Capítulo 2

Marco teórico

2.1. Formulaicidad lingüística

Lingüística "es la ciencia del lenguaje" que estudia la facultad comunicativa de los humanos y las lenguas naturales involucradas en ese proceso, a través del análisis de los elementos que las constituyen, sus relaciones, leyes de su funcionamiento y sus formas, así como sus funciones para expresar el pensamiento humano dentro de un contexto social [5].

Los numerosos modelos de procesamiento del lenguaje en humanos, desarrollados dentro de esta ciencia, se pueden clasificar en dos enfoques principales, que son el **sistema analítico** y el **sistema holístico** [30]. La ventaja del sistema analítico es que nos permite construir nuevas frases basados en reglas, además de interpretar entradas nuevas o inesperadas. El sistema holístico, por otro lado, reduce el esfuerzo de procesamiento: hablando de eficiencia y eficacia, es mejor recuperar una cadena prefabricada que crear una nueva.

Los hablantes adultos por lo general presentan un balance entre estos dos sistemas, aunque en la práctica parece que la balanza se inclina un poco más por el holístico, al menos en ciertos dominios temáticos o situacionales de comunicación. Tales dominios

reciben el nombre de formulaicos, según la nueva teoría formulaica de la lingüística.

2.1.1. Teoría formulaica

La teoría formulaica es una nueva corriente de la lingüística teórica del idioma inglés, no ha pasado a formar parte de la lingüística aplicada, no se han formalizado sus conceptos matemáticamente, y tampoco se han desarrollado métodos de lingüística computacional basados en esa teoría. La razón por la que esta teoría no está siendo ampliamente utilizada aunque sí ha sido difundida y aceptada, radica en lo reciente de su aparición. La primera mención de esa teoría se da en el año 2000 en [31], y aún no ha sido trabajada por otros distintos a sus autores y no se le ha adaptado a otros idiomas. Actualmente esta teoría representa el estado de arte en la teoría lingüística del idioma inglés.

Los términos usados en el idioma en inglés para referirse a aspectos de esa teoría son *formulaic* y *formulaicity*, los cuales traducimos como **formulaico** y **formulaicidad** respectivamente (Véase sección 2.1.2).

"La **teoría formulaica** es un intento por Wray y Perkins de sintetizar e integrar en un único modelo lingüístico todas las teorías previas concernientes al tópico de la economía de esfuerzo en la comunicación humana" [21]. Combina el sistema analítico y el sistema holístico.

La teoría formulaica es una teoría lingüística reciente que establece que nosotros utilizamos fórmulas o expresiones preformadas como parte de lo que decimos y escribimos [31]; donde una **fórmula** es una oración completa o un grupo de palabras, o una palabra, o parte de una palabra, pero que sin importar su estructura debe ser siempre una unidad semántica dentro del discurso, es decir, no puede ser descompuesta y analizada por partes sino debe ser interpretada como un todo [30].

Una sola palabra es considerada fórmula para un dominio específico si forma parte de su terminología o léxico especializado, es decir, dentro de éste se le da preferencia a este término excluyendo el uso de sus sinónimos. Denominaremos este concepto como **palabra formulaica**.

Nos referiremos bajo el nombre de **frase** o **secuencia formulaica** a fórmulas constituidas por un grupo de palabras. Según Wray en [30], una secuencia formulaica es una secuencia continua o discontinua de palabras u otros elementos, que es, o parece ser, prefabricada: esto es, almacenada y recuperada como un todo de la memoria en el tiempo de uso, en lugar de ser generada o analizada por la gramática del lenguaje. En estás definiciones el concepto de frase "alude a un grupo de palabras que posee una independencia relativa dentro de la oración" [3].

2.1.2. Terminología

Formalismo

Hasta la fecha no existen trabajos en español que discutan la teoría formulaica, bajo este nombre. La única ocurrencia fue en [21], sin embargo en este trabajo no hay discusión acerca de por qué se utilizó la terminología en español. Por lo tanto, a continuación revisaremos la terminología relacionada utilizada en la lingüística del idioma español para determinar si alguno de los términos ya existentes puede ser utilizado en referencia a esta nueva teoría.

Los términos **formalista** y **formalismo** son asociados con la rama de lingüística llamada lingüística estructural, la cual hace énfasis en el análisis gramatical de la estructura de los enunciados, sin tomar en cuenta la semántica inherente en esas estructuras y el uso o la función comunicativa de las mismas [10]. Es decir, el término formalismo lingüístico es "aplicado a los lenguajes artificiales utilizados por los modelos

teóricos para la representación formal de sus análisis" [15]. Así, "en lingüística se llama formalismo a la tendencia a hacer descripciones formales del lenguaje, esto es, a la formalización de sus unidades y niveles, mediante la presentación explícita de su organización general y abstracta como código o sistema", donde la formalización se refiere a la notación formal o simbolización y "tiene la ventaja de exponer los conceptos con una brevedad relativa, contribuyendo a una presentación más clara y a un desarrollo más exacto de las nociones teóricas, es decir, tiene una simple función ilustradora". "El significado de 'formalismo' y 'formalista' nace de la acepción [...] del término forma, la cual se opone a función, es decir, a la finalidad [...] del lenguaje, entendido como instrumento de comunicación" [5]. A su vez, el nombre de **formulización** se le da a una formalización excesiva que "ha empobrecido las descripciones lingüísticas al convertirlas en pura forma matemática vacía de todo contenido, que va más allá de lo que en lingüística es útil y aceptable, por lo que algunos lingüistas se han resistido a caer en ella" [5].

Formulismo

Se dice que existe **formulismo** cuando existe demasiado apego a las fórmulas [11]; por otro lado, el lenguaje **formulista**, es aquel que pone en práctica el formulismo [11], es decir, se refiere al lenguaje formal. Basados en la definición de lenguaje formulaico, se puede decir que el lenguaje formal es un subconjunto de éste; el cual, además de incluir el lenguaje formal, también incluye a aquella lista de términos que describen o caracterizan un dominio (incluyendo a los dominios orales que no necesariamente son formales). Tal es el caso de las secuencias formulaicas usadas como herramientas de interacción social. Por ejemplo, en el caso del idioma inglés con el uso de expresiones ya definidas como *happy new year* y *merry new year* o *happy christmas* y *merry christmas*: aunque en ambos ejemplos, las dos variaciones son gramaticalmente correctas, no todas

se escuchan bien, porque se acostumbra decirlo de una forma solamente.

Formulaicidad

Después de analizar todos los términos que podrían ser utilizados para conformar la terminología de este proyecto, y así elegir a los que describan mejor y de preferencia de manera única los conceptos de la teoría lingüística que aplicaremos (eliminando así cualquier ambigüedad), tomamos prestados del idioma inglés los términos **formulaicidad** (*formulaicity*) y **formulaico** (*formulaic*), en lugar de utilizar los términos españoles formulismo y formulista, para marcar esta diferencia, debido a la confusión que existe entre formulismo y formalismo. Al utilizar los términos ingleses, nos aseguramos que nos referimos a un lenguaje basado en expresiones o palabras formulaicas y no a un lenguaje basado en expresiones formales.

Formulaicidad es manifestada en cadenas de palabras lingüísticas donde la relación de cada palabra con el resto está relativamente predefinida y donde la sustitución de una palabra por otra de la misma categoría está relativamente restringida.

Lenguaje formulaico

Término usado en algunos estudios teóricos y descriptivos de la gramática para referirse a las expresiones que carecen de características sintácticas o morfológicas normales. También puede ser utilizado para referirse al lenguaje que contiene fórmulas o símbolos especiales, como en la escritura científica [10]. Este término fue considerado para hacer referencia claramente al conjunto de fórmulas que describen un dominio. Aunque en nuestro trabajo se prefirió el término *secuencia formulaica* debido a que lenguaje formulaico, como ya vimos, ya es usado en la literatura para referirse también a lenguaje formal, no sólo al lenguaje que comprende secuencias formulaicas [30].

2.1.3. Dominios formulaicos

Dominio formulaico

Se refiere a los dominios que tienen un vocabulario cerrado o restringido. Aunque en la teoría formulaica aún no se ha llegado a desarrollar una definición formal de este concepto, como nos podemos dar cuenta al revisar el trabajo de Wray [30].

Corpus

"El nombre de **corpus** se aplica a toda colección de textos compilados según unos criterios explícitos que hacen que sea suficientemente representativo como para constituir, en si mismo, un modelo a escala de los aspectos lingüísticos que el investigador quiere examinar" [5]. Generalmente se refiere a una colección de documentos en una base de datos electrónica [18].

Córpora

Es el plural de la palabra corpus [18].

Lingüística de corpus

"La rama de la lingüística que estudia las lenguas basándose en grandes repertorios lingüísticos llamados córpora" [5].

2.1.4. Detección de formulaicidad

En el intento de establecer un criterio para detectar la formulaicidad de un dominio, se han realizado conteos asistidos por computadora de frecuencias de palabras y frases, se ha analizado la estructura interna (*compositionality*) de las oraciones escritas y la forma fonológica de las orales, etc. En todos los casos, se han encontrado

muy pocas bases para establecer un criterio robusto que nos permita calcular el grado exacto de formulaicidad. Entre los criterios que pueden ser de ayuda en este proceso, sin que ninguno de ellos por sí solo sea criterio suficiente, se han detectado semántica, estructura, fonética y frecuencia, los cuales se explican a continuación. Sin embargo, aun combinándolos todos, no se forma un criterio inequívoco, por lo cual la intuición juega un rol importante en la detección de formulaicidad [30].

Semántica

El criterio semántico se refiere al significado implícito de palabras, es decir que en toda fórmula debe haber presente por lo menos una palabra llena. "Las **palabras llenas** son palabras con significado denotativo, es decir que puede relacionarse con entidades, procesos, cualidades del mundo extralingüístico. En este grupo se incluyen los nombres y los adjetivos, los verbos y los adverbios" [5].

Por el contrario, "las **palabras vacías** son palabras cuyo significado es gramatical, es decir, un significado que expresa una relación, la pertenencia a una clase, etc. Son palabras vacías las preposiciones, las conjunciones, los artículos y los pronombres" [5]. Estas pueden estar presentes en las fórmulas a manera de conectores entre las palabras llenas.

Estructura

Estructura gramatical también se conoce como sintáctica debido a que es "integrada por funciones sintácticas que asumen las unidades integrantes de una construcción oracional" [5]. "Se llama **función** al papel que desempeña un término [...] en la estructura gramatical del enunciado, considerando que cada miembro de la oración contribuye a su sentido general. [...] En gramática generativa, la función es la relación gramatical que los elementos de una estructura (las categorías) mantienen entre sí dentro de esta estruc-

tura" [13]. La función de la palabra en la oración está regida por la categoría implícita de dicha palabra. "Por ejemplo, sustantivo, adjetivo, verbo son términos que designan 'categorías' (más específicamente conocidas como **categorías léxicas**) puesto que hacen referencia a las clases en que, tradicionalmente, se han agrupado las palabras que comparten determinadas propiedades" [5]. Así con este criterio se analizan las palabras por las funciones léxicas que cumplen dentro de una oración.

Fonética

Este criterio está basado en la coherencia fonológica, para identificar secuencias formulaicas. Está restringido a lenguaje hablado, aunque en los textos escritos se pueden ver algunas de sus características en la puntuación y su disposición. La hipótesis de este criterio es que los oyentes deben ser capaces de distinguir secuencias que no son formulaicas de las que sí lo son, por ejemplo, distinguir entre el significado literal de una frase y su uso en un sentido figurado. La hipótesis anterior fue probada en una serie de experimentos de Van Lancket y Canter. Algunas características fonológicas que son usadas para identificar secuencias formulaicas son: fluidez, tensión y articulación [30]. No se detallarán porque no son relevantes a nuestra aplicación, ya que nosotros vamos a tratar con textos escritos.

Frecuencia

En lingüística de corpus, las búsquedas por computadora se conducen para establecer patrones en la distribución de palabras dentro del texto. Este proceso se basa en contar la frecuencia, que revela con qué otras palabras ocurre más a menudo una palabra dada. Estos patrones de colocación resultan estar lejos del azar. Esto es, si se toma una cadena de palabras que es indiscutiblemente formulaica, se le puede buscar a través de un gran corpus y se puede ver que tiene una frecuencia alta y también es idiomática.

Estas asociaciones nos invitan a ver la frecuencia como una posibilidad, quizás un factor determinante en la identificación de secuencias formulaicas [30].

Existen algunos problemas asociados a tomar la frecuencia como único criterio para identificar la formulaicidad. No es capaz de diferenciar entre las ocurrencias de una configuración cuando es formulaica y una secuencia de palabras que no constituye una unidad semántica sino que se repite con la misma configuración por coincidencia. Además, no todo lo que es formulaico tiene alta frecuencia y tampoco se puede decir que todo lo prefabricado es formulaico.

Intuición

El método de extracción de secuencias formulaicas menos científico y más usado es la intuición. Para empezar, debemos recordar que existe una relación directa entre formulaicidad y modismos (*idiomacity*). Un modismo en la mayoría de los contextos debe ser interpretado en su sentido figurado y no en su sentido literal, pero sólo es identificado en términos de la intuición de los miembros de la comunidad para la que el discurso es relevante, es decir, una expresión es un modismo o expresión idiomática si suena natural y es considerada como una unidad semántica para esta comunidad. A manera de ejemplo, considérese el caso de la expresión en español de "ponte en mis zapatos" que, en la mayoría de los casos, no debe ser interpretada literalmente sino que se refiere a que alguien trate de comprender la situación por la que uno atraviesa. Aunque se aplican otras medidas, la intuición todavía guía el diseño de experimentos, la interpretación de resultados y la elección de ejemplos para hacer pruebas [30].

En este caso, la intuición influye directamente en la elección de los dominios que serán usados como casos de prueba, debido a que, hasta la fecha, no hay una metodología formal establecida para discernir entre un dominio formulaico y otro que no lo es.

2.2. Lingüística computacional

La tarea de la lingüística computacional consiste en describir cómo debe resolverse un problema lingüístico; para ello debe detallar los diferentes pasos que deben seguirse para llegar al resultado esperado [5] y finalmente aplicar métodos computacionales. "Esta ciencia es una combinación de dos ciencias más grandes; la lingüística, que estudia las leyes del lenguaje humano, y la inteligencia artificial, que investiga los métodos computacionales para el manejo de sistemas complejos" [33]. El problema que se busca resolver en la lingüística computacional es la comprensión del lenguaje.

La **comprensión del lenguaje** intenta representar de manera formal el lenguaje oral o escrito. Por lo general, para resolver este problema se opta por construir un procesador lingüístico conformado por los módulos que se describen a continuación.

1. El **módulo morfológico**: en éste las palabras son reconocidas, es decir, se identifica su tiempo, género y número.

2. El **módulo sintáctico**: en éste las oraciones son reconocidas, es decir, se estructuran gráficamente, estableciendo las relaciones entre las palabras.

3. El **módulo semántico**: es el último paso; en éste se reconoce la estructura completa del texto y se convierte a una "red semántica". El concepto de red semántica se define posteriormente (Véase sección 2.6.4).

Además de tratar la comprensión del lenguaje, la lingüística computacional también explora otras áreas; una de las más grandes es el procesamiento automático de textos que abarca desde parsers hasta minería de textos [33].

2.3. Recuperación de información

También llamada recuperación de documentos, es el proceso sistemático en el que se manipula la información textual con la única finalidad de que pueda ser encontrada de nuevo con facilidad (recuperación). Particularmente son sistemas orientados a información textual no estructurada, para la cual se lleva a cabo un proceso de indizado con el fin de darle estructura.

El proceso de RI se vale de varias técnicas para buscar documentos relevantes a una consulta dentro de un repositorio. Para esto, los documentos almacenados deben estar clasificados y agrupados. El proceso de clasificación se lleva a cabo actualmente con métodos no automatizados en donde se asignan palabras claves relevantes al tema alrededor del cual gira el texto. Los métodos automatizados existen pero son complejos o no son confiables: complejos porque están basados en técnicas de vectorización o probabilísticas y se necesita estructurar los textos (preprocesamiento), por otro lado no son confiables porque en algunos casos sólo toman en cuenta la repetición de palabras en el texto, y éste no es un criterio definitivo (Véase sección 2.1.4).

Al final este proceso informa al usuario si existen o no documentos, dentro del repositorio, que contengan la información solicitada en la consulta.

El área de RI comparte aspectos similares con otras áreas relacionadas con procesamiento de información, por ejemplo los sistemas expertos y los sistemas manejadores de bases de datos.

2.4. Minería de textos

“La minería de texto se enfoca en el descubrimiento de patrones interesantes y nuevos conocimientos en un conjunto de textos, es decir, su objetivo es descubrir cosas

tales como tendencias, desviaciones y asociaciones entre la "gran" cantidad de información textual" [33]. Según Tan, citado en [33], ésta se realiza a través de dos etapas principales que se describen a continuación.

1. **Etapa de preprocesamiento:** para facilitar su análisis es recomendable que se les de cierta estructura, es decir que los textos se representen de manera estructurada o semi-estructurada.

2. **Etapa de descubrimiento:** en esta etapa se analiza la representación y se buscan patrones o conocimiento.

Los métodos que se ocupan durante el preprocesamiento influyen en la forma en que se representan los textos, y el tipo de patrones encontrados depende de esta representación [33]. Algunas de las tareas de la minería te textos son categorización, agrupamiento (*clustering*) y asociación.

2.4.1. Categorización

Es el proceso de encontrar un conjunto de modelos o funciones que describan y, sobre todo, que distingan clases; en el caso particular de textos, las clases se refieren al nivel temático [16]. A este proceso también se le conoce como **clasificación**.

2.4.2. Sumarización

Es un proceso automatizado que sirve para crear resúmenes breves de documentos. La idea general es obtener oraciones claves que reflejen el tema del texto. Idealmente se obtiene la idea general del texto y esto es útil para reducir problemas de comprensión, además de evitar la pérdida de información clave.

2.4.3. Agrupamiento (*Clustering*)

El agrupamiento (*clustering*) de documentos intenta asociarlos por contenido para reducir el espacio de búsqueda requerido para responder a una consulta [14]. En general se intenta tener grupos, donde cada grupo contendrá un objeto que es similar al resto de los objetos y será diferente de los objetos de otros grupos; la similitud entre objetos por lo general está calculada en base a una función de distancia. Para formar un *cluster*, se pretende maximizar la semejanza entre objetos de ese grupo; por el contrario, se intenta minimizar la semejanza con objetos de grupos distintos [16]. En este caso, este método se refiere a encontrar patrones dentro de la estructura de los textos.

2.4.4. Reglas de asociación

Se pretende encontrar dependencias entre las variables que describen los objetos (por ejemplo textos) almacenados en un repositorio con el fin de inferir o predecir el valor de estas variables en diferentes circunstancias.

2.5. Algoritmos

2.5.1. Algoritmo de reducción a raíces léxicas (*stemming*)

Este tipo de algoritmo tiene como objetivo reducir las palabras a su raíz léxica. Uno de los más populares es el algoritmo de Porter.

Algoritmo de Porter

El algoritmo de Porter es un proceso para remover las terminaciones morfológicas e inflexionales de palabras en inglés (originalmente). Su uso principal es como parte de un proceso de normalización que se aplica generalmente en sistemas de recuperación

de información [22]. En la actualidad existen adaptaciones de este algoritmo a varios idiomas.

2.5.2. Algoritmo de colocación

Éste es un algoritmo que nos permite analizar la colocación de las palabras con respecto a otra que se escoge como palabra clave, es decir, la función que toma una palabra de acuerdo a su posicionamiento en la oración. El funcionamiento de este algoritmo consiste en tomar una palabra clave; en base a ésta se tomarán las n palabras a su izquierda y n a su derecha y se contará cuántas veces aparece cada palabra en cada una de las posiciones. Este algoritmo toma como principio de evaluación la colocación de las palabras y la frecuencia de ésta, no realiza análisis semántico ni sintáctico.

2.5.3. *N*-gramas

El modelo *n-grama* es un modelo de predicción de palabras, usa las n-1 palabras previas para predecir la siguiente. Es un modelo probabilístico donde se asigna probabilidad a cadenas de palabras, de tal manera que permite obtener la probabilidad de una oración completa o dar una predicción probabilística de cuál es la próxima palabra en una secuencia [17]. Al igual que el algoritmo de colocación éste checa colocación, frecuencia y no realiza análisis gramatical ni sintáctico.

2.5.4. Algoritmo de distancias

Este algoritmo compara cadenas y calcula la distancia entre éstas. Donde la distancia entre dos cadenas está definida como el número de operaciones necesarias para transformar una cadena x en una cadena y. Las operaciones posibles incluyen sustitución, inserción y eliminación, y cada una de ellas tiene el mismo costo [7]. Éste es el

algoritmo más común para la comparación de cadenas; existen extensiones de este algoritmo que intentan reducir su complejidad y mejorar la exactitud. Aplica los mismos criterios de los dos algoritmos anteriores.

2.5.5. Vectorización

Es una de las estrategias de recuperación de información, en donde la consulta y cada documento son representados como vectores. En todos los casos se encuentra una medida de similaridad entre los dos vectores para encontrar los documentos relevantes a la consulta [14]. Aplica un modelo matemático, en donde se asignan pesos a términos del documento basados en sus frecuencias. No aplica ningún análisis semántico ni sintáctico.

2.5.6. Redes neuronales

Es una secuencia de neuronas o nodos interconectados en una red a través de una serie de ligas. Los nodos finales están conectados a los documentos almacenados en un repositorio. Una consulta activa la red enviando un impulso eléctrico que viaja entre los nodos y cuyo valor al final del trayecto es igual a los pesos acumulados asociados a las ligas por las que viajó; dicho valor se interpreta como coeficiente de semejanza entre la consulta y el documento encontrado al final de cada trayecto. Las redes son entrenadas para ajustar los pesos en las ligas y así distinguir entre los documentos relevantes e irrelevantes para una búsqueda dada [14]. Aplica un modelo matemático para asociar pesos a las ligas entre los nodos, estos pesos se van modificando de acuerdo a la experiencia. No aplica ninguno de los criterios convencionales (frecuencia, colocación, análisis sintáctico y semántico).

2.5.7. Indizado

Este proceso consiste en asociar un peso a los términos de un texto, tales pesos están en proporción con la relevancia que éstos tienen dentro del texto, es decir, aquellos términos que en realidad describen el contenido o el tema en torno al cual gira el texto tendrán mayor relevancia. Generalmente el indizado es un proceso no automatizado que se realiza con ayuda de un experto humano, pues el criterio de selección de este algoritmo es la relevancia semántica de las palabras en un contexto.

2.5.8. Método bayesiano

Es un método probabilístico que determina un valor de similitud entre una consulta y un documento; este valor se determina la probabilidad de que el documento sea relevante a la consulta. Se calcula para cada término de la consulta [14]. En éste se aplica el teorema de Bayes, el cual tiene como objetivo determinar la dependencia entre eventos. Funciona básicamente así, se conoce la probabilidad de que ocurra un evento E_1, por otro lado tenemos el evento E_2 y se quiere saber si su ocurrencia afecta al evento E_1. En otras palabras, y para nuestro caso mientras mayor sea esta dependencia, mayor es su relevancia. No se aplica ningún análisis semántico ni sintáctico.

2.5.9. SUBDUE (Substructure discovery system)

Es un sistema de descubrimiento de conocimiento, que encuentra subestructuras interesantes y repetitivas (subgrafos) en la información de entrada representada como un grafo etiquetado [8]. Permite elegir entre tres principios de evaluación, los llamados MDL (*Minimum Description Length*), *"set cover"* y *"size"*; aunque el más común es MDL. Los principios de evaluación anteriores pueden ser estudiados a detalle en [23].

SUBDUE representa la información como un grafo etiquetado, donde los vértices

representan objetos o atributos, y las aristas representan las relaciones entre los vértices. Entonces, como entrada al sistema tenemos vértices y aristas, y como salida, los patrones descubiertos e instancias de los mismos [8].

A grandes rasgos, SUBDUE lleva a cabo su búsqueda empezando con un solo vértice y se expande repetidamente por una arista o un nodo y una arista. Su espacio de búsqueda son todos los subgrafos del grafo de entrada y utiliza una heurística de compresión [8].

Según [8], el algoritmo, a grandes rasgos, sigue la secuencia de pasos presentada a continuación:

1. crear una subestructura por cada etiqueta única de vértice;
2. expandir la mejor subestructura por una arista o por una arista más un vértice vecino;
3. terminar cuando la cola esté vacía o el número de subestructuras descubiertas sea mayor o igual a un límite dado;
4. comprimir el grafo y repetir para generar descripción jerárquica.

SUBDUE maneja múltiples variables con las que podemos experimentar para obtener información más detallada, lo cual se refleja en resultados más exactos. En general, las variables nos ayudan a configurar el formato de resultados de SUBDUE: por ejemplo, la cantidad de subgrafos arrojados en cada iteración, el tamaño de los subgrafos que queremos como resultados, etc.

MDL (*Minimum Description Length*)

Minimiza el DL de un conjunto de datos, es decir, busca la subestructura que representa la máxima compresión de un grafo, donde DL de un grafo es el número de bits

necesarios para describir a un grafo completamente.

2.6. Representación

2.6.1. Grafos

Es un conjunto de puntos llamados vértices, y un conjunto de parejas de vértices llamadas aristas.

2.6.2. Estructuras *trie*

"Es un árbol n-ario" [...] "donde n es el grado máximo de los nodos del árbol" [19]. Representan claves como secuencias de dígitos o caracteres alfabéticos. Es útil cuando se necesita hacer una búsqueda caracter por caracter. Cada nodo representa un estado, al cual se llega recorriendo los caractéres que este estado está representando; para marcar el final de una cadena se usa el caracter especial '#'[19].

2.6.3. Estructuras *two-trie*

"El *two-trie* intenta resolver el problema de compactación de un *trie* sin perder la unicidad de las claves"[19]. Consiste de dos *trie*: uno para prefijos y otro para terminaciones.

2.6.4. Redes semánticas

Se puede decir que es una forma de representar conocimiento; las redes semánticas tienen asociada una sintaxis y una semántica. A través de círculos etiquetados que se llaman nodos, se representan objetos. Además puede haber relaciones entre cada par de objetos, representadas con arcos etiquetados. La diferencia entre una red semántica

y una red cualquiera es que tiene asociada la semántica, es decir, un nodo representa conceptos y existen relaciones de herencia o instanciación entre los objetos.

2.6.5. Vectores

Un vector es una estructura matemática específica. Tiene numerosas aplicaciones en física y geometría, debido a que tiene la característica de representar magnitud y dirección simultáneamente. Por lo general, la ubicación de un punto en el plano cartesiano está expresado como un par de coordenadas *(x,y)*, el cuál es ejemplo de un vector. Este vector *(x,y)* tiene una distancia (magnitud) y un ángulo (dirección) con respecto al origen (0,0). Los vectores son útiles para simplificar problemas en geometría de tres dimensiones [2].

2.6.6. Árboles gramaticales

Llamados también *parse trees*, se tratan de una estructura de árbol que describe la derivación de una oración con un lenguaje formalista de acuerdo a las reglas de la gramática. La raíz del árbol es un símbolo que marca el comienzo de la expresión, mientras que los nodos-hoja representan los lexemas de la entrada. Los nodos intermedios representan las relaciones de producción/reducción de la entrada, descritas por la gramática [1].

2.7. Trabajos relacionados

2.7.1. Proyectos en lingüística y lingüística computacional

Existen varios proyectos en lingüística para realizar análisis que pueden ser hechos usando una computadora. Principalmente se tiene interés en buscar:

- **Lista de palabras**: si se está interesado en conocer cuáles palabras ocurren más frecuentemente en los textos.
- **Concordancia**: si se quiere conocer qué orden de palabras tiende a ocurrir con respecto a una palabra dada.
- **Distribución** de palabras a través de varias partes del texto.
- **Colocación**: si se quiere saber cómo está asociado un conjunto de palabras con otras palabras.

Existen varios programas disponibles para el análisis de textos, tal es el caso de WordSmith Tools, MicroConcord, WordCruncher, este último para indexar textos.

Algunos ejemplos de proyectos que aplican lingüística computacional se presentan a continuación:

- **Generador de diccionario**. El primer diccionario generado completamente por computadora fue el *Collins COBUILD English Language Dictionary*, que está basado en un conjunto de textos, conocido como el *Bank of English* [32]. "Se trata de un diccionario diseñado para el aprendizaje de la lengua inglesa cuya naturaleza, calidad y presentación lo convierten en una obra relevante. Está basado en un corpus representativo de textos ingleses de gran variedad del que se extrajo información lexicográfica sobre el significado y uso de las palabras, patrones sintácticos que caracterizan las diferentes acepciones, colocaciones más frecuentes, etc." [20].
- **Clasificación automática de textos de desastres naturales en México**. Este trabajo fue desarrollado recientemente por el Laboratorio de Tecnologías del Lenguaje del INAOE como parte de un proyecto más grande que tiene como finalidad crear automáticamente un repositorio donde se almacene información

relevante a desastres naturales en México. El proyecto más grande consiste en combinar métodos de búsqueda de información y de clasificación de textos para crear el repositorio mencionado anteriormente. En este trabajo, utilizando técnicas de clasificación de textos, se filtraron páginas Web que tuvieran información relevante al dominio de los desastres naturales.

Todo el proceso se llevo a cabo en varias etapas, la primera fue de preprocesamiento donde se removieron las etiquetas html, eliminando palabras vacías, puntuación y reduciendo a raíces léxicas. La siguiente etapa es de indizado para posteriormente aplicar el método vectorial. Por último, se aplican técnicas clasificadoras entre ellas el método bayesiano.

Al final, se obtuvieron resultados satisfactorios, concluyendo que es posible clasificar una página Web dentro de categorías como: huracán, inundación, sequía y no relevante con un 97 % de precisión [27].

- **Ejemplo de algoritmo de colocación** Se analizó la obra de Thomas Hardy, *Far from the Madding Crowd*, se estudió la múltiple aparición de la palabra 'red' y qué secuencia de palabras la antecedían y proseguían. En este caso, se tomaron las cinco palabras a su derecha y las cinco palabras a su izquierda y se mostró con qué frecuencia se encontró cada una en cada posición [32].

2.7.2. La Web semántica

Es una extensión de la Web actual en la cual se le da a la información un significado bien definido; la Web semántica es la representación abstracta de datos sobre el World Wide Web, basada en los estándares de RDF y otros estándares [29]. En realidad se define la información basándose en ontologías definidas por expertos que permiten hacer inferencias con el fin de, por ejemplo, clasificar la información.

La Web semántica provee un marco que permite que la información en Internet sea compartida y reutilizada a través de aplicaciones, etc. Es un esfuerzo colaborativo dirigido por W3C con la participación de muchos investigadores y patrocinadores industriales. Está basado en RDF (*Resource Description Framework*), el cual integra una variedad de aplicaciones usando XML para la sintaxis y URIs para el nombrado [28].

Existen muchos trabajos relacionados con la Web semántica; aquí en particular mencionaremos Arkquakt, un sistema que extrae conocimiento de documentos en la Web dentro de un dominio restringido (artistas). Artequakt extrae automáticamente conocimiento acerca de artistas de la Web, construye una base de conocimientos y la usa para generar biografías. Este proyecto relaciona herramientas de extracción de conocimiento con una ontología para lograr un soporte de conocimiento continuo y una guía de extracción de información. La herramienta de extracción busca documentos en línea y extrae conocimiento que coincide con la estructura de clasificación dada. Provee este conocimiento en un formato legible por la máquina y que será depositado y mantenido automáticamente en una base conocimiento.

Durante la primera parte de Artequakt, se crea una ontología para el dominio de artistas y pintores. Se aplican varias herramientas y técnicas de extracción de información que automáticamente llenan la ontología con la información extraída de los documentos en línea. El sistema almacena la información en la base de conocimiento y la analiza para las duplicaciones. En la segunda parte, se desarrollaron herramientas para construcción de narrativa para consultar la base de conocimientos a través de un servicio de ontologías que busca y recupera textos o hechos relevantes y genera una biografía específica [4].

2.7.3. Minería de textos

Sumarización

Existen en el mercado varias aplicaciones que realizan sumarización de documentos, en particular el llamado *Subject Search Summarizer*, el cual tiene la función de crear breves resúmenes de cualquier documento o página Web. Esta herramienta genera y despliega un resumen formado por una lista de oraciones clave que son obtenidas de los documentos usando un algoritmo. Como salida produce oraciones que reflejan el tema de un documento particular, lo cual permite al usuario reducir el tiempo necesario para entender el significado general de los documentos, leer sin perder información clave y familiarizarse con la estructura de los documentos. Esta aplicación funciona independientemente del formato del archivo. Su desempeño es independiente del tamaño y del idioma, soporta un poco más de 30 lenguajes, detectando automáticamente el idioma [26].