



UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

“Detección del grado de aceptación de un
texto de acuerdo al contexto y dominio”

Para Obtener el Título de
Ingeniero en Software

Presenta
Iván Hernández Martínez

Asesor:
Dr. René Arnulfo García Hernández

TIANGUISTENCO, MÉX. JULIO 2018



El comité revisor designado por el Departamento Académico de la Unidad Académica Profesional Tianguistenco de la Universidad Autónoma del Estado de México, aprobó la tesis: **Detección del grado de aceptación de un texto de acuerdo al contexto y dominio** y autorizó la impresión de la misma del C. Iván Hernández Martínez el día **10 de julio 2018**.

**ATENTAMENTE
PATRIA, CIENCIA Y TRABAJO**

"2018, Año del 190 Aniversario de la Universidad Autónoma del Estado de México"



Dr. José Luis Tapia Fabela



Dra. Yulia Nikolaevna Ledeneva



**Dr. René Arnulfo García
Hernández**



Dra. Adriana Fonseca Munguía
Jefa del Departamento Académico de
la UAP Tianguistenco
Vo.Bo.



Declaración de originalidad del trabajo escrito

Mediante esta carta hago constar que el trabajo de tesis presentado en este documento es original porque cita debidamente los contenidos utilizados como soporte a la investigación presentada, por lo que exonero a la Universidad Autónoma del Estado de México de cualquier problema de derechos de propiedad intelectual.



Iván Hernández Martínez

Resumen

El lenguaje natural lo usan los humanos para comunicar una idea, sentimiento o pensamiento con su entorno. El lenguaje natural evoluciona para adaptarse a las nuevas necesidades de comunicación de los humanos. Esta evolución genera cambios en diferentes contextos y dominios y además en cada dominio se usan terminologías diferentes. Los correctores ortográficos y gramaticales son útiles para detectar errores en un dominio y contexto específico del español bien escrito. Sin embargo, no son útiles para detectar la aceptación que tiene un texto con un contexto y dominio específico. Debido a esto, publicaciones de Facebook que contienen un mensaje para un grupo específico se detectan como error ortográfico o gramatical. Sin embargo, para el grupo que lee el mensaje y lo decodifica no contiene un error.

Para ayudar a detectar palabras poco frecuentes en un texto se han desarrollado métodos basados en el análisis estadístico. Los métodos basados en análisis estadístico demuestran buenos resultados, pero requieren de un corpus muy grande para detectar palabras poco frecuentes en el español. Además, esto no garantiza que el corpus contenga todas las combinaciones de una palabra en español. Por esta razón se propone el uso de un modelo de frecuencia de n -gramas para detectar el grado de aceptación de un texto de acuerdo a un contexto y dominio. Además, para evaluar nuevas palabras que no se tienen en un corpus se propone el uso de un modelo de aprendizaje automático que evalúe nuevas palabras.

Contenido

Página

ÍNDICE DE FIGURAS	VIII
ÍNDICE DE TABLAS	IX
ÍNDICE DE GRÁFICAS	X
CAPÍTULO 1. INTRODUCCIÓN.....	1
1.1 Planteamiento del problema	8
1.2 Justificación.....	8
1.3 Hipótesis.....	9
1.4 Alcances y limitaciones	9
1.5 Objetivo general	9
1.6 Estructura de la tesis.....	10
CAPÍTULO 2. MARCO TEÓRICO.....	11
2.1 Lenguaje natural	11
2.2 Procesamiento del lenguaje natural (PLN).....	12
2.2.1 Modelo.....	12
2.2.2 Lingüística	13
2.2.2.1 Morfología.....	13
2.2.2.2 Sintaxis	13
2.2.2.3 Semántica	13
2.3 Corpus	14
2.4 Modelos de lenguaje estadísticos.....	14
2.5 Aprendizaje	16
2.6 Aprendizaje automático	17
2.6.1 Redes neuronales	18
2.6.1.1 Neurona	18
2.6.1.2 Perceptrón.....	20
2.6.1.3 Función de activación perceptrón.....	21
2.6.1.4 Redes neuronales multicapa	24
2.6.1.4 Entrenamiento de la red neuronal multicapa	25
2.6.1.4 Backpropagation.....	26
2.8 Resumen	28
CAPÍTULO 3. ESTADO DEL ARTE	29

3.1 Trabajos que evalúan la detección de errores gramaticales	29
3.2 Trabajos que detectan errores gramaticales en el idioma inglés	30
3.3 Trabajos que detectan errores gramaticales en español	31
3.4 Resumen	33
CAPÍTULO 4. MÉTODO PROPUESTO	34
4.1 Descripción del método propuesto	34
4.2 Etapas del método propuesto	36
4.2.1 Preprocesamiento	38
4.2.2 División de las secuencias de texto	38
4.2.3 Separación de palabras y signos	38
4.2.4 Representación de las oraciones con diferentes modelos de texto	38
4.2.5 Representación de los modelos de n-gramas con su frecuencia por medio de grafos.....	39
4.2.5.1 Representación de las oraciones y frecuencias de unigrama por medio de grafos.....	39
4.2.5.2 Representación de las oraciones y frecuencias de bigrama contiguos por medio de grafos	40
4.2.5.3 Representación de las oraciones y frecuencias de bigrama con salto en s por medio de grafos	41
4.2.5.4 Datos de entrenamiento.	43
4.2.6 Entrenamiento de la red neuronal	43
4.2.6.3 Creación de las capas de entrada, oculta y salida con diferente número de neuronas	43
4.2.6.4 Inicio de entrenamiento	43
4.2.6.5 Detección del grado de aceptación de un texto de acuerdo al contexto y dominio.....	44
CAPÍTULO 5. EXPERIMENTACIÓN.....	45
5.1 Datos de entrada	45
5.2 Preprocesamiento	45
5.2.1 Obtener artículos HTML de Wikipedia	46
5.2.3 Eliminar etiquetas HTML de artículos de Wikipedia	46
5.3 División de las secuencias de texto	48
5.4 Separación de letras y signos	49
5.5 Representación de las oraciones con diferentes modelos de texto.....	50
5.4 Representación de las oraciones y las frecuencias de modelos de n-gramas por medio de grafos	51
5.4 Entrenamiento de la red neuronal	52
5.4.1 Entrenamiento de la red neuronal con unigrama	53
5.4.2 Entrenamiento de la red neuronal con bigrama contiguo (salto igual a 0).....	54
5.4.3 Entrenamiento de la red neuronal con bigrama con salto en 1	55
5.5 Detección del grado de aceptación de un texto de acuerdo al contexto y dominio.....	56
5.5.1 Detección del grado de aceptación de un fragmento de un artículo de <i>Wikipedia</i> en español.....	56
5.5.2 Detección del grado de aceptación de un fragmento de poemas	57
5.5.3 Detección del grado de aceptación de un fragmento de un libro en español	58
5.5.4 Detección del grado de aceptación de un fragmento de una noticia es español.....	59

5.5.5 Detección del grado de aceptación de un fragmento de publicaciones de Facebook	60
5.5.6 Detección del grado de aceptación de un fragmento de Wikipedia en idioma inglés	60
5.5.7 Detección del grado de aceptación de un fragmento de Wikipedia en idioma alemán	61
5.6 Resumen	64
CAPÍTULO 6. CONCLUSIONES	65
6.1 Conclusiones	65
6.2 Aportaciones	66
6.3 Trabajo futuro	66
BIBLIOGRAFÍA	67

Índice de Figuras

FIGURA 1.1 PROCESO COMUNICATIVO.....	3
FIGURA 1.2 ORACIÓN ETIQUETADA.....	5
FIGURA 1.3 ORACIÓN SIN ERROR SINTÁCTICO PERO CON ERROR SEMÁNTICO	5
FIGURA 2.1 PARTES DE LA NEURONA (CRUZ, 2011).	19
FIGURA 2.2 DIAGRAMA DEL PERCEPTRÓN	20
FIGURA 2.3 FUNCIÓN PASO.....	22
FIGURA 2.4 REPRESENTACIÓN GRÁFICA DE LA TABLA DE VERDAD.	23
FIGURA 2.5 REPRESENTACIÓN XOR.....	24
FIGURA 2.6 PERCEPTRÓN MULTICAPA	25
FIGURA 2.7 FUNCIÓN SIGMOIDEA.....	26
FIGURA 2.8 RED NEURONAL <i>BACKPROPAGATION</i>	27
FIGURA 4.1 MÉTODO PROPUESTO GENERAL	37
FIGURA 4.2 MODELO DE FRECUENCIA DE UNIGRAMA	39
FIGURA 4.3 MODELO DE BIGRAMA	40
FIGURA 4.4 MODELO DE BIGRAMA CON SALTO EN S.....	41
FIGURA 5.1 INTERFAZ DE KIWIX	46
FIGURA 5.2 ETIQUETAS HTML QUE CONTIENE LOS ARTÍCULOS.....	47
FIGURA 5.3 SECUENCIAS DE TEXTO	48
FIGURA 5.4 ORACIONES FORMADAS DE LAS SECUENCIAS DE TEXTOS	49
FIGURA 5.5 PALABRAS Y SIGNOS SEPARADOS POR UN ESPACIO	50

Índice de Tablas

TABLA 1.1 FRECUENCIA DE PALABRAS.....	7
TABLA 2.1 MODELO DE UNIGRAMA Y BIGRAMA.....	15
TABLA 2.2 MODELO DE BIGRAMA SALTO N	16
TABLA 2.3 ANALOGÍA DE UNA NEURONA	20
TABLA 2.4 TABLA DE VERDAD	22
TABLA 3.1 TRABAJOS QUE USAN MODELOS DE N -GRAMAS Y ANÁLISIS ESTADÍSTICO.....	33
TABLA 4.1 MODELOS DE N -GRAMAS	35
TABLA 4.2 REPRESENTACIÓN DE LA ORACIÓN CON LOS MODELOS DE UNIGRAMA, BIGRAMA CONTIGUOS Y BIGRAMA CON SALTO EN S .	38
TABLA 4.3 MODELO DE FRECUENCIA DE UNIGRAMA	40
TABLA 4.4 MODELO DE FRECUENCIA DE BIGRAMA	40
TABLA 4.5 MODELO DE BIGRAMAS CON SALTO EN S	41
TABLA 4.6 FRECUENCIA DE 3 N -GRAMAS NORMALIZADOS	42
TABLA 4.7 LETRAS DE UNA ORACIÓN NORMALIZADA.....	42
TABLA 4.8 FORMATO DE PATRONES DE ENTRADA Y SALIDA	43
TABLA 5.1 ETIQUETAS HTML ELIMINADAS.....	47
TABLA 5.2 MODELO DE FRECUENCIA DE UNIGRAMA PARA 5 PALABRAS	52
TABLA 5.3 MODELO DE FRECUENCIA DE BIGRAMA CONTIGUO PARA 5 PALABRAS.....	52
TABLA 5.4 MODELO DE FRECUENCIA DE BIGRAMA CON SALTO EN 1 PARA 5 PALABRAS	52
TABLA 5.5 GRADO DE ACEPTACIÓN PARA EL FRAGMENTO DE WIKIPEDIA EN ESPAÑOL.....	57
TABLA 5.6 GRADO DE ACEPTACIÓN PARA EL FRAGMENTO DE POEMAS.....	58
TABLA 5.7 GRADO DE ACEPTACIÓN PARA EL FRAGMENTO DE LIBRO EN ESPAÑOL	59
TABLA 5.8 GRADO DE ACEPTACIÓN PARA EL FRAGMENTO DE NOTICIA EN ESPAÑOL.....	59
TABLA 5.9 GRADO DE ACEPTACIÓN PARA EL FRAGMENTO DE PUBLICACIONES DE FACEBOOK.....	60
TABLA 5.10 GRADO DE ACEPTACIÓN PARA EL FRAGMENTO DE WIKIPEDIA EN IDIOMA INGLÉS	61
TABLA 5.11 GRADO DE ACEPTACIÓN PARA EL FRAGMENTO DE WIKIPEDIA EN IDIOMA ALEMÁN	62
TABLA 5.12 SUMATORIA DE LAS SALIDAS DE DIFERENTES FRAGMENTOS DE TEXTO	63

Índice de Gráficas

GRÁFICA 1.1 FRECUENCIA NORMALIZA DE LAS PRIMERAS 100 PALABRAS DE <i>WIKIPEDIA</i> EN ESPAÑOL	8
GRÁFICA 5.1 FRECUENCIA MÁS ALTA DE LOS 100 MODELOS DE UNIGRAMA, BIGRAMA CONTIGUO Y BIGRAMA CON SALTO EN 1.....	51
GRÁFICA 5.2 SALIDA DE LA RED NEURONAL CON UNIGRAMA PARA LAS 100 PALABRAS MÁS FRECUENTAS CON 3 ÉPOCAS DIFERENTES .	54
GRÁFICA 5.3 SALIDA DE LA RED NEURONAL CON BIGRAMA CONTIGUO PARA LAS 100 PALABRAS MÁS FRECUENTAS CON 3 ÉPOCAS DIFERENTES	55
GRÁFICA 5.4 SALIDA DE LA RED NEURONAL CON BIGRAMA CON SALTO EN 1 PARA LAS 100 PALABRAS MÁS FRECUENTAS CON 3 ÉPOCAS DIFERENTES	56
GRÁFICA 5.5 SUMATORIA DE LOS DIFERENTES CONTEXTOS Y DOMINIOS	62



CAPÍTULO 1.

Introducción

El lenguaje permite expresar y comunicar los pensamientos que se tienen en nuestro entorno con alguien más (Vásquez, Quispe, & Huayna, 2009). La necesidad de comunicación de uno o más individuos para transmitir sus pensamientos genera la necesidad de tener un medio de transmisión. Para ello, los humanos han desarrollado el lenguaje natural para poder transmitir sus pensamientos.

El lenguaje natural permitió crear las primeras sociedades humanas (Bolshakov & Gelbukh, 2004). El lenguaje natural evoluciona con el paso del tiempo, para mejorar y facilitar la manera de transmitir una idea, conocimiento o una emoción (Vásquez, Quispe, & Huayna, 2009). La constante evolución que tiene el lenguaje natural le ayuda adaptarse a nuevas necesidades de comunicación de un cierto grupo de personas. Los cambios son buenos porque indica que la lengua se encuentra sana, actualizada y preparada para nuevos cambios que sufra con el paso del tiempo. Las únicas lenguas que no evolucionan son las lenguas muertas y las lenguas artificiales (de Guevara, 1980).

La evolución del lenguaje natural ha creado los diferentes idiomas que conocemos actualmente. Por ejemplo, el idioma español que usamos surgió de esta evolución, el cual sufre cambios, porque un gran número de palabras foráneas se agregaron con el tiempo. Estas palabras fueron tomadas de otro idioma y actualmente son usadas con mucha naturalidad sin conocer su origen, algunas de estas palabras son;

- *jardín y chimenea del francés.*
- *pijama del japonés.*
- *coche del húngaro*

Otros ejemplos como *coronel, plantel, vergel, capitán y fusil* tienen un origen foráneo (de Guevara, 1980).

Los cambios que tiene el español son normados por la real académica española (RAE) desde 1713. La real académica española fue creada de la necesidad de proteger el idioma español. Para dar solución a esta necesidad se tomaron las grandes obras de la época como referencia de la lengua "bien hablada" (Melgar, 1987).

La RAE publicó su primer diccionario en 1726 y una actualización con el sexto y último volumen en 1739 (Melgar, 1987). Este diccionario se conoce actualmente como diccionario de autoridades y contiene las palabras usadas en esa época, además de una definición y un ejemplo de cómo usarse en diferentes contextos. Este diccionario es indispensable para estudiar textos en español anteriores al siglo XVIII (Melgar, 1987). Uno de los problemas que enfrentó la RAE para crear el primer diccionario fueron los constantes cambios en el significado de las palabras.

El esfuerzo de la RAE para normar el español dio como resultado que cualquier hablante del español se comunique con otro sin problema. Esto es, debido a que se tienen reglas gramaticales que norman la comunicación. En este sentido, una de las ramas de la gramática es la ortografía y se define como el arte de escribir bien. La ortografía se encarga del uso correcto de signos de puntuación y acentos de las palabras (Palma Cruz, 20012).

Como resultado, el español que usamos en diferentes países es entendido por todos, lo que ayuda a compartir la información de una manera más fácil. La información compartida genera un proceso comunicativo, donde participan dos o más personas. En el proceso comunicativo participa un emisor, receptor y un mensaje. El emisor es la persona encargada de transmitir el mensaje. El receptor recibe el mensaje y trata de entenderlo. El mensaje contiene información que se intercambia entre el emisor y receptor. El resultado del proceso comunicativo depende del contexto y conocimiento que tiene cada participante, en la figura 1.1 se muestra el proceso comunicativo descrito anteriormente.

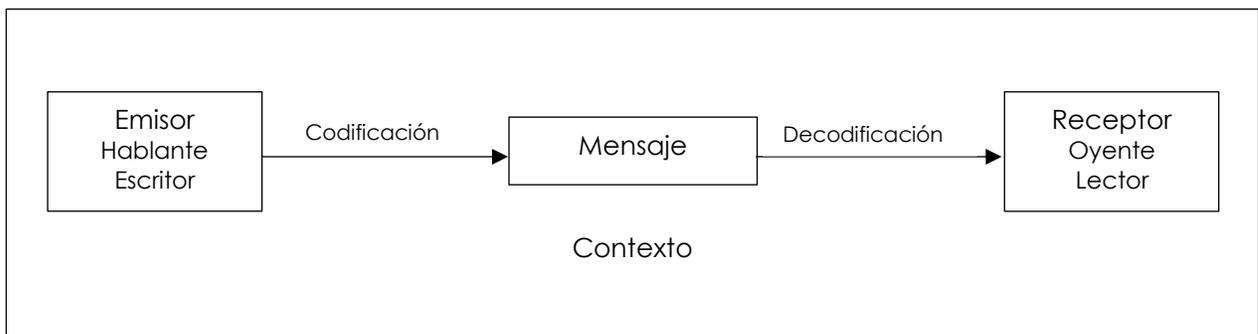


Figura 1.1 Proceso comunicativo

El contexto y conocimiento tiene gran importancia, porque si alguno de los participantes no tiene conocimiento sobre el tema, el mensaje se interpreta incorrectamente y la comunicación se ve afectada.

Con el desarrollo de las primeras redes (internet) se mejoró la comunicación y organización de la información por medio de computadoras. Las computadoras e internet cambiaron la forma de comunicación entre personas. Así, el internet conectó a muchas personas permitiendo el intercambio y acceso de información digital en diferentes idiomas y dominios. El intercambio y acceso de información digital da lugar a que se generen grandes volúmenes de información. En el año 2011, el mundo generó 1.8 zettabytes de información digital y para el año 2020 la cantidad de información será 50 veces más grande (Hernández & Gómez, 2013).

Actualmente debido a varios factores sociopolíticos, económicos y tecnológicos hacen que la lengua esté en constante cambio lo que da lugar a nuevas terminologías (Posteguillo, 2002).

Estos cambios producen que muchas palabras de otro idioma se agregan a otro con facilidad. Como se mencionó anteriormente, en internet se tienen grandes volúmenes de información digital que se produce en redes sociales y áreas específicas del conocimiento. Dentro de estos dominios se tienen anglicismos y tecnicismos que muchas personas no conocen. Debido a esto la RAE trata de normar el uso de anglicismos en el español. Para normar estos fenómenos la RAE toma como referencia los medios de comunicación para efectuar cambios y añadir normas en diferentes niveles de la lengua (Devís, 2006). Sin embargo, muchos medios de comunicación digital, por ejemplo, las bibliotecas digitales de cada especialidad cuentan con millones de artículos cada una. Cada biblioteca digital tiene diferentes temas y trata cada uno con diferentes niveles de formalidad y terminología. Esto hace imposible que la RAE genere un diccionario y reglas actualizadas para cada uno de estos subdominios.

Sin embargo, las nuevas palabras de cada área no afectan de manera negativa al español. El español siempre ha recibido de buena manera las palabras nuevas de otro idioma tratando de incorporarlas sin que tengan una influencia negativa sobre el buen uso de la lengua. Además, con el paso del tiempo muchas de estas palabras son aceptadas y otras olvidadas (Devís, 2006).

Los procesadores de texto que tiene una computadora ayudan a dar una mejor presentación y formalidad al texto que se redacta. Los procesadores de texto son programas que permiten manipular texto de manera fácil y rápida. Permite copiar, cortar y pegar el texto para dar una mejor presentación a los textos redactados sin mucho esfuerzo además incluyen correctores ortográficos y gramaticales.

Por ejemplo, el procesador de texto de *Microsoft Word* señala errores ortográficos en color rojo y gramaticales en color verde (García-Heras Muñoz, 2007). El procesador de texto de *Microsoft Word* detecta y corrige la mayoría de errores ortográficos, pero en la detección y corrección de los errores gramaticales se tiene un grado de precisión bajo. Para detectar errores gramaticales a nivel sintáctico se hace un análisis estructurado. Primero se identifica cada palabra de manera individual asignándole una etiqueta como; verbo, sujeto y adverbio.

Después se forma la estructura con las etiquetas reconocidas, en la figura 1.2 se muestra una oración de tres palabras con sus etiquetas.

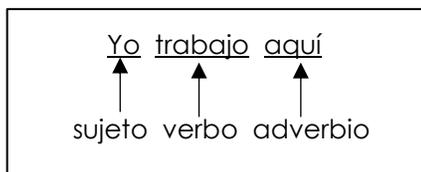


Figura 1.2 Oración etiquetada

El siguiente paso es analizar la estructura formada a un nivel sintáctico sin tomar en cuenta el significado semántico solo si la estructura sintáctica es correcta (Garcia , 2012). Como resultado se tienen oraciones con una estructura sintáctica correcta pero que pueden contener un error semántico que no se detecta, como se ve en la figura 1.3.

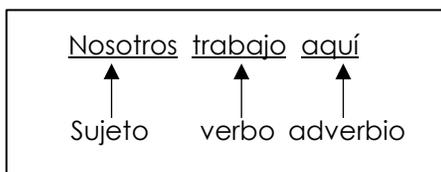


Figura 1.3 Oración sin error sintáctico pero con error semántico

El español tiene un gran número de palabras y algunas de estas palabras tienen origen foráneo y son usadas frecuentemente para comunicarnos. Las palabras que no están en un diccionario se detectan como error o simplemente no se detecta. Así, Los métodos estadísticos muestran buenos resultados para la detección de errores (San Mateo, 2016). Estos métodos estadísticos hacen uso de grandes volúmenes de información para detectar y corregir posibles errores. Para tener buenos resultados necesitan tener diccionarios con millones de palabras que representen solo una parte del lenguaje y esto genera que la información léxica que necesitan crezca exponencialmente. Más aún, estos diccionarios no garantizan que la información léxica contenga todas las palabras usadas en el español y su posible combinación con otras. Dando como resultado, que algunas combinaciones correctas se señalen como error.

En el año 2016, en el trabajo de San Mateo (San Mateo, 2016) se detectan errores en español por medio de métodos estadísticos. San Mateo (San Mateo, 2016) utiliza modelos de unigramas y bigramas para el análisis de palabras adyacentes, además como referencia para buscar la frecuencia de cada modelo usa un corpus de cien millones de palabras. De esta manera, en este trabajo se señalan los pares de palabras (bigramas) poco o muy poco frecuentes.

Como se mencionó anteriormente, el proceso comunicativo se produce cuando se intercambia información, en diferentes contextos y dominios. Sin embargo, en los diferentes contextos y dominios no se cuentan con normas claras que establezcan el uso correcto de las palabras. Esto es debido a que el grupo de personas a quien se dirige el mensaje tiene el conocimiento para entenderlo. Un ejemplo claro se encuentra en las redes sociales donde la mayoría del público entiende un mensaje como este "No se q decir asique voy a empezar con los memes un buen meme xd".

El procesador de *Microsoft Word* detecta un error ortográfico en la palabra 'xd' debido a que esta palabra no está en su diccionario. Otros correctores ortográficos y gramaticales detectan más errores como *Stilus®* que detecta 3 errores. Detecta 2 errores ortográficos en la palabra 'q' y 'asique'. Además, *Stilus®* detecta la palabra 'xd' como un error de tipografía (confusión en el uso de mayúsculas/minúsculas). Sin embargo, *SpanishChecker®* detecta 3 errores ortográficos en las palabras 'asique', 'memes' y 'xd'. Además, *SpanishChecker®* detecta en la palabra 'q' un error gramatical. Los 3 correctores ortográficos y gramaticales antes mencionados detectan errores ortográficos, gramaticales y de tipografía. Sin embargo, en el contexto de redes sociales esta forma de escribir un mensaje no se considera un error. En redes sociales se trata de transmitir una idea, pensamiento o información con el menor número de palabras. Debido a esto, los usuarios de redes sociales forman nuevas palabras para transmitir una emoción como "LOL o LMAO" que significan algo así como reírse.

En otro contexto, con un nivel de mayor formalidad la oración anterior es considerada como un error, pero esto depende del contexto donde se evalué.

En áreas específicas del conocimiento el uso de terminología técnica ayuda a nombrar diferentes modelos o métodos que son usados y conocidos por el usuario de esa área. El uso de estas terminologías diferentes para cada contexto y dominio produce que un corpus no

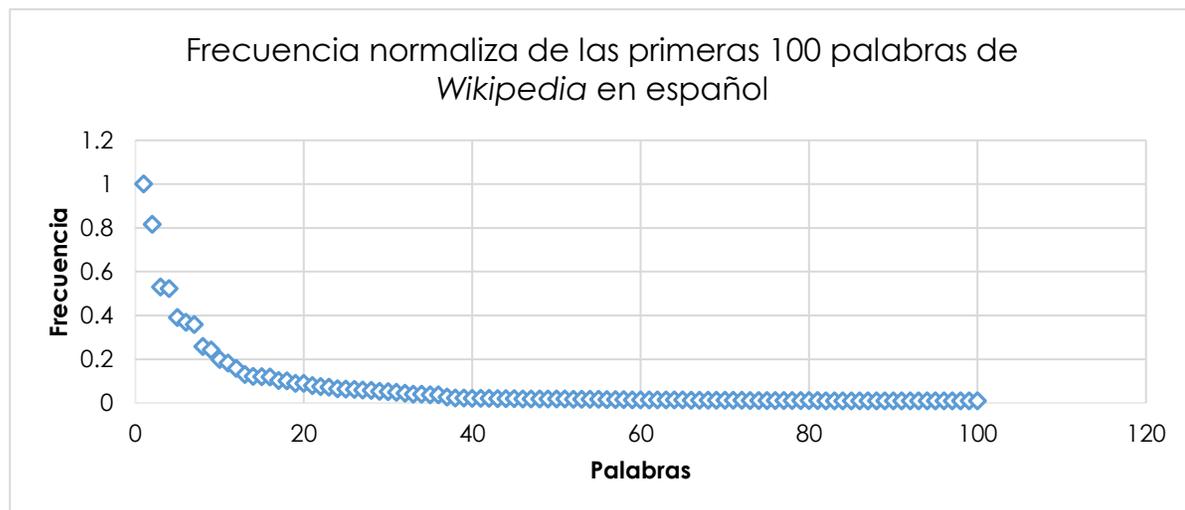
tenga la mayoría de palabras que un grupo de personas usa frecuentemente y por lo tanto no cuenta con todos los grados formalidad y dominio que usan los diferentes grupos para comunicarse.

Para analizar la distribución de la frecuencia de cada palabra en el lenguaje la ley de *Zipf* analiza la probabilidad de aparición de una palabra en un corpus (Piantadosi, 2014). En la tabla 1.1 se muestran las frecuencias que tienen las 10 primeras palabras y signos de puntuación de Wikipedia en español. El total de palabras tomadas para este ejemplo es de 65,195 palabras y las frecuencias están normalizadas.

Tabla 1.1 Frecuencia de palabras

Palabra	Frecuencia
de	1
,	0.81556228
.	0.52870425
la	0.521768
en	0.39034829
el	0.36845115
y	0.35749336
que	0.25708383
a	0.24110832
los	0.19808884

La ley de *Zipf* establece que la segunda palabra más frecuente de un corpus o idioma tiene la mitad de frecuencia que la palabra más frecuente y así con cada palabra siguiente (Piantadosi, 2014). En la gráfica 1.1 se muestran las primeras 100 palabras y signos de puntuación.



Gráfica 1.1 Frecuencia normaliza de las primeras 100 palabras de *Wikipedia* en español

1.1 Planteamiento del problema

El problema de normar la lengua con los constantes cambios que sufre en un período de tiempo muy corto hace que la RAE tenga un proceso de normar lento. El uso de correctores ortográficos y gramaticales de los procesadores de texto solo ayuda en la detección de errores sencillos. Los métodos estadísticos ofrecen buenos resultados, pero necesitan de grandes volúmenes de información para tener buenos resultados. Pero esto no ayuda en diferentes contextos y dominios con nuevas combinaciones de palabras.

Con lo anterior se plantea la siguiente pregunta de investigación.

¿Cómo aprender las combinaciones de palabras correctas que regularmente forman un texto que está escrito de acuerdo a un nivel de dominio y de formalidad?

1.2 Justificación

Los correctores ortográficos que se han desarrollado dan buenos resultados para la detección y corrección de estos errores. Sin embargo, los correctores gramaticales basados en el análisis sintáctico como Microsoft Word no dan buenos resultados en la detección de errores más complejos que tienen que ver con el significado de la palabra u oración en un contexto. Como resultado de esto se produce una mala corrección que confunden al usuario o simplemente

no detecta el error cometido (Lawley & Martin, 2006). Los métodos estadísticos para detección de errores gramaticales en español ofrecen buenos resultados para detectar una combinación de palabras poco frecuente. Si la combinación de palabras no es encontrada o tiene un número muy bajo de coincidencias se considera un error. Sin embargo, como resultado de esto un gran número de combinaciones de palabras correctas son detectadas como error, porque cada palabra tiene un número exponencial de combinaciones con diferentes frecuencias. Además, el factor que no consideran los correctores ortográficos y gramaticales es el grado de pertenencia de las palabras en diferentes contextos y dominios que hay en un lenguaje natural. Por ejemplo, para comunicarse en forma escrita en la redacción de artículos, escritura de tesis, chats o publicaciones de Facebook el nivel de formalidad y terminología cambia. Como resultado se obtienen combinaciones de palabras nuevas que son muy poco frecuentes, pero que no contiene error.

1.3 Hipótesis

Si se utiliza la frecuencia de los modelos de n -gramas continuos y no continuos de un corpus, que luego se use para entrenar una red neuronal *backpropagation* que aprenda a codificar y decodificar el aprendizaje, entonces se puede detectar el grado de aceptación del texto utilizando la salida de la red neuronal.

1.4 Alcances y limitaciones

- Solo se usará para idioma español.
- No se detectaran combinaciones poco frecuentes.
- El modelo de aprendizaje es una red neuronal *backpropagation*.
- Las características obtenidas del corpus de Wikipedia son las frecuencias de los modelos de unigrama, bigrama contiguo y bigrama con salto en s.

1.5 Objetivo general

Comprobar si el método propuesto generaliza y asocia el aprendizaje contenido en el corpus de Wikipedia.

Objetivos específicos:

- Obtener los artículos de Wikipedia.
- Eliminar las etiquetas HTML de artículos de Wikipedia para solo obtener las secuencias de texto.
- Representar las secuencias de texto en oraciones.
- Usar los modelos de n -gramas para representar las oraciones y obtener las regularidades y combinaciones de las palabras de la oración.
- Generar una muestra de entrenamiento con las oraciones e información obtenida de los modelos de n -gramas.
- Entrenar la red neuronal con los datos obtenidos del corpus de Wikipedia.

1.6 Estructura de la tesis

La estructura usada para el desarrollo de esta tesis es la siguiente:

En este capítulo se da una breve introducción a los conceptos fundamentales para entender el problema a resolver y se describen los objetivos a seguir para dar solución a este problema.

En el capítulo 2 se presentan los conceptos necesarios para resolver el problema planteado en esta tesis y describen de manera general las modelos que proporciona el procesamiento del lenguaje natural.

En el capítulo 3 se explican los trabajos desarrollados para la detección y corrección de errores gramaticales y ortográficos para el idioma español e inglés, además su funcionamiento y limitaciones que tienen con cierto tipo de errores.

En el capítulo 3 se desarrolla y explica el método propuesto para dar solución al problema planteado.

En el capítulo 5 se presenta el desarrollo del método propuesto junto con los resultados y análisis obtenidos.

En el capítulo 5 se presenta conclusiones obtenidas del método propuesto y trabajo futuro.



CAPÍTULO 2.

Marco Teórico

En el presente capítulo se explica los conceptos usados para dar solución al problema planteado. Los conceptos se describen de manera general solo tomando en cuenta una pequeña parte de cada tema.

2.1 Lenguaje natural

El lenguaje natural es un medio que permite comunicar información con diferentes personas. El lenguaje natural está formado por un conjunto finito de signos, donde cada signo individualmente no transmite un mensaje completo. Al combinar los signos se forman palabras, y al combinar las palabras se puede transmitir un mensaje completo a otra persona.

Las nuevas tecnologías ayudan a compartir la información de manera digital. Actualmente las computadoras almacenan la información digital, pero no son capaces de interpretarla. Para dotar a una computadora con esta capacidad, el procesamiento del lenguaje natural crea programas que le ayuden a realizar esta tarea.

2.2 Procesamiento del lenguaje natural (PLN).

El procesamiento del lenguaje natural es una rama de la inteligencia artificial con el objetivo de crear programas que ayuden a una máquina a tener inteligencia. Para considerar a una máquina inteligente debe poder realizar las tareas que un humano realiza cotidianamente en su hogar o trabajo. Una actividad cotidiana que realiza el ser humano es la comunicación usando un lenguaje natural. Sin embargo, debido a la complejidad que tiene el lenguaje natural es una tarea difícil de implementar.

La computadora no interpreta el lenguaje natural como un conjunto de letras que forman palabras. Los conjuntos de palabras forman una cadena de letras que a su vez forman una oración. Esta es la manera más sencilla de analizar el lenguaje natural, como un conjunto de oraciones que forman textos. Sin embargo, debido a que cada oración se relaciona con objetos y acciones del mundo, estas son completamente ignoradas por esta forma de análisis. Si bien es una forma bastante radical de analizarlo, es útil (Clark, Fox, & Lappin, 2013).

Por lo tanto, el PLN trata de dotar a las computadoras con el poder de entender oraciones que un humano escribe o habla cotidianamente (Vásquez, Quispe, & Huayna, 2009). Para dotar a una computadora con esta capacidad el PLN crea programas que puedan entender y procesar las palabras de un texto en lenguaje natural (Bolshakov & Gelbukh, 2004). Para crear los programas el PLN desarrolla modelos del lenguaje con cierto grado de formalismo (Ledeneva & Sidorov, 2010). Los programas que desarrolla el PLN ayudan, por ejemplo, en:

- *corrección gramatical*
- *desambiguación del sentido de la palabra*
- *compilación de diccionarios y corpus*
- *recuperación de información inteligente*
- *traducción automática de un idioma a otro*

2.2.1 Modelo

Un modelo es una representación mental de un objeto del mundo real, de este objeto del mundo real solo se toman las características importantes que lo componen. Sin embargo, con

estas características no hacen posible tener un modelo con una representación exacta del objeto real. Los modelos del lenguaje solo toman algunas características lingüísticas de acuerdo a la parte del lenguaje natural que trata de modelar (Gelbukh & Sidorov, 2010).

2.2.2 Lingüística

Las características del lenguaje natural son estudiadas por la lingüística. La lingüística estudia las diferentes lenguas que existen y da a conocer las reglas con las que se norman, además tiene diferentes ramas que le ayudan a este estudio (Bolshakov & Gelbukh, 2004) como lo son la: *morfología, la sintaxis y la semántica*.

2.2.2.1 Morfología

La morfología estudia las palabras que forman una lengua y las analiza a nivel de letras. La morfología estudia la estructura interna de las palabras y de cómo se relacionan las letras para poder formar una nueva palabra (Bolshakov & Gelbukh, 2004).

2.2.2.2 Sintaxis

La sintaxis es otra forma de estudiar el lenguaje natural, la cual trabaja a nivel de palabra para verificar la estructura de una oración. Por lo tanto, la sintaxis estudia la relación de cada palabra que forma una oración, porque las palabras se rigen por reglas para su uso correcto dentro de la oración (Di Tullio, 2005) (Bolshakov & Gelbukh, 2004).

2.2.2.3 Semántica

En los dos análisis anteriores solo se verifica la estructura de una palabra o de una oración, pero esto no es suficiente para entender el significado de una palabra u oración. Para entender el significado que transmite cada palabra dentro de una oración se realiza un análisis semántico. La semántica estudia el significado de las palabras dependiendo de la relación con las demás palabras dentro de la oración. Este análisis también es aplicado a nivel oración. La pragmática estudia cómo se relaciona una oración con el texto donde está escrita (Bolshakov & Gelbukh, 2004).

Como se mencionó anteriormente, los modelos del lenguaje ayudan a modelar una parte del lenguaje que puede ser morfológica, sintáctica o semántica. Para poder analizar características sintácticas del lenguaje natural se hace uso del etiquetado POS. El etiquetado

POS es un método donde se etiqueta las palabras como verbos, sustantivos, preposiciones, adverbios, adjetivos, pronombres además de palabras que no entran en las categorías antes mencionadas (Lin, Soe, & Thein, 2011). Para etiquetar una palabra en una categoría como verbo o sustantivo se usa un corpus etiquetado.

2.3 Corpus

Los corpus son colecciones de texto que representan solo una pequeña parte del lenguaje. Por lo tanto, un corpus es una gran colección de texto que representan los fenómenos que ocurren en el lenguaje. Los corpus se usan para obtener información lingüística (morfológica, sintáctica y semántica). Además de que permite estudiar la relación que tienen una palabra con otras en diferentes contextos, esto debido a la gran cantidad de información léxica que lo compone (Gelbukh & Sidorov, 2010).

2.4 Modelos de lenguaje estadísticos

Para analizar la información léxica contenida en el corpus no se analiza todo el texto como una cadena de caracteres debido a la complejidad que esto tiene. Para representar la información se usan los modelos de lenguaje estadístico. El modelo de lenguaje estadístico define la probabilidad de distribución dada una cadena de caracteres dentro de un conjunto finito de caracteres (Fink, 2014), además usa cadenas de tamaños diferentes, por lo tanto trabaja solo con pequeños trozos de la cadena de caracteres (Clark, Fox, & Lappin, 2013).

En los modelos de texto estadístico para la representación de una cadena de caracteres de tamaño n se utiliza los n -gramas. Los n -gramas establecen la probabilidad de ocurrencia de elementos como letras o palabras. La probabilidad de ocurrencia se establece a partir de los elementos predecesores conocidos, además los elementos tienen un orden cronológico (Fink, 2014). Este modelo de texto analiza la dependencia de las palabras dado un conjunto de estas, esto permite estudiar la parte estructural del lenguaje. Permite ver la diferencia de probabilidades entre estas 2 frases que tienen las mismas palabras, por ejemplo; “como estas hoy” y “estas como hoy”.

Un n -grama simple permite ver la probabilidad de solo una palabra “como”, “estas”, “hoy”, en este caso el n -grama sería de tamaño 1 unigrama. Para modelar la dependencia de 2 palabras

se usa el n -grama de tamaño 2 bigrama; esto permite ver la dependencia de 2 palabras “como estas” o “estas como”. Para modelar la dependencia de 3 palabras se usa el n -grama de tamaño 3 (trigrama), esto permite diferenciar “como estas hoy” de “estas como hoy” (Liu & Özsu, 2009).

Por su parte, San Mateo (San Mateo, 2016) utiliza modelo de unigrama y bigrama para la representar la siguiente cadena de texto de longitud 13: “y su tiene intención de visitar la tumba del que fue su amigo”. La representación de esta cadena se muestra en la tabla 2.1 con unigrama y bigrama.

Tabla 2.1 Modelo de unigrama y bigrama

Unigrama	Bigrama
y	y su
su	su tiene
tiene	tiene intención
intención	intención de
de	de visitar
visitar	visitar la
la	la tumba
tumba	tumba del
del	del que
que	que fue
fue	fue su
su	su amigo
amigo	

Con el modelo de bigrama usado por San Mateo (San Mateo, 2016) se modela la relación que tienen dos palabras de la cadena, pero esta relación es solo con la palabra siguiente. Debido a esto, el modelo de trigrama mostraría la misma relación que un bigrama, pero con tres palabras.

Para modelar la relación que hay con “y” con “tiene” o “y” con “intención” del ejemplo anterior se puede hacer uso del bigrama con salto en s. El salto está definido por tamaño de

n palabras secuenciales, lo que permite ver la relación que tiene la primera palabra con la tercera, cuarta e incluso quinta palabra de una cadena de texto. La oración anterior se representa en la tabla 2.2 con el modelo de bigramas con salto en s .

Tabla 2.2 Modelo de bigrama salto n

Bigrama salto en $n=1$	Bigrama salto en $n=2$
y tiene	y intención
su intención	su de
tiene de	tiene visitar
intención visitar	intención la
de la	de tumba
visitar tumba	visitar del
la del	la que
tumba que	tumba fue
del fue	del su
que su	que amigo
fue amigo	

Para que una computadora pueda hacer uso de la información contenida en los n -gramas el PLN puede hacer uso modelos de aprendizaje automático.

2.5 Aprendizaje

Los seres humanos y algunas especies de animales tienen la capacidad de resolver problemas mediante la adquisición de información (Matich, 2001). La información que adquiere con el paso del tiempo y logra adaptar además de mejorar para dar solución a diferentes problemas se conoce como aprendizaje. El aprendizaje ayuda a reconocer y comprender problemas muy complejos, debido a esto se busca la manera más eficiente de resolverlos. (Kasabov, 1996). En un problema que requiere mucha fuerza física desarrollo maquinas o instrumentos que ayuden a realizarlo.

Con el paso del tiempo las máquinas evolucionaron para mejorar la manera en que realizaban un proceso. Las primeras máquinas solo se diseñaron para resolver ciertos cálculos y cuando

se requería cambiar el tipo de cálculo se tenía que modificar o cambiar toda la estructura de la máquina (hardware). Con el tiempo las máquinas superaron la limitante de no poder reprogramarse y actualmente tienen la capacidad para reprogramarse, adaptarse y aprender (Cruz, 2011).

Una de las máquinas que actualmente tiene la capacidad de reprogramarse es la computadora. Con la implementación de las computadoras se pudo resolver una gran cantidad de problemas que eran repetitivos y largos de hacer para un humano.

La implementación de algoritmos ayudó a esta tarea, pero cuando los problemas que se tratan de resolver con un algoritmo no tenían una secuencia lineal y lógica que seguir éstos no ayudaban. Como consecuencia, surgió la necesidad de dotar a las computadoras con métodos y modelos que ayudarán a resolver estos problemas.

2.6 Aprendizaje automático

Como se mencionó anteriormente, la inteligencia artificial investiga como dotar a una computadora con la capacidad entender el lenguaje natural. Además de esto, la inteligencia artificial crea métodos y modelos que tratan de imitar la forma en que aprendemos. Los modelos de aprendizaje automático tratan de resolver diferentes problemas con el aprendizaje y experiencia con el que son entrenados (Kasabov, 1996).

El aprendizaje automático lo definen estos autores de la siguiente manera:

- *Descubrimiento totalmente autónomo o manera automática de las regularidades y relaciones que existen en un conjunto de datos, sin la necesidad de tener un gran diccionario y reglas que ayuden a este proceso. (Gelbukh, 2010)*
- *Mecanismo que le permiten a una computadora tener la capacidad de aprender a partir de experiencias, aprender de ejemplos o aprender a partir de analogías, las 2 principales ramas de este enfoque son las redes neuronales y algoritmos genéticos. (Michael, 2005)*

Las ramas que menciona Michael (Michael, 2005) de redes neuronales y algoritmos genéticos están inspirados en modelos biológicos.

2.6.1 Redes neuronales

Las redes neuronales no son sino otro intento de emular la capacidad que tienen los humanos de aprender. Las redes neuronales están basadas en el cerebro humano a un nivel muy básico tratando emular las conexiones e intercambio de información que tienen lugar en cada neurona del cerebro.

Las redes neuronales se iniciaron en 1943, en un principio solo eran aproximaciones a funciones booleanas. En los inicios de los años 70 se tenían expectativas muy altas de lo que una red neuronal podía lograr, pero esto no fue posible debido a que el poder computacional era muy limitado dejándolas abandonadas.

En la actualidad algunas aplicaciones utilizan una red neuronal, pero aún se tiene una limitante que es el poder computacional que se tiene actualmente debido a que un cerebro humano tiene entre 10^{10} y 10^{11} neuronas (Cruz, 2011).

2.6.1.1 Neurona

Como se mencionó, las redes neuronales artificiales tratan de imitar la manera en que el cerebro aprende. El aprendizaje en el cerebro comienza con las células situadas en la corteza cerebral que tienen el nombre de neuronas.

La neurona es una célula que se encuentra en el cerebro de la mayoría de los seres vivos y su función principal es tener una entrada un procesamiento y una salida (Russell & Norvig, 2004). Por lo tanto, la neurona es una célula especializada en procesar información. Para realizar el procesamiento de información tiene tres elementos básicos. La neurona tiene un cuerpo (soma) y dos tipos de ramificaciones. Una de las ramificaciones tiene el nombre de dendritas y su función es recibir información. La otra ramificación tiene el nombre de axón, y su función es comunicarse con otras neuronas por medio de sus dendritas (Cruz, 2011). A esta conexión de axón y dendritas se le llama sinapsis.

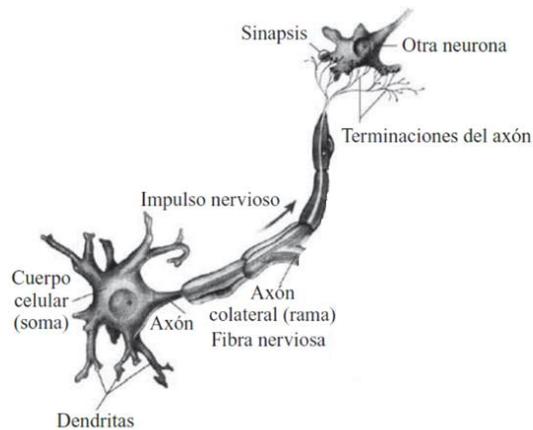


Figura 2.1 Partes de la Neurona (Cruz, 2011).

En la sinapsis se intercambian señales electroquímicas entre cada neurona conectada para reaccionar a un estímulo externo (Michael, 2005). El intercambio de señales electroquímicas se realiza cuando las dendritas reciben una entrada por medio del axón de otra neurona. Esta entrada es procesada en el soma para generar un nivel de excitación. El nivel de excitación está definido por un umbral que determina si la entrada puede generar una excitación en cada conexión de la neurona. En el caso contrario donde no se genere una excitación no se intercambiarán señales electroquímicas, por lo tanto, no se genere una excitación o salida en ninguna neurona. Debido a esto las neuronas forman redes que le permiten comunicarse con otras neuronas para generar una salida como un conjunto a un estímulo externo que reciben de entrada. Como resultado podemos formar nuevas conexiones entre cada neurona. Las conexiones neuronales en el cerebro humano son aproximadamente 10^{15} conexiones (Cruz, 2011).

Esta son las principales características de una neurona que le permiten comunicarse para formar complicadas conexiones entre ellas (Cruz, 2011). Estas conexiones entre cada neurona y el procesamiento de información que reciben como entrada es lo que nos permite aprender nueva información.

Para el desarrollarlo de una red neurona artificial se tomaron las siguientes características de una red neurona biológica. La tabla 1.1 muestra la analogía de una neurona biológica y artificial.

Tabla 2.3 Analogía de una neurona

Neurona biológica	Neurona artificial
Soma	Neurona
Dendrita	Entrada
Axón	Salida
Sinapsis	Pesos

2.6.1.2 Perceptrón

La primera red neuronal tiene el nombre de perceptrón. El perceptrón contiene una neurona que recibe entradas de 0 a 1 y una salida de 0 a 1 (Michael, 2005). La figura 2.2 representa las partes del perceptrón.

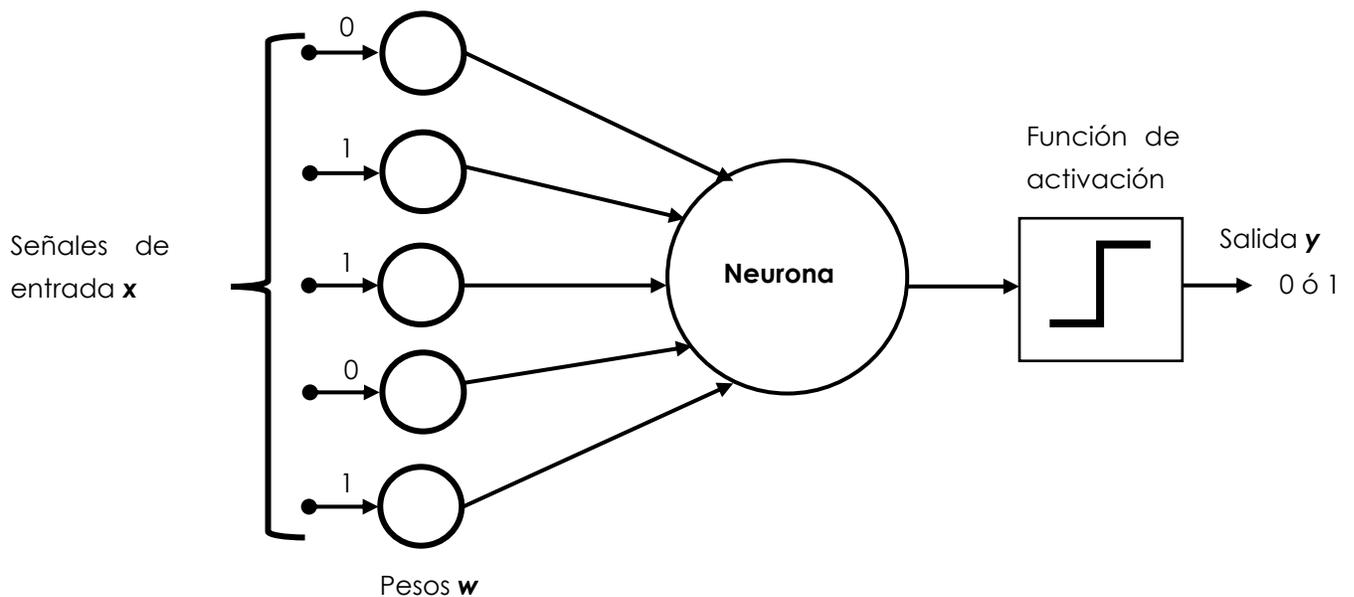


Figura 2.2 Diagrama del perceptrón

El perceptrón tiene N número de pesos conectados a una neurona con una función de activación que genera una salida directamente. Cada salida es independiente y no es

afectada por otra salida (Russell & Norvig, 2004). La función de activación trata de imitar el proceso que una neurona biológica cuando recibe una o varias entradas. Las neuronas biológicas tienen 2 estados de activación uno es el estado inactivo (no excitado) y el otro estado es activo (excitado). Las neuronas artificiales tienen estos 2 estados. La función de activación mide su estado de actividad, la cual depende de la entrada global que tenga, los valores que puede tomar son de $\{0,1\}$ o de $\{-1,1\}$, en el cual los valores de $\{0,-1\}$ son inactivos y los valores de $\{1,1\}$ son activos (Matich, 2001).

2.6.1.3 Función de activación perceptrón

La función de activación del perceptrón es una operación muy sencilla que ocurre cuando una neurona recibe los pesos de entrada. La entrada de una neurona solo acepta un valor como entrada y para poder realizar esta operación hace una entrada global donde $x_i = (x_{i1} * w_{i1}) + (x_{i2} * w_{i3}) + \dots (x_{in} * w_{in})$. La operación anterior tiene el nombre de sumatoria del perceptrón.

Donde x_i es la salida que tiene la neurona, y cada x_{in} representa el patrón de entrada que recibe. El valor de entrada de x_{in} es multiplicado por el peso de w_{in} correspondiente para entrada. Los valores que se le asignan a cada w_{in} son de -1 a 1 generalmente los valores de x_{in} son aleatorios.

Con la entrada global de x_i , la neurona la puede procesarla con la función de activación para definir si la salida tiene valor 1 o 0. Existen diferentes tipos de funciones de activación las cuales toman valores de salidas diferentes. (Michael, 2005)

La función de activación del perceptrón es la función paso, que recibe el valor de x_i para poder asignar un valor de salida a y .

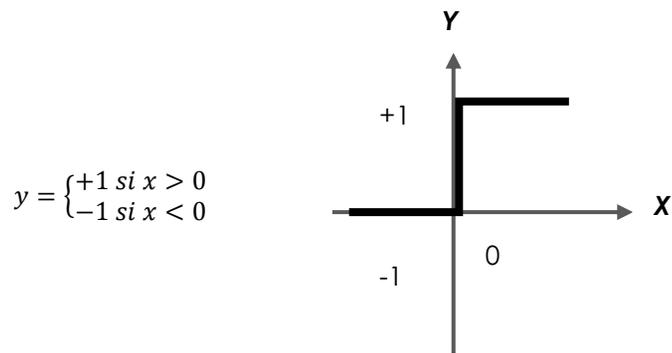


Figura 2.3 Función paso

Las características antes mencionadas del perceptrón hacen posible que tenga la capacidad de aprender. El perceptrón tiene la capacidad de aprender las combinaciones de compuertas AND y OR ya que contiene valores de entrada y salida que son linealmente separables en una clasificación (Michael, 2005). En la tabla 2.4 y figura 2.4 se muestra lo mencionado anteriormente.

Tabla 2.4 Tabla de verdad

Entrada	Entrada	And	Or	XOR
x_1	x_2	$x_1 \cap x_2$	$x_1 \cup x_2$	$x_1 \oplus x_2$
0	0	0	0	0
0	1	0	1	1
1	0	0	1	1
1	1	1	1	0

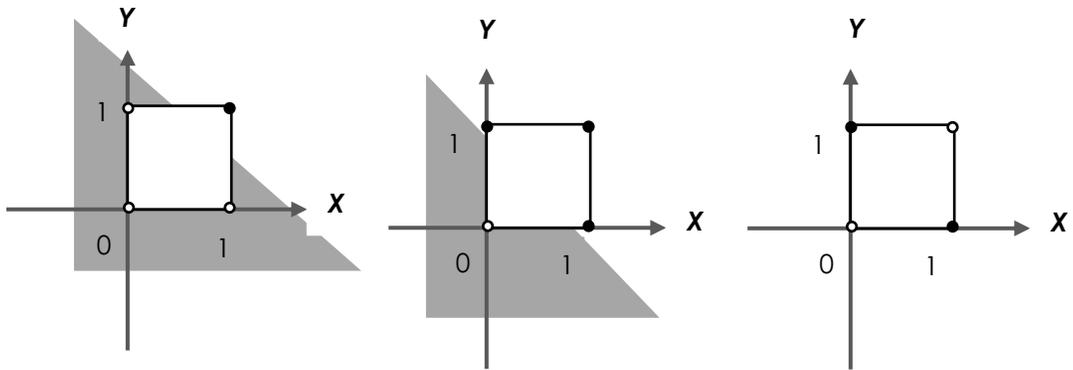


Figura 2.4 Representación gráfica de la tabla de verdad.

El proceso mencionado anteriormente lo describe el autor Cruz de la siguiente manera:

“El algoritmo de aprendizaje del perceptrón funciona para aprender funciones binarias linealmente separables, ya que de otra manera el algoritmo no convergería ni produciría la mejor solución. Debido a que las salidas son binarias, se emplean unidades lineales de umbral. Cada unidad calcula la suma con pesos de las N entradas x_j , $j = 1 \dots N$, y genera una salida binaria (Cruz, 2011).”

Como se muestra en la tabla 2.2 el perceptrón puede aprender los valores de salida de la compuerta AND y OR porque son linealmente separables por una recta. Sin embargo, el perceptrón no es capaz de aprender la salida de la compuerta XOR porque sus valores no son linealmente separables por una recta. Para que el perceptrón pueda aprender la compuerta XOR tiene que tener 2 rectas que separen la salida como se muestra en la figura 2.5.

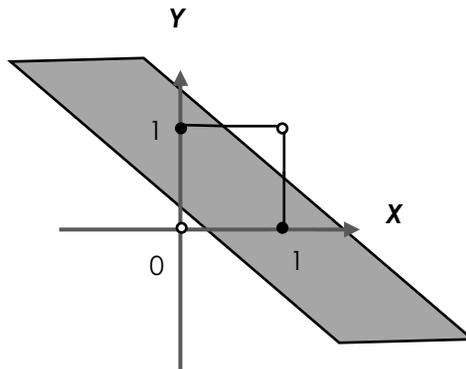


Figura 2.5 Representación XOR

2.6.1.4 *Redes neuronales multicapa*

El perceptrón no tiene la capacidad de aprender problemas muy complejos donde los datos son muy cercanos. Los problemas requieren de funciones más complejas para poder representarlos en diferentes clases. Para poder representar una función más compleja se tienen las redes neuronales multicapa. Las redes neuronales multicapa están formadas por un conjunto de perceptrones.

Los perceptrones están distribuidos en diferentes capas y tiene el nombre de capa de entrada, capa oculta y capa de salida. Cada capa cuenta con un número definido de perceptrones (Michael, 2005).

Capa de entra: contiene las neuronas que reciben los patrones de entrada por lo general no tienen ninguna función o proceso que realicen ya que solo reciben los datos.

Capa oculta: en esta capa es donde los patrones de entrada son procesados y se realizan los procesos para calcular una entrada en la neurona y una salida con la entrada procesada. La capa oculta es la encargada de la representación de las funciones.

El autor Michael lo define de la siguiente manera:

“Con una capa oculta, podemos representar cualquier función continua de las señales de entrada, y con dos capas ocultas incluso se pueden representar funciones discontinuas (Michael, 2005).”

Capa de Salida: aceptan las funciones de activación de la capa oculta, las procesan para dar una salida. La salida de la red neuronal tiene el nombre de patrón de salida.

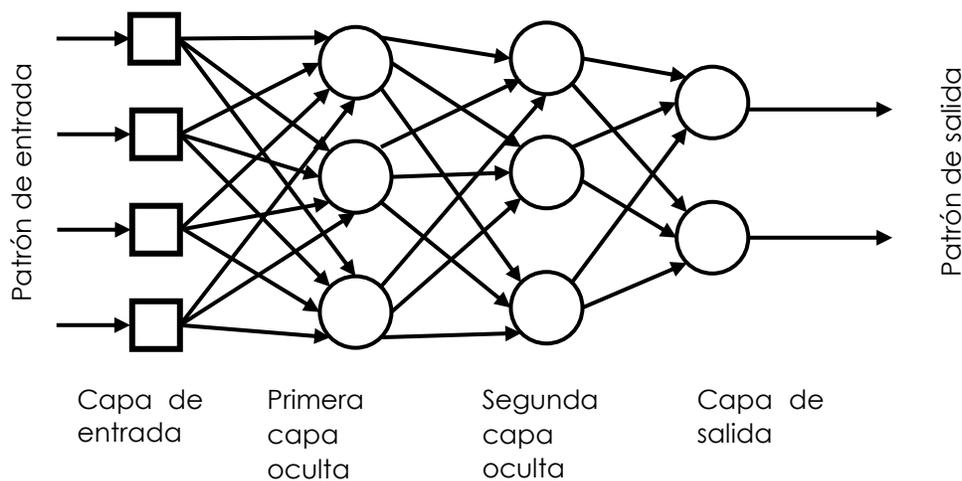


Figura 2.6 Perceptrón multicapa

2.6.1.4 Entrenamiento de la red neuronal multicapa

El entrenamiento de una red neuronal artificial es similar al proceso de aprendizaje que tenemos en nuestro cerebro. Para aprender nueva información en nuestro cerebro se realizan miles de conexiones y desconexiones entre las neuronas. El proceso de conexión y desconexión también es usado en las redes neuronales artificiales. Para hacer este proceso cada neurona que pertenece a una capa sigue el siguiente proceso:

- Recibe un conjunto de patrones con valores de 0 y 1 que recibe en la capa de entrada.

- Los patrones de entrada son recibidos de la capa de entrada para ser procesados en la capa oculta.
- Los patrones procesados en la capa oculta son recibidos por una capa de salida que tiene como salida un 0 o 1.

Este proceso se repite para cada conjunto de patrones que se presente. Cuando a una neurona se le asigna el valor de 0, la neurona queda desconectada de la red. En cada época (iteración) los pesos de cada conexión cambian su valor para poder dar la salida deseada (*adaptación de los pesos*). Cuando la capa de salida puede asignar el patrón de salida deseado para cada conjunto de patrones que se le presenta, se concluye que la red neuronal aprendió correctamente la muestra de entrenamiento (Matich, 2001).

2.6.1.4 *Backpropagation*

La red neuronal *backpropagation* (BP) pertenece a las redes neuronales multicapa. Es una red neuronal no lineal su funcionamiento más simple y funcional es el perceptrón. La combinación de perceptrones en cada capa hace posible que supere la función lineal del perceptrón. Tiene la capacidad de aprender problemas de clasificación muy complejos.

La red neuronal BP minimiza el error que tiene para aprender. Tiene una función de activación diferente al perceptrón llamada función sigmoidea en la figura 2.7 se muestra esta función (Hernández & Gómez, 2013).

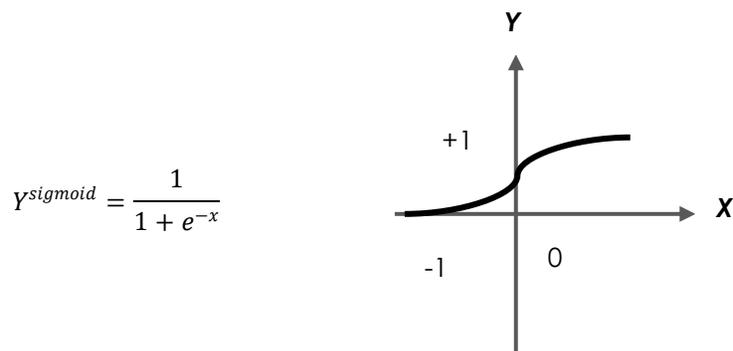


Figura 2.7 Función sigmoidea

Los valores de salida que proporciona esta función están comprendidos dentro de un rango que va de 0 a 1 (Matich, 2001).

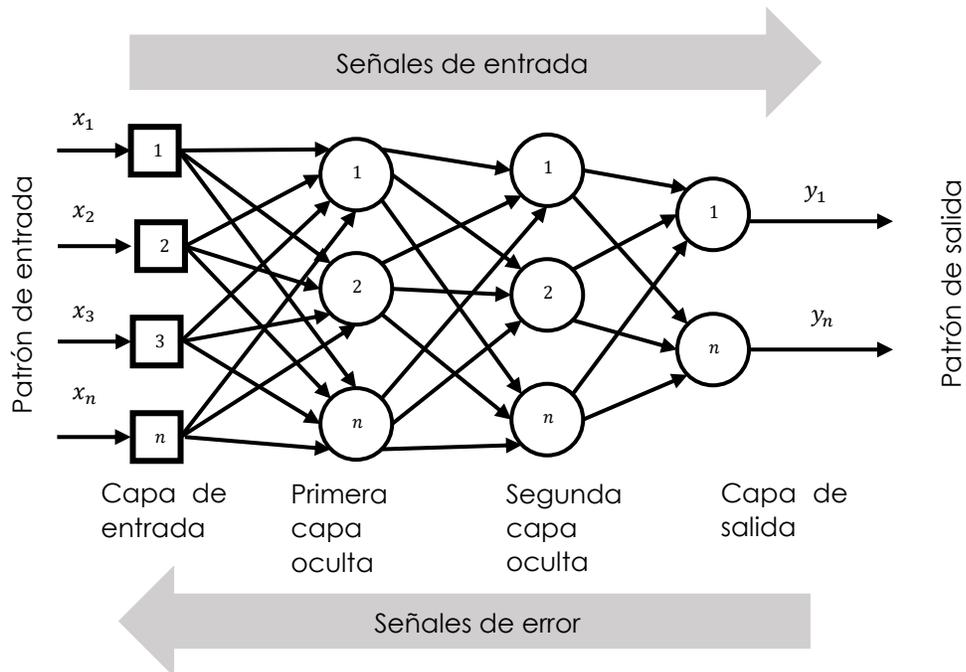


Figura 2.8 Red neuronal *backpropagation*

En la figura 2.8 se muestra la estructura de la red neuronal BP. Se muestran todas las neuronas conectadas de cada capa. Cuenta con una capa de entrada, dos capas ocultas y una de salida.

El proceso de aprendizaje de la red neuronal BP comienza cuando recibe un patrón de entrada en la capa de entrada. El patrón genera una señal de entrada se propaga neurona por neurona y avanza a través de la red. El avance que tiene en cada capa genera una señal de salida en la última capa de la red neuronal. Con esta salida se calcula el error que tiene la red. Este error es propagado hacia atrás iniciando en la capa de salida y se utiliza para ajustar los pesos de cada conexión neuronal. Esto permite que los pesos de cada conexión ubicadas en las capas ocultas cambien durante el entrenamiento. Como resultado de esto, la red neuronal BP asocia la información de entrada y salida.

En conclusión, las redes neuronales son memorias que almacenan información. Esta información se guarda en los pesos asociados a las conexiones entre neuronas. Por lo tanto, la red neuronal es una memoria que graba información en forma estable (Matich, 2001).

2.8 Resumen

En el presente capítulo se describe brevemente los conceptos de lenguaje natural, procesamiento del lenguaje natural y aprendizaje automático. El lenguaje natural lo usamos los humanos para comunicarnos e intercambiar información y el procesamiento del lenguaje natural crea métodos que ayudan al humano a comunicarse con una computadora. El procesamiento del lenguaje natural crea los modelos del lenguaje que ayudan a esta representar una pequeña parte del lenguaje natural. Sin embargo, estos modelos no son representaciones completas porque solo toman una o varias características importantes que componen al lenguaje natural. Las características del lenguaje natural son estudiadas por la lingüística. Para estudiar el lenguaje natural la lingüística tiene varias ramas, cada rama está enfocada en una parte del lenguaje natural. Además, se describen los modelos de texto estadístico y un modelo de aprendizaje automático. El modelo de aprendizaje automático descrito es una red neuronal *backpropagation*.



CAPÍTULO 3.

Estado del Arte

En este capítulo se presentan los trabajos más relacionados al problema de determinar el grado de pertenencia de cadenas de texto a un lenguaje por lo que se aborda normalmente como detección de errores ortográficos o gramaticales.

3.1 Trabajos que evalúan la detección de errores gramaticales

Uno de los procesadores de texto más populares es *Microsoft Word*. Este procesador de texto ayuda en la detección de errores ortográficos y gramaticales. Como resultado, diversos trabajos evalúan el rendimiento en detección y corrección de errores ortográficos y gramaticales. En las primeras versiones los procesadores de texto solo detectaban y corregían errores ortográficos. Los procesadores de *Microsoft Word* y *WordPerfect* en su versión 6.0 fueron analizados por Gracia, Tapia Poyota y Ana (García, Tapia Poyato, & Ana, 1997). En estos procesadores de texto para la detección de errores ortográficos contaban con 2 bases de datos. La primera base de datos es un diccionario abreviado de la Real Academia Española. La segunda base de datos es un diccionario creado por el usuario con los términos que más usaba y que no contenía el primer diccionario. La función principal de estos 2 diccionarios solo es reconocer las palabras de un texto una a una, si una palabra no se encontraba se sugiere una parecida como corrección. En una versión posterior de *Microsoft Word*, García-Heras Muñoz (García-Heras Muñoz, 2007) evalúa la detección que tiene en errores ortográficos y

gramaticales. Como resultado se obtienen buenos resultados en detección y corrección de errores ortográficos, pero no detecta la mayoría de errores gramaticales.

3.2 Trabajos que detectan errores gramaticales en el idioma inglés

Debido a esto, se crearon diversos métodos para mejorar la detección de errores gramaticales. Lawley y Martin (Lawley & Martin, 2006) para la detección de errores gramaticales hacen uso de Microsoft Word y un corpus de errores. En este trabajo, primero se usa a Microsoft Word para detectar errores ortográficos. Después usa el corpus que contiene frases incorrectas y palabras problemáticas para señalar posibles errores. Lawley y Martin (Lawley & Martin, 2006) tienen un 60% de efectividad al buscar errores. Para mejorar la efectividad de detección propone incorporar más frases incorrectas y palabras problemáticas además de implementar expresiones regulares.

Por su parte, More (Moré, 2006) propone como corrector gramatical la frecuencia de un buscador web. El método busca partes de las oraciones escritas por el usuario en internet, para determinar si son frecuentes o poco frecuente. Con esto ayuda a detectar errores en partes de la oración que son poco frecuentes. Como problema tiene que muchas frases de la oración que contienen errores son encontradas frecuentemente en internet.

También se tiene el método de Sjöbergh (Sjöbergh, 2006) que busca la frecuencia de frases en un corpus de referencia. Para analizar cada oración hace un *chunker* en la oración. Con cada fragmento de la oración obtenido del *chunker* es buscado en el corpus de referencia. Con esta búsqueda se obtiene la frecuencia de cada fragmento de la oración. Los fragmentos de oración son más grandes que un trigramo. La frecuencia debe ser muy baja o nula para determinar si el fragmento analizado contiene un error. Para detectar diferentes tipos de errores se cambia el tamaño del *chunker*. Se compara con *Microsoft Word 2000* y obtiene los siguientes resultados; *Microsoft Word* detecta 14 errores y 3 falsas alarmas: por parte de su método se detectan 31 errores y 13 falsas alarmas.

Los trabajos mencionados anteriormente están enfocados en la detección de errores gramaticales para el idioma inglés.

3.3 Trabajos que detectan errores gramaticales en español

Existen diferentes enfoques usados para la detección y corrección de errores gramaticales para el idioma español. Algunos de estos enfoques se usan para detectar malapropismos entre dos palabras (Bolshakov, Galicia-Haro, & Gelbukh, 2005). Para detectar un malapropismo primero se analiza su estructura sintáctica y después se verifica a nivel semántico. El malapropismo se encuentra a nivel semántico como lo explican con el siguiente ejemplo; *mañana sopeada* para *mañana soleada*. La estructura sintáctica es correcta, pero semánticamente el significado es incorrecto. Para determinar si un bigrama es un malapropismo proponen como posible solución hacer uso de un corpus grande y del índice de compatibilidad semántica. El índice de compatibilidad semántica verifica el bigrama a nivel sintáctico y semántico. El índice de compatibilidad semántica se calcula a partir de las estadísticas obtenidas del buscador de *Google*.

Como corpus de referencia utiliza el motor de búsqueda de *Google*. Con el motor de búsqueda de *Google* se busca el bigrama y el bigrama con salto en s. Se usa el bigrama con salto en s debido a que un malapropismo afecta a una colocación. Una colocación es un conjunto de dos palabras que forman una frase y donde esta frase se esperaría encontrarse menos de lo habitual, pero se encuentran más de lo esperado en la lengua (Bolshakov, Galicia-Haro, & Gelbukh, 2005). Como resultado obtienen una detección y corrección de 125 malapropismos propuestos por ellos.

En el trabajo de Nazar y Renau (Nazar & Renau, 2012) utilizan la frecuencia n -gramas para la detección y corrección de errores gramaticales. Obtiene la frecuencia de n -gramas de tamaño n . Los n -gramas se construyeron a partir del corpus de *Google Books*. La idea de Nazar y Renau (Nazar & Renau, 2012) es utilizar el corpus de *Google Books* como verificador gramatical. Para hacer esto, primero se analiza un texto de entrada y se detecta los n -gramas que son poco frecuentes. La hipótesis de este trabajo es obtener las reglas gramaticales contenidas en la información de los n -gramas.

Para la detección de errores gramaticales usa el modelo de bigrama. Además, utiliza el modelo de trigrama para completar frases de acuerdo al contexto de las 2 palabras que conoce. El método propuesto por Nazar y Renau (Nazar & Renau, 2012) es comparado con el

corrector ortográfico y gramatical de *Microsoft Word*. Como resultado se obtiene que su método detecta más errores que el corrector de *Microsoft Word*. Sin embargo, en el proceso de detección el método señala bigramas correctos. En cambio, el corrector de *Microsoft Word* detecta menos errores, pero deja muchos errores sin ser detectados. Para solucionar este problema se propone el uso de trigramas en la etapa de detección o un modelo más grande de n -gramas. Además, propone hacer uso del etiquetado *POS* y también detectar nombres propios. Como conclusión menciona que los algoritmos estadísticos podrían ser un complemento útil para un sistema basado en reglas.

El método de San Mateo (San Mateo, 2016) estudia el uso de n -gramas de tamaño 1 y 2 para la detección de errores ortográficos y gramaticales en español. Solo trabaja con unigramas y bigramas aplicando métodos estadísticos como lo hace Nazar y Renau (Nazar & Renau, 2012). Los errores los detecta analizando la frecuencia de cada unigrama y bigrama. La frecuencia la obtiene de un corpus de 100 millones de palabras; el corpus está formado por textos en español, no incluye textos técnicos. La mayor parte de los textos que forman el corpus son;

- *textos escritos por nativos*
- *textos narrativos (cuentos, novelas y ensayos)*
- *noticias y artículos del año 2012*

Trataron de formar un corpus en español que represente el uso actual de la lengua para obtener mejores resultados.

El funcionamiento del algoritmo de detección toma la probabilidad de combinación de palabras y la frecuencia de cada bigrama para calcular el umbral. El umbral determina si el bigrama tiene un error, si el umbral tiene un valor mayor a 1 el bigrama no contiene error de acuerdo a la información obtenida del corpus. El umbral con un valor menor a 1 ó 0 indica que el bigrama es poco frecuente o no se encuentra en el corpus, además de aportar información de que el bigrama no es usado comúnmente por los nativos de la lengua.

Con los trabajos de San Mateo, Nazar y Renau (San Mateo, 2016) (Nazar & Renau, 2012) se concluye que la detección de errores ortográficos y gramaticales implica más que solo

obtener la frecuencia de las palabras con el modelo de bigramas. Esto se produce debido a que cada palabra tiene un número exponencial de combinaciones. Como resultado se obtiene que la combinación de ciertas palabras en un contexto da como resultado un error, pero en otro contexto no lo tenga. Además, otra limitante es el tamaño del corpus como lo describe Nazar y Renau" a pesar de que el corpus utilizado es probablemente el corpus más extenso jamás compilado, hay bigramas que no están presentes en él". En la tabla 3.1 se muestra los 3 trabajos que usan n -gramas y análisis estadístico para el idioma español.

Tabla 3.1 Trabajos que usan modelos de n -gramas y análisis estadístico

Trabajo	Características	Corpus	Modelo
(Bolshakov, Galicia-Haro, & Gelbukh, 2005)	Detección de malapropismos en español	Buscador de Google	n -gramas
(Nazar & Renau, 2012)	Detección de y corrección de errores gramaticales en español	Google Books Corpus	n -gramas
(San Mateo, 2016)	Detección de errores gramaticales en español	corpus español de 100 millones de palabras	n -gramas

3.4 Resumen

En este capítulo se describen los trabajos relacionados con la detección de errores ortográficos y gramaticales. Los primeros trabajos evalúan el funcionamiento de los procesadores de texto en la detección y corrección de errores ortográficos y gramaticales. Como resultado de la evaluación se concluye que solo detectan una pequeña parte de errores gramaticales, pero tienen buenos resultados en la detección y corrección de errores ortográficos. En la segunda parte se describen el uso de métodos estadísticos para detectar errores gramaticales en inglés y obtienen resultados aceptables. Los métodos desarrollos para inglés usan como corpus de referencia pequeños corpus y la web para buscar la secuencia analizada. Para el idioma español se describen tres trabajos que detectan y dos corrigen los errores detectados.



Capítulo 4.

Método propuesto

En el presente capítulo se describe el método propuesto para aprender la combinación de palabras correctas que regularmente aparecen en el contexto y dominio de un texto. Se describe mediante un diagrama que muestra los pasos que tiene el método propuesto. Los pasos que se describen son para resolver el problema mencionado anteriormente, ¿Cómo aprender las combinaciones de palabras correctas que regularmente forman un texto que está escrito de acuerdo a un nivel de dominio y de formalidad?

4.1 Descripción del método propuesto

Como se explicó, la frecuencia de modelos de n -gramas obtenidos de un corpus de referencia es útil para detectar errores gramaticales. El modelo basado en unigramas se usa para verificar si una palabra existe en el corpus. Como resultado de la verificación de cada palabra se detectan errores ortográficos. El modelo basado en bigramas se usa para verificar si una combinación de 2 palabras existe dentro del corpus. Con el bigrama se verifica la estructura sintáctica de 2 palabras y como resultado se detectan algunos errores sintácticos. Para mejorar la detección de errores gramaticales San Mateo, Nazar y Renau (San Mateo, 2016), (Nazar & Renau, 2012) proponen hacer uso de modelos de n -gramas mayores a dos, pero como lo menciona Nazar y Renau (Nazar & Renau, 2012) la frecuencia obtenida en n -gramas

mayores a dos sería menor. Esto es debido a que cada combinación de palabras mayor a 3, 4 o 5 tiene una frecuencia más baja. Como ejemplo de esto, en la tabla 4.1 se muestra la búsqueda realizada en el motor de búsqueda de *Google*. La búsqueda en *Google* se hace para obtener la frecuencia de diferentes modelos *n*-gramas.

Tabla 4.1 Modelos de *n*-gramas

Modelo	Palabras	Frecuencia
Unigrama	La	19,280,000,000
Bigrama	La casa	534,000,000
Trigrama	La casa es	31,600,000
Cuatrigrama	La casa es roja	58,800

Como se muestra en la tabla 4.1 la frecuencia del modelo de unigrama es la más alta, pero para el modelo de bigrama o cuatrigrama la frecuencia es menor. Para un *n*-grama de tamaño 6 como el siguiente "La casa es roja y grande" la frecuencia que obtiene del motor de búsqueda de *Google* es de 9. Sin embargo, con el *n*-grama de tamaño 6 no se puede analizar la relación que tiene la primera palabra con la última palabra.

Para analizar la relación que tiene "La" con "roja" o "La" con "grande" de la oración de tamaño 6 se propone el uso de bigramas con salto en *s* como Bolshakov, Galicia-Haro, y Gelbukh (Bolshakov, Galicia-Haro, & Gelbukh, 2005).

Para obtener la frecuencia de cada modelo de unigrama, bigrama y bigrama con salto en *n* se usa un corpus de artículos de *Wikipedia* en español. Otro problema que se tiene es que alguna combinación no se encuentra dentro del corpus como lo menciona Nazar y Renau (Nazar & Renau, 2012). Como solución a este problema, se propone el uso de la red neuronal *backpropagation* para que aprenda el modelo de texto que se propone en esta tesis y asigne una salida a las nuevas combinaciones que se le presentan.

4.2 Etapas del método propuesto

El primer paso es obtener un corpus que contenga información léxica de un dominio y contexto. Del corpus se eliminarán caracteres y secuencias de texto que no aporten información relevante, para solo obtener secuencias de texto que aporten información del contexto y dominio al que pertenecen. El segundo paso es dividir las secuencias de texto en "oraciones" y obtener una nueva representación del texto. El tercer paso es separar las palabras y signos que contiene cada oración, esto se hace agregando un espacio entre cada palabra, signo y número de la oración.

El siguiente paso es representar las oraciones con los diferentes modelos de n -gramas y obtener la frecuencia cada modelo de n -grama del corpus de referencia. Con la frecuencia obtenida de cada unigrama, bigrama contiguos y bigrama son salto en s se crea el modelo de frecuencia de n -gramas usando un grafo. El grafo contiene los diferentes modelos n -gramas y la frecuencia de cada modelo de n -gramas. Los grafos son guardados y procesados para el entrenamiento de la red neuronal *backpropagation*. Las etapas del método propuesto se presentan en la figura 4.1 mediante un diagrama de flujo.

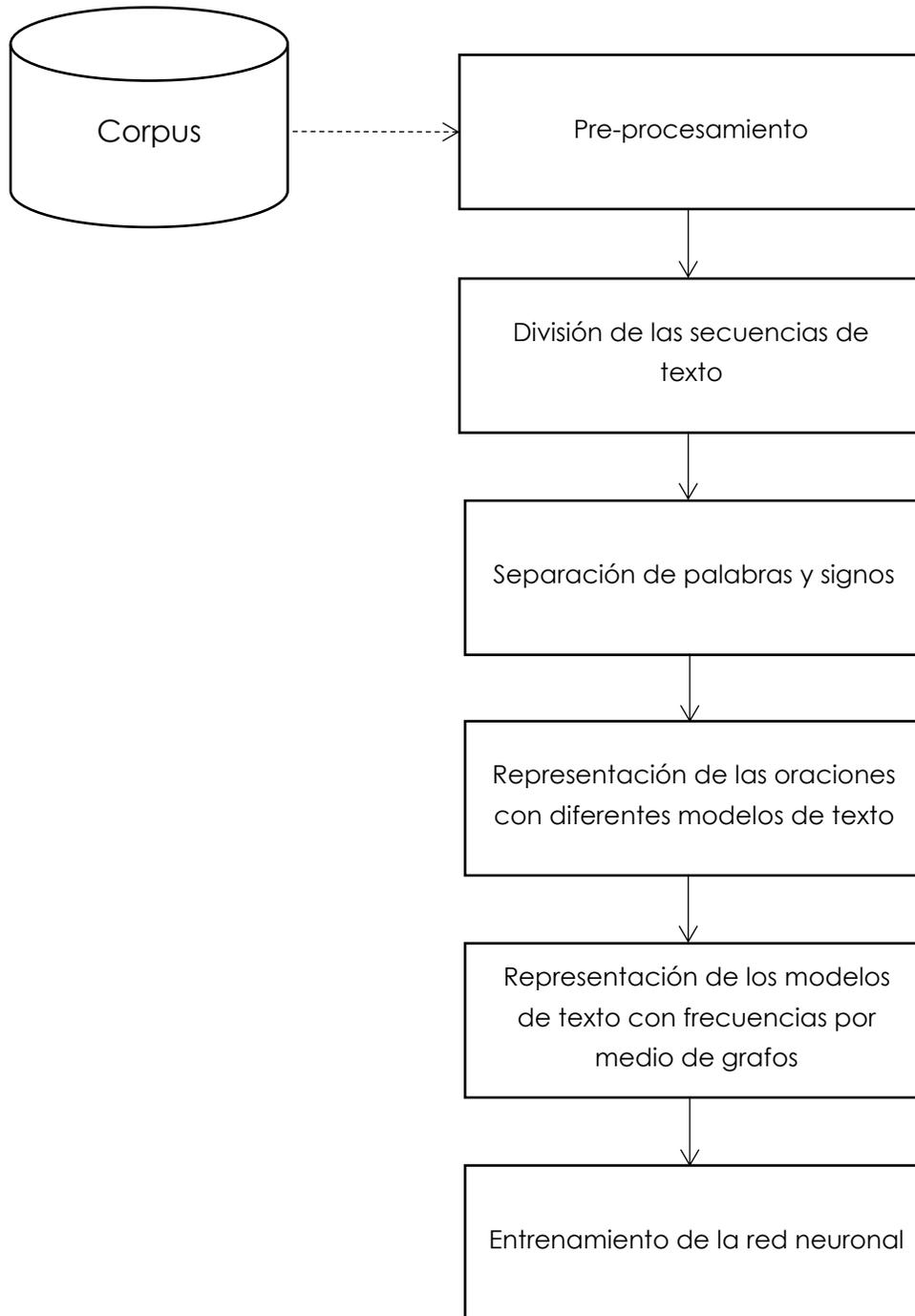


Figura 4.1 Método propuesto general

4.2.1 Preprocesamiento

En esta etapa se obtiene y limpia cada texto del corpus a usar.

4.2.2 División de las secuencias de texto

En esta etapa las secuencias de texto de cada artículo son separadas para formar oraciones.

4.2.3 Separación de palabras y signos

En esta etapa las palabras y signos que contiene cada oración son separados. Para separar las palabras y signos solo se agrega un espacio entre cada uno.

4.2.4 Representación de las oraciones con diferentes modelos de texto

En esta etapa se representan las oraciones con los modelos diferentes modelos de texto para obtener la frecuencia del corpus, primero se representa cada oración como un grafo. Los vértices representan las palabras que contiene cada oración. Como ejemplo, para representar la siguiente oración "La calle estaba mojada" la cual contiene 4 vértices $V = \{La, calle, estaba, mojada\}$. Primero se analiza la estructura que tiene la oración y las palabras, signos y números son representados como vértices. Cada vértice se representa como un modelo de unigrama, bigrama contiguos y bigrama con salto en s.

Tabla 4.2 Representación de la oración con los modelos de unigrama, bigrama contiguos y bigrama con salto en s

Palabra	Vértice	Unigrama	Bigrama	Bigrama con salto en 1
La	v_1	v_1	$v_1 v_2$	$v_1 v_3$
calle	v_2	v_2	$v_2 v_3$	$v_2 v_4$
estaba	v_3	v_3	$v_3 v_4$	
mojada	v_4	v_4		

Como se muestra en la tabla 4.2 la oración contiene 4 vértices y cada uno de estos representa a 4 modelos de unigrama, 3 modelos de bigramas contiguos y 2 bigramas con salto en 1. Para el modelo de unigrama se obtiene un total de 4 frecuencias que se obtienen del corpus. Para

los modelos de bigramas y bigramas con salto en 1 también se obtiene la frecuencia. La frecuencia de cada modelo de n -gramas se guarda.

4.2.5 Representación de los modelos de n -gramas con su frecuencia por medio de grafos

En esta sección se presenta el modelo de texto que se propone en esta tesis de manera formal.

4.2.5.1 Representación de las oraciones y frecuencias de unigrama por medio de grafos

Para representar los grafos de oraciones con el modelo de frecuencia de unigrama primero se modela cada grafo de oración como se muestra en la figura 4.2. Se obtiene el número total de vértices y por cada vértice se genera un lazo. Los vértices representan el modelo de unigrama que contiene una palabra. Los lazos representan la frecuencia para el modelo de unigrama que se guardó en el paso anterior. El proceso es igual para cada grafo de oración que se presenta, el proceso se muestra a continuación.

Se tiene los vértices $v = \{v_1, v_2, v_3, \dots, v_n\}$ y las aristas $f(e) = \{f(e_1), f(e_2), f(e_3), \dots, f(e_n)\}$

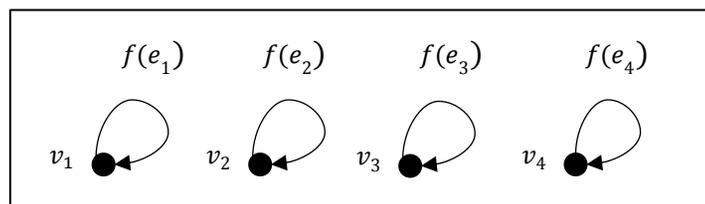


Figura 4.2 Modelo de frecuencia de unigrama

En la figura 4.2 se representa el grafo para el modelo de frecuencias de unigrama. La arista e_1 forma un lazo como el siguiente $f(e_{1\{v_1, v_1\}}) = frecuencia(unigrama_1)$, esto es igual para las aristas $f(e_2), f(e_3)$ y $f(e_4)$. En la tabla 4.4 se representan los modelos de unigramas obtenidos.

Tabla 4.3 Modelo de frecuencia de unigrama

Lazo	Vértices	Palabra	Frecuencia
e_1	$\{v_1, v_1\}$	v_1	$f(\text{unigrama}_1)$
e_2	$\{v_2, v_2\}$	v_2	$f(\text{unigrama}_2)$
e_3	$\{v_3, v_3\}$	v_3	$f(\text{unigrama}_3)$
e_4	$\{v_4, v_4\}$	v_4	$f(\text{unigrama}_4)$

4.2.5.2 Representación de las oraciones y frecuencias de bigrama contiguos por medio de grafos

Para representar los grafos de oraciones con el modelo de frecuencia de bigramas contiguos el proceso es similar al modelo de frecuencia de unigrama. Se tienen el conjunto de vértices $v = \{v_1, v_2, v_3, \dots, v_n\}$ y de aristas $f(e) = \{f(e_1), f(e_2), f(e_3), \dots, f(e_n)\}$. En la figura 4.2 se muestra el grafo para obtener el modelo de bigrama.

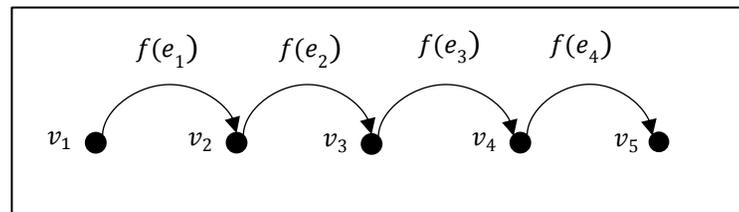


Figura 4.3 Modelo de bigrama

Cada arista asocia los vértices de la figura 4.3, como resultado la arista e_1 tiene la siguiente asociación $e_{1\{v_1, v_{1+1}\}} = \text{frecuencia}(\text{bigrama}_{1,2})$. Cada arista une dos palabras adyacentes de la oración como se muestra en la figura 4.3 y como resultado se forman cuatro bigramas que se muestran en la tabla 4.4.

Tabla 4.4 Modelo de frecuencia de bigrama

Arista	Vértices	Bigrama	Frecuencia
e_1	$\{v_1, v_2\}$	$v_1 v_2$	$\text{frecuencia}(\text{bigrama}_{1,2})$
e_2	$\{v_2, v_3\}$	$v_2 v_3$	$\text{frecuencia}(\text{bigrama}_{2,3})$
e_3	$\{v_3, v_4\}$	$v_3 v_4$	$\text{frecuencia}(\text{bigrama}_{3,4})$
e_4	$\{v_4, v_5\}$	$v_4 v_5$	$\text{frecuencia}(\text{bigrama}_{4,5})$

Para oraciones que tienen una longitud par de 22, 28 o 50 palabras se forma un modelo de bigrama para cada palabra. No obstante, si la longitud del grafo de oración es impar la última palabra no forma un nuevo bigrama contiguo.

4.2.5.3 Representación de las oraciones y frecuencias de bigrama con salto en s por medio de grafos

La representación de los grafos de oraciones en un modelo de bigrama con salto en s se muestra en la figura 4.4. El modelo de bigrama con salto en s se usa para analizar la relación que tiene el vértice v_1 con el vértice v_3 de la figura 4.4 y esto se usa para cada vértice que lo compone. Con este modelo de texto se analiza la relación que tiene la primera palabra con la tercera, cuarta o quinta de una oración. Para modelar oraciones de mayor tamaño se tiene los vértices $v = \{v_1, v_2, v_3, \dots, v_n\}$ y las aristas $f(e) = \{f(e_1), f(e_2), f(e_3), \dots, f(e_n)\}$ que forman los nuevos modelos de bigramas con salto en s .

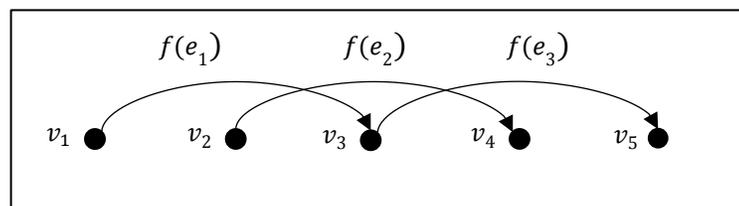


Figura 4.4 Modelo de bigrama con salto en s

Cada arista une dos palabras de la oración como se muestra en la figura 4.4 y como resultado se forman 3 bigramas con salto en 1 que se muestran en la tabla 4.5, además las aristas $f(e_1)$, $f(e_2)$ y $f(e_3)$ tienen la siguiente asociación $e_{\{v_i, v_{i+1}\}} = frecuencia(bigrama_{i, i+2})$

Tabla 4.5 Modelo de bigramas con salto en s

Arista	Vértices	Bigrama con salto en 1	Frecuencia
e_1	$\{v_1, v_2\}$	$v_1 v_3$	$frecuencia(bigrama_{1,3})$
e_2	$\{v_2, v_3\}$	$v_2 v_4$	$frecuencia(bigrama_{2,4})$
e_3	$\{v_3, v_4\}$	$v_3 v_5$	$frecuencia(bigrama_{3,5})$

La frecuencia de los modelos de n -gramas se normalizan dividiendo la frecuencia obtenida entre la frecuencia más alta del modelo evaluado. En la tabla 4.6 se muestran las frecuencias más altas para cada modelo. En la tabla 4.6 se muestra las frecuencias más altas para cada modelo de n -gramas.

Tabla 4.6 Frecuencia de 3 n -gramas normalizados

Modelo	Palabra	Frecuencia más alta	Frecuencia normalizada
Unigrama	de	495	1
Bigrama	de la	79	1
Bigrama con salto en 1	la de	67	1

También se normalizan las letras de cada palabra. Para normalizar las letras se les asigna un valor fijo a cada letra, signo y número. Después, se divide el valor de cada letra, signo y número entre el valor más grande que fue asignado. Como ejemplo, en la tabla 4.7 se muestra la palabra 'Esto', la letra E tiene el valor de 69 y para normalizarlo se divide entre 256 y el resultado es de 0.26953125.

Tabla 4.7 Letras de una oración normalizada

Letra	Valor	Valor normalizado
E	69	0.26953125
s	115	0.44921875
t	116	0.453125
o	111	0.43359375

4.2.5.4 Datos de entrenamiento.

En esta etapa se genera el archivo o carpeta que contiene los datos de entrenamiento. Los datos de entrenamiento están formados por palabras como patrones de entrada y las frecuencias como patrones de salida.

4.2.6 Entrenamiento de la red neuronal

En esta etapa se inicia el entrenamiento de la red neuronal *backpropagation* con los modelos de frecuencia de *n*-gramas. Para iniciar el entrenamiento se reciben los patrones de entrada y salida que la red neuronal *backpropagation* debe generalizar y asociar. Como entrada tiene las palabras normalizadas y las frecuencias de cada modelo de *n*-gramas normalizados. En la tabla 4.8 se muestra un ejemplo de entradas y salidas que la red neural tiene.

Tabla 4.8 Formato de patrones de entrada y salida

Patrones de entrada	Patrones de salida
0.302,0.306,0.0,0.0,0.0,0.0	1.0
0.053,0.0,0.0,0.0,0.0,0.0,0.0	0.834
0.337,0.288,0.0,0.0,0.0,0.0	0.610
0.0622,0.0,0.0,0.0,0.0,0.0	0.560
0.306,0.346,0.0,0.0,0.0,0.0	0.446
0.306,0.337,0.0,0.0,0.0,0.0	0.386

4.2.6.3 Creación de las capas de entrada, oculta y salida con diferente número de neuronas

En esta etapa se crea la capa de entrada, capa oculta y capa de salida con el número de neuronas necesario para el patrón de entrada y salida. El número de neuronas se modifica para los diferentes archivos de entrenamiento.

4.2.6.4 Inicio de entrenamiento

Se inicia el entrenamiento de la red neuronal con los patrones de entrada y salidas que recibió en el archivo de entrenamiento. En esta etapa inicia la aproximación al valor de salida deseado en cada época que realiza la red neuronal. Dentro de este proceso de entrenamiento se evalúa el error que genera cada patrón y se ajusta los pesos de todas las

neuronas de cada capa. Este proceso es un ciclo y para detenerse se debe alcanzar un error mínimo o un número de épocas máximo definido antes de iniciar el entrenamiento.

4.2.6.5 Detección del grado de aceptación de un texto de acuerdo al contexto y dominio

En esta etapa se detecta el grado de aceptación de diferentes de texto de contextos y dominios diferentes. La detección del grado de aceptación se realiza con la salida que tiene la red neuronal después de la etapa de entrenamiento.



CAPÍTULO 5.

Experimentación

En este capítulo se muestran los experimentos y resultados obtenidos del método propuesto. Cada paso se desarrolla siguiendo el orden descrito en el método propuesto. Para el funcionamiento del método propuesto los datos de entrada son obtenidos del programa de Kiwix. Kiwix es un programa que permite visualizar y descargar artículos de Wikipedia en formato HTML.

5.1 Datos de entrada

Wikipedia es una enciclopedia digital multilingüe que contiene más de 46 millones de artículos en 288 idiomas, los artículos son redactados conjuntamente por voluntarios de todo el mundo y cualquier persona puede editarlos. Wikipedia en español cuenta con 1,385,604 artículos hasta esta fecha y crece cada día gracias a la participación de gente de todo el mundo, esto lo hace el mayor proyecto de recopilación de conocimiento jamás realizado en la historia de la humanidad (Wikipedia, s.f.).

5.2 Preprocesamiento

Se obtiene y limpia cada texto del corpus a usar para el desarrollo de la tesis.

5.2.1 Obtener artículos HTML de Wikipedia

Los artículos de Wikipedia son obtenidos de la aplicación Kiwix. La aplicación Kiwix permite visualizar y descargar artículos de Wikipedia sin conexión a internet. Con la aplicación de Kiwix se descargan artículos aleatorios para formar el corpus de entrada. En la figura 4.1 se muestra la interfaz de la aplicación.



Figura 5.1 Interfaz de Kiwix

Se descargaron 40,749 artículos de forma aleatoriamente. Todos los artículos están en el idioma español y tienen una extensión .htm que asigna el programa de Kiwix.

5.2.3 Eliminar etiquetas HTML de artículos de Wikipedia

En esta etapa se eliminan las etiquetas HTML que contienen los artículos de Wikipedia, en la tabla 5.1 se muestran las etiquetas eliminadas. En la figura 5.2 se muestra un artículo con la estructura y etiquetas HTML que contiene.

```

<html class=""><head>
<meta http-equiv="content-type" content="text/html; charset=UTF-8"><!-- base
href="zim://A/M%C3%A9xico.html" -->
  <meta charset="UTF-8">
  <title>México</title>
  <meta name="viewport" content="width=device-width, initial-scale=1.0">
  <link rel="stylesheet" href="M%C3%A9xico_files/a">
  <script src="M%C3%A9xico_files/a_002"></script><script src="M%C3%A9xico_files/a_004"></script>
</head>
<body class="mw-body mw-body-content mediawiki" style="background-color: white; margin: 0; border-
width: 0px; padding: 0px;">
  <div id="content" class="mw-body" style="padding: 1em; border-width: 0px; max-width: 55.8em; margin: 0
auto 0 auto">
    <a id="top"></a>
    <h1 id="titleHeading" style="background-color: white; margin: 0;">México</h1>
    <div id="mw-content-text">
    <table class="infobox geography vcard" style="width:22.7em; line-height: 1.4em; text-align:left;

```

Figura 5.2 Etiquetas HTML que contiene los artículos

Como resultado de eliminar las etiquetas HTML se obtiene una secuencia de texto de cada artículo, la secuencia de texto contiene letras, signos y números. En la figura 5.3 se muestra la secuencia de texto obtenida. Además, los artículos que contienen menos de 50 palabras son eliminados en este paso y se obtienen 28,705 artículos sin etiquetas HTML.

Tabla 5.1 Etiquetas HTML eliminadas

Etiquetas
head
tbody
h1,h2,h3,h4,h5,h6
table
br

México, cuyo nombre oficial es Estados Unidos Mexicanos, es un país de América, ubicado en la parte meridional de América del Norte. Su capital es la Ciudad de México. Políticamente es una república democrática, representativa y federal compuesta por 32 entidades federativas (31 estados y la Ciudad de México).

El territorio mexicano tiene una superficie de 1 964 375 km², por lo que es el decimocuarto país más extenso del mundo y el tercero más grande de América Latina. Limita al norte con los Estados Unidos de América a lo largo de una frontera de 3 118 km, mientras que al sur tiene una frontera de 956 km con Guatemala y 193 km con Belice; las costas del país limitan al oeste con el océano Pacífico y al este con el golfo de México y el mar Caribe, sumando 11 593 km, por lo que es el tercer país americano con mayor longitud de costas.

México es el undécimo país más poblado del mundo, con una población estimada de 119 millones de personas en 2015, la mayoría de las cuales tienen como lengua materna el español, al que el estado reconoce como lengua nacional junto a 67 lenguas indígenas propias de la nación. En el país se hablan alrededor de 287 idiomas; debido a las características de su población, es el país hispanohablante más poblado, así como el séptimo país con mayor diversidad lingüística en el mundo.

La presencia humana en México se remonta a 14 000 años antes del presente. Después de miles de años de desarrollo cultural, surgieron en el territorio mexicano las culturas mesoamericanas, aridoamericanas y

Figura 5.3 Secuencias de texto

5.3 División de las secuencias de texto

Las secuencias de texto de cada artículo se dividen para formar oraciones, en la figura 5.4 se muestran las oraciones formadas de las secuencias de texto de los artículos de *Wikipedia* en español.

México, cuyo nombre oficial es Estados Unidos Mexicanos, es un país de América, ubicado en la parte meridional de América del Norte.
Su capital es la Ciudad de México.
Políticamente es una república democrática, representativa y federal compuesta por 32 entidades federativas (31 estados y la Ciudad de México).
El territorio mexicano tiene una superficie de 1 964 375 km², por lo que es el decimocuarto país más extenso del mundo y el tercero más grande de América Latina.
Limita al norte con los Estados Unidos de América a lo largo de una frontera de 3 118 km, mientras que al sur tiene una frontera de 956 km con Guatemala y 193 km con Belice; las costas del país limitan al oeste con el océano Pacífico y al este con el golfo de México y el mar Caribe, sumando 11 593 km, por lo que es el tercer país americano con mayor longitud de costas.
México es el undécimo país más poblado del mundo, con una población estimada de 119 millones de personas en 2015, la mayoría de las cuales tienen como lengua materna el español, al que el estado reconoce como lengua nacional junto a 67 lenguas indígenas propias de la nación.
En el país se hablan alrededor de 287 idiomas; debido a las características de su población, es el país hispanohablante más poblado, así como el séptimo país con mayor diversidad lingüística en el mundo.
La presencia humana en México se remonta a 14 000 años antes del presente.

Figura 5.4 Oraciones formadas de las secuencias de textos

5.4 Separación de letras y signos

En esta etapa las palabras y signos que contiene cada artículo son separados. En la figura 5.4 se muestra las secuencias de texto sin separación. En esta etapa solo se agrega un espacio entre cada palabra, signo y número como se muestra en la figura 5.5.

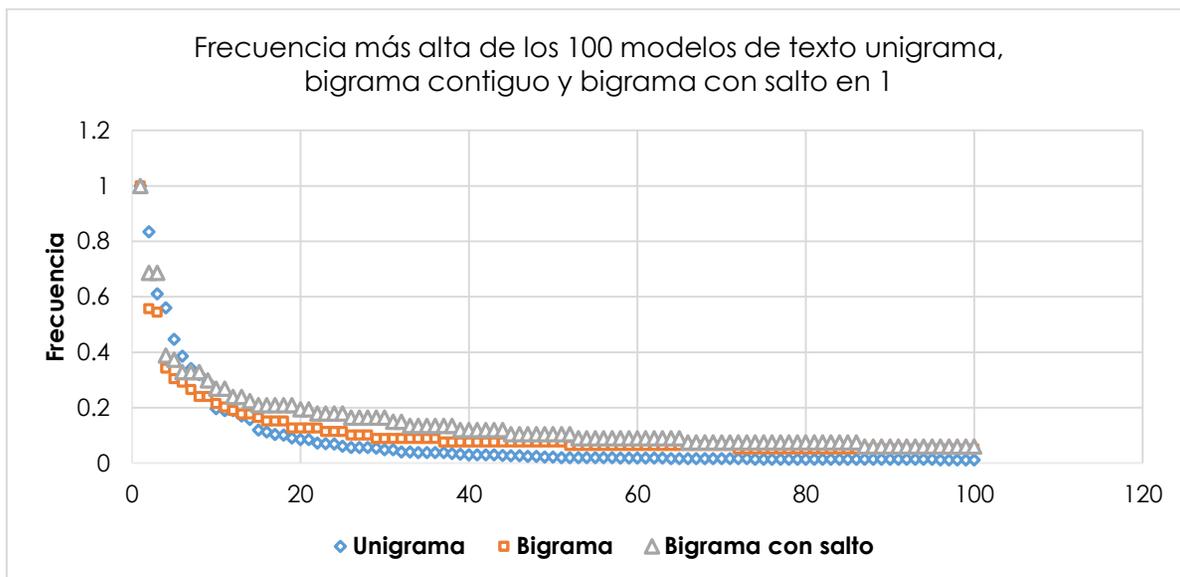
México , cuyo nombre oficial es Estados Unidos Mexicanos , es un país de América , ubicado en la parte meridional de América del Norte .
Su capital es la Ciudad de México .
Políticamente es una república democrática , representativa y federal compuesta por 32 entidades federativas (31 estados y la Ciudad de México) .
El territorio mexicano tiene una superficie de 1 964 375 km² , por lo que es el decimocuarto país más extenso del mundo y el tercero más grande de América Latina .
Limita al norte con los Estados Unidos de América a lo largo de una frontera de 3 118 km , mientras que al sur tiene una frontera de 956 km con Guatemala y 193 km con Belice ; las costas del país limitan al oeste con el océano Pacífico y al este con el golfo de México y el mar Caribe , sumando 11 593 km , por lo que es el tercer país americano con mayor longitud de costas .
México es el undécimo país más poblado del mundo , con una población estimada de 119 millones de personas en 2015 , la mayoría de las cuales tienen como lengua materna el español , al que el estado reconoce como lengua nacional junto a 67 lenguas indígenas propias de la nación .
En el país se hablan alrededor de 287 idiomas ; debido a las características de su población , es el país hispanohablante más poblado , así como el séptimo país con mayor diversidad lingüística en el mundo .
La presencia humana en México se remonta a 14 000 años antes del presente .

Figura 5.5 Palabras y signos separados por un espacio

5.5 Representación de las oraciones con diferentes modelos de texto

Las oraciones se representan con el modelo de texto de unigrama, bigrama y bigrama con salto en 1. Además, para cada modelo de n -gramas se obtiene la frecuencia que tiene en el corpus de muestra. Para formar la muestra del corpus se seleccionan 240 oraciones de forma aleatoria de los 28,705 artículos de *Wikipedia* en español. Con las 240 oraciones se forman 2,774 modelos de unigrama. Para el modelo de bigrama contiguo se obtiene un total de 6,217 bigramas y para el modelo de bigramas con salto en 1 se obtiene un total de 6,005 bigramas. En el modelo de unigrama la palabra más frecuente es la palabra 'de' y para el modelo de bigrama contiguo el bigrama más frecuente es 'de la'. En el modelo de bigrama con salto en 1 el bigramas más frecuente es 'la de'. Las frecuencias se obtienen de la muestra de 240 oraciones de *Wikipedia* en español y las frecuencias se guardan.

En la gráfica 5.1 se muestran la frecuencia más alta de los 100 modelos de texto unigrama, bigrama contiguo y bigrama con salto en 1.



Gráfica 5.1 Frecuencia más alta de los 100 modelos de unigrama, bigrama contiguo y bigrama con salto en 1

Las frecuencias de unigramas, bigramas contiguos y bigramas con salto en 1 de la gráfica 5.1 están normalizados.

5.4 Representación de las oraciones y las frecuencias de modelos de n-gramas por medio de grafos

La frecuencia que se obtiene de las 240 oraciones de *Wikipedia* en español con los diferentes modelos de *n*-gramas se representa con grafos para formar el modelo de texto que se usara para entrenar la red neuronal. Para formar el modelo de frecuencia de los diferentes modelos *n*-gramas se usan las 240 oraciones y las frecuencias que se tienen guardadas de los diferentes modelos de *n*-gramas. Cada palabra de la oración es representada como un vértice y la frecuencia se representa como un lazo para el modelo de unigrama y como arista para el modelo de bigrama. Como ejemplo, en la tabla 5.2 se muestra el modelo de frecuencia de unigrama para la siguiente frase “Luego de un mal arranque”.

Tabla 5.2 Modelo de frecuencia de unigrama para 5 palabras

Vértice	Lazo
$v_1 =$ Luego	$e_1 = 4$
$v_2 =$ de	$e_2 = 495$
$v_3 =$ un	$e_3 = 78$
$v_4 =$ mal	$e_4 = 2$
$v_5 =$ arranque	$e_5 = 1$

En la tabla 5.3 se muestra se muestra la misma frase, pero con el modelo de frecuencia de bigrama contiguo.

Tabla 5.3 Modelo de frecuencia de bigrama contiguo para 5 palabras

Vértice	Lazo
$v_1, v_2 =$ Luego de	$e_1 = 3$
$v_2, v_3 =$ de un	$e_2 = 10$
$v_3, v_4 =$ un mal	$e_3 = 1$
$v_4, v_5 =$ mal arranque	$e_4 = 1$

En la tabla 5.4 se muestra se muestra la frase anterior con el modelo de frecuencia de bigrama con salto en 1.

Tabla 5.4 Modelo de frecuencia de bigrama con salto en 1 para 5 palabras

Vértice	Lazo
$v_1, v_3 =$ Luego un	$e_1 = 2$
$v_2, v_4 =$ de mal	$e_2 = 1$
$v_3, v_5 =$ un arranque	$e_3 = 1$

5.4 Entrenamiento de la red neuronal

En esta sección se presentan los resultados que tiene cada experimento con diferentes muestras de entrenamiento y diferentes configuraciones en la red neuronal *backpropagation*.

5.4.1 Entrenamiento de la red neuronal con unigrama

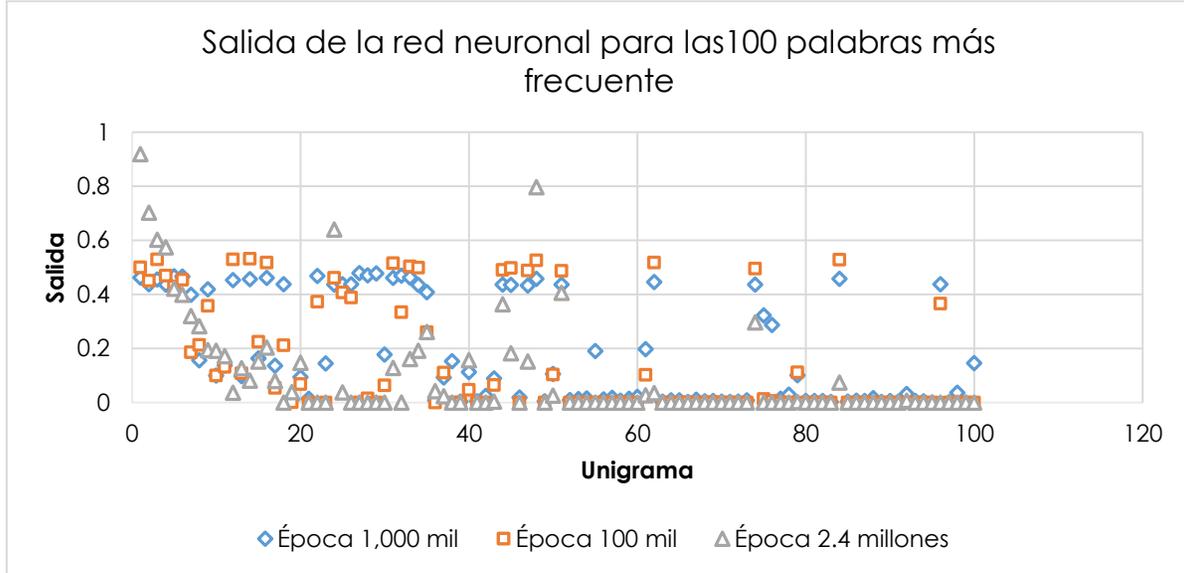
La muestra de entrenamiento para este experimento contiene el modelo de frecuencia de unigrama. El tamaño para los patrones de entrada es de 18 caracteres y como salida se tiene un solo patrón. Los patrones de entrada contienen palabras, números y signos. Los patrones de salida son la frecuencia de cada unigrama. La muestra de entrenamiento contiene 7,857 modelos de frecuencias de unigramas. Esta muestra de entrenamiento se formó de las 240 oraciones de *Wikipedia* en español.

La configuración de la red neuronal *backpropagation* para iniciar el entrenamiento se realiza con los siguientes parámetros:

- 18 neuronas en la capa de entrada
- 200 neuronas en la primera capa oculta
- 1 neurona en la capa de salida

Las 18 neuronas en la capa de entrada representan el número de caracteres máximo que contiene la palabra con mayor longitud y la neurona en la capa de salida representa la frecuencia del modelo de unigrama.

La salida que tiene la red neuronal con 3 épocas diferentes para las 100 primeras palabras se muestra en la gráfica 5.2. Las salidas obtenidas en la última época de la red neuronal de la gráfica 5.2 muestran un comportamiento similar a la gráfica 5.1 que contiene las frecuencias reales. Algunas salidas tienen un valor alto, pero como se observa en la gráfica 5.2 con cada época los valores altos disminuyen.



Gráfica 5.2 Salida de la red neuronal con unigrama para las 100 palabras más frecuentes con 3 épocas diferentes

5.4.2 Entrenamiento de la red neuronal con bigrama contiguo (salto igual a 0)

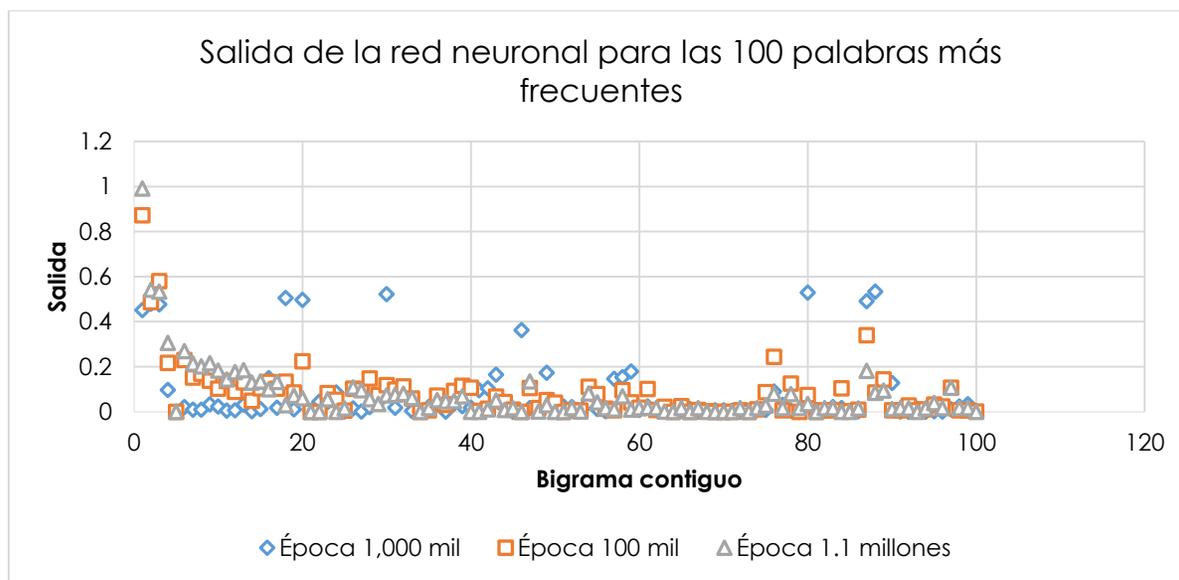
La muestra de entrenamiento para este experimento contiene el modelo de frecuencia de bigrama contiguo. El tamaño para los patrones de entrada es de 26 caracteres y como salida se tiene solo un patrón de salida. Los patrones de entrada contienen una combinación 2 palabras, números o signos. Los patrones de salida son la frecuencia de cada bigrama contiguo. La muestra de entrenamiento contiene 7,617 modelos de frecuencias de bigrama contiguo. La muestra de entrenamiento se formó de las 240 oraciones de *Wikipedia* en español.

El entrenamiento de esta red neuronal *backpropagation* se realiza con la siguiente configuración:

- 26 neuronas en la capa de entrada
- 200 neuronas en la primera capa oculta
- 1 neurona en la capa de salida

El número de neuronas en la capa de entrada está formado por 26 neuronas y las 26 neuronas representan los caracteres que forman el bigrama contiguo.

La salida de la red neuronal en la época 500 mil muestra una salida cercana al valor real de las primeras 100 palabras. En la gráfica 5.3 se muestran la salida de la red neuronal para las 100 palabras más frecuentes.



Gráfica 5.3 Salida de la red neuronal con bigrama contiguo para las 100 palabras más frecuentes con 3 épocas diferentes

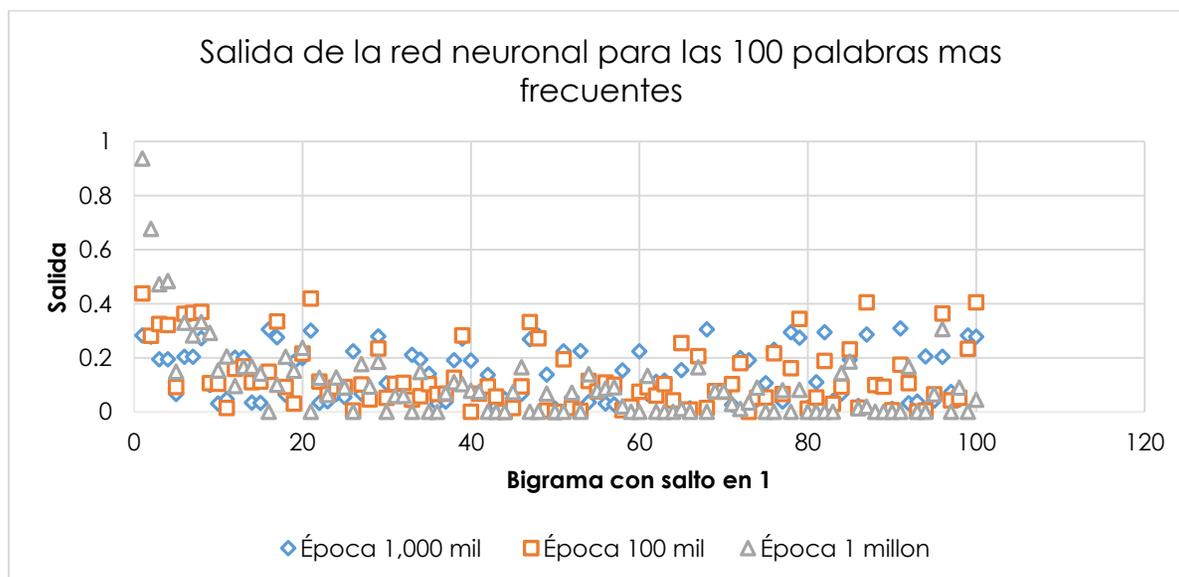
5.4.3 Entrenamiento de la red neuronal con bigrama con salto en 1

La muestra de entrenamiento para este experimento está formada por 7,377 bigramas con salto en 1. El número de neuronas en la capa de entrada es de 30 neuronas. Los bigramas con salto en 1 se construyeron a partir de 240 oraciones de Wikipedia en español. La neurona de salida representa la frecuencia del modelo de bigrama con salto en 1.

El entrenamiento de esta red neuronal *backpropagation* se realiza con la siguiente configuración:

- 30 neuronas en la capa de entrada
- 200 neuronas en la primera capa oculta
- 1 neurona en la capa de salida

La salida que se obtiene para el bigrama con salto en 1 en las diferentes épocas se muestra en la tabla 5.4 para las primeras 100 palabras más frecuentes.



Gráfica 5.4 Salida de la red neuronal con bigrama con salto en 1 para las 100 palabras más frecuentes con 3 épocas diferentes

5.5 Detección del grado de aceptación de un texto de acuerdo al contexto y dominio

Para evaluar los resultados obtenidos del entrenamiento de la red neuronal se detecta el grado de aceptación de diferentes fragmentos de textos en español, inglés y alemán. Los fragmentos de textos están formados por un total de 100 palabras, signos y números.

5.5.1 Detección del grado de aceptación de un fragmento de un artículo de Wikipedia en español

El fragmento de texto de Wikipedia en español está formado por archivo diferente al de entrenamiento y el fragmento de texto a evaluar es el siguiente: "El objetivo de Gipuzkoa

Sarean era mejorar la competitividad y el bienestar de Gipuzkoa a través del desarrollo del capital social . La cuestión principal que guió la iniciativa era la siguiente : históricamente Gipuzkoa ha disfrutado de un gran capital social organizado en la acción colectiva , compromiso cívico y desarrollo de empresas cooperativas , que en el pasado ha configurado una importante ventaja competitiva . No obstante , en la actualidad , se percibe una situación de declive del capital social derivada del predominio de valores individualistas que pueden afectar al bienestar , la cohesión social y la”.

Las salidas que asigna la red neuronal al fragmento de texto se suman. Para el modelo de unigrama la suma da un total de 22.069. El modelo de bigrama contiguo da un total de 3.928 y el modelo de bigrama con salto en 1 da un total de 4.467. La suma de los 3 modelos se muestra en la tabla 5.5.

Tabla 5.5 Grado de aceptación para el fragmento de texto de Wikipedia en español

Modelo	Sumatoria
Unigrama	22.069
Bigrama contiguo	3.928
Bigrama con salto en 1	4.467
Total	30.464

5.5.2 Detección del grado de aceptación de un fragmento de poemas

El fragmento de texto a usar pertenece a varios poemas y se usan las 100 palabras que se muestran a continuación: “Quiero y no quiero querer a quien no queriendo quiero . He querido si querer y sigo , sin querer , queriendo . si porque tú me quieras quieres que te quiera más : te quiero más que me quieres , ¿ qué más quieres ? , ¿ quieres más ? No me mires , que miran que nos miramos . Miremos la manera de no mirarnos . Nos miraremos , y cuando nadie nos mire , nos miraremos . Comprendo que tus besos jamás han de ser míos ; comprendo que en tus ojos no me he de ver”.

La suma de los 100 modelos de unigramas da un total de 19.185. El modelo de bigrama contiguo da un total de 3.163 y el modelo de bigrama con salto en 1 da un total de 5.379. La suma de los 3 modelos se muestra en la tabla 5.6.

Tabla 5.6 Grado de aceptación para el fragmento de poemas

Modelo	Sumatoria
Unigrama	19.185
Bigrama contiguo	3.163
Bigrama con salto en 1	5.379
Total	27.727

5.5.3 Detección del grado de aceptación de un fragmento de un libro en español

El fragmento de texto pertenece al libro de 100 años de soledad y solo se usan 100 palabras del primer. El fragmento de texto a evaluar es el siguiente: *“Muchos años después , frente al pelotón de fusilamiento , el coronel Aureliano Buendía había de recordar aquella tarde remota en que su padre lo llevó a conocer el hielo . Macondo era entonces una aldea de veinte casas de barro y cañabrava construidas a la orilla de un río de aguas diáfanas que se precipitaban por un lecho de piedras pulidas , blancas y enormes como huevos prehistóricos . El mundo era tan reciente , que muchas cosas carecían de nombre , y para mencionarlas había que señalarías con el dedo . Todos los años , por el mes”*.

Las salidas que asigna la red neuronal al fragmento de texto se suman. Para el modelo de unigrama la suma da un total de 20.422. El modelo de bigrama contiguo da un total de 2.582 y el modelo de bigrama con salto en 1 da un total de 3.835. La suma de los 3 modelos se muestra en la tabla 5.7.

Tabla 5.7 Grado de aceptación para el fragmento de libro en español

Modelo	Sumatoria
Unigrama	20.422
Bigrama contiguo	2.582
Bigrama con salto en 1	3.835
Total	26.839

5.5.4 Detección del grado de aceptación de un fragmento de una noticia es español

El fragmento de texto a usar pertenece a una noticia tomada del portal de noticia "El Universal" y se usan 100 palabras de la noticia: "Suman 13 las personas fallecidas por el accidente ocurrido esta mañana en la autopista México - Pachuca ; uno de los lesionados falleció cuando era trasladado al hospital Magdalena de las Salinas , en la Ciudad de México . Se prevé para el Pacífico Sur cielo nublado con tormentas puntuales muy fuertes acompañadas de actividad eléctrica en Guerrero y Oaxaca , y tomentas fuertes en Chiapas , además un ambiente caluroso y viento de dirección variable de 15 a 30 kilómetros por hora en la región . En la Mesa del Norte se estima cielo con tormentas puntuales muy fuertes".

La suma de la salida de los 100 modelos de unigrama da un total de 18.046. La suma del modelo de bigrama contiguo da un total de 3.588 y el modelo de bigrama con salto en 1 da un total de 4.385. La suma de los 3 modelos se muestra en la tabla 5.8.

Tabla 5.8 Grado de aceptación para el fragmento de noticia en español

Modelo	Sumatoria
Unigrama	18.046
Bigrama contiguo	3.588
Bigrama con salto en 1	4.385
Total	26.019

5.5.5 Detección del grado de aceptación de un fragmento de publicaciones de Facebook

El fragmento de texto a usar pertenece a varias publicaciones de una página de Facebook y se usa un total de 100 palabras de las siguientes publicaciones: "Que onda raza alguien que juegue nos falta 1 en flex ? Una cuenta bronze - plata que me presten para subirla XD it's free La cambio por una de lol de \$ 200 Hermanos , busco ayuda para crecer en el mundo de twitch . . . puesto que es un proyecto que tengo y la verdad agradecería que pasaras a platicar y darme una oportunidad te juro que sera divertido c: pd : no soy muy bueno en el juego pero soy buen pedo alv Jugando Solo Q en Platinoob *en* Como jugar con los codos League of".

La suma de la salida para el modelo de unigrama da un total de 15.621. El modelo de bigrama contiguo da un total de 3.983 y el modelo de bigrama con salto en 1 da un total de 4.210. La suma de los 3 modelos se muestra en la tabla 5.9.

Tabla 5.9 Grado de aceptación para el fragmento de publicaciones de Facebook

Modelo	Sumatoria
Unigrama	15.621
Bigrama contiguo	3.983
Bigrama con salto en 1	4.210
Total	23.814

5.5.6 Detección del grado de aceptación de un fragmento de Wikipedia en idioma inglés

El fragmento de texto a usar pertenece a un artículo de Wikipedia en idioma inglés y se usan las siguientes 100 palabras: "November 3 The Yale Bulldogs traveled to West Point , and finally yielded some points , with the Army Cadets taking a 6 - 0 lead at halftime . Yale made no first downs , but won the game anyway . Clarence Alcott blocked a punt and returned it for a touchdown to tie the game 6 - 6 on the point after . With two minutes left , Bigelow of Yale kicked a 35 - yard field goal (for 4 points) from a steep angle , and a 10 - 6 win . Although there was no".

La suma de las salidas de los modelos de unigrama da un total de 14.783. El modelo de bigrama contiguo da un total de 1.644 y el modelo de bigrama con salto en 1 da un total de 2.049. La suma de los 3 modelos se muestra en la tabla 5.10.

Tabla 5.10 Grado de aceptación para el fragmento de Wikipedia en idioma inglés

Modelo	Sumatoria
Unigrama	14.783
Bigrama contiguo	1.644
Bigrama con salto en 1	2.049
Total	18.476

5.5.7 Detección del grado de aceptación de un fragmento de Wikipedia en idioma alemán

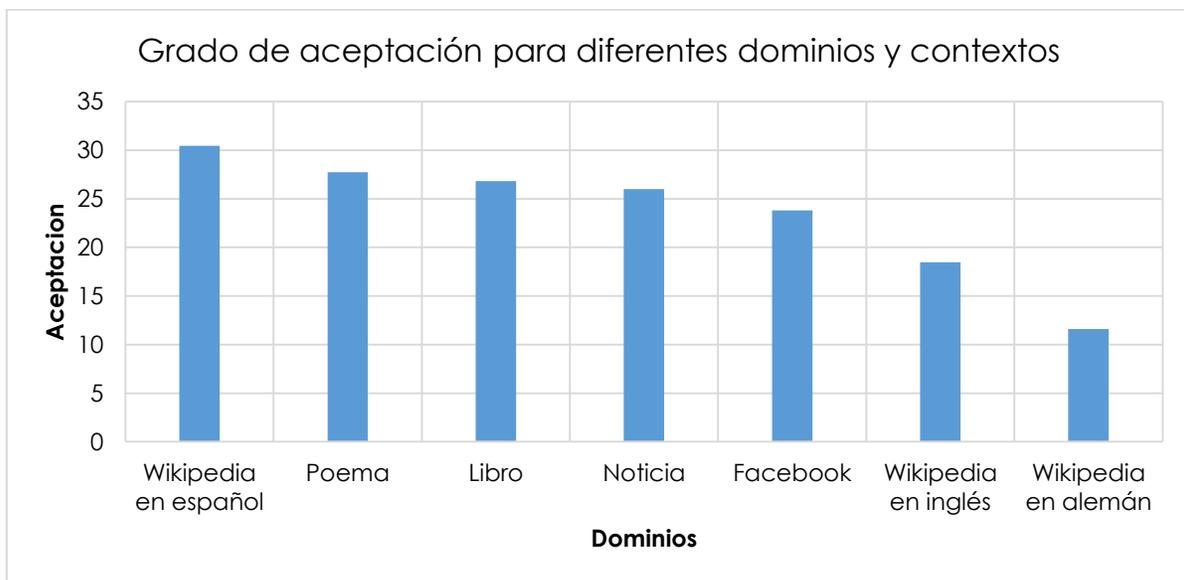
El fragmento de texto a usar pertenece a un artículo de Wikipedia en idioma alemán y se usan 100 palabras siguientes: *“Strecker behauptete 1911 gegenüber Rudolf Franke , dass für Schwachstromtechnik keinerlei Interesse bestehe und führte als Beweis seine bestenfalls 5 Hörer der letzten 20 Jahre an . Durch diese Beratertätigkeit machte Strecker die Bekanntschaft mit der Leitung der Deutschen Edison Gesellschaft und ließ sich für die Jahre 1885 bis 1886 anwerben . Dort in Berlin heiratete er auch am 10 . Oktober 1885 Luise Julie Wilhelmine Anna Sandberger ; mit ihr hatte er eine Tochter . Anlässlich seines 50 . Geburtstages ernannte man Strecker 1908 zum Geheimen Oberpostrat . Am 1 . Juli 1910 wählt man ihn zum Präsidenten des”*.

La suma de la salida para el modelo de unigrama total de 9.078. El modelo de bigrama contiguo da un total de 1.108 y el modelo de bigrama con salto en 1 da un total de 1.414. La suma de los 3 modelos se muestra en la tabla 5.11.

Tabla 5.11 Grado de aceptación para el fragmento de Wikipedia en idioma alemán

Modelo	Sumatoria
Unigrama	9.078
Bigrama contiguo	1.108
Bigrama con salto en 1	1.414
Total	11.6

Los resultados obtenidos de la evaluación de los diferentes fragmentos de texto con la red neuronal se analizan en la gráfica 5.5 y la tabla 5.3. El artículo de *Wikipedia* en español obtiene la sumatoria más alta de todos los fragmentos de texto evaluados. La suma de las salidas de la noticia, libro y poema tiene valores cercanos al fragmento de *Wikipedia* en español, pero no son superiores. Los valores de la noticia, libro y poema tienen estos valores porque contienen un nivel de formalidad parecido a *Wikipedia* en español. Sin embargo, no son del dominio de *Wikipedia* en español. La sumatoria de frecuencia de las publicaciones de *Facebook* obtiene un valor de 23.814 por el nivel de formalidad con el que están escritas las publicaciones.



Gráfica 5.5 Sumatoria de los diferentes contextos y dominios

Los fragmentos de texto de los artículos de *Wikipedia* en alemán e inglés obtienen una sumatoria menor a las publicaciones de *Facebook*. Esto indica que no tienen un nivel de pertenencia en el idioma español. La mayoría de palabras que tiene el artículo en idioma alemán e inglés no están presentes en el archivo de entrenamiento. Sin embargo, la red neuronal evalúa las palabras nuevas con frecuencias muy bajas.

Tabla 5.12 Sumatoria de las salidas de diferentes fragmentos de texto

Fragmentos de texto	Suma de salidas	Ranking
Wikipedia en español	30.464	1
Poema	27.727	2
Libro	26.839	3
Noticia	26.019	4
Facebook	23.814	5
Wikipedia en Inglés	18.476	6
Wikipedia en alemán	11.6	7

En la tabla 5.12 se muestra el ranking que obtuvo cada fragmento de texto evaluado. *Wikipedia* en español tiene el *ranking* 1 por el grado de aceptación y pertenencia que tiene con *Wikipedia* en español. Sin embargo, los fragmentos de texto en español del poema, libro, noticia y *Facebook* tienen un grado de pertenencia y aceptación menor a *Wikipedia* en español obteniendo un ranking de 2 a 5. Además, los fragmentos en español no tienen una suma de salidas menor a 20 por el grado de pertenencia que tienen con el idioma español. Los artículos de *Wikipedia* en inglés y alemán tienen el ranking 6 para inglés y 7 para alemán. Además, los fragmentos de *Wikipedia* en idioma inglés y alemán tienen una suma menor a 20 por el nivel de pertenencia que tienen para el idioma español. Los resultados que se obtienen para los fragmentos de *Wikipedia* en otro idioma son bajos y el contexto que tienen de *Wikipedia* no influye en los resultados. Además, el fragmento de *Wikipedia* en el idioma inglés tiene un grado de aceptación mayor en español que el fragmento de *Wikipedia* en el idioma alemán.

Con los experimentos realizados se concluye que la red neuronal *backpropagation* puede generalizar y asociar las palabras con su frecuencia mediante un modelo de frecuencia de n-gramas.

5.6 Resumen

En el presente capítulo se muestran los experimentos para el entrenamiento de una red neuronal *backpropagation*. En la primera parte se muestra el corpus de donde se obtuvieron los textos que contienen la información a usar. Además, se muestra el preprocesamiento para eliminar etiquetas que contiene cada texto. En la segunda parte se muestra el modelo de representación para los textos obtenidos del preprocesamiento. El modelo que se obtiene es usado para entrenar la red neuronal *backpropagation*. Además, se presentan los diferentes entrenamientos y resultados que se obtienen de la red neuronal. Se evalúa el entrenamiento de la red neuronal con diferentes fragmentos de texto para demostrar que clasifica nuevas palabras que no tiene en su archivo de entrenamiento y se detecta el grado de aceptación de un texto de acuerdo al contexto y dominio.



CAPÍTULO 6.

Conclusiones

6.1 Conclusiones

El método propuesto en este trabajo demostró detectar el grado de aceptación de un texto de acuerdo al contexto y dominio. Sin embargo, la muestra de entrenamiento usada para detectar el grado de aceptación solo contiene 240 oraciones aleatorias de diferentes artículos de *Wikipedia* en español. La muestra de 240 oraciones es usada para entrenar 3 redes neuronal *backpropagation* con diferentes configuraciones obteniendo buenos resultados. Las muestras de entrenamiento formadas con las 240 oraciones ayudaron a identificar los problemas que tiene la red neuronal *backpropagation* para asociar y generalizar la información contenida en un corpus. Esto ayudo a crear diferentes modelos de frecuencia de n -gramas y diferentes muestras de entrenamiento para la red neuronal *backpropagation*. Además, el modelo de frecuencia de n -gramas es independiente del lenguaje porque solo requiere de un corpus para obtener la información y crear los diferentes modelos de frecuencia de n -gramas.

La red neuronal *backpropagation* demostró buenos resultados solo con una capa oculta que tiene 200 neuronas. Además, el tiempo que tardo para realizar cada entrenamiento fue de 81 horas para el modelo de frecuencia de unigrama. Para el modelo de bigrama contiguo y bigrama con salto en 1 el tiempo de que tardo fue de 128 horas.

6.2 Aportaciones

Las aportaciones de esta tesis son las siguientes:

- Se construyó el corpus de Wikipedia 2017 en español.
- Se desarrolló el método para limpiar artículos de *Wikipedia*.
- Se desarrolló el método para representar oraciones con el modelo de frecuencia de n -gramas.
- Se demostró que la red neuronal *backpropagation* generaliza y asocia el modelo de frecuencia de n -gramas.
- Se demostró que la salida red neural *backpropagation* detecta el grado de aceptación de diferentes contextos y dominios.
- Se demostró que la red neuronal *backpropagation* clasifica nuevas palabras de diferentes idiomas, dominios y contextos.

6.3 Trabajo futuro

Con los resultados obtenidos en el desarrollo del presente trabajo, se genera como trabajo futuro:

- Realizar experimentos con un número mayor de oraciones.
- Usar diferentes corpus para entrenar la red neuronal *backpropagation*.
- Usar diferentes configuraciones en las capas de la red neuronal *backpropagation*.
- Obtener un número mayor de características para el modelo de frecuencia de n -gramas.
- Realizar experimentos con el aprendizaje de la red neuronal para detectar combinaciones de palabras poco frecuentes.
- Usar el aprendizaje de la red neuronal para diferentes problemas en el procesamiento del lenguaje natural.

Bibliografía

- Bolshakov, I. A., Galicia-Haro, S. N., & Gelbukh, A. (2005). Detection and Correction of Malapropisms in Spanish by means of Internet Search. (S. B. Heidelberg, Ed.) *International Conference on Text, Speech and Dialogue*, 115-122.
- Bolshakov, I., & Gelbukh, A. (2004). Computational linguistics models, resources, applications. (C. d. Computación, Ed.)
- Clark, A., Fox, C., & Lappin, S. (2013). *The Handbook of Computational Linguistics and Natural Language Processing*. John Wiley & Sons.
- Cruz, P. P. (2011). Inteligencia artificial con aplicaciones a la ingeniería. (Alfaomega, Ed.)
- David, K. (5 de 2007). A brief Introduction on Neural Networks.
- de Guevara, A. L. (1980). Consideraciones sobre el español actual. *18*, 5-61.
- Devís, A. (2006). El español en la red¿destrucción o reforma del lenguaje? A. Cancellier, M. Caterina, y L. Silvestri (Coords.). *Actas del XXI Congreso Aispi= Atti del XXII Convegno Aispi*, 71-88.
- Di Tullio, Á. (2005). Manual de gramática del español. (L. i. luna, Ed.) Buenos Aires.
- Fink, G. A. (2014). *Markov Models for Pattern Recognition: From Theory to Applications*. Springer Science & Business Media.
- García , F. H. (2012). Palabras problemáticas y frases incorrectas: una solución autónoma para detectar lo indetectable. *RAEL: revista electrónica de lingüística aplicada*(11), 41-55.
- García, A., Tapia Poyato, A., & A. M. (1997). El corrector ortográfico y la presentación del texto escrito. 375-412.
- García-Heras Muñoz, A. (2007). Programas informáticos de corrección gramatical en el aprendizaje de una lengua extranjera (inglés): expresión escrita. 14-15. (E. d.-L. Mancha, Ed.)
- Gelbukh, A. (2010). Procesamiento de lenguaje natural y sus aplicaciones. *1*, 6-11.
- Gelbukh, A., & Sidorov, G. (2010). *Procesamiento automático del español con enfoque en recursos léxicos grandes* (Segunda ed., Vol. 2). México.
- Hernández, M. B., & Gómez, J. M. (31 de 7 de 2013). Aplicaciones de Procesamiento de Lenguaje Natural. *32*.
- Kasabov, N. K. (1996). Foundations of neural networks, fuzzy systems, and knowledge engineering. (M. Alencar, Ed.)

- Lawley, J., & Martin, R. (2006). Corrector de gramática para estudiantes autodidactas de inglés como lengua extranjera. *340*, 1171-1191.
- Ledeneva, Y., & Sidorov, G. (2010). Recent advances in computational linguistics. *Informatica*, *34*(1).
- Lin, N. Y., Soe, K. M., & Thein, N. L. (2011). Developing a Chunk-based Grammar Checker for Translated English Sentences. 245-254.
- Liu, L., & Özsu, M. T. (Edits.). (2009). *Encyclopedia of Database Systems*. Springer US.
- Matich, D. J. (3 de 2001). Redes Neuronales: Conceptos básicos y aplicaciones. *Cátedra de Informática Aplicada a la Ingeniería de Procesos–Orientación I*.
- Melgar, R. L. (1987). La Real Academia Española: pasado, realidad presente y futuro. *67*(242), 327-346. (R. A. Española, Ed.)
- Michael, N. (2005). Artificial intelligence a guide to intelligent systems. Addison Wesley.
- Moré, J. (2006). *A grammar checker based on web searching* (Vol. 8). Digithum.
- Nazar, R., & Renau, I. (23 de 4 de 2012). Google books n-gram corpus used as a grammar checker. 27-34. (A. f. Linguistics, Ed.)
- Palma Cruz, D. L. (Diciembre de 20012). Uso de estrategias didácticas para la enseñanza de la ortografía (escritura de palabras) a partir de situaciones comunicativas concretas, en el cuarto grado de la escuela primaria de aplicación musical de San Pedro Sula. *Tesis de Maestría*, 66.
- Piantadosi, S. T. (1 de 10 de 2014). Zipf's word frequency law in natural language: A critical review and future directions. *21*, *5*, 1112-1130. (S. US, Ed.)
- Posteguillo, S. (2002). Influencia del inglés de internet en la lengua española. *Revista de Investigación Lingüística*, *5*(2), 117-139.
- Russell, S., & Norvig, P. (2004). Inteligencia Artificial Un Enfoque Moderno. 2.
- San Mateo, A. (3 de 2016). Un corpus de bigramas utilizado como corrector ortográfico y gramatical destinado a hablantes nativos de español. *49*(90), 94-118. (P. U. Valparaíso, Ed.) España.
- Sjöbergh, J. (2006). Chunking: an unsupervised method to find errors in text. *Proceedings of the 15th NODALIDA conference*, 180-185.
- Vásquez, A. C., Quispe, J. P., & Huayna, A. M. (2009). Procesamiento de lenguaje natural. *6*(2), 45-54.
- Wikipedia. (s.f.). *Wikipedia*. Recuperado el 09 de 07 de 2018, de <https://es.wikipedia.org/wiki/Wikipedia>