**Universidad de Colima**

**Facultad de Ingeniería Mecánica y Eléctrica**

**Ingeniería en Computación Inteligente**

**Procesamiento de Lenguaje Natural**

**Contenido**

**1.- Conceptos básicos**

1.1 El Procesamiento del Lenguaje Natural (PLN)

1.2 Niveles de Procesamiento del Lenguaje Natural

1.2.1 Nivel fonético

1.2.2 Nivel fonémico

1.2.3 morfológico

1.2.4 léxico

1.2.5 sintáctico

1.2.6 semántico

1.2.7 del discurso

1.2.8 pragmático.

**2. Representación del conocimiento**

2.1 Etapas de compilación del procesamiento del lenguaje natural

2.1.1 Análisis morfológico o léxico

2.1.2 Análisis sintáctico.

2.1.3 Análisis semántico

2.1.4 Análisis pragmático

2.2 Sistemas informáticos

**3. Procesamiento computacional de contenido lingüístico**

3.1 Técnicas de lingüística computacional

3.1.1 Reglas Lingüísticas y Análisis Morfosintácticos

3.1.2 Modelos de Lenguaje Pre-entrenados

3.1.3 Modelos de Machine Learning

**4. Aplicaciones del procesamiento de lenguaje natural**

4.1 Búsqueda avanzada de información

4.2 detección de entidades (personas, lugares, marcas u otros términos) con aprendizaje automático

4.3 Detección de tópicos, similitudes o anomalías en los textos

4.4 asistentes de voz o sistemas conversacionales

4.5 Clasificación automática de documentos y mensajes

4.6 Análisis de sentimiento y de la opinión

**Bibliografia**

La Siguiente Gran Revolución: NLP (Procesamiento del Lenguaje Natural), <https://www.youtube.com/watch?v=cTQiN9dewIg> , url: <https://youtu.be/cTQiN9dewIg>

CC6205 - Procesamiento de Lenguaje Natural: Introducción parte I, <https://www.youtube.com/watch?v=HEKTNOttGvU&list=PLppKo85eGXiXIh54H_qz48yHPHeNVJqBi> , urL: <https://youtu.be/HEKTNOttGvU?list=PLppKo85eGXiXIh54H_qz48yHPHeNVJqBi> ,

Investigar editores para programar paginas web

Nombre, versiones, tipo de lenguajes que trabaja, características

Hosting, tipos de hosting, características

Filezilla client

**1.- Conceptos básicos**

**1.1 El Procesamiento del Lenguaje Natural**

El procesamiento de lenguaje natural (NLP del idioma inglés Natural Language Processing) es una disciplina que trata la interacción entre la computadora y el lenguaje humano. El NLP está compuesto por un conjunto de técnicas computacionales que permiten el análisis y la representación de los textos. Esta área de investigación involucra diferentes disciplinas de las Ciencias de la Computación, como la Inteligencia Artificial y la Lingüística, y a partir de conceptos referidos a cada una de ellas se logra que una máquina pueda intentar comprender lenguaje natural e identificar su significado.

es el campo de conocimiento de la Inteligencia Artificial que se ocupa de investigar la manera de comunicar las máquinas con las personas mediante el uso de lenguas naturales, como el español, el inglés o el chino.

En la búsqueda de saber las posibilidades de que una computadora pueda comprender y manipular el lenguaje natural escrito u oral, de manera que se pueda hacer uso del mismo, se lograría saber aplicando **el Procesamiento de Lenguaje Natural** el cual incluye diferentes disciplinas de las Ciencias de la Computación, como la Inteligencia Artificial y la Lingüística, y a partir de conceptos referidos a cada una de ellas se logra que una máquina pueda intentar comprender lenguaje natural e identificar su significado.

Virtualmente, cualquier lengua humana puede ser tratada por los ordenadores. Lógicamente, limitaciones de interés económico o práctico hace que solo las lenguas más habladas o utilizadas en el mundo digital tengan aplicaciones en uso.

El Procesamiento del Lenguaje Natural (PLN) es un área de investigación que explora las posibilidades de una computadora para comprender y manipular el lenguaje natural escrito u oral, de manera que se pueda hacer uso del mismo [Manning1999, Kumar2011]. Esto se realiza con el objetivo de desarrollar técnicas y herramientas que permitan la implementación de sistemas capaces de interpretar y utilizar el lenguaje natural para desempeñar las tareas deseadas, como, por ejemplo, un clasificador de noticias, o un identificador de correo indeseado.

**El Procesamiento del Lenguaje Natural (PLN)** o Natural Language Processing (NLP) se centra en el análisis de las comunicaciones humanas y, en concreto, de su lenguaje.

Ante la gran cantidad de información en texto que generamos actualmente, surge la posibilidad de analizarla y aprovecharla. Las **técnicas de PLN** permiten **extraer** información disponible en cualquier sector.

Dentro de la Inteligencia Artificial, el PLN es la principal tarea de la lingüística computacional. Un análisis automático sobre cualquier tipo de texto permite clasificar, organizar, buscar o descubrir información no explícita, agilizando tareas que se realizan manualmente e identificando los elementos más relevantes de un escrito.

**1.2 Niveles de Procesamiento del Lenguaje Natural**

Las técnicas de NLP pueden ser organizadas en diferentes niveles de procesamiento. Cada uno de estos niveles representa un tipo de análisis que se debe efectuar al texto de entrada para extraer información específica.

Los diferentes niveles de procesamiento existentes son: Nivel fonético, Nivel fonémico, morfológico, léxico, sintáctico, semántico, del discurso y pragmático.

En un principio, el procesamiento del lenguaje natural se pensó como un modelo secuencial que repetía una y otra vez la misma secuencia de procesos para lograr la comprensión de la entrada. Desde el punto de vista psicolingüístico, se admite que el procesamiento del lenguaje es mucho más simple si se plantea una vista de niveles que pueden interactuar en diferentes órdenes. Partiendo del conocimiento existente sobre el lenguaje, un análisis se alimenta de los niveles más altos de procesamiento para asistir a los de más bajo nivel. Por ejemplo, el conocimiento pragmático de que un documento que se está leyendo habla de informática será utilizado cuando una palabra en particular tenga varios sentidos o significados. De esta manera, esta palabra se interpretará con un sentido “informático”.

a continuación, se dará una breve explicación de cada uno de los niveles que forman parte del procesamiento del lenguaje.

**Nivel fonético**

Se encarga de la interpretación del sonido dentro de las palabras.

**Nivel fonémico**

Se llama fonemas a las unidades teóricas básicas postuladas para estudiar el nivel fonológico de la lengua humana. Es decir, estudian la variación en la pronunciación cuando las palabras están conectadas. Estos niveles componen el análisis fonológico del lenguaje natural, realizado para los casos en los que la entrada es verbal/sonora.

**Nivel morfológico**

Se encarga de analizar la composición de las palabras. El análisis de este nivel consiste en determinar la forma, clase o categoría gramatical de cada palabra dentro de una oración. Teniendo en cuenta este nivel, un sistema NLP es capaz de desglosar una palabra y obtener el significado a través del significado de cada uno de sus morfemas. Por ejemplo, si se busca interpretar la palabra “Libros”, se obtiene que “Libr-” es el lexema, que “-o-” y “-s” son morfemas dependientes flexivos de masculino y plural.

**Nivel léxico**

Se encarga del significado individual de cada palabra. Para poder realizar el procesamiento léxico, se debe tratar cada palabra por sí misma, y etiquetarla con una parte del discurso dependiendo del contexto en la cual se encuentra. Cada palabra que compone un texto lleva asociada un conjunto de datos morfológicos, sintácticos y semánticos. El objetivo específico de este nivel es analizar cada una de estas palabras para saber su significado y función dentro de una oración. Por ejemplo, la palabra “lima” puede definirse como un derivado del verbo “limar”, pero también puede ser un sustantivo si se refiere al “fruto del limero”. Para poder determinar cuál es el rol de cada palabra en un texto en necesario resolver algunas ambigüedades que se presentan en este nivel como lo son la sinonimia, antonimia, entre otras. Por ejemplo**, la sinonimia** es la relación entre palabras diversas que comparten un mismo significado como puede ser cerca y próximo, suave y terso, dulce y azucarado. En el caso de **la antonimia**, se trata de una relación de opuestos entre dos palabras; cuando se presenta una antonimia se está haciendo un contraste o comparación entre dos palabras completamente contrarias como son cerca y lejos, suave y áspero, dulce y salado.

**Nivel sintáctico**

Se encarga de analizar la función de cada palabra dentro de una oración, descubriendo así la estructura u organización de la misma. El resultado de este procesamiento será una representación de la oración analizada que mostrará la relación entre las palabras que la conforman. Por ejemplo:

Se puede observar mediante el ejemplo anterior, como dentro del nivel sintáctico se identifican los componentes de una oración permitiendo conocer la función que cumple el mismo. La sintaxis de esta oración identifica dos artículos “el”, dos sustantivos como son “hombre” y “auto”, y un verbo “lava”. A partir de esta clasificación de palabras, el procesamiento permite determinar el sujeto y el predicado de la oración.

**Nivel semántico**

Se encarga de obtener el sentido de una oración a partir de la interacción entre las palabras que la conforman. El procesamiento semántico admite sólo un sentido a las palabras con varios significados, y así incluir el sentido en la representación semántica de la oración. Para ilustrar el análisis realizado en este nivel con un ejemplo, se plantea la desambiguación de la palabra “vela” teniendo en cuenta los siguientes dos significados:

Dos oraciones pueden contener esta palabra con cualquiera de los dos significados. Por ejemplo:

A partir del análisis realizado por este nivel, se puede determinar el significado que tiene la palabra “vela” en cada caso. En la frase n°1, se puede ver que la palabra “vela” está acompañada del nombre propio “San Pancracio”, que se refiere a un santo. Teniendo en cuenta esto, la definición que mejor aplica es la n°1, ya que son conceptos relacionados con la religión. En la oración n°2, se puede ver que se encuentra la palabra “barcos” acompañando a la palabra en cuestión. De esta manera, se determina que corresponde mejor con la definición n°2 de vela, ya que hace referencia a un objeto que impulsa una nave.

**Nivel de discurso**

Se encarga de trabajar con unidades de texto más grande que los niveles anteriores. Hace foco en el texto como un todo para alcanzar el significado haciendo conexiones entre las oraciones. Dos de los procedimientos que son realizados por este nivel son **la resolución de anáforas**, y el **reconocimiento de la estructura del texto**. La resolución de **anáforas** consiste en reemplazar pronombres con la entidad a la que hacen referencia. En este caso, el procedimiento resuelve que el pronombre “Él” de la segunda oración, hace referencia a “Juan”, presente en la oración anterior:

El reconocimiento de la estructura del discurso trata de identificar la función que cumple cada oración en el texto, sumando información al significado del texto completo. Por ejemplo, un artículo de diario puede estar compuesto por una iniciativa, una historia principal, eventos previos, una evaluación, y las expectativas. Conociendo la intención de cada uno de estos componentes del discurso, resulta más sencillo comprender la idea que se quiere transmitir.

**Nivel pragmático**

El nivel pragmático se encarga de analizar las diferentes variables relevantes para la comprensión de un texto o para explicar la elección de determinadas formas de llevarlo a cabo en función de los factores contextuales. Entre las variables se pueden mencionar: la situación, el emisor, el receptor, el enunciado, y en caso de tratarse de lenguaje verbal, el tono en el que se está expresando. Este nivel utiliza el contexto por encima de los contenidos del texto para la comprensión. Para comprender mejor la función del nivel pragmático, se tiene el siguiente ejemplo:

En este caso, se puede observar como la palabra “ellos” en la primera oración hacer referencia a los “concejales”, mientras que, en la segunda oración, referencia a los “manifestantes”. La única manera de poder determinar el sentido de esa palabra dentro del texto, es mediante el contexto. En ambos casos, a partir del conocimiento de cómo se desarrollan las manifestaciones políticas, se puede determinar a quién hacer referencia “ellos”.

**2. Representación del conocimiento**

**2.1 Etapas de compilación del procesamiento del lenguaje natural**

componentes del procesamiento del lenguaje natural. No todos los análisis que se describen se aplican en cualquier tarea de PLN, sino que depende del objetivo de la aplicación.

2.1.1 Análisis morfológico o léxico. Consiste en el análisis interno de las palabras que forman oraciones para extraer lemas, rasgos flexivos, unidades léxica compuestas. Es esencial para la información básica: categoría sintáctica y significado léxico.

2.1.2 Análisis sintáctico. Consiste en el análisis de la estructura de las oraciones de acuerdo con el modelo gramatical empleado (lógico o estadístico).

2.1.3 Análisis semántico. Proporciona la interpretación de las oraciones, una vez eliminadas las ambigüedades morfosintácticas.

2.1.4 Análisis pragmático. Incorpora el análisis del contexto de uso a la interpretación final. Aquí se incluye el tratamiento del lenguaje figurado (metáfora e ironía) como el conocimiento del mundo específico necesario para entender un texto especializado.

**2.2 Los sistemas informáticos de comprensión del lenguaje natural**

requieren una base de conocimiento provista de representaciones conceptuales que reflejen la estructura del sistema cognitivo de los seres humanos. Aunque la semántica superficial puede ser suficiente en algunas otras aplicaciones computacionales, la construcción de una base de conocimiento robusta garantiza su reutilización en la mayoría de las tareas de procesamiento del lenguaje natural. En este escenario, FunGramKB se presenta como una base de conocimiento multipropósito cuyo modelo ha sido diseñado de manera específica para tareas de comprensión del lenguaje natural.

Precisamente, uno de los elementos que han contribuido en forma notable al éxito de esta base de conocimiento ha sido el poder expresivo de su sistema notacional. El propósito de este artículo es describir la gramática, junto con su fundamentación teórica, del lenguaje de representación conceptual utilizado en FunGramKB.

**3. Procesamiento computacional de contenido lingüístico**

**3.1 Técnicas de lingüística computacional**

**3.1.1. Técnicas del Procesamiento de Lenguaje Natural**

Para lograr el procesamiento de lenguaje natural, existen un conjunto de técnicas mediante las cuales se extrae del texto información determinada. A continuación, se describirán algunas de las técnicas más comunes utilizadas por los diferentes sistemas NLP para procesar texto escrito en lenguaje natural. Cada técnica de NLP puede implicar uno o varios niveles de procesamiento explicados en la sección anterior.

**Detección de oraciones**

La detección de oraciones es una de las técnicas básicas correspondiente al nivel de procesamiento sintáctico. Si bien parece una tarea simple, detectar oraciones tiene ciertas dificultades a la hora de procesar títulos, abreviaturas, lista de elementos, y otros componentes que no siguen un patrón de texto plano. Esta técnica funciona recortando una secuencia de caracteres entre dos signos de puntuación. El signo debe estar acompañado por un espacio en blanco. Excluyendo el caso de la primera frase y en posibles ocasiones la última frase. Para determinar las abreviaciones en el texto se utiliza palabras cargadas en el modelo. Por esto la técnica utiliza un modelo por idioma, ya que tiene símbolos o abreviaturas necesarias para detectar las sentencias. En el siguiente ejemplo se puede observar la separación en oraciones de un fragmento de texto.

Se puede ver en el ejemplo la delimitación de oraciones dentro de un párrafo indicando el caracter de inicio y final. Un caso especial que se puede encontrar en el fragmento de texto es la abreviatura “Sr.”. El modelo en español determina que el punto que existe en “Sr.” es por una abreviatura a la palabra “Señor”, ignorando de esta manera el signo de puntuación como final de oración.

**Segmentación por palabras**

Una vez identificadas cada una de las oraciones que componen el texto, el siguiente paso es la segmentación por palabras, más conocida como analizador léxico o “Tokenizer”. Esta técnica pertenece al nivel léxico y consiste en la identificación de tokens, los cuales son unidades lingüísticas como palabras, puntuación, números, caracteres alfanuméricos, etc. Una forma de identificar tokens en idiomas modernos que utilizan un sistema de escritura basado en el Griego, como el Inglés y otros idiomas Europeos, se realiza delimitando espacios en blanco con límites de palabra, entre comillas, paréntesis y puntuación.

El trato con las abreviaciones es similar a la detección de oraciones, ya que no existen normas universalmente aceptadas para muchas abreviaturas y acrónimos. El enfoque más adoptado para el reconocimiento de abreviaturas es mantener una lista de palabras recortadas reconocidas.

Algunos ejemplos que pueden traer problemas son las direcciones de emails, palabras con apostrofes, URLs, ciudades, etc. Un ejemplo de esta problemática es el siguiente. Suponiendo que dentro de un fragmento de texto se encuentran los valores decimales 7.1 ó 82.4, al segmentar por cada valor específico se obtendrán los tokens “7”, “1”, “82”, y “4”. Este resultado no es el esperado si se trataba de extraer el valor decimal como un único token. Lo mismo ocurre para el valor “$2,023.74”, ya que el analizador admite que tanto el punto como la coma son delimitadores, y por lo tanto dividirá el número en tres partes. Asimismo, se debe tener en cuenta el idioma con el cual se está trabajando, ya que por ejemplo en idiomas como el Chino Mandarín, no se definen límites tan claros entre las palabras como en el Español o Inglés, y puede tornar esta tarea más compleja. Para ilustrar el funcionamiento de esta técnica, considere el siguiente ejemplo:

En el ejemplo se puede ver la separación por palabras indicada tanto por los espacios en blanco como por signos de puntuación. De esta manera, se obtiene el listado de palabras que componen el párrafo.

Etiquetado gramatical o Part-of-Speech(POS)-tagging

Una vez ejecutadas las dos técnicas previamente explicadas, se puede realizar el proceso de etiquetar las palabras según el rol que cumplen dentro de una oración. Este proceso de NLP se conoce como Etiquetado gramatical o Part-of-Speech(POS tagging), etiquetado de partes del discurso. Este proceso se encarga de asignar a cada una de las palabras de un texto su categoría gramatical de acuerdo a la definición de la misma o el contexto en que aparece, por ejemplo, sustantivo, adjetivo, adverbio, etc. Para ello es necesario establecer las relaciones de una palabra con sus adyacentes dentro de una frase o de un párrafo. Un mismo token puede tener múltiples etiquetas POS, pero solo una es válida dependiendo del contexto.

Son numerosos los sistemas que automatizan la asignación de partes del discurso ("tagging"). Muchos de ellos utilizan técnicas tales como modelos ocultos de Markov (Brants 2000), enfoque de máxima entropía (Ratnaparkhi 1996), y el aprendizaje basado en la transformación (Brill 1994).

Sin embargo, la gran mayoría de estos métodos utilizan la misma información para determinar las etiquetas POS, por lo que obtienen niveles de desempeño similares. Un ejemplo de esta técnica es el siguiente:

En primer lugar, las técnicas de etiquetado de partes del discurso poseen un conjunto de textos mediante la cual se entrenan los modelos de predicción. Luego, el etiquetador recibe un texto de entrada para analizar, y busca la secuencia de etiquetas más probable a partir del modelo matemático aprendido. Este punto depende exclusivamente de la técnica en sí, ya que utilizan cada una un modelo distinto.

**Segmentación morfológica**

Otra técnica de NLP muy utilizada en el procesamiento de texto, es la segmentación morfológica o en morfemas. Un **morfema** es el fragmento mínimo capaz de expresar el significado de una palabra, es decir, es la unidad significativa más pequeña de un idioma. Un morfema no es idéntico a la palabra, ya que este puede estar acompañado, por ejemplo, por prefijos o sufijos, mientras que una palabra, por definición, es independiente. Cuando una palabra se sostiene por sí misma, se considera una raíz porque tiene un significado propio y cuando depende de otro morfema para expresar una idea, es un afijo porque tiene una función gramatical. Cada palabra puede comprender uno o más morfemas.

**Los morfemas se clasifican en 2 categorías**. Los morfemas “independientes” admiten cierta libertad fonológica del lexema. En algunos casos pueden formar por sí solos una palabra:

Pronombres: cuíde-se, di-le, él, ella.

Preposiciones: desde, a, con, de.

Conjunciones: y, e, o, pero, aunque.

Determinantes: él, ella, ese, un, una.

Los morfemas “dependientes” van unidos o fusionados a otra unidad mínima dotada de significado, también conocidos como monema, para completar su significado. En ciertos casos provocan cambios de acento, cambios fonéticos en los fonemas adyacentes y sólo pueden aparecer en un orden secuencial concreto. Hay dos tipos de morfemas:

**Derivativos:** Estos morfemas son facultativos, es decir, añaden algunos matices al significado de los lexemas.

-Prefijos

-Sufijos

-Interfijos

**Flexivos**: Estos morfemas son constitutivos, es decir, señalan las relaciones gramaticales y sus accidentes entre los diferentes agentes de una acción verbal o una expresión nominal.

-Género

-Número

-Persona

-Modo y tiempo

La identificación de monemas permite el análisis en profundidad de una palabra dentro de un fragmento de texto, proporcionando información específica como puede ser el género, modo y tiempo, entre otras. Mediante el análisis realizado con esta técnica se puede ubicar de manera precisa cada palabra de cada oración.

**Eliminación de “Stop Words”**

La técnica “Stop Word” es utilizada para excluir palabras muy comunes que suelen tener poco valor para recuperar información que necesita el usuario. La cantidad de ocurrencias de una palabra en el texto determina si es o no una “stop word”, dado que cuanto más ocurrencias existan menos relevancia tiene en el texto. Dentro de este grupo se encuentran los artículos, los pronombres, las preposiciones, y las conjunciones. Esta técnica permite reducir el tamaño del texto para analizar, eliminando aproximadamente el 30% o 40% de dichas palabras. Además, se mejora la eficiencia, ya que la selección de palabras claves es más precisa. A continuación se plantea una identificación de “stop words” para poder observar de manera clara cuál es la ventaja de esta técnica.

**Detección de Stopwords.**

Esta detección de stop words está realizada con un listado en español que se encuentra en la Web1. Estos listados proveen un conjunto de palabras recurrentes en un idioma específico. La Figura 2.3, ilustra cómo a partir de un listado de palabras stopwords, estas son detectadas dentro del texto indicando que es posible su eliminación. En este caso en particular, al listado de palabras de uso común, se le agrega un conjunto de palabras propias del documento que se desea analizar. Esto se lleva a cabo mediante la medida numérica TF-IDF (Term Frequency - Inverse Document Frequency), que permite determinar que palabras son importantes para un documento dado de acuerdo a la frecuencia de aparición dentro del texto. Estas es una de las técnicas que se pueden utilizar para generar una lista de stop words, como así también lo es el Modelo de Espacio Vectorial (MSV) para determinar la relevancia de una palabra, entre otros.

Reconocimiento de Entidades Nombradas (NER)

Es una subtarea de la extracción de información que busca y clasifica elementos del texto que pertenecen a categorías predefinidas como pueden ser nombres de personas, entidades, organizaciones, lugares, expresiones temporales, cantidades, porcentajes, etc.

Para poder reconocer las diferentes entidades se utilizan una serie de aproximaciones. En primer lugar, algunas entidades simples se pueden reconocer mediante patrones codificados con expresiones regulares para encontrar entidades de fecha, tiempo, velocidad, etc. También suelen haber técnicas que utilizan una lista ordenada para reconocer nombres de personas, lugares, organizaciones, etc. Por último, existen reconocedores de entidades que utilizan un algoritmo de entropía máxima para clasificar cada uno de los tokens como un tipo de entidad particular en caso de que así sea.

Figura 2.4. Reconocimiento de entidades nombradas.

Como se puede ver en el ejemplo de la Figura 4, muestra el funcionamiento de una técnica de NER. En este caso se extrae la entidad “Truman” de un texto de entrada la cual permite realizar varias interpretaciones. En este caso “Truman” puede referirse a “Truman Capote”, “Harry S. Truman” o “Truman, Minnesota”. Para poder realizar el análisis es necesario tener una noción del contexto en el cual se encuentra esta entidad para poder determinar a qué se refiere. Dentro de las posibles entidades se realiza una asociación con los conceptos del contexto dentro de una base de datos de conocimiento. En este caso, “Harry S. Truman” se relaciona con el concepto de “Inauguración”, con la ciudad de “Potsdam” y con el apellido “Eisenhawer”, de manera que la entidad “Truman” hallada en el texto se asocia con esta entidad.

Stemming

Las palabras están morfológicamente estructuradas en prefijos, sufijos y una raíz. La técnica de Stemming busca un concepto de la palabra eliminando tanto prefijos como sufijos y obteniendo la raíz. De esta manera, se efectúa una reducción de la palabra a su mínimo elemento con significado. Un término que es reducido a su común denominador simplifica la recuperación de documentos cuyas palabras tenga la misma raíz. Por ejemplo:

Como se puede ver en este ejemplo, todos los términos derivan de la raíz “catalog”, haciendo posible la recuperación de ocho palabras que comparten una misma raíz como derivados con el mismo significado. Aun así, esta técnica no siempre funciona correctamente ya que hay palabras que poseen raíces compartidas por más de un significado, como se puede ver en la siguiente Tabla

Término con prefijo Raíz/Stem Término con el que causaría confusión Prevalencia valenc Valencia, valencia, valenciano, ambivalencia, polivalencia, Precatalogar catalog Descatalogar, catalogo,

Tabla 2.1.

Uno de los métodos más conocidos para llevar a cabo la reducción morfológica es el algoritmo de Martin Porter 2. También existe un lenguaje llamado Snowball, que permite el desarrollo de reglas para la extracción de stems de manera sencilla. Una vez que se tienen los algoritmos, estos son compilados por Snowball traduciendo el contenido a C o Java, permitiendo así incluir el análisis desarrollado en un proyecto desarrollado en dichos lenguajes.

**3.1.1 REGLAS LINGÜÍSTICAS Y ANÁLISIS MORFOSINTÁCTICOS**

El análisis se basa en exponer la estructura de relaciones entre las palabras del texto, y en diseñar reglas que permitan relacionar estas estructuras con rasgos como la temática o el tono del documento.

**3.1.2 MODELOS DE LENGUAJE PRE-ENTRENADOS**

Analizando repositorios de textos masivos de fuentes abiertas, se consiguen modelos del funcionamiento del lenguaje. Estos representan cada palabra mediante embeddings o vectores numéricos, que codifican su significado y función en las oraciones. En el caso de los modelos de lenguaje más modernos, estos embeddings son sensibles al contexto en el que se utiliza cada palabra, y se entrenan empleando corpus de miles de millones de palabras.

**3.1.3 MODELOS DE MACHINE LEARNING**

Aprenden a automatizar una tarea de análisis del texto a partir de corpus representativos de la misma. Estos pueden estar anotados, de forma que el sistema aprenda a reproducir el proceso de anotación, o no anotados, buscando entonces patrones y relaciones en el texto.

**4. Aplicaciones del procesamiento de lenguaje natural**

**4.1 Búsqueda avanzada de información**

El análisis de texto permite detectar y recuperar automáticamente información específica en documentos de texto libre de cualquier sector.

una aplicación para buscar información dentro de un documento sólo necesita obtener las palabras relevantes para generar un diccionario e indexar dicha información.

una aplicación que identifica la temática de un texto determinado, la cual requiere de un conjunto de técnicas de NLP especializadas para llevar a cabo su objetivo.

**4.2 detección de entidades**

Named-entity recognition (NER).

La detección de entidades (personas, lugares, marcas u otros términos) con aprendizaje automático es útil para detectar en qué contextos se mencionan determinadas palabras, por ejemplo, en documentos clínicos o legales.

**4.3 Detección de tópicos, similitudes o anomalías en los textos**

Con el análisis lingüístico, se detectan temas o patrones en la información, que nos indican ideas relevantes, relaciones, coincidencias o errores. Algo útil, por ejemplo, para la detección de plagio o el control de calidad de documentos.

**4.4 asistentes de voz o sistemas conversacionales**

Chatbots

El PLN es el primer paso en el desarrollo de los asistentes de voz o sistemas conversacionales, siendo esencial en la parte de comprensión del lenguaje.

**4.5 Clasificación automática de documentos y mensajes**

Se pueden etiquetar automáticamente textos según su temática u otras características. Es especialmente útil en ámbitos donde se maneja mucha información o se necesita hacerlo con rapidez, como el sector legal o el de la atención al cliente.

**4.6 Análisis de sentimiento y de la opinión**

Por las palabras que utilizamos, se pueden detectar opiniones acerca de un tema, una persona o un producto en publicaciones de redes sociales, comentarios de clientes o encuestas de clima.

Trabajos con procesamiento de lenguaje natural

1.- García-Reina, L. F. (2018). Asistente virtual de tipo ChatBot. Trabajo de Grado. Universidad Católica de Colombia. Facultad de Ingeniería. Programa de Ingeniería de Sistemas. Bogotá, Colombia

Resumen : El presente es un trabajo de grado con el objetivo de optar al título de Ingeniero de sistemas, el cual comprende el proceso de gestión de PQR´s adoptado por la corporación San Isidro – Colegio Anglo Americano, y el cual busca mediante las técnicas de inteligencia artificial reducir la intervención manual de este proceso, diseñando e implementando un prototipo inteligente capaz de categorizar los mensajes recibidos automáticamente, con esto se logra semi-automatizar el actual proceso. Esta herramienta se obtuvo a partir de la integración de técnicas y tecnologías como: Web services, algoritmos de aprendizaje automático, ChatBot, procesamiento de lenguaje natural, programación web, Bases de datos relacionales (MySQL) y no relacionales (RedisDB) e integración de API´s, mediante el uso del lenguaje de programación Python y una arquitectura de software del tipo Modelo Vista Controlador.

Palabras: GESTIÓN DE CALIDAD,APRENDIZAJE,ASISTENTE,AUTOMÁTICO, CHATBOT, GESTIÓN DE CALIDAD, PQR´S, PROCESAMIENTO DE LENGUAJE NATURAL, SVM, TELEGRAM, VIRTUAL

<https://repository.ucatolica.edu.co/handle/10983/17726>

2. toma de decisiones basadas en evidecnias

<https://www.gobiernodigital.gov.co/623/articles-74967_recurso_15.pdf>

definición de palabras

3. Búsqueda de imágenes basada en el Procesamiento de Lenguaje Natural

aplicada en una Agenda Visual orientada a Personas con Trastorno del Espectro

Autista <https://economicas.bdigital.uncu.edu.ar/objetos_digitales/13286/23-tecnologas-de-la-informacin-y-la-comunicacin-zorrilla-andrs-uni..pdf>

Bibliografia

Pagina sobre php con PLN

<https://programmerclick.com/article/84732095403/>

Corrector ortográfico

<http://www.revistasignos.cl/index.php/signos/article/view/134/40>

amazon usando PLN

<https://aws.amazon.com/es/comprehend/>