similitud del coseno cuando el espacio es positivo, donde el resultado de la métrica se limita al intervalo [0,1]. Ejemplo de la utilización de esta métrica es en la recuperación de información, en término de documentos, estos son representados como vectores, la similitud de dos documentos corresponde al ángulo del coseno (Huang, 2008).
La similitud del coseno esta dada por:

$$Sim_{cos}(x,y) = \frac{x \cdot y}{||x||\,||y||} \tag{2.1}$$

Donde:
$||x||$ y $||y||$ son vectores bajo la norma euclideana $x = (x_1, x_2, ..., x_i)$, $y = (y_1, y_2, ..., y_i)$, definido como $\sqrt{x = x_1^2, x_2^2, ..., x_i^2}$, $\sqrt{y = y_1^2, y_2^2, ..., y_i^2}$. Conceptualmente es el ángulo de los vectores (Han et al., 2011).

2.15. Métricas semánticas

Antes de definir las métricas, es necesario saber las propiedades que definen una métrica. Una función de distancia D con valores reales no negativos, definida en el producto cartesiano $X \bullet X$ del conjunto X es llamada una métrica de X, si para cada valor $x, y, z \in X$ se cumple:

- $d(x, y \geq 0$.
- $d(x, y) = 0$.
- $d(x, y) > 0$ cuando $x \neq y$.
- $d(x, y) = d(y, x)$ (simetría).
- $d(x, z) \leq d(x, y) + d(y, z)$ (desigualdad triangular).

2.15.1. Métrica Wu and Palmer

La métrica Wu and Palmer se centra en el impacto que tienen los verbos en los sistemas de traducción automática (Wu and Palmer, 1994).
Dada una ontología, formados por un nodo raíz R y un conjunto de nodos, n_1 y n_2 son elementos de la ontología a los cuáles se medirá la similitud. La métrica de similitud Wu and Palmer se define con la siguiente formula:

$$Sim_{wp} = \frac{2d_3}{d_1 + d_2 + 2d_3} \tag{2.2}$$

Donde:
d_1 y d_2 son las distancias entre R y n_1, n_2.
NC es el nodo común que comparten n_1 y n_2.
d_3 es la distancia entre el nodo común NC y el nodo raíz R.

2.15.2. Métrica *PathLength*

Pathlength es una métrica que se basa en la longitud del camino entre un concepto n_1 y un concepto n_2. Entre más cerca está el concepto n_1 del concepto n_2, mayor será su similitud. Cabe destacar que la longitud entre los conceptos está dada por el número de aristas (Pedersen et al., 2004). La métrica PathLength está dada por la siguiente formula:

$$Sim_{pathlength} = -\log pathlength(n_1, n_2) \tag{2.3}$$

Donde:
$pathlength(n_1, n_2)$ es el número de aristas del camino más corto entre los conceptos n_1 y n_2.

2.15.3. Métrica Lin

Lin es una métrica de similitud semántica que se basa en la comparación de dos conceptos A y B en una taxonomía (Lin, 1998); apoyándose en las siguientes ideas intuitivas:

- Intuición1: "La similitud entre A y B está relacionada con sus elementos comunes. Entre más elementos comunes comparten, más similares son."
- Intuición2: "La similitud entre A y B está relacionada con las diferencias entre ellos. Entre más diferencias tienen, menos similares son."
- Intuición3: "La máxima similitud entre A y B se alcanza cuando A y B son idénticos, no importa cuánto puntos en común comparten."

Luego, la fórmula de similitud de Lin está dada por:

$$Sim_{lin} = \frac{2\log P(n_0)}{\log P(n_1) + \log P(n_2)} \tag{2.4}$$

Donde:
se asume que la taxonomía es un árbol, $P(n_0)$, $P(n_1)$ y $P(n_2)$ son probabilidades de los nodos.
n_1 y n_2 son los nodos elegidos a comparar.
n_0 es el nodo específico que subsume en tanto a n_1 y n_2.

2.15.4. Métrica Resnik

La métrica de similitud de Resnik considera solo la relación taxonómica "Es-Un", donde se basa en la noción del contenido de información.

La argumentación estándar de la teoría de la información, dice que el contenido de información de un concepto c se puede cuantificar como el logaritmo de verosimilitud negativa.

A través de aquella argumentación, la cuantificación del contenido de información toma forma intuitiva en aquél contexto: a medida que aumenta la probabilidad, el carácter informativo disminuye, por lo que el concepto más abstracto, menor será su contenido de

información. Por otra parte, si hay un concepto superior único, su contenido de información es 0 (Resnik, 1995).

$$Sim_{resnik} = -\log P(n) \tag{2.5}$$

$$P(n) \in (c_1, c_2)$$

Donde:

$P(n)$ es la probabilidad del nodo dominador, más bien, la probabilidad del nodo que subsume a los conceptos c_1 y c_2.

2.15.5. Métrica Jiang & Conrath

La métrica propuesta por Jiang y Conrath (Jiang and Conrath, 1997), trata de un enfoque donde se basa en la noción de aristas. Jian y Conrath dicen: "La distancia semántica entre dos nodos es la diferencia de su masa semántica si están en el mismo eje, o la adición de las dos distancias calculadas a partir de cada nodo a un nodo común, en el que dos ejes cumplen si los dos nodos originales están en diferentes ejes. Es fácil demostrar que la medida de distancia propuesta también satisface las propiedades de una métrica".

$$Sim_{jc} = \frac{1}{Dist(c_1, c_2)} \tag{2.6}$$

Donde:

$Dist(c_1, c_2) = \log(n_1) + \log(n_2) - 2\log(n_0)$.

n_0 es el nodo común que subsume en tanto a n_1 y n_2.

c_1 y c_2 son los conceptos que se comparan, n_1 y $n_2 \in (c_1, c_2)$.

2.15.6. Métrica Leacock & Chodorow *Similarity*

La métrica propuesta por Leacock y Chodorow (Leacock and Chodorow, 1998) se basa en la medida de distancia entre los conceptos A y B, lo cuál implica seleccionar el camino más

corto en una taxonomía. Para calcular la longitud del camino de similitud entre los conceptos A y B, se ocupa la siguiente formula:

$$Sim_{lc} = -\log(\frac{n}{2d}) \tag{2.7}$$

Donde:

n es el número de nodos en el camino más corto entre los conceptos A y B.

d es la profundidad máxima en la taxonomía.

2.15.7. Métrica Similitud de palabras

Todas las métricas de similitud a nivel de sentido, pueden convertirse en una medida de similitud de palabras calculando la máxima similitud entre todos los sentidos posibles (Chavez et al., 2014).

$$WS(w1,w2) = max_{\substack{s1 \in sentidos(w1) \\ s2 \in sentidos(w2)}} sim(s1,s2) \tag{2.8}$$

Donde:

$sim(s1,s2)$ es una de las métricas semánticas a nivel de sentidos previamente descritas.

2.15.8. Máxima similitud de palabras

En (Chavez et al., 2014) proponen 2 modelos agrupando algunas métricas ya descritas. La máxima similitud de palabras se define como:

$$MaxSim(w1,w2) = \begin{cases} 1 & \text{si} \quad QGDistance(w1,w2) = 1 \\ Max\left(Sim_{hal}\left(w1,w2\right), Sim_{wup}\left(w1,w2\right)\right) \end{cases} \tag{2.9}$$

Donde:

$QGDistance(w1,w2)$ es la distancia léxica QGram entre las palabras $w1$ y $w2$.

$Sim_{hal}(w1,w2)$ es el análisis semántico latente previamente descrito entre las palabras $w1$ y $w2$.

$Sim_{wup}(w1,w2)$ es la métrica Wup and Palmer para las palabras $w1$ y $w2$.

2.15.9. Métrica Estadística y relación de peso

Para el cálculo de relación de peso se utilizó la siguiente métrica (Chavez et al., 2014):

$$StaWeiRat(w1,w2) = \frac{\left(Sim_{hal}(w1,w2) + \left(\frac{1}{WeiRat(w1,w2)}\right)\right)}{2} \tag{2.10}$$

Donde: $Sim_{hal}(w1,w2)$ es el análisis semántico latente previamente descrito entre las palabras $w1$ y $w2$.

$WeiRat(w1,w2)$ toma los valores basados en la relación entre las palabras $w1$ y $w2$. Esta relación se basa en la siguiente tabla:

Tabla 1: Tabla valores Weight Ratio.

Valor	**Relación entre w1 y w2**
10	Antónimo
1	Sinónimo
2	Hiperónimo
3	Hipónimo
3	Una palabra se encuentra frecuentemente en la glosa de otra
9	Otro

2.16. Rasgos léxicos

Los rasgos léxicos o métricas léxicas, para la similitud textual, son atributos que se basan en medidas de distancias entre palabras. A continuación se presenta una breve descripción de cada rasgo ocupado en esta investigación. La librería que ocupó esta investigación es SimMetrics library v1.5 for .NET 2.0.

2.16.1. *Dice-Similarity*

Con el objetivo de medir dos cadenas, *Dice-similarity*, también conocida como Sorensen-Dice coefficient (Sørensen, 1948), calcula el coeficiente a partir de dos secuencias de caracteres ocupando bi-gramas:

$$d_{dice} = \frac{2 * n_t}{n_x + n_y} \tag{2.11}$$

Donde:

n_t es el total de caracteres bi-gramas encontrados en las cadenas x e y.

n_x el número de bi-gramas en la cadena x.

n_y es el número de bi-gramas en la cadena y.

2.16.2. *Euclidean Distance*

Una de las distancias más comunes es la distancia Euclideana (*Euclidean Distance*) (Cha, 2007), se define como:

$$d_E = \sqrt{\sum_{i=0}^{n} (p_i - q_i)^2} \tag{2.12}$$

Donde:

p y q son los puntos de cada objeto.

2.16.3. Jaccard

Otra distancia común es la distancia de Jaccard que se define como el tamaño de la intersección de la muestra dividido entre el tamaño de la unión de la muestra (Sun et al., 2015).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.13}$$

Donde:

A y B son las muestras a comparar.

2.16.4. Jaro

Jaro es una métrica que dada dos cadenas x e y, la distancia entre ellas se puede calcular a través de la ecuación:

$$d_j(x,y) = \frac{1}{3}\left(\frac{m}{|x|} + \frac{m}{|y|} + \frac{m-t}{m}\right) \tag{2.14}$$

Donde:

m: número de caracteres que coinciden.

t: número de transposiciones.

2.16.5. Jaro-Winkler

Winkler propone una variante a la métrica Jaro, dice que dos cadenas tienen un prefijo común l (Sun et al., 2015).

$$d_w(x,y) = d_j(x,y) + \big(lp(1 - d_j(x,y))\big) \tag{2.15}$$

Donde:

$d_j(x,y)$: distancia de Jaro entre las cadenas x e y.

p: Es una variable, y usualmente p < 0.25.

2.16.6. Levenshtein

La distancia de Levenshtein es una métrica que toma el valor de la diferencia entre dos cadenas (Hirschberg, 1997). La distancia de Levenshtein se define de acuerdo a:

- Inserciones, sustituciones y borrados, operaciones que permiten convertir una cadena A en una cadena B.
- Asignado un costo a cada tipo de operación, se busca la secuencia de operaciones para convertir de A a B con el menor costo posible.

2.16.7. *Overlap Coefficient*

La métrica Coeficiente de Solapamiento, más conocido como *Overlap Coefficient*, es una medida de similitud que se relaciona con el índice Jaccard. Se ocupa en el análisis de redes sociales (Matsuo et al., 2004). Esta medida calcula los solapamientos entre dos conjuntos:

$$OverlapCoefficient(X,Y) = \frac{|X \cap Y|}{min(|X|,|Y|)} \tag{2.16}$$

Donde:
X e Y son los conjuntos de palabras a medir.

2.16.8. *QGrams*

Otra distancia en la que se han obtenido buenos resultados es la llamada *QGrams Distance*, estos son simples subcadenas (n-gramas) de longitud q de una palabra dada (Ukkonen, 1992).

Considerando el siguiente ejemplo, los q-grams de longitud q=3 para la cadena “play soccer” son: f(1,##p), (2,#pl), (3,pla), (4,lay), (5,ay), (6,y s), (7, so), (8,soc), (9,occ), (10,cce),

(11,cer), (12,er%), (13,r%%), donde # y % indican el inicio y el fin de las cadenas respectivamente.

2.16.9. Smith Waterman

Smith Waterman es una modificación a la distancia de Levenshtein. Fue creada para identificar el alineamiento óptimo entre cadenas de ADN y secuencias de proteínas. Esta distancia penaliza el error de inserción y borrado de caracteres (Smith and Waterman, 1981). Posteriormente, surge una modificación a la distancia de Smith Waterman, esta modificación permite que existan caracteres no alineados en la secuencia (Gotoh, 1982).

2.16.10. *Block distance*

Block distance, más conocida como *Manhattan distance*, es definida la distancia entre dos puntos, como la suma de las diferencias absolutas entre sus puntos (Krause, 2012).

$$d_{ij} = \sum_{k=1}^{n} |p_{ik} - q_{jk}| \tag{2.17}$$

2.16.11. *Chapman Length Deviation y Chapman Mean Length*

Otras dos medidas simples, que se basan en la longitud de las cadenas comparadas son las llamadas *Chapman Length Deviation y Chapman Mean Length* (Chapman and Parkinson, 2006).

Chapman Mean Length es la diferencia entre las longitudes de las cadenas en comparación.

Chapman Length Deviation es la longitud media de Chapman, entregando una medida de similitud entre dos cadenas a partir del tamaño de la longitud media de los vectores.

2.16.12. Nedleman Wunch

Nedleman Wunch es una métrica similar a Levenshtein, la cual al igual que Smith Waterman, penaliza el error de inserción y borrado de caracteres en la secuencia de alineamiento. Esta métrica fue desarrollada como un método para calcular la similitud entre dos proteínas (Needleman and Wunsch, 1970).

2.16.13. Monge Elkan

Monge Elkan (Monge et al., 1996), es una métrica de emparejamiento recursivo, donde comparan la cadena x y la cadena y, ambas se dividen en sub-cadenas y cada sub-cadena de x se compara con cada sub-cadena de y, se define como:

$$match_{(x,y)} = \frac{1}{|x|}\sum_{i=1}^{|x|} max_{j=1}^{|y|} match\left(x_i, y_j\right) \tag{2.18}$$

2.16.14. *Simple Matching Coefficient*

Simple Matching Coefficient es una métrica que fue propuesta para ser utilizada en una taxonomía numérica (Sokal, 1958). Se puede utilizar sólo cuando se comparan conjuntos con tres o más variables a contrastar.

La ecuación que define a SMC es:

$$SMC = \frac{a+d}{a+b+c+d} \tag{2.19}$$

La tabla 2 (extraída de (Schuetz, 2011)) muestra el esquema de asignación que utiliza Simple Matching Coefficient :

Tabla 2: Tabla de asignación SMC.

	Objeto 1	
Objeto 2	*Número de variables con categoría 1*	*Número de variables con categoría 2*
Número de variables con categoría 1	*a*	*b*
Número de variables con categoría 2	*c*	*d*

Capítulo 3

3. Estado del arte

La similitud semántica textual ha tenido diversas aplicaciones a lo largo del tiempo, como en la recuperación de información, web semántica y en general en el proceso del lenguaje natural. Este problema ha sido abordado por diversos enfoques, como también competiciones, una de ellas es SemEval, que es un *Workshop*, competición internacional, donde invita a participar proponiendo diferentes temas de investigación. En general, alrededor de 30 grupos participantes promedio por tema obtiene SemEval cada año, donde cada grupo aborda con diferentes enfoques los temas propuestos.

En (Corley and Mihalcea, 2005) se propuso un enfoque combinando las 6 métricas que entrega *WordNet* (Wup[2], PathLength, Lin, Resnik, J&C[3], L&C[4]), basado en un sistema de aprendizaje automático. El pre-procesamiento de los textos consistió en tokenizar, taggear y colocar las palabras en un conjunto de clases. Luego, realizaron la búsqueda de la similitud semántica entre verbos y sustantivos, y solo coincidencia léxica entre adjetivos y adverbios. Para las pruebas, utilizaron el corpus de parafraseo de Microsoft, que consta de 4.076 pares de entrenamiento y 1.725 pares de pruebas, y el corpus PASCAL, consistente en 1.380 pares de hipótesis de prueba (580 pares de desarrollo y 800 pares de prueba). Los resultados obtenidos variaron dependiendo del conjunto de datos y el tipo de aprendizaje empleado. Para un modelo no supervisado los resultados variaron entre 0.583 a 0.688 y para un modelo supervisado los resultados variaron entre 0.589 a 0.715. Con esto, Rada y Courtney plantean que la combinación es un buen indicador para la similitud semántica textual.

En (Torres and Gelbukh, 2009) se realizó un enfoque, para determinar el grado de similitud semántica entre dos palabras, con 2 métricas entregadas por *WordNet*: Lin y J&C,

[2]Wu and Palmer

[3]Jiang & Conrath

[4]Leacock & Chodorow Similarity

desambiguando las frases con el algoritmo de Lesk y combinando todo lo mencionado anteriormente. Para evaluar la implementación, utilizaron 4 corpus diferentes: SENSEVAL-2, SENSEVAL-3, SEMEVAL y SEMCOR. Además, en los experimentos consideraron 2 estrategias "back-off" para aquellas palabras no abarcadas en la implementación, la primera consistió tomar el primer sentido como el mas frecuente y la segunda en tomar un sentido aleatorio. La precisión global se midió con la formula de el número de instancias correctas dividido por el número total de instancia.

Los resultados variaron, para cada estrategia en cada corpus, el resultado mas bajo para la estrategia "back-off" tomando el primer sentido como el mas frecuente fue de 37.4 %, en el corpus de SEMEVAL. Para la estrategia "back-off" tomando el sentido aleatorio, el resultado mas bajo fue de 36.3 %, en el corpus SEMEVAL. El resultado más alto, tomando el primer sentido como el mas frecuente fue de 61.6 %, en el corpus SEMCOR y el resultado mas alto, tomando el sentido aleatorio, fue de 55 %. Concluyeron que la combinación de las métricas utilizadas tienen más precisión que cada una por separado.

En (Bär et al., 2012), se utilizó un modelo de regresión logarítmico lineal y combina algunas medidas de similitud semántica. Este sistema utilizó las métricas de similitud de Jiang and Conrath (1997), Lin (1998), y Resnik (1995), además, utilizó análisis y medidas que se nombran a continuación:

1. Medidas simples basado en cadenas.
    - Medidas de similitud de cadenas.
    - Carácter / n-gramas de palabras.
2. Medidas de similitud semántica.
    - Análisis semántico explícito.
    - Pruebas de implicación.
    - Distribución de Thesauro.

3. Mecanismo de expansión de texto.
    - Sistema de sustitución léxica.
    - Traducción automática estadístico.
4. Medidas relacionadas con la estructura y estilo.

Para la combinación de todas las medidas que utilizaron en esta tarea, utilizaron el modelo de log-linear regression de WEKA, con una *cross-validation* de 10. El resultado que obtuvieron en términos de correlación de Pearson fue 0.857.

En (Šarić et al., 2012), se propuso un sistema para medir la similitud semántica, similar al primer lugar, donde combina múltiples medidas de similitud. Este sistema utiliza una máquina de aprendizaje supervisado, el vector de regresión de soporte (SVR), para combinar una gran cantidad de características calculadas a partir de pares de frases. Obtuvo un resultado en términos de correlación de Pearson de 0.8569. Las métricas utilizadas son Pedersen et al. (2004), Leacock and Chodorow (1998), y Lin (1998).

Las características que además se combinaron en el sistema, son las siguientes:

1. Características de superposición n-gramas.
2. WordNet - Aumento de superposición de la palabra.
3. Características sintácticas.

En (Croce et al., 2013), el sistema llamado UNITOR modela la similitud semántica textual como un problema de regresión combinando las características en el modelo vector de soporte (SV). El resultado que obtuvo fue de 0.7981 en términos de correlación de Pearson. El sistema ocupó diferentes estimaciones de los cuáles destaca:

1. Superposición léxica: Esta es una función de similitud básica que modela las frases como superposición léxica. Dado los conjuntos W_a y W_b de palabras que aparecen en dos textos genéricos ta y tb, LO es estimado como la similitud de Jaccard.
2. Semántica de composición distributiva: Otra de las funciones de similitud, se obtiene

al tener en cuenta la composición sintáctica de la información léxica de las frases. La información léxica básica se obtiene en un espacio de co-ocurrencia de palabras. Las palabras que aparecen en una frase se proyectan en un espacio. Una frase puede ser representada mediante la aplicación de una combinación lineal. La función de similitud entre dos frases es entonces el coseno de similitud entre sus correspondientes vectores.

Otra propuesta que combina rasgos léxico-semánticos es la descrita en (Chávez et al., 2013), utilizó el modelo *bagging* usando *REPtree* para el entrenamiento. Se consideraron métricas léxicas extraídas de la librería SimMetrics (Chapman and Parkinson, 2006). Ocuparon Needleman Wunch, Smith Waterman, Smith Waterman Gotoh, Smith Waterman Gotoh Windowed Affine, Jaro, Jaro-Winkler, Chapman Length Deviation, Chapman Mean Length, *QGram Distance, Block Distance, Cosine Similarity, Dice Similarity, Euclidean Distance, Jaccard Similarity, Matching Coefficient*, Monge Elkan y *Overlap Coefficient*. Los resultados variaron dependiendo del corpus empleado, el resultado más bajo en términos de correlación de Pearson fue de -0.00065 y el más alto fue de 0.6168. El sistema ocupó 4 estrategias diferentes para la extracción de rasgos, entre ellos:

1. Medidas de similitud basadas en cadenas.
2. Medidas de similitud semántica.
3. Alineamiento léxicos-semántico.
4. Alineamiento semántico.

En (Chavez et al., 2014), los autores propusieron algunas modificaciones, que incluyeron el empleo de SVM[5], *Dagging* y 25 rasgos léxicos - semánticos. Este sistema alcanzó el primer lugar para el idioma español en la competición SemEval 2014, pero en inglés, el mejor lugar que obtuvo fue el número 16. Se concluyó que este sistema obtuvo resultados importantes y puede ser aplicado en diferentes escenarios, como se hizo el 2014, participando en 3 tareas de SemEval. Dentro de los rasgos léxicos, el sistema ocupó *Dice-Similarity,*

[5]*Support Vector Machine*

Euclidean Distance, JaccardSimilarity, Jaro-Winkler, *Levenstein Distance, Overlap Coefficient, QGrams Distance*, Smith Waterman, Smith Waterman Gotoh, SmithWaterman Gotoh Windowed Affine. Los resultados en idioma inglés estuvieron en el rango de 0.4752 a 0.8127 en términos de correlación de Pearson, resultados que dependían del corpus empleado. En español, los resultados estuvieron en el rango de 0.78021 a 0.82539 en términos de correlación de Pearson.

En (Buscaldi et al., 2015), se propuso un sistema llamado SOPA, el cual mezcla diferentes rasgos. Utilizaron 3 modelos diferentes, *Support Vector Regression, Multi-Layer Perceptron y Random Forest*, obteniendo mayores resultados en *Random Forest*. Los rasgos totales que utilizaron fueron 16 para corpus en idioma inglés y 14 para corpus en idioma español, de las cuales destaca la métrica Wup, similitud basada en n-grama, coseno y levenshtein. Los resultados en términos de correlación obtenidos para el idioma inglés estuvieron en el rango de 0.5914 a 0.8414 con el modelo *Random Forest*, en el idioma español, los resultados obtenidos estuvieron en el rango de 0.5637 a 0.5655. Concluyeron que el sistema SOPA aún necesita ser analizado para poder ser mejorado.

3.1. Discusión

Tabla 3: Tabla resumen estado del arte.

Autor	**Resumen**	**Resultados**
(Corley and Mihalcea, 2005)	Enfoque combinando 6 métricas que entrega WordNet	Los resultados variaron entre 0.583 a 0.688 R
(Torres and Gelbukh, 2009)	Desambiguación por Lesk. Similitud semántica entre 2 palabras, Lin y J&C	61.6 % precisión
(Bär et al., 2012)	J&C, Lin, Resnik, Regresión logarítmico lineal	0.857 R
(Šarić et al., 2012)	PathLength, L&C, Lin, SVR	0.856 R
(Chávez et al., 2013)	17 métricas léxicas	0.616 R
(Chavez et al., 2014)	25 métricas léxicas y semánticas, SVM y *Dagging*	0.475 a 0.812 R
(Buscaldi et al., 2015)	SVR, *Multi-Layer Perceptron y Random Forest*, Wup, n-grama, coseno y levenshtein	0.841 R

La tabla 3 muestra un resumen con resultados y conceptos Según la revisión bibliográfica, investigaciones que se basan en la combinación tanto como métricas semánticas, léxicas, n-gramas y características de análisis para las frases, además de participar en SemEval, la mayoría obtuvo buenos lugares en la competencia de similitud textual. Lo que concluyeron Corley and Mihalcea (2005), la combinación es un buen indicador de similitud y además de los resultados obtenidos por la mayoría de los trabajos de combinación, se deduce que experimentar con más métricas se pueden obtener mayores resultados en términos de correlación.

Además, los modelos que se mencionan, como *Random Forest*, Regresión Logarítmico Lineal, *Support Vector Machine y Dagging* entre otros, se tomaron en cuenta por obtener buenos resultados. El sistema a seguir y que se asemeja más a lo que va de esta investigación es el expuesto en (Chavez et al., 2014), donde utiliza gran cantidad de rasgos léxicos y métricas semánticas. Además, los nuevos rasgos léxicos que se tomaron en cuenta para enriquecer el modelo, son tomados del mismo sistema que participó el 2013, donde se destacan 7 rasgos léxicos que no fueron tomados en cuenta el año 2014. Estos rasgos son Needleman Wunch, Jaro, Chapman Length Deviation, Chapman Mean Length, Block Distance, Matching Coefficient, Monge Elkan. Todo se justifica por los trabajos anteriores donde la combinación de métricas parece brindar buenos resultados. Además, se quiere probar el sistema desambiguando con el algoritmo de Lesk, para observar el efecto que pudiese tener el algoritmo mencionado.

Capítulo 4

4. Descripción de la propuesta

En este capítulo, se presentan las actividades que se realizaron para la preparación de los corpus y obtener los datos para luego ser entrenados con algún modelo basado en aprendizaje automático, siguiendo la metodología definida previamente. Esta investigación trabajó con herramientas entregadas por NLTK (*Natural Language Toolkit*).

De acuerdo a los trabajos reflejados en el estado del arte, muestran que combinando métricas, se podría obtener buenos resultados, hipótesis que siguió esta investigación.

Para este estudio, las métricas a nivel de palabra se utilizaron para definir las métricas a nivel de frase. Esto quiere decir, palabra a palabra se midieron para formar una matriz en función de costo y luego entregar un valor a través del alineamiento húngaro. Esto llevó a que cada métrica a nivel de frase toma un par de frases y representan su similitud mediante un número.

Como hay varias métricas a nivel de frase $(m1, m2, \ldots, mn)$ entonces una frase quedó representada por un vector de números, cada uno calculado con una métrica diferente. A cada vector se le añade además un número que representa la similitud textual de la frase. Con esos vectores se entrenan los algoritmos de aprendizaje (*Random Forest, Dagging, Linear Regression y SMOreg*).

Las métricas para la investigación de la similitud textual, se obtuvieron siguiendo el sistema del 2014 UMCC (Chavez et al., 2014) más 7 métricas léxicas nuevas, definidas previamente en el marco teórico y obtenidas del sistema UMCC 2013 (Chávez et al., 2013), estas se dividen en semánticas y léxicas. Además, para el estudio se usaron tres rasgos basados en n-gramas (bi-gramas, tri-gramas y tetra-gramas).

Las métricas semánticas que aportan información dependiendo de la estructura de la taxonomía, en este caso *WordNet*, son Wup and Palmer, *PathLength* y Leacock and Chodorow, se basan en el camino que tiene un concepto a otro. Estas proveen información semántica de que si los conceptos están en la estructura "Es-Un" o "Parte de".

Las métricas semánticas que se basan en el contenido de información son Lin, Resnik y Jiang & Conrath. Estas resultan útiles cuando las palabras son diferentes, pero tienen significado similar.

Para los rasgos léxicos, existen aquellos que comparan las cadenas en forma de n-gramas o subcadenas, como son *Dice similarity*, QGrams y Monge Elkan.

Aquellos que nos entregan información de las frases como solapamiento, son Jaccard y *Overlap Coefficient*.

Los rasgos que entregan información de similitud palabra a palabra son Levenshtein, Smith Waterman, Smith Waterman Gotoh, Smith Waterman Gotoh Windowed Affine, Nedleman Wunch y *Matching Coefficient*.

Rasgos léxicos que entregan información midiendo la longitud de las frases, palabra a palabra o distancia entre frases, son Chapman Mean Length, Chapman Length Deviation, *Euclidean Distance*, *SentenceLength* y *Block Distance*.

4.1. *Sense-phrase*

Se destaca que se eligen métricas semánticas debido a que algunas frases pueden ser iguales, pero ocurren en diferentes contextos, algo que las métricas léxicas no toman en consideración el contexto en que ocurren las frases. Se puede dar el problema de que entre dos frases iguales, solamente al cambiar de posición la palabra de una frase, puede cambiar el contexto y por ende los sentidos de las palabras. Para esto se agregaron métricas léxicas, las cuales aunque cambien las palabras de posiciones, estás serán las mismas para estas métricas.

En la sección Estado del arte, deben desambiguar cada frase para obtener sentidos y

de acuerdo al modelo UMCC (Chavez et al., 2014) la desambiguación se realizó tomando el primer sentido de la palabra como el más probable, pero no siempre el primer sentido tiene que ser el correcto, por ende también se realizó el estudio del efecto del desambiguador de Lesk como enfoque.

Uno de los principales aportes de esta investigación es la aplicación de sense-phrase, para esto se consideró el problema que se da en los rasgos léxicos, como miden la distancia entre las palabras sin tomar en cuenta si las palabras están relacionadas semánticamente, se realizó el proceso de pasar sentidos a los rasgos léxicos para evitar el problema de que su distancia léxica sea alta cuando están relacionadas semánticamente, o pueda ser baja cuando no están relacionadas semánticamente. En la tabla 4 se muestra un ejemplo de distancia léxica entre palabras, cabe destacar que es solo un ejemplo, no todas las distancias léxica funcionan de la misma manera.

Tabla 4: Tabla ejemplo léxicos.

hola	ola	Distancia = 1
campo	campesino	Distancia = 4

La palabra ola y hola son dos palabras que no tienen relación semántica entre ellas, pero léxicamente las separa una sola letra, las palabras campo y campesino son dos palabras que semánticamente tienen relación, pero que léxicamente las separan cuatro letras. Para que esto no ocurra, se probó entregar sentidos a los rasgos léxicos, campo y campesino tendrían sentidos cercanos o iguales, en cambio ola y hola tendrían sentidos diferentes. Esto aporta en el estudio.

Ejemplo de como queda una frase con sentidos, con *stopwords* removidos, se muestra de la siguiente forma:

- **Frase normal:** two woman mix something food processor
- **Frase con sentidos:** womanhood.n.02 shuffle.v.03 food.n.02 processor.n.01

  Destacar que no todas las palabras tienen sentidos en el contexto en que se da la frase.

Para la combinación de todas las métricas que entraron en el estudio, se ocuparon 4 modelos:

1. Dagging
2. Linear Regression
3. Random Forest
4. SMOreg

Estos modelos fueron elegidos por entregar buenos resultados y/o ser los mas ocupados en el estado del arte, tienen la característica de que su entrenamiento se basa en función de la cantidad de datos que tengan para entrenar, es decir, entre más datos de entrenamiento, mayores resultados se podrían obtener. SMOreg es un algoritmo que se implementa como máquina de soporte de vectores, diferente a los que se mencionan en el estado del arte, pero que cumple la función de soporte de vectores.

Las herramientas o tecnologías que se utilizaron para esta investigación fueron NLTK, herramienta diseñada para el procesamiento del lenguaje natural, WordNet para la extracción de sentidos y las métricas semánticas que tiene, Python para realizar el código del sistema y acoplar todas las herramientas, Java para extraer las distancias léxicas de la librería (Chapman and Parkinson, 2006), Lucene para realizar el análisis semántico latente con la base de datos indexada proporcionada por (Chavez et al., 2014), librería numpy para ejecutar el algoritmo húngaro, WEKA para realizar los entrenamientos con los algoritmos de aprendizaje automático.

El proceso para obtener los resultados de cada métrica, se refleja en la figura 1, lo cuál consiste en una serie de pasos para obtener un vector de resultados para cada frase.

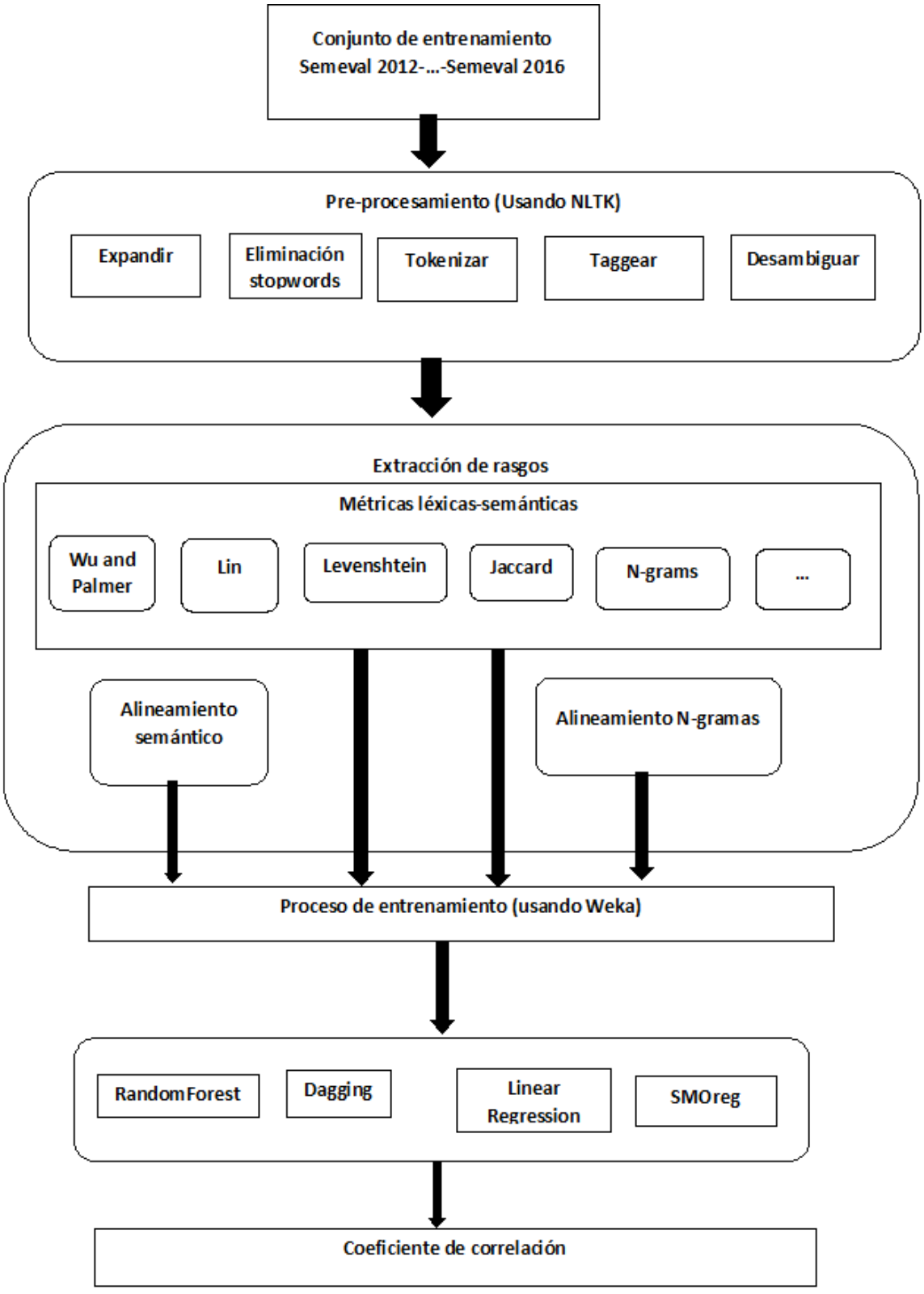

Figura 1: Esquema modelo estudio.

4.2. Pre-procesamiento

De acuerdo al modelo ya definido en la sección anterior, el sistema se constituyó de una serie de pasos para la realización del estudio de la similitud textual y obtener un corpus de entrenamiento.

El pre-procesamiento de las frases consistió en preparar las frases para la medición con cada métrica léxica y semántica.

- El primer paso consistió en convertir todas las letras mayúsculas en minúsculas.
- Todas las abreviaciones fueron expandidas, debido a las contracciones que existen en el idioma inglés.
- Las frases de los corpus que entrega SemEval traen caracteres que no son parte de una frase, por ejemplo *"Imagine a place that's % white and % black"*, los caracteres % no aporta información en la frase, por ende se borró todo carácter que no aportaba información.
- Todas las puntuaciones fueron removidas a excepción de los números decimales.
- Las frases fueron tokenizadas y luego etiquetadas.
- Se identificaron los *stopwords* y se removieron de las frases.
- Se creó una nueva frase para la medición con métricas léxicas sin los *stopwords* y etiquetas.

4.2.1. Extracción de sentidos

Para la medición con métricas semánticas, se desambiguó la frase para extraer los sentidos de cada palabra, dependiendo del contexto de la frase. Todo estos sentidos se extrajeron de *WordNet*.

La desambiguación se realizó de dos formas. La primera, tomó el primer sentido que

entrega *WordNet* como el más probable, debido a que entrega la lista de sentidos, siendo el primero como el más probable. La segunda, se ocupó el desambiguador de Lesk, donde se tomó la frase para entregar los sentidos dependiendo del contexto que entregaba Lesk.

4.2.2. Similitud semántica

A través de los sentidos extraídos, se procedió a comparar los sentidos, de las palabras a comparar, con cada métrica semántica ya definida. Cada métrica entrega una matriz como función de costo formada por el grado de similitud arrojado por cada palabra de la frase que fue comparada.

4.2.3. Similitud léxica

Con los nuevos pares de frases creados en el pre-procesamiento (sin *stopwords* y etiquetas), se procedió a comparar las frases con las métricas léxicas ya definidas. Cada una entrega el grado de similitud a nivel de frase.

4.2.4. N-Gramas

En el cálculo de los n-gramas, los pares de frases creados en el pre-procesamiento (sin *stopwords* y etiquetas), fueron comparados palabra a palabra, para cada n-grama ya definido (bi-gramas, tri-gramas, tetra-gramas). Cada n-grama entrega una matriz como función de costo.

4.2.5. Alineamiento de sentencias

Para las métricas semánticas y n-gramas, los resultados entregados por cada una debieron alinearse a nivel de frase. Existen varios métodos de alineamiento que pueden ser usados, todos con un distinto punto de vista. Cabe recordar que las métricas semánticas y n-gramas

se miden a nivel de palabra, entregando una matriz de costo. Se utilizó el algoritmo húngaro como forma de alinear dos frases y reducir así su costo, debido a que las métricas semánticas funcionan a nivel de palabra.

4.3. Enfoques propuestos

Para los experimentos, hay que destacar que se realizaron cuatro enfoques distintos, cada uno probado en los cuatro modelos ya mencionados. Además, se realizaron pruebas con las métricas con las que trabajó el modelo UMCC (Chavez et al., 2014) para comparar resultados con las nuevas métricas agregadas y estudiar si agregar las métricas tienen un efecto positivo o negativo.

Los enfoques propuestos son:

1. **Enfoque A:** La desambiguación se realiza mediante el primer sentido entregado por WordNet como el más frecuente y no se emplean *sense-phrase*.
2. **Enfoque B:** La desambiguación se realiza mediante el algoritmo de Lesk y no se emplean *sense-phrase*.
3. **Enfoque C:** La desambiguación se realiza mediante el primer sentido entregado por WordNet como el más frecuente y se emplea *sense-phrase*.
4. **Enfoque D:** La desambiguación se realiza mediante el algoritmo de Lesk y se emplea *sense-phrase*.

Capítulo 5

5. Experimentos

En esta sección, se muestra los estudios que se realizaron basados en el modelo UMCC 2014 (Chavez et al., 2014), bajo los 4 enfoques propuestos en esta investigación. Los parámetros de todos los modelos quedaron estándar, una validación cruzada de 10 particiones y los diferentes enfoques se probaron con un total de 11.105 instancias. Para el modelo *Dagging*, el modelo utilizado a combinar fue *Random Forest*, en todos los enfoques. Además, en la tabla 5 se muestran las métricas empleadas de los primeros cinco experimentos realizados y se clasificaron alfabéticamente los experimentos de la siguiente manera:

A) Experimento basado en modelo UMCC.

B) Experimento basado con modelo UMCC más 7 métricas nuevas.

C) Experimento solo con rasgos léxicos.

D) Experimento solo con rasgos semánticos.

E) Experimento con rasgos léxicos-semánticos, sin n-gramas.

F) Experimento modelo UMCC más Block Distance.

G) Experimento modelo UMCC más Chapman Length Deviation.

H) Experimento modelo UMCC más Nedleman Wunch.

I) Experimento modelo UMCC más Chapman Mean Length.

J) Experimento modelo UMCC más Matching Coefficient.

K) Experimento modelo UMCC más Monge Elkan.

L) Experimento modelo UMCC más Jaro.

M) Experimento modelo base más 3 métricas.

Tabla 5: Tabla de experimentos y métricas empleadas.

Experimentos / Métricas	A	B	C	D	E
Wu and Palmer	X	X		X	X
PathLength	X	X		X	X
Lin	X	X		X	X
Jiang & Conrath	X	X		X	X
Leacock & Chodorow	X	X		X	X
Similitud de palabra	X	X		X	X
Máxima similitud de palabras	X	X		X	X
Estadística y relación de peso	X	X		X	X
Dice similarity	X	X	X		X
Euclidean Distance	X	X	X		X
Jaccard	X	X	X		X
Jaro	X	X	X		X
Jaro-Winkler	X	X	X		X
Levenshtein	X	X	X		X
Overlap Coefficient	X	X	X		X
QGrams	X	X	X		X
Smith Waterman	X	X	X		X
Smith Waterman Gotoh	X	X	X		X
Smith Waterman Gotoh Windowed Affine	X	X	X		X
Block Distance	X	X	X		X
Chapman Mean Length	X	X	X		X
Chapman Length Deviation	X	X	X		X
Monge Elkan	X	X	X		X
Matching Coefficient	X	X	X		X
Nedleman Wunch	X	X	X		X
SentenceLength	X	X	X		X
Bi-gramas	X	X			
Tri-gramas	X	X			
Tetra-gramas	X	X			

5.1. Corpus

Existen varios corpus disponibles sobre los que puede evaluarse la propuesta (SEMCOR, SENSEVAL, Microsoft, PASCAL). Sin embargo, se utilizaron los datos de las competiciones SemEval 2012 al 2016 debido a que han sido empleados por más de 30 autores quienes han participado en la tarea de similitud semántica textual.

En la tabla 6 se muestran los corpus de cada año ocupados en esta investigación, con la cantidad de pares de frases que contiene cada uno. Cabe destacar que todos estos fueron sometidos a la fase de pre-procesamiento previamente señalado en la seccion 4.2. En total, se generó un conjunto de entrenamiento de 11.105 instancias, es decir, un total de 11.105 pares de frases que formaron el conjunto de entrenamiento de cada enfoque.

Tabla 6: Corpus utilizados.

Año	**Nombre (archivo .txt)**	**Pares**
2012	MSRpar	1500
2012	MSRvid	1500
2012	OnWN	750
2012	SMTnews	750
2012	SMTeuroparl	750
2013	HDL	750
2013	FNWN	189
2013	OnWN	561
2013	SMT	750
2014	HDL	750
2014	OnWN	750
2014	Deft-forum	450
2014	Deft-news	300
2014	Images	750
2014	Tweets-news	750
2015	HDL	750
2015	Images	750
2015	Ans. student	750
2015	Ans. forum	375
2015	Belief	375
2016	HDL	249
2016	Plagiarism	230
2016	Postediting	244
2016	Ans.-Ans,	254
2016	Quest.-Quest.	209

5.2. Experimentos basados en modelo UMCC

En esta sección, se muestra los estudios que se realizaron basados en las métricas que ocupó el modelo UMCC 2014 (Chavez et al., 2014), bajo los 4 enfoques propuestos en esta investigación.

Tabla 7: Tabla de coef. correlación UMCC.

Enfoque / Modelo	A	B	C	D
Random Forest	**0.7953**	0.7689	0.6716	0.6469
Dagging	0.7404	0.7431	0.6755	0.6095
Linear Regression	0.7081	0.7185	0.6298	0.5632
SMOreg	0.7064	0.7195	0.5774	0.5596

En la tabla 7 se muestran los resultados de cada enfoque, en cada modelo probado. Los resultados del modelo base UMCC (Chavez et al., 2014) son buenos para probar un aumento de métricas en el modelo y corroborar la hipótesis, en cada enfoque con cada modelo.

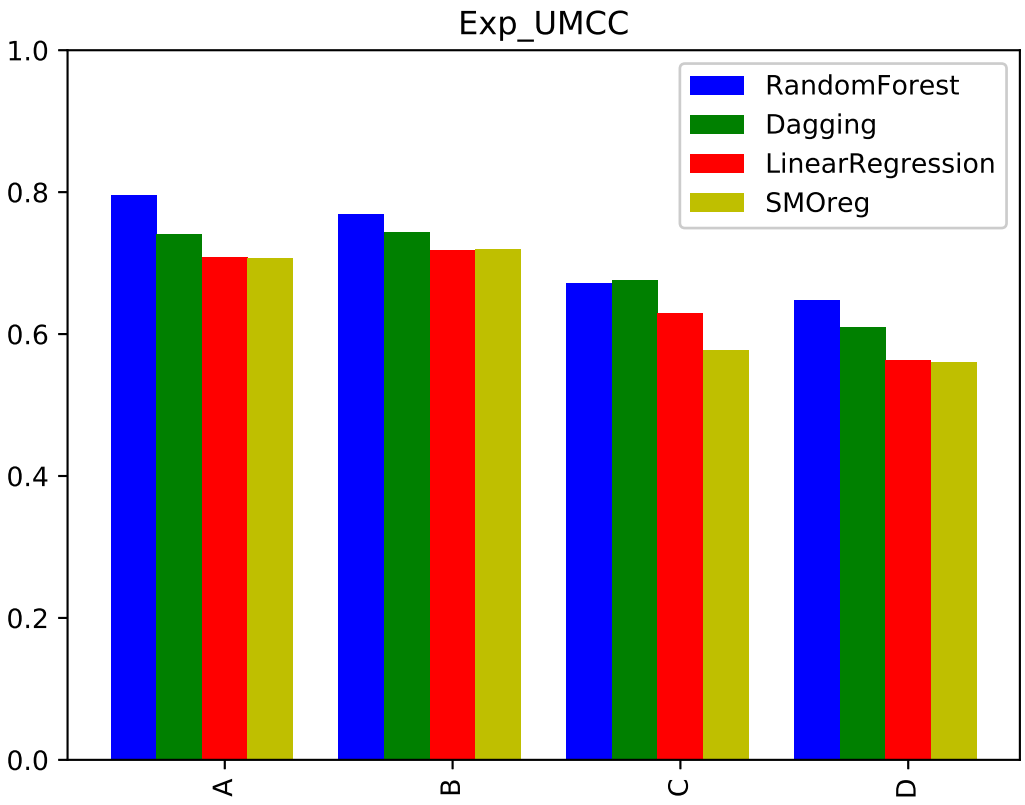

Figura 2: Resultados correlación modelo UMCC.

En el caso de los modelos, el que mejor resultado obtuvo fue *Random Forest* en 3 de los 4 enfoques propuestos, solo en el enfoque C, *Random Forest* fue superado por *Dagging*, por una diferencia de 0.0039. *Linear Regression* no obtuvo mejor resultado al igual que SMOreg en comparación a *Random Forest y Dagging*. En general, la prueba con las métricas del modelo base, el enfoque A muestra buenos resultados, el enfoque B también muestra buenos resultados, no más que el enfoque A, para corroborar la hipótesis previamente explicada en la sección 4 y los enfoques C y D muestran una diferencia alta en comparación de los enfoques A y B, resultados que no son mayores a los de A y B. La figura 2 muestra los gráfico de los resultados, donde se ve claramente que *Random Forest* y el enfoque A tienen, en mayor parte, los resultados más altos en esta prueba. Para corroborar la hipótesis antes mencionada,

se realizó un nuevo experimento con la propuesta que se planteó en esta investigación.

5.3. Experimento basado con modelo UMCC más 7 métricas nuevas.

En esta sección, se muestra los estudios que se realizaron bajo la propuesta que se empleó en esta investigación. Se creó un nuevo modelo de entrenamiento para cada enfoque y poder corroborar la hipótesis previamente descrita en la sección 4.

Tabla 8: Tabla de coef. correlación.

Enfoque / Modelo	A		B		C		D	
Random Forest	**0.7953**	0.7950	0.7689	0.7705	0.6716	0.6716	0.6469	0.6424
Dagging	0.7404	0.7479	0.7431	0.7429	0.6755	0.6261	0.6095	0.6055
Linear Regression	0.7081	0.7113	0.7185	0.7215	0.6298	0.5843	0.5632	0.5656
SMOreg	0.7064	0.7100	0.7195	0.7161	0.5774	0.5802	0.5596	0.5614

En la tabla 8 se muestran los resultados de cada enfoque, en cada modelo probado. Los resultados de la nueva propuesta para esta investigación, destacando que las 7 nuevas métricas son rasgos léxicos, no son mejores, debido a que en comparación a los resultados mostrados en la tabla 7, no todos los modelos mostraron buenos resultados en términos de correlación. La primera columna debajo de cada enfoque de la tabla 8, muestra los resultados de los experimentos de la sección 5.2, la segunda columna debajo de cada enfoque, muestra los resultados de la propuesta de esta investigación.

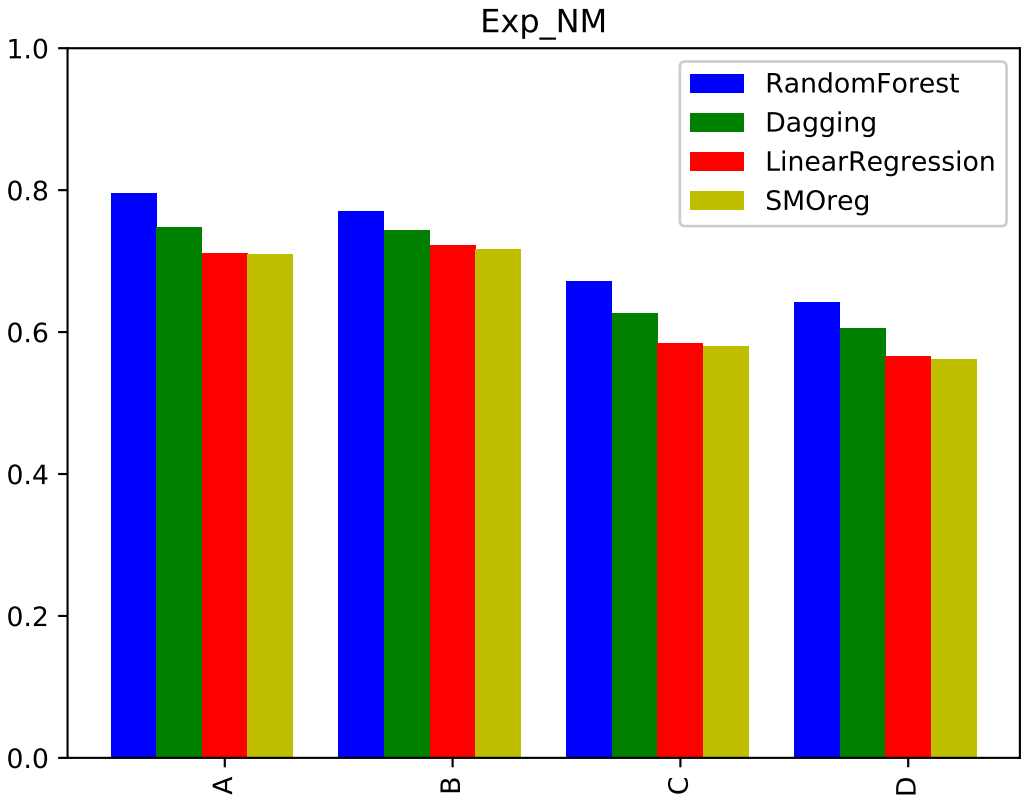

Figura 3: Resutado correlación todas las métricas.

En el enfoque A, *Random Forest* bajó un 0.0003, un valor que puede ser considerado nulo en comparación a los resultados de la tabla 7, en cambio *Dagging* sufrió un aumento de 0,0075. *Linear Regression* con el enfoque A también sufre un aumento, al igual que SMOreg.

En el enfoque B, *Random Forest* y *Linear Regression* sufren un aumento, pero *Dagging* y SMOreg bajan, todo en términos de correlación.

En el enfoque C, en *Random Forest* el resultado se mantuvo respecto al experimento en base al modelo UMCC (Chavez et al., 2014), pero *Dagging, Linear Regression* y SMOreg, bajaron su valor.

En el enfoque D, *Random Forest, Dagging* y SMOreg bajan su valor respecto al primer

experimento, pero *Linear Regression* aumenta.

5.4. Discusión

Si se comparan los resultados gráficamente reflejados en la figura 4, entre los resultados de la sección 5.2 y la sección 5.3, del modelo *Random Forest* de cada enfoque, se aprecia que la diferencia es nula. Solo en el enfoque B se aprecia que la correlación es mejor para la propuesta de esta investigación en comparación al modelo UMCC, donde se deduce que en el enfoque B con el modelo *Random Forest*, se corrobora la hipótesis previamente explicada.

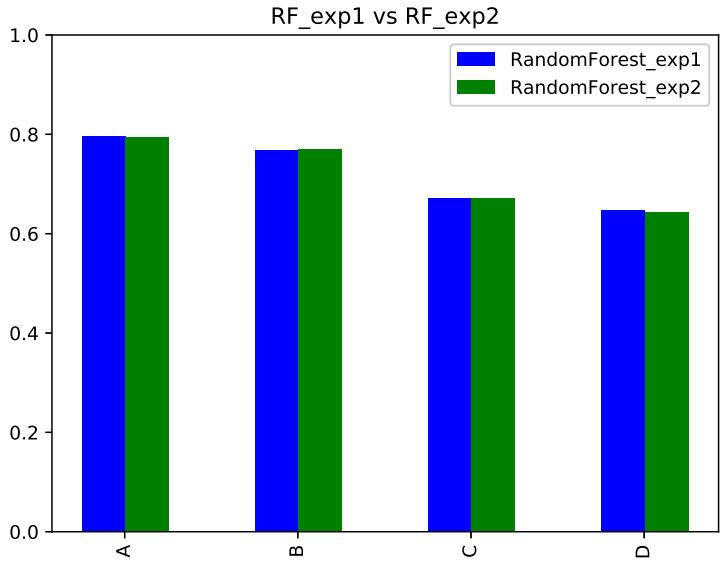

Figura 4: Resutados para *Random Forest* entre experimento 1 y 2.

Si se compara los gráficos de la figura 5, entre el los resultados de la sección 5.2 y la sección 5.3, del modelo *Dagging* de cada enfoque, la diferencia es nula a excepción del enfoque C, donde existe una diferencia. Se puede concluir que para *Dagging*, en el enfoque C, no impacta el aumento de métricas y corroborar la hipótesis que sigue esta investigación. En comparación a *Random Forest, Dagging* en el enfoque A tuvo un mínimo aumento en términos de correlación.

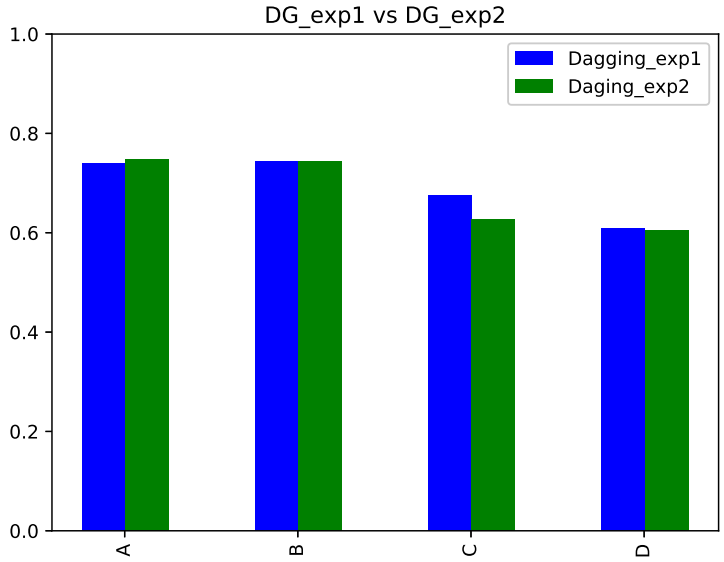

Figura 5: Resutados para Dagging entre experimento 1 y 2.

Si se compara los gráficos de la figura 6, entre los resultados de la sección 5.2 y la sección 5.3, del modelo *Linear Regression* de cada enfoque, se puede apreciar que la diferencia es nula a excepción del enfoque C, donde existe diferencia al igual que *Dagging*. Se puede concluir que para *Linear Regression*, en el enfoque C, el aumento de métricas no impacta de manera positiva para obtener un mejor resultado en términos de correlación.

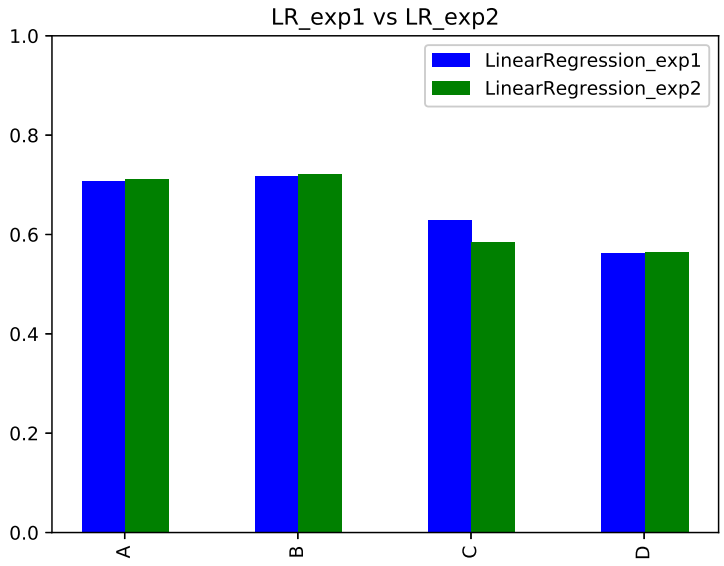

Figura 6: Resutados para Linear Regression entre experimento 1 y 2.

Si se compara los gráficos de la figura 7, entre los resultados de la sección 5.2 y la sección 5.3, del modelo SMOreg de cada enfoque, la diferencia es nula. En el enfoque A, se aprecia un mínimo aumento en términos de correlación, pero en el enfoque B se aprecia una mínima baja. En el enfoque A se corrobora la hipótesis previamente explicada, pero con un mínimo aumento.

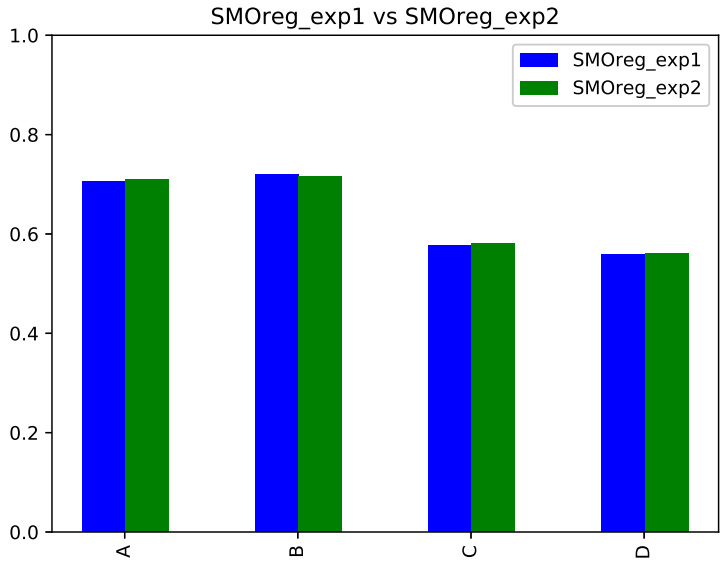

Figura 7: Resutados para SMOreg entre experimento 1 y 2.

En términos generales, el modelo *Random Forest* fue el modelo que mejores resultados obtuvo, pero basándose en el estado del arte, mayormente se ha ocupado un algoritmo de máquina de soporte de vectores, en este caso, SMOreg es el modelo de soporte de vectores, que no da buenos resultados en comparación de *Dagging y Random Forest*. Además se reflejó que aumentando las métricas, en forma general, no hubo mejores resultados para corroborar con exactitud la hipótesis que siguió esta investigación.

5.5. Experimentos solo con rasgos léxicos

En esta sección se muestran los resultados de todos los enfoques, donde se realizó el estudio de la correlación solo con rasgos léxicos previamente definidos.

En la tabla 9 se puede apreciar los resultados del experimento realizado solo con rasgos (métricas) léxicos. Se destaca que el enfoque A en Random Forest, varía casi en 0,01 en términos de correlación con respecto a los experimentos de la sección 5.2.

Tabla 9: Tabla de coef. correlación rasgos léxicos.

Enfoque / Modelo	A		B		C		D	
Random Forest	**0.7953**	0.7862	0.7689	0.7589	0.6716	0.6313	0.6469	0.6013
Dagging	0.7404	0.7375	0.7431	0.7416	0.6755	0.5984	0.6095	0.5790
Linear Regression	0.7081	0.7096	0.7185	0.7200	0.6298	0.5705	0.5632	0.5489
SMOreg	0.7064	0.7087	0.7195	0.7188	0.5774	0.5666	0.5596	0.5455

Gráficamente se ve que el mejor resultado lo da el modelo *Random Forest*.

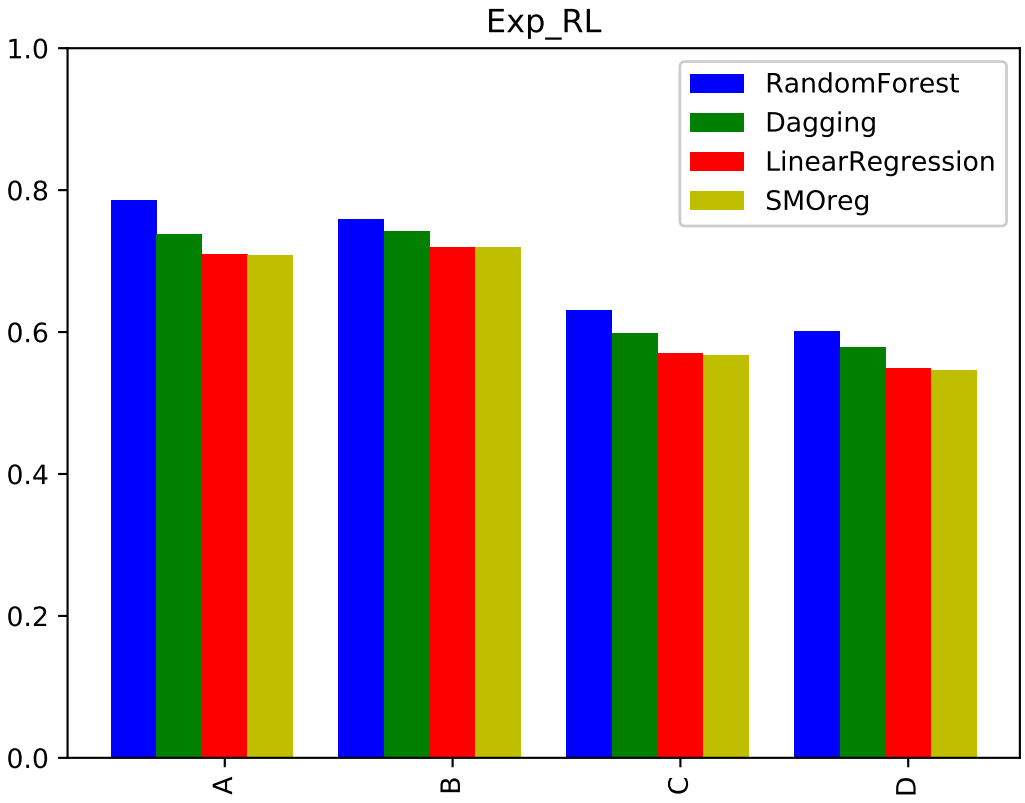

Figura 8: Resultados correlación solo rasgos léxicos.

Al ver los resultados de la tabla 9, el conjunto de entrenamiento generado solo por métricas léxicas muestra un resultado alto en términos de correlación, no más que los resultados de las secciones 5.2 y 5.3. El problema que existe en este conjunto de entrenamiento es que no entrega información semántica, esto quiere decir, no hay información sobre el contexto de las palabras, el significado de cada una, solo entrega información de la distancia entre cada palabra. Por ende, se realizó una prueba solo con rasgos semánticos.

5.6. Experimentos solo con rasgos semánticos

En esta sección se muestran los resultados de todos los enfoques, donde se realizó el estudio de la correlación solo con rasgos semánticos, previamente definidos.

La tabla 10 muestra los resultados obtenidos solo con rasgos semánticos en términos de correlación.

Tabla 10: Tabla de coef. correlación rasgos semánticos.

Enfoque / Modelo	A		B		C		D	
Random Forest	**0.7953**	0.5382	0.7689	0.4112	0.6716	0.4224	0.6469	0.3903
Dagging	0.7404	0.3829	0.7431	0.3528	0.6755	0.3448	0.6095	0.3409
Linear Regression	0.7081	0.2289	0.7185	0.2345	0.6298	0.2276	0.5632	0.2314
SMOreg	0.7064	0.2311	0.7195	0.2368	0.5774	0.2298	0.5596	0.2331

Gráficamente sigue que el mejor resultado lo da el modelo *Random Forest*, pero no son buenos, debido a que los resultados son menores a los resultados de la propuesta de esta investigación.

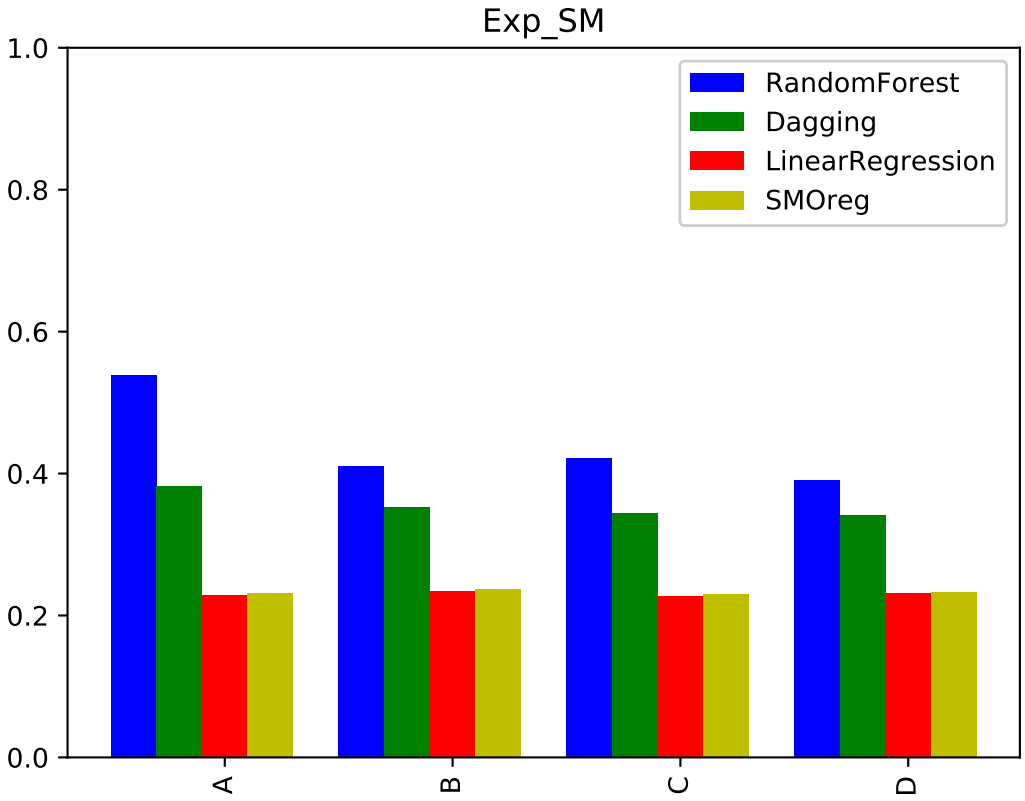

Figura 9: Resultados correlación solo rasgos semánticos.

Como ha ocurrido en todas las pruebas, el enfoque A es el que mejor correlación obtiene en el modelo *Random Forest*. Obtiene una gran diferencia con los demás modelos. Existe una diferencia mínima en los enfoques B y C con el modelo *Random Forest*, donde C supera al enfoque B. Se puede ver gráficamente que en los modelos *Linear Regression* y SMOreg obtienen una igualdad en todos los enfoques, pero al ver la tabla se puede ver que la diferencia es mínima.

Al comparar con los resultados de la tabla 9, el conjunto de entrenamiento generado

por rasgos léxicos obtiene mejor correlación que el conjunto de rasgos semánticos, pero en esta prueba se obtiene la ventaja de que los rasgos semánticos si entregan información sobre el contexto de las palabras en las frases, aunque en términos de correlación, es bajo en comparación a los experimentos de la sección 5.5, se destaca que hay diferencia en el número de métricas empleadas para cada prueba.

5.7. Experimentos con rasgos léxicos-semánticos, sin n-gramas

En esta sección, se muestran los resultados de las pruebas de todo los enfoques, con rasgos léxicos-semánticos, pero sin n-gramas, lo que da un total de 28 rasgos.

La tabla 11 muestra los resultados obtenidos de este experimento, en términos de correlación.

Tabla 11: Tabla de coef. correlación rasgos léxicos-semánticos.

Enfoque / **Modelo**	**A**		**B**		**C**		**D**	
Random Forest	0.7953	**0.7955**	0.7689	0.7701	0.6716	0.6669	0.6469	0.6421
Dagging	0.7404	0.7410	0.7431	0.7435	0.6755	0.6219	0.6095	0.6028
Linear Regression	0.7081	0.7116	0.7185	0.7215	0.6298	0.5813	0.5632	0.5620
SMOreg	0.7064	0.7102	0.7195	0.7196	0.5774	0.5767	0.5596	0.5614

Gráficamente en la figura 10, se puede ver que *Random Forest* con el enfoque A, sigue siendo mejor en comparación a los demás enfoques.

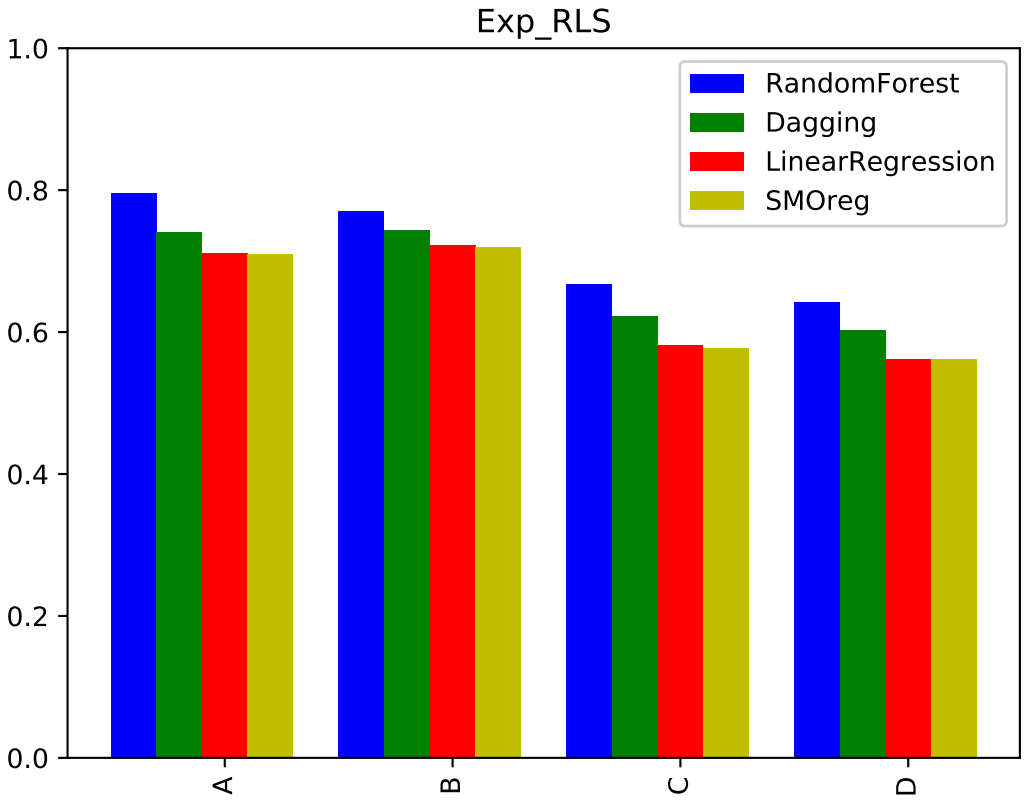

Figura 10: Resultados correlación rasgos léxicos-semánticos, sin n-gramas.

Comparando los resultados con la tabla 7 de la sección 5.2, para el enfoque A en todos los modelos, la correlación sufre un mínimo aumento.

Para el enfoque B, solo en *Linear Regression* la correlación disminuye, los otros modelos tienen un aumento mínimo.

Para el enfoque C, en todos los modelos, la correlación disminuye.

Para el enfoque D, en todos los modelos, la correlación disminuye.

En términos generales, la correlación aumentó o disminuyó de forma mínima, que podría denotar una diferencia nula con los resultados de la tabla 7.

5.8. Experimentos con cada métrica agregada

Para esta sección, las pruebas se realizaron tomando el modelo UMCC y agregando solamente una métrica de las 7 nuevas que se propuso en esta investigación a todos los enfoques, para ver el impacto que puede tener cada una en el resultado final de correlación en cada modelo. En la tabla 12 muestra en resumen las métricas empleadas en cada experimento. Cabe destacar que la comparación de los resultados en esta sección, se compararon con los resultados del modelo base UMCC (Chavez et al., 2014), donde la primera columna de cada enfoque, en cada tabla de los experimentos, representa los resultados de los experimentos de la sección 5.2 y la segunda columna muestra los resultados de estos experimentos.

Tabla 12: Tabla de experimentos y métricas empleadas.

Experimentos / Métricas	F	G	H	I	J	K	L	M
Wu and Palmer	X	X	X	X	X	X	X	X
PathLength	X	X	X	X	X	X	X	X
Lin	X	X	X	X	X	X	X	X
Jiang & Conrath	X	X	X	X	X	X	X	X
Leacock & Chodorow	X	X	X	X	X	X	X	X
Similitud de palabra	X	X	X	X	X	X	X	X
Máxima similitud de palabras	X	X	X	X	X	X	X	X
Estadística y relación de peso	X	X	X	X	X	X	X	X
Dice similarity	X	X	X	X	X	X	X	X
Euclidean Distance	X	X	X	X	X	X	X	X
Jaccard	X	X	X	X	X	X	X	X
Jaro							X	
Jaro-Winkler	X	X	X		X	X	X	X
Levenshtein	X	X	X	X	X	X	X	X
Overlap Coefficient	X	X	X	X	X	X	X	X
QGrams	X	X	X	X	X	X	X	X
Smith Waterman	X	X	X	X	X	X	X	X
Smith Waterman Gotoh	X	X	X	X	X	X	X	X
Smith Waterman Gotoh Windowed Affine	X	X	X	X	X	X	X	X
Block Distance	X							
Chapman Mean Length				X				X
Chapman Length Deviation		X						
Monge Elkan						X		X
Matching Coefficient					X			
Nedleman Wunch			X					X
SentenceLength	X	X	X	X	X	X	X	X
Bi-gramas	X	X	X	X	X	X	X	X
Tri-gramas	X	X	X	X	X	X	X	X
Tetra-gramas	X	X	X	X	X	X	X	X

5.8.1. Enfoques modelo UMCC más Block Distance

En la tabla 13 se puede ver los resultados en conjunto de todos los enfoque y modelos. Al comparar estos resultados, se muestra que con la métrica *Block Distance*, los resultados variaron en una mínima diferencia.

Tabla 13: Tabla de coef. correlación modelo base más *Block Distance*.

Enfoque / Modelo	A		B		C		D	
Random Forest	**0.7953**	0.7949	0.7689	0.769	0.6716	0.6747	0.6469	0.6456
Dagging	0.7404	0.7400	0.7431	0.7425	0.6755	0.6302	0.6095	0.6106
Linear Regression	0.7081	0.7082	0.7185	0.7187	0.6298	0.5815	0.5632	0.5636
SMOreg	0.7064	0.7066	0.7195	0.7164	0.5774	0.5786	0.5596	0.5599

Gráficamente en la figura 11, se puede ver que el enfoque A, con el modelo *Random Forest*, tiende a ser el mejor coeficiente de correlación.

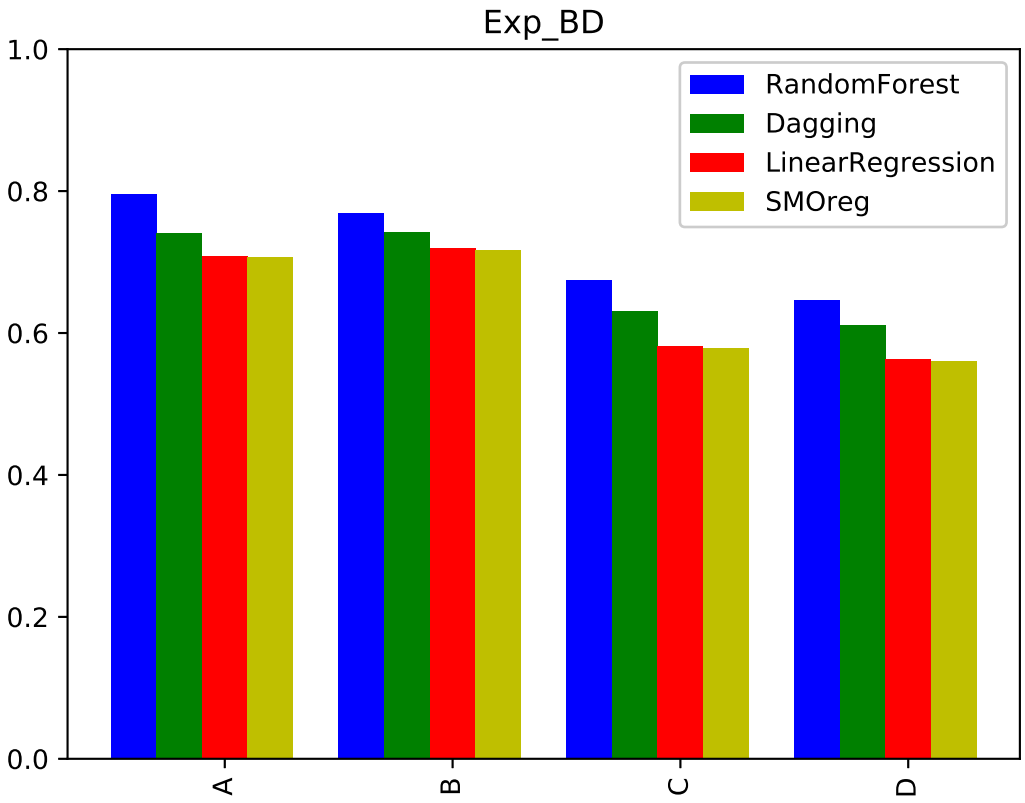

Figura 11: Resultados correlación modelo base más *Block Distance*.

Los gráficos revelan que los enfoques A y B son ampliamente superiores que los enfoques C y D. En todos los enfoques el modelo *Random Forest* es el que obtiene mejor correlación, seguido por *Dagging* y hay una igualdad general en los modelos *Linear Regression* y SMOreg.

Al comparar resultados, para el enfoque A, el coeficiente de correlación en *Random Forest y Dagging* disminuyó, pero en *Linear Regression* y SMOreg mejora.

Para el enfoque B, el coeficiente de correlación en *Random Forest y Dagging* disminuyó, pero en *Linear Regression* y SMOreg mejora.

Para el enfoque C, el coeficiente de correlación en *Random Forest* y SMOreg mejora, pero en *Dagging y Linear Regression* disminuyó.

Para el enfoque D, el coeficiente de correlación en *Random Forest* mejora, pero en *Dagging, Linear Regression* y SMOreg Disminuyó.

En términos generales, el impacto que produjo *Block Distance* en el enfoque A con el modelo *Random Forest* (por ser el enfoque con el modelo de mayor coef. de correlación), es bajar el resultado en términos de correlación. No existe una tendencia de buenos o malos resultados de manera general, todo depende del modelo y enfoque con el cual se trabaje.

5.8.2. Enfoques modelo UMCC más Chapman Length Deviation

En la tabla 14 se puede ver los resultados en conjunto de todos los enfoque y modelos. Al comparar estos resultados, se refleja que con la métrica *Chapman Length Deviation*, los resultados variaron en un mínimo porcentaje.

Tabla 14: Tabla de coef. correlación modelo base más *Chapman Length Deviation*.

Enfoque / **Modelo**	**A**		**B**		**C**		**D**	
Random Forest	**0.7953**	0.7945	0.7689	0.7680	0.6716	0.6762	0.6469	0.6431
Dagging	0.7404	0.7395	0.7431	0.7419	0.6755	0.6282	0.6095	0.6082
Linear Regression	0.7081	0.7087	0.7185	0.7188	0.6298	0.5806	0.5632	0.5629
SMOreg	0.7064	0.7072	0.7195	0.7165	0.5774	0.5774	0.5596	0.5596

En la figura 12 se muestran los resultados gráficamente de los experimentos de esta sección.

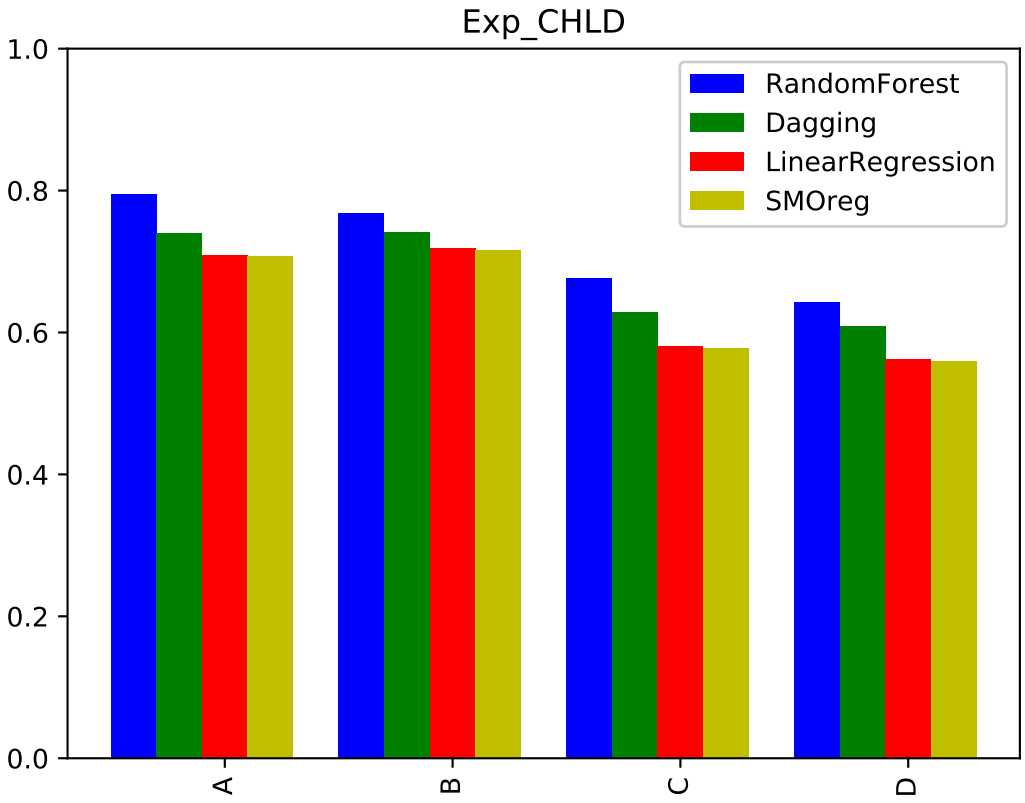

Figura 12: Resultados correlación modelo base más Chapman Length Deviation.

Los gráficos revelan que los enfoques A y B son ampliamente superiores que los enfoques C y D. En todos los enfoques el modelo *Random Forest* es el que obtiene mayor correlación, seguido por *Dagging* y hay una igualdad general en los modelos *Linear Regression* y SMOreg.

Al comparar resultados, para el enfoque A, en los modelos *Random Forest y Dagging* los resultados disminuyeron en términos de correlación. *Linear regression* y SMOreg, la correlación mejoró.

Para el enfoque B, en los modelos *Random Forest, Dagging* y SMOreg, el coef. de

correlación disminuyó. Para *Linear Regression*, la correlación subió.

Para el enfoque C, en los modelos *Random Forest, Dagging y Linear Regression*, los coef. de correlación disminuyeron. En SMOreg la correlación se mantuvo.

Para el enfoque D, en los modelos *Random Forest, Dagging*, los coef. de correlación disminuyeron, en cambio, en *Linear Regression* subió. En SMOreg el coef. de correlación se mantuvo.

En general, los resultados varían levemente, en forma insignificante, no existió un mejor o peor resultado en todos los enfoques, en cada modelo, todo en términos de correlación. Para el enfoque A y *Random Forest* (por ser el mejor resultado), el impacto que produce *Chapman Length Deviation* es bajar el resultado en términos de correlación.

5.8.3. Enfoques modelo UMCC más Nedleman Wunch

En la tabla 15 se puede ver los resultados en conjunto de todos los enfoque y modelos. Al comparar estos resultados, se refleja que con la métrica Nedleman Wunch, los resultados variaron en un mínimo porcentaje.

Tabla 15: Tabla de coef. correlación modelo base más Nedleman Wunch.

Enfoque / **Modelo**	**A**		**B**		**C**		**D**	
Random Forest	0.7953	**0.7972**	0.7689	0.7698	0.6716	0.6744	0.6469	0.6454
Dagging	0.7404	0.7417	0.7431	0.7436	0.6755	0.6294	0.6095	0.6099
Linear Regression	0.7081	0.7088	0.7185	0.7194	0.6298	0.5812	0.5632	0.5644
SMOreg	0.7064	0.7069	0.7195	0.7171	0.5774	0.5788	0.5596	0.5608

En la figura 13 se muestran los resultados gráficamente de los experimentos de esta sección.

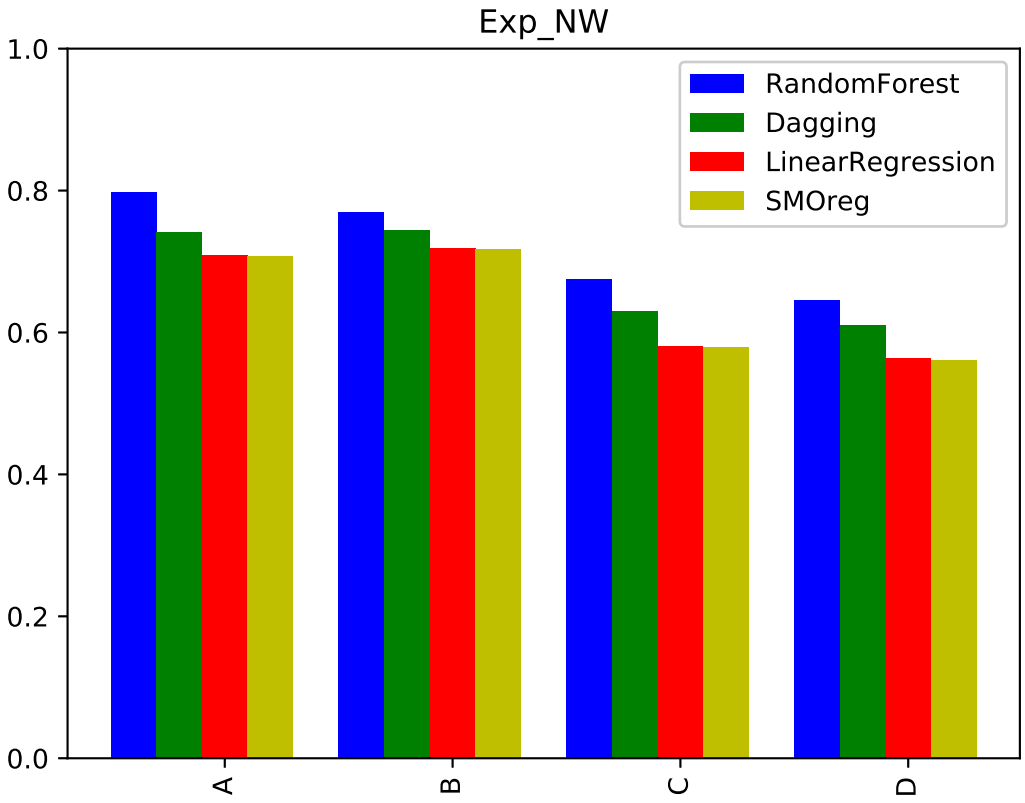

Figura 13: Resultados correlación modelo base más Nedleman Wunch.

Los gráficos revelan que los enfoques A y B son ampliamente superiores que los enfoques C y D. En todos los enfoques el modelo *Random Forest* es el que obtiene mayor correlación, seguido por *Dagging* y hay una igualdad general en los modelos *Linear Regression* y SMOreg. Al comparar los resultados, todos los modelos, en el enfoque A mejoraron.

Para el enfoque B, los resultados en todos los modelos mejoraron a excepción de SMOreg, donde el resultado disminuyó.

Para el enfoque C, *Dagging y Linear Regression*, los resultados disminuyeron, una diferencia aproximadamente de 0.05 en términos de correlación, pero *Random Forest* y SMOreg

mejoraron.

Para el enfoque D, *Random Forest y Dagging*, los resultados disminuyeron, pero *Linear Regression* y SMOreg, los resultados mejoraron.

En general, la métrica Nedleman Wunch, en mayor parte, mejoró el coef. de correlación de la mayoría de los enfoques, en la mayoría de los modelos. Para el enfoque de mayor correlación (A) y con el modelo que dio mayor resultados (*Random Forest*), en comparación a los resultados del modelo UMCC, el impacto que se produce es dar mejor resultado, que se define como el mayor resultado que se ha dado en comparación a los demás experimentos.

5.8.4. Enfoques modelo UMCC más ChapmanMeanLength

En la tabla 16 se puede ver los resultados en conjunto de todos los enfoque y modelos. Al comparar estos resultados, se refleja que con la métrica *Chapman Mean Length* los resultados variaron en un mínimo porcentaje.

Tabla 16: Tabla de coef. correlación modelo base más *Chapman Mean Length*.

Enfoque / **Modelo**	**A**		**B**		**C**		**D**	
Random Forest	0.7953	**0.7976**	0.7689	0.7707	0.6716	0.6760	0.6469	0.6477
Dagging	0.7404	0.7425	0.7431	0.7439	0.6755	0.6294	0.6095	0.6097
Linear Regression	0.7081	0.7087	0.7185	0.7190	0.6298	0.5821	0.5632	0.5643
SMOreg	0.7064	0.7066	0.7195	0.7168	0.5774	0.5792	0.5596	0.5611

En la figura 14 se muestran los resultados gráficamente de los experimentos de esta sección.

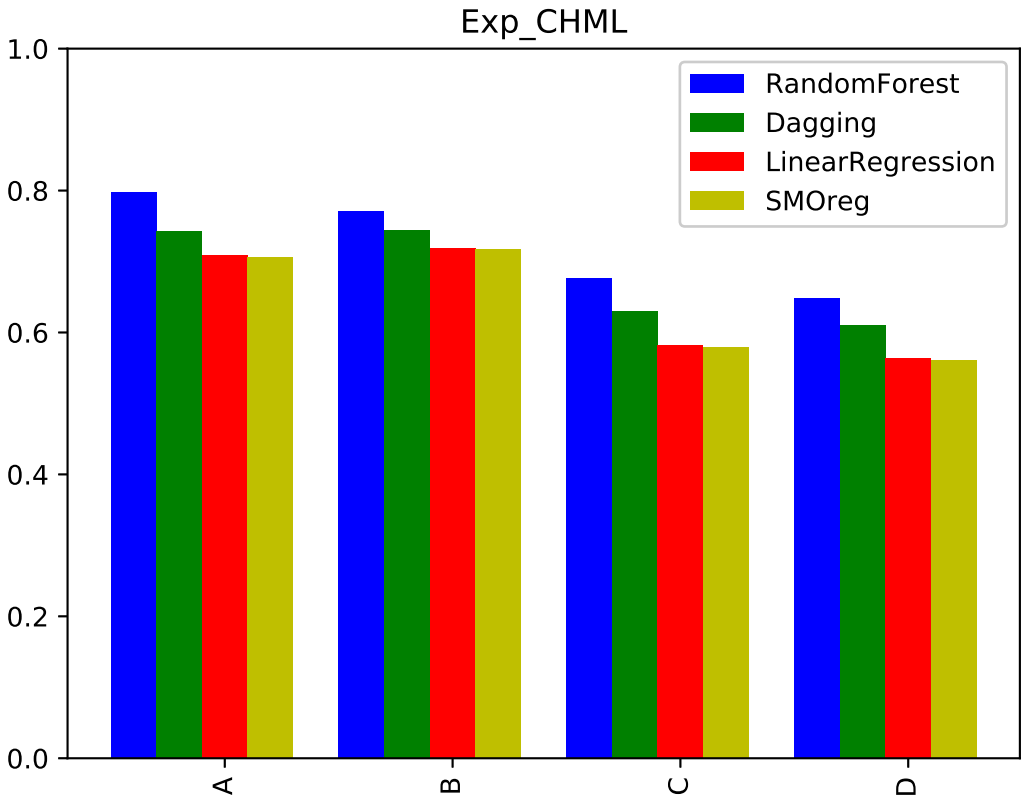

Figura 14: Resultados correlación modelo base más *Chapman Mean Length*.

Los gráficos revelan que los enfoques A y B son ampliamente superiores que los enfoques C y D. En todos los enfoques el modelo *Random Forest* es el que obtiene mayor correlación, seguido por *Dagging* y hay una igualdad general en los modelos *Linear Regression* y SMOreg.

Al comparar los resultados del enfoque A en todos los modelos, estos resultados mejoraron.

Los resultados del enfoque B en los modelos *Random Forest, Dagging y Linear Regression*, mejoraron, pero en SMOreg, el resultado disminuyó.

Para el enfoque C, en los modelos *Random Forest* y SMOreg, los resultados mejoraron, pero para *Dagging y Linear Regression*, los resultados disminuyeron.

Para el enfoque D, en los modelos *Random Forest y Dagging*, los resultados mejoraron, pero en los modelos *Linear Regression* y SMOreg, los resultados disminuyeron.

En resumen, para el enfoque A, que dio mayor correlación en el modelo *Random Forest*, mostró un mejor resultado lo que se traduce como la métrica que tuvo el mejor resultado, anteriormente fue Nedleman Wunch. En general, en todos los enfoques, la métrica *Chapman Mean Length* impacta de forma positiva al generar un buen resultado en términos de correlación.

5.8.5. Enfoques modelo UMCC más Matching Coefficient

En la tabla 17 se puede ver los resultados en conjunto de todos los enfoque y modelos. Al comparar estos resultados, se puede ver que con la métrica *Matching Coefficient*, los resultados variaron.

Tabla 17: Tabla de coef. correlación modelo base más *Matching Coefficient*.

Enfoque / **Modelo**	**A**		**B**		**C**		**D**	
Random Forest	**0.7953**	0.7944	0.7689	0.7687	0.6716	0.6741	0.6469	0.6459
Dagging	0.7404	0.7395	0.7431	0.7418	0.6755	0.6297	0.6095	0.6096
Linear Regression	0.7081	0.7081	0.7185	0.7185	0.6298	0.5798	0.5632	0.5632
SMOreg	0.7064	0.7062	0.7195	0.7161	0.5774	0.5774	0.5596	0.5594

En la figura 15 se muestran los resultados gráficamente de los experimentos de esta sección.

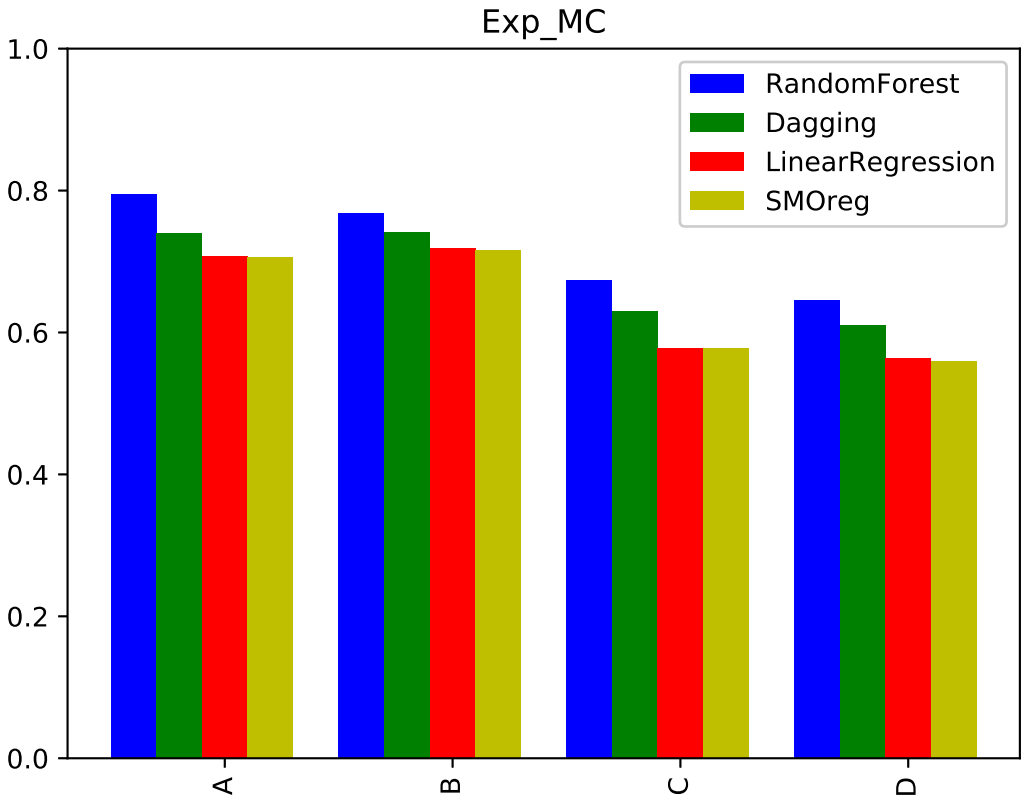

Figura 15: Resultados correlación modelo base más *Matching Coefficient*.

Los gráficos revelan que los enfoques A y B son ampliamente superiores que los enfoques C y D. En todos los enfoques el modelo *Random Forest* es el que obtiene mayor correlación, seguido por *Dagging* y hay una igualdad general en los modelos *Linear Regression* y SMOreg.

Al comparar los resultados, para el enfoque A, en los modelos *Random Forest, Dagging* y SMOreg, los resultados en términos de correlación disminuyeron. *Linear Regression* mantuvo su correlación.

Para el enfoque B, en los modelos *Random Forest, Dagging* y SMOreg, los resultados

en términos de correlación disminuyeron. *Linear regression* mantuvo su correlación.

Para el enfoque C, en los modelos *Dagging y Linear Regression*, los resultados disminuyeron en términos de correlación, *Random Forest* mejoró y SMOreg se mantuvo.

Para el enfoque D, *Random Forest* y SMOreg, los resultados disminuyeron, *Linear Regression* mantuvo el resultado y *Dagging* mejoró, todo en términos de correlación.

En resumen, se dio que en 3 de los 4 enfoques, en el modelo *Linear Regression* la correlación se mantuvo. Para el enfoque A, en *Random Forest* (por obtener el mayor resultado), el resultado disminuyó en comparación a los resultados del modelo UMCC, y en general, *Matching Coefficient* impacta de forma negativa a los resultados en términos de correlación.

5.8.6. Enfoques modelo UMCC más MongeElkan

En la tabla 18 se puede ver los resultados en conjunto de todos los enfoque y modelos. Al comparar estos resultados, se puede ver que con la métrica Monge Elkan, los resultados variaron.

Tabla 18: Tabla de coef. correlación modelo base más Monge Elkan.

Enfoque / **Modelo**	**A**		**B**		**C**		**D**	
Random Forest	0.7953	**0.7969**	0.7689	0.7704	0.6716	0.6763	0.6469	0.6448
Dagging	0.7404	0.7422	0.7431	0.7444	0.6755	0.6291	0.6095	0.6089
Linear Regression	0.7081	0.7086	0.7185	0.7193	0.6298	0.5811	0.5632	0.5633
SMOreg	0.7064	0.7064	0.7195	0.7168	0.5774	0.5780	0.5596	0.5599

En la figura 16 se muestran los resultados gráficamente de los experimentos de esta sección.

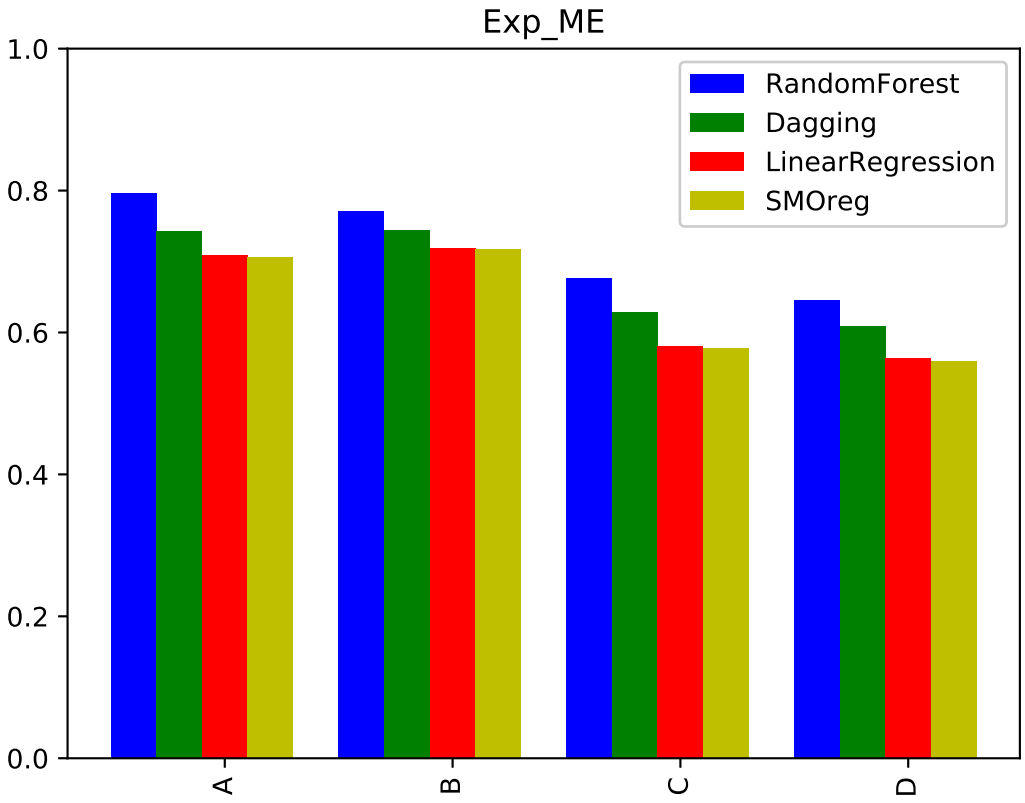

Figura 16: Resultados correlación modelo base más Monge Elkan.

Los gráficos revelan que los enfoques A y B son ampliamente superiores que los enfoques C y D. En todos los enfoques el modelo *Random Forest* es el que obtiene mayor correlación, seguido por *Dagging* y hay una igualdad general en los modelos *Linear Regression* y SMOreg.

Al comparar resultados, para el enfoque A, en los modelos *Random Forest, Dagging y Linear Regression*, los resultados en términos de correlación mejoraron, en SMOreg se mantuvo.

Para el enfoque B, en los modelos *Random Forest, Dagging y Linear Regression*, los

resultados en término de correlación mejoraron, en SMOreg, el resultado disminuyó.

Para el enfoque C, en los modelos *Random Forest* y SMOreg, los resultados en términos de correlación mejoraron, en cambio, en los modelos *Dagging y Linear Regression*, los resultados disminuyeron, en la que se destacó una baja considerable en el modelo *Dagging*.

Para el enfoque D, en los modelos *Random Forest y Dagging*, los resultados en términos de correlación disminuyeron, en cambio, en *Linear Regression* y SMOreg, los resultados mejoraron.

En resumen, los resultados en términos de correlación varían en un valor relativamente bajo. Para el enfoque A, en el modelo *Random Forest* (por ser el mayor resultado), el resultado mejoró y en general la métrica Monge Elkan impacta de forma positiva en todos los enfoques.

5.8.7. Enfoques modelo UMCC más Jaro

En la tabla 19 se puede ver los resultados en conjunto de todos los enfoque y modelos. Al comparar estos resultados, se puede ver que con la métrica Jaro, los resultados variaron.

Tabla 19: Tabla de coef. correlación modelo base más Jaro.

Enfoque / **Modelo**	**A**		**B**		**C**		**D**	
Random Forest	**0.7953**	0.7944	0.7689	0.7687	0.6716	0.6741	0.6469	0.6459
Dagging	0.7404	0.7395	0.7431	0.7418	0.6755	0.6297	0.6095	0.6096
Linear Regression	0.7081	0.7081	0.7185	0.7185	0.6298	0.5798	0.5632	0.5632
SMOreg	0.7064	0.7062	0.7195	0.7161	0.5774	0.5774	0.5596	0.5594

En la figura 17 se muestran los resultados gráficamente de los experimentos de esta sección.

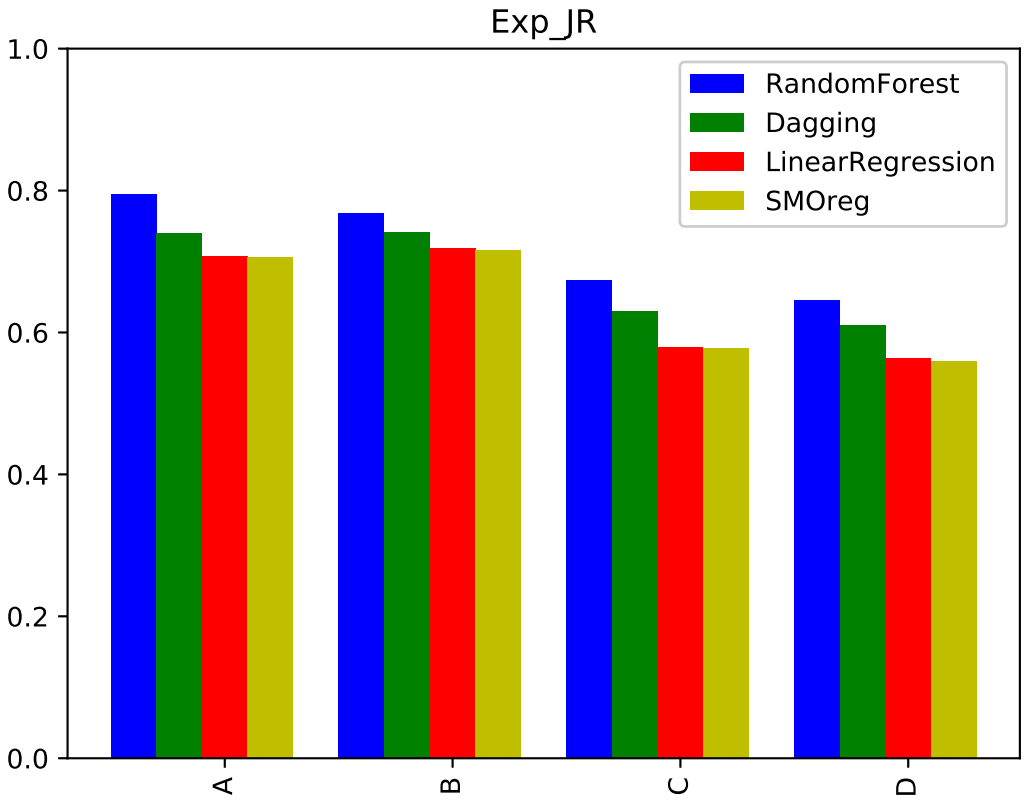

Figura 17: Resultados correlación modelo base más Jaro.

Los gráficos revelan que los enfoques A y B son ampliamente superiores que los enfoques C y D. En todos los enfoques el modelo *Random Forest* es el que obtiene mayor correlación, seguido por *Dagging* y hay una igualdad general en los modelos *Linear Regression* y SMOreg.

Al comparar los resultados, para el enfoque A, *Random Forest, Dagging* y SMOreg, los resultados en términos de correlación disminuyeron. *Linear Regression* se mantuvo.

Para el enfoque B, *Random Forest, Dagging* y SMOreg, los resultados en términos de correlación disminuyeron. *Linear Regression* se mantuvo.

Para el enfoque C, el resultado de *Random Forest* mejoró, los resultados de *Dagging y Linear Regression* disminuyeron y SMOreg mantuvo el resultado.

Para el enfoque D, los resultados de *Random Forest* y SMOreg disminuyeron, el resultado de *Dagging* mejoró, el resultado de *Regression Linear* se mantuvo, todo en términos de correlación.

En resumen general, el impacto que tuvo la métrica Jaro es negativo, en la mayoría de los enfoques y modelos, en términos de correlación. Para el enfoque con el modelo de mayor resultado (A y *Random Forest*), la correlación disminuyó, en comparación a los resultados del modelo UMCC.

5.8.8. Discusión

Para esta sección de experimentos, no hubo una métrica que impactara de forma positiva en todos los enfoques y modelos. Para una competencia como SemEval, destacamos que el enfoque con mayor resultado de correlación, enfoque A, en el modelo *Random Forest*, se destacan 3 métricas en la cuál impactaron de forma positiva en la mayoría de los enfoques, Nedleman Wunch, *Chapman Mean Length* y Monge Elkan. Por ende se procedió a realizar una nueva prueba, basado en el modelo UMCC más estas 3 métricas mencionadas.

5.8.9. Experimentos modelo base más 3 métricas

La realización de esta prueba tomó el modelo UMCC más las 3 métricas que impactaron de forma positiva en la mayoría de los modelos, Nedleman Wunch, *Chapman Mean Length* y Monge Elkan.

En la tabla 20 se puede ver los resultados en conjunto de todos los enfoque y modelos. Al comparar estos resultados, se puede observar que con las 3 métricas, los resultados mostraron, en forma general, buenos resultados.

Tabla 20: Tabla de coef. correlación modelo base más 3 métricas.

Enfoque / **Modelo**	**A**		**B**		**C**		**D**	
Random Forest	0.7953	**0.7968**	0.7689	0.7702	0.6716	0.6714	0.6469	0.6446
Dagging	0.7404	0.7422	0.7431	0.7439	0.6755	0.6271	0.6095	0.6076
Linear Regression	0.7081	0.7087	0.7185	0.7193	0.6298	0.5828	0.5632	0.5653
SMOreg	0.7064	0.7069	0.7195	0.7167	0.5774	0.5792	0.5596	0.5615

En la figura 18 se muestran los resultados gráficamente de los experimentos realizados en esta sección.

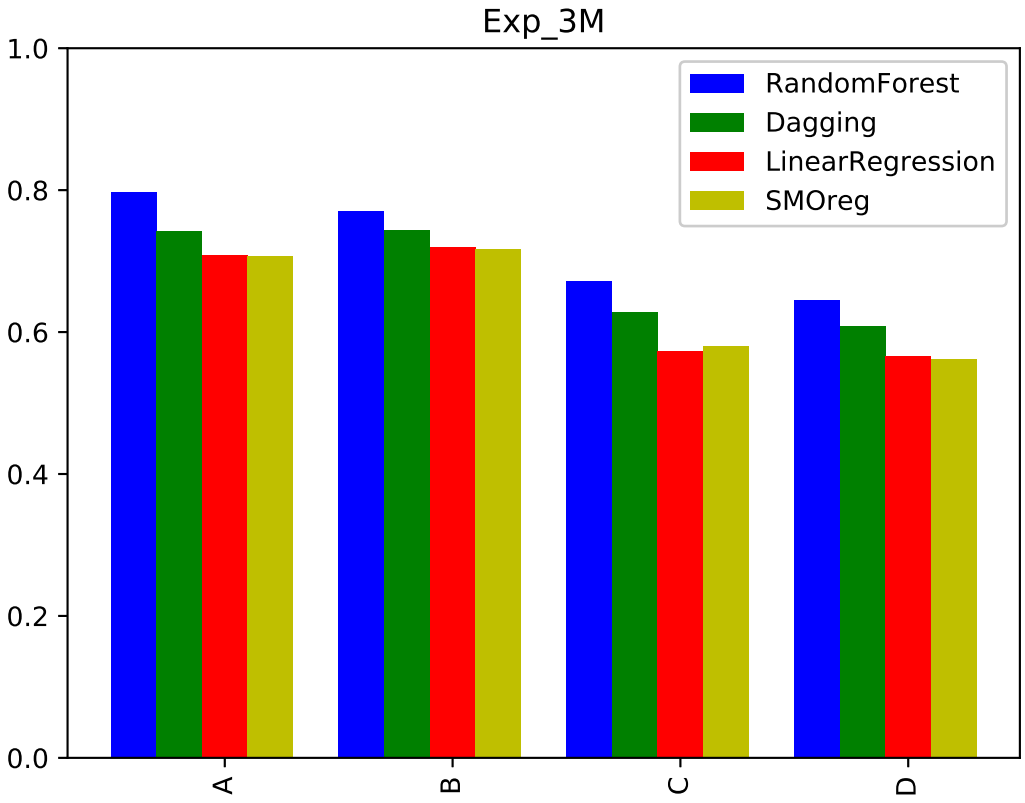

Figura 18: Resultados correlación modelo base más 3 métricas.

Los gráficos revelan que los enfoques A y B son ampliamente superiores que los enfoques C y D. En todos los enfoques el modelo *Random Forest* es el que obtiene mayor correlación, seguido por *Dagging* y hay una igualdad general en los modelos *Linear Regression* y SMOreg.

Al comparar los resultados, para el enfoque A, en todos los modelos, los resultados, en términos de correlación, mejoraron.

Para el enfoque B, en los modelos *Random Forest y Dagging*, los resultados mejoraron, en *Linear Regression* y SMOreg, los resultado disminuyeron, todo en términos de correlación.

Para el enfoque C, en los modelos *Random Forest, Dagging y Linear Regression*, los

resultados en términos de correlación, disminuyeron, pero en SMOreg, el resultado mejoró.

Para el enfoque D, en los modelos *Random Forest y Dagging*, los resultados disminuyeron, pero en *Linear Regression* y SMOreg, los resultados mejoraron, todo en términos de correlación.

En resumen, al dejar el modelo UMCC con las 3 métricas previamente mencionadas en esta sección, los resultados en la mayor parte de los enfoques, mejoraron. Para el enfoque con mayor resultado en términos de correlación (enfoque A), los resultados presentaron un mejora, no es el mayor resultado que se dio en todas los experimentos realizados en esta investigación. No se dio, de forma mayoritaria, que si las 3 métricas impactan de forma positiva o de forma negativa, varía en todos los enfoques, pero si se observa por el enfoque A, de mayor resultado, las 3 métricas impactan de forma positiva en los resultados.

5.9. Discusión general

Con todos los experimentos ya concluidos, se realizó un gráfico para cada modelo, con todos los enfoques, en todos los experimentos. Los experimentos están clasificados alfabéticamente en la sección 5.

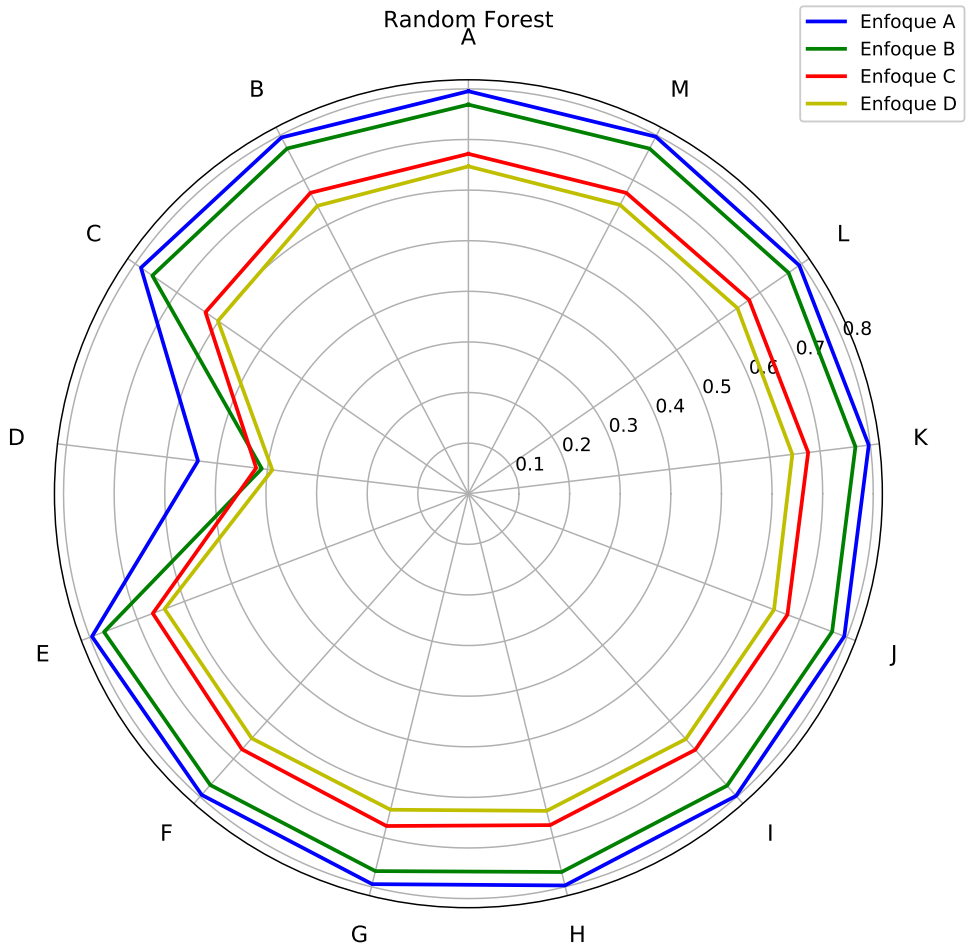

Figura 19: Resultados correlación todos los experimentos en *Random Forest*.

Al observar la figura 19, se deduce que el enfoque con mejor correlación en todos los experimentos es el enfoque A. Para todos los experimentos, a excepción del experimento D, el coeficiente de correlación bordea los 0.8. Además, el enfoque B sigue detrás del enfoque A, pero en el experimento D, el enfoque B fue sobrepasado por el enfoque C, por una distancia mínima. Los enfoques C y D se encuentran a una mayor distancia del enfoque A, no obstante, no sobrepasan la línea del 0.7 de coeficiente de correlación, por ende se concluyó que el mejor enfoque en términos de correlación, en el modelo *Random Forest*, es el enfoque A.

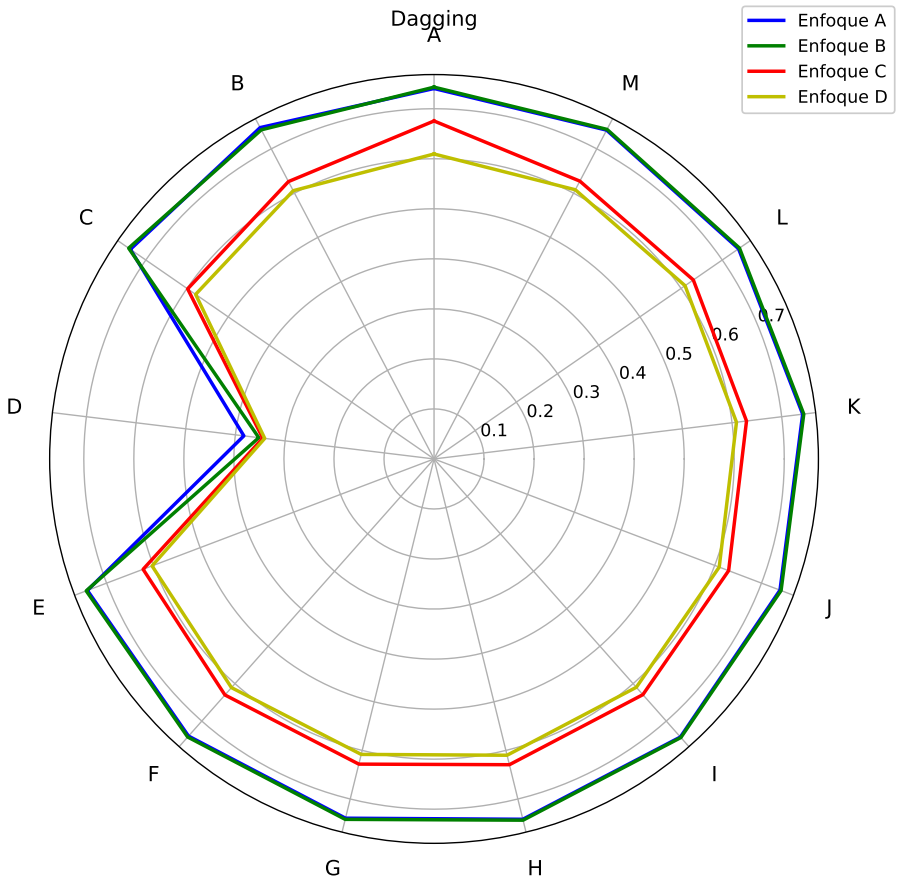

Figura 20: Resultados correlación todos los experimentos en *Dagging*.

Al observar la figura 20, existe una igualdad entre el enfoque A y el enfoque B en todos los experimentos, a excepción del experimento D, donde el enfoque A tomó distancia del enfoque B. Si se observan los datos de cada tabla, el enfoque B en la mayor parte de los experimentos, superó por una mínima diferencia al enfoque A, pero en general, ambos enfoques no sobrepasaron el rango de los 0.78 de coeficiente de correlación. Los enfoques C Y D se encuentran a una mayor distancia de los enfoques A y B, ambos no superan la línea del 0.7 de coeficiente de correlación, por ende, se concluyó que los enfoques A y B son los mejores enfoques en términos de correlación, en el modelo *Dagging*.

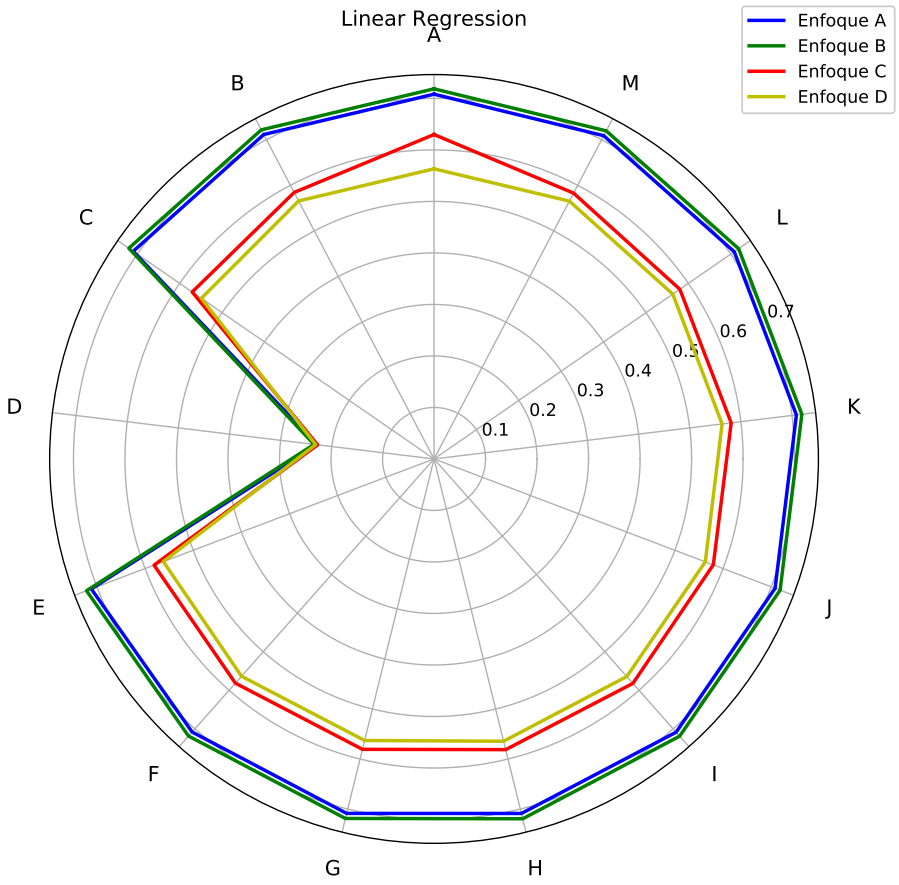

Figura 21: Resultados correlación todos los experimentos en *Linear Regression*.

Al observar la figura 21, existe una diferencia mínima entre el enfoque A y el enfoque B, siendo el enfoque B quien tomó mayor coeficiente de correlación. En el experimento D, la gráfica refleja una igualdad en todos los enfoques, pero en todos los demás experimentos, refleja una distancia de casi 0.1 de los enfoques C y D a los enfoques A y B. Ningún enfoque superó la línea de 0.7 de coeficiente de correlación. Se concluyó que el enfoque B es el mejor enfoque en términos de correlación en el modelo *Linear Regression*.

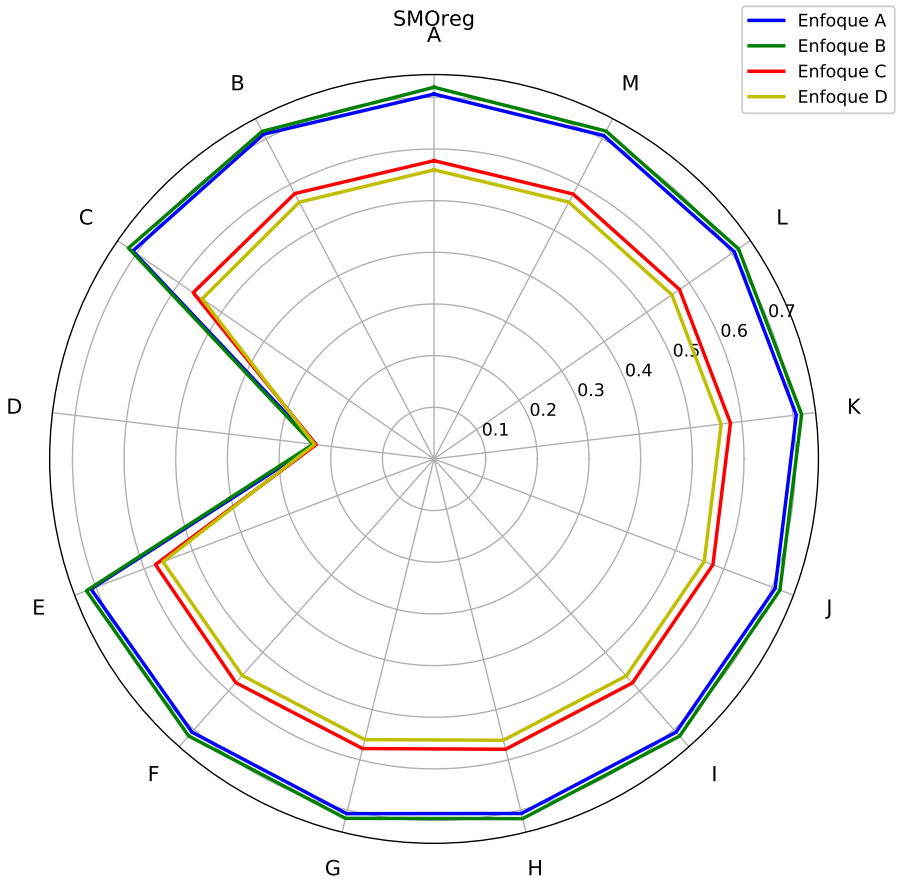

Figura 22: Resultados correlación todos los experimentos en SMOreg.

Al observar la figura 22, existe una diferencia mínima entre el enfoque A y el enfoque B, siendo el enfoque B quien tomó mayor coeficiente de correlación. En el experimento D, la gráfica refleja una igualdad en todos los enfoques, pero en todos los demás experimentos, refleja una distancia de casi 0.1 de los enfoques C y D a los enfoques A y B. Ningún enfoque superó la línea de 0.7 de coeficiente de correlación. Se concluyó que el enfoque B es el mejor enfoque en términos de correlación modelo en SMOreg.

En resumen, se probó que el modelo propuesto en esta investigación no es mejor al modelo de UMCC en términos de correlación, pero al experimentar con cada métrica propuesta en el modelo de esta investigación, se obtuvo 3 métricas que dan mejor resultado que el modelo UMCC. El modelo propuesto por esta investigación no responde de manera positiva a la hipótesis planteada.

5.10. Prueba suma de rangos Wilcoxon

Para saber que enfoque obtuvo resultados significativos, se realizó la prueba Wilcoxon, debido que aunque numéricamente un enfoque haya tenido resultados altos en términos de correlación, estos pueden no ser significativos. Dado que asumió que los datos no tienen distribución normal y los enfoques son independientes. Para la prueba se eligió en orden para comparar, desde el enfoque que dio mejores resultados al enfoque que dio peores resultados. Esta prueba se realizó para un nivel de 0.05 de significancia, siguiendo las hipótesis planteadas por cada comparación entre enfoques.

5.10.1. Enfoque A y Enfoque B

A) H_0: Mediana del enfoque A $\leq$ Mediana del enfoque B.

B) H_1: Mediana del enfoque A $>$ Mediana del enfoque B.

El valor p encontrado entre los datos de los enfoques A y B es de 0.7156, mayor a 0.05, por lo que no se rechaza la hipótesis nula. Por ende, los resultados del enfoque A no son más significativos que los del enfoque B.

5.10.2. Enfoque A y Enfoque C

A) H_0: Mediana del enfoque A $\leq$ Mediana del enfoque C.

B) H_1: Mediana del enfoque A $>$ Mediana del enfoque C.

El valor p encontrado entre los datos de los modelos enfoques A y C es de $3.4965e^{-10}$, menor a 0.05, por lo que se rechaza la hipótesis nula. Por ende, los resultados del enfoque A son más significativos que los del enfoque C.

5.10.3. Enfoque A y Enfoque D

A) H_0: Mediana del enfoque A $\leq$ Mediana del enfoque D.

B) H_1: Mediana del enfoque A $>$ Mediana del enfoque D.

El valor p encontrado entre los datos de los modelos enfoques A y D es de $4.1653e^{-10}$, menor a 0.05, por lo que se rechaza la hipótesis nula. Por ende, los resultados del enfoque A son más significativos que los del enfoque D.

5.10.4. Enfoque B y Enfoque C

A) H_0: Mediana del enfoque B $\leq$ Mediana del enfoque C.

B) H_1: Mediana del enfoque B $>$ Mediana del enfoque C.

El valor p encontrado entre los datos de los modelos enfoques B y C es de $4.4188e^{-10}$, menor a 0.05, por lo que se rechaza la hipótesis nula. Por ende, los resultados del enfoque B son más significativos que los del enfoque C.

5.10.5. Enfoque B y Enfoque D

A) H_0: Mediana del enfoque B $\leq$ Mediana del enfoque D.

B) H_1: Mediana del enfoque B $>$ Mediana del enfoque D.

El valor p encontrado entre los datos de los modelos enfoques B y D es de $3.4958e^{-10}$, menor a 0.05, por lo que se rechaza la hipótesis nula. Por ende, los resultados del enfoque B son más significativos que los del enfoque D.

5.10.6. Enfoque C y Enfoque D

A) H_0: Mediana del enfoque C $\leq$ Mediana del enfoque D.

B) H_1: Mediana del enfoque C $>$ Mediana del enfoque D.

El valor p encontrado entre los datos de los modelos enfoques C y D es de $4.1636e^{-10}$, menor a 0.05, por lo que se rechaza la hipótesis nula. Por ende, los resultados del enfoque C son más significativos que los del enfoque D.

5.11. Discusión

Al realizar la prueba suma de rangos de Wilcoxon entre todos los enfoques, se pudo obtener que los resultados del enfoque A no es más significativo que el enfoque B, es decir, la desambiguación por Lesk es igual o mejor que la desambiguación tomando el primer sentido como el más probable. Para los enfoques C y D, las *sense-phrase* no tuvieron significancia en los resultados, en la tabla 21 se ordenan los enfoques del más significativo al menos significativo

Tabla 21: Tabla orden de modelos.

Orden de enfoques
A-B
C
D

Capítulo 6

6. Conclusiones

6.1. Objetivo 1

Para la revisión bibliográfica sobre similitud semántica, se estudiaron los conceptos y se buscaron artículos científicos relacionados con el problema extraídos de algunas fuentes como ACM (*Association for Computing Machinery*), *Sciencedirect*, *Google scholar*, IEEE y por sobre todo, las publicaciones hechas por SemEval, *workshop* que invita a participar en este tipo de problema.

6.2. Objetivo 2

Para definir un esquema para combinar las métricas de similitud léxica y semántica, se tomó el modelo del 2014 UMCC (Chavez et al., 2014) que participó en la competencia SemEval. Además, la propuesta en esta investigación fue agregar 7 métricas léxicas nuevas al modelo para corroborar la hipótesis planteada en la sección 4. Los métodos para combinar todas las métricas, fue a partir de 4 modelos supervisados, esto quiere decir que los algoritmos necesitan ser entrenados con los datos de tal manera que puedan entregar en sus salidas una predicción de su variable dependiente. Estos modelos son *Random Forest, Dagging, Linear Regression* y SMOreg. Además, se realizaron 4 enfoques para ver el impacto que tenía cada enfoque con sus propias características de desambiguación de frase. Al final del proceso de cada frase, se obtiene un vector de datos que luego son entrenados en los modelos ya mencionados.

6.3. Objetivo 3

La sección 5 describe todos los experimentos realizados, en total 12 experimentos, para los 4 enfoques propuestos, en cada modelo. Se realizaron experimentos probando cada métrica nueva en el modelo UMCC, para ver el impacto de cada una, en cada enfoque. Esto permitió dar cuenta que las métricas impactan de forma distinta en cada enfoque. Además, dio cuenta de que modelo y enfoque da mejor resultado en términos de correlación.

6.4. Conclusiones generales y trabajos futuros

Como resultado de la investigación realizada para el estudio de la similitud semántica textual a través de la combinación de métricas léxica-semánticas demostró que, si bien la combinación entrega buenos resultados, la propuesta de esta investigación no genera un impacto positivo al aumentar las métricas léxicas, en términos de correlación, visto desde el modelo UMCC, con la desambiguación de tomar el primer sentido como el más probable (Enfoque A), ya que en los demás enfoques el impacto varía, en algunos mejora el resultado de correlación y en otros disminuye el resultado de correlación. Por ende la hipótesis expuesta en esta investigación no se corrobora con el modelo propuesto, pero al experimentar con cada métrica agregada en la propuesta, se puede corroborar la hipótesis planteada al mejorar los resultados con algunas métricas (Nedleman Wunch, *Chapman Mean Length* y Monge Elkan). Además, los modelos también juegan un rol importante, *Random Forest* es el que entregó mejores resultados en términos de correlación, lo que se destaca como un buen modelo para combinar datos, en este caso, métricas para medir el grado de similitud, aunque tampoco quiere decir que los demás modelos sean malos, pero no generan mayor resultado que *Random Forest*.

La desambiguación también se destaca, el enfoque B, que se desambiguó por Lesk, impactó de manera positiva, la prueba de Wilcoxon refleja que el enfoque B tiene una significancia igual o mejor que el enfoque A en los modelos probados en esta investigación.

El problema de las palabras con los rasgos léxicos descrito en la sección 4 fue tomado en cuenta y se trató en los enfoques C y D, se probó el impacto que genera el que los rasgos léxicos midieran sentidos y no palabras (*sense-phrase*), desambiguando tanto con Lesk (enfoque D) como el primer sentido como el más probable (enfoque C). Para estos enfoques, el impacto es negativo en todos los modelos, en términos de correlación, es baja en comparación a los enfoques A y B, por ende se determinó que el problema que se podría generar con las palabras en rasgos léxicos no impactan como para obtener un peor resultado. Si bien el problema que se genera con los rasgos léxicos al medir solo palabras, estos obtuvieron mayor correlación que las métricas semánticas, hay evidencia (Chavez et al., 2014) de que los rasgos léxicos son buenos para este problema, con solo rasgos léxicos el sistema UMCC obtuvo el primer lugar en la prueba de similitud semántica textual en español.

Si observamos los experimentos de la sección 5.8, la métrica léxica que mayor resultado obtuvo fue *Chapman Mean Length*. En esa sección se ve reflejado como impacta cada métrica en el modelo base, por ende, para un trabajo futuro se debiese probar cada métrica léxica del modelo en general y ver que impacto tiene cada una, Además, se debiese aumentar las métricas semánticas, ya que entregan información del contexto de la frase, que es fundamental en este tipo de problemas y tomar un número igual tanto de rasgos léxicos y métricas semánticas y comparar los resultados.

Al realizar la prueba suma de rangos de Wilcoxon, estadísticamente se concluyó que el enfoque A no es mejor que el enfoque B, los resultados en todos los modelos del enfoque A no reflejan mayor significancia que el enfoque B, debido a que A fue el enfoque con mayores resultados.

Referencias

Bär, D., Biemann, C., Gurevych, I., and Zesch, T. (2012). Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 435–440, Montréal, Canada. Association for Computational Linguistics.

Burgess, C., Livesay, K., and Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, 25(2-3):211–257.

Buscaldi, D., Garcia Flores, J., Meza, I. V., and Rodriguez, I. (2015). Sopa: Random forests regression for the semantic textual similarity task. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 132–137, Denver, Colorado. Association for Computational Linguistics.

Cavnar, W. B., Trenkle, J. M., et al. (1994). N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1.

Chapman, S. and Parkinson, C. (2006). *SimMetrics library v 1.5 for .NET 2.0 System and Reference Manual*. Sam Chapman, Department of Computer Science, University of Sheffield, Sheffield, S.Yorks, United Kingdom.

Chávez, A., Dávila, H., Gutiérrez, Y., Collazo, A., Abreu, J. I., Fernández Orquín, A., Montoyo, A., and Muñoz, R. (2013). Umcc_dlsi: Textual similarity based on lexical-semantic features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 109–118, Atlanta, Georgia, USA. Association for Computational Linguistics.

Chavez, A., Dávila, H., Gutiérrez, Y., Fernández-Orquín, A., Montoyo, A., and Muñoz, R. (2014). Umcc_dlsi_semsim: Multilingual system for measuring semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 716–721, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Corley, C. and Mihalcea, R. (2005). Measuring the Semantic Similarity of Texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, EMSEE '05, pages 13–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Croce, D., Storch, V., and Basili, R. (2013). Unitor-core_typed: Combining text similarity and semantic filters through sv regression. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 59–65, Atlanta, Georgia, USA. Association for Computational Linguistics.

Fellbaum, C. (2005). WordNet and wordnets. In Brown, K., editor, *Encyclopedia of Language and Linguistics*, pages 665–670, Oxford. Elsevier.

Gómez-Gómez, M., Danglot-Banck, C., and Vega-Franco, L. (2003). Sinopsis de pruebas estadísticas no paramétricas. cuándo usarlas. *Revista Mexicana de Pediatría*, 70(2):91–99.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of molecular biology*, 162(3):705–708.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Hirschberg, D. (1997). Serial computations of levenshtein distances.

Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of*

the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, pages 49–56.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Krause, E. F. (2012). *Taxicab geometry: An adventure in non-Euclidean geometry*. Courier Corporation.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Leacock, C. and Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Matsuo, Y., Tomobe, H., Hasida, K., and Ishizuka, M. (2004). Finding social network for trust calculation. In *Proceedings of the 16th European Conference on Artificial Intelligence*, pages 510–514. IOS Press.

Monge, A. E., Elkan, C., et al. (1996). The field matching problem: Algorithms and applications. In *KDD*, pages 267–270.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet:: Similarity: measuring the

relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics.

Perkins, J. (2014). *Python 3 Text Processing with NLTK 3 Cookbook*. Packt Publishing Ltd.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Schuetz, T. (2011). A concise guide to market research: the process, data and methods using ibm spss statistics.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.

Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, 38:1409–1438.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.*, 5:1–34.

Sun, Y., Ma, L., and Wang, S. (2015). A comparative evaluation of string similarity metrics for ontology alignment. *JOURNAL OF INFORMATION &COMPUTATIONAL SCIENCE*, 12(3):957–964.

Ting, K. M. and Witten, I. H. (1997). Stacking bagged and dagged models. In Fisher, D. H., editor, *Fourteenth international Conference on Machine Learning*, pages 367–375, San Francisco, CA. Morgan Kaufmann Publishers.

Torres, S. and Gelbukh, A. (2009). Comparing similarity measures for original wsd lesk algorithm. *Research in Computing Science*, 43:155–166.

Ukkonen, E. (1992). Approximate string-matching with q-grams and maximal matches. *Theoretical computer science*, 92(1):191–211.

Šarić, F., Glavaš, G., Karan, M., Šnajder, J., and Dalbelo Bašić, B. (2012). Takelab: Systems for measuring semantic text similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 441–448, Montréal, Canada. Association for Computational Linguistics.

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.