Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Sistema de Traducción Automática Español-Náhuatl-Español

por

Rodolfo Alberto Martínez Torres

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

Septiembre 2021

Asesor:
Dr. David Eduardo Pinto Avendaño

Agradecimientos

A mi alma mater, la Benemérita Universidad Autónoma de Puebla, por brindarme nuevos conocimientos que me han hecho crecer profesionalmente.

Al Consejo Nacional de Ciencia y Tecnología le agradezco la beca recibida, lo que me permitió realizar estudios de posgrado.

A mi tutor David Eduardo Pinto Avendaño, por ser un gran investigador, guía, por su tiempo, por sus conocimientos. Pero sobre todo, por ser un gran profesor, excelente ser humano y un amigo.

A los miembros del jurado, por tomarse el tiempo en la corrección de tesis y examen de grado.

A todos por el interés, apoyo, observaciones y regaños, necesarios pare este trabajo de tesis. Agradezco tener ese privilegio.

A mi familia, que son mi faro, inspiración y pilar en la vida.

Resumen

El uso de sistemas computacionales para convertir un lenguaje a otro, es lo que denominamos como Traducción Automática, y esta forma parte importante del procesamiento del lenguaje natural. Además, la traducción automática estadística forma parte de la traducción automática, que usa el paradigma del aprendizaje automático para la traducción de textos. La traducción automática estadística hace uso de un modelo de lenguaje, un modelo de traducción y un decodificador.

En esta tesis se ha desarrollado un sistema web de traducción automática estadística español-náhuatl-español. Los desarrollos del modelo de lenguaje, modelo de traducción y el decodificador se realizan utilizando software opensource disponible en el entorno Linux. Se hace uso de KenLM para el modelo de lenguaje, GIZA++ y mkcls para el modelo de traducción y finalmente se hace uso de SMT Moses para la decodificación. Siendo el modelo de lenguaje el encargado de calcular la probabilidad de las frases del idioma de destino, el modelo de traducción ve la probabilidad de las frases de destino dada la frase de origen y el decodificador maximiza la probabilidad del texto traducido del idioma de destino.

En la formación del sistema se utilizó un corpus paralelo de 23,513 oraciones en español y náhuatl.

Índice general

Capítulo 1. Introducción

La gran diversidad que hay en México la hace una de las áreas más importantes del planeta, tanto en términos biológicos, (flora y fauna), como culturales (ideas, tradiciones, costumbres y lenguas), y dentro de esta diversidad (De Ávila Blomberg (2008)), se encuentra su amplia cultura lingüística. Por lo tanto, considerando lo último, y tomando el número de lenguas referenciadas en el *Ethnologue*[1] y para comparar la diversidad de lenguas existentes en México como con los diversos países de América, así como otras zonas geográficas del mundo, actualmente México tiene 291 lenguas vivas, 1,008 en América y 6,912 en el mundo; estas lenguas representan el 28.9 % y 4.2 % respectivamente.

En México hay una gran diversidad de clasificaciones en lo que a familias lingüísticas se refiere. Una de estas divide al territorio mexicano en 3 áreas, Norte, Centro y Sur, y una cuarta para el náhuatl. Siendo el náhuatl perteneciente a la familia yutonahua, que durante el imperio azteca tuvo su mayor apogeo, y mediante un sistema tributario y conquistas territoriales que tenían, lograron difundir su

[1] https://www.ethnologue.com

lengua a lo largo de todo el país (García Aymerich (2017)).

En particular, el Instituto Nacional de Estadística y Geografía (INEGI (2020)), en el estado de Puebla, identificó la existencia de 453,162 hablantes de dicha lengua, es por tal motivo, que en esta investigación se desarrolla e implementa un sistema para la traducción del español al náhuatl y viceversa usando modelos estadísticos de traducción basados en un corpus paralelo.

1.1. Definición del problema

Uno de los propósitos de esta investigación es ayudar a cerrar la brecha entre hablantes de náhuatl y aquellos que no la hablan, mediante el uso de una herramienta computacional; en particular para la variante que se utiliza en la zona del centro de Puebla.

En los últmos años, las lenguas indígenas han estado presentando problemas de marginación debido a que se enfrentan a una castellanización de su lengua (De la Luz Ramírez (2019)), y una forma de hacer frente a esto, es la de brindar herramientas computacionales para poder lograr la comunicación y así estar más cerca de otras culturas y zonas geográficas al interactuar con muchas personas mediante la interpretación del significado de un texto y su traducción en dicha lengua, en particular, del español al náhuatl y viceversa. Debido al problema antes mencionado, se tiene como consecuencia el desplazamiento de innumerables lenguas indígenas (IAMT (2019)).

La traducción automática se define como la traducción automatizada median-

te la cual se utiliza herramientas digitales para traducir un texto de un lenguaje natural a otro (Bhattacharyya (2015)), y para el caso particular de este trabajo de investigación se pretende obtener un sistema web de acceso libre, que dada una palabra/frase en la lengua náhuatl obtenga su correspondiente traducción en español o bien que la palabra/frase se encuentre en español y obtenga su correspondiente traducción al náhuatl.

1.1.1. Objetivos

Los objetivos, se presentan, primero el objetivo general y posteriormente los objetivos particulares que se plantearon:

Objetivo general: Diseñar e implementar un sistema computacional para la traducción automática español-nahuatl-español usando modelos estadísticos.

Objetivos particulares:

1. Seleccionar los tipos de técnicas necesarias en la construcción de un modelo computacional que ayude en la traducción automática entre dos idiomas.

2. Construir o reutilizar un corpus paralelo para las lenguas español y nahuatl y que sea eficiente en la construcción del modelo planteado de traducción.

3. Diseñar e implementar un sistema de traducción utilizando los modelos propuestos y el corpus .

4. Evaluar el rendimiento del modelo de traducción construido mediante un conjunto de pruebas.

1.1.2. Preguntas de investigación

Una vez identificados los problemas actuales sobre el desplazamiento de las lenguas indígenas, se establecen las siguientes preguntas de investigación:

- ¿Es posible construir un corpus paralelo del náhuatl del centro de Puebla, que sirva para la traducción automática?
- ¿Se tendrán técnicas para el desarrollo de un modelo de traducción para el náhuatl del centro de Puebla?

1.1.3. Hipótesis

Con base a las preguntas que se proponen, se establecen las siguientes hipótesis: *Crear un sistema traductor del náhuatl para su variante del centro del estado de Puebla, a partir de un corpus paralelo construido..*

1.1.4. Impacto

El modelo que se desarrollará e implementará, tendrá una serie de beneficios sociales importantes al poder usarse en las comunidades de personas que tienen en el náhuatl como su idioma nativo o de uso frecuente. Se espera un impacto dentro de la zona conurbada de la ciudad de Puebla, como lo son Cholula, Atlixco y La Resurrección. Se consideran al menos estas zonas ya que los pobladores mantienen una constante interacción entre personas hablantes y no hablantes del náhuatl.

1.1.5. Resultados esperados

- Un corpus paralelo para la traducción automática español-náhuatl-español.
- Un sistema computacional que sirva para la traducción de frases español-náhuatl-español utilizando el corpus paralelo.

1.2. Estructura de la tesis

Este trabajo de investigación se encuentra estructurado de la siguiente manera: En el capítulo II, se ofrece el marco teórico necesario para comprender el trabajo de investigación. En el capítulo III se hace un recorrido entre las diferentes trabajos que representan el estado del arte de las propuestas de traducción automática. El capítulo IV se muestra la metodología de solución del sistema de traducción automática estadística. En el capítulo V se proporcionan los resultados alcanzados. Finalmente, se tienen las conclusiones, trabajo a futuro y las referencias consultadas.

Capítulo 2. Marco Teórico

En esta sección se presentan los conceptos teóricos necesarios para la comprensión del estado del arte que se requieren dentro del presente trabajo de investigación.

2.1. Procesamiento del Lenguaje Natural

El *procesamiento del lenguaje natural* (PLN) es una rama que se encuentra dentro de la Inteligencia Artificial y que mediante el uso de *Lenguajes Naturales* permite la comunicación humano-maquina, esto a través del campo de la investigación (utilizados en la comunicación humana, ya sea oral, escrito o signado) (Martín Mateos and Ruiz Reina (2013)).

También forma parte de una rama de la ingeniería lingüística computacional; en donde se emplea tanto el manejo y definición de las reglas gramaticales de los lenguajes naturales y del uso de las computadoras (Moreira et al. (2021)). El PLN, en sí, son programas especialmente diseñados que un sistema recibe, y que permitan la comunicación en el propio lenguaje humano. En la actualidad el PLN ha crecido exponencialmene debido a que presenta diversas aplicaciones, tales como poder co-

rregir documentos, el uso de la traducción automática, sistemas de recuperación de la información, extracción y análisis de información y resúmenes, sistemas inteligentes educativos, búsqueda en textos, así como el apoyo en el análisis de sentimientos, entre otras cosas. El uso del PLN ha servido en el desarrollo de sistemas para el reconocimiento del habla o la corrección ortográfica en documentos (Benavides Cañón and Rodríguez Correa (2013)).

La capacidad de una computadora para lograr el procesamiento de información compleja, no solamente considerar los elementos por separado como letras o sonidos, es lo que se define como PLN. De esta manera, Gelbukh (2010) ejemplifica la definición antes mencionada como:

Un perico no es un animal parlante, a pesar de emitir palabras o frases; así, una contestadora de teléfono común, una impresora o un programa para procesar texto, aunque trabajan con letras o sonidos, tampoco son mecanismos o software de PLN, mientras que un traductor automático claro que lo es.

2.1.1. Sistemas de traducción automática

Los sistemas de traducción automática se dividen según Oliver Gonzalez (2014) en:

Traducción directa: Esta técnica de traducción tiene como base la consulta a diccionarios bilingües para decidir la traducción de las palabras o expresiones multipalabra del documento origen. En esta técnica se realizan los procesos de análisis y generación de textos los cuales suelen limitarse al análisis morfológico

y la lematización. Es importante mencionar que la lematización permite una búsqueda más fácil y rápida en el diccionario bilingüe. Una vez que que se ha generado una traducción palabra a palabra se aplican diferentes reglas gramaticales que sean necesarias, como por ejemplo, el cambio de orden de las palabras.

Sistemas de transferencia: A partir de un análisis sintáctico del texto original se genera una estructura considerando las reglas gramaticales del idioma al que se quiere traducir y de esta forma se obtiene la oración traducida. Es importante mencionar que el análisis sintáctico puede llegar a ser completo o solamente superficial.

Sistemas de interlingua: Esta forma de traducción transforma el texto original en una representación abstracta o intermedia, llamada interlingua, que es independiente del idioma, la traducción es generada directamente a partir de esta representación, permitiendo que se necesiten menos componentes para relacionar el idioma origen con su traducción en el idioma destino, así como para añadir una nuevo idioma. La parte de análisis y creación son generados con conocimiento de un solo idioma. Este tipo de sistemas permiten construir sistemas de traducción automática para dos idiomas totalmente diferentes.

El definir el idioma intermedio es una de las principales desventajas que presenta esta estrategia. Ya que éste debe ser abstracto e independiente de los idiomas origen y destino. Además, otro inconveniente es la complejidad que se tiene al obtener del texto original que información se requiere para poder ser

representada de manera intermedia; y así poder generar la traducción en el idioma destino considerando que pueden existir diferentes posibilidades de traducción. Teniendo en cuenta lo mencionado anteriormente, existe una pérdida de información al tener un idioma intermedio; ya que este proceso se lleva a cabo de manera secuencial, en lugar de realizar una traducción directa.

Sistemas estadísticos(*Statistical Machine Translation*): Para realizar una traducción mediante modelos estadísticos se utilizan corpus bilingües los cuales son analizados, para lo cual se requiere de una gran número de textos originales con su traducción que sirven de parámetros evaluados estadísticamente y permiten de esa manera su traducción.

Para trabajar los sistemas de traducción estadística se manejan las dos siguientes dos probabilidades:

- Considerar una mejor probabilidad de que dada una palabra o serie de palabras en el idioma origen se obtenga su traducción en una o más palabras en el idioma destino.
- La probabilidad de la validez de una cadena traducida.

Es decir, se calcula la probabilidad de que una oración en el idioma a traducir (t-target) sea la traducción de una frase en el idioma de origen (s-source), esto se escribe como $p(t|s)$. Esta misma probabilidad se calcula como:

$$p(t|s) = p(s|t)p(t)$$

donde:

- $p(s|t)$ es el modelo de traducción que define la probabilidad de que el texto traducido sea generado por la oración en el idioma de entrada.
- $p(t)$ es el modelo del idioma y define con que probabilidad una oración traducida sea correcta en ese idioma.

Dividiendo el problema en dos subproblemas; como se consideran los parámetros para deducir las traducciones, el sistema estadístico genera un número enorme de las mismas, por lo que se calcula aquella traducción con la mayor probabilidad:

$$argmax\ p(t|s) = argmax\ p(s|t)p(t)$$

Es decir, encontrar la máxima probabilidad no es factible realizarla calculando todas las probabilidades de las opciones de traducción, esto ya que el número de opciones de traducciones es elevado, por lo que el tiempo de ejecución sería grande. Debido a lo anterior, partiendo de métodos heurísticos se puede hacer una selección de la mejor opción de traducción, esto sin perder la calidad de los resultados y reduciendo el espacio de la búsqueda de opciones. Con estas heurísticas se obtiene una de las traducciones con una alta probabilidad, sin garantizar que sea la mayor, este proceso se lleva a cabo en el módulo *decodificador*.

En los sistemas usados para la traducción automática basados en estadísticas no es requerido que se generen reglas lingüísticas que conllevan a un esfuerzo humano y que pueden llegar a empobrecer el sistema. Además dichos sistemas

pueden entrenarse y utilizarse para distintas lenguas, ya que al no generar reglas lingüísticas los módulos son independientes de los idiomas. El único problema que este tipo de sistemas presenta, es cuando no se tienen corpus bilingües que sean lo suficientemente completos para que el sistema estadístico pueda ser entrenado y considerando que la generación del corpus puede llegar a ser muy costoso.

La clasificación de los principales sistemas de traducción automática estadística son los siguientes:

- **Traducción a partir de palabras**: Este sistema esta basado en la traducción léxica, donde, cada palabra se traduce de manera independiente del resto de las palabras. Por eso es necesario contar con un diccionario bilingüe entre las lenguas origen y destino.
- **Traducción a partir de frases**: Este modelo hace uso de traducciones de pequeños conjuntos/grupos de palabras, los cuales no necesariamente deben de completar una idea; entre los modelos de traducción automática estadística, este es el que obtiene mejores resultados.
- **Traducción a partir de sintaxis**: El modelo a partir de sintaxis hace uso de traducciones que poseen estructuras de las oraciones. Estas traducciones son llevadas a cabo mediante el uso de *parsers*[1] potentes que nos dan el análisis de la secuencia gramatical de la traducción.
- **Modelos de traducción factorizados**:Estos modelos hacen uso de fra-

[1]http://www.alegsa.com.ar/Dic/parser.php

ses que son separadas en *tokens*[2] dentro de un vector, el cual tiene como finalidad ordenar y contener por niveles los datos de la forma, el lema, la estructura gramatical y la morfología o semántica de estas frases.

Sistemas de traducción automática basada en ejemplos: A partir de una serie de traducciones previamente dadas, se puede obtener una nueva traducción que sea similar; se puede decir que es un repositorio de traducciones. Cuando se tiene una traducción asistida, a partir de una frase similar, se genera una nueva traducción producida por un traductor humano; de manera análoga para el caso de la traducción automática, la traducción se lleva a cabo de la misma manera, dada una oración similar, el sistema se encarga de generar una nueva traducción.

Sistemas híbridos

Es cuando la traducción se ejecuta en paralelo por varios sistemas de traducción automática. De las traducciones generadas por los diferentes sistemas y mediante el uso de técnicas estadísticas se puede escoger la mejor traducción que generó el sistema híbrido. O bien, en algunos sistemas la traducción se obtiene a través de la unión de las mejores salidas generadas por el sistema híbrido. Este método es usado regularmente para unir las traducciones generadas por sistemas basados en reglas con sistemas estadísticos.

[2]http://www.alegsa.com.ar/Dic/token.php

2.1.2. Evaluación de la traducción automática

La calidad de la traducción generada por los sistemas de traducción automática juegan un papel importante, ya que la idea central del texto debe de ser la misma de origen a la de destino, evitando añadir o eliminar conceptos. El proceso para medir la calidad se le conoce como *evaluación de la traducción automática*. En donde se definen métodos que cuentan el número de errores basados en palabras y/o frases dependiendo del método de traducción utilizado.

2.1.2.1. Manual

Dentro de la evaluación de la traducción automática, la *evaluación humana* o manual, es la opción más común para juzgar y medir la calidad de la traducción, ya que está realizada por expertos tanto del idioma origen como destino. Esta evaluación es juzgada por expertos en traducción y lingüística desde dos perspectivas diferentes. La primera perspectiva es el grado de adhesión al texto destino y a las normas del idioma destino, refiriéndose, por ejemplo, a las características como la gramática y la claridad. Esta perspectiva de evaluación de la calidad se conoce como fluidez. Los evaluadores sólo tienen acceso a la traducción que se está juzgando y no a los datos de origen. La fluidez requiere que un experto hable con fluidez sólo el idioma de destino. La segunda perspectiva, la adhesión al texto de origen la cual se juzga en función de las normas y el significado del texto de origen, en términos de lo bien que el texto de destino representa el contenido informativo del texto de origen. Esto se conoce como exactitud. Los evaluadores tienen acceso al texto fuente y a las

traducciones que se están juzgando. Con frecuencia, también se tiene en cuenta el contexto de una oración. Los evaluadores deben ser bilingües tanto en las lenguas origen y destino.

Esta forma de evaluación lleva mucho tiempo, ya que, es inherentemente subjetiva, y para disminuir el problema de la subjetividad, generalmente se pide a más expertos que evalúen las traducciones en el mismo conjunto de evaluación, y sus evaluaciones, finalmente, están justificadas estadísticamente, pero esto hace que sea costosa.

2.1.2.2. Automática

Los sistemas de traducción automática mejoran a medida que los recursos crecen y los errores se arreglan, por lo que la evaluación debe repetirse muchas veces, y aparte, las métricas de evaluación automática son alternativas gratuitas a la evaluación humana. Estas métricas evalúan el resultado de los sistemas de traducción automática comparándolo con traducciones de referencia. Las métricas de evaluación proporcionan puntuaciones de evaluación basadas en la traducción de referencia más similar Maučec and Donaj (2019).

A continuación se muestran alguna métricas automáticas para evaluar la traducción automática.

BLEU Esta métrica (Bilingual Evaluation Understudy) fue una de las primeras métricas en reportar una alta correlación con las evaluaciones humanas. Bleu es actualmente referencia bibliográfica obligada en el campo de la evaluación

de la traducción automática.

La idea central detrás de la métrica es que "*cuanto más cerca esté una traducción automática de una traducción humana profesional, mejor será*". La métrica calcula puntuaciones para segmentos individuales, generalmente frases, entonces promedia estas puntuaciones sobre todo el corpus para una puntuación final. Y para ello utiliza una forma de precisión modificada para comparar una traducción candidata con una o múltiples traducciones de referencia. Papineni et al. (2002).

En varias referencias, por ejemplo Denkowski and Lavie (2010) y/o KantanMT (2020), citan que, el *score* obtenido por Bleu esta en un rango de 0 a 1, y que, *scores* arriba de `0.25` generalmente reflejan traducciones entendibles. Y para *scores* que sobrepasan los `0.50` hacen referencia a muy buenas traducciones fluidas, listas para ser publicadas como referencias de la lengua.

LEPOR Esta métrica surge como la combinación de muchos factores de evaluación, incluidos los ya existentes y los modificados. LEPOR está diseñado con los factores de penalización por longitud, precisión, penalización por orden de palabras de *n-gramas* y memoria. La penalización de longitud mejorada asegura que la traducción de la hipótesis, sea castigada si es más larga o más corta que la traducción de referencia. La puntuación de precisión refleja la exactitud de la traducción de la hipótesis. La puntuación de recuerdo refleja la lealtad de la traducción de la hipótesis a la traducción de referencia o al idioma de origen. El factor de penalización del orden de las palabras basado en *n-gramas*

está diseñado para los diferentes órdenes de posición entre la traducción de la hipótesis y la traducción de referencia Han (2017).

METEOR La métrica se basa en la media armónica ponderada de la precisión de *unigrama* y la memoria de *unigrama*, y así resolver algunas de las deficiencias inherentes a la métrica del *bleu*. También incluye la coincidencia de sinónimos, donde en lugar de coincidir sólo en la forma exacta de la palabra, la métrica también coincide en los sinónimos, característica que no se encuentran en otras métricas. La métrica *meteor* lematiza[3] las palabras y coincide en las formas lematizadas. La aplicación de la métrica es modular, en la medida en que los algoritmos que hacen coincidir las palabras se aplican como módulos, y pueden añadirse fácilmente nuevos módulos que apliquen diferentes estrategias de emparejamiento Banerjee and Lavie (2005).

TER *Translation Error Rate* (TER) es una métrica de error que calcula el número de modificaciones que un experto tendrá que realizar para cambiar la traducción destino de un sistema de traducción automática, de tal manera que esta coincida con la traducción origen. Esta herramienta fue diseñada para correlacionar con los juicios humanos, esta métrica se basa en la distancia de *Levenshtein* como un paso de edición. Ésta métrica sufre de una deficiencia, la salida del sistema de traducción puede ser una traducción aceptable, pero diferente de la traducción de referencia, por lo tanto, esta métrica marcará error en la traducción. Snover et al. (2006).

[3]Relacionar una palabra flexionada o derivada con su forma canónica o lema.

WER *Word Error Rated* (WER) es una métrica basada en la distancia de *Levenshtein*, a nivel de palabra. Obtiene el número de operaciones mínimas que se deben de hacer para obtener la traducción de referencia a partir de la traducción generada, esto mediante sustituciones, inserciones y eliminaciones. En el caso de que se proporcionen varias traducciones de referencia para una oración origen, se calcula la distancia mínima a este conjunto de referencias Nießen et al. (2000).

2.1.3. Modelos

Los sistemas de traducción basados en frases utilizan modelos de entrenamiento para alinear corpus, estos modelos se encargan de extraer frases o inferir reglas de sintaxis.

Un modelo se define como una distribución de probabilidad sobre el conjunto de las cadenas de caracteres o de palabras, a partir del análisis de un corpus (Martín Mateos and Ruiz Reina (2013)).

Se conoce como modelos de IBM a cinco modelos que realizan la alineación de palabras de una frase origen y una frase destino.

Modelo de traducción

Los modelos de traducción se originan en IBM (Collins (2011)) estos modelos dan origen a lo que es la traducción automática, esto mediante la obtención de las probabilidades que se le da a cada traducción de ser generada a partir de una oración

dada en un idioma diferente, estas probabilidades dependen directamente del corpus paralelo utilizado. Dentro de las limitaciones que presentan estos modelos, es que solo modelan alineamientos del lenguaje *1:N* del lenguaje origen al lenguaje destino, es decir en la traducción *source-target* varias palabras origen pueden conectarse a una en el idioma destino, pero no al revés, la información sintáctica se limita a clases de palabras, no es posible hacer re-ordenamientos de larga distancia, la gramaticalidad de las traducciones depende exclusivamente del modelo de lenguaje usado, se tiene como consecuencia de estas restricciones los textos mal o no traducidos.

Alineamiento de textos

Dado un corpus paralelo bilingüe, alinear textos consiste en conectar cada palabra que se va a traducir a su equivalente semántico en la lengua destino. Teniendo fenómenos como alineamientos multiples (1:N) o alineamientos no conectados a nada (1:0) (Estrella (2018)).

Modelos de alineación IBM

Son una secuencia de modelos utilizados en la traducción automática estadística para entrenar un modelo de traducción y a su vez el modelo de alineación a nivel de palabras para así poder mapear las oraciones de origen con las oraciones de destino correspondientes; comenzando con las probabilidades de traducción léxica y pasando a la reordenación y la duplicación de palabras (Brown et al. (1993)).

Modelo 1 de IBM Este modelo no considera el orden de las palabras, toma cada frase como una bolsa de palabras[4]. No modela alineamientos 1:*N*, 1:*0*, sólo 1:1, es decir, solo toma los casos cuando una palabra de entrada se traducirá en

[4]Las palabras que aparecen en un archivo son usadas como términos índices para ese archivo.

una sola palabra en la lengua destino, pero no cuando palabras producirán varias palabras o incluso se descartarán (no producirán palabras) (Wołk and Marasek (2014)).

Modelo 2 de IBM Este modelo incorpora el concepto de re-ordenamiento absoluto de palabras, es decir, la probabilidad de una conexión entre palabras en la lengua destino y palabras en lenguaje origen dependerá de la distribución de probabilidad de alineación $a(i|j, l_s, l_t)$, donde i es la posición de una palabra origen y j su posición en la lengua destino. l_s es la longitud de la oración de origen y l_t la longitud de la oración destino.

Sí $p(s|t)$ es la probabilidad de traducción y $a(i \vee j, l_s, l_t)$ la probabilidad de alineación, entonces el Modelo 2 de IBM se define como (Wołk (2019)):

$$p(s, a|t) =\in \prod_{j=1}^{l_s} f(s_j|t_{a(j)})a(a(j)|j, l_s, l_t)$$

mapeando cada una de la palabras la traducción generada a una en la lengua que se desea traducir, dado por $a(j)$.

Modelo 3 de IBM Incorpora el concepto de *fertilidad*, que se define como la probabilidad de que una palabra en el origen se conecte a n palabras del destino. Por ejemplo:

Nochipan ipan in metztli Diciembre tequin ***sehua****.*

Siempre ***hace mucho frío*** *en el mes de Diciembre.*

La traducción de *"sehua"* es *"hace mucho frío"*, y decimos que la fertilidad de *sehua* es 3.

Además del concepto de *distorsión*, que es la probabilidad de que dada una palabra del origen en una posición i se conecte a una palabra en posición j, dado el largo de la oración origen y la oración destino.

También permite la asignación nula,que es la alineación de la palabra nula en el idioma origen s_0 con el texto en el lado del idioma destino, es decir,no existe traducción para esa palabra. Sin embargo, la diferencia importante es que ahora *nulo* puede correlacionarse con varias palabras. Es decir, una palabra que no representa nada también tiene fertilidad, (Estrella (2018)), ya que todas las palabras del destino podrían no estar conectadas a alguna palabra en el origen, es decir:

- Para cada palabra en el origen $s_i(i = 1, 2, ..., l)$ se elige la *fertilidad* Φ_i con probabilidad $n(\Phi_i|s_i)$.
- Elegir el número Φ_0 de palabras destino que se generan a partir de $s_0 = null$, usando la probabilidad p_1 y la suma de fertilidades obtenidas en el paso anterior.
- Sea m la suma de las fertilidades para todas las palabras, incluido *null*.
- Para cada $i = 1, 2, ..., l$, y cada $k = 1, 2, ..., \Phi_i$, se elige a una palabra en el destino τ_{ik} con probabilidad $t(\tau_{ik}|s_i)$.
- Para cada $i = 1, 2, ..., l$, y cada $k = 1, 2, ..., \Phi_i$, se elige dentro del destino la posición π_{ik} con probabilidad $d(\pi_{ik}|i, l, m)$.

- Para cada $k = 1, 2, ..., \Phi_0$, se escoge una posición π_{0k} de las $(\Phi_0 - k + 1)$ posiciones vacantes en $1, 2, ..., m$, para una probabilidad total de $1/\Phi_0!$.
- Obteniendo una oración en el idioma destino con τ_{ik} palabras en las posiciones π_{ik} $(i = 1, 2, ..., l$, y $k = 1, 2, ..., \Phi_i)$.

Modelo 4 de IBM En este modelo, cada palabra depende de la palabra previamente alineada y de las clases a las que pertenecen las palabras circundantes. Algunas palabras tienden a re-ordenarse durante la traducción más que otras. En algunos idiomas, como el inglés, dependiendo del contexto, los adjetivos a menudo se mueven antes del sustantivo que los precede. Las clases de palabras introducidas en este modelo, resuelven este problema condicionando las distribuciones de probabilidad de estas clases.

Para cada conjunto de palabras s_i alineado por lo menos a una palabra en t se define como un *cept*, si hay palabras con fertilidad distinta de cero y los *cept* multipalabra son excluidos, es decir, un *cept* es un subconjunto de palabras en la oración s con las posiciones de las palabras que generan (Koehn (2009)). Dicha distribución se puede definir de la siguiente manera.

- Para la palabra inicial en el *cept*: $d_1(j - \odot_{[i-1]}|\mathcal{A}(f_{[i-1]}), \mathcal{B}(s_j))$
- Para palabras adicionales: $d_1(j - \Pi_{i,k-1}|\mathcal{B}(s_j))$

Donde $\mathcal{A}$ y $\mathcal{B}$ son funciones que dependen de los vocabularios, además s_j y $f_{[i-1]}$ son distribuciones de probabilidad de distorsión de las palabras.

El *cept* se forma alineando cada palabra de entrada f_i con al menos una palabra de salida.

Modelo 5 de IBM Este modelo re formula el modelo de alineación con más parámetros de entrenamiento para superar la deficiencia del modelo anterior (Knight (1999)), además de que solo se pueden colocar palabras en posiciones libres, esto se realiza rastreando el número de posiciones libres y permitiendo la colocación solo en dichas posiciones. El modelo de distorsión es similar al modelo 4, pero se basa en posiciones libres. Sí v_j denota el número de posiciones libres en el destino, las probabilidades de distorsión se define como (Brown et al. (1993)):

- Para la palabra inicial en el *cept*: $d_1(v_j|\mathcal{B}(s_j), v_{\odot i-1}, v_{max})$
- Para palabras adicionales: $d_1(v_j - v_{\pi_{i,k-1}}|\mathcal{B}(s_j), v_{max'})$

Usa el primer modelo para definir de forma iterativa la alineación. Primero, considera que los parámetros del modelo tienen una probabilidad uniforme. En cada iteración, asigna probabilidades a los datos faltantes y asigna parámetros para completar los datos.

Modelo 6 de IBM Se define como sigue (Wołk (2019)):

$$p_6(t, a|s) = \frac{p_4(t,a|s)^{\alpha} \cdot p_{HMM}(t,a|s)}{\sum_{a',t',p_4}(t',a'|s)^{\alpha} \cdot p_{HMM}(t',a'|s)}$$

Donde, el parámetro de interpolación α se utiliza para contar el peso del modelo IBM 4 en relación con el modelo oculto de Markov. Una combinación

logarítmica lineal de varios modelos se puede definir como $p_k(t, a|s)$ para $k = 1, 2, ..., K$ como:

$$p_6(t, a|s) = \frac{\prod_{k=1}^{K} p_k(t,a|s)^{\alpha_k}}{\sum_{a',t'} \prod_{k=1}^{K} p_k(t',a'|s)^{\alpha_k}}$$

La combinación logarítmica lineal se usa en lugar de la combinación lineal porque los $p_k(t, a|s)$ valores son típicamente diferentes en términos de sus órdenes de magnitud.

Este modelo produce las mejores alineaciones entre los modelos de IBM, con tasas de error de alineación más bajas (Och and Ney (2003)).

Modelo de lenguaje

Se define como Modelo de Lenguaje, al conjunto de reglas que permiten estructurar un lenguaje, de tal forma que el orden de las sentencias lingüísticas pueda ser restringidas en base a probabilidades. Estas reglas sirven de herramienta en aplicaciones que muestran una sintaxis y/o semántica compleja (Olarte (2019)).

Si un modelo acepta mayoritariamente oraciones correctas y elimina aquellas oraciones que presenten palabras incorrectas, se dirá que es un buen modelo de lenguaje.

N-grama

En los campos de la Inteligencia Artificial, procesamiento del lenguaje natural, Bioinformática, recuperación de información se define como *n-grama* aquella subcadena de n elementos sucesivos, es decir (Manning and Schutze (1999)):

Dados $s_1 s_2 s_3 ... s_k ...$ elementos de una secuencia ordenada S, se define como *n-grama* a cualquier subsecuencia $A = s_{i+1} s_{i+2} ... s_{i+n}$, donde i es un número entre 0 y $|S| - n$, garantizando de esta forma que la subsecuencia A tenga un tamaño de n o de manera equivalente $|A| = n; n > 1$.

Si *n=2* se denominan bigramas; n=3, trigramas; cuando n>=4 se dirá en general $n - gramas$.

Los *n-gramas* son de mucho provecho para el procesamiento de documentos, en la toma de decisión del lenguaje vertido en textos y en el análisis estadístico de estos.

2.1.4. Corpus

Los textos contenidos dentro de un corpus no necesariamente se encuentran en una sola lengua. Puede ser dos (corpus bilingüe) o en más lenguas (corpus multilingüe). En tales casos dichos textos no se encuentran reunidos de manera arbitraria, sino que están considerados según criterios idénticos de selección en las distintas lenguas.

Hoy en día se considera que el corpora deben cumplir los siguientes requisitos:

1. Formato electrónico: El corpus se debe encontrar digitalizado para que pueda ser una herramienta útil para el lingüista y especialistas en el área.

2. Autenticidad de los datos: Para considerar un corpus, se deben de tener que los textos son tomados a partir de muestras reales de la lengua objeto a estudiar.

3. Criterios de selección: En el momento de la creación de un corpus se designan

diferentes criterios dependiendo de un objetivo en particular que se tenga, además se consideran apreciaciones lingüísticas y/o extralingüísticas.

4. Representatividad: Los textos a considerar deben observar parámetros de tipo estadístico, en donde, se pueda garantizar la representatividad de la variabilidad de la lengua objeto en dichos textos, sin olvidar los criterios antes mencionados.

5. Tamaño: Como primer paso se considera la recopilación del corpus, y antes de reunirlo se fija un tamaño, el cual puede estar dado en millones de unidades (entendiendo por unidades palabras, frases, formas o n-gramas) y se inicia el proceso de recopilación hasta llegar el tamaño previamente definido.

Lingüística de corpus

Utilizando muestras de datos que sean reales, para estudiar una lengua objeto a partir de un corpus, se define una metodología empírica de trabajo. De esta manera, el *corpus* se puede definir como el conjunto de datos, esto en una acepción muy general. Se establece como *lingüística de corpus* al uso de sistemas que han sido de utilidad para reunir, procesar y organizar datos dentro de un corpus, proveyendo actualidad a dicha tarea.

Según Cendejas Castro (2013) los corpus pueden ser:

- *Bases de datos de árboles*: Se tiene que los textos se encuentran etiquetado de manera sintáctica y el análisis sintáctico se estructura de forma arbórea, tomando de ahí su nombre.

- *Corpus orales*: Éstos corpus se conforman de señales de voz y en algunas ocasiones se consideran la transcripción de su correspondiente anotación fonética.

- *Corpus multimodales*: Éstos corpus se conforman de diferentes datos orales, entre los que se pueden considerar: la prosodia, los gestos, los movimientos de la boca, grabaciones sonoras y fílmicas en donde se incluye noticias y documentales.

- *Corpus textuales*: Se consideran aquellos textos que contienen la lengua escrita o bien la transcripción de la lengua oral.

- *Corpus sincrónicos vs. diacrónicos*: Un corpus sincrónico es aquel corpus que contiene textos representativos de la lengua actual, mientras que el corpus diacrónico refiere a textos de distintas etapas de la historia con el propósito de registrar modificaciones que se presentan en estas.

- *Corpus monolingües vs. multilingües*: Se define al corpus monolingüe como aquel que contiene textos en una única lengua mientras que el corpus multilingüe contienen textos traducidos en varias lenguas y que sean de utilidad para comparar en diferentes lenguas, así como para componer diccionarios y motores de traducción, siendo éstos últimos casos.

- *Corpus históricos vs. textuales modernos*: Un corpus histórico requiere que los documentos sean digitalizados a través de un reconocimiento óptico de caracteres (ROC/OCR[5]) mientras que en un corpus moderno los textos ya se

[5]http://www.alegsa.com.ar/Dic/ocr.php

encuentran digitalizados para su uso.

- *Corpus de referencia vs. monitor*: La principal diferencia en éstos corpus es que el primero tiene un tamaño previamente determinado y el corpus monitor puede aumentar su dimensión en todo momento.

- *Corpus dialectales*: Son aquellos corpus que únicamente se encuentran de manera oral.

Corpus paralelos

Se define como corpus paralelo aquellos documentos que contienen el mismo contenido semántico, pero escrito en diferentes idiomas, a estos documentos se les conoce como corpus bilingües o bitextos[6]. Sin embargo, estos textos no necesariamente tienen una equivalencia entre palabras, oraciones y/o párrafos; por ejemplo, dos documentos pueden tener oraciones completamente desordenados con relación al documento original, sin dejar de ser documentos paralelos.

Se puede decir entonces que, un corpus paralelo es un reportorio de documentos escritos en una o más lenguas y que tienen el mismo concepto semántico, por tanto se tiene un documento original y su traducción en otras lenguas. Es de importancia hacer mención que este tipo de corpus contiene documentos traducidos de la lengua **fuente** (source) a la lengua **destino** (target) y documentos traducidos de manera inversa.

Decodificador

Es el encargado de considerar la traducción con mayor probabilidad de todas

[6]Un documento presentado con dos columnas (original y traducción).

aquellas traducciones obtenidas. Por lo tanto, considerando tanto un modelo de lenguaje como uno de traducción se obtienen todas las traducciones y se ofrece aquella con la mayor probabilidad.

2.2. Aplicación web

Estas aplicaciones se ejecutan mediante una intranet o bien por internet y utilizan un navegador, y permiten realizar alguna actividad esto con la ayuda del acceso a un servidor web. Para lograr su ejecución se requiere que la aplicación sea codificada en un lenguaje que sea interpretado por los navegadores.

Una ventaja es que la aplicación web se maneja como un cliente ligero dentro del navegador, mostrando ser independiente de los sistemas operativos, además de ser fácil tanto para las actualizaciones como para el mantenimiento, sin necesidad de la distribución o instalación de algún software, este es el motivo por el cual dichas aplicaciones son populares (Lerma-Blasco et al. (2013)).

Servidor web

Es un programa conformado en dos partes, en la primer parte, que es la del servidor, y es donde se procesa la información y en la segunda parte, la del cliente se traduce dicha información para dar respuesta, dicha respuesta generalmente utiliza un lenguaje. La información que puede procesarse es desde páginas con hipertexto, pasando por texto, ligas a otros sitios, formularios hasta figuras botones e incluso multimedia.

HTML

Es un lenguaje que hace uso del marcado de hipertexto o "*HyperText Markup Language*" por sus iniciales en inglés. Cabe hacer mención que HTML no se considera como un lenguaje de programación, ya que el manejo de HTML es mediante etiquetas, contenido y atributos, y así escribe todos sus elementos.

HTML es un lenguaje web que es interpretado en el navegador permitiendo desplegar los diferentes sitios web e incluso soporta las aplicaciones web.

CSS

Las siglas CSS (*Cascading Style Sheets*) cuyo significado es «Hojas de estilo en cascada» se aplica a los *estilos* que presentan las páginas web o HTML tanto en color como forma y márgenes entre otros, de modo que sea simple, permitiendo que con facilidad se tenga un estilo semejante en las diferentes páginas de un sitio.

Estos estilos son aplicados siguiendo un patrón, iniciando de arriba abajo, y que mediante el uso de reglas se evitan ambigüedades en estos estilos.

PHP

PHP (Arce (2018)), que significa en inglés Personal Home Page Tools, aunque actualmente se traduce como *Hypertext Preprocessor* (preprocesador de hipertexto. Este lenguaje de programación es usado en la elaboración de paginas HTML. Además, su ejecución puede ser mediante el acceso a un servidor web, desde línea

de comandos y por medio de un cliente GUI[7].

Una de las principales ventajas de este lenguaje de programación, es que puede ser ejecutado en todos los sistemas operativos y en varios servidores web.

Además de tener diversas bibliotecas que sirven en la programación de procesos frecuentes, como es el caso tener soportar de una gran cantidad de bases de datos.

Las paginas programadas en lenguaje PHP son procesados por un servidor web que se encarga de interpretar el código HTML y los comandos PHP que se encuentran en estas, y finalmente estas son visualizadas por los navegadores web.

2.3. Python

Python es un lenguaje de programación cuya sintaxis favorece a una interpretación mas fácil y esto genera un código más limpio (González Duque (2019)).

Se dice que este lenguaje de programación es un lenguaje interpretado, debido a que mediante el uso de un programa interprete no es necesario compilar a lenguaje máquina, ya que sus líneas de código se ejecutan directamente en la computadora.

Tiene un tipado dinámico, con lo que la declaración de los tipos de variables a utilizar puede ser asignados mientras el programa se está ejecutando, y su tipo de variable puede cambiar dependiendo del valor asignado.

Fuertemente tipado, para poder hacer uso de una variable como una nueva variable de distinto tipo, se debe cambiar previamente a esta variable al nuevo tipo deseado.

[7]Aplicación mediante la cual un usuario puede interactuar con la computadora.

La programación en el lenguaje Python se puede llevar a cabo en multitud de plataformas, como por ejemplo sistemas UNIX, Solaris, Linux, DOS, Windows, OS/2, Mac OS, etc., esto debido a que el interprete de este lenguaje puede correr en estos sistemas, y los programas no presentan problemas de compatibilidad, al menos que se haga uso de bibliotecas específicas a un sistema operativo.

Dentro de los paradigmas que permite Python, se encuentra la programación imperativa, programación funcional y también la programación orientada a aspectos.

Capítulo 3. Estado del arte

En la actualidad existe una diversidad de diccionarios del Náhuatl que pueden ser encontrados en el Internet, tales como AULEX[1], un diccionario pensado para autodidactas e investigadores que deseen conocer el vocabulario nuevo del idioma náhuatl, el cual contiene voces del náhuatl coloquial y culto. Al ser un diccionario en línea, es posible encontrar definiciones de palabras o ocurrencias de subcadenas en las entradas del diccionario. Este mismo diccionario ha sido vaciado en FREELANG[2] y puede nuevamente consultarse entradas al diccionario usando una interfaz web.

Por otro lado, la Academia Veracruzana de las Lenguas Indígenas presenta una página web con su Diccionario náhuatl – español en línea[3]. Tiene una variedad interesante de entradas, sin embargo, no deja de ser un diccionario en línea. Existe también el conocido Gran Diccionario Náhuatl de la UNAM[4] que está disponible en línea. También podemos encontrar frecuentemente cursos en línea para el

[1] http://aulex.org/nah-es/

[2] https://es.freelang.net/enlinea/nahuatl.php

[3] http://aveligob.mx/traductores/nahu-espa/Buscespnah_2.php

[4] http://www.gdn.unam.mx

aprendizaje, tales como el mostrado en el Curso de Náhuatl en línea[5] que contiene materiales adicionales a un simple diccionario. Más aún, existen aplicaciones (APP) para dispositivos móviles para el aprendizaje del Náhuatl como lo es Metstlisoft[6], o *Tozcatl*, aplicación para sistemas Android[7], y que fue programada por estudiantes de Ciencias de la Computación de la Benemérita Universidad Autónoma de Puebla. Si bien, indican una traducción automática del epañol al nahuatl, según se ha observado, no usa modelos de traducción estadísticos, sino más bien, empatamientos exactos de cadenas de texto basadas en un diccionario clásico, por lo cual, no es posible llevar a cabo traducciones de oraciones complejas.

Los métodos de traducción estadísticos son los que en la literatura han mostrado tener los mejores rendimientos para los sistemas de traducción automática. Sin embargo, tienen el inconveniente de necesitar corpus bilingües paralelos para llevar a cabo la estimación de los diccionarios de traducción estadísticos.

En este caso, tenemos la ventaja de contar con un desarrollo importante por parte de Gutierrez-Vasques et al. (2016). Los autores han dedicado tiempo para construir un corpus paralelo español-náhuatl, el cual puede ser usado para estimar los tan necesarios diccionarios estadísticos de traducción. Los modelos de traducción que tienen como base secuencias de palabras son mucho más efectivos para llevar a cabo la traducción de una lengua origen a una lengua destino y han ganado popularidad debido a dos cosas: a) la generación de corpora paralelo bilingüe tales

[5]http://mexica.ohui.net/glosarios/2/

[6]http://www.conacytprensa.mx/index.php/tecnologia/tic/3488-crean-app-para-traducir-al-nahuatl

[7]https://www.android.com

como EUROPARL[8], y b) el acceso a recursos computaciones con mayores cantidades de memoria.

En la literatura podemos encontrar muchos artículos que tratan el problema de traducción estadístico bilingüe para el caso de múltiples idiomas, tales en Stentiford and Steer (1998), en donde se presenta un sistema de traducción de idiomas para traducir frases de un primer idioma a un segundo idioma que comprende una colección de frases en el segundo idioma. Las frases introducidas en el primer idioma se caracterizan cada una por una o más palabras clave, y se genera la frase correspondiente en el segundo idioma. Tal enfoque de frases permite lo que es una traducción efectivamente rápida y precisa, incluso del habla. Dado que las frases en el segundo idioma se preparan con anticipación y se guardan, no debe haber problemas de mala traducción o construcción incompleta. El resultado puede estar en texto, o, usando síntesis de voz, en forma sonora. Con la elección adecuada de palabras clave, es posible caracterizar una gran cantidad de frases relativamente largas y complejas con solo unas pocas palabras clave.

En ciertos trabajos han atendido el problema de la generación de herramientas de apoyo como, por ejemplo, en Ortiz-Martínez et al. (2019), que mediante la herramienta desarrollada Thot, estudian y aplican modelos estadísticos de sucesiones de palabras, y a partir de estas cadenas se pueden crear diccionarios estadísticos con traducciones correctas.

Una tesis que aborda la traducción Náhuatl - Español, es Ríos Dolores (2016), mostrando buenos avances, hacen uso de un corpus de textos muy diversos y de

[8]https://www.statmt.org/europarl/

muchas regiones del país, haciéndolo un traductor de propósito muy general.

A pesar de todas estas valiosas contribuciones, no hemos encontrado un sistema de este tipo para el caso del Español-Náhuatl, y que sea de la variante del centro de Puebla, la razón, probablemente sea que hasta el momento no existían corpus paralelo para ambos idiomas. De ahí la importancia de construir un modelo computacional basado en traducción estadística que permita encontrar probabilidades de traducción basadas en secuencias de palabras y que el sistema sea de acceso libre.

En el caso particular de los sistemas automático estadísticos se han desarrollado diversos sistemas, en la siguiente sección se explican algunos de ellos.

3.1. Sistemas de Traducción Automática Estadística

El auge que tuvo la traducción automática permitió que estos sistemas no solo fueran utilizados para fines de estudio, si no que también han llegado a encargados y usuarios que realizan traducción, haciendo que su uso se convirtiera en algo más frecuente, permitiendo así el incremento de más motores de traducción automática (Parra Escartín (2018)).

Estos nuevos sistemas tenían grandes mejores con relación a sus predecesores, esto debido a los progresos en las investigaciones, obteniendo traducciones muy aceptables, haciendo de estos, una herramienta confiable para aquellos que se dedican a la traducción.

Los motores de traducción automática fueron los principales componentes que tuvieron mejoras con el avance de los sistemas de traducción, y esto se vio reflejado

en la forma de hacer los ordenamientos suboracionales[9], esto como una mejora a los modelos ya existentes de traducción, así como la integración de datos lingüísticos al entrenamiento, para finalmente también sumar métodos de procesamiento del lenguaje natural y aprendizaje automático. La eficacia de estos motores de traducción está relacionado directamente a la buena condición que presenten los corpus, a los documentos lingüísticos y otros factores que intervienen durante el entrenamiento de las traducciones.

Surgiendo proyectos de sistemas de traducción automática estadística como los siguientes:

KantanMT

KantanMT[10], es una plataforma Irlandesa de traducción automática estadística basada en la nube. Los miembros pagan una tarifa de suscripción mensual para una cuenta en la plataforma KantanMT, permitiendo a sus usuarios construir fácilmente motores de traducción en más de 750 combinaciones de idiomas.

Lilt

Lilt[11] es una plataforma de origen estadounidense, que hace uso de la traducción automática estadística para dar soluciones a peticiones de traducción; además en este sistema con su modulo de traducción automática interactiva, se puede optar por alguna de las traducciones que genera el sistema, esto dependiendo de alguna de las palabras que se seleccione generadas en la traducción, obteniendo así, una nueva salida de traducción.

[9] Aquellas en las que parte de la estructura de la oración se elimina.

[10] https://www.kantanmt.com

[11] https://lilt.com

MateCAT

La plataforma MateCAT[12] es el resultado de proyectos de investigación europeos. Estos sistemas de traducción automática asistida tienen la ventaja de que pueden ser accedidos las veinticuatro horas vía Internet.

El apogeo de esta herramienta se ha visto favorecida debido al apoyo de grandes grupos, como lo son la Fundación Bruno Kessler de Trento (Italia), Translated Srl., la universidad de Le Mans (Francia) y la universidad de Edimburgo (Reino Unido), cuyas investigaciones e inversiones económicas, han dado como resultado este sistema.

Estas herramientas antes mencionadas, aparte de ser software con licencia comercial, muestran la limitación cuando se trata de dos lenguas con excasos recursos lingüisticos digitales, o son basados en el siguiente sistema.

Moses

El Sistema de Traducción Automático Estadístico Moses[13], o por sus siglas en ingles SMT Moses (*Statistical Machine Translation Moses*), es un sistema bajo licencia LGPL[14], y por ende puede ser utilizado por otros sistemas. El sistema Moses tuvo como idea basa a EuroMatrix y EuroMatrixPlus, auspiciados por la Comisión Europea, además de un sinfín de organismos como las Universidad de Praga, Edimburgo, Maryland, el Instituto Tecnológico de Massachusetts, así como la agencia de DARPA, RWTH Aachen, la Fundación Bruno Kessler, NFS, el departamento de defensa de los EEUU y el proyecto TC-Star.

[12]https://www.matecat.com

[13]http://www.statmt.org/moses/

[14]Licencia de software libre (Lesser General Public License).

Dado un Corpus paralelo de dos lenguas, SMT Moses realiza de forma automática pruebas de modelos de traducción, para así hacer más rápido la búsqueda de la traducción correcta. Este sistema mediante el *Teorema de Bayes* hace cálculos a partir de la cadena de entrada (t), obteniendo la cadena (s) de mayor probabilidad de ser la traducción idónea. Y mediante el calculo de $p(s|t) \cdot p(t)$ podemos obtener $p(t|s)$, que es la cadena con mayor posibilidad de ser la traducción de la oración a traducir(modelo de traducción), y $p(t)$ es aquella cadena con mayor probabilidad de ser la cadena de traducción (modelo de lenguaje de la lengua a la que se desea traducir). Obtener la traducción más adecuada $\tilde{\boldsymbol{O}}$ Obtener la traducción más adecuada X se hace calculando la sentencia de caracteres con la máxima probabilidad correcta de traducción:

$$\tilde{\boldsymbol{O}} = argmax\ p(t|s) = argmax\ p(s|t) \cdot p(t)$$

SMT Moses puede utilizar algunos de los siguientes módulos SRILM (*SRI Languages Modeling Toolkit*), IRSTLM (*IRST Languages Modeling Toolkit*), KENLM (*KenLM Language Model Toolkit*), RANDLM (*Randomised Language Modelling Toolkit*) para generar el modelo de lenguaje tipo n-grama[15].

Teniendo como base un corpus bilingüe, esta herramienta crea el modelo de traducción mediante el módulo GIZA++, generando alineamientos de traducciones en ambas direcciones, es decir, del texto que se desea traducir al texto generado, y del texto generado al texto que se desea traducir, estos modelos de traducción están basados en palabras, y cuyos alineamientos resultantes son combinados.

[15] Dadas n-1 palabras, el modelo de n-grama da la probabilidad de obtener la n-sima palabra.

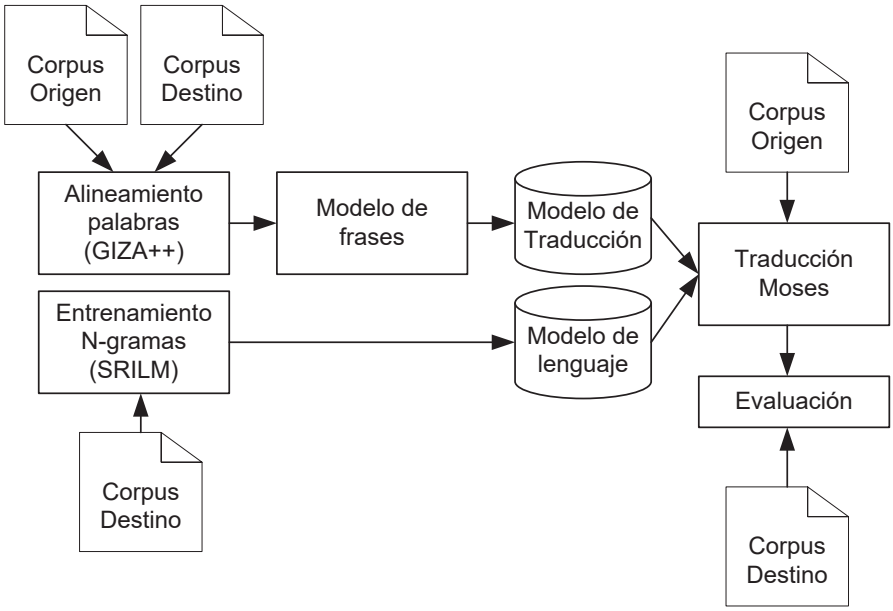

Figura 3.1: Arquitectura Moses (Koehn (2019)).

En la figura 3.1 se presenta la arquitectura de SMT Moses.

Slate

Slate Toolkit[16] es una herramienta de uso comercial, que hace uso de traducción estadística, tiene soporte para diccionario de frases, re-ordenamiento léxico, modelado de idiomas.

Varias de sus herramientas son modificaciones de SMT Moses.

[16]https://www.slate.rocks

Capítulo 4. Lenguas indígenas

4.1. Lengua

El proceso de la comunicación humana involucra un sinfín de procesos, desde el entender, comprender y analizar los signos de comunicación empleados entre diferentes personas, o tan complejos como los códigos lingüísticos que tiene cada lengua, incluyendo conductas culturales y sociales que rigen la interacción de las personas. La actividad de comunicación combina además dentro lo social, cuestiones biológicas de los individuos hasta cuestiones institucionales y de credo, desde puntos de vista individual y colectivos, haciendo muy cambiante la forma de expresar entre los humanos Rojas (2017).

Para Coelho (2019) **lengua**, es una cita fundamental a la comunicación humana, establecido como un sistema de intercambio verbal, mediante el uso de símbolos orales y palabras descritas de forma gestual, y escrito, mediante el uso de papel y mensajes.

4.2. Lenguas indígenas

La Organización de las Naciones Unidas (ONU (2019)) registra que de las casi 6,700 lenguas que se tiene conocimiento, solo el 3 % de la población mundial hace uso del 96 % de estas lenguas. Otro dato importante, es que existen al menos 4,000 lenguas indígenas, pero que solo es hablado por el 6 % de la población mundial.

América Latina posee una grandeza inmensurable en el mundo en cuestión de familias lingüísticas. El 87 % de la población indígena que radica en América Latina y el Caribe se encuentran en los países de México, Perú, Guatemala, Bolivia y Colombia.

Estas lenguas significan más que un medio de comunicación, lenguas que a través de siglos se han ido desarrollando, hasta ser bastos y completos sistemas de transmisión de conocimientos de culturas que predominaron en la antigüedad. Estos lenguajes son símbolo de identidad de estas comunidades indígenas que a través de miles de años se han preservado, dando mediante estas, una representación a sus culturas, ideas y manifestación de sus tradiciones.

4.3. Lenguas indígenas en México

La gran diversidad existentes de lenguas indígenas hacen de México, un país con una gran riqueza lingüística, estas lenguas han sido habladas desde la época prehispánica, información perteneciente al Instituto Nacional de Lenguas Indígenas (INALI), indica que dentro de los países que reúnen a la mitad de las lenguas

indígenas hablantes en el mundo, México forma parte importante de este grupo. Y de las cuales, muchas están en peligro de desaparecer debido a factores demográficos, es decir, que las lenguas son habladas por una mínima cantidad de individuos, propiciando la sustitución de estas por la lengua más prestigiosa, en nuestro caso el español.

Claro registro de estas lenguas indígenas en México es el de las pequeñas poblaciones aztecas, que se encuentran presentes desde antes de la conquista, y cuya relación comercial con otros pueblos, hacia que sus lenguas tuvieran un alto crecimiento. Gracias a estelas, construcciones, esculturas, pergaminos, entre otras cosas que estos pueblos usaban para anotar su escritura, lengua y conocimientos a través de estas, es así como se puede apreciar información valiosa como su forma de comunicación, gobierno, religión, cultura, forma de vida, etc.

Después de la conquista, 1521, los conquistadores debían de resolver la forma de comunicación con los pueblos conquistados. España no podía entender este gran problema, pues desconocía de la gran diversidad lingüística que dominaba en las nuevas tierras. Fue de importancia crear diccionarios y gramáticas, pues, debido a las guerras y los vínculos comerciales que tenían los aztecas, la lengua náhuatl se había propagado por varios territorios, teniendo una etapa geolingüística muy importante, ya que, a lo largo del tiempo, esta lengua desarrollo variantes, esto debido al sistema tributario azteca; además de la existencia de otras lenguas (García Aymerich (2017)), y cuyo usos siguen hasta nuestros días.

4.3.1. Náhuatl

En México, afirma Hernández (2011), el náhuatl es la segunda lengua con la mayor cantidad de personas que lo hablan. Además, de tener una fuerte influencia en el español de México, ya que derivado de esta lengua, se tienen una gran cantidad de nahuatlismos de uso diario, ya que de estos se mencionan diferentes herramientas, gastronomía, nombres personales, animales, expresiones de uso cotidiano, tales como apapachar, aguacate, ezquite, metate, comal, chocolate, atole, jícara, papalote, popote, tlapalería, tianguis, Izel, Neli, entre otros.

De igual manera, varios lugares de la república mexicana, su nombre tiene un origen y significado en la lengua náhuatl, y estos lugares se encuentran a lo largo del territorio nacional. Se pueden observar desde nombres de lugares, pueblos, ciudades, municipios y hasta estados, tales como México, Toluca, Jalisco, Michoacán, Oaxaca, Tlaxcala, Cuernavaca,Tzintzingareo, Tepoztlan, Tlacuilotepec, Mazatlán, Tlatlauquitepec, Chalco, Tlalnelhuyocan, Tlaquiltenango, Azcapotzalco, Ecatepec, Tequixquiac, Popocatepetl, entre otros sitios.

Hasta hoy en día, pueblos nahuas siguen conservando sus tradiciones y costumbres como parte de su identidad comunitaria y regional, esto a pesar de los intentos de homologar costumbres, lenguas y culturas que se presento el siglo pasado, más sin embargo estos pueblos siguen conservando su patrimonio milenario, incluyendo las variantes de su lengua. A la llegada de los españoles, estos se percataron de la gran variedad de lenguas nunca escuchadas por ellos, y comprender estas significo un gran problema que tuvieron que resolver. En su investigación descubrieron que había

diversas lenguas generales y que había personas que podían servir de interpretes o nahuatlatos.

El gran imperio azteca contaba con una de esas lenguas generales, el náhuatl o mexicana, además de ser hablada por pueblos vecinos al pueblo azteca. Y cuya expansión de esta lengua por Mesoamérica se debió principalmente a las conquistas de la Triple Alianza, teniendo una gran influencia cultural, incluso sobre otras lenguas generales que desde los olmecas habían visto pasar sucesivamente varias entidades políticas (Valle Perez (1998)). A principios del siglo XVI la lengua náhuatl tenía una influencia que abarcaba desde los territorios del actual estado de Sonora, hasta la península de Nicoya en la actual Republica de Costa Rica. Entre las lenguas generales, la lengua náhuatl pasó a ser la de mayor relevancia, lengua franca. Esta lengua sirvió como lenguaje de conquista y de evangelización hacia los pueblos conquistados, además de que se impuso como medio de escritura alfabética, propiciando de esta forma la producción de infinidad de textos escritos en esta lengua, esto sin dejar de lado la escritura pictoglífica. Y durante el apogeo que vivió el imperio azteca, su lengua, el náhuatl se adecuó a reglas dentro de un corpus de gramáticas y un enriquecimiento en su glosario, haciendo que esta lengua forme parte importante dentro de la cronología universal de la lingüística.

Se sabe (Hernández de León-Portilla (2011)) que los primeros hablantes de yutonahua llegaron de muy al norte, muy probable del actual estado costero de Oregón, en EE. UU., esto hace miles de años. Después siguieron migrando hacia el sur, teniendo como primer paso el actual estado de Sonora (2,500 a.C.), para finalmente tener un primero encuentro con el imperio teotihuacano, esto hacia el 400

a.C., y asentarse en esta zona, denominada Mesoamérica, que ya estaba habitada por grandes civilizaciones de alto nivel cultural, su largo caminar ha quedado registrado por camino de la lengua yutonahua que formaron un camino desde las Montañas Rocosas del norte, hasta la Sierra Madre Occidental.

Náhuatl de Puebla

De las 31 lenguas náhuatl que se hablan en el país, nueve se encuentran en el estado de Puebla.

1. Sierra Noreste de Puebla
2. Noroeste Central
3. Sierra Negra, Sur
4. Sierra Negra, Norte
5. Sierra Oeste de Puebla
6. Norte de Puebla
7. Centro de Puebla
8. Oriente de Puebla
9. Mexicano de Guerrero

De estas nueve lenguas náhuatl, una comparte territorio con el estado de Guerrero, el *mexicano de Guerrero*, el cual mayoritariamente se habla en ese estado, y al que la gente que lo habla le llama mexicano.

El *náhuatl* del centro de Puebla. Se emplea en varios municipios poblanos (Acajete, Acteopan, Atlixco, Atoyatempan, Calpan, Cohuecán, Huaquechula, Huejotzingo, Huatlatlahuca, La Resurrección, Nealticán, Ocoyucan, Puebla, San Andrés Cholula, San Juan Atzompa, Teopantlán, Tepatlaxco de Hidalgo, Tepeojuma, Tianguismanalco, Tlapanalá, Tochimilco, entre otros (INALI (2009))). Siendo San Miguel Canoa, que junto con la Resurrección y Azumiatla en donde se están establecidos el mayor número de personas que hablan el náhuatl en la capital de Puebla, y cuya variedad cuenta con una tipología no tonal[1] y sus palabras son largas con muchos afijos[2].

[1]Lengua que usa diferentes tonos para una misma palabra consiguiendo así significados distintos.

[2]https://definicion.de/afijo/

Capítulo 5. Desarrollo

En este capítulo se describirá los pasos seguidos en el desarrollo del sistema.

Corpus

Para la creación del Corpus Paralelo se hizo uso del Diccionario náhuatl del estado de Puebla (Brockway and Hershey (2018)) y de la ayuda de personas originarias de la Comunidad La Resurrección, así como de la coordinación de Náhuatl, de la Escuela Primaria Federal Bilingüe "Fuertes de Guadalupe y Loreto". Obteniendo $23,513$ oraciones náhuatl-español, los textos en náhuatl presentan caracteres latinos y sin variaciones ortográficas al español. Las oraciones estan dirigidos al tema *Náhuatl para la vida y el trabajo.*

Los archivos utilizados en el entrenamiento del sistema de traducción son generados a partir del corpus alineado, el cual está formado por tres pares de documentos, y que contienen oraciones alineadas, estos documentos están organizados por pares, tres documentos en la lengua que se desea traducir y otros tres en la lengua a traducir. Siendo el primer par, el que servirá para entrenar al sistema de traducción (20,000 oraciones), para la optimización se utiliza el segundo par (3,413 oraciones),

el ultimo par (100 oraciones) es ocupado para llevar a cabo las evaluaciones a tráves de alguna métrica, como por ejemplo BLEU.

Los archivos $frases.nh$ y $frases2.nh$ para las frases en náhuatl y $frases.es$ y $frases2.es$ para el español. A los archivos se le hizo una *tokenisation* (todas las palabras y signos quedaran separadas entre espacios en blanco). Todo en minúsculas, generando nuevos archivos a partir de los existentes (esto através de un programa realizado en `Python`):

$frases.nh \rightarrow frases.clean.nh$

$frases2.nh \rightarrow frases2.clean.nh$

$frases.es \rightarrow frases.clean.es$

$frases2.es \rightarrow frases2.clean.es$

La tabla 5.1 muestra las características principales del corpus, ya con los procesos realizados antes de ser introducidos al sistema.

Corpus	*frases.clean.nh*	*frases.clean.es*	*frases2.clean.nh*	*frases2.clean.es*
Oraciones	20,000		3,513	
Tokens	379,056	448,474	90,993	107,657

Tabla 5.1: Características del corpus

Y en la tabla 5.2 se muestra parte del corpus utilizado.

Implementación

El sistema implemento el modelo de Traducción Automática Estadística a partir de frases, usando el software de código abierto *SMT Moses*, sobre un sistema

operativo Linux, **CentOS 7**.

Modelo de lenguaje

Se generó el modelo de lenguaje mediante el programa *lmplz* de la herramienta KENLM, incluida en SMT Moses, que extrae a partir del mismo corpus, se considero trigramas como la máxima longitud de *n-gramas* para contar. Los resultados se muestran en la tabla 5.2, y en las figuras de la tabla 5.3 parte de los archivos generados.

Entrenamiento del sistema de traducción

Usando SMT Moses, se analiza y extrae la información estadística necesaria para realizar traducciones.

El proceso tardó cinco horas, obteniendo el diccionario estadístico para ambas lenguas.

Se necesita de un último paso para calibrar y optimizar el sistema de traducción utilizando los dos últimos archivos del corpus, los archivos *frases2.nh* y *frases2.es*.

En este proceso (español y náhuatl), se genera el archivo de configuración "`moses.ini`". El cual almacena las rutas de sistema generadas para el modelo, además de las configuraciones de parámetros necesarios para que se ejecuten las instrucciones. En la tabla 5.5 se muestran parte de los diccionarios estadísticos obtenidos.

Aplicación web

La traducción automática estadística se trasladó a una aplicación web desarrollada en HTML combinado con CSS y PHP siguiendo el algoritmo siguiente 5.1,que es utilizado para generar la interfaz, permitiendo un uso fácil y adecuado para cual-

Oraciones en Náhuatl	Oraciones en Español
In Julian mostli miaquiya inon itlanemac.	Julián aumenta su mercancía todos los días.
Ihcuac miquihque, quiihtohuahque sequin yahue mictlan, cani ahyic sehuis in tletl.	Cuando las personas se mueren, dicen que algunas van al infierno, donde nunca se apaga el fuego.
Inin michi tequin hueyi huan mochihuas tlacuali para nochin ichanihcau in michtemohuani.	Este pescado es muy grande y da comida para toda la familia del pescador.
Inon tlacatl ic motecpanohua, nochipan michquitzquiya.	Ese hombre se mantiene pescando todos los días.
In tlamachtiyani oyahca michtemohua ihuan inon itenonotzcau.	El maestro había ido a pescar con su amigo.
Inon tlacatl in itequiu michtemohuani.	Ese hombre trabaja como pescador.
In michtlapictli inon pihpictos ica in totomochtli.	Se enrollan los tamales de pescado en hojas de mazorca.
Inon tlacatl, ic miexcomitl, huelis cocoxqui inon iihte.	Ese hombre es pedorro; a lo mejor está enfermo del estómago.
Ihcuac tequin tlacua, inin telpocatl miexi.	Este niño ventosea cuando come mucho.
In miexini tlamati quen colotl huan ahmo quipiya ihuetz.	El vinagrillo se parece al alacrán, pero no tiene aguijón.
Inon tlacatl nisic nepic mihcuaniya; ahmo momati san cani.	Ese hombre siempre se muda de casa; no se halla en un solo lugar.
Inon telpocatl omihmat. Ayacmo ijconon oquichiu ica in ahmo cuali.	Ese muchacho ya se acostumbró; ya no hace esas maldades.
Inon tlacatl mihmati ihcuac huinti; ahmo huetzi san tlaluis.	Ese hombre se controla cuando se emborracha; no se cae tanto.
Inon tlacatl quihuica mihminqui tlen quicuas nepantla.	Ese hombre lleva tacos de frijol para comer al mediodía.
Inon tecutli mihmiqui ica itztic.	Ese anciano se entumece por el frío.
Ihcuac cuepontos inon chacalxochitl, mihnecui tequin ahhueyac.	Cuando ya abrió la flor de mayo huele bonito.
In conetl mihpotza tequin tleca ochichic miac.	El bebé eructa mucho porque tomó mucha leche.
In Pedro omihsotlac ipan in autobús tleca omocococ.	Pedro vomitó en el autobús, porque se enfermó.
In nosihtzin mihtohua quen san ijconon ma ye inon tlahtoli.	Mi abuela quiere decir que así se quedan las cosas.
In mihtotilistli tequin cualtzin.	La danza es muy bonita.
Yehhua mihtotiya ihcuac quinequi.	Él baila cuando quiere.
Inon mihtotiyani ahhuel mihtotiya cualtzin.	Ese danzante no sabe bailar muy bien.
Ihcuac nicpiya tzompili, nimihtzomiya cada rato.	Tengo que sonarme las narices a cada rato cuando tengo catarro.
Inon masehualten, ihcuac omononotzque ica tlen panohua, in Santiago mihyohuiya. Ahmo onahuat.	Cuando los vecinos platicaban de lo que sucedió, Santiago se abstuvo de hablar. No chistó.
In yolcamen noyojque mihyotiyahque quen in tlacamen.	Los animales también respiran como la gente.
Inon conetl ic tequin momohsihuiya, quemanian mijcahuetzi.	Cuando se molesta ese niño, a veces se priva.
Ihcuac in Juana otlatecac ica in totoltetl, omijcalac macuili.	Cuando Juana colocó una nidada de huevos para que empollara la gallina, se murieron cinco dentro del cascarón.

Tabla 5.2: Parte del corpus utilizado.

modelo	*español*	*náhuatl*
unigramas	55,747	69,858
bigramas	221,946	220,162
trigramas	361,294	328,068

Tabla 5.3: Modelos de lenguaje generados.

```
-0.59006023 ic oquintenihtzon inon
-0.59006023 tleca oticou inon
-0.59006023 ic timopalehuis inon
-0.59006023 ihcuac tesehualnextiya inon
-0.59006023 ihuan tetehui inon
-0.59006023 aquin tetlahtlamacas inon
-0.59006023 tlahmo quianasque inon
-0.59006023 tlahquitqui quinchihchihua inon
-0.7297615  in misa inon
-0.59006023 ihcuac onictlahpalo inon
-0.59006023 ihcuac yoquitlali inon
-0.59006023 noyojque quincua inon
-0.59006023 ic quintzacuas inon
-0.59006023 huehca tlacuitlapa inon
-0.59006023 ic quisasque inon
-0.59006023 tequin tlahmachtic inon
-0.59006023 tequihuahten oquitemohque inon
-0.59006023 juana yoquisiaumictihque inon
-0.82112503 ahmo quinnotza inon
-0.59006023 ica tlahueli inon
-0.59006023 ica ipaquilis inon
-0.59006023 oquiitac quiichcuatoc inon
-0.59006023 ma tlaoya inon
-0.59006023 tla quinamiquis inon
-0.59006023 ma quihuiquili inon
-0.59006023 con reverencia inon
-0.7297615  yica tlatlapehpentli inon
-0.59006023 sihuatl quinyecpaca inon
-0.59006023 cuali ticchihchihuas inon
-0.59006023 ma mitzcua inon
-0.85885745 axan tompilihui inon
```

Modelo de lenguaje en Náhuatl

```
-1.1806557  peon para vigilar
-1.1753289  milpa a vigilar
-2.6669087  va a vigilar
-0.87847996 ira a vigilar
-1.7111652  las mujeres cuelgan
-1.4787304  ropa para secarla
-3.1299546  ese hombre fecundo
-2.2531276  hombre le colgo
-2.9929771  para que pueda
-1.0484517  este fierro prensa
-2.5476885  con el tejolote
-0.8820144  tejolote se machuca
-2.0837953  les gusta espolvorear
-1.1618545  hombres estan esparciendo
-1.5717175  le ponen suficiente
-0.8806863  chiquillo tiene contiene
-2.519992   el agua contiene
-0.88078195 encomendar el encargo
-1.8679175  el le encargo
-0.87459284 buscar mi encargo
-0.8812769  sacara su cartilla
-2.0837953  les gusta asar
-1.1807057  carne para comerla
-0.881186   lana para cardarla
-1.6566256  agua que formaba
-2.1560383  la mujer desplumo
-1.356108   gallina para cocerla
-0.88189733 vereda que pasaba
-2.0173979  la autoridad multo
-1.3574144  persona que acarrea
-0.8674487  agua solo zambulle
```

Modelo de lenguaje en Español

Tabla 5.4: Parte de los modelos de lenguaje obtenidos.

```
se parece a ||| ixnesi quen ||| 0.166667
se parece a ||| ixnesi ||| 0.166667
se parece a ||| tlamati quen ||| 0.0909091
se parece a ||| tlamati ||| 0.142857
se parece al alacran   pero no ||| tlamati quen colotl huan ahmo ||| 1
se parece al alacran   pero ||| tlamati quen colotl huan ||| 1
se parece al alacran   ||| tlamati quen colotl ||| 1
se parece al cerdo   ||| tlamati quen pitzotl . ||| 0.5
se parece al cerdo   ||| tlamati quen pitzotl ||| 0.5
se parece al cerdo ||| tlamati quen pitzotl . ||| 0.5
se parece al cerdo ||| tlamati quen pitzotl ||| 0.5
se parece al fruto del sauco   ||| ixnesi quen itlaquilyo in xometl . ||| 0.5
se parece al fruto del sauco   ||| ixnesi quen itlaquilyo in xometl ||| 0.5
se parece al fruto del sauco ||| ixnesi quen itlaquilyo in xometl . ||| 0.5
se parece al fruto del sauco ||| ixnesi quen itlaquilyo in xometl ||| 0.5
se parece al fruto del ||| ixnesi quen itlaquilyo in ||| 1
se parece al fruto ||| ixnesi quen itlaquilyo ||| 1
se parece al fuego ||| nesi quen tletl   ||| 0.5
se parece al fuego ||| nesi quen tletl   ||| 0.5
se parece al gato domestico   ||| tlamati quen in mistli . ||| 0.5
se parece al gato domestico   ||| tlamati quen in mistli ||| 0.5
se parece al gato domestico ||| tlamati quen in mistli . ||| 0.5
se parece al gato domestico ||| tlamati quen in mistli ||| 0.5
se parece al gavilan ||| tlamati quen cuixi ||| 0.5
se parece al gavilan ||| tlamati quen ||| 0.0909091
se parece al ||| nesi quen tletl   ||| 0.5
se parece al ||| nesi quen tletl   ||| 0.5
se parece al ||| tlamati quen cuixi ||| 0.5
se parece al ||| tlamati quen in ||| 1
se parece al ||| tlamati quen ||| 0.272727
se parece mucho a su hermano ||| quiixuica yec in icniu ||| 1
```

Español–Náhuatl

```
inon sihuatl tequin tlahtlacolyo ||| esa mujer es muy ||| 0.25
inon sihuatl tequin ||| a esa mujer le ||| 1
inon sihuatl tequin ||| a esa mujer ||| 0.2
inon sihuatl tequin ||| esa mujer es muy pecadora ||| 0.5
inon sihuatl tequin ||| esa mujer es muy ||| 0.75
inon sihuatl tequin ||| esa mujer ||| 0.011236
inon sihuatl tlahtzonqui, quichihchihua ||| esa mujer es costurera, hace ||| 0.25
inon sihuatl tlahtzonqui. ||| esa mujer es costurera. ||| 1
inon sihuatl tlahtzonqui ||| esa mujer es costurera ||| 1
inon sihuatl yones quen pilua. ||| esa mujer esta embarazada. ||| 0.5
inon sihuatl yones quen pilua ||| esa mujer esta embarazada ||| 0.5
inon sihuatl yones quen pilua. ||| nota que esa mujer esta embarazada. ||| 0.5
inon sihuatl yones quen pilua ||| nota que esa mujer esta embarazada ||| 0.5
inon sihuatl yones quen pilua. ||| que esa mujer esta embarazada. ||| 0.5
inon sihuatl yones quen pilua ||| que esa mujer esta embarazada ||| 0.5
inon sihuatl yones quen pilua ||| se nota que esa mujer esta embarazada ||| 0.5
inon sihuatl yones quen pilua ||| esa mujer esta embarazada. ||| 0.5
inon sihuatl yones quen pilua ||| esa mujer esta embarazada ||| 0.5
inon sihuatl yones quen pilua ||| nota que esa mujer esta embarazada. ||| 0.5
inon sihuatl yones quen pilua ||| nota que esa mujer esta embarazada ||| 0.5
inon sihuatl yones quen pilua ||| que esa mujer esta embarazada. ||| 0.5
inon sihuatl yones quen pilua ||| que esa mujer esta embarazada ||| 0.5
inon sihuatl yones quen pilua ||| se nota que esa mujer esta embarazada ||| 0.5
inon sihuatl ||| a esa mujer. ||| 1
inon sihuatl ||| a esa mujer ||| 0.8
inon sihuatl ||| de esa mujer ||| 1
inon sihuatl ||| del pie a esa mujer. ||| 1
inon sihuatl ||| del pie a esa mujer ||| 1
inon sihuatl ||| el tobillo de esa mujer ||| 1
inon tecutli, oquichihchihuilihque ||| ese anciano , le arreglaron ||| 1
```

áhuatNl–Español

Tabla 5.5: Parte de los diccionarios estadísticos obtenidos.

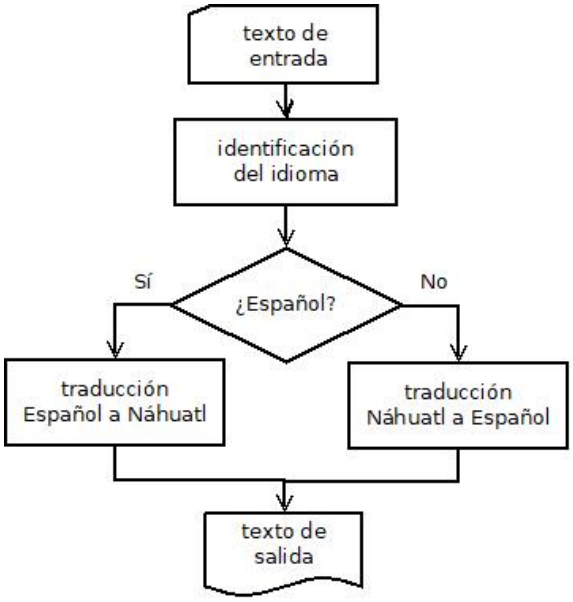

Figura 5.1: Diagrama sistema web.

quier usuario.

Se accede a la aplicación mediante la dirección IP [148.228.103.215], actualmente esta dirección es temporal, y se usa para pruebas. La aplicación, mediante código de programación PHP, valida si el texto introducido esta en español o en náhuatl. Y dependiendo del caso, hace la traducción correspondiente, NAH-ESP o ESP-NAH.

El código siguiente muestra la parte de programación de la selección del lenguaje.

```
//--------------------
<?php
$ruta_entrada=' > /traductor/traducciones/entradas.txt';
$agregar='echo ';
if(empty($_POST["source"]))
```

```
{
   $source="";
   $target="";
}
else{
   $source=$_POST["source"];
   $traducir=$agregar.$source.$ruta_entrada." 2>&1";
   $a=shell_exec($traducir);
   $b=shell_exec('python3 /traductor/traducciones/deteccion.py 2>&1');
   $idioma=shell_exec('cat /traductor/traducciones/idioma.txt 2>&1');
   if($idioma == 1){
      shell_exec('/traductor/mosesdecoder/bin/moses -f \\
                  /traductor/esp-nah/mert-work/run.moses.ini -i \\
                  /traductor/traducciones/entradas.txt > \\
                  /traductor/traducciones/salidas.txt');
      target=shell_exec('cat /traductor/traducciones/salidas.txt');
   }
   else{
      shell_exec('/traductor/mosesdecoder/bin/moses -f \\
                  /traductor/nah-esp/mert-work/run.moses.ini -i \\
                  /traductor/traducciones/entradas.txt > \\
                  /traductor/traducciones/salidas.txt');
      $target=shell_exec('cat /traductor/traducciones/salidas.txt');
```

```
    }

}

?>

//--------------------
```

Capítulo 6. Conclusiones

En este apartado se presentan brevemente los resultados obtenidos de la traducción automática estadística que se ha desarrollado, así como las conclusiones y trabajo a futuro.

Pruebas y resultados

Las pruebas se realizaron, gracias a personas oriundas de la comunidad La Resurrección, las cuales hicieron uso de esta aplicación web. Además de textos sacados de Internet, de frases simples.

En la figuras 6.1 y 6.2 se muestran ejemplos de uso del sistema para ambos casos de traducción, ESP-NAH, NAH-ESP. Se tiene en cuenta el uso simple de la herramienta mediante su uso intuitivo.

Evaluación

Para poder determinar la calidad generada por el sistema de traducción se hizo usó de la evaluación automática mediante la métrica BLEU. Para ambos casos se uso un corpus de 100 frases, en la tala 6.1 se muestra el resultado obtenido.

De los resultados obtenidos, la tabla 6.2 muestra la mejor traducción, así como

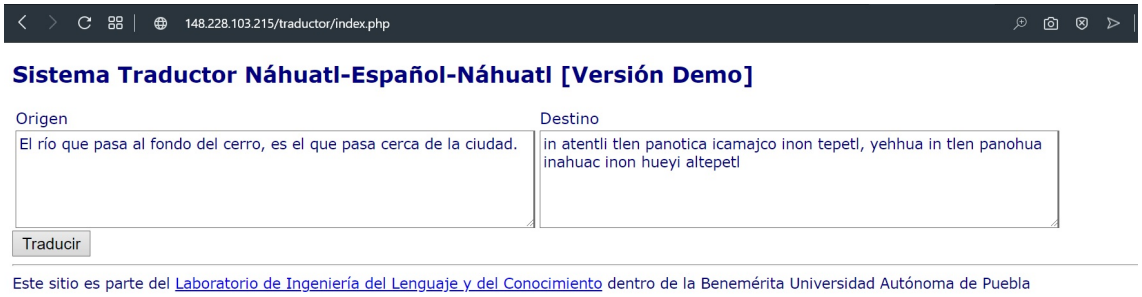

Figura 6.1: Traducción Español-Náhuatl.

Figura 6.2: Traducción Náhuatl-Español.

	BLEU
traducción español - náhuatl	0.0907098
traducción náhuatl - español	0.1360565

Tabla 6.1: Evaluación de la traducción automática.

la peor arojada por el sistema.

Score	Referencia	Traducción
1.0	El estudiante de la BUAP es muy inteligente y nunca falta a las clases.	el estudiante de la buap es muy inteligente y nunca falta a las clases.
7.262123179505913e-78	Él se dedica a hacer ejercicio en las mañanas.	el dedica hacer in las mañanas.

Tabla 6.2: Valoración de la traducciones.

Limitaciones

Durante la traducción de diferentes textos, se encontraron problemas de falta de correspondencias léxicas, así como de correspondencia gramatical y malas equivalencias.

Con textos en lengua Náhuatl extraídos de Internet, los problemas antes mencionados se incrementaron, esto se debe a lo siguiente.

Una de las limitaciones que causaron mayor problema, es la escasez de documentos digitales en la lengua náhuatl, disponiendo de un corpus muy limitado. Y estando este proyecto dirigido a un caso particular de las variantes del náhuatl, esta escasez de corpus se evidencia. Por esta importante razón, las traducciones generadas por el sistema, muestran un resultado no del todo exacto. Generar una gran cantidad de documentos supone un coste adicional de capital humano y tiempo.

Conclusión

Esta tesis gira en torno al desarrollo de una herramienta tecnológica que ayude en la traducción español-náuatl-español.

Al inicio de este proyecto, una de las dificultades que encontré fue la instalación, configuración y entendimiento de los diferentes módulos que conforman al

sistema estadístico SMT Moses. El análisis de las herramientas con las que cuenta y escoger la más adecuada, además del entendimiento de pasos a seguir para la creación del modelo de traducción a través de un corpus con lenguas de forma paralela y realizar pruebas, fue algo que también tomo tiempo.

Una segunda, y a la par de la primera, es la obtención del corpus, la cuál se logró, a partir de la transcripción literal de la lengua Náhuatl y su correspondiente traducción, lo cual se debe de agradecer, nuevamente, a las personas que brindaron su valioso tiempo y conocimientos en la lengua Náhuatl. Y otra parte, a partir de la extración de oraciones de textos en formato digital.

Una de las aportaciones generadas a partir de esta tesis se encuentra lo siguiente, la digitalización, generación y disposición de un corpus paralelo español-náhuatl de la zona centro del estado de Puebla; una aplicación en línea para la traducción del español-náhuatl de fácil acceso mediante un navegado web, y que no necesite de la asistencia de un experto en las lenguas.

Si bien, la traducción automática no sustituye la necesidad de los expertos en traducción. También se debe de comprender que la traducción realizada por expertos en idiomas tiene limitaciones. La velocidad con que los expertos en el área realizan estas tareas, no se pueden comparar con la velocidad y volumen que los sistemas de traducción pueden realizar en poco tiempo. A demás de que estos sistemas pueden ser consultados en cualquier momento, mediante el uso de servidores especializados en traducción.

En cuanto a la difusión de la aplicación, se puede afirmar que el sistema no ha tenido una divulgación. Esto se debe a que todo ha sido de manera local, en un

servidor de pruebas con solo acceso a intranet y no se ha hecho de uso de ningún otro material o escenario adicional, por ejemplo dependencias gubernamentales que brindan diferentes servicios a comunidades con lengua Náhuatl.

Líneas futuras

Como resultado de la implementación y empelo del sistema de traducción, se han detectado cumplir con más requisitos que darán mayor calidad a las traducciones generadas y que ayudarán a entregar mejores resultados,

- Implementar nuevos métodos de traducción automática, tal es el caso de la traducción automática neuronal, por ejemplo, utilizar la herramienta OpenNMT, o utilizar modelos como *Seq2seq* que también se utiliza para traducir textos cortos. Esto para hacer un análisis de las mejoras posibles en comparación a esta propuesta.

- Hacer uso de técnicas para solucionar la dificultad de escasez de datos en el corpus para la lengua náhuatl del centro del estado de Puebla, como lo es la técnica de *Data augmentation* y ver su aplicación en este tipo de corpus.

Bibliografía

Arce, A. (2018). Programación PHP. Technical report, UTN, Argentina.

Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Benavides Cañón, P. A. and Rodríguez Correa, S. (2013). Procesamiento del lenguaje natural en la recuperación de información. Technical report, Universidad de la Salle, Sistemas de Información y Documentación.

Bhattacharyya, P. (2015). *Machine Translation*. CRC Press, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL.

Brockway, E. and Hershey, T. (2018). *Diccionario Náhuatl, del estado de Puebla*, volume 42 of *vocabularios y diccionarios indígenas*. Instituto Lingüístico de Verano A.C., Tlalpan, CDMX, México, 2 edition.

Brown, P. F., Della-Pietra, S. A., Della-Pietra, V. J., and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation. *Computational Linguistics*, 19(2):263–313.

Cendejas Castro, E. A. (2013). *Alineación automática de textos paralelos a nivel de palabras usando información lingüística diversa*. PhD thesis, Instituto Politécnico Nacional, CDMX, México.

Coelho, F. (2019). Lengua. `https://www.significados.com/lengua/`.

Collins, M. (2011). Statistical Machine Translation : IBM Models 1 and 2.

De la Luz Ramírez, C. (2019). La importancia de las lenguas indígenas.

De Ávila Blomberg, A. (2008). La diversidad lingüística y el conocimiento etnobiológico. *Capital natural de México*, 1:497–556.

Denkowski, M. J. and Lavie, A. (2010). Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks.

Estrella, P. (2018). Traducción Automática. Slides.

García Aymerich, C. (2017). Lenguas Indígenas en México. Technical report, Universidad Autómoma Metropolitana, Unidad Azcapotzalco.

Gelbukh, A. (2010). Procesamiento de Lenguaje Natural y sus Aplicaciones. *Komputer Sapiens*, 1(2):1–32.

González Duque, R. (2019). Python para todos. `http://mundogeek.net/tutorial-python/`.

Gutierrez-Vasques, X., Sierra, G., and Pompa, I. H. (2016). Axolotl: a Web Accessible Parallel Corpus for Spanish-Nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Han, A. L.-F. (2017). Lepor: An augmented machine translation evaluation metric.

Hernández, N. (2011). Presencia contemporánea de los nahuas. *Arqueología Mexicana*, 19(109):53–57.

Hernández de León-Portilla, A. (2011). El náhuatl y el tronco lingüístico yutonahua. *Arqueología Mexicana*, 19(109):32–37.

IAMT (2019). International association for machine translation.

INALI (2009). Catálogo de las lenguas indígenas nacionales. variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas.

INEGI (2020). Diversidad en puebla. `http://www.cuentame.inegi.org.mx/monografias/informacion/pue/poblacion/diversidad.aspx?tema=me&e=21`.

KantanMT (2020). What is a good BLEU Score? `https://kantanmt.com/whatisbleuscore.php`. [Online; accessed 13-Enero-2020].

Knight, K. (1999). A Statistical MT Tutorial Workbook. Manuscript prepared for the 1999 JHU Summer Workshop.

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P. (2019). *MOSES Statistical Machine Translation System, User Manual and Code Guide*. University of Edinburgh.

Lerma-Blasco, R. V., Murcia Andrés, J. A., and Mifsud Talón, E. (2013). *Aplicaciones Web*. McGraw-Hill/Interamericana de España, S.L., Edificio Valrealty (Madrid), 2 edition.

Manning, C. D. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, 2 edition.

Martín Mateos, F. J. and Ruiz Reina, J. L. (2013). Procesamiento del lenguaje natural. Technical report, Universidad de Sevilla, Dpto. Ciencias de la Computación e Inteligencia Artificial.

Maučec, M. and Donaj, G. (2019). *Machine Translation and the Evaluation of Its Quality*.

Moreira, D., Cruz, I., Gonzalez, K., Quirumbay, A., Magallan, C., Guarda, T., Andrade, A., and Castillo, C. (2021). Analysis of the Current State of Natural Language Processing. *Iberian Journal of Information Systems and Technologies*, 42:126–136.

Nießen, S., Och, F. J., Leusch, G., and Ney, H. (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Olarte, C. (2019). Modelos del lenguaje. Slides.

Oliver Gonzalez, A. (2014). *Traducción y Tecnologías: procesos, herramientas y recursos*. UOC Publisher, Universitat Oberta de Catalunya.

ONU (2019). Lenguas indígenas. `https://www.un.org/indigenous`. Foro permanente para las cuestiones indígenas.

Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. (2019). Estimación de modelos de traducción de secuencias de palabras a partir de corpus muy grandes mediante THOT. *Semantic Scholar*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Parra Escartín, C. (2018). Evolución de la traducción automática. *La Linterna del Traductor*, 1(16):20–28.

Rojas, C. (2017). Review: Victoria escandell vidal. la comunicación: lengua, cognición y sociedad. pages 347 – 349.

Ríos Dolores, J. C. (2016). Traducción automática náhuatl-español: variables que influyen en la calidad de la traducción. Master's thesis, Universidad Nacional Autónoma de México, CDMX, México.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation.

Stentiford, F. W. and Steer, M. G. (1998). Language translation system and method. United States Patent.

Valle Perez, P. (1998). Pedro Carrasco, Estructura político-territorial del Imperio Tenochca. La Triple Alianza de Tenochtitlan, Tetzonco y Tlacopan. *Dimensión Antropológica*, 12:129–134. `http://www.dimensionantropologica.inah.gob.mx/?p=1356`.

Wołk, K. (2019). *Machine Learning in Translation Corpora Processing.*

Wołk, K. and Marasek, K. (2014). Real-time statistical speech translation. In *New Perspectives in Information Systems and Technologies, Volume 1*, pages 107–113. Springer International Publishing.