

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación



“TRADUCCIÓN AUTOMÁTICA DE PICTOGRAMAS A TEXTO”

Tesis para obtener el grado de
Maestra en Ciencias de la Computación

presenta:

Karen Leticia Vazquez Flores

Asesor de Tesis:

Dr. David Eduardo Pinto Avendaño

Co-asesor de Tesis:

Dra. Darnes Vilariño Ayala

Puebla
Junio 2019

Dedicatoria

*Con todo mi amor dedico esta tesis a mi Papá, Mamá y Hermana por la hermosa
vida que me brindan*

ustedes son la luz, la paz y la motivación de mis días.

*La dedico también a Ma. Susana Concepción Bonilla y Fuentes (Mamá Conchita) †
que desde el cielo iluminas y guías mi camino.*

Y principalmente a Dios.

Gracias por estar conmigo en todo momento y apoyarme en esta etapa

Agradecimientos

Agradezo al *Consejo Nacional de Ciencia y Tecnología* (CONACYT) por la beca otorgada a lo largo de la elaboración de la presente, del proyecto titulado “Traducción Automática de Pictogramas a Texto”.

Un gran agradecimiento a mi asesor de tesis *Dr. David Eduardo Pinto Avendaño* por todo el tiempo dedicado en este proyecto así como su paciencia, entrega y motivación para que yo sea una mejor persona y estudiante. A mi co-asesora la Dra. Darnes Vilariño Ayala y a la Dra. Maria Somodevilla García por la motivación a seguir adelante.

Principalmente mi mayor agradecimiento a mi papá, José Carlos Vazquez Hernandez por siempre enseñarme el sentido de la responsabilidad, por ser el mejor padre y hombre en el mundo, a mi mamá, Leticia Flores Bonilla por la gran lucha que hizo para vivir y estar aquí, gracias por enseñarme a nunca rendirme. Gracias también a mi hermana, Karla Susana Vazquez Flores por su paciencia, por ser mi amiga y sobre todo gracias por tu apoyo incondicional. Gracias a toda mi familia por su amor, comprensión, consejos y apoyo incondicional en cada momento de mi vida y durante mi desarrollo profesional.

Por último un agradecimiento especial a mi gran y mejor amigo Luis Gustavo Olivares Soto, gracias por todo tu apoyo y generosidad, sin ti mi vida y desarrollo profesional no sería igual.

Gracias a cada uno de ustedes por toda la paciencia y el amor brindado.

Karen Leticia Vazquez Flores GRACIAS!

Índice general

Dedicatoria	I
Agradecimientos	I
Lista de figuras	V
Lista de tablas	VI
Resumen	VII
1. Introducción	1
1.1. Planteamiento	1
1.2. Preguntas de investigación	2
1.2.1. Objetivos	2
1.3. Hipótesis	3
1.4. Impacto	3
1.5. Resultados esperados	3
1.6. Organización de la Tesis	4
2. Estado del arte	5
3. Marco teórico	9
3.1. Procesamiento de Lenguaje Natural	9
3.2. Pictogramas	10
3.3. Modelo de predicción	11
3.3.1. N-grams	11

3.4. Taxonomía	12
3.5. Etiquetador de partes de la oración	12
3.6. Lenguajes de programación	13
3.6.1. Python	13
3.6.2. Aplicación web	14
3.6.3. PHP	14
3.6.4. HTML	14
3.6.5. MySQL	15
4. Metodología de solución	16
4.0.1. Diseño de la metodología	16
5. Resultados experimentales	27
5.1. Conjunto de datos	27
5.2. Resultados	28
6. Conclusiones	35
Bibliografía	37

Índice de figuras

4.1. Proceso general para la solución.	16
4.2. Taxonomía representada gráficamente.	18
4.3. Subclases de la clase abierta.	18
4.4. Subclases de la clase cerrada.	19
4.5. Desarrollo del etiquetador propio representado gráficamente.	21
4.6. Arquitectura de algoritmo para el modelo de predicción.	24
4.7. Proceso para realizar la traducción de pictogramas a texto.	25
5.1. Pantalla principal que contiene las clases. (1-2)	28
5.2. Pantalla principal que contiene las clases. (2-2)	29
5.3. Interfaz que muestra las subclases dada la selección de la clase anterior.	29
5.4. Interfaz que muestra las subclases dada la selección de la subclase anterior.	30
5.5. Se muestra los pictogramas posibles dado el pictograma anterior. (1-4)	31
5.6. Muestra de los pictogramas posibles dado el pictograma anterior. (2-4)	31
5.7. Se muestra los pictogramas posibles dado el pictograma anterior. (3-4)	32
5.8. Muestra los pictogramas posibles dado el pictograma anterior. (4-4)	32
5.9. Obtenido con base a las selecciones anteriores.	33
5.10. Frases frecuentes que traduce cada usuario.	33
5.11. Pantalla para agregar pictogramas personalizados por cada usuario.	34

Índice de tablas

4.1. Cantidad de clases y subclases en la taxonomía.	17
4.2. Subclases de taxonomía.	21
4.3. Sub-Subclases de taxonomía.	22
4.4. Sub-sub-subclases de taxonomía.	23

Resumen

Los pictogramas son imágenes o signos que representan de manera más o menos realista un cierto significado. Su objetivo es traspasar la expresión lingüística para expresar figurativamente un objeto o un significado. Dado que los pictogramas son contruidos para ser comprendidos en a lo más tres miradas evitándose detalles superfluos, los mismos pueden ser utilizados como una forma primitiva de comunicación entre personas que tengan dificultades para entender o manejar un lenguaje más amplio o complejo como el lenguaje natural.

En este proyecto de tesis se abordó el problema de traducir pictogramas a texto, con la finalidad de apoyar a personas con problemas de desarrollo intelectual en su proceso de comunicación. La idea es crear un puente entre la expresión limitada de este tipo de personas para habilitar su expresión oral. Se diseñó un corpus basado en pictogramas que permita generar un modelo estadístico de traducción automática que dará la posibilidad de generar una oración textual a partir de una secuencia de pictogramas. Más aún, con la secuencia textual obtenida se alimentó un sintetizador de voz que genera entonces secuencias de voz asociadas a pictogramas de entrada.

Capítulo 1

Introducción

La comunicación es un proceso muy importante utilizado a diario, se lleva a cabo entre dos o mas personas y existen muchas formas de hacerlo, una forma de comunicación es mediante el uso de pictogramas, los cuales son imágenes que representan una frase, acción sentimientos, etc.

En esta tesis se presenta un traductor automático de pictogramas a texto enfocado a ayudar a personas que presentan alguna discapacidad en su proceso de comunicación y así crear un puente de comunicación con otras, para esto se utilizan técnicas de recuperación de información, procesamiento de lenguaje natural, modelo de predicción con bigramas y la creación de una Taxonomía para poder obtener resultados precisos al momento de la traducción de pictogramas a texto.

Existen varias investigaciones orientadas al proceso de comunicación con pictogramas y sus efectos en el ámbito educativo, sin embargo no se han desarrollado sistemas que sean de utilidad a estas personas.

1.1. Planteamiento

El problema que dio origen a este proyecto es la falta de apoyo a personas con problemas ó discapacidades de desarrollo intelectual que podrán beneficiarse de un sistema de traducción automático de pictogramas a texto.

Existen en la actualidad personas que presentan dificultades para comunicarse ya sea de manera oral y/o escrita y para las cuales sera de mucho beneficio el utilizar los

pictogramas (imágenes) con la finalidad de expresar ideas, conceptos, sentimientos, etc. Habilita a la persona a poseer una manera de interpretar, comprender y transformar su realidad en imágenes, para qué, a través de éstas exprese y transmita su pensamiento al interlocutor.

1.2. Preguntas de investigación

En esta sección se presentan las principales cuestiones que se realizaron para iniciar el desarrollo de esta investigación.

¿Cuáles son los problemas específicos que tiene una persona con discapacidad en su proceso de comunicación?

¿Ayudaría una taxonomía con las partes de la oración al proceso de traducción de pictogramas a texto?

¿Por que los Pictogramas pueden ayudar al proceso de comunicación?

¿Cuáles serian las limitaciones que se tiene al momento de usar pictogramas para el proceso de comunicación?

¿Qué efecto tiene los pictogramas en niños pequeños que presentan alguna discapacidad intelectual en la comunicación?

¿Ayudaría la traducción de pictogramas a texto a personas que tienen problemas para comunicarse?

1.2.1. Objetivos

Los objetivos se plantearon en objetivo general y objetivos específicos, se describen a continuación.

Objetivo general

- Diseñar y constriuir un modelo computacional para la traducción automática de pictogramas a texto.

Objetivos específicos

- Determinar el tipo de técnicas necesarias para la construcción de un modelo computacional para la traducción de pictogramas a texto.
- Construir un corpus de pictogramas con la calidad necesaria para que sea útil en el proceso de construcción del modelo de traducción propuesto.
- Analizar las características del corpus de pictogramas construido para determinar su uso en poblaciones objetivo, considerando niveles de utilidad basados en el tipo de morfología lingüística asociada a los pictogramas.
- Diseñar un conjunto de pruebas que evalúen el rendimiento del modelo de traducción construido.

1.3. Hipótesis

Con las preguntas planteadas en la sección anterior se define la siguiente hipótesis:

Es posible crear un modelo de lenguaje de predicción asociado a los pictogramas que nos permitan predecir cual es el siguiente pictogramas que utilizaría un usuario dado uno previo, con el fin de ayudar a personas que presenta alguna discapacidad intelectual en su proceso de comunicación a establecer un dialogo con aquellas que no lo presentan.

1.4. Impacto

El impacto principal de esta investigación es contribuir con aquella parte de la sociedad que no puede mantener una comunicación fluida debido a problemas de discapacidades intelectuales.

1.5. Resultados esperados

- Desarrollar un sistema adecuado en diseño y funcionalidad.
- Implementación correcta de una taxonomía de las partes de la oración.

- Creación y aplicación de un modelo de predicción automática para la formación de oraciones entendibles.

1.6. Organización de la Tesis

La presente tesis se encuentra organizada en capítulos para su mejor comprensión y se estructura de la siguiente manera.

- Capitulo 1 Introducción se describe el planteamiento del problema la hipótesis y los objetivos esperados.
- Capitulo 2 Se realiza un estudio del estado del arte en el cual se puede leer algunas investigaciones que se han realizado desde hace algunos años sobre el tema tratado “Traducción automática y modelo de traducción”.
- Capitulo 3 Se presenta un marco teórico para encontrar o entender el significado para unos temas o términos usados en este documento.
- Capitulo 4 Se exponen las propuestas de solución para el problema planteado.
- Capitulo 5 Se muestra el detalle de los resultados obtenidos con los enfoques propuestos.

Y finalmente se presentan las conclusiones obtenidas durante el desarrollo de esta tesis así como las referencias consultadas

Capítulo 2

Estado del arte

En los últimos años se ha analizado la forma de ayudar a persona con discapacidades en lectura o escritura, las investigaciones muestran que usar pictogramas en lugar de texto es una herramienta valiosa para ayudar a estas personas.

En el estudio presentado en [23] se trata de ayudar a personas con discapacidades intelectuales a través de un recurso en el que se vincula un conjunto de 5,710 pictogramas a conceptos léxico-semánticos en Cornetto, una base de datos similar a Wordnet para holandés. Se muestra un sistema de traducción de texto a pictografía, es decir, las palabras se convierten en pictografías solo si la palabra es igual al nombre del archivo.

También en [24] se genera lenguaje natural a partir de pictografías con un sistema para personas con discapacidades de desarrollo intelectual. El sistema traduce mensajes de conjuntos de uno o más pictogramas en idioma holandés usando enlaces de Wordnet y un modelo de lenguaje basado en ngramas. Así mismo, proporcionan varios métodos de introducción de pictogramas que ayudan a los usuarios a seleccionar los pictogramas apropiados.

En el caso de [7], desarrollaron una aplicación android de chat con pictogramas para personas con discapacidades intelectuales (PID). Esta aplicación utiliza una capa de accesibilidad para traducir texto escrito en pictogramas, lo cual permite que el PID cree mensajes usando pictogramas y cuidadores para escribir texto sin formato. En este trabajo, se han evaluado dos versiones tanto por los PID como por sus cuidadores y el concepto de enviar mensajes usando pictogramas les parece una

buena herramienta

En [22] se investiga cómo los objetos visuales pueden ayudar a niños con autismo; se aborda el uso de estrategias de instrucción para pequeños que pueden caracterizarse como aprendices visuales comenzando con una investigación en la que se supervisan las diferencias de estilos de aprendizaje asociadas con el autismo y concluye con la descripción de cómo la instrucción con claves visuales se puede aplicar en varios contextos.

Por otro lado, en [12] se aborda una ayuda interactiva para la comunicación expresiva con símbolos pictográficos en personas que no pueden adquirir habilidades en la comunicación del lenguaje natural pero que a menudo pueden encontrar un lenguaje simbólico para ser un medio de expresión alternativo. En la investigación descrita se discuten características del proceso de comunicación y una ayuda basada en computadora para la comunicación con unidades pictográficas que proporciona flexibilidad para la definición de símbolos por parte de usuario, mapeo entre símbolos y expresión multimodo. El sistema es una herramienta potencialmente muy poderosa para el desarrollo y la utilización de procesos de comunicación entre grupos de usuarios que no necesariamente comparten el acceso a una base común de conocimiento lingüístico.

En [8] se tratan a las imágenes como texto con pictogramas y pictografía retórica, es decir, proponen una retórica abstracta de las relaciones para las formas gráficas. Esta retórica proporciona la base fundamental para conceptualizar el diseño de la información en entornos visuales y textuales, no sobre la base de su contenido o forma específicos, sino a un nivel de estructura organizacional que por lo general no queda explícito. Su premisa es que el concepto de retórica pictográfica, o retórica metagráfica, es una forma útil de hacer explícito el orden abstracto de elementos en textos, imágenes y formas gráficas. Estos principios, críticos para los sistemas de escritura, pictográficos y de otro tipo, también subyacen en la estructura de gráficos y diagramas de varios tipos. Además, pueden usarse para hacer explícitos los efectos retóricos de la estructura de los textos inscritos o impresos en sus términos corporales y visuales. También se ha trabajado en la creación de herramientas automáticas para convertir pictogramas en textos en lenguaje natural para tres idiomas diferentes (holandés, inglés, español) y dos idiomas pictográficos (Beta y Sclera), los cuales están

actualmente involucrados en el proceso de traducción². Desarrollando sistemas para traducir de pictografías a textos, lo que hace que los sitios web y la corriente principal aplicaciones, como las redes sociales y el correo web, accesibles, la CCL contribuirá a la inclusión de las personas con discapacidad intelectual. Además, respaldar el texto escrito con imágenes simplificadas permite a las organizaciones y empresas llegar fácilmente a un gran grupo de personas que anteriormente no tenían o tenían muy poco acceso al contenido escrito digital.

En [16] se menciona que los pictogramas son una ayuda que puede usarse para facilitar la lectura y la comprensión de la información. Realizaron un estudio, cuyo objetivo fue evaluar si los niños entienden los pictogramas desarrollados por la Farmacopea de los Estados Unidos (USP)³. Y especialmente, si los pictogramas mejoran la comprensión de los niños de la información del prospecto del medicamento. Obteniendo como resultado que la mayoría de los niños entendieron correctamente los significados la proporción de las explicaciones correctas varía de 30 a 99pictograma. Incluso los pictogramas bien entendidos no ayudaban a los niños a comprender la información del folleto, aunque reducían la necesidad de sondeos. Este estudio muestra que el contexto en el que se prueban los pictogramas hace una diferencia en los resultados.

[21] Presenta un sistema para la predicción de palabras y expansiones, que permite al usuario abreviar o no cada palabra y así reducir la carga cognitiva requerida para su utilización, el sistema consta de tres bloques, diccionarios, modulo de predicción e interfaz de usuario. Implementan dos algoritmos para la predicción y además realizan una combinación de ambos para obtener los mejores resultados de su investigación.

En el caso de [17] cuantificaron el impacto de adoptar técnicas de deducción de orden superior que se basan en un mayor contexto de palabras, establecieron 21 condiciones experimentales mediante la combinación de tres ordenes ngram diferentes (unigrama, bigrama y trigramas) y utilizan el ahorro de pulsaciones de tecla como una medida única del rendimiento predictivo.

En el trabajo de [11] presentan la predicción de palabras que se basa en el aprendizaje de una representación para cada palabra en función de las palabras y los lingüísticos que se encuentran en su contexto, se utiliza SNW para la predicción de palabras, es un red dispersa de unidades lineales sobre un espacio de características

común predefinido o aprendido de forma incremental. Dada la tarea de predicción, la palabra representación se evalúa en ella y compite por la palabra más probable para completar la oración.

También en [20] describen los diferentes métodos de predicción de palabras en castellano que se utilizan en una aplicación para discapacitados motrices, empleando un modelo híbrido que conjuga métodos estadísticos con reglas declarativas y consiguiendo incrementar la velocidad en la escritura desde 2-3 palabras por minuto a 8-10. Presentan métodos estadísticos de predicción de palabras, chart parser extendido y como perspectiva inmediata se plantea el incremento del número de reglas.

En [2] se describe una propuesta para la arquitectura de un avanzado sistema de predicción de palabras, la arquitectura es modular y sumamente flexible, con bloques independientes e interfaces claras, de forma que permite utilizar fácilmente distintos algoritmos de predicción, e incluso predecir en varios idiomas. Para el método de predicción se basan en gramáticas probabilistas de palabras, el sistema puede predecir correctamente más del 94 % de las palabras del texto, y reducir más del 50 % el número de pulsaciones necesarias para escribir el texto.

Para la predicción de palabras en muchas investigaciones realizadas se utilizan modelos de lenguaje en [3] se expone el kit de herramientas de modelado de lenguaje estadístico de CMU para facilitar la construcción y prueba de los modelos de lenguaje de bigrama y trigramas, esta herramienta se puede utilizar para crear un vocabulario, construir un modelo de lenguaje, evaluar el modelo, el kit de herramientas representa un significado software para la eficiencia computacional.

Capítulo 3

Marco teórico

En este apartado se encuentra la descripción de conceptos utilizados a lo largo de la presente tesis. Se encuentran términos principales relacionados con el área de investigación para su desarrollo.

3.1. Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (*PLN*) es un área o campo de la inteligencia artificial que combina las tecnologías de la ciencia computacional con la lingüística aplicada.

Segun [19] el procesamiento de lenguaje natural es una subdisciplina de la inteligencia artificial y rama de la ingeniería lingüística computacional que tiene como razón principal construir sistemas y mecanismos que permitan la comunicación entre personas y maquinas por medio de lenguajes naturales; el *PLN* busca crear programas que puedan analizar, entender y generar lenguajes que los humanos utilizan habitualmente, de manera que el usuario pueda llegar a comunicarse con la maquina o computadora de la misma forma que lo haría con un ser humano. El objetivo principal de este campo es conseguir que las computadoras realicen tareas que impliquen el lenguaje humano, tareas como comunicacion humano-computadora, mejorando la comunicacion entre personas, o simplemente hacer uso del procesamiento de texto en el lenguaje [18].

El *PLN* consiste en la utilizacion de un lenguaje natural para comunicarnos con la

computadora, la cual debe entender las oraciones que le sean proporcionadas, el uso de estos lenguajes naturales, facilita el desarrollo de programas que realicen tareas relacionadas con el lenguaje o bien, desarrollar modelos que ayuden a comprender los mecanismos humanos relacionados con el lenguaje [29].

Se entiende, entonces que *PLN* es la habilidad que tiene la maquina para procesar la informacion transmitida, no simplemente de las letras o los sonidos del lenguaje, es decir, en este sentido por ejemplo un perico no es un animal parlante; así, una contestadora telefonica comun, una impresora, un procesador de palabras tampoco son dispositivos o *software* de *PLN*, mientras que un traductor automatico sin duda lo es [13].

El *PLN* es la parte de la Inteligencia Artificial (I.A.) que se define como: “La ciencia y la ingenieria de hacer maquinas inteligentes” [10].

3.2. Pictogramas

Las Tecnologías de la Información y las Comunicaciones (TICs) y la WWW deben ser una fuente de oportunidades para la integración, el aprendizaje, el empleo y no un conjunto de nuevas barreras que aumente la exclusión y la discriminación.

Un pictograma es una herramienta que permite la representación del lenguaje mediante dibujos, fotos e imágenes. Los símbolos gráficos abarcan desde sistemas muy sencillos basados en dibujos o fotografías hasta sistemas progresivamente más complejos como los sistemas pictográficos. Tambien resultan muy útiles también para las personas que no están alfabetizadas a causa de la edad o la discapacidad. Debe considerarse que en muchos casos las personas con discapacidades motoras y/o cognitivas, disponen de profesionales o acompañantes terapéuticos que los asisten y con los cuáles generalmente establecen y organizan estos sistemas de símbolos gráficos.[15].

Estos sistemas son aptos no sólo para personas con discapacidad intelectual o Trastornos del Espectro Autista (TEA), sino también por personas con discapacidades motoras como Parálisis Cerebral (PC), Esclerosis Lateral Amiotrófica (ELA), Edad Mental (EM)[14].

3.3. Modelo de predicción

La predicción de palabras consiste en intentar adivinar la o las palabras que un usuario está escribiendo, basándose en las letras que están ya escritas, y en una información adicional proveniente del texto creado previamente. Es uno de los métodos más comúnmente utilizados para ayudar, tanto en la comunicación personal como en la escritura, a personas con discapacidades motrices. Las ventajas que se obtienen de su uso dependen del grado de discapacidad, de su capacidad lingüística, etc. Las personas con problemas lingüísticos también pueden beneficiarse del uso de la predicción, ya que pueden irse guiando con las palabras predichas en su proceso de escritura, obteniendo un apoyo gramatical [20].

Los métodos de predicción, a partir del texto escrito por el usuario, proponen restricciones que deben cumplir las palabras siguientes (categoría gramatical y su probabilidad, concordancias, etc.). Los métodos de predicción disponibles están basados en secuencias de hasta 6 palabras (n-gramas), hasta 3 categorías (n-POS) y un analizador basado en una gramática independiente del contexto, cuya potencia ha sido aumentada de forma importante para soportar: gestión de probabilidades de reglas, ambigüedad (gramatical) de las palabras, posibilidad de que en la regla haya elementos (terminales o no terminales) opcionales, posibilidad de que los símbolos no terminales sean tanto categorías gramaticales como significantes o lemas (imponiendo las reglas de concordancia de rasgos adecuadas), posibilidad de prohibir un determinado significante o lema en una posición determinada de una regla, y un potente sistema de tratamiento de rasgos, que permite tanto controlar la concordancia entre los distintos símbolos (terminales y no terminales), como imponer o prohibir rasgos en cualquier símbolo de una regla [21].

3.3.1. N-grams

Un n-gram es una secuencia de N palabras: un 2-gram (o bigrama) es una secuencia de dos palabras como hola como”, tu tarea”, o tu turno” [?]. En este trabajo se usa n-gram para estimar la probabilidad de la siguiente palabra dada la palabra anterior. Una manera para estimar la probabilidad, es contar las frecuencias: tomar un corpus demasiado grande , contar el número de veces del n-gram y de la pala-

bra a estimar. Un modelo de bigramas, por ejemplo, aproxima la probabilidad de una palabra a todas las palabras anteriores $P(W_n|W_1^{n-1})$ usando solo la probabilidad condicional de una palabra anterior. En otras palabras, en lugar de calcular la probabilidad, se aproxima con la probabilidad

$$P(el|que) \quad (3.1)$$

Cuando se utiliza un modelo de bigrama para predecir la probabilidad condicional de la siguiente palabra, se realiza la siguiente aproximación:

$$P(W_n|W_1^{n-1}) \approx P(W_n|W_{n-1}) \quad (3.2)$$

Por ejemplo, para calcular una probabilidad de un bigrama particular de una palabra y dada una palabra anterior x , se calcula el recuento del bigrama $C(xy)$. La suma de todos los conteos de bigram que comienzan con una palabra dada w_{n-1} debe ser igual al conteo del unigram para la palabra w_{n-1} :

$$P(W_n|W_{n-1}) = \frac{C(W_{n-1}W_n)}{C(W_{n-1})} \quad (3.3)$$

3.4. Taxonomía

La Taxonomía es la ciencia que trata de los principios de clasificación y esta muy relacionada con la sistemática, se refiere a la clasificación o agrupamiento sistemático y los organismos y grupos a categorías. El propósito de la taxonomía es desarrollar un ordenamiento lógico de los organismos basándose en su afinidad natural [27].

3.5. Etiquetador de partes de la oración

Un etiquetador de parte de la oración es un sistema que utiliza el contexto para asignar partes de la oración.

El etiquetado automático de texto es un primer paso importante para descubrir la estructura lingüística de los corpus de texto grandes. La información de parte del discurso facilita el análisis de alto nivel, como reconocer frases nominales y otros

patrones en el texto [6].

Una etiqueta de parte del discurso es una categoría gramatical como verbos, sustantivos, adjetivos, adverbios, etc. El etiquetador de parte del discurso es una herramienta esencial en muchas aplicaciones de procesamiento de lenguaje natural, como la desambiguación de sentido de palabra, el análisis, la respuesta a preguntas y la traducción automática [28].

3.6. Lenguajes de programación

En cuanto al desarrollo de programación, para llevar acabo la solución propuesta se utiliza como lenguaje de programación Python y el desarrollo de software en HTML con ayuda de CSS y base de datos en MySQL. En los siguientes párrafos se define cada uno.

3.6.1. Python

Es un lenguaje de programación poderoso y fácil de aprender. Cuenta con estructuras de datos eficientes y de alto nivel y un enfoque simple pero efectivo a la programación orientada a objetos. Contiene una extensa biblioteca estándar en forma binaria y de código fuente para las principales plataformas, como también distribuciones y enlaces de muchos módulos libres de Python de terceros, programas y herramientas, y documentación adicional [26].

Cabe mencionar que Python es un lenguaje de programación simple, pero potente con una funcionalidad excelente para el procesamiento de datos lingüísticos [1], tiene una curva de aprendizaje superficial, su sintaxis y semántica son transparentes, y tiene una buena funcionalidad manipulando cadenas.

Natural Language Toolkit (NLTK)

NLTK citeloper2002nltk, es el kit de herramientas de lenguaje natural, es un conjunto de módulos de programa de código abierto, tutoriales y conjuntos de problemas. NLTK cubre el procesamiento del lenguaje natural simbólico y estadístico, es un kit de herramientas con distintos módulos que definen estructuras de datos e

interfaces para realizar áreas específicas del Procesamiento de Lenguaje Natural. En esta tesis se utiliza principalmente el modulo tagger, que define cada token (palabra) de un texto con información complementaria, con su parte de la oración o su etiqueta de sincronización de WordNet.

3.6.2. Aplicación web

Las aplicaciones web se han convertido en pocos años en complejos sistemas con interfaces de usuario [?]. El usuario interacciona con las aplicaciones web a través del navegador, el sistema se distribuye en tres componentes: el navegador, que representa la interfaz de usuario, la aplicación, que se encarga de realizar las operaciones necesarias según las acciones llevadas a cabo por éste y la base de datos, donde la información relacionada con la aplicación se hace persistente.

3.6.3. PHP

PHP es un lenguaje de programación usado generalmente para la creación de contenido para sitios Web. Se trata de un lenguaje interpretado usado para la creación de aplicaciones para servidores. Permite la conexión a diferentes tipos de servidores de bases de datos tales como MySQL, Postgres, Oracle, DB, Microsoft SQL Server, entre otros; lo cual permite la creación de Aplicaciones Web muy robustas. PHP también tiene la capacidad de ser ejecutado en la mayoría de los sistemas operativos tales como UNIX (y de ese tipo, como Linux), Windows y Mac OS X, y puede interactuar con los servidores de Web más populares ya que existe en versión CGI, módulo para Apache [5].

3.6.4. HTML

HTML es un lenguaje muy sencillo que permite describir hipertexto, es decir, texto presentado de forma estructurada y agradable, con enlaces (hyperlinks) que conducen a otros documentos o fuentes de información relacionadas, y con inserciones multimillonaria (gráficos, sonido...). La descripción se basa en especificar en el texto la estructura lógica del contenido (títulos, párrafos de texto normal, enumeraciones,

definiciones, citas, etc.) así como los diferentes efectos que se quieren dar (cursiva, negrita, o un gráfico determinado) [9].

3.6.5. MySQL

MySQL se presenta como una herramienta con altas prestaciones para el desarrollo de bases de datos, especialmente apropiados para ser usada por pequeñas organizaciones o empresas. La utilización conjunta de ambos: PHP y MySQL permite llegar a desarrollar interesantes aplicaciones web que puedan cubrir las necesidades de pequeñas empresas que quieran fortalecer su presencia en internet o usuarios individuales que quieran generar verdaderas paginas dinámicas[4].

Capítulo 4

Metodología de solución

Para la solución de esta tesis, se propuso y probaron varios métodos que permiten solucionar el problema de traducción. En este capítulo se pretende dar una descripción del sistema desarrollado, así como el proceso y las técnicas empleadas para alcanzar los objetivos de este trabajo.

En la figura 4.1 se muestra la estructura del algoritmo realizado para la solución.



Figura 4.1: Proceso general para la solución.

Como se muestra en la figura, la solución es dividida en diferentes procesos los cuales serán detallados a continuación.

4.0.1. Diseño de la metodología

Para el desarrollo de un modelo de predicción es necesario crear una taxonomía sobre las partes de la oración (*POS* por sus siglas en ingles) y su desarrollo consta de las siguientes fases.

Taxonomía

La taxonomía es el proceso más importante para la solución de esta tesis, fue creada con base a las partes de la oración (*POS* por sus siglas en inglés), describiendo una jerarquía de ellas. Para su desarrollo se realizó lo que sigue:

1. Búsqueda *POS*: Se localizaron las partes de la oración (*POS*) con una búsqueda exhaustiva y se sustrajeron las más importantes y utilizadas en el idioma inglés y español.
2. Clases Principales: De acuerdo a los resultados obtenidos de la búsqueda, se considero como clases principales la separación entre palabras abiertas y cerradas, teniendo así la **Clase abierta** y la **Clase cerrada**.
3. Subclases: Obtenidas las clases principales, se realiza otra clasificación, es decir, la clase abierta contiene 6 clases y cada una se subdivide en categorías más pequeñas, la clase cerrada contiene 5 clases y así mismo se realiza otra clasificación si el caso es necesario. También se considera el tipo de palabra: Normal o Concepto.

El resultado de la taxonomía se muestra en la figura 4.0.1, y en las figuras 4.0.1, 4.0.1 se muestra el detalle de las clases.

En la tabla 4.0.1 se presentan la jerarquía de la taxonomía.

Tabla 4.1: Cantidad de clases y subclases en la taxonomía.

Tipo	Cantidad
Clase	2
Sub-Clase	15
Sub-sub-clase	53
Sub-sub-sub-clase	3
Total	71

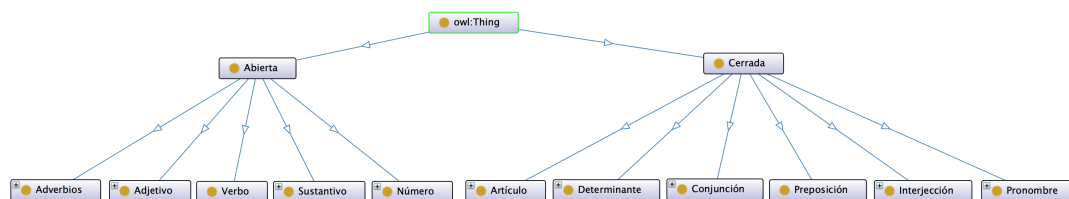


Figura 4.2: Taxonomía representada gráficamente.



Figura 4.3: Subclases de la clase abierta.



Figura 4.4: Subclases de la clase cerrada.

Desarrollo de etiquetador automático de palabras

Un etiquetador de palabras por partes de la oración es el paso más importante para alimentar la taxonomía y para la traducción de pictogramas a texto. En este trabajo se considero importante el desarrollo de un etiquetador propio, se decidió utilizar diccionarios del lenguaje español, como el Diccionario de la Lengua Española (RAE)¹, Diccionario online WordReference², recursos de ARASAAC³ y el etiqueta-

¹<http://www.rae.es>

²<https://www.wordreference.com>

³<http://www.arasaac.org>

dor automático de Natural Language Toolkit (NLTK por sus siglas en inglés)⁴.

1. Extracción de partes de la oración: Para este primer paso, se utilizó el diccionario RAE y WordPress, en los cuales, sus estructuras constan de palabra y definición, así mismo, la definición consta a su inicio de una abreviación que indica que parte de la oración puede ser la palabra, posterior a esto se encuentra su significado de la palabra. Se fue alimentando una base de datos en MySQL con palabras de los diccionarios para tener un mejor alcance y calidad de etiquetado.
2. Definición de abreviaturas obtenidas: Como siguiente paso, era necesario entender el significado de las abreviaturas, para lo cual se utilizó WordReference que además de contar con un diccionario de palabras, proporciona una lista de las abreviaciones de partes de la oración y el significado de cada una de ellas, lo cual fue de utilidad para traducir las abreviaciones y seguir al siguiente paso, un ejemplo a lo descrito anteriormente es:

adj. y s Que computa o efectúa operaciones matemáticas.

3. Etiquetar palabras: Es el paso final para este proceso, una vez reunidos todas las posibles etiquetas de acuerdo a los recursos mencionados, se etiquetaron todas los pictogramas con la ayuda de Python e involucrando la herramienta NLTK, además de la base de datos creada anteriormente, con el fin de obtener mejores resultados y que una sola palabra pudiera tener más de una etiqueta por ejemplo:

alta: -adjetivo -sustantivo femenino

En las siguientes tablas se representa la cantidad de pictogramas que hay por cada POS, precisamente en la tabla 4.2 se presentan las subclases principales, en la tabla 4.3 se detallan cada una de las subclases de acuerdo al nivel anterior y por ultimo en la tabla 4.4 trata de las sub-sub-sub clases.

⁴<https://www.nltk.org>

Tabla 4.2: Subclases de taxonomía.

Categorías	Cantidad
Adjetivo	1706
Adverbio	90
Artículo	10
Conjunción	8
Determinante	5
Número	177
Preposición	44
Pronombre	7
Sustantivo	295
Verbo	2456

En la figura 4.0.1 se presenta la arquitectura gráfica para alimentar la taxonomía dado el etiquetador.

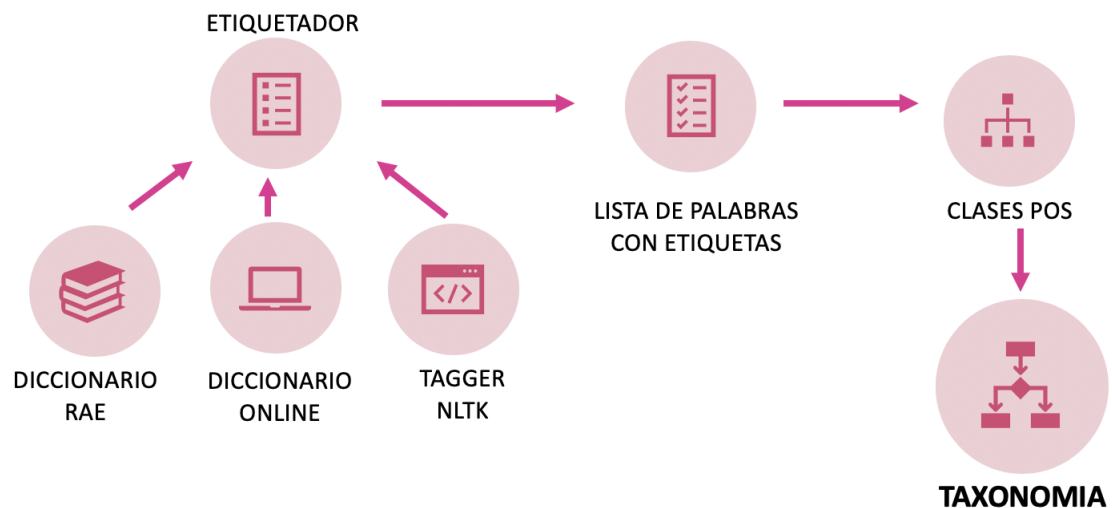


Figura 4.5: Desarrollo del etiquetador propio representado gráficamente.

Modelo de predicción

Para el desarrollo del modelo de predicción se utilizó probabilidad por bigramas, en la figura ?? se expone gráficamente los pasos del modelo y se describen a continuación.

Tabla 4.3: Sub-Subclases de taxonomía.

Sub-Clase	Cantidad
Adjetivo calificativo	57
Adjetivo comparativo	11
Adjetivo demostrativo	1
Adjetivo específico	16
Adjetivo explicativo	33
Adjetivo indefinido	7
Adjetivo numeral	25
Adjetivo ordinal	7
Adjetivo posesivo	4
Adverbio afirmativo	4
Adverbio calificativo	60
Adverbio cantidad	26
Adverbio comparativo	5
Adverbio distributivo	1
Adverbio duda	4
Adverbio exclamativo	1
Adverbio lugar	32
Adverbio modo	78
Adverbio negación	6
Adverbio orden	6
Adverbio tiempo	48
Artículo determinado	10
Artículo indeterminado	1
Conjunción adversativa	8
Conjunción causal	1
Conjunción consesiva	2
Conjunción continuativa	1
Conjunción coordinada	4
Conjunción copulativa	8
Conjunción distributiva	3
Conjunción disyuntiva	4
Conjunción ilativa	2
Conjunción subordinante	1
Conjunción temporal	3
Determinante demostrativo	12
Determinante exclamativo	2
Determinante indefinido	1
Determinante posesivo	11
Interjección	22
Número entero	49
Número decimal	0
Pronombre correlativo	1
Pronombre demostrativo	5
Pronombre exclamativo	3
Pronombre indefinido	22
Pronombre posesivo	12
Sustantivo común	160
Sustantivo concreto	1
Sustantivo femenino	3798
Sustantivo masculino	4779
Sustantivo propio	235

Tabla 4.4: Sub-sub-subclases de taxonomía.

Sub-sub-clase	Cantidad
Artículo determinado masculino	2
Artículo determinado neutro	2
Artículo determinado femenino	4
Total	8

■ Pre-procesamiento de texto

El procesamiento de texto es un paso importante en el desarrollo de la predicción de palabras, debido a que el texto que no sea pre-procesado cuidadosamente puede obtener malas predicciones y antes de generar un modelo es importante analizar el texto de forma correcta, en esta tesis se consideró un corpus de 185,000 mil noticias lo cual es bastante amplio y adecuado para el desarrollo del modelo, en esta fase se realizó lo siguiente:

- Eliminación de basura: Limpieza de caracteres especiales como lo son . , ‘!’ ·\$%()=?¿-¡¿; etc
 - Tokenización: Es un proceso en el cual se divide un texto por palabras.
- **N-gramas** Un N-gram es una secuencia de N caracteres o términos (tokens) consecutivos, con la previa eliminación de caracteres aislados, palabras vacías o stop words, stemmer y ordenación de términos, es importante destacar que para este modelo de predicción sólo se eliminaron los caracteres aislados y se utilizó bigramas.
- **Cálculo de probabilidad** Para el modelo de probabilidad basado en bigramas se usó la fórmula de estimación de probabilidad. La ecuación se muestra en ??
- Donde W_n es la palabra siguiente, W_{n-1} la palabra anterior. En la fórmula, la probabilidad varía entre 0 y +1, no se predice si es o no la siguiente palabra, si no, proporciona un grado de probabilidad y entre más se acerque a 1 mejor es la probabilidad con la palabra anterior y mejores resultados al momento de formar oraciones

Enseguida se muestra un ejemplo de aplicación.

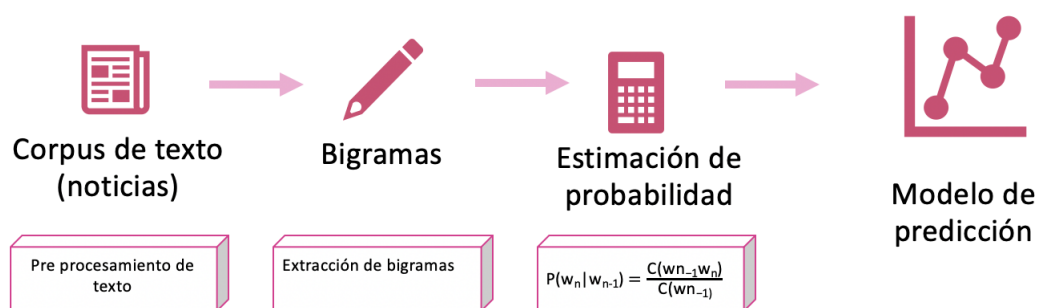


Figura 4.6: Arquitectura de algoritmo para el modelo de predicción.

Traducción

La traducción automática de pictogramas a texto es lo principal en esta investigación, se describe lo realizado a continuación.

Etiquetado de pictogramas

Preprocesamiento:

- Extracción: Se obtienen los títulos (nombres) de cada pictograma, pueden ser de una o más palabras.
- Depuración: Obtenidos los títulos se eliminó los signos de puntuación si era el caso.
- Tokenización: En caso de un título con más de una palabra se realizó una tokenización para obtener un mejor resultado en el etiquetado pos.
- Etiquetado POS: Cada nombre asociado al pictograma fue etiquetado con su parte de la oración utilizando el etiquetador descrito anteriormente. Es importante mencionar que cada pictograma puede tener más de una etiqueta.
- Asociación de pictograma con id y nombre: Se realizó una tabla en la base de datos previamente construida, que permitiera asociar cada pictograma con un id y su palabra o concepto representativo.
- Almacenamiento: Se alojaron los pictogramas en la base de datos de acuerdo a su id, nombre, definición y ruta de almacenamiento del pictograma.

En la figura 5.2 se presenta de forma gráfica la funcionalidad del modelo de predicción desarrollado.

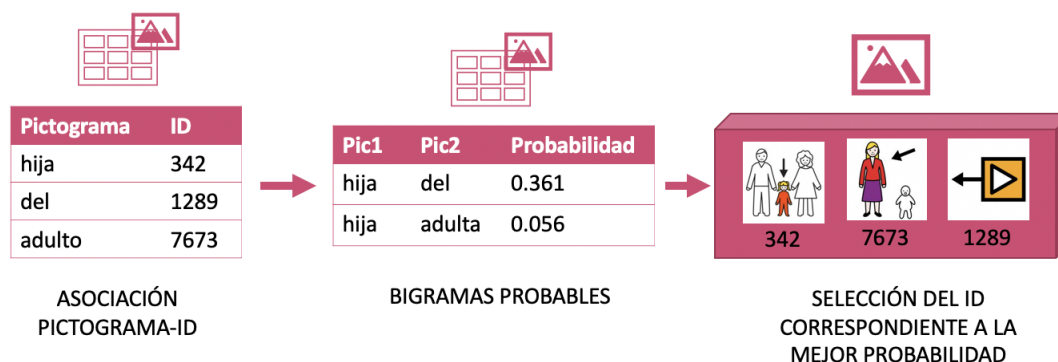


Figura 4.7: Proceso para realizar la traducción de pictogramas a texto.

Desarrollo de sistema

La traducción automática de pictogramas a texto con modelo de predicción se transportó a una sistema web desarrollado con html combinado con css y php que permite tener una interfaz amigable para un uso fácil y adecuado para cualquier usuario, cuya finalidad es que personas con discapacidades tengan un entorno para comunicarse.

El sistema aporta lo siguiente:

- Traducción de pictogramas a texto con predicción del siguiente pictograma, es la fase principal y la descrita anteriormente.
- Creación de usuario: Permite tener una cuenta personalizada para que cada usuario pueda modificarlo de acuerdo a sus necesidades.
- Aumento de pictogramas: A cada usuario se le permite agregar sus propios pictogramas o imágenes en su defecto, para incrementar su lenguaje y poder traducir mejor sus ideas obteniendo una mejor funcionalidad.
- Sección de traducciones frecuentes: El sistema permite mostrar las traducciones que cada usuario utiliza con más frecuencia con la finalidad de ahorrar tiempo.

-
- Selección de traducciones favoritas: El usuario puede guardar sus traducciones favoritas y traducirlas con mayor rapidez.
 - Voz: Si el usuario no logra entender la idea con los pictogramas mostrados, todos los pictogramas y las traducciones realizadas pueden ser traducidas a voz para un mejor entendimiento.

Capítulo 5

Resultados experimentales

En este capítulo, se exponen los resultados experimentales con la solución propuesta

5.1. Conjunto de datos

Un corpus [25] es un conjunto homogéneo de muestras de lengua de cualquier tipo (orales, escritos, literarios, coloquiales, etc.) los cuales se toman como modelo de un estado o nivel de lengua predeterminado. El conjunto de enunciados incluidos en un corpus, una vez analizados, debe permitir mejorar el conocimiento de las estructuras lingüísticas de la lengua que representan.

En este trabajo se utilizaron dos con gran cantidad de datos, cada uno de ellos con un fin en particular. A continuación se describen cada uno de ellos.

- Corpus de Pictogramas. Contiene más de 16,000 imágenes que representan números, letras, sentimientos, acciones etc. De los cuales aproximadamente 10,000 representan un concepto único, siendo este corpus la parte fundamental de esta investigación.
- Corpus de Noticias. Este conjunto de datos fue de gran utilidad para obtener el modelo de predicción anteriormente descrito, contiene más de 187,000 mil noticias ayudando así a obtener resultados óptimos al momento de formar oraciones.

5.2. Resultados

A continuación se presentan los resultados obtenidos al implementar la soluciones expuestas anteriormente, se presenta el resultado final como una aplicación web que permite traducir de manera exitosa pictogramas a texto formando oraciones o ideas coherentes ya que permite la predicción del siguiente pictograma dado la selección anterior.

En las imágenes 5.1 y 5.2 se muestra la pantalla principal al iniciar la sesiona de usuario presentando el listado de todas las clases que se encuentran en la taxonomía (Nivel 1). También en la 5.3 se presentan las subclases dadas la selección de la clase anterior en el ejemplo al seleccionar como clase principal sustantivo se despliegan las subclases de tal parte de la oración. Así mismo al continuar con la siguiente opción se muestran las sub clase de la sub clase anterior como se muestra en la figura 5.4.

En la 5.5 , 5.6 , 5.7 y 5.8 se detalla la continuidad del proceso de predicción dado el pictograma anterior con el fin de formar una oración lógica como aprecia en la figura 5.9.

Por ultimo se observa que la figura 5.10 y 5.11 la aplicación web tiene otros herramientas que facilitan el uso a los usuarios.

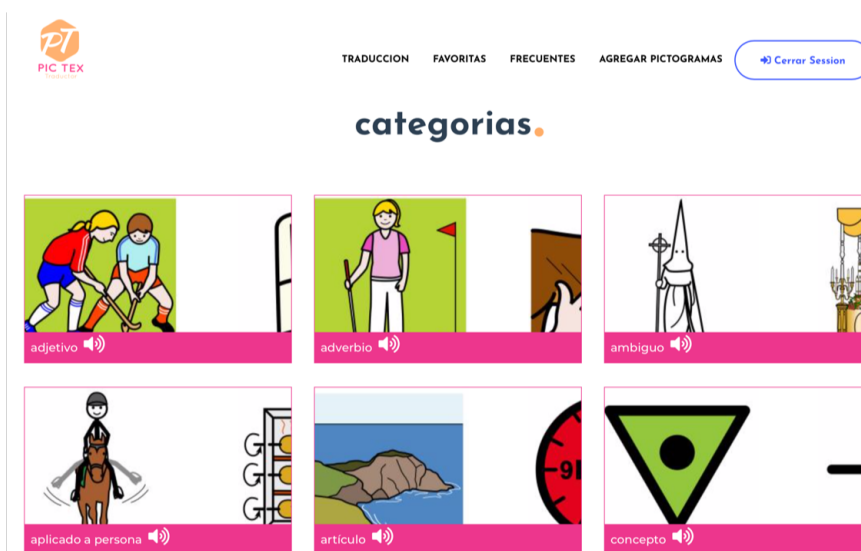


Figura 5.1: Pantalla principal que contiene las clases. (1-2)

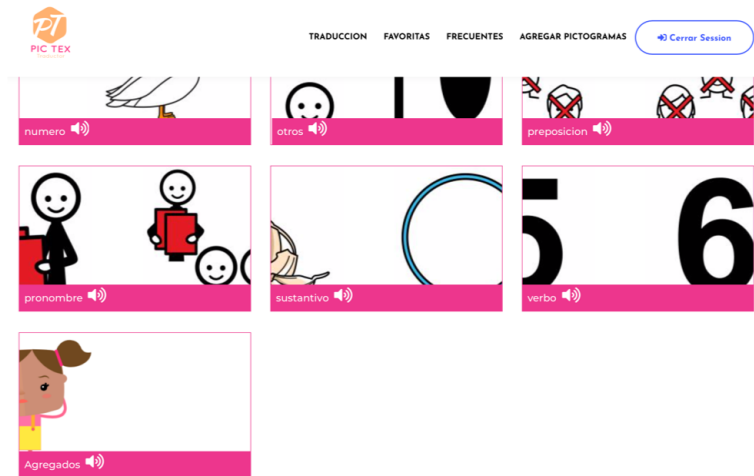


Figura 5.2: Pantalla principal que contiene las clases. (2-2)

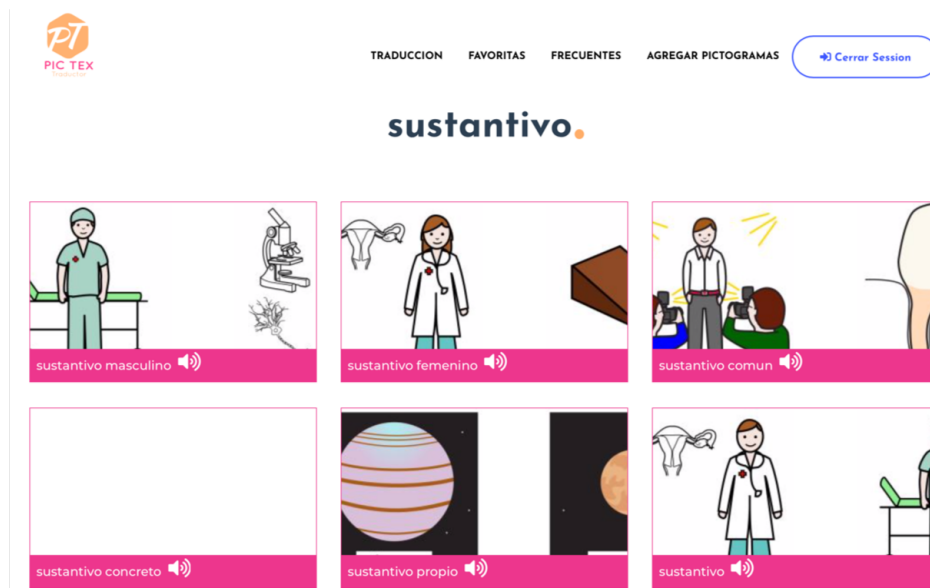


Figura 5.3: Interfaz que muestra las subclases dada la selección de la clase anterior.



Figura 5.4: Interfaz que muestra las subclases dada la selección de la subclase anterior.

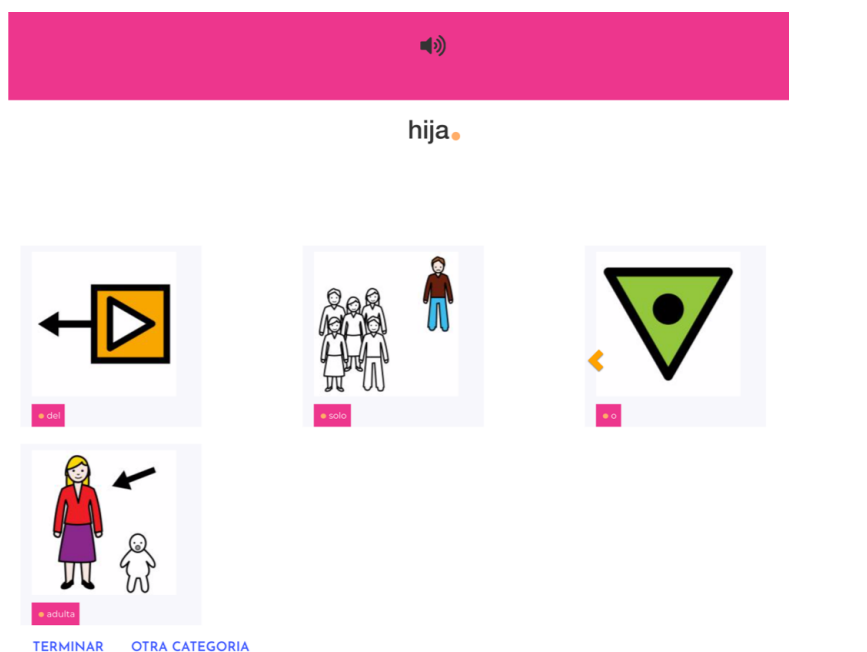


Figura 5.5: Se muestra los pictogramas posibles dado el pictograma anterior. (1-4)



Figura 5.6: Muestra de los pictogramas posibles dado el pictograma anterior. (2-4)



Figura 5.7: Se muestra los pictogramas posibles dado el pictograma anterior. (3-4)



Figura 5.8: Muestra los pictogramas posibles dado el pictograma anterior. (4-4)

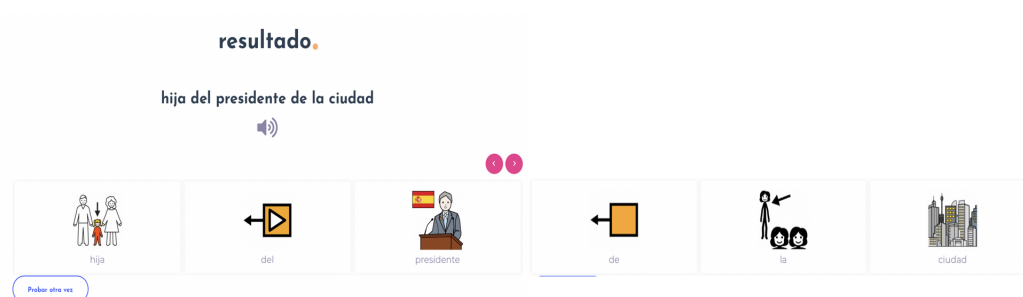


Figura 5.9: Obtenido con base a las selecciones anteriores.



Figura 5.10: Frases frecuentes que traduce cada usuario.



The screenshot shows a web form titled "agregar pictogramas." in a bold, dark blue font. The form is contained within a light gray border. It includes two text input fields: "Nombre del pictograma" and "Definición del pictograma". Below these is a dropdown menu labeled "Selecciona una categoría" with "concepto" selected. At the bottom left is a "Seleccionar archivo" button, and to its right is the text "ningún archivo seleccionado". A blue "Enviar" button is located at the bottom center of the form.

agregar pictogramas.

Nombre del pictograma

Definición del pictograma

Selecciona una categoría

concepto

Seleccionar archivo ningún archivo seleccionado

Enviar

Figura 5.11: Pantalla para agregar pictogramas personalizados por cada usuario.

Capítulo 6

Conclusiones

En esta tesis, se presenta la propuesta de solución al problema de traducción automática de pictogramas a texto con la implementación de un modelo de predicción, que permite elegir el próximo pictograma dado el seleccionado anteriormente, cumpliendo uno de los principales objetivos, el cual es ayudar a personas con alguna discapacidad a comunicarse.

Los objetivos inicialmente propuestos se cumplieron y se alcanzaron mejoras que permitieron lograr objetivos no planteados.

El algoritmo propuesto para la traducción de pictogramas a texto, incluye principalmente el desarrollo de una taxonomía por partes de la oración, la cual se realizó por medio de un etiquetador POS (*PartOfSpeech* por sus siglas en inglés), dicho etiquetador también fue desarrollado a lo largo de esta tesis con el fin de obtener mejor calidad de etiquetas por palabra y así mismo una mejor taxonomía. El etiquetador implementado (*KarGus* nombre dado al etiquetador) contiene gran variedad de palabras que pueden ser etiquetadas por más de una parte de la oración.

Se obtuvo y mejoro un corpus en español que cuenta con más de 15,000 pictogramas, de los cuales, 10,000 representan una palabra, es decir, una palabra puede ser representada por mas de un pictogramas, con la finalidad de entender mejor cada una.

Como se menciona anteriormente se implemento un modelo de predicción que permite elegir entre varias posibilidad el siguiente pictograma dado el anterior, ayudando así a obtener mejor estructura de oraciones para traducir, este modelo de

predicción se realizó por medio de un corpus de noticias masivamente amplio y utilizando bigramas y la formula de estimación de probabilidad.

Todo esto se desarrollo en el lenguaje de programación Python y para obtener una interfaz cómoda al usuario se utilizo HTML, MySql y PHP. Obteniendo así una aplicación web fácil y agradable, para el uso de personas con discapacidades y/o otras personas.

Como parte del trabajo a futuro y mejoras a la traducción, se considera la agrandar el corpus de pictogramas para obtener mas opciones de traducción, también desarrollar otras características para el modelo de predicción utilizando otros n-gramas con la finalidad de una predicción precisa o mas opciones de predicción. Se considera también poder realizar algunos cambios en el sistema que permitan una mejor navegación por personas con discapacidades y así ser mas utilizado.

Bibliografía

- [1] Steven Bird, Ewan Klein, y Edward Loper. *Natural language processing with Python*. “O’Reilly Media, Inc.”, 2009.
- [2] Sira Elena Palazuelos Cagigas, José Luis Martín Sánchez, Lisset Hierrezuelo Sabatela, y Javier Macías Guarasa. Hacia una arquitectura flexible para sistemas de predicción de palabras: propuesta de diseño y evaluación. *Procesamiento del Lenguaje Natural*, 35, 2005.
- [3] Philip Clarkson y Ronald Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. En *Fifth European Conference on Speech Communication and Technology*. 1997.
- [4] Ángel Cobo. *PHP y MySQL: Tecnología para el desarrollo de aplicaciones web*. Ediciones Díaz de Santos, 2005.
- [5] L.I. Anabell Comas. *JAVA o PHP*. Revista Digital Universitaria, 2015.
- [6] Doug Cutting, Julian Kupiec, Jan Pedersen, y Penelope Sibun. A practical part-of-speech tagger. En *Third Conference on Applied Natural Language Processing*. 1992.
- [7] Jo Daems, Nele Bosch, Steven Solberg, Jan Dekelver, y Marina Kultsova. Ablechat: Development of a chat app with pictograms for people with intellectual disabilities. *Engineering4Society 2016*, pág. 25, 2016.
- [8] Johanna Drucker y Jerome McGann. Images as the text: Pictographs and pictographic rhetoric. *Information design journal*, 10(2):95–106, 2000.

-
- [9] edu. *INTRODUCCION AL LENGUAJE HTML*. 2016. URL <http://www.eduxunta.gal>.
- [10] Andrea Esuli y Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss classification. En *Proceedings of the 14th ACM international conference on Information and knowledge management*, págs. 617–624. ACM, 2005.
- [11] Yair Even-Zohar y Dan Roth. A classification approach to word prediction. En *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, págs. 124–131. Association for Computational Linguistics, 2000.
- [12] MC Fairhurst y C Stephanidis. An interactive aid for expressive communication with pictographic symbols. *International journal of bio-medical computing*, 17(3-4):177–184, 1985.
- [13] Alexander Gelbukh. Procesamiento de lenguaje natural y sus aplicaciones. *Komputer Sapiens. Sociedad Mexicana de inteligencia aritificial*, 1, 2010.
- [14] España. Gobierno de Aragon. Arasaac (centro aragonees de tecnologias para la educacion). URL <<http://www.catedu.es/arasaac/aac.php>>.
- [15] Matías González, Hernán Sosa, y Adriana Elba Martin. Sistemas de comunicación no verbales. *Informes Científicos-Técnicos UNPA*, 6(2):30–56, 2014.
- [16] Katri Hämeen-Anttila, Kati Kemppainen, Hannes Enlund, J Bush Patricia, y Airaksinen Marja. Do pictograms improve children’s understanding of medicine leaflet information? *Patient education and counseling*, 55(3):371–378, 2004.
- [17] Gregory W Lesh, Bryan J Moulton, D Jeffery Higginbotham, et al. Effects of ngram order and training text size on word prediction. En *Proceedings of the RESNA’99 Annual Conference*, págs. 52–54. Citeseer, 1999.
- [18] James H Martin y Daniel Jurafsky. Speech and language processing. *International Edition*, 710, 2000.

-
- [19] Paula Andrea Benavides Ca non y Sandra Rodríguez Correa. Pln procesamiento del lenguaje natural en la recuperación de informción. 2007.
- [20] Sira Elena Palazuelos Cagigas, Rodrigo Mateos, José Lázaro, Juan Ignacio Godino Llorente, Francisco Aliaga García, Martín Sánchez, José Luis, y Santiago Aguilera Navarro. Predicción de palabras en castellano. *Procesamiento del lenguaje natural*, n^o 25 (septiembre 1999); pp. 151-158, 1999.
- [21] Sira Elena Palazuelos Cagigas, Martín Sánchez, José Luis, y Javier Macías Guarsa. Evaluación automática de un sistema híbrido de predicción de palabras y expansiones. *Procesamiento del lenguaje natural*, n^o 39 (sept. 2007), pp. 147-154, 2007.
- [22] Kathleen A Quill. Instructional considerations for young children with autism: The rationale for visually cued instruction. *Journal of autism and developmental disorders*, 27(6):697–714, 1997.
- [23] Leen Sevens, Gilles Jacobs, Vincent Vandeghinste, Ineke Schuurman, y Frank Van Eynde. Improving text-to-pictograph translation through word sense disambiguation. En *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, págs. 131–135. 2016.
- [24] Leen Sevens, Vincent Vandeghinste, y Frank Van Eynde. Improving the precision of synset links between cornetto and princeton wordnet. En *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, págs. 120–126. 2014.
- [25] Joan Torruella y Joaquim Llisterri. Diseño de corpus textuales y orales. *Filología e informática. Nuevas tecnologías en los estudios filológicos*, págs. 45–77, 1999.
- [26] Guido Van Rossum y Fred L Drake Jr. Tutorial python. *Disponible gratuitamente en python.org*, 2004.
- [27] Hugo Montoya Villafañe. Taxonomía: Clasificación de los seres vivos. *Revista Facultad de Odontología Universidad de Antioquia*, 8(2), 2016.

-
- [28] Alfian Farizki Wicaksono y Ayu Purwarianti. Hmm based part-of-speech tagger for bahasa indonesia. En *Fourth International MALINDO Workshop, Jakarta*. 2010.
- [29] Augusto Cortez Vasquez y Hugo Vega Huerta y Jaime Pariona Quispe y Ana Maria Huayna. Procesamiento de lenguaje natural. *Revista de investigacion de Sistemas e Informatica*, 6(2):45–54, 2014. ISSN 1816-3823.