Universidad Nacional Autónoma de México

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

Traducción automática náhuatl-español: variables que influyen en la calidad de la traducción

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

Maestro en Ciencia e Ingeniería de la Computación

P R E S E N T A :

Julio César Ríos Dolores

TUTOR

Dr. GERARDO EUGENIO SIERRA MARTÍNEZ
INSTITUTO DE INGENIERÍA

Ciudad Universitaria, CD. MX., septiembre 2019

Para todos la luz. Para todos todo.
EZLN, Cuarta Declaración de la Selva Lacandona

A mi familia por su amor incondicional. A mi padre por su paciencia. A mi madre por su alegría. A mi hermana por su comprensión. A Romina y a Renata por su existencia.

Agradecimientos

A la Universidad Nacional Autónoma de México y al Consejo Nacional de Ciencia y Tecnología por permitirme realizar estudios de posgrado. En específico, agradezco a la beca CONACYT 482138.
A mi tutor Gerardo Eugenio Sierra Martínez, y a los miembros del jurado: Héctor Jiménez Salazar, Ivan Vladimir Meza Ruiz, Gemma Bel Enguix y Sofía Natalia Galicia Haro, por tomarse el tiempo para contribuir en la realización de este trabajo.
A todos los que con sus comentarios y conocimiento me brindaron la ayuda necesaria para poder realizar esta tesis.
Por último, y especialmente, a mi familia, que siempre me ha apoyado. Sin su respaldo el desarrollo de este trabajo no hubiese sido posible.

Resumen

Usualmente, en el campo de traducción automática se trabaja a partir de un corpus paralelo y un modelo de traducción automática. Cuando las lenguas a traducir no tienen grandes corpus paralelos para el entrenamiento de los modelos, o los corpus con los que se cuenta no cumplen con las características necesarias para garantizar buenos resultados, se tiene que recurrir a otros recursos y a métodos más sofisticados. Algunos trabajos sugieren que es posible aumentar el desempeño de los modelos de traducción automática realizando cierto tipo de preprocesamiento en los corpus de entrenamiento. Siguiendo esta línea de investigación, en la presente tesis se realizan varios experimentos, con el objetivo de identificar si es posible aumentar el desempeño de los modelos de traducción automática estándar, es decir, de la traducción automática estadística basada en frases y la traducción automática neuronal modelo secuencia a secuencia, cuando se trabaja en un escenario de escasos recursos (corpus paralelos pequeños), con un par de lenguas tipológicamente distantes (náhuatl y español). Esto a partir de la segmentación morfológica del corpus, su normalización ortográfica, el aumento del tamaño del corpus mediante un algoritmo de alineamiento automático, y el uso de un corpus monolingüe de gran tamaño.
Los resultados obtenidos muestran que es posible mejorar el desempeño de los modelos de traducción automática estadística a partir del enfoque antes mencionado, no obstante, se tienen que tomar en cuenta varios factores para lograr un aumento significativo en los resultados. En cuanto a la traducción automática neuronal, los experimentos realizados mostraron que la segmentación morfológica del corpus y la normalización ortográfica perjudican el desempeño del modelo para este caso específico.

Abstract

Usually, to work in the field of machine translation we start from a parallel corpus and a machine translation model. When the languages to translate don't have big parallel corpus for the model training, or when the corpus doesn't have the necessary characteristics to achieve good results, we need to involve more sophisticated methods and extra resources. Some works suggest that it's possible to enhance the performance of the machine translation models doing some kind of preprocessing over the training corpus. Following this line of research, in this thesis we perform several experiements with the objective to identify if it's possible to enhance the performance of the standard machine translation models, especifically, phrase-based statistical machine translation and neural sequence-to-sequence models, when we work in low resources conditions (with small parallel corpus), with a couple of typological distant languages (nahuatl and spanish). This from the morphological segmentation over the corpus, its orthographic normalization, the increase of the corpus size through an automatic alignment algorithm, and the use of a big monolingual corpus.
The results show that it's possible to achieve better results in statistical machine translation from this approach, nevertheless, must be taken into account several factors to achieve a significant increase in the results. As for the neural model, the experiments show that morphological segmentation and ortografic normalization are detrimental for the performance of the model for this specific case.

Índice

1. Introducción **8**
- 1.1. Relevancia y contribución del trabajo . . . 9
- 1.2. Objetivo general . . . 10
  - 1.2.1. Objetivos particulares . . . 10
- 1.3. Preguntas de investigación . . . 10
- 1.4. Estructura de la tesis . . . 11

2. Marco conceptual **12**
- 2.1. Procesamiento del lenguaje natural . . . 12
  - 2.1.1. Corpus lingüísticos . . . 12
    - Corpus monolingües . . . 13
    - Corpus multilingües . . . 13
  - 2.1.2. Alineamiento automático . . . 14
    - Alineamiento a nivel oración . . . 14
  - 2.1.3. Modelos de lenguaje . . . 18
  - 2.1.4. Corrección ortográfica automática . . . 18
    - Tipos de corrección ortográfica automática . . . 19
- 2.2. Redes neuronales artificiales . . . 20
  - Redes neuronales recurrentes . . . 21
- 2.3. Fundamentos lingüísticos . . . 23
  - 2.3.1. Morfología . . . 23
  - 2.3.2. Variación lingüística . . . 24

3. Traducción automática y lenguas indígenas en México **26**
- 3.1. Traducción automática . . . 26
  - 3.1.1. Traducción automática basada en reglas . . . 26
  - 3.1.2. Traducción automática basada en ejemplos . . . 26
  - 3.1.3. Traducción automática estadística . . . 27
  - 3.1.4. Traducción automática neuronal . . . 27
  - 3.1.5. Evaluación . . . 28
    - BLEU (*Bilingual Evaluation Understudy*) . . . 28
- 3.2. Lenguas indígenas en México . . . 30
  - 3.2.1. Lengua náhuatl . . . 30
  - 3.2.2. Desarrollo tecnológico para lenguas indígenas en México . 31

4. Modelos de traducción automática basados en datos **34**
- 4.1. Modelos de traducción automática estadística . . . 34
  - Planteamiento del problema . . . 34
  - Traducción automática estadística basada en palabras . . 35
  - Modelos IBM . . . 36
  - Traducción automática estadística basada en frases . . . 41
- 4.2. Modelos de traducción automática neuronal . . . 43
  - Modelo *secuencia a secuencia* . . . 43
  - Mecanismos de atención . . . 44

Atención global 44
Atención local 46
4.3. Traducción automática basada en datos y escenarios con escasos recursos digitales 47

5. Métodos y herramientas empleadas **49**
5.1. Marco teórico 49
5.2. Obtención y preprocesamiento del corpus Español-Náhuatl 49
5.3. Modelos de traducción automática 50
5.4. Implementación y experimentación 50
5.4.1. Normalización ortográfica 50
Algortimo de normalización ortográfica 51
Evaluación de la normalización ortográfica 52
5.4.2. Segmentación morfológica 52
Evaluación de la segmentación morfológica 53
5.4.3. Traducción automática 54
Traducción automática estadística (software) 54
Traducción automática neuronal (software) 55
5.5. Evaluación y análisis de resultados 55

6. Resultados **56**
6.1. Corpus paralelo español-náhuatl 56
6.1.1. Segmentación del corpus 57
6.2. Corpus monolingües 58
6.3. Alineamiento automático 59
6.4. Normalización ortográfica 59
6.4.1. Evaluación de la normalización ortográfica 60
6.4.2. Conglomerados de referencia 60
6.5. Segmentación morfológica 61
6.5.1. Evaluación de la segmentación morfológica 62
6.6. Traducción automática 64
6.6.1. Evaluación de la traducción automática 65
6.7. Discusión 68

7. Conclusiones **72**

A. Anexo: Corpus anotado de variantes ortográficas **84**

1. Introducción

Dentro de la naturaleza existe gran diversidad en la forma de comunicarse entre las especies. Esta comunicación puede darse, por ejemplo, a través de señales visuales en los chimpancés, señales químicas en las plantas, sonoras entre el humano y el perro, entre otras. No obstante, de entre todas las formas de comunicación existentes quizá la más destacada sea el lenguaje humano. Esto se debe no sólo a la cantidad de información que es posible transmitir a través de este sistema de comunicación, sino también a los procesos neurológicos involucrados, sus implicaciones históricas y un largo etcétera. No cabe duda que una de las características que nos separa del resto de las especies es el lenguaje. Como señala Concepción Company (2017), un hecho bastante fascinante que no se repite en ninguna otra especie es nuestra capacidad de hacer frente a expresiones sin precedentes en nuestra experiencia lingüística, es decir, nuestra capacidad de *sintaxis libre*. Además de esto, el lenguaje nos transforma en seres históricos, capaces de conocer indirectamente al mundo, ya que nos permite echar un vistazo a través del tiempo, y saber no sólo que ha ocurrido en el pasado, sino también expresar nuestras predicciones acerca del futuro.

El lenguaje también implica en el ser humano una forma de crear identidad. Inmediatamente después de nacer se hace imperiosa la necesidad de comunicarse con el mundo, por cuestiones simplemente de supervivencia, pero conforme pasa el tiempo, surge la necesitad de expresar ideas cada vez más complejas, por lo que el aprendizaje de un idioma es una consecuencia natural en nuestro desarrollo. El idioma o la lengua poco a poco va forjando al individuo ya que lo involucra dentro de un contexto social, sin embargo, esta adopción al mismo tiempo lo aísla de otros; y es precisamente este punto el que da lugar a esta tesis.

Un fenómeno bastante estudiado es el de las *lenguas en contacto*. Para un estudiante de nivel superior en México resulta una necesidad tener conocimiento del inglés, ya que muchos de los libros o artículos que consultará durante su formación estarán redactados en esta lengua; el mismo obstáculo se le presenta a un hablante de náhuatl que tiene que ir a la Ciudad de México en busca de trabajo; o a un hablante de Otomí que comercia con comunidades Nahuas y Matlatzincas. Dicho fenómeno, si bien es cierto, es una consecuencia natural del mundo sumamente interrelacionado en el que nos desenvolvemos, también representa un impedimento que no siempre es fácil de sortear. En México existe un escenario bastante peculiar en lo que se refiere a esta cuestión. Éste es un territorio en donde un sector importante de la población tiene por lengua materna alguna lengua indígena, lo que quiere decir que el contacto entre hablantes de distintos idiomas es inevitable. A pesar de esto, las condiciones para que dichos hablantes se desenvuelvan con facilidad dentro del territorio nacional aún no existen. De hecho sucede todo lo contrario, ya que éstos se encuentran aislados y con poco apoyo en un ambiente de discriminación y rechazo. Lo que los mantiene en una situación de desventaja, desplazamiento e invisibilización. Pese a lo anterior, han existido algunos intentos para enfrentar este problema, como por ejemplo: la promulgación de leyes, creación de institutos, y en última

instancia el desarrollo científico y tecnológico.
Es claro que la necesidad de comunicación entre personas que no hablan la misma lengua ha sido un problema importante dentro de la ciencia. Éste se ha abordado a partir de distintos enfoques y uno de ellos es el desarrollo de traductores automáticos, tema que específicamente interesa a esta tesis. En el caso particular de México, a pesar de que la investigación en este campo tiene una aplicación directa en los problemas que se presentan en el día a día, la inversión en el desarrollo de tecnologías del lenguaje, y específicamente para lenguas indígenas, es bastante limitada. Como se verá más adelante, esta tesis busca abonar en esta área, a partir de la experimentación en el campo de la traducción automática con las lenguas náhuatl y español, lo cual se espera no sólo sea una contribución para el área sino también motivo de futuras investigaciones.

1.1. Relevancia y contribución del trabajo

El desarrollo de tecnologías de la información para las lenguas indígenas es un tema poco atendido que plantea grandes retos para la comunidad científica. Este trabajo busca abordarlo de manera parcial limitándose a la traducción automática y a la normalización ortográfica automática.
Es importante mencionar que la mayoría de las lenguas indígenas en México no cuentan con una norma de escritura, por lo tanto, al tratar de abordarlas desde el procesamiento del lenguaje natural (PLN), la variación ortográfica en los recursos necesarios para el desarrollo de tecnología representa un problema importante que no se ha trabajado lo suficiente, es por esto que aquí se propone una variante de un algoritmo de normalización ortográfica y la evaluación de su desempeño.
Por otra parte, en lo que se refiere al área de traducción automática para estas lenguas, si bien existen trabajos que buscan atacar los obstáculos en este tema (Mager, 2017; Mager & Meza, 2018; Gutierrez-Vasques, 2017), a partir del preprocesamiento del corpus con base en hipótesis fundamentadas en las características de la lengua, no existe un trabajo que realice un análisis cuantitativo de la variación en el desempeño de los modelos de traducción, específicamente para el español y el náhuatl clásico, cuando dichos modelos son entrenados a partir de corpus preprocesados de maneras distintas o con diferentes características como pueden ser: tamaño, distintos tipos de segmentación morfológica, etc. En esta investigación se plantea realizar el análisis antes mencionado con el propósito de determinar bajo qué circunstancias vale la pena o no realizar el preprocesamiento del corpus, modificar ciertas variables o buscar recursos complementarios para mejorar la calidad de la traducción.
Esta tesis no sólo pretende realizar una contribución en el campo del procesamiento del lenguaje natural, sino también busca señalar la falta de atención hacia un sector de la población y sumarse a la corriente que busca un desarrollo científico favorable para todos.

1.2. Objetivo general

Este trabajo busca realizar un análisis con respecto a la calidad de las traducciones obtenidas a partir de distintos modelos de traducción automática cuando se trabaja con un par de lenguas tipológicamente distintas bajo un esquema de escasos recursos digitales.
Lo que se busca cuantificar es el impacto que tiene en el rendimiento de los modelos de traducción automática la modificación de algunas variables, que de acuerdo con otros trabajos podrían influir en el resultado.

1.2.1. Objetivos particulares

- Evaluar el desempeño de distintos modelos de traducción automática, modificando algunas variables y preprocesando de distintas maneras el corpus de entrenamiento, trabajando con las lenguas español y náhuatl[1],
- proponer una variante para un algoritmo de normalización ortográfica,
- crear un corpus anotado de variantes ortográficas para náhuatl clásico,
- analizar el impacto de la normalización ortográfica en los modelos de segmentación morfológica,
- realizar segmentación morfológica y normalización ortográfica en el corpus para medir su impacto en los modelos de traducción,
- realizar un análisis cuantitativo de los datos.

1.3. Preguntas de investigación

- Cuando se trabaja en el campo de traducción automática, usualmente se parte de un corpus paralelo y cierto preprocesamiento, ¿es posible aumentar la calidad de la traducción utilizando información lingüística en dicho preprocesamiento?
- Al tratar de realizar traducciones para un par de lenguas, donde una de ellas es polisintética, ¿la segmentación morfológica de ésta ayuda a obtener mejores resultados en cualquier escenario?
- ¿Qué tanto impacta la variación ortográfica en algunos de los modelos de traducción automática más populares, cuando se trabaja en un escenario de escasos recursos?
- En el contexto de lenguas de escasos recursos digitales, particularmente para las lenguas indígenas en México, ¿los modelos de traducción automática estándar son capaces de brindar resultados suficientemente buenos como para ser utilizados en la práctica?

[1] Variante: Náhuatl clásico.

1.4. Estructura de la tesis

Este trabajo está conformado por siete capítulos. El primero plantea las motivaciones e interrogantes que dieron pie a la tesis, así como los objetivos que desea alcanzar. El segundo capítulo corresponde al marco conceptual, y en éste se compendian los requerimientos básicos de lingüística y procesamiento del lenguaje natural necesarios para la fácil lectura de este trabajo. En el tercer capítulo es donde se entra de lleno al tema específico que interesa a la tesis, haciendo un recuento general de los paradigmas que han guiado el área de la traducción automática, así como un breviario acerca del contexto en el que las lenguas indígenas en México se desenvuelven. Dado que el tema principal de la tesis es la traducción automática, el cuarto capítulo la aborda de manera más específica, profundizando en los *modelos de traducción automática basados en datos* que se utilizaron en esta tesis. El capítulo cinco expone tanto los métodos como las herramientas utilizadas para la construcción de un traductor automático, que van desde la obtención y el preprocesamiento de los datos, hasta el entrenamiento de los modelos de traducción automática y su evaluación. En el sexto capítulo se muestran los resultados obtenidos, así como su análisis. Por último se tienen las conclusiones y la bibliografía.

2. Marco conceptual

En este capítulo se busca compendiar los requerimientos básicos de lingüística y procesamiento del lenguaje natural, necesarios para la fácil lectura de este trabajo.

2.1. Procesamiento del lenguaje natural

De acuerdo con Jurafsky y Martin (2006) el procesamiento del lenguaje natural tiene por objetivo lograr que las computadoras sean capaces de realizar tareas en donde el lenguaje humano juega el papel principal, como pueden ser: agentes capaces de simular conversaciones (agentes conversacionales), sistemas que permitan realizar traducciones en distintas lenguas (traducción automática), etcétera. Para esto se suele trabajar a partir de un enfoque multidisciplinario en donde se intersectan áreas como la lingüística, ciencias de la computación y matemáticas, por mencionar algunas.
El procesamiento del lenguaje natural es una tarea difícil ya que debe lidiar con un conjunto de estructuras del lenguaje bastante complejas, lo que ocasiona retos monumentales al momento de desarrollar modelos o sistemas propios del área. Esto sucede, ya que trabajar con dichas estructuras de manera eficiente y sin caer en el reduccionismo no es una tarea trivial. Además de esto, el lenguaje humano tiene otras características, como por ejemplo: su ambigüedad, que plantean limitantes continuas dentro de la investigación.
Actualmente, algunos autores consideran que de lograr que las computadoras procesen el lenguaje con un dominio tan amplio como que el que tienen los humanos, se estaría dando uno de los pasos más importantes en pro de la inteligencia artificial (Jurafsky & Martin, 2006), sin embargo, aún existe dentro del área una gran cantidad de problemas abiertos, por lo que el camino a seguir aún es largo.

2.1.1. Corpus lingüísticos

Un recurso muy importante dentro del PLN son los corpus lingüísticos. Éstos suelen servir como materia prima para el desarrollo científico y tecnológico en esta área, y usualmente cuando se trabaja bajo este enfoque, la calidad de dicho recurso tiene un papel importante tanto en los resultados como en su escalabilidad.
Un corpus lingüístico se define como un conjunto de textos de materiales escritos y/o hablados, debidamente recopilados para realizar ciertos análisis lingüísticos (Sierra Martínez, 2017).
Antes de conformar un corpus lingüístico se tienen que tomar en cuenta varios factores con el fin de garantizar la calidad de dicho recurso. Torruella y Llisterri (1999) ponen atención en doce aspectos, que de acuerdo con ellos son primordiales al momento de conformar un corpus, éstos son: finalidad, límites del corpus, tipo de corpus, proporciones de los diferentes grupos temáticos del corpus, po-

blación y muestra, número y longitud de los textos de la muestra, captura de los textos y etiquetado, procesamiento del corpus, crecimiento del corpus, *hardware* y *software*, aspectos legales, y por último, presupuesto y etapas.
Los corpus lingüísticos pueden clasificarse de acuerdo con distintos criterios, como por ejemplo: por el origen de los datos, la especificidad de los elementos, su temporalidad, la lengua, la representatividad, entre otros (Sierra Martínez, 2017).
De acuerdo con Sierra Martínez (2017), al clasificar los corpus lingüísticos tomando como criterio la lengua de sus elementos, podemos dividirlos en dos grandes grupos: corpus monolingües y corpus multilingües.

Corpus monolingües

Como su nombre lo indica, los corpus monolingües recopilan información de una sola lengua, haciéndolos más comunes que los corpus multilingües y más fáciles de recopilar. Dentro de esta categoría podemos además entrar en una subclasificación que nos lleva a los corpus monolingües según la variedad dialectal, es decir, en este caso los elementos del corpus se diferencian por dialectos o variedades lingüísticas; y a los corpus monolingües comparables, los cuales se componen de textos originales de una lengua y traducciones de otros textos semejantes en la misma lengua (Sierra Martínez, 2017).
La información lingüística de los corpus monolingües suele ser explotada en distintas áreas. En el PLN son ampliamente utilizados en la construcción de modelos de lenguaje, representación vectorial de palabras, recuperación de información y muchas otras más.

Corpus multilingües

Los corpus multilingües pueden ser textos en distintos idiomas o paralelos. En el primer caso el corpus es un conjunto de textos en varios idiomas que se recopilan con base en criterios bastante diversos; y en el segundo, los corpus son una misma colección de textos recopilada en más de una lengua (Sierra Martínez, 2017).
Los corpus paralelos son de interés en esta tesis, por lo que vale la pena ahondar un poco más al respecto.

Corpus Paralelos

De manera más específica, los corpus paralelos son una colección de textos traducidos a una o varias lenguas, de modo que el más sencillo de éstos consta del texto original y su traducción a otra lengua. Es importante mencionar que este tipo de corpus puede contener textos traducidos de la lengua **A** a la lengua

B y textos traducidos en la dirección contraria (Torruella & Llisterri, 1999). Resultan de utilidad ya que contienen, de manera implícita, una gran cantidad de información lingüística que puede ser aprovechada tanto en tareas del PLN como en otras áreas. La traducción automática, extracción léxica automática, desambiguación lingüística y traducción humana son sólo algunos ejemplos en donde este recurso es utilizado.

2.1.2. Alineamiento automático

En principio, al trabajar con un corpus paralelo sólo podemos garantizar que el contenido semántico de un texto corresponde con el de su traducción. No obstante, en algunas ocasiones se desea más información con respecto a las correspondencias dentro de los textos, o dicho de otra manera, se busca que estén alineados hasta cierto nivel. Existen corpus paralelos construidos tomando en cuenta estas necesidades y podemos encontrarlos con un alineamiento desde un nivel de texto completo hasta niveles más granulares como por ejemplo: alineamientos a nivel de oración e incluso a nivel de palabra.
Para tareas especificas dentro del PLN, como por ejemplo la traducción automática, los corpus paralelos suelen utilizarse con un alineamiento a nivel de oración, sin embargo, muchas veces es difícil encontrarlos alineados a ese nivel de granularidad dentro de cierto dominio, es por esto que para abordar dicho problema se han desarrollado distintos métodos de alineamiento automático.

Alineamiento a nivel oración

Según Manning y Schütze (1999) el alineamiento de texto es el primer paso al momento de utilizar corpus paralelos. Este problema no es trivial ya que las correspondencias no siempre son uno a uno y el orden también suele jugar un papel importante al momento de alinear los textos.
En el alineamiento a nivel de oración, lo que se busca es poder encontrar qué grupo de oraciones en una lengua corresponde en contenido con otro grupo de oraciones en otra lengua. Cabe mencionar que dichos grupos pueden no corresponder con ninguna oración en la otra lengua, por lo que también se necesita tomar en cuenta tanto la eliminación de oraciones como su inserción. El caso más común dentro de este alineamiento es la correspondencia uno a uno (1:1), no obstante, suelen también estar presentes alineamientos 2:1, 1:2, 2:2 e incluso 1:3 o 3:1, sin ser éstos los únicos posibles casos (Manning & Schütze, 1999).
Los métodos de alineamiento automático pueden clasificarse de acuerdo con varios criterios. En este trabajo me limitaré a abordarlos siguiendo el trabajo de Manning & Schütze, el cual los divide en dos categorías: métodos basados en longitud y métodos basados en léxico.

Métodos basados en longitud

Como su nombre lo índica estos métodos toman en cuenta comparaciones entre las longitudes de cada oración para formar el alineamiento, tomando como hipótesis el supuesto de que oraciones cortas serán traducidas a oraciones cortas y oraciones largas serán traducidas en oraciones largas (Manning & Schütze, 1999).
Trabajar bajo este paradigma ofrece además de eficiencia computacional, una alternativa para aquellas lenguas que cuentan con pocos recursos.
Es importante mencionar que los métodos clasificados dentro de esta categoría por Manning & Schütze toman como base el algoritmo de Gale & Church, y a partir de éste buscan obtener un mejor alineamiento utilizando distintos recursos. Los autores argumentan que se realizó la clasificación de esta manera, ya que en estos métodos la base del alineamiento recae en un algoritmo basado en longitud, y los recursos utilizados para aumentar su desempeño son secundarios. Por otro lado, en los métodos basados en léxico dichos recursos tienen el papel principal.

Gale & Church (1993)

El objetivo de este algoritmo es encontrar el alineamiento más probable **A** dados dos textos paralelos **S** y **T**, es decir:

$$argmax_A P(A|S,T)$$

Para calcular la probabilidad de cierto tipo de alineamiento, Gale y Church proponen medir la longitud de las oraciones en términos de caracteres. El algoritmo supone que sólo son posibles alineamientos del tipo 1:1, 1:0, 0,1, 2:1, 1:2 y 2:2, por lo que cualquier otro tipo de alineamiento no es considerado.
En su artículo (Gale & Church, 1993) demuestran que es posible calcular el alineamiento más probable dentro de los supuestos planteados utilizando programación dinámica. Esto a partir de la definición de una medida de costo y su minimización.

Brown et al. (1991)

Este enfoque es bastante parecido al propuesto por Gale y Church con la diferencia de que las comparaciones entre la longitud de las oraciones son medidas en términos de palabras y no de caracteres. En su artículo Gale y Church argumentan que la varianza del número de palabras en las traducciones es mayor a la varianza del número de caracteres, por lo que medir la longitud de las oraciones en términos de palabras empeorará los resultados del alineamiento (Gale & Church, 1993), sin embargo, la diferencia con el trabajo de Brown y otros radica en que este último no busca el alineamiento total de los textos, sino más bien producir un subconjunto del corpus alineado (Brown et al., 1991).
Ellos proponen hacer un alineamiento más preciso utilizando léxico como punto

de referencia entre las oraciones alineadas e ignorar las secciones que no se alinearon de manera correcta.

Wu (1994)

En su artículo Wu aplica el algoritmo de Gale y Church a un corpus paralelo inglés-cantonés y con base en los experimentos ahí expuestos argumenta que las suposiciones hechas por Gale y Church no son evidentes cuando se trabaja con estas lenguas.
Los resultados obtenidos en este experimento son similares a los reportados por Gale y Church, no obstante, Wu propone además incorporar información léxica dentro del modelo original de Gale y Church para mejorar su rendimiento (Manning & Schütze, 1999). Esto se realiza partiendo de una nueva hipótesis, en la cual se supone que el alineamiento ya no dependerá sólo de la longitud de las oraciones, sino también de un conjunto de pistas léxicas (*lexical cues*).
Bajo esta nueva hipótesis el desarrollo teórico es bastante parecido al mostrado por Gale y Church; y el alineamiento final se obtiene utilizando programación dinámica, sin embargo, los requerimientos en tiempo y en memoria crecen linealmente en función con el número de pistas léxicas utilizadas, por lo que resulta importante utilizar el menor número de pistas léxicas posibles sin comprometer el desempeño del nuevo modelo (Wu, 1994). Wu menciona dos principales factores a tener en cuenta al momento de definir las pistas léxicas: el primero hace referencia a su precisión; y el segundo a la frecuencia, lo que quiere decir que este proceso de elección depende fuertemente del dominio bajo el que se está trabajando.

Métodos basados en léxico

A diferencia de los métodos anteriores, los métodos basados en léxico utilizan información léxica como herramienta principal para hacer los alineamientos (Manning & Schütze, 1999). Estos métodos suelen ser más robustos, sin embargo, para aplicarlos es necesario contar con información léxica para las lenguas involucradas, lo que en algunos casos no es sencillo de conseguir.

Kay & Röscheisen (1993)

En este método se utiliza un alineamiento parcial del léxico para inducir el alineamiento de las oraciones. En concreto involucra un proceso de convergencia donde un alineamiento parcial a nivel palabra induce un alineamiento de máxima verosimilitud a nivel de oración, que es utilizado para refinar el alineamiento a nivel palabra y repetir el proceso recurrentemente. El alineamiento a nivel de palabra se basa en la suposición de que dos palabras empatan si su distribución es la misma (Manning & Schütze, 1999; Kay & Röscheisen, 1993).

Chen (1993)

Chen propone hacer el alineamiento mediante la construcción de un modelo estadístico básico de traducción palabra por palabra (Chen, 1993). En donde el mejor alineamiento será aquel que maximice la probabilidad de generar el corpus con el que se está trabajando, dado el modelo de traducción, o dicho de otra manera, lo que se desea es que una vez parametrizado este modelo de traducción, sea posible inducir el alineamiento más probable a partir del cálculo de las probabilidades de traducción entre oraciones bajo distintos alineamientos, estableciendo como alineamiento final a aquel que maximiza la probabilidad conjunta de tener el corpus dado y un alineamiento específico.
En este artículo, el autor argumenta que los métodos basados en longitud no son suficientemente robustos, y los métodos basados en léxico disponibles en ese momento no son suficientemente rápidos como para un uso practico, por lo que con su propuesta logra atacar estos problemas y además obtener mejores resultados que los métodos antes descritos (Manning & Schütze, 1999; Chen, 1993).

Haruno & Yamazaki (1996)

Haruno y Yamazaki argumentan que los métodos anteriores no funcionan como deberían cuando se trabaja con lenguas tipológicamente distintas por lo que proponen un método que puede ser considerado una variante del propuesto por Kay y Röscheisen. Ellos mencionan que para lenguas de este estilo, como pueden ser el japones y el inglés, incluir palabras funcionales en la extracción léxica dificulta el alineamiento, por lo que los autores las excluyen, lo cual se logra mediante el uso de etiquetadores (un recurso extra con el que se tiene que contar). Además de esto, los autores mencionan que cuando se desea alinear textos cortos, no hay suficiente repetición en las palabras como para que la suposición hecha por Kay y Röscheisen acerca de la distribución de las mismas se cumpla, por lo que usan un diccionario (Manning & Schütze, 1999; Haruno & Yamazaki, 1996).
Es importante mencionar que este método comienza a moverse en un esquema dentro del PLN en donde el conocimiento lingüístico involucrado en los métodos comienza a acrecentarse de manera importante (Manning & Schütze, 1999).

2.1.3. Modelos de lenguaje

Otro problema con grandes implicaciones dentro del PLN consiste en poder estimar una distribución de probabilidad sobre una secuencia de palabras, o para ponerlo en términos propios del área, estimar un modelo de lenguaje.
Esta tarea resulta de utilidad en varias áreas, entre las que destacan: reconocimiento del habla, traducción automática, etiquetado gramatical, entre otros (Goodman, 2001).
El objetivo de los modelos de lenguaje es estimar la probabilidad de una secuencia de palabras $p(w_1...w_i) = p(w_1)p(w_1|w_2)...p(w_i|w_{i-1}, ...w_1)$, y se ha abordado desde varias perspectivas. Una de las más populares es trabajar a partir de *n-gramas*. Bajo este enfoque se asume una propiedad de Markov de orden n para las secuencias de palabras, es decir, la probabilidad de la palabra $n+1$ dentro de la secuencia depende únicamente de las n anteriores, y la estimación de dicha probabilidad suele realizarse mediante el conteo de los *n-gramas* en un corpus de entrenamiento y algunas técnicas de suavizamiento o *smoothing* (Goodman, 2001; Katz, 1987; Ney et al., 1994; Jelinek et al., 1991).
Por otro lado, un enfoque que ha tomado relevancia en los últimos años es trabajar en este problema usando redes de neuronas artificiales. Uno de los mayores obstáculos para los modelos basados en *n-gramas* es el efecto Hughes (*curse of dimensionality*). Un ejemplo claro es mostrado en el artículo *A Neural Probabilistic Language Model* (Bengio et al., 2003), en donde se menciona que al buscar la distribución conjunta de tan solo 10 palabras consecutivas, tomando en cuenta un vocabulario V de tamaño $10,000$, existen hasta $10,000^{10} - 1$ parámetros libres. Además de lo anterior también hay otros inconvenientes como por ejemplo la dispersión de los datos o las dependencias de largo alcance dentro de los contextos que usualmente no son tomadas en cuenta por estos modelos.
En el caso neuronal usualmente se busca aprender tanto una representación vectorial de las palabras como la distribución de probabilidad de las secuencias de palabras en términos de dichas representaciones (Bengio et al., 2003). Esto busca enfrentar los problemas antes mencionados, ya que las secuencias de palabras que no fueron observadas en el entrenamiento tendrán una probabilidad alta cuando dichas palabras sean similares (en términos de sus representaciones vectoriales) a palabras ya observadas, y además el uso de estructuras neuronales recurrentes diseñadas para capturar dependencias a largo plazo resultan de utilidad al modelar contextos con este tipo de dependencias (Khandelwal et al., 2018; Mikolov et al., 2010). Actualmente los modelos de lenguaje neuronales son un área muy activa dentro de la investigación y han demostrado un desempeño superior con respecto a los modelos clásicos en algunos escenarios (Dauphin et al., 2016; Józefowicz et al., 2016; Melis et al., 2017).

2.1.4. Corrección ortográfica automática

Un tema bastante estudiado en el procesamiento del lenguaje natural es la corrección ortográfica automática. Esta es una tarea que busca detectar errores

ortográficos dentro de un conjunto de palabras, y usualmente suele anclarse en dos enfoques: en el primero se busca hacer la corrección utilizando la opción valida (correctamente escrita) más *cercana* a la palabra en cuestión, lo cual hace necesario definir, de alguna manera, una forma medir dicha cercanía; en el segundo, lo que se busca es realizar la corrección con la palabra más frecuente (y en términos frecuentistas, más probable) que se relacione con la palabra analizada (Manning et al., 2008).
Esta tarea se usa de manera importante en áreas como minería de textos, recuperación de información y como se verá más adelante, el desarrollo en este campo puede además ser aprovechado en el área de la normalización ortográfica.

Tipos de corrección ortográfica automática

La corrección ortográfica automática pude clasificarse en dos tipos: por palabras aisladas y palabras en contexto (Manning et al., 2008). En el primer caso, como su nombre lo indica, se realizarán las correcciones de las palabras de manera aislada, es decir, tomando únicamente la información provista por la palabra a corregir; en el segundo caso, además de la información brindada por la palabra, se toma en cuenta su contexto[2].
Por motivos prácticos en este trabajo se abordará sólo la corrección ortográfica de palabras aisladas, ya que es precisamente este enfoque el que tiene implicaciones directas en los resultados de esta tesis.

Corrección ortográfica de palabras aisladas

Como ya se mencionó, cuando se trabaja la corrección ortográfica de palabras asiladas, solamente se toma en cuenta la información que nos brinda la palabra a analizar. Para estos fines existen varias maneras de enfrentar el problema, en este trabajo se abordarán dos: la primera se basa en una distancia de edición (*edit distance*); y la segunda en el superposicionamiento de *k-gramas* (*k-gram overlap*).

Distancia de edición

Dadas dos palabras o cadenas de caracteres, supóngase s_1 y s_2, la distancia de edición entre éstas es el mínimo número de operaciones de edición que se requieren para transformar a s_1 en s_2. Las operaciones de edición más comunes suelen ser: insertar un carácter en la cadena; borrar un carácter; remplazar un carácter por otro. Si la distancia de edición está dada por estas operaciones, se le conoce como distancia *Levenshtein*. Una generalización de esta distancia podría

[2]El contexto de una palabra se refiere al conjunto de palabras que la "rodean", es decir, una vecindad de radio variable con centro en la palabra.

ser a través de la ponderación de las operaciones tomando como referencia la probabilidad de realizar cada operación.

Superposicionamiento de *k-gramas*

Un *k-grama* es una subsecuencia de longitud k de una secuencia dada, por ejemplo los *2-gramas* o *bigramas* de la palabra *tesis*, corresponderían al conjunto: $\{te, es, si, is\}$. Bajo el enfoque de superposicionamiento de *k-gramas* podemos definir que tan "cercanas" son las palabras mediante la comparación del número de *k-gramas* que tienen en común. Una medida de distancia basada en este enfoque es la distancia *Sorensen* o coeficiente *Sorensen-Dice*.
Bajo el contexto de comparación de secuencias o como medida de similitud entre cadenas de caracteres, ésta se define a través del uso de bigramas de la siguiente manera:

$$s = \frac{2n_t}{n_x + n_y}$$

Donde:

- n_t es la cardinalidad de la intersección de *bigramas* entre las secuencias x y y,
- n_x y n_y representan la cardinalidad del conjunto de *bigramas* de la secuencia x y la secuencia y, respectivamente.

Una vez establecido lo anterior, ya tenemos una medida formal para establecer que tan "cercanas" son las palabras, ya sea bajo un enfoque u otro, y con base en esto hacer la corrección ortográfica correspondiente, intercambiando la palabra *mal escrita* con la palabra correcta más *cercana*.

2.2. Redes neuronales artificiales

Debido a la enorme capacidad de cómputo y a las grandes cantidades de información con las que contamos actualmente, las redes neuronales artificiales han tenido un gran auge durante los últimos años en un sinnúmero de áreas y el procesamiento del lenguaje natural no es la excepción. Problemas como la traducción automática, modelado del lenguaje, reconocimiento del habla, resumen de texto automático, por sólo mencionar algunos, han sido abordados utilizando estas herramientas, logrando en algunos casos resultados superiores con respecto a los modelos tradicionales (Vaswani et al., 2017; Dauphin et al., 2016; Graves et al., 2013; Dong, 2018).
En lo que respecta a esta tesis, es importante ahondar en las redes neuronales recurrentes (RNN) con el fin de facilitar la comprensión del modelo *secuencia a secuencia* que será abordado posteriormente.

Redes neuronales recurrentes

Una RNN es un arquitectura neuronal en donde hay retroalimentación en la información que se está procesando, es decir, este tipo de red toma en cuenta los datos que está recibiendo en cierto momento, pero también los datos que ya procesó (vease figura 1).
Como se puede apreciar en la figura 1, una RNN puede ser vista como un conjunto de múltiples copias de una misma celda, donde cada una de éstas pasa información a la siguiente mientras se van procesando las entradas de una en una. En este esquema **A** representa la celda básica, x_t la t-ésima entrada y h_t corresponde al t-ésimo valor de salida.

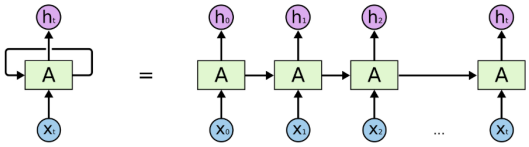

Figura 1: Ilustración de una RNN desplegada tomada de (Olah, 2015).

Un problema bastante conocido que se presenta con este tipo de redes es el problema de dependencias a largo plazo o *long-term dependency problem*, éste surge cuando el espacio entre la información relevante para el problema y el momento en el que ésta se tiene que utilizar es relativamente grande, lo que quiere decir que la red tiene que ser capaz de aprender a resolver el problema en cuestión utilizando tanto la información que se acaba de procesar como la información que se procesó tiempo atrás. Entre más grande es este espacio, más difícil es que la red aprenda. Varios modelos han surgido para tratar de dar solución a este problema entre los que destacan las redes *Long-Short Term Memory* o **LSTM** (Hochreiter & Schmidhuber, 1997) y las *Gated Recurrent Unit* o **GRU** (Cho et al., 2014).

Long-short term memory (LSTM)

Las redes **LSTM** son un tipo de RNN propuestas por Hochreiter y Schmidhuber (1997) que buscan aprender dependencias a largo plazo dentro de los datos a partir de la incorporación de ciertos elementos dentro de las celdas.
Al igual que en el modelo estándar de una RNN la estructura de esta red puede ser vista en cierto modo como una cadena donde cada entrada es procesada por una celda, la cual obtiene una salida que a su vez sirve para alimentar a la siguiente celda (vease figura 2) (Olah, 2015).
Lo interesante de este tipo de red es la definición de las celdas que procesan la información. En éstas se incorporan el estado t-ésimo de la celda C_t, en el cual se modifican los elementos del estado anterior a partir de una transformación de la t-ésima entrada x_t; la salida h_{t-1}; una compuerta de olvido f_t, la cual

aprenderá en que proporción los datos del estado anterior siguen siendo relevantes con respecto a la entrada actual y a la salida anterior; y por último una compuerta de entrada i_t, la cual será la encargada de realizar el nuevo aporte de información al estado de la red con base en un nuevo valor candidato $\tilde{C}_t$. Dichos componentes son definidos por los autores de la siguiente manera:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = tanh(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Donde:

- σ representa la función sigmoide,
- $tanh$ representa la función tangente hiperbólica,
- $W.$ y $b.$ la matriz de pesos y el sesgo de una red completamente conectada (Goodfellow et al., 2016),
- C_t es el nuevo estado de la celda.

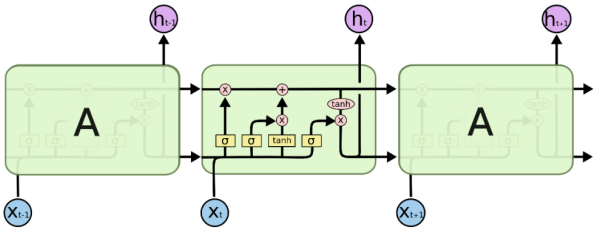

Figura 2: Ilustración de una LSTM tomada de (Olah, 2015).

2.3. Fundamentos lingüísticos

Por último, y antes de entrar de lleno al tema de traducción automática, será necesario dar una breve introducción a algunos conceptos propios de la lingüística, ya que serán de utilidad para la comprensión de las secciones posteriores.

2.3.1. Morfología

La morfología en términos lingüísticos se encarga del estudio de la estructura interna de las palabras. Se dice que existe estructura morfológica cuando hay grupos de palabras que presentan regularidades tanto en forma como en significado de manera sistemática, es decir, dichos patrones no ocurren de manera aleatoria (Haspelmath, 2002).
El análisis morfológico consiste en la identificación de las partes de la palabra, o dicho en términos lingüísticos, sus constituyentes. A los constituyentes más pequeños con significado se les denomina morfemas, así que en estos términos se podría decir también que la morfología es el estudio de la combinación de morfemas para producir palabras.
Es importante mencionar que el término que se está abordando no sólo se utiliza para referirse a una rama de la lingüística sino también para denotar una parte del sistema de lenguaje, por lo que se puede hablar de la morfología de las lenguas, como por ejemplo: la morfología del náhuatl.
Dicho lo anterior resulta evidente que la morfología no es igual en todas las lenguas. Lo que en una lengua se expresa morfológicamente en otra puede ser expresado usando varias palabras o de manera implícita, es por esto que dentro de la lingüística suelen usarse términos como morfología analítica y morfología sintética para describir el grado en el que la morfología es usada en las lenguas. La distinción entre las lenguas analíticas y sintéticas es bastante difusa, lo que provoca un continuo que va desde las lenguas más analíticas también llamadas lenguas aislantes hasta las lenguas en alto grado sintéticas o mejor conocidas como polisintéticas (Haspelmath, 2002).
Además de la tipología morfológica, esta disciplina nos ofrece los elementos necesarios para estudiar la formación de palabras mediante los que se han denominado procesos morfológicos de flexión, derivación y composición. El primero consiste en una modificación sistemática de la raíz[3] por medio de morfemas flexivos [4] con el fin de agregar información gramatical (Manning & Schütze, 1999), por ejemplo, agregar "-as" como sufijo de la raíz "niñ-", agrega información de género y cantidad sin modificar la categoría gramatical ni el significado. La derivación es menos sistemática y usualmente resulta en un cambio más radical del significado, formando palabras a partir de una raíz y por lo menos un morfema derivativo[5]; y por último, la composición se refiere a la formación de palabras usando al menos dos raíces (Muñoz Basols & Gironzetti).
En resumen, y de manera muy general, se puede decir que el objetivo de esta

[3] También llamada base léxica o morfema radical, es el morfema con significado léxico de una palabra (Muñoz Basols & Gironzetti).
[4] Afijo que no cambia la categoría gramatical de una palabra. (Muñoz Basols & Gironzetti)
[5] Morfema que cambia el significado de una palabra (Muñoz Basols & Gironzetti).

disciplina es describir y explicar los patrones morfológicos de los lenguajes humanos (Haspelmath, 2002).

2.3.2. Variación lingüística

Otro aspecto importante a tomar en cuenta si se está trabajando con lenguaje, y más aún en el contexto de esta tesis, es la variación lingüística. Como señala Christian Lehmann (2018), la variación es *un hecho fundamental del lenguaje y propiedad de toda lengua*. De acuerdo con el autor, ésta consiste en la existencia de formas alternativas de hablar o escribir, a las cuales se les denomina variantes, y a su vez al conjunto de variantes que suelen concurrir en un modo de hablar o escribir se les denomina variedad.
Las variantes suelen agruparse en dos niveles: el primero obedece a factores sociolingüísticos; y el segundo a reglas específicas del sistema lingüístico (Lehmann, 2018).
En el primer nivel *la variación se despliega a lo largo de ciertas dimensiones que constituyen la arquitectura de la lengua* (Lehmann, 2018). Estas dimensiónes reciben los nombres de: diafásica, diastrática, diatópica y diacrónica (Muñoz Basols & Gironzetti). La variación diafásica se refiere a los diferentes tipos de registro[6] que usa un hablante según la situación en la que se desenvuelve en relación con el momento y el contexto de la enunciación; la variación diastrática es aquella que surge de las diferencias entre los hablantes según su estatus socioeconómico y/o cultural; la variación diatópica se centra en las diferencías geográficas; y por último la variación diacrónica hace referencia a la temporalidad, es decir, una variación histórica.
Por otro lado, en el segundo nivel se consideran todos los dominios del sistema lingüístico, es decir, en este caso podremos hablar de variación fonética, fonológica, morfológica, sintáctica y léxica (Lehmann, 2018).
Un factor importante dentro de este fenómeno es la ortografía. La escritura es *un factor unificador muy potente que ha demostrado su alta eficacia en procesos de disgregación lingüística* (Bravo García, 2015). Por poner un ejemplo, sabemos que el español se puede pronunciar de forma distinta según las coordenadas geográficas del hablante, sin embargo, siempre se escribe igual, y esto se convierte en un apoyo vital para el afianzamiento de la norma lingüística del español (Bravo García, 2015), no sucediendo así con muchas otras lenguas como el náhuatl, el tsotsil, el chol, entre otras, las cuales no cuentan con una norma de escritura generalmente aceptada.
La variación en el lenguaje representa además un desafio en muchos aspectos. Por ejemplo, resulta complejo trabajar en pro de la enseñanza de lenguas sin una norma de escritura como sucede con las lenguas indígenas en México, ya que resulta muy complicado el proceso de creación de programas de estudio o libros de texto dado el alto grado de variación en estas lenguas. El mismo problema se presenta en la creación de tecnologías de lenguaje, planificación del lenguaje

[6] Adecuación del hablante al uso de la lengua que requiere una situación o un contexto determinado (Muñoz Basols & Gironzetti).

[7], etc.

[7]En términos de Kaplan y Baldauf (1997), la planificación del lenguaje puede entenderse como sigue: *Language planning is a body of ideas, laws and regulations (language policy), change, rules, beliefes, and practices intended to achieve a planned change (or to stop change from happening) in the language use in one or more communities.*

3. Traducción automática y lenguas indígenas en México

La siguiente sección contiene un recuento general de los paradigmas que han regido el área de traducción automática durante los últimos años, así como un breviario acerca del contexto en el que las lenguas indígenas en México se desenvuelven.

3.1. Traducción automática

Una tarea sin duda imperante dentro de las tecnologías de lenguaje es la traducción automática (TA). De acuerdo con Battacharyya (2015), el campo de la TA tiene sus orígenes durante la guerra fría. En sus inicios, debido al constante desarrollo tecnológico de aquellos días, la comunidad científica tenía grandes esperanzas de poder resolver el problema en unos cuantos años (Brown et al., 1993), sin embargo, después de más de 50 años de desarrollo el problema sigue abierto.
Tomando como base el hecho de que el lenguaje humano es tan complejo que resultaría imposible hacer un análisis general e inmutable del mismo, los enfoques actuales buscan crear modelos que sean capaces de descubrir reglas de traducción a partir de un conjunto de datos (Brown et al., 1993).
Históricamente el desarrollo en el campo de la traducción automática ha sido guiado principalmente por los siguientes paradigmas: traducción automática basada en reglas (RBMT), traducción automática basada en ejemplos (BEMT), traducción automática estadística (SMT) y más recientemente traducción automática neuronal (NMT), es por esto que a continuación expongo, de manera muy general, las ideas que están detrás de cada uno de estos enfoques.

3.1.1. Traducción automática basada en reglas

Este paradigma tiene sus fundamentos en el conocimiento lingüístico que se tenga para las lenguas que se deseen traducir. En este caso, lo que se busca es generar reglas que permitan el análisis de la oración en la lengua de origen, reglas para transferir la representación de la oración que resulta del primer análisis a una representación en la lengua destino y reglas para generar la nueva oración a partir de la nueva representación (Bhattacharyya, 2015). Estos sistemas tienen la ventaja de realizar traducciones precisas y poder dar una explicación de cómo es que son generadas, sin embargo, están supeditados a la habilidad y conocimiento de los expertos que realizan las reglas, son muy costosos y resultan muy poco generalizables.

3.1.2. Traducción automática basada en ejemplos

La traducción automática basada en ejemplos puede ser vista como un punto intermedio entre la traducción basada en reglas y la traducción automática

estadística. Desde este enfoque, los patrones de traducción provienen de los datos, sin embargo, la identificación de dichos patrones recae en un conjunto de reglas (Bhattacharyya, 2015). Este paradigma suele también ser llamado traducción automática basada en analogía y se basa principalmente en la idea de que las personas al traducir no suelen hacer un análisis lingüístico profundo de la oración, sino que primero descomponen la oración en fragmentos (frases fragmentarias), y después traducen estos fragmentos, para que al final, haciendo una adecuada composición, se obtenga la oración traducida. La traducción de las frases fragmentarias se hace utilizando el principio de traducción por analogía (Nagao, 1984).
Los dos principales problemas de este enfoque son: el problema de definición de frontera (*boundary definition*) y el problema de fricción en la frontera (*boundary friction*). Donde el primero describe el problema que surge cuando las frases fragmentarias tienen errores sintácticos (por cómo fueron escogidas), y el segundo surge cuando al hacer la combinación de las frases fragmentarias traducidas no se considera el contexto y esto ocasiona que la oración final presente errores gramaticales (Way, 2003).

3.1.3. Traducción automática estadística

A diferencia de los enfoques anteriores, en este caso no se requiere conocimiento lingüístico *a priori*, ya que los patrones de traducción son aprendidos a partir de un corpus paralelo, utilizando para esto modelos estadísticos (Bhattacharyya, 2015).
La ventaja de trabajar bajo este enfoque es que los métodos resultan generalizables y se puede prescindir de un experto que conozca ambas lenguas. Por otro lado, la calidad de los resultados depende completamente del corpus (Al-Onaizan et al., 2000).

3.1.4. Traducción automática neuronal

La traducción automática neuronal es un enfoque relativamente nuevo en el campo de traducción automática que está basado completamente en redes neuronales artificiales. Los modelos clasificados dentro de este enfoque usualmente se componen de un codificador y un decodificador. El codificador se encarga de obtener una representación vectorial de tamaño fijo de una oración de longitud variable, y el decodificador genera la traducción de tal representación (Cho et al., 2014).
Estos modelos requieren sólo una fracción de la memoria que necesita un modelo de traducción estadística, sin embargo, se ha demostrado que obtienen peores resultados que estos modelos cuando se trabaja con escasos recursos (Cho et al., 2014; Koehn & Knowles, 2017).

3.1.5. Evaluación

El proceso de evaluación de un traductor automático resulta un problema bastante complejo debido a que resulta imposible establecer de manera certera y precisa cómo tendría que traducirse una oración, ya que no existe una sola respuesta correcta, es decir, una oración en una lengua L_1 puede ser traducida en una gran variedad de oraciones en una lengua L_2, donde todas ellas resultan igualmente validas.
Una forma de evaluación que resulta bastante evidente es realizar este proceso de manera manual, y que una o varias personas, capacitadas para hacerlo, emitan un juicio con respecto a las traducciones del sistema y con base en esto medir qué tan bueno es; sin embargo, el hecho de que sean humanos los encargados de estar revisando las traducciones implica un costo muy alto en tiempo y recursos, haciendo de este método una forma de evaluación bastante ineficaz.
También se ha buscado realizar la evaluación de las traducciones de manera automática a través de la creación de distintas métricas. Las ventajas de abordar la evaluación desde un enfoque automático se hacen evidentes cuando la cantidad de oraciones que se utiliza para medir el desempeño del modelo es bastante grande. Existe además evidencia de que estas métricas de evaluación suelen estar correlacionadas positivamente con la evaluación manual (Banerjee & Lavie, 2005; Papineni et al., 2002), no obstante, el desarrollo dentro de esta área sigue siendo objeto de debate, debido a las críticas en donde se menciona que al trabajar bajo este paradigma se suelen ignorar varios aspectos importantes de la lengua, lo que conlleva a evaluaciones bastante limitadas, por lo que este problema sigue estando abierto (Koehn, 2010, 2004).

BLEU (*Bilingual Evaluation Understudy*)

La métrica de evaluación más popular dentro de la literatura, hablando de traducción automática, es BLEU. Fue propuesta por Papineni y otros en 2002 y se basa en frases de referencia y *n-gramas* (Papineni et al., 2002).
En su artículo, Papineni y otros argumentan que entre más parecida sea la oración obtenida por un traductor automático a una o varias traducciones hechas o validadas por un especialista, más alto es el desempeño de dicho traductor. En concreto, ellos proponen realizar estas comparaciones a través de una medida de precisión modificada basada en *n-gramas* y un factor de penalización que toma en cuenta la longitud de las oraciones.
En cuanto a la precisión modificada basada en *n-gramas* (p_n), ésta busca resolver los problemas que surgen al utilizar la medida precisión (Número de palabras en la oración traducida que ocurren en la frase de referencia dividida por el total de palabras de la oración traducida) al comparar qué tanto se parece la oración traducida o candidata a las oraciones de referencia (Papineni et al., 2002). Para calcular p_n en todo el corpus, primero se obtiene el número de *n-gramas* que concuerdan entre las oraciones candidato y las oraciones de referencia, acotado

por el número máximo de veces que éstos aparecen en las oraciones de referencia. Una vez que se tiene la suma total de este número, se divide entre el número total de *n-gramas* presentes en las oraciones del corpus que se está evaluando. Matemáticamente los autores lo expresan de la siguiente manera:

$$p_n = \frac{\sum_{C \in candidatos} \sum_{n-grama \in C} Cuenta_{clip}(n-grama)}{\sum_{C' \in candidatos} \sum_{n-grama' \in C'} Cuenta(n-grama')} \quad (1)$$

Donde:

- $Cuenta(n-grama)$ corresponde al número de apariciones de *n-grama* en la oración candidata,
- $Cuenta_{clip}(n-grama)$ corresponde al número de apariciones de *n-grama* en la oración candidata acotado por el máximo número de veces que aparece dicho *n-grama* en la oración de referencia, es decir, $Cuenta_{clip}(n-grama) = min(Cuenta(n-grama)$, Max-ref-cuenta$(n-grama))$ y,
- Max-ref-cuenta(n-grama) es igual al total de veces que aparece *n-grama* en la oración de referencia.

Por otra parte, para tomar en cuenta la longitud de las oraciones, ellos proponen un factor de penalización BP basado en la longitud total de las oraciones candidatas en el corpus (c) y la longitud efectiva[8] del corpus de referencia (r). Dicho factor fue definido por los autores como sigue:

$$BP = \begin{cases} 1 & \text{si c} > \text{ r} \\ e^{\frac{1-r}{c}} & \text{si c}\leq r \end{cases}$$

Una vez definido lo anterior la métrica BLEU se puede expresar de acuerdo con los autores de la siguiente manera:

$$BLEU = BP * exp \sum_{n=1}^{N} w_n log p_n$$

BLEU toma valores entre 0 y 1, donde valores cercanos a 0 corresponderán a malas traducciones de acuerdo con la métrica, y valores relativamente altos corresponderán a traducciones parecidas a las oraciones de referencia.
Por último es importante destacar que en este trabajo se utilizó esta métrica con los valores de $N = 4$ y $w_n = \frac{1}{N}$, además también se están reportando los resultados utilizando la notación $\%BLEU$[9] (Koehn, 2004).

[8] Corresponde a la suma sobre las longitudes de las oraciones de referencia de modo tal que si hay más de una oración de referencia se escoge para la operación aquella cuya longitud tiene mayor similitud con la longitud de la oración candidata.

[9] Utilizando la notación % BLEU un puntaje de 0.12 se reportará como 12 % BLEU.

3.2. Lenguas indígenas en México

México es un país con una amplia diversidad lingüística, éste se encuentra entre los diez primeros con más lenguas originarias [10], y en la segunda posición si sólo se toman en cuenta países en América latina (Secretaría de cultura, México, 2018). De acuerdo con el Instituto Nacional de Lenguas Indígenas (INALI), en territorio mexicano se hablan 364 variantes de las 68 agrupaciones lingüísticas[11] reportadas por la misma entidad.

Más de 6,000,000 de personas son consideradas hablantes de alguna lengua indígena en México, sin embargo, el avance con respecto a los derechos de este sector de la población ha sido lento (INEGI, 2015). En 2001 se reformó el artículo segundo de la Constitución Política de los Estados Unidos Mexicanos con el fin de brindar reconocimiento y asumir el compromiso federal con las comunidades indígenas. Posteriormente, en 2003 se publicó la Ley General de Derechos Lingüísticos de los Pueblos Indígenas, en la cual se busca *regular el reconocimiento y protección de los derechos lingüísticos, individuales y colectivos de los pueblos y comunidades indígenas, así como la promoción del uso cotidiano y desarrollo de las lenguas indígenas, bajo un contexto de respeto a sus derechos.* A pesar de esto, dichas comunidades permanecen en desventaja y bajo un ambiente de discriminación e indiferencia gubernamental como lo señala Felipe Canuto Castillo en su artículo: *Las lenguas indígenas en el México de hoy: política y realidad lingüísticas* (Canuto Castillo, 2013). En este mismo trabajo Canuto afirma que la condición de las lenguas indígenas en México es de minorizadas y que *la discriminación es el factor que más presión ha ejercido sobre las poblaciones indígenas, no sólo en el plano de la lengua, sino de toda la cultura.* Por otro lado, Terborg y García (2011) contribuyen con evidencia puntual, a partir de distintos estudios en comunidades específicas, al análisis sobre el desplazamiento de varias lenguas indígenas por el español, mostrando una disminución en la vitalidad de éstas.

Es importante mencionar que la mayor parte de las lenguas indígenas en México no cuentan con academias lingüísticas que contribuyan a la realización de una correcta planificación del lenguaje y que sean capaces de brindar a los hablantes y a las instituciones un marco de referencia sobre el cual apoyarse.

En la actualidad, las lenguas indígenas en México se encuentran en un estado de deterioro continuo, tanto por la disminución de la cantidad de hablantes, como por la falta de esfuerzos para generar un escenario propicio para su conservación.

3.2.1. Lengua náhuatl

De acuerdo con los datos de la Encuesta Intercensal 2015 realizada por el INEGI, las 5 lenguas indígenas más habladas en el país son: náhuatl con 1,725,620 hablantes, maya con 859,607 hablantes, tseltal con 556,720 hablantes,

[10] Lengua cuyo origen histórico se remonta a épocas anteriores a la conquista de América (Ministerio de cultura, Perú).

[11] Conjunto de variantes lingüísticas comprendidas bajo el nombre dado históricamente a un pueblo indígena.

mixteco con 517,665 hablantes y tsotsil con 487,898 hablantes (INEGI, 2015). Esto quiere decir que el náhuatl (macrolengua[12] pertinente a la familia lingüística yuto-nahua) es la lengua con el mayor número de hablantes en México después del español, y resulta importante, no sólo por la cantidad de hablantes, sino también por su distribución geográfica. Estados como Puebla, Guerrero, Veracruz, San Luis Potosí, Oaxaca, Jalisco, Michoacán, Estado de México, Ciudad de México, entre otros, albergan herederos de sus más de 30 variantes (INALI). Esta lengua es en alto grado aglutinante[13], lo cual le confiere grandes posibilidades de expresión (Sullivan, 2014). En lo referente a la fonética, las vocales: *a, e, i, o, u* corresponden a los mismos sonidos en castellano, al igual que las consonantes *p, t, k, s, l, n, m*. El saltillo [14] ' o *h* es ligeramente aspirado, la *ch* corresponde a la *ch* castellana, *tl* se pronuncia como un solo sonido, *tz* corresponde a *ts*, *sh* corresponde a la *sh* inglesa , *ll* se pronuncia como la *ll* italiana y el acento por lo regular cae en la penúltima sílaba (Sullivan, 2014). Por último, la sintaxis del náhuatl prefiere un orden VSO (Verbo-Sujeto-Objeto) y las relaciones entre el verbo y sus argumentos se ajustan a un sistema nominativo-acusativo (Ramírez Celestino & Herrera Meza).

3.2.2. Desarrollo tecnológico para lenguas indígenas en México

Al igual que en el contexto político, el desarrollo científico y tecnológico en torno a las lenguas originarias en general, y en particular a aquellas habladas en territorio mexicano, ha sido lento, y para algunos nimio.

La innovación e investigación en tecnologías de lenguaje se centra de manera predominante en lenguas mayoritarias, lo que priva a un sector de la población de los beneficios directos que implican los avances en esta área, no obstante, existen algunos esfuerzos que buscan poner el dedo sobre el renglón en este tema. Mager y otros (2018) en su artículo *Challenges of language technologies for the indigenous languages of the Americas* hacen un compilado de dichos esfuerzos y presentan las principales lineas de investigación en el desarrollo de tecnologías de lenguaje para estas lenguas, además de un análisis con respecto a los principales desafíos que esto conlleva. Los autores hacen un compendio general de información referente a los recursos disponibles para las lenguas indígenas de América. En particular, para el caso de las lenguas indigenas en México se mencionan en este trabajo los corpus y recursos digitales con los que se cuenta actualmente, herramientas para realizar segmentación y análisis morfológico, sistemas de traducción automática, reconocimiento óptico de caracteres, entre otros (véase Cuadro 1).

Este trabajo también destaca el incremento en los últimos años del estudio de las lenguas originarias, tanto en el área puramente lingüística como en el desarrollo

[12]Lengua que existe en forma de diferentes variedades, no necesariamente intercomprensibles, pero que por ciertas razones se consideran formas dialectales de una misma lengua.

[13]Es una lengua en que se unen dos o más raíces con afijos, o sin ellos, para formar una nueva palabra.

[14]Saltillo desde la lingüística se define como una consonante sorda producida por una interrupción del flujo pulmonar de aire en la glotis.

de tecnologías de lenguaje, sin embargo, de igual manera señala que aunque los esfuerzos van en aumento, sigue siendo un trabajo insuficiente dada la diversidad lingüística del continente.

Recursos	Referencia
Corpus paralelo español-náhuatl Axolotl	(Gutierrez-Vasques et al., 2016)
Gran diccionario Nahuatl	(UNAM)
Corpus paralelo español-wixarika	(Mager & Meza, 2018)
Base de datos sobre clases flexivas en otomangue	(Feist & Palanca, 2015)
Método para la alienación automática de textos entre los idiomas mixteco y español	(Santiago, 2017)
The LMU system for the CoNLL-SIGMORPHON 2017 shared task on universal morphological reinflection	(Kann & Schütze, 2017)
Chachalaca	(Thouvenot, 2011)
Probabilistic finite-state morphological segmenter for the Wixarika (Huichol) language	(Mager et al., 2018)
Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages	(Kann et al., 2018)
Exploring bilingual lexicon extraction for Spanish-Nahuatl	(Gutierrez-Vasques, 2017)
Un experimento de reconocimiento automático de la derivación léxica en el ralámuli	(Medina-Urrea & Alvarado García, 2006)
Affix discovery by means of corpora: Experiments for Spanish, Czech, Ralámuli and Chuj	(Medina-Urrea, 2007)

Affix discovery based on entropy and economy measurements	(Medina-Urrea, 2008)
Traductor estadístico wixarika-español usando descomposición morfológica	(Mager et al., 2016)
Traductor híbrido wixarika-español con escasos recursos bilingües	(Mager, 2017)
Bilingual lexicon extraction for a distant language pair using a small parallel corpus	(Gutierrez-Vasques, 2015)
Microsoft Translator	(Microsoft)
Hacia La Traducción Automática De Las Lenguas Indígenas De México	(Mager & Meza, 2018)
Traductor Automático español–purépecha mediante OpenNMT	(Soriano, 2018)
Creating a massively parallel bible corpus	(Mayer & Cysouw, 2014)
An unsupervised model of orthographic variation for historical document transcription	(Garrette & Alpert-Abrams, 2016)
Labeling the languages of words in mixed-language documents using weakly supervised methods	(King & Abney, 2013)
Towards the speech synthesis of raramuri: A unit selection approach based on unsupervised extraction of suffix sequences	(Medina-Urrea et al., 2009)
Endangered data for endangered languages: Digitizing print dictionaries	(Maxwell & Bills, 2017)

Cuadro 1: Recursos y tecnologías del lenguaje para lenguas indígenas en México, realizado con información de Mager et al. (Mager et al., 2018).

4. Modelos de traducción automática basados en datos

De los paradigmas ya mencionados, dos de éstos destacan por basar su funcionamiento únicamente en los datos o corpus con los que se cuenta, esto les permite mantener cierta independencia con respecto a las lenguas involucradas, haciéndolos menos costosos y más generalizables. Estos enfoques son: traducción automática estadística y traducción automática neuronal. En esta tesis se plantea trabajar a partir de éstos por varios motivos:

1. Se busca que el trabajo aquí realizado resulte generalizable, lo cual permitiría que los resultados en cuestión de traducción y análisis puedan ser aplicados a otras variantes lingüísticas del náhuatl o incluso a otras lenguas con características similares.

2. Se planea evaluar si es posible aumentar el desempeño de estos modelos a partir de cierto tipo de preprocesamiento y analizar en qué casos vale la pena o no trabajar de esta manera.

3. Es de interés realizar un comparativo con respecto al desempeño de ambos modelos entrenados sobre un mismo corpus en un par de lenguas tipológicamente distantes, en un escenario de escasos recursos.

4.1. Modelos de traducción automática estadística

Como su nombre lo indica, en el paradigma de traducción automática estadística las traducciones se obtienen a partir de un modelo estadístico cuyos parámetros se ajustan con base en un corpus paralelo.

Planteamiento del problema

Un oración e en una lengua L_o puede ser traducida a una lengua L_d de diferentes maneras. Desde un enfoque estadístico, se plantea que cada oración f en L_d es una posible traducción de e. Entonces, es posible calcular $p(f|e)$, esto es, la probabilidad de que un traductor obtenga la oración f a partir de la oración e (Brown et al., 1993).
Dada una oración f, el objetivo del traductor es encontrar la oración e que el emisor tenía en mente cuando produjo f. Los autores afirman que es posible minimizar las posibilidades de equivocarse si se escoge la oración $\hat{e} \in L_o$, tal que $p(e|f) \leq p(\hat{e}|f),\ \forall e \in L_o$.
Usando el teorema de Bayes se puede escribir $p(e|f)$ de la siguiente manera:

$$p(e|f) = \frac{p(f|e)p(e)}{p(f)}$$

Dado que el denominador es independiente de e, encontrar $\hat{e}$ se traduce a encontrar e tal que maximice $p(f|e)p(e)$, es decir:

$$\hat{e} = argmax_e p(f|e)p(e) \tag{2}$$

La ecuación anterior plantea tres problemas computacionales: estimar las probabilidades asociadas a $p(e)$ o modelo de lenguaje, estimar la probabilidad asociada a $p(f|e)$ o modelo de traducción, y plantear un método de búsqueda eficiente que nos permita encontrar una solución sub-óptima para el problema de búsqueda.

Traducción automática estadística basada en palabras

Un enfoque bastante ingenioso dentro de este paradigma fue desarrollado en 1993 por Brown y otros (Brown et al., 1993), y consiste en trabajar las traducciones tomando como base las palabras dentro de las oraciones. Los modelos de traducción estadística basados en palabras parten del planteamiento anterior, por lo que buscan encontrar $\hat{e} = argmax_e\ p(e|f) = argmax_e p(f|e)p(e)$, es decir, se busca el elemento ê, el cual es el argumento que maximiza la probabilidad de tener una oración $e = e_1^l = e^1 e^2 ... e^l$ en el lenguaje destino L_d dada la oración $f = f_1^m = f^1 f^2 ... f^m$ en el lenguaje de origen L_o.
Además de lo anterior, al trabajar bajo estos supuestos los autores también introducen algunos otros factores (un factor de alineamiento a y un factor de longitud m) que tienen impacto dentro del modelo de traducción. Dichos factores implican suposiciones clave dentro del desarrollo de los modelos y tienen un papel preponderante al momento su entrenamiento.

Factor de alineamiento

Al introducir un factor de alineamiento a al modelo, Brown y otros lográn escribir $p(f|e)$ como:

$$p(f|e) = \sum_i p(f, a^i|e)$$

Donde a^i nos dirá a qué posición de la oración e se mapea la palabra i de la oración f. Por ejemplo: si se trabaja con el par de oraciones: The balance was the territory of the aboriginal people y *Le reste appartenait aux autochtones.* El factor de alineamiento quedaría definido de la siguiente manera: $a^1 = 1$, ya que *Les* tiene una correspondencia con *The*; $a_2 = 2$ ya que *reste* tiene una correspondencia con *balance*; $a^3 = 3$, $a^3 = 4$ y $a^3 = 5$ ya que *appartenait* tiene una correspondencia con *was the territory*; $a^4 = 6$ y $a^4 = 7$ ya que *aux* tiene una correspondencia con *of the*; y por último $a^5 = 8$ y $a^5 = 9$ ya que *autochtones* tiene una correspondencia con *aboriginal people.*

Factor de longitud

El otro factor importante a considerar está relacionado con la longitud de las oraciones. En este caso, Brown y otros definen un factor de longitud m que corresponde a una variable aleatoria que indica la longitud de la oración f y es introducido al modelo de la siguiente manera:

$$p(f, a|e) = \sum_{m} p(f, a, m|e)$$

Donde:

$$\begin{aligned} p(f, a, m|e) &= p(m|e)p(f, a|e, m) = p(m|e)p(f^1, a^1, f^2, a^2, ..., f^m, a^m|e, m) \\ &= p(m|e) \prod_{j=1}^{m} p(f^j, a^j|f_1^{j-1}, a_1^{j-1}, e, m) \\ &= p(m|e) \prod_{j=1}^{m} p(a^j|f_1^{j-1}, a_1^{j-1}, e, m) \prod_{j=1}^{m} p(f^j|f_1^{j-1}, a_1^{j-1}, e, m) \end{aligned}$$

Una vez incorporados los factores anteriores en el modelo, los autores lo reescriben como se muestra a continuación:

$$p(f|e) = \sum_{a} \sum_{m} \Big[p(m|e) \prod_{j=1}^{m} p(a^j|f_1^{j-1}, a_1^{j-1}, e, m) \prod_{j=1}^{m} p(f^j|f_1^{j-1}, a_1^{j}, e, m) \Big] \quad (3)$$

En el caso en que se tiene f fijo de longitud m, y además el factor de longitud m no depende del factor de alineamiento a, la ecuación (3) se puede escribir como:

$$p(f|e) = p(m|e) \sum_{a} \Big[\prod_{j=1}^{m} p(a^j|f_1^{j-1}, a_1^{j-1}, e, m) \prod_{j=1}^{m} p(f^j|f_1^{j-1}, a_1^{j}, e, m) \Big] \quad (4)$$

Esto quiere decir que:

$$p(f, a|e) = p(m|e) \prod_{j=1}^{m} p(a^j|f_1^{j-1}, a_1^{j-1}, e, m) p(f^j|f_1^{j-1}, a_1^{j}, e, m) \quad (5)$$

Modelos IBM

En concreto al hablar de traducción estadística basada en palabras hablamos de los modelos IBM desarrollados por Brown y otros (1993). Éstos son modelos generativos de traducción que se derivan de la ecuación (5) al hacer ciertas suposiciones (Bhattacharyya, 2015).

IBM 1

En el modelo IBM 1 se asume que $p(m|e)$ es independiente de e y de m, $p(a^j|f_1^{j-1}, a_1^{j-1}, e, m)$ sólo depende de la longitud de la oración en el lenguaje destino (l), por lo tanto es igual a $(l+1)^{-1}$ y $p(f^j|f_1^{j-1}, a_1^j, e, m)$ depende sólo de f^j y e^{a^j}.
Los parámetros del modelo dadas las suposiciones anteriores son los siguientes:

- $\epsilon \equiv p(m|e)$
- $t(f^j|e^{a^j}) \equiv p(f^j|f_1^{j-1}, a_1^j, e, m)$ (probabilidad de traducción de f^j dado e^{a^j}).

Por lo tanto:

$$p(f, a|e) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f^j|e^{a^j})$$

y dado que el alineamiento está dado por la especificación de los valores de a^j para j desde 1 a m, entonces:

$$p(f|e) = \frac{\epsilon}{(l+1)^m} \sum_{a^1=0}^{l} ... \sum_{a^m=0}^{l} \prod_{j=1}^{m} t(f^j|e^{a_j})$$

IBM 2

En este modelo los autores realizan las mismas suposiciones que en el modelo IBM 1, con la diferencia de que $p(a^j|f_1^{j-1}, a_1^j, e, m)$ en este caso depende de j, a^j, m y l. Por lo tanto se introduce nuevo conjunto de probabilidades de alineamiento dado por la siguiente ecuación:

$$a(a^j|j, m, l) = p(a^j|f_1^{j-1}, a_1^j, m, l)$$

Sujeto a la restricción:

$$\sum_{i=0}^{l} a(i|j, m, l) = 1$$

Obteniendo por último:

$$p(f|e) = \epsilon \sum_{a_1=0}^{l} ... \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e^{a^j}) a(a^j|j, m, l)$$

IBM 3

En este modelo se introduce explícitamente el concepto de fertilidad el cual es definido por los autores para alguna palabra e^i dentro de la oración e, como el número de palabras en la oración f que están relacionadas con la palabra e^i en un alineamiento elegido de manera aleatoria. A esta nueva variable se le denota Φ_{e^i}. Por ejemplo, en las oraciones: *Le programme a été mis en application* y *And the program has been implemented*, podemos decir que la fertilidad para la palabra *implemented* es 3, ya que se encuentra relacionada con las palabras *mis, en* y *application*. En los modelos 1 y 2, la distribución de esta variable era implícitamente elegida al momento de escoger los parámetros, para este y los siguientes modelos se realiza la parametrización directamente.
De acuerdo con los autores, dada una oración e, se define la fertilidad de cada una de las palabras en e y una lista (*tablet*) de las palabras en f que están relacionadas (dicha tablet puede estar vacia). La colección de *tablets* es una variable aleatoria T que se denomina *tableau* de e; la *tablet* de la i-ésima palabra en la oración e es denotada como T_i y la k-ésima palabra en la i-ésima *tablet* es otra variable aleatoria denotada como T_{ik}. Siguiendo con el ejemplo anterior, el *Tableu* de la oración *And the program has been implemented* quedaría definido de la siguiente manera: ($T_1 = ()$, $T_2 = (Le)$, $T_3 = (programme)$, $T_4 = (a)$, $T_5 = (été)$ y $T_6 = (mis,\ en,\ application)$), en donde, T_2 corresponde a la *tablet* de *the*, y *en* a la variable $T_{2,6}$.
Para obtener f se tiene que escoger el *tableau* correspondiente y permutar sus palabras, esta permutación es una variable aleatoria Π.
La posición en f de la k-ésima palabra en la i-ésima *tablet*, es la variable aleatoria Π_{ik}, y la probabilidad conjunta de un *tableau* τ y una permutación π, dada una oración e es:

$$\begin{aligned} p(\tau,\pi|e) = & \prod_{i=1}^{l} p(\phi_i|\phi_1^{i-1},e)p(\phi_0|\phi_1^l,e) \text{ x} \\ & \prod_{i=0}^{l}\prod_{k=1}^{\phi_{ik}} p(\tau_{i1}^{k-1}|\tau_0^{i-1},\phi_0^l,e) \text{ x} \\ & \prod_{i=1}^{l}\prod_{k=1}^{\pi_{ik}} p(\pi_{i1}^{k-1}|\pi 1^{i-1},\tau_0^l,\phi_0^l,e) \text{ x} \\ & \prod_{i=1}^{l} p(\pi_{0k}|\pi_{01}^{k-1},\pi_1^l\tau_0^l,\phi_0^l,e) \end{aligned} \tag{6}$$

Donde:

$\tau_{i1}^{k-1} = \tau_{i1}, \tau_{i2}, ..., \tau_{ik-1}$,
$\pi_{i1}^{k-1} = \pi_{i1}, \pi_{i2}, ..., \pi_{ik-1}$,
$\phi_i = \phi_{e_i}$.

Conociendo τ y π es posible determinar una oración f y un alineamiento a, sin embargo, varios pares (τ, π) pueden conducirnos al mismo par (f, a). Al conjunto de esos pares los autores los denotan $< f, a >$.

Dicho lo anterior $p(f, a|e)$ puede expresarse de la siguiente manera:

$$p(f, a|e) = \sum_{(\tau,\pi)\in<f,a>} p(\tau, \pi|e)$$

Las suposiciones hechas en este caso son:

- Para $1 \leq i \leq l$, $p(\phi_i|\phi_1^{i-1}, e)$ sólo depende de ϕ_i y e_i,
- $\forall i$; $p(\tau_{i1}^{k-1}|\tau_0^{i-1}, \phi_0^l, e)$ depende sólo de τ_{ik} y e_i, i, m y l.

Los parámetros del modelo son los siguientes:

- $n(\phi|e^i) \equiv p(\phi|\phi_1^{i-1}, e)$ (Probabilidades de fertilidad)
- $t(f|e^i) \equiv p(T_{ik} = \tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^l, e)$ (Probabilidades de traducción)
- $d(j|i, m, l) \equiv p(\pi_{ik} = j|\pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \phi_0^l, e)$ (Probabilidades de distorsión)

Por último se tiene que:

$$\begin{aligned} p(f|e) &= \sum_{a_0=0}^{l} \ldots \sum_{a_m=0}^{l} p(f, a|e) \\ &= \sum_{a_0=0}^{l} \ldots \sum_{a_m=0}^{l} \binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0} \prod_{i=1}^{l} \phi_i! n(\phi_i|e^i) \mathrm{x} \\ &\qquad \prod_{j=1}^{m} t(f_j|e^{a^j}) d(j|a^j, m, l) \end{aligned} \tag{7}$$

Sujetos a las restricciones:

$$\sum_f t(f|e) = 1,\ \sum_j d(j|i, m, l) = 1,\ \sum_\phi n(\phi|e) = 1 \text{ y } p_0 + p_1 = 1.$$

Con p_0 y p_1 reales positivos cuya suma es igual a 1.

IBM 4

Las probabilidades de distorsión del tercer modelo no toman en cuenta que muchas veces varias palabras de la oración de origen suelen traducirse en una sola en la oración destino, y que a su vez la posición relativa en la oración final también cambia. En el modelo 4 se propone cambiar el tratamiento de $p(\pi_{ik} = j|\pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \phi_0^l, e)$ para solventar ese problema.

Como se describe en el artículo (Brown et al., 1993), cada conjunto de palabras e^i alineado por lo menos a una palabra en f define un *cept*, si hay palabras que no se alinean con ninguna palabra se dice que el *cept* está vacio y los *cept* multipalabra son excluidos, es decir, un *cept* es un subconjunto de palabras en la oración e con las posiciones de las palabras que generan. Para las oraciones *J'applaudis à la décision* y *I applaud the decision* existen 5 *cepts*: (I), $(applaud)$, (the), $(decision)$, incluido el *cept* vacio el cual genera la palabra *à* de la oración en fránces, lo cual podría verse gráficamente de la siguiente manera: (*J'applaudis à la décision* $|e^0(3)$ I(1) applaud(2) the(4) decision(5)), en donde, por convención, $e^0(3)$ se pone al inicio indicando que es un *cept* vacio el que da lugar a la tercer palabra de la oración en fránces; I da lugar a la primera; *applaud* a la segunda; *the* a la cuarta y *decision* a la quinta.
En el modelo 3 el esquema *ceptual* es definido por la fertilidad de las palabras, una palabra es un *cept* si su fertilidad es mayor a 0. En IMB 4 los autores definen $[i]$ el cual denota la posición en la oración e del i-éstimo *cept*, también definen el centro del i-ésimo *cept* como $\odot_i$ para que sirva de cota superior para el número promedio de palabras correspondientes a la *tablet* en f, y por último definen la cabeza del *cept* como la palabra en su *tablet* para la cual la posición en la oración f es menor. Lo anterior es facilmente ejemplificado con las siguientes oraciones: *Le programme a été mis en application* y *And the program has been implemented*, en donde se tienen 5 *cepts*: (*Le programme a été mis en application| the(l) program(2) has(3) been(4) implemented(5,6,7)).* El centro del quinto *cept* en este ejemplo es 6, y su cabeza sería la quinta palabra de la oración en fránces, es decir, *miss.*
Por último, en este modelo remplazan $d(j|i,m,l)$ por dos factores: uno para posicionar la cabeza de cada *cept* y otro para las palabras restantes. Para $[i] > 0$ se necesita que la cabeza del i-ésimo *cept* sea $\tau_{[i]1}$, y se asume que $p(\Pi_{[i]1} = j|\pi_1^{[i]-1}, \tau_0^l, \phi_0^l, e) = d_1(j - \odot_{i-1}|\mathcal{A}(e^{[i-1]}), \mathcal{B}(f^j))$.

Donde:

$\mathcal{A}$ y $\mathcal{B}$ son funciones que dependen de los vocabularios.

IBM 5

IBM 3 y 4 resultan ser modelos deficientes, y ese problema se ataca con IBM 5. En el modelo 4, no sólo varias palabras pueden sobreponerse, sino que también pueden colocarse antes de la primera posición o después de la última en la oración f, por lo tanto, después de colocar las palabras τ_1^{i-1} y $\tau_{[i]1}^{k-1}$, en la oración f existirán lugares vacantes, y es precisamente en dichos lugares donde se deberá colocar $\tau_{[i]k}$. Los modelos 3 y 4 son deficientes ya que fallan en cumplir esta condición.
En este caso, los autores definen $v(j, \tau_1^{[i]-1}, \tau_{[i]1}^{k-1})$ como el número de lugares va-

cantes hasta la posición j, justo antes de colocar $\tau_{[i]k}$, y asumen dos parámetros de distorsión d_1 y $d_{>1}$. Al igual que en el modelo 4, se supone que para $[i] > 0$,

$$p(\Pi_{[i]1} = j|\pi_1^{[i]-1}, \tau_0^l, \phi_0^l, e) = d_1(v_j|\mathcal{B}(f^j), v_{\odot_{i-1}}, v_m - \phi_{[i]} + 1)(1 - \delta(v_j, v_{j-1})).$$

Donde δ es la función delta de *Kronecker*, que toma el valor de 1 si ambos argumentos son iguales y 0 en otro caso.
El número de lugares vacantes por encima de j es el mismo que el número de lugares vacantes por encima de $j-1$ sólo cuando j no es en si mismo un lugar vacante, por lo tanto, el último factor toma el valor de 1 cuando j es un lugar vacante, y 0 en otro caso.
En el último de los parámetros de d_1, v_n es el número de lugares vacantes hasta el momento en la oración f. Si $\phi_{[i]} = 1$, entonces, $\tau_{[i]1}$ puede ser colocada en cualquiera de las vacantes, si $\phi_{[i]} = 2$, entonces, $\tau_{[i]1}$ puede ser colocada en cualquiera de las vacantes excepto en la última. En general, $\tau_{[i]1}$ puede ocupar cualquier vacante que se encuentre a la izquierda de las $\phi_{[i]} - 1$ vacantes disponibles.
Como en el modelo 4, se permite que d_1 dependa del centro del *cept* previo y de f^j, pero se elimina la dependencia sobre $e^{[i-1]}$.
Entonces, para $[i] > 0$:

$$p(\Pi_{[i]1} = j|\pi_1^{[i]-1}, \tau_0^l, \phi_0^l, e) = d_{>1}(v_j - v_{\pi_{[i]k-1}}|\mathcal{B}(f^j), v_m - v_{\phi_{[i]k-1}} - \phi_{[i]} + k) * (1 - \delta(v_j, v_{j-1})) \quad (8)$$

Esto ocasiona que el último factor garantice que $\tau_{[i]k}$ sea asignado a una de las posiciones vacantes, asumiendo de nuevo que esta probabilidad sólo depende de f^j.

Traducción automática estadística basada en frases

Otro enfoque dentro del paradigma consiste en utilizar frases como base para el modelo de traducción, esto da pie a la traducción automática basada en frases. Cabe mencionar que los mejores resultados en traducción automática estadística se obtienen cuando se trabaja bajo este esquema.
Los modelos de traducción automática estadística basados en frases parten de un modelo estándar y las diferentes variantes pueden ser vistas como una extensión de éste (Koehn, 2010).

Modelo estándar

Como ya se mencionó, bajo el modelo estándar ê corresponde a la mejor traducción en la lengua L_d para una oración f en la lengua L_o, es decir:

$$\hat{e} = argmax_e\ p(e|f) = argmax_e p(f|e)P(e)$$

Ahora bien, como lo plantea Koehn (2010), si se supone que las oraciones pueden ser divididas en frases, es decir, $f = (\overline{f}_1, \overline{f}_2, ..., \overline{f}_I) = \overline{f}_1^I$ y analogamente para e, entonces $p(f|e)$ se puede ver de la siguiente manera:

$$p(\overline{f}_1^I|\overline{e}_1^I) = \prod_{i=1}^{I} \phi(\overline{f}_i|\overline{e}_i) d(inicio_i + fin_{i-1} - 1)$$

Donde:

- $f = (\overline{f}_1, \overline{f}_2, ..., \overline{f}_I) = \overline{f}_1^I$, es decir, f puede descomponerse como un conjunto de I frases $\overline{f}_i$, análogamente para e.
- $\phi(\overline{f}_i|\overline{e}_i)$ corresponde a la probabilidad de traducción de la frase $\overline{e}_i$ a la frase $\overline{f}_i$ y se calcula a partir del corpus paralelo.
- $d(.)$ es la probabilidad de distorsión.
- $inicio_i + fin_{i-1} - 1$ es una una medida del reordenamiento de las frases inducido por la traducción que sólo depende de i, donde:
  - $inicio_i$: representa la posición de la primer palabra de la frase $\overline{f}_k$ que se traduce a la i-ésima frase $\overline{e}_i$.
  - fin_i: representa la posición de la última palabra de la frase $\overline{f}_k$.

Es importante mencionar que en la práctica $d(.)$ suele aproximarse con una función de costo de la siguiente forma: $d(x) = \alpha^{|x|}$, en vez de estimarse de los datos, y parte de la idea de que movimientos de las frases a través de grandes distancias tendrían que ser más costosos que movimientos más pequeños o que incluso no realizar movimiento alguno (Koehn, 2010).

4.2. Modelos de traducción automática neuronal

Los modelos de traducción automática neuronal constituyen el estado del arte dentro de la traducción automática para las principales lenguas estudiadas dentro del campo del PLN, esto debido principalmente a la capacidad de cómputo con la que se cuenta actualmente y a la disponibilidad de recursos para estas lenguas.
Originalmente, estos modelos fueron desarrollados bajo un esquema de *secuencia a secuencia* (Sutskever et al., 2014; Cho et al., 2014), y posteriormente fueron potenciados a través de la utilización de mecanismos de atención (Luong et al., 2015; Bahdanau et al., 2014; Klein et al., 2017).

Modelo *secuencia a secuencia*

El modelo *secuencia a secuencia* busca abordar el problema en donde una secuencia tiene que ser mapeada a otra que probablemente tiene una longitud distinta, un ejemplo obvio es la traducción automática.
En este contexto, el modelo busca aproximar la probabilidad condicional $p(y|x)$ de traducir una oración en la lengua de origen $x = x_1, x_2, ..., x_n$ a una oración en la lengua destino $y = y_1, y_2, ..., y_m$, y está constituido básicamente por dos componentes: un codificador encargado de obtener una representación c para cada oración a traducir; y un decodificador, el cual genera la traducción de las oraciones palabra por palabra (veáse figura 3), lo que quiere decir que descompone la probabilidad condicional de la siguiente manera:

$$\log p(y|x) = \sum_{j=1}^{m} \log p(y_j|y_{<j}, c)$$

Una forma de modelar esta descomposición es utilizar una red neuronal recurrente, por ejemplo una **GRU** (Cho et al., 2014) o una **LSTM** (Hochreiter & Schmidhuber, 1997), y parametrizar la probabilidad de decodificar cada palabra y_j con la siguiente ecuación:

$$p(y_j|y_{<j}, c) = softmax(g(h_j))$$

Donde:

- g es una función que regresa un vector del tamaño del vocabulario de la lengua destino.
- h_j corresponde a la unidad oculta de la red neuronal recurrente.
- $softmax$ es la función softmax [15].

[15] $softmax : \mathbb{R}^k \to \mathbb{R}^k$, tal que, $softmax(\overline{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}}$ para $i = 1, ..., k$ y $\overline{x} \in \mathbb{R}^k$

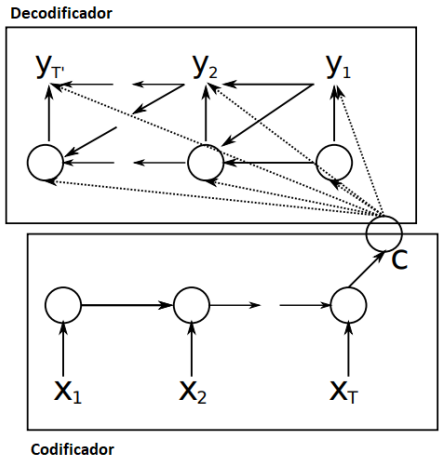

Figura 3: Ilustración del modelo secuencia a secuencia tomada de (Cho et al., 2014).

Mecanismos de atención

Dentro del área de traducción automática neuronal los mecanismos de atención han aumentado el desempeño de los sistemas mediante un enfoque parcial en las oraciones durante el proceso de traducción. Uno de los mecanismos más populares es el propuesto en (Luong et al., 2015). En su artículo, Luong y otros proponen dos tipos de mecanismos de atención, uno global en donde todas las palabras de la oración de origen son considerados, y uno local en donde sólo un subconjunto de las palabras en la oración de origen se toma en consideración. El propósito de estos mecánismos es encontrar un vector c_t que contenga información útil para poder realizar la predicción de la palabra y_t, y la diferencia entre ambos modelos radica en cómo se está obteniendo este vector, también conocido como vector de contexto (Luong et al., 2015).

Atención global

La idea de un modelo de atención global es considerar todos los estados ocultos del codificador para obtener el vector de contexto C_t, en este modelo una variable vectorial de alineamiento α_t, cuya dimensión es igual al número de pasos de tiempo que constituye la entrada, se obtiene comparando el estado oculto actual h_t en el decodificador, con cada uno de los estados ocultos del codificador $\bar{h}_s$. De acuerdo con los autores, esta variable se puede definir de la siguiente manera:

$$\alpha_t(s) = align(h_t, \bar{h_s}) = \frac{\exp(score(h_t, \bar{h_s}))}{\sum_{s'} \exp(score(h_t, \bar{h_{s'}}))} \tag{9}$$

Score se refiere a una función basada en contenido (*content-based function*). Algunas alternativas propuestas en el artículo son:

$$score(h_t, \bar{h}_s) = \begin{cases} \mathrm{h}_t^\top \bar{h}_s & \text{punto} \\ \mathrm{h}_t^\top W_\alpha \bar{h}_s & \text{general} \\ \mathrm{v}_\alpha^\top tanh(W_\alpha[h_t; \bar{h}_s)] & \text{aditiva} \end{cases}$$

Además de las propuestas anteriores también se expone en el artículo una función basada en locación (*location-based function*) en donde los puntajes o *scores* del alineamiento son calculados únicamente a partir del estado oculto h_t de la siguiente manera:

$$\alpha_t = softmax(W_\alpha h_t)$$

Por lo que dado un vector de alineamiento, éste puede utilizarse para calcular el vector de contexto c_t como un promedio ponderado de los estados ocultos en el codificador (Luong et al., 2015).

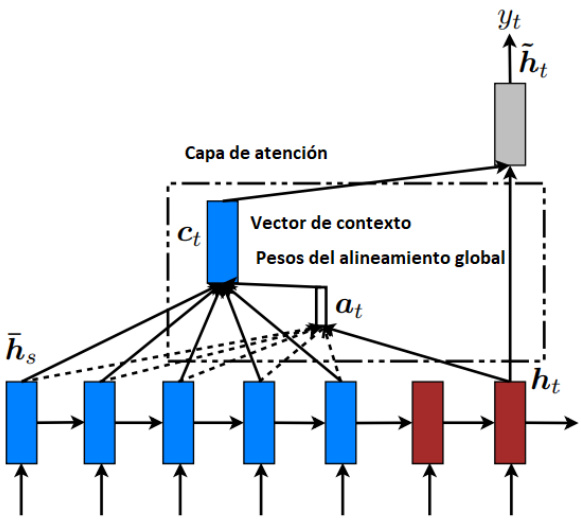

Figura 4: Ilustración del modelo de atención global adaptada de (Luong et al., 2015).

Atención local

Un problema que surge con el mecanismo de atención global es que para cada palabra predicha se tienen que tomar en cuenta todas las palabras en la oración a traducir, lo cual es costoso. Para enfrentar este problema Luong y otros proponen un mecanismo de atención local, el cual se enfoca en un subconjunto pequeño de posiciones en la oración destino para cada una de las palabras predichas (Luong et al., 2015).
En este caso el mecanismo de atención local se enfoca en una pequeña ventana o contexto. Primero genera una posición de alineamiento p_t para cada palabra en el tiempo t, y el vector de contexto c_t se calcula mediante un promedio ponderado sobre el conjunto de estados ocultos en el codificador dentro de la ventana $[p_t - D, p_t + D]$, donde D es seleccionado de manera empírica. A diferencia del modelo global, la dimensión de α_t ahora es fija (Luong et al., 2015). Los autores proponen dos variantes de este modelo. La primera consiste en un alineamiento monótono (*monotonic alignment*) en donde $p_t = t$ y el vector α_t se define conforme a la ecuación 9; y la segunda corresponde a un alineamiento predictivo (*predictive algnment*), en donde no se asumen alineamientos monótonos y se realiza una predicción de las posiciones de alineamiento de la siguiente manera:

$$p_t = S \cdot sigmoid(v^\top tanh(W_p h_t)) \tag{10}$$

Donde W_p y v_p son los parámetros del modelo encargados de aprender a predecir las posiciones y S corresponde a la longitud de la oración a traducir.
Para lograr que los puntos de alineamiento esten cerca de p_t, ellos utilizan una distribución Gausseana centrada en p_t, lo que permite escribir el alineamiento como se muestra a continuación:

$$\alpha_t = align(h_t, \bar{h}_s) exp(-\frac{(s - p_t)^2}{2\sigma^2})$$

Como se puede observar, se está utilizando la función *align* definida en 9 y se fija de manera empírica la desviación estándar como $\sigma = \frac{D}{2}$. Nótese que p_t es un número real y s es un entero dentro de una ventana centrada en p_t (Luong et al., 2015).

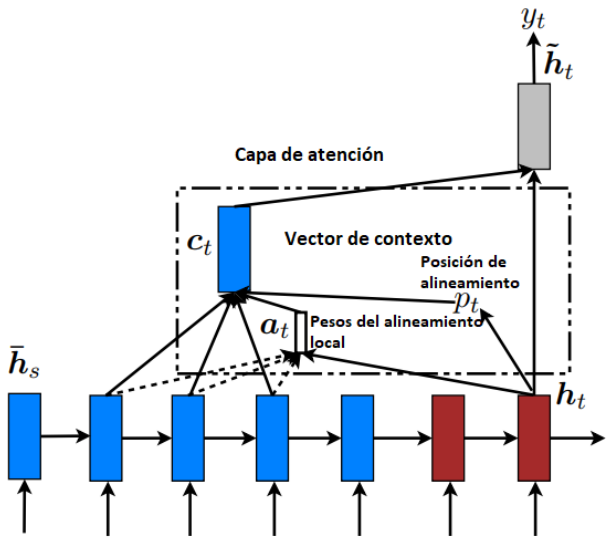

Figura 5: Ilustración del modelo de atención local adaptada de (Luong et al., 2015).

4.3. Traducción automática basada en datos y escenarios con escasos recursos digitales

Resulta obvio decir que bajo un enfoque basado en datos lo que se busca es que los patrones de traducción sean aprendidos de los mismos datos, esto quiere decir que la estimación de las probabilidades necesarias tanto para los modelos estadísticos como para los modelos neuronales recaen completamente en un corpus de entrenamiento.

Las lenguas mayoritarias como el inglés, francés, español, etcétera, cuentan con un conjunto de corpus importantes con los que se puede trabajar, por ejemplo: LCD (*Linguistic Data Consortium*), Gigaword, Europarl, OPUS, Acquis Communitaire, etcétera. No obstante, la cantidad de recursos con los que se cuenta para lenguas minoritarias es mucho menor, y aún más pequeña es la cifra si hablamos de lenguas minorizadas. Esto claramente restringe el alcance de los modelos de traducción basados en datos.

Existen trabajos que ponen en evidencia el desempeño de los modelos de traducción estadística y neuronal cuando se enfrentan a cantidades de recursos limitadas. Khoen y otros (2003) por ejemplo, muestran como el índice BLEU incrementa de manera importante conforme el corpus aumenta y cae drásticamente cuando cuando el tamaño del corpus se reduce. En el artículo *Six Challenges for Neural Machine Translation* (Koehn & Knowles, 2017) se muestra la misma tendencia tanto para los modelos de traducción estadística como para los modelos de traducción neuronal, obteniendo los últimos peores resultados cuando los recursos eran escasos.

A pesar de su complejidad este problema no se ha dejado de lado y existen algunos trabajos que buscan abordarlo. En el caso de los modelos estadísticos, se ha buscado aumentar su desempeño a través del uso de distintos recursos como por ejemplo: aumentar el tamaño del corpus utilizando paráfrasis (Marton et al., 2009; Callison-Burch et al., 2006), utilizar diccionarios e información morfo-sintáctica (Nießen & Ney, 2004; Popovic & Ney, 2005; Mager, 2017), aprovechar la existencia de corpus comparables para estimar similitudes entre palabras y frases (Irvine & Callison-Burch, 2013), explotar las similitudes entre las lenguas a traducir (Mikolov et al., 2013), entre otros. Dichos trabajos buscan aprovechar recursos ya existentes, no obstante, cuando las lenguas en cuestión son poco estudiadas, no se cuenta con suficiente información morfológica y sintáctica, y además si resultan ser tipológicamente distintas, las metodologías ahí planteadas se ven limitadas. En el caso de los modelos neuronales, este tema ha tomado bastante relevancia en los últimos años, y el enfoque paradigmático ha sido la transferencia de conocimiento[16]. Para realizar transferencia de conocimiento se suele entrenar un modelo *padre* para posteriormente inicializar los parámetros de un modelo *hijo* que es entrenado con el par de lenguas para las que no se cuenta con suficientes recursos. Cabe destacar que tanto existen trabajos que utilizan lenguas mayoritarias y bien estudiadas para entrenar el modelo padre (Zoph et al., 2016; KocmiTom & Bojar, 2018), como algunos otros que utilizan lenguas con escasos recursos pero que están relacionadas con las lenguas del modelo hijo (Nguyen & Chiang, 2017).

Además de la transferencia de conocimiento han surgido otros intentos en NMT que van desde incorporar al modelo grandes corpus monolingües (Çaglar et al., 2015) hasta prescindir por completo de los corpus paralelos y trabajar únicamente a partir de corpus monolingües (Lample et al., 2017).

Sólo para dar una comparación entre la cantidad de recursos con los que se cuenta según sea el caso, se puede decir que mientras para las lenguas mayoritarias los corpus constan de millones de oraciones alineadas, para lenguas minoritarias se llega a trabajar con unos cuantos miles.

[16]La transferencia de conocimiento consiste en utilizar conocimiento obtenido a partir de una tarea aprendida para mejorar el rendimiento de una tarea que este relacionada (Torrey & Shavlik, 2009).

5. Métodos y herramientas empleadas

Este capítulo expone tanto los métodos como las herramientas utilizadas para la construcción de un traductor automático, que van desde la obtención y el preprocesamiento de los datos, hasta el entrenamiento de los modelos de traducción automática y su evaluación.

5.1. Marco teórico

Fue necesario hacer una revisión de la literatura centrándose en aquellos trabajos cuyo objetivo es proponer alguna solución para enfrentar el bajo rendimiento que presentan los traductores automáticos cuando la cantidad de recursos con la que se cuenta es escasa. Y aunque existen algunos trabajos dentro de la literatura que abordan esta problemática, es importante tomar en cuenta que las características de las lenguas con las que se trabaja, por ejemplo, rumano (Gu et al., 2018), pueden ser muy diferentes a las de las lenguas indígenas en México; por lo tanto, resulta ineludible poner especial atención a los trabajos que abordan estas lenguas. En específico para el área de traducción automática los trabajos más relacionados con esta tesis son: *Traductor híbrido wixarika-español con escasos recursos bilingües* (Mager, 2017), *Extracción léxica bilingüe automática para lenguas de bajos recursos digitales* (Gutierrez-Vasques, 2018) y *Hacia la traducción automática de las lenguas indígenas de México* (Mager & Meza, 2018). Así que se tomaron como punto de partida estos trabajos, con el fin de analizar cómo se ha abordado el problema y comprobar si algunas de las técnicas utilizadas resultan de utilidad para esta investigación.

5.2. Obtención y preprocesamiento del corpus Español-Náhuatl

Para poder realizar este trabajo se necesitó conformar un corpus paralelo con la mayor cantidad de información posible para la variante de náhuatl a trabajar, en este caso náhuatl clásico.
Una vez obtenido un corpus con las características deseadas, realizar un correcto preprocesamiento en éste podría ayudar a mejorar el desempeño de los modelos de traducción, como se ha sugerido en otros trabajos (Mager, 2017; Gutierrez-Vasques, 2018). Por esta razón, en esta tesis se realizó un proceso de segmentación morfológica y normalización ortográfica, además del alineamiento a nivel oración de manera automática de algunos textos para aumentar el tamaño del corpus, buscando en todo momento que dichos procesos resulten generalizables.

5.3. Modelos de traducción automática

La definición de los modelos de traducción con los que se va a trabajar es clave dentro de la investigación. En este caso se propone utilizar el modelo de traducción automática estadística basado en frases (Koehn, 2010) y el modelo neuronal *secuencia a secuencia* (Sutskever et al., 2014). Dado que su popularidad y evidencia de buenos resultados, por lo menos para las lenguas mayoritarias (Koehn, 2010; Klein et al., 2017), los han hecho accesibles a todo el público. Además de esto, los experimentos realizados con estos modelos ayudarán a dar solución a la pregunta de investigación acerca de si los modelos de traducción automática estándar son capaces de brindar resultados suficientemente buenos como para ser utilizados en la práctica cuando se trabaja con lenguas tipológicamente distantes y además en un escenario de escasos recursos.

5.4. Implementación y experimentación

Dados los modelos de traducción automática y el corpus a utilizar, lo primero que se tiene que abordar es la creación de los programas necesarios para el preprocesamiento antes mencionado. En particular se necesitó programar lo referente a la normalización ortográfica, y una implementación del algoritmo de Gale y Church para el caso del alineamiento a nivel oración.
La existencia de software libre (GIZA++, MOSES, OpenNMT, Morfessor, etc.) dentro del procesamiento de lenguaje natural es aprovechada en este trabajo, en primer término como una manera de optimizar recursos; y en segundo, para abordar la pregunta acerca de si los modelos de traducción automática estándar y las implementaciones disponibles para el público en general son capaces de brindar resultados suficientemente buenos como para ser utilizados en la práctica.
Por último, la evaluación de los resultados obtenidos en todo el proceso se realizó a partir de las métricas de evaluación automática más usadas en la literatura (BLEU, Purity, Precisión, Exhaustividad, valor F.)

5.4.1. Normalización ortográfica

Como ya se mencionó, un aspecto importante a considerar al momento de conformar un corpus es la variación lingüística, y en particular la variación ortográfica que éste pueda presentar.
La literatura registra varias propuestas para lidiar con este problema, las cuales van desde definir un conjunto de reglas basadas en la lengua (UNAM, 2005), hasta procesos independientes del lenguaje que utilizan herramientas estadísticas para identificar las variantes ortográficas [17] (Dasigi & Diab, 2011).
El método de normalización ortográfica empleado en este trabajo se basa en gran

[17] Variante ortográfica se entenderá en esta tesis como una variación en la escritura de una palabra o una expresión, y que además compartan el mismo significado.

medida en el artículo *CODACT: Towards Identifying Orthographic Variants in Dialectal Arabic* (Dasigi & Diab, 2011), en el cual se realiza un análisis de conglomerados utilizando distintas métricas (*Levenshtein edit distance, biased edit distance, contextual string similarity*, etc.) para reconocer las variantes ortográficas de algunas palabras. En dicho trabajo los autores plantean el problema como un *problema de clustering*, en donde el objetivo es identificar las cadenas de caracteres (palabras) que sean similares, y con esto agrupar las distintas variantes ortográficas. Para realizar este agrupamiento, en el artículo se mapean las palabras en un espacio vectorial, y posteriormente se identifican grupos de posibles variantes ortográficas a través de la medida de similitud coseno [18]. Una vez hecho lo anterior utilizan las métricas antes mencionadas para obtener lo que de acuerdo con el algoritmo son variantes ortográficas (Dasigi & Diab, 2011).
La propuesta que aquí se presenta plantea utilizar un algoritmo *soundex* estándar (Manning et al., 2008) para realizar el primer agrupamiento y con base en éste realizar un análisis de conglomerados utilizando las métricas *sorensen* y *lenveshtein*, esto con el fin enfrentar los inconvenientes que surgen al buscar una buena representación vectorial de las palabras, lo cual es difícil de lograr cuando los recursos con los que se cuenta son escasos (Jiang et al., 2018).
A continuación se presenta el algoritmo de normalización ortográfica propuesto en este trabajo.

Algortimo de normalización ortográfica

1. Aplicar algoritmo soundex al corpus,
2. Agrupar las palabras a partir de su código soundex,
3. Para cada grupo hacer:

    - Si el grupo contiene sólo una palabra:
      pasar al siguiente grupo.
    - De lo contrario:
      realizar un análisis de conglomerados jerárquico tomando como medida de distancia alguna métrica de distancia entre palabras (Levenshtein, Sorensen, etc.).

    - Seleccionar una palabra representante[19]

[18] Medida de similitud entre dos vectores que se encuentran en un espacio dotado de un producto interno.

[19] El criterio utilizado en este trabajo para elegir a dicha representante es: la palabra de longitud mínima dentro del grupo. No obstante, este algoritmo es suficientemente generalizable, como para trabajar con lenguas para las que se han logrado consensos entre los lingüistas y las comunidades, es decir, para aquellas lenguas que cuentan con academias lingüísticas y un proceso de normalización lingüística.

- Para cada conglomerado dentro del grupo, sustituir en el texto original todas las apariciones de las palabras contenidas en éste por la palabra representante del conglomerado.

Evaluación de la normalización ortográfica

Dado que el propósito principal de este trabajo radica en la traducción automática, para evaluar la calidad del algoritmo de normalización ortográfica se plantea hacerlo a partir de dos enfoques: el primero corresponde a una evaluación indirecta a partir de los resultados en la traducción; y el segundo será una evaluación de *clustering* utilizando la métrica *purity* y un corpus anotado de variantes ortográficas creado con la ayuda de una especialista.

Purity

Purity es un criterio de evaluación de *clustering* externo, lo que quiere decir que dado un *gold standard* o un conjunto de conglomerados de referencia, esta métrica va a comparar qué tanto se parecen los conglomerados obtenidos con los conglomerados de referencia, creados idealmente por uno o varios especialistas. Para realizar el cálculo de esta métrica, cada conglomerado es asociado a la clase contenida en él con mayor frecuencia, y la precisión de dicha asociación será medida a partir del número de asociaciones correctas dividido por el total de elementos (Manning et al., 2008). Matemáticamente esto puede expresarse de la siguiente manera:

Sean $\Omega = \{w_1, w_2, ..., w_k\}, \Gamma = \{c_1, c_2, ..., c_j\}$, el conjunto de conglomerados y el conjunto de clases respectivamente; y N el número de elementos a agrupar.

$$Purity(\Omega, \Gamma) = \frac{1}{N} \sum_{k} max_j |\{w_k \cap c_j\}|$$

Esta métrica toma valores entre 0 y 1, donde valores cercanos a 0 implican conglomerados de mala calidad y valores cercanos a uno implican conglomerados muy parecidos al *gold standard*.

5.4.2. Segmentación morfológica

Además de los inconvenientes presentados en secciones anteriores, trabajar con dos lenguas tipológicamente distantes, como lo son el náhuatl y el español, sin tomar en cuenta las características particulares de cada una, podría conducir a resultados deficientes como lo muestran distintos trabajos (Gutierrez-Vasques, 2015; Mager et al., 2016).

La segmentación morfológica consiste en dividir una palabra en morfemas[20], y ha sido ampliamente estudiada desde el punto de vista del procesamiento del lenguaje natural (Smit et al., 2014; Wang et al., 2016; Demberg, 2007). Para el caso especifico de la traducción automática, algunos trabajos enfocados al estudio en lenguas aglutinantes y morfológicamente ricas han optado por realizar este tipo de segmentación en el corpus como parte de un preprocesamiento con el fin de mejorar los resultados obtenidos (Mager et al., 2016; Mager & Meza, 2018; Al-Haj & Lavie, 2012; Badr et al., 2008). Siguiendo esta linea de investigación, este trabajo plantea realizar segmentación morfológica en el corpus de entrenamiento y evaluar el impacto en la calidad de la traducción.
En el caso de la lengua náhuatl, la segmentación morfológica automática se ha abordado tanto a partir de reglas (Thouvenot, 2011), como a partir de modelos no supervisados (Gutierrez-Vasques, 2015). Uno de los objetivos de este trabajo es que los métodos utilizados resulten generalizables, y es por esto que para realizar la tarea en cuestión se plantea utilizar Morfessor, la cual es una herramienta de segmentación morfológica automática que en principio no requiere conocimiento *a priori* de la lengua a segmentar, y que ha sido ampliamente utilizada en distintas lenguas (incluido el náhuatl) obteniendo resultados relativamente buenos (Smit et al., 2014; Gutierrez-Vasques, 2017; Creutz et al., 2006).
Morfessor es una herramienta de segmentación morfológica basada en algoritmos probabilísticos de aprendizaje de máquina cuyo objetivo es encontrar la segmentación que mejor se ajuste a los datos (Smit et al., 2014). En específico, se utilizará Morfessor 2.0 el cual ofrece la ventaja de poder realizar el entrenamiento de los modelos de manera semisupervisada y no supervisada. El proceso de segmentación morfológica, haciendo uso de esta herramienta, puede ser dividido en 2 etapas. La primera corresponde a la etapa de entrenamiento, y la segunda a la etapa de inferencia. En la etapa de entrenamiento, un conjunto de componentes (palabras) D_W y (opcionalmente) un conjunto anotado de componentes $D_{W \to A}$ es utilizado para realizar el ajuste de los parámetros Θ del modelo de segmentación. La etapa de inferencia consiste en realizar la segmentación morfológica de componentes nuevas dados los parámetros ajustados del modelo (Virpioja et al., 2013).

Evaluación de la segmentación morfológica

Para la evaluación de la segmentación morfológica se utilizarán las métricas más populares (en el contexto de identificación de fronteras[21]): precisión (*precision*), exhaustividad (*recall*) y el valor F (*F-score*). Dichas métricas son definidas de la siguiente manera:

$$\text{precisión} = \frac{\text{número de fronteras encontradas de manera correcta}}{\text{total de fronteras encontradas}}$$

[20] Unidad mínima con significado.
[21] En este caso con frontera me refiero al punto en donde se realiza la separación de un morfema.

$$\text{exhaustividad} = \frac{\text{número de fronteras encontradas de manera correcta}}{\text{total de fronteras correctas}}$$

$$\text{valor F} = 2 * \frac{\text{precisión*exhaustividad}}{\text{precisión+exhaustividad}}$$

5.4.3. Traducción automática

Para realizar el entrenamiento de los modelos de traducción automática se trabajó con software de código abierto. En el caso de la traducción automática estadística se utilizó: Moses (Koehn et al., 2007), GIZA++(Och & Ney, 2003) y KenLM (Heafield, 2011); para el modelo neuronal: OpenNMT (Klein et al., 2017).

Traducción automática estadística (software)

Moses es una implementación de modelos de traducción automática estadística bastante popular, que implicó grandes beneficios para el PLN, dado que puso al alcance de todos la capacidad de entrenar estos modelos y realizar experimentación de una manera más eficiente. El proceso de entrenamiento utilizando esta herramienta consiste básicamente en tomar un corpus paralelo y usar las coocurrencias de las palabras o segmentos (frases) para inferir las correspondencias de traducción entre dos lenguas. En el caso de la traducción automática estadística basada en frases estas correspondencias se dan entre secuencias consecutivas de palabras (Koehn et al., 2007).
Los dos principales componentes dentro de Moses son un *pipeline* de entrenamiento (*training pipeline*) y un decodificador. El *pipeline* de entrenamiento consiste en una colección de herramientas que toman el corpus paralelo para construir el modelo de traducción; el decodificador se encarga de traducir una oración en la lengua de origen a una oración en la lengua destino, dado el modelo de traducción ya entrenado (Koehn et al., 2007).
En el *pipeline* de entrenamiento, el corpus tiene que ser preprocesado (*tokenizing*, remover oraciones no alineadas, etc.) para posteriormente realizar un alineamiento a nivel palabra y a partir de éste inferir la traducción de las frases, además de las estadísticas del corpus necesarias para la estimación de las probabilidades (Koehn, 2010). Dicho alineamiento puede realizarse utilizando GIZA++, la cual es una implementación de los modelos originales IBM (Och & Ney, 2003; Brown et al., 1993).
Para la construcción del modelo de lenguaje necesario en los modelos estadísticos, es posible utilizar varias herramientas (SRILM (Stolcke, 2002), IRSTLM (Federico et al., 2008), KenLM (Heafield, 2011), etc.). Moses incluye por *default* la librería KenLM para la construcción de dichos modelos (Koehn et al., 2007).

Para el proceso de decodificación lo que se busca es la oración en la lengua destino con más alta probabilidad dada la oración a traducir y el modelo entrenado. Para esto el decodificador en Moses implementa un algoritmo de búsqueda *Beam*, donde la oración final es traducida de izquierda a derecha y los costos estimados necesarios en este tipo de búsqueda están en función de las probabilidades de traducción, distorsión y del modelo de lenguaje, mencionadas en la sección 4.1 (Koehn, 2010).

Traducción automática neuronal (software)

Para el modelo de traducción neuronal se utilizó OpenNMT (Klein et al., 2017). Éste es un *framework* de código abierto para la traducción automática neuronal que permite al usuario entrenar de manera sencilla los modelos más populares dentro de este paradigma de traducción priorizando la eficiencia computacional, la modularidad del código y la extensión de los modelos (Klein et al., 2017). Al igual que en Moses, el proceso podría definirse en dos etapas: entrenamiento, donde se preprocesa el corpus, se escoge el tipo de red a utilizar, y se procede a realizar el ajuste de los parámetros; y la decodificación, en donde al tener la red ya entrenada se realiza la traducción de las oraciones, en la forma descrita en la sección 4.2, usualmente utilizando una búsqueda *Beam* para generar de izquierda a derecha y palabra por palabra la oración en la lengua destino.

5.5. Evaluación y análisis de resultados

La evaluación se realizó para cada etapa dentro del proceso utilizando las métricas correspondientes, mencionadas en las secciones 3 y 5. Además de esto se realizó un análisis con respecto a los resultados buscando entender sus causas y conjeturar acerca de sus implicaciones. En particular se realizó un análisis cuantitativo de los datos buscando una evaluación cabal y realista de los resultados, poniendo énfasis en las limitaciones y singularidades del trabajo.

6. Resultados

En esta sección se muestran los resultados obtenidos en cada una de las etapas del proceso, así como la evaluación y discusión de los experimentos realizados.

6.1. Corpus paralelo español-náhuatl

Para la construcción del corpus paralelo español-náhuatl, el presente trabajo se limitó a utilizar *Axolotl*, un corpus paralelo español-náhuatl que recopila un conjunto de textos muy diversos, entre los cuales se encuentran textos históricos, didácticos, cuentos, recetarios, musicales, etc. Este corpus cuenta hasta el momento con 38 libros y tiene la característica de presentar variación dialectal, diacrónica y ortográfica, por este motivo, sólo se utilizó un subconjunto del total, el cual se compone de los 15 libros que se listan a continuación:

Texto	Dialecto Náhuatl	Dialecto Español	Nivel de alineamiento
Anales de Tepeteopan	Clásico	Actual	Párrafo
Chimalpain Cuauhtlehuanitzi	Clásico	Actual	Oración
Documentos nahuas de la Ciudad de México del siglo XVI	Clásico	Actual	Párrafo
Vida económica de Tenochtitlan	Clásico	Actual	Oración
Historia de México narrada en náhuatl y español	Clásico	Actual	Oración
La llave del náhuatl	Clásico	Actual	Párrafo
La Tinta negra y roja	Clásico	Actual	Párrafo
Primer Amoxtli Libro	Clásico	Actual	Párrafo
Revista: La lengua y cultura Náhuatl	Clásico	Actual	Párrafo
Testimonios de la Antigua Palabra	Clásico	Actual	Párrafo
La voz profunda	Clásico	Actual	Oración
De porfirio Diaz a Zapata	Clásico	Actual	Oración
Nican mopohua	Clásico	Actual	Oración

Recetario nahua de milpa alta	Clásico	Actual	Oración
La tierra nos escucha	Clásico	Actual	Oración

Cuadro 2: Conjunto de textos utilizados para este trabajo.

Como se puede observar en el cuadro 2, los textos que se utilizarón en este trabajo cumplen dos principales requerimientos: el primero es que están escritos en las variantes: náhuatl clásico y español actual; y el segundo es que están alineados por lo menos a nivel de párrafo.

6.1.1. Segmentación del corpus

El tamaño del corpus es uno de los factores que influyen en el entrenamiento de los modelos de traducción automática. Tanto el número de oraciones como el número de palabras dentro de éste suelen jugar un papel bastante importante al momento de buscar buenos resultados.
En este trabajo se realizaron distintos experimentos utilizando dos corpus de diferente tamaño, esto con el objetivo de cuantificar la influencia de esta variable en el resultado final. El primer corpus (C1) está constituido por los textos: *Vida económica de Tenochtitlan, La voz profunda, De porfirio Diaz a Zapata, Nican mopohua, Historia de México narrada en Náhuatl y Español, Llave del náhuatl, Recetario nahua de milpa alta, Testimonios de la antigua palabra, Revista: la lengua y cultura nahuatl,* y *La tierra nos escucha.* Es de suma importancia mencionar que C1 fue un recurso que se encontró dentro de la literatura y fue previamente constituido para la tesis doctoral de Ximena Gutiérrez Vasques (Gutierrez-Vasques, 2018), esto implica que el corpus fue normalizado ortográficamente a partir de un conjunto de 270 reglas, y además la definición de oración dentro de dicha tesis toma en cuenta características semánticas, no siendo así en este trabajo, en donde la definición de las oraciones parte de un enfoque más simplista y común dentro del procesamiento del lenguaje natural, es decir, éstas quedaron determinadas a partir de la separación establecida por la puntuación, lo cual quiere decir que C1 no representa un subconjunto de C2 (en términos de oraciones) estrictamente hablando como se verá más adelante.
Por otro lado, el segundo corpus (C2) contiene textos alineados de manera automática, y los elementos que lo componen son: *Anales de Tepeteopan, Chimalpain Cuauhtlehuanitzi, Documentos nahuas de la Ciudad de México del siglo XVI, Historia de México narrada en Náhuatl y Español, Llave del náhuatl, Primer Amoxtli Libro, Revista: la lengua y cultura nahuatl Vida económica de Tenochtitlan, Testimonios de la antigua palabra, La Tinta negra y roja,* y *Vida económica de Tenochtitlan.*
Es posible observar que si bien C1 y C2 tienen una amplia intersección, no es del

todo cierto decir que uno es una extensión del otro. Esto se debe al hecho de que además del corpus, Ximena tenía, dado su conocimiento de la lengua náhuatl, un criterio más especifico de selección, no siendo así en mi caso, en dónde yo me limité a trabajar con los datos que encontré en las estadísticas de Axolotl, es decir, yo únicamente seleccioné textos que estuviesen clasificados dentro de las variantes náhuatl clásico y español actual, sin tomar en cuenta ninguna otra información del texto.
En lo referente a la segmentación de los corpus para el entrenamiento de los modelos, los corpus fueron divididos en tres subconjuntos: entrenamiento (*train*), desarrollo (*dev*) y prueba (*test*) (véase cuadro 3). Para C1 estos subconjuntos se escogieron de manera aleatoria y corresponden, al 85 %, 10 % y 5 % del corpus, respectivamente. Para el caso de C2 esta división no fue directa.
El inconveniente que se presentó al momento de escoger específicamente el subconjunto de prueba para C2 se debe al alineamiento automático involucrado. En este corpus, si se escogía de manera aleatoria el subconjunto de prueba existía la posibilidad de que contuviese errores de alineamiento, lo cual implicaría a su vez una evaluación irreal de la traducción. Fue por esto que se decidió conformar un subconjunto de 500 oraciones elegidas de manera aleatoria y someterlas a una evaluación manual. De estas 500 oraciones se conservaron las primeras 100 que cumpliesen con la condición de estar correctamente alineadas y el resto fue desechado, conformando así el subconjunto de prueba. Los otros dos subconjuntos (entrenamiento y desarrollo) se escogieron de manera aleatoria y corresponden aproximadamente al 85 % y 8 %, respectivamente, del corpus original. Se tomó la decisión de trabajar únicamente con 100 oraciones debido a lo tardado que resultaba el proceso de evaluación manual.

	Oraciones			Español		Náhuatl	
	Train	Dev	Test	Tokens	Tipos	Tokens	Tipos
C1	4,975	586	293	117,495	13,034	81,629	21,106
C2	6,560	663	100	139,568	13,774	101,178	21,822

Cuadro 3: Características generales de los corpus C1 y C2 utilizados en este trabajo.

6.2. Corpus monolingües

Para la generación de los modelos de lenguaje, necesarios en la traducción automática estadística, me limité a utilizar cinco corpus monolingües: los primeros cuatro están constituidos por la parte en español y náhuatl de los corpus C1 y C2 respectivamente; y el quinto corresponde a la parte en español del corpus multilingüe *Europarl* constituido por 1,029,155 oraciones y 30,007,569 palabras (Koehn, 2005).
Europarl es un corpus multilingüe de acceso libre para 11 lenguas, con más de

20 millones de palabras para cada una. Además de esto tiene la ventaja de haber sido preprocesado para su uso en tareas de traducción automática.
Por otro lado, para el entrenamiento de los modelos de segmentación morfológica utilicé dos corpus monolingües, CS y CS-N, en donde el primero está constituido por 135,092 tokens y 27,404 tipos, y corresponde a la parte en náhuatl de todos los textos descritos en el cuadro 2; y CS-N corresponde a la normalización ortográfica de CS utilizando el algoritmo de la sección 5.4.1 (vease cuadro 4).

Corpus	Oraciones	Tokens (español)	Tokens (náhuatl)
C1	4,975	117,495	81,629
C2	6,560	139,568	101,178
Europarl	1,029,155	30,007,569	-
CS	-	-	135,092
CS-N	-	-	135,092

Cuadro 4: Características generales de los corpus monolingües utilizados en este trabajo.

6.3. Alineamiento automático

Como ya mencioné, no todos los textos utilizados estaban alineados *a priori* a nivel de oración, en concreto sólo 8 de los 15 textos recopilados cumplen con esta restricción, por este motivo se procedió a alinear el resto de manera automática utilizando el algoritmo de Gale & Church (1993). Éste fue utilizado sin modificación alguna debido a la evidencia de buenos resultados dentro de la literatura (Singh & Husain, 2005), la facilidad de implementación, y principalmente su *independencia* del lenguaje.

6.4. Normalización ortográfica

Axolotl está conformado por una gran variedad de textos, y dada la ausencia de una norma de escritura para el náhuatl la variación ortográfica es inevitable. Para esta tesis se aplicó en los corpus un proceso de normalización ortográfica, empleando el algoritmo propuesto en la sección 5.4.1, lo cual implicó una reducción importante del número de tipos en los corpus utilizados. Como se puede observar en el cuadro 5, realizar este proceso condujo a una reducción en el número de palabras únicas en náhuatl dentro del corpus C1, es decir, los 21,106 tipos para náhuatl en este corpus se transformaron en 14,607 cuando se realizó la normalización ortográfica. Un resultado similar se obtuvo en C2, donde el número de tipos se redujo de 21,822 a 14,750.
Se esperaba que esta disminución ayudase a reducir el problema de dispersión de los datos. Es bien sabido que la traducción de palabras poco frecuentes dentro del corpus de entrenamiento conducirá a traducciones malas dada la naturaleza

de los modelos y una de las causas de esta dispersión es la variación ortográfica.

Corpus	Náhuatl	
	Tokens	Tipos
C1	81,629	21,106
C1N	81,629	14,607
C2	101,178	21,822
C2N	101,178	14,750

Cuadro 5: Características generales de los corpus C1, C1N, C2 y C2N.

6.4.1. Evaluación de la normalización ortográfica

Para evaluar la calidad de esta tarea se plantearon dos formas de hacerlo: la primera corresponde a una evaluación indirecta realizada a partir de los resultados en la traducción automática y será abordada posteriormente; y la segunda es una evaluación del *clustering* utilizando la métrica *purity* y un corpus anotado de variantes ortográficas, el cual se construyó para esta tesis con la ayuda de una especialista.
Como se muestra en el cuadro 6 al utilizar las métricas *Sorensen* y *Levenshtein* los resultados obtenidos al hacer la evaluación sobre el *gold standard* (muestra correspondiente al 2.7 %) fueron bastante buenos y con base en éstos se decidió realizar el proceso de normalización ortográfica automática utilizando la métrica *Sorensen* con un umbral de 0.25 para realizar un análisis de conglomerados, en específico: un análisis de conglomerados jerárquico aglomerativo de similitud máxima, es decir, se aplicó la métrica a todas las posibles parejas con el mismo código *soundex*, y se agruparon las palabras cuya distancia entre sí fuese menor al umbral referido, para hacer la normalización ortográfica.

Métrica	Umbral	Purity
Sorensen	0.25	0.97
Levenshtein	0.4	0.96

Cuadro 6: Evaluación del proceso de normalización ortográfica.

6.4.2. Conglomerados de referencia

Para crear el conjunto de conglomerados de referencia o *gold standard*, se escogieron de manera aleatoria 100 grupos obtenidos a partir del corpus **CS**, del cual se hablará posteriormente, y a partir de éstos una especialista fue la

encargada de realizar las anotaciones necesarias e identificar las variantes ortográficas dentro de cada grupo.
Dicho *gold standard* puede ser consultado en el Anexo A.

6.5. Segmentación morfológica

Para realizar el entrenamiento de los modelos de segmentación morfológica en Morfessor se utilizaron dos corpus monolingües en náhuatl y dos corpus anotados de segmentación morfológica. El primer corpus monolingüe en náhuatl (CS) está constituido por 135,092 tokens y 27,404 tipos, y corresponde a la parte en náhuatl de todos los textos descritos en el cuadro 2. Cabe destacar que el número de tokens en este caso resulta mayor al de la parte en náhuatl de C2, ya que en ciertos textos algunos fragmentos no fueron alineados y por lo tanto no se encuentran contenidos en C2 pero sí en CS; el segundo corpus monolingüe (CS-N) contiene 135,092 tokens y 15,617 tipos. Es importante mencionar que CS-N corresponde a la normalización ortográfica de CS utilizando el algoritmo de normalización propuesto en este trabajo con una métrica de similitud *Sorensen* y un umbral de 0.25 para un análisis de conglomerados jerárquico aglomerativo de similitud máxima.
En lo que respecta a los corpus anotados de segmentación morfológica se hizo uso de dos recursos encontrados en la literatura. El primero (CA1) contiene palabras segmentadas de la variante *náhuatl clásico* y fue recopilado para la tesis doctoral de Ximena Gutiérrez Vasques (Gutierrez-Vasques, 2018) con la ayuda del investigador Leopoldo Valiñas Coalla; y el segundo (CA2) corresponde a la variante *náhuatl del noreste central* y es parte del archivo de lenguas indígenas de México bajo el título de *nahuatl de acaxochitlán*, utilizado por Khan y otros (Kann et al., 2018) para realizar experimentos de segmentación morfológica. Las características generales de estos corpus se pueden observar en el cuadro 7.

	Número de componentes (palabras)	
Corpus anotado	Entrenamiento	Prueba
CA1	1,666	288
CA2	488	294

Cuadro 7: Características generales de los corpus anotados de segmentación morfológica.

6.5.1. Evaluación de la segmentación morfológica

La evaluación del proceso de segmentación morfológica se realizó utilizando las métricas: precisión, exhaustividad y valor F en el contexto particular de identificación de fronteras dentro de las palabras (Virpioja et al., 2013). Esto se puede apreciar en el cuadro 8.
En lo referente a los modelos de segmentación, los experimentos aquí realizados reportan 5 diferentes, todos obtenidos utilizando Morfessor 2.0 con un entrenamiento de tipo batch y un algoritmo de decodificación recursivo (Virpioja et al., 2013). El primero y *baseline* (NS), corresponde al modelo de segmentación obtenido a partir de un entrenamiento no supervisado utilizando el corpus monolingüe CS; RSSX corresponde al modelo de segmentación obtenido a partir de un entrenamiento semisupervisado utilizando el corpus monolingüe CS y el corpus anotado CA1; NSSX corresponde al modelo de segmentación obtenido a partir de un entrenamiento semisupervisado utilizando el corpus monolingüe CS-N y el corpus anotado CA1; NSSM corresponde al modelo de segmentación obtenido a partir de un entrenamiento semisupervisado utilizando el corpus monolingüe CS-N y el corpus anotado CA2; y por último RSSM corresponde al modelo de segmentación obtenido a partir de un entrenamiento semisupervisado utilizando el corpus monolingüe CS y el corpus anotado CA2.

Modelo	Precisión	Exhaustividad	Valor F	test
NS	0.679	0.612	0.644	Test CA1
RSSX	0.644	0.707	0.674	Test CA1
RSSM	0.608	0.642	0.625	Test CA1
NSSX	0.637	0.784	0.703	Test CA1
NSSM	0.505	0.736	0.599	Test CA1
NS	0.6	0.57	0.584	Test CA2
RSSM	0.725	0.867	0.79	Test CA2
RSSX	0.592	0.634	0.612	Test CA2
NSSM	0.73	0.872	0.795	Test CA2
NSSX	0.562	0.612	0.586	Test CA2

Cuadro 8: Evaluación del proceso de segmentación morfológica.

Al analizar el cuadro 8 podemos notar que la evaluación de los modelos de segmentación se hizo sobre dos subconjuntos de prueba: test CA1 y test CA2, los cuales hacen referencia al subconjunto de prueba de los corpus anotados CA1 y CA2 respectivamente; por ejemplo, la precisión del modelo NS es de 0.679 cuando se evalúa en el subconjunto de prueba de CA1, y es de 0.6 cuando lo hace en el subconjunto de prueba de CA2. Esto se realizó con el objetivo de conjeturar con respecto a la escalabilidad de los modelos dentro de las distintas variantes de la lengua náhuatl.
La calidad de los modelos de segmentación medida a través de los tres índices ya mencionados mostró algunos patrones:

1. En primer lugar podemos observar que en cuestión de exhaustividad y valor F, en casi todos los casos, entrenar los modelos de manera semisupervisada ayuda a mejorar su desempeño. Esto era de esperarse cuando la evaluación se realiza sobre un mismo corpus anotado, es decir, cuando el subconjunto de prueba y de entrenamiento provienen del mismo corpus; sin embargo, resulta digno de atención el hecho de que incluso cuando se entrena con un corpus anotado de naturaleza distinta[22], por lo menos uno de los índices aumenta.

2. Al evaluar los modelos RSSX y NSSX sobre test CA2, la exhaustividad y el valor F toman valores por encima de los que presenta el modelo no supervisado. En principio dichos modelos fueron entrenados con un corpus anotado en náhuatl clásico, no obstante, cuando se evalúa en un corpus de náhuatl del noreste central, éstos logran obtener mejores resultados por lo menos en términos de exhaustividad y valor F en comparación con el modelo entrenado de manera no supervisada (NS). Por otra parte, al evaluar los modelos RSSM y NSSM en test CA1 vemos un aumento sólo en la exhaustividad; la precisión y el valor F comparados con el modelo no supervisado son menores.

3. Otro factor que parece jugar un papel importante dentro del proceso es la normalización ortográfica. Cuando se realizó la evaluación de los modelos en el subconjunto de prueba de CA1, los mejores resultados en exhaustividad y valor F se obtuvieron cuando se realizó un entrenamiento semisupervisado con el corpus anotado CA1 y se utilizó el corpus monolingüe normalizado, el mismo patrón se observó para el subconjunto de prueba de CA2 con la peculiaridad de que en este caso fueron los tres índices los que mostraron valores por encima de los demás.

4. Un resultado bastante destacado es el obtenido por el modelo NS, ya que presenta la mayor precisión cuando se evalúa en el corpus CA1, lo cual resulta de interés ya que en ese caso sólo bastó con el corpus monolingüe para que el modelo obtuviese el mejor resultado en esos términos; sin

[22]Es importante centrarse en el hecho de que aunque la variantes de náhuatl que se están utilizando dentro de los corpus anotados son diferentes (náhuatl clásico y náhuatl del noroeste central), éstas no dejan de pertenecer a la misma agrupación lingüística.

embargo, tomando en cuenta que la precisión de los otros modelos en ese escenario en realidad no resulta inferior de manera importante, se puede decir de manera general, limitando la evaluación a la media armónica de la precisión y la exhaustividad, que para estos experimentos realizar un proceso de normalización ortográfica en los corpus monolingües y entrenar los modelos de manera semisupervisada ayudará a mejorar la segmentación morfológica.

5. Por último, en cuestión de la escalabilidad, si bien se observó un aumento en la exhaustividad al hacer una evaluación cruzada de los modelos, esto también implicó una disminución en la precisión.

6.6. Traducción automática

Para realizar los experimentos referentes a la traducción automática se realizaron distintos tipos de preprocesamiento en los corpus C1 y C2 con el fin de realizar el entrenamiento de los modelos bajo distintos escenarios. Como se puede observar en el cuadro 9, el corpus C1 fue dividido en tres subconjuntos: entrenamiento (*train*), desarrollo (*dev*) y prueba (*test*). Posteriomente se realizó sobre éste una segmentación del tipo no supervisada utilizando el modelo NS descrito anteriormente, a este escenario se le denotará C1-NS; los escenarios C1-RSSM y C1-RSSX corresponden al corpus C1 segmentado con los modelos RSSM y RSSX, respectivamente; C1N es el corpus C1 después del proceso de normalización ortográfica antes mencionado; C1NE corresponde al corpus C1N con la distinción de que para el modelo del lenguaje en este caso no se utilizará la parte en español de C1 sino la parte en español del corpus Europarl (Koehn, 2005); y por último C1N-NSSM y C1N-NSSX denotan al corpus C1N segmentado con los modelos NSSM y NSSX, respectivamente. Un proceso análogo se aplicó al corpus C2 y se puede observar en el cuadro 10.

	Oraciones			Español		Náhuatl	
	Train	Dev	Test	Tokens	Tipos	Tokens	Tipos
C1	4,975	586	293	117,495	13,034	81,629	21,106
C1-NS	4,975	586	293	117,495	13,034	160,437	3,084
C1-RSSM	4,975	586	293	117,495	13,034	151,747	6,230
C1-RSSX	4,975	586	293	117,495	13,034	158,604	6,954
C1N	4,975	586	293	117,495	13,034	81,629	14,607
C1NE	4,975	586	293	117,495	13,034	81,629	14,607
C1N-NS	4,975	586	293	117,495	13,034	157,514	2,903
C1N-NSSM	4,975	586	293	117,495	13,034	193,328	3,136
C1N-NSSX	4,975	586	293	117,495	13,034	184,444	3,045

Cuadro 9: Características generales de los corpus generados a partir del preprocesamiento de C1.

	Oraciones			Español		Náhuatl	
	Train	Dev	Test	Tokens	Tipos	Tokens	Tipos
C2	6,560	663	100	139,568	13,774	101,178	21,822
C2-NS	6,560	663	100	139,568	13,774	172,106	3,933
C2-RSSM	6,560	663	100	139,568	13,774	176,931	6,678
C2-RSSX	6,560	663	100	139,568	13,774	188,972	7,133
C2N	6,560	663	100	139,568	13,774	101,178	14,750
C2NE	6,560	663	100	139,568	13,774	101,178	14,750
C2N-NS	6,560	663	100	139,568	13,774	180,164	2,981
C2-NSSM	6,560	663	100	139,568	13,774	169,947	5,248
C2-NSSX	6,560	663	100	139,568	13,774	173,294	5,474

Cuadro 10: Características generales de los corpus generados a partir del preprocesamiento de C2.

Entre las características más destacables de los experimentos realizados utilizando el corpus C1, resalta la variación tan drástica en el número de tipos dentro de cada escenario. Los 21,106 tipos encontrados en el corpus original se reducen drásticamente a 3,084 cuando se realiza una segmentación morfológica no supervisada. Cuando esta segmentación se realiza de manera semisupervisada la reducción se mantiene aunque no de una manera tan drástica, en este caso el número de tipos se reduce a 6,230 y 6,954, dependiendo del modelo utilizado (RSSM y RSSX). Cuando se realiza el proceso de normalización ortográfica, si bien aún hay una disminución en el número de tipos, ésta es más moderada y como se abordará más adelante, podría ser una de las causas que explican la variación del desempeño de los modelos de TA. Por último, se observa además el mismo patrón cuando se realiza la segmentación morfológica del corpus normalizado, es decir, hay una drástica disminución en el número de tipos cuando se trata de una segmentación no supervisada y este número es un poco más grande y se mantiene en un rango similar cuando la segmentación es semi-supervisada. En lo que respecta a los experimentos realizados con el corpus C2, la variación en el número de tipos muestra un patrón bastante similar al descrito en el párrafo anterior.

6.6.1. Evaluación de la traducción automática

La evaluación de la TA se realizó a través de la métrica BLEU en cada uno de los escenarios descritos. En concreto se entrenaron dos modelos de traducción automática. El primero corresponde a un modelo de traducción automática estadística basado en frases utilizando MOSES (Koehn et al., 2007) como la implementación de dicho modelo; el alineador GIZA++ (Och & Ney, 2003) y KenLM (Heafield, 2011) para el modelo de lenguaje. El segundo es el modelo *secuencia a secuencia* implementado en OpenNMT (Klein et al., 2017), y se

utilizó para el codificador una red recurrente bidireccional con 4 capas y celdas del tipo LSTM con 512 unidades; y para el decodificador una red recurrente con 4 capas del tipo LSTM con 512 unidades, un mecanismo de atención global (Luong et al., 2015) y un optimizador Adam (Kingma & Ba, 2015) con una tasa de aprendizaje de 0.001. Se utilizaron estos parametros y arquitectura ya que fueron los que obtuvieron mejores resultados después de varios experimentos.
Los modelos fueron entrenados para realizar las traducciones tanto de náhuatl a español como de español a náhuatl[23] tomando como *baseline* los modelos entrenados utilizando corpus sin ningún tipo de preprocesamiento.
Como se puede observar en el cuadro 11 para el corpus C1, el modelo estadístico en el peor de los escenarios (BLEU: 10.04) supera al modelo neuronal en su mejor escenario (BLEU: 8.36), situación esperada de antemano por la cantidad de recursos que se está utilizando. Ahora bien, enfocándose únicamente en los modelos estadísticos, el *baseline* cuando se traduce de náhuatl a español es superado cuando se segmenta el corpus con el modelo RSSX y cuando se normaliza ortográficamente. Cuando se traduce de español a náhuatl la normalización ortográfica también implicó un aumento en BLEU para el modelo. Por último es relevante hacer notar que en los experimentos utilizando el modelo *secuencia a secuencia*, el preprocesamiento perjudicó el desempeño.
El cuadro 12 muestra los resultados obtenidos utilizando un corpus más grande (C2), y resulta relevante que en este caso absolutamente en todos los escenarios el modelo estadístico logra superar el *baseline* tanto para la traducción de náhuatl a español como de español a náhuatl[24], obteniendo una vez más el mejor desempeño para ambos casos cuando se aplica únicamente el proceso de normalización ortográfica. No obstante, con este corpus la brecha entre el desempeño del modelo neuronal y el estadístico es mucho mayor, aunque se muestra el mismo patrón en donde el preprocesamiento en el corpus perjudica el desempeño del modelo neuronal.
Los experimentos usando C2 muestran cierta consistencia con los experimentos realizados utilizando C1. Por ejemplo, el proceso de normalización ortográfica resulta ser el más útil al momento de aumentar la calidad de la traducción medida en términos de BLEU, mostrando un aumento importante en este índice cuando se utiliza C2. Esto podría ser debido al tamaño del corpus, pero también a que en C1 ya se había aplicando un proceso de normalización ortográfica basado en reglas (Gutierrez-Vasques, 2018), lo cual probablemente hace que este nuevo proceso de normalización tenga un menor impacto. Otro punto importante es el hecho de que al utilizar el modelo RSSX para la segmentación morfológica se logra aumentar el desempeño del modelo estadístico en ambos casos, siendo tanto para C1 como para C2 el segundo mejor escenario.

[23]La traducción de español a náhuatl sólo se realizó en escenarios en donde no hay segmentación morfólogíca para el náhuatl, ya que en esos casos las oraciones traducidas estarían conformadas por morfemas (de acuerdo con la segmentación morfológica automática), lo cual en términos prácticos resulta inútil a menos que se puedan reconstruir las palabras a partir de dichos morfemas, lo cual nos lleva a otro problema de investigación fuera del alcance de esta tesis.

[24]Los experimentos de traducción de español a náhuatl sólo se realizaron en escenarios sin segmentación.

Por otro lado, es digno de atención el hecho de que al combinar el proceso de segmentación morfológica y normalización ortográfica el modelo estadístico usando C1 empeora, contrario a lo que se podría pensar antes de realizar el experimento; en lo referente a C2, aplicar estos procesos de manera combinada si bien logran aumentar el BLEU en comparación con el *baseline*, los resultados resultan inferiores a los obtenidos cuando se realizan dichos procesos de manera aislada.
Con respecto a las divergencias, resalta que utilizar un corpus más grande como puede ser *Europarl* para el modelo del lenguaje, no siempre resulta de utilidad. Podemos observar que sólo con C2 este recurso representó un beneficio para el modelo sin llegar a destacar más que la segmentación semisupervisada. Otra divergencia importante se muestra en los resultados del modelo neuronal. Si bien realizar segmentación morfológica y normalización ortográfica empeora el desempeño en ambos casos, para C1 el primer proceso implica un mayor deterioro del resultado, no sucediendo así con C2.

	%BLEU	
Corpus C1	N-E	E-N
SMT		
Baseline	11.43	7.43
C1-NS	10.83↓	-
C1-RSSM	11.25↓	-
C1-RSSX	11.52↑	-
C1N	11.77↑	10.25↑
C1NE	11.02↓	-
C1N-NSN	10.48↓	-
C1N-NSSX	11.34↓	-
C1N-NSSM	10.04↓	-
NMT		
Baseline	8.36	-
C1N	7.18↓	-
C1-RSSX	7.28↓	-

Cuadro 11: Resultados del proceso de TA utilizando C1.

Corpus C2	%BLEU N-E	E-N
SMT		
Baseline	10.62	11.59
C2-NS	12.33↑	-
C2-RSSM	12.58↑	-
C2-RSSX	13.64↑	-
C2N	14.28↑	12.87↑
C2NE	12.53↑	-
C2N-NS	12.22↑	-
C2N-NSSM	13.32↑	-
C2N-NSSX	11.90↑	-
NMT		
Baseline	5.35	-
C2N	3.73 ↓	-
C2-RSSX	2.02 ↓	-

Cuadro 12: Resultados del proceso de TA utilizando C2.

6.7. Discusión

Como se mostró en las secciones anteriores, el proceso para llegar a tener un traductor automático a partir de este enfoque es bastante largo, y en los resultados obtenidos existen algunos indicios que muestran que no en todos los casos el preprocesamiento con el que se experimentó es de utilidad. También hay algunas limitaciones y puntos específicos que requieren ahondar más al respecto, ya que aunque su evaluación fue realizada rigurosamente con base en índices respaldados por la teoría, existen algunos sesgos que no hay que dejar de lado. Abordaré esta sección en el mismo orden en que presenté los resultados, conjeturando al respecto de sus posibles causas y poniendo énfasis en aquellas partes que desde mi análisis corresponden a las principales limitaciones de este trabajo. En lo que respecta a la normalización ortográfica, su evaluación directa mostró resultados (*purity*) bastante buenos; no obstante, esta métrica se basa en un corpus de referencia, y dado que este corpus fue obtenido a partir de un algoritmo *soundex* estándar (lo que quiere decir que se conformó a partir de un algoritmo fonológico diseñado para el inglés), estos resultados no reflejan el hecho de que los grupos dentro del corpus de referencia no contienen todas las posibles variantes ortográficas de una palabra y además contienen palabras que no comparten la misma pronunciación (en términos fonológicos) del náhuatl, por lo que dicha evaluación refleja solo parcialmente el desempeño del algoritmo de normalización ortográfica. Valdría la pena en trabajos futuros ahondar más en la construcción del corpus de referencia y quizá utilizar un algoritmo fonológico diseñado específicamente para el náhuatl. Aunque por otro lado, haciendo una evaluación indirecta en los resultados de traducción automática, este proce-

so logró aumentar el desempeño del modelo estadístico de manera consistente, mostrando que el algoritmo de normalización ortográfica aquí propuesto resulta útil cuando el objetivo primordial es la traducción.
La segmentación morfólogíca del corpus para aumentar el desempeño de los modelos de traducción fue otra hipótesis basada en la literatura que se buscó explorar. En principio para la tarea *per se*, se mostró en los experimentos una tendencia a obtener mejores resultados cuando el corpus monolingüe estaba normalizado ortográficamente y se realizó un entrenamiento semisupervisado, no obstante, al momento de entrenar los modelos de TA, no fueron los mejores modelos de segmentación los que obtuvieron el mejor desempeño. Como se puede observar en el cuadro 11, sólo un modelo de segmentación logró que el modelo estadístico superara el *baseline* y éste fue el modelo RSSX. Entre lo que se puede observar, es que en términos de precisión dicho modelo es el segundo mejor cuando se evalúa sobre test CA1[25]; en cuestión de exhaustividad dicho modelo en realidad no resulta importante, ya que éste ocupa la tercera posición; y por último el valor F lo sitúa como el segundo mejor, sólo después de NSSX. Con lo anterior, se podría suponer (y sería demasiado aventurado asegurar) que en cuestión de traducción automática estadística entre náhuatl clásico y español, al realizar segmentación morfológica en el corpus, el hecho de que el modelo de segmentación tenga una precisión alta tiene un mayor impacto y se puede dejar de lado la exhaustividad hasta cierto punto (aunque no demasiado, como pone en evidencia el modelo NS, el cual tiene mayor precisión que RSSX, pero exhaustividad menor, y no logra superar el *baseline*). Esto es consistente con los resultados del corpus C2 en donde el modelo de segmentación RSSX es el que obtiene el segundo mejor resultado. Siguiendo este análisis podemos inmediatamente conectarlo con otro resultado el cual ya se había mencionado, y era el hecho de que el uso de un corpus anotado de naturaleza distinta, en este caso náhuatl del noreste central, podía aumentar el desempeño del modelo en términos de la exhaustividad comprometiendo su precisión, Siguiendo con nuestra hipótesis, esto implicaría que si bien se podría aumentar el desempeño del modelo de TA a través de esta practica, los modelos de TA entrenados bajo estas circunstancias no tendrían que situarse por encima de los modelos en donde la segmentación morfológica presenta una mayor precisión y una exhaustividad no tan dispar. Evidencia de esto se puede observar en el cuadro 12, en donde al segmentar el corpus C2 utilizando el modelo RSSM se logra superar el *baseline*, sin embargo, el indice BLEU se encuentra por debajo del obtenido por el modelo NSSX el cual tiene una precisión más alta y exhaustividad similar. Otro ejemplo de esto se observa en el cuadro 11, en donde aunque no logran superar el *baseline*, los modelos RSSM y NSSM muestran un patrón consistente con lo anterior, o dicho de manera especifica, el modelo RSSM tiene una mayor precisión que NSSM pero menor exhaustividad y muestra mejores resultados en

[25] En esta sección me limitaré a hablar únicamente de la evaluación sobre el subconjunto de prueba de CA1, ya que este corpus fue construido con palabras extraídas de *Axolot* (Gutierrez-Vasques et al., 2016), y corresponden con la variante utilizada en este trabajo. El análisis hecho con CA2 fue hecho principalmente para medir la capacidad de generalización sobre otras variantes de la lengua.

la TA.
Después del análisis anterior salta a la vista el resultado obtenido en el escenario C2N-NSSM. Para este caso, el modelo de traducción automática estadística muestra el tercer mejor resultado, lo cual contradice la hipótesis del párrafo anterior. En efecto, en este caso dicho modelo de segmentación tiene una precisión bastante pobre y aunque en términos de exhaustividad es el segundo mejor, nuestro supuesto era que una precisión alta y una exhaustividad razonable conducirían a mejores resultados, no siendo así para este escenario. Una posible explicación podría ser el problema de dispersión de lo datos. Como se aparecia en el cuadro 13, los tipos en el corpus de prueba que se observaron en el entrenamiento representan un porcentaje menor cuando no se realiza ningún tipo de preprocesamiento. Este porcentaje está entre el 66 % y 67 % dependiendo del corpus (C1 o C2), lo que implica que los modelos tendrán problemas al traducir alrededor del 34 % de las palabras (palabras que no se observaron); sin embargo, este porcentaje va aumentando, cuando se comienza a preprocesar el corpus. La normalización ortográfica implica un aumento de alrededor del 10 % y la segmentación morfológica un aumento que va desde el 22.03 % hasta el 31.2 %. El modelo de segmentación que ocasiona una mayor intersección entre el corpus de entrenamiento y el corpus de prueba, es precisamente el modelo NSSM. Para C1N este modelo hace posible que el 98.6 % de los tipos dentro del corpus de prueba esté presente en el corpus de entrenamiento, es decir, casí todas las palabras se observaron en el entrenamiento. Con C2N se aprecia un resultado parecido ya que en este caso esta segmentación morfológica ocasiona que el 94.5 % de los tipos en el corpus de prueba se observen en el entrenamiento. Esto podría ser la causa de los buenos resultados en el escenario C2N-NSSM en cuestión de TA, pero, ¿por qué no sucedió así en corpus C1? Esta pregunta me lleva a considerar otra variable dentro de las hipótesis antes planteadas: el tamaño del corpus. Es interesante el hecho de que aunque en el escenario C1N-NSSM se observó el 98.6 % de las palabras en el entrenamiento, en realidad los resultados obtenidos son peores que el *baseline*.
Como ya se mencionó varias veces, los modelos basados en datos necesitan grandes cantidades de información para encontrar los patrones de traducción correctos, y esto se pone en evidencia en el comparativo de los resultados entre C1 y C2. En el corpus más pequeño, sólo la normalización ortográfica, y una segmentación morfológica relativamente buena fueron capaces de lograr que el modelo de TA superara el *baseline*, no sucediendo así cuando se trabajó con C2. En el corpus más grande incluso la peor de las segmentaciones logró aumentar el desempeño del modelo estadístico, por lo que se podría pensar que este tipo de preprocesamiento resulta prometedor siempre y cuando se tenga un corpus de suficiente tamaño (teniendo en mente que se está trabajando bajo la restricción de estar un escenario con escasos recursos).
La conjunción de los argumentos expuestos, así como los resultados apuntan a considerar el tamaño del corpus como uno de los factores preponderantes al momento de entrenar los modelos, ya que tiene una influencia clara en los resultados obtenidos después del preprocesamiento. Por lo que antes de aplicar alguno de estos métodos (segmentación morfológica y normalización ortográfica), valdría

la pena evaluar si existen las condiciones para que dicho preprocesamiento represente un aumento en la calidad de la traducción, es decir, se tiene que tener en cuenta si los modelos de segmentación que se pueden obtener son relativamente buenos, si existe gran variación ortográfica dentro del corpus, si el tamaño de éste es suficientemente grande, entre otros, antes de invertir tiempo y recursos.

Tipos						
	Entrenamiento		Prueba		Intersección	
Corpus	Español	Náhuatl	Español	Náhuatl	Español	Náhuatl
C1	12,188	19,168	765	740	88.4 %	67.4 %
C1SSX	12,188	3,055	765	470	88.4 %	98.1 %
C1N	12,189	13,395	765	676	88.4 %	77 %
C1NSSX	12,188	2,901	765	462	88.4 %	98.2 %
C1NSSM	12,188	3,015	765	470	88.4 %	98.6 %
C2	12,907	20,110	794	775	84.8 %	66 %
C2SSX	12,907	6,732	794	652	84.8 %	88.03 %
C2N	12,907	13,714	794	714	84.8 %	75.4 %
C2NSSX	12,907	5,217	794	612	84.8 %	91.33 %
C2NSSM	12,907	5,543	794	658	84.8 %	94.5 %

Cuadro 13: Este cuadro está conformado por el número de tipos dentro del corpus de entrenamiento y del corpus de prueba en cada uno de los escenarios de TA, así como su intersección en términos de porcentaje.

7. Conclusiones

Como se mostró en los capítulos anteriores, la construcción de un traductor automático en un escenario de escasos recursos utilizando este enfoque es un proceso que involucra distintas etapas, lo que da lugar conclusiones especificas, propias de cada una.

En lo que respecta al corpus utilizado, es importante resaltar que una de las limitantes a las que se tuvo que hacer frente, fue el hecho de que los recursos a los que se tenía acceso eran pocos y con una gran variación lingüística. Insistir en esta cuestión es importante, ya que si se desea trabajar en el tema de traducción automática de lenguas indígenas, este escenario será inevitable dentro de la investigación. La forma en que se abordó este problema durante el desarrollo de la tesis fue a través de la aplicación de un algoritmo de alineamiento automático y con la propuesta de una variante de un algoritmo de normalización ortográfica, ambos *independientes del lenguaje*. Con esto se buscó aprovechar al máximo los recursos con los que se contaba, y aportar evidencia empírica acerca de si esta forma de abordar el problema resulta útil en el contexto de traducción automática.

El alineamiento automático ayudó a aumentar el tamaño del corpus con el que se entrenaron los modelos de traducción. Como se muestra en los resultados, este proceso, en conjunto con el preprocesamiento propuesto, implicó un mejor desempeño en la traducción automática, lo cual es un aliciente, ya que si bien, existen corpus paralelos para algunas lenguas indígenas, encontrarlos alineados a un nivel de granularidad tan específico, como puede ser a nivel de oración, es difícil; y el hecho de que un algoritmo de alineamiento automático *independiente del lenguaje* sea útil, en particular para dos lenguas tipológicamente distantes como lo son el náhuatl y el español, en este contexto, traza un posible camino para optimizar el uso de los recursos con los que se cuenta.

La variación ortográfica desde mi punto vista, fue uno de los puntos más importantes abordados en esta tesis. El escenario actual en el que se encuentran las lenguas indígenas en México hace urgente trabajar en este *problema* u obstáculo práctico desde las tecnologías del lenguaje, ya que tiene grandes implicaciones no sólo en la traducción automática sino en aspectos educativos, como puede ser, el desarrollo de material didáctico para las distintas variantes lingüísticas; políticos, como pueden ser los programas de planificación del lenguaje, etc. En este trabajo se propuso una variante de un algoritmo de normalización ortográfica *independiente del lenguaje* que ayudó consistentemente a mejorar el desempeño, no sólo de los modelos de traducción automática estadística (un aumento de hasta 3.66 puntos en el indice BLEU), que son el objetivo final de la tesis, sino también de los modelos de segmentación morfológica involucrados (un aumento de la exhaustividad de hasta 0.172).

Ya que se está trabajando con una lengua polisintética como es el náhuatl, la segmentación morfológica como un medio para aumentar el desempeño de los modelos de traducción automática es otra cuestión tratada en esta tesis. Los resultados obtenidos en este tema fueron bastante más amplios, ya que los experimentos aquí realizados me permitieron hacer ciertas observaciones no sólo

en el resultado final sino también en los procesos intermedios. En lo que se refiere a la segmentación morfológica *per se*, se pudo observar que si bien, estos modelos entrenados de manera no supervisada a partir de este enfoque no obtienen buenos resultados (exhaustividad, precisión y valor F de 0.679, 0.612, 0.644 respectivamente), es posible aumentar su desempeño utilizando un corpus anotado de segmentación morfológica (aumento de 0.95 en la exhaustividad y 0.03 en el valor F), y no sólo eso, sino que también el proceso de normalización ortográfica automática resulta de utilidad (aumento de 0.172 en la exhaustividad y 0.059 en el valor F). Es importante mencionar que los resultados obtenidos indican además que es posible aumentar el desempeño de dichos modelos al utilizar conocimiento lingüístico (corpus anotado de segmentación morfológica) de una variante distinta al náhuatl que se está trabajando (aumento de 0.124 en la exhaustividad durante la evaluación cruzada). Lo cual implica un mayor alcance para este trabajo, ya que se sugiere que los corpus aquí utilizados podrían ser aprovechados para otras variantes de náhuatl. Ahora bien, en lo que se refiere al impacto de este proceso en la traducción automática, los experimentos realizados muestran que hay una relación no sólo con la calidad de los modelos de segmentación morfológica, es decir, qué tan buenos son, sino también con el tamaño del corpus de entrenamiento. Por un lado observamos que en el corpus más pequeño (C1) se necesitó que la segmentación morfológica fuese relativamente buena (valor F-0.703) para aumentar el desempeño en la traducción (aumento de 0.09 en el indice BLEU), no sucediendo así en el corpus más grande (C2), lo que lleva a suponer que al tener pocos datos, si se desea aumentar el desempeño de los modelos de traducción automática estadística utilizando segmentación morfológica, se necesita de una buena segmentación. No obstante, conforme la cantidad de datos aumenta se puede relajar este requerimiento. Esto quiere decir, que aunque se esté trabajando con una lengua polisintética, no necesariamente la segmentación morfológica será de utilidad en todos los casos. Es importante mencionar, que los resultados obtenidos por los modelos estándar de traducción automática neuronal fueron bastante pobres, lo cual se esperaba *a priori* dada la cantidad de datos con la que se está trabajando. Sin embargo, lo destacable aquí es que los experimentos realizados dan muestra de que el preprocesamiento planteado resulta ser un detrimento para estos modelos a diferencia de lo que ocurre con los modelos estadísticos, por lo que si se desea abordar estos escenarios utilizando redes de neuronas artificiales, quizá se deba voltear más hacia los trabajos de transferencia de conocimiento en lugar de a este tipo de preprocesamientos.

En resumen, lo anterior se podrían compendiar en los siguientes puntos:

1. Una manera de enfrentar la falta de recursos cuando se trabaja el tema de traducción automática, en específico: la falta de corpus paralelos (español-náhuatl) alineados a nivel oración, es el uso de algoritmos de alineamiento automático independientes del lenguaje, como por ejemplo Gale & Church (1993).

2. Al trabajar con un corpus que presenta variación lingüística, y en específico variación ortográfica, una forma de hacer frente a esta cuestión es a partir

del algoritmo de normalización ortográfica aquí propuesto.

3. Lo experimentos reflejan que es posible aumentar la calidad de la segmentación morfológica a partir de un proceso de normalización ortográfica y utilizando información lingüística, tanto de la variante de náhuatl con la que se está trabajando (náhuatl clásico), como de una variante diferente (náhuatl del noreste central).

4. La calidad de la traducción automática estadística, medida en términos del índice BLEU, puede ser aumentada consistentemente a través del proceso de normalización ortográfica propuesto en esta tesis, lo cual implica que la variación ortográfica dentro del corpus de entrenamiento tiene un impacto importante en los modelos de traducción automática estadística cuando se trabaja en un escenario de escasos recursos.

5. La segmentación morfológica es capaz de aumentar la calidad de la traducción en términos de BLEU, siempre y cuando se tenga una segmentación de calidad o una cantidad de datos suficientes, es decir, no resulta de utilidad en todos los escenarios.

6. El preprocesamiento aquí realizado no resulta de utilidad cuando se trabaja bajo el enfoque estándar de traducción automática neuronal.

Por último, me gustaría mencionar que si bien los experimentos realizados lograron un aumento en la calidad de las traducciones, y esto representa un avance dentro de la investigación, esta tesis únicamente se limita a ser una contribución en el ámbito académico, ya que en mi opinión, los resultados obtenidos por los modelos de traducción automática estándar se encuentran lejos de ser idóneos para su uso práctico.

Referencias

Al-Haj, H. & Lavie, A. (2012). The impact of arabic morphological segmentation on broad-coverage english-to-arabic statistical machine translation. *Machine Translation, Vol. 26, No. 1/2, pp. 3-24, Springer. doi: 10.1007/s10590-011-9101-1.*

Al-Onaizan, Y., Germann, U., Hermjakob, U., Knight, K., Koehn, P., Marcu, D., & Yamada, K. (2000). Translating with scarce resources. *American Association for Artificial Intelligence.*

Badr, I., Zbib, R., & Glass, J. (2008). Segmentation for english-to-arabic statistical machine translation. *Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 153–156, Association for Computational Linguistics.*

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR, abs/1409.0473.*

Banerjee, S. & Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72.*

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic model. *Journal of Machine Learning Research, pages 1137–1155.*

Bhattacharyya, P. (2015). *Machine translation.* Broken Sound Parkway, Boca Raton: CRC Press.

Bravo García, E. (2015). Variación lingüística y unidad ortográfica en la lengua española. `http://evabravogarcia.com/variacion-linguistica-y-unidad-ortografica-en-la-lengua-espanola/`. Online; visto el 22 de febrero del 2019.

Brown, P., Lai, J., & Mercer, R. (1991). Aligning sentences in parallel corpora. *ACL '91 Proceedings of the 29th annual meeting on Association for Computational Linguistics, pages 169-176.*

Brown, P. E., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2).

Çaglar, G., Firat, O., Xu, K., Cho, K., Loïc, B., Lin, H., Bougares, F., Holger, S., & Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *CoRR, abs/1503.03535.*

Callison-Burch, C., Koehn, P., & Osborne, M. (2006). Improved statistical machine translation using paraphrases. *Proceedings NAACL2006.*

Canuto Castillo, F. (2013). Las lenguas indígenas en el méxico de hoy: política y realidad lingüísticas. *Lenguas modernas, páginas 31-45*, (42).

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. *Proceedings of the 31st annual meeting on Association for Computational Linguistics, pages 9–16, Morristown, NJ, USA.*

Cho, K., Merrienboer, B. v., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *CoRR, abs/1409.1259.*

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp. 1724–1734)., Doha, Qatar. Association for Computational Linguistics.

Company, C. C. (2017). Lengua, cultura y visión del mundo. la identidad del español de méxico. `https://descargacultura.unam.mx/lengua-cultura-y-vision-del-mundo-la-identidad-del-espanol-de-mexico-6533185`. Online; visto el 09 de mayo del 2019.

Creutz, M., Lagus, K., & Glass, J. (2006). Morfessor in the morpho challenge. *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes.*

Dasigi, P. & Diab, M. (2011). Codact: Towards identifying orthographic variants in dialectal arabic. *Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 318–326.*

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2016). Language modeling with gated convolutional networks. *CoRR, abs/1612.08083.*

Demberg, V. (2007). A language-independent unsupervised model for morphological segmentation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 920–927.*

Dong, Y. (2018). A survey on neural network-based summarization methods. *CoRR, abs/1804.04589.*

Federico, M., Bertoldi, N., & Cettolo, M. (2008). IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association.*

Feist, T. & Palanca, E. L. (2015). Oto-manguean inflectional class database. Economic and Social Research Council (ESRC) and Arts and Humanities Research Council (AHRC).

Gale, W. A. & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics, 19*(1).

Garrette, D. & Alpert-Abrams, H. (2016). An unsupervised model of orthographic variation for historical document transcription. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 467–472.*

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

Goodman, J. (2001). A bit of progress in language modeling. *CoRR, cs.CL/0108005.*

Graves, A., Mohamed, A., & Hinton, G. E. (2013). Speech recognition with deep recurrent neural networks. *CoRR, abs/1303.5778.*

Gu, J., Hassan, H., Devlin, J., & Li, V. O. (2018). Universal neural machine translation for extremely low resource languages. *Proceedings of NAACL-HLT 2018, pages 344–354.*

Gutierrez-Vasques, X. (2015). Bilingual lexicon extraction for a distant language pair using a small parallel corpus. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 154–160.*

Gutierrez-Vasques, X. (2017). Exploring bilingual lexicon extraction for spanish-nahuatl. *ACL Workshop in Women and Underrepresenting Minorities in Natural Language Processing.*

Gutierrez-Vasques, X. (2018). Extracción léxica bilingüe automática para lenguas de bajos recursos. *Thesis for: PhD Computational Linguistics, UNAM.*

Gutierrez-Vasques, X., Sierra, G., & Pompa, I. H. (2016). Axolotl: a web accessible parallel corpus for spanish-nahuatl. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. European Language Resources Association (ELRA).*

Haruno, M. & Yamazaki, T. (1996). High-performance bilingual text alignment using statistical and dictionary information. *Proceedings of the 34th conference of the Association for Computational Linguistics, pp. 131 – 138, Santa Cruz, California.*

Haspelmath, M. (2002). *Understanding morphology.* New York: Oxford University Press Inc.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. *Proceedings of the Sixth Workshop on Statistical Machine Translation (pp. 187-197). Association for Compu-tational Linguistics.*

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation 9, 9.*

INALI. Catálogo de las lenguas indígenas nacionales: Variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas. `https://www.inali.gob.mx/clin-inali/html/v_nahuatl.html`. Online; visto el 19 de diciembre del 2018.

INEGI (2015). Encuesta Intercensal 2015. `https://www.inegi.org.mx/programas/intercensal/2015/`. Online; visto el 16 de diciembre del 2018.

Irvine, A. & Callison-Burch, C. (2013). Combining bilingual and comparable corpora for low resource machine translation. *Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 262–270.*

Jelinek, F., Merialdo, B., Roukos, S., & Strauss, M. (1991). A dynamic lm for speech recognition. *Proc. ARPA Workshop on Speech and Natural Language, pages 293–295.*

Jiang, C., Yu, H.-F., Hsieh, C.-J., & Chang, K.-W. (2018). Learning word embeddings for low-resource languages by pu learning. *Proceedings of NAACL-HLT 2018, pages 1024–1034.*

Józefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. *CoRR, abs/1602.02410.*

Jurafsky, D. & Martin, J. H. (2006). *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Prentice Hall.

Kann, K., Mager, M., Meza, I., & Shütze, H. (2018). Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), volume 1, pages 47–57.*

Kann, K. & Schütze, H. (2017). The lmu system for the conll-sigmorphon 2017 shared task on universal morphological reinflection. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection.*

Kaplan, R. B. & Baldauf, R. B. (1997). *Language planning: from practice to theory.* Multilingual Matters.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the langauge model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-35(3):400–401.*

Kay, M. & Röscheisen, M. (1993). Text-translation alignment. *Association for Computational Linguistic.*

Khandelwal, U., He, H., Qi, P., & Jurafsky, D. (2018). Sharp nearby, fuzzy far away: How neural language models use context. *CoRR, abs/1805.04623.*

King, B. & Abney, S. (2013). Labeling the languages of words in mixed-language documents using weakly supervised methods. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1110–1119.*

Kingma, D. P. & Ba, J. L. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations.*

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *CoRR, abs/1701.02810.*

KocmiTom & Bojar, O. (2018). Trivial transfer learning for low-resource neural machine translation. *Proceedings of the Third Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 244–252.*

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. *Computer Science and Artificial Intelligence Laboratory Massachusetts Institute of Technology.*

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, (pp. 79–86)., Phuket, Thailand. AAMT, AAMT.

Koehn, P. (2010). *Statistical Machine translation.* New York: Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, Nicola, C. B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.*

Koehn, P. & Knowles, R. (2017). Six challenges for neural machine translation. *Proceedings of the First Workshop on Neural Machine Translation.*

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. *Proceeding NAACL '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 30.*

Lample, G., Denoyer, L., & Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *CoRR, abs/1711.00043.*

Lehmann, C. (2018). Variación y normalización de la lengua maya. *Cuadernos de Lingüística de El Colegio de México, 5*, 331 – 387.

Luong, M., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *CoRR, abs/1508.04025.*

Mager, J. M. (2017). Traductor híbrido wixarika-español con escasos recursos bilingües. Master's thesis, Universidad Autónoma Metropolitana.

Mager, J. M., Barron Romero, C., & Meza Ruíz, I. V. (2016). Traductor estadístico wixarika-español usando descomposición morfológica. *COMTEL*.

Mager, M., Carrillo, D., & Meza, I. (2018). Probabilistic finite-state morphological segmenter for the wixarika (huichol) language. *Journal of Intelligent Fuzzy Systems*.

Mager, M., Gutierrez-Vasques, X., Sierra, G., & Meza, I. (2018). Challenges of language technologies for the indigenous languages of the americas. *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*.

Mager, M. & Meza, I. (2018). Hacia la traducción automática de las lenguas indígenas de méxico. Congreso DH2018, The Association of Digital Humanities Organizations (ADHO), El Colegio de México y la Universidad Nacional Autónoma de México (UNAM).

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. The MIT Press.

Marton, Y., Callison-Burch, C., & Resnik, P. (2009). Improved statistical machine translation using monolingually-derived paraphrases. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 381–390*.

Maxwell, M. & Bills, A. (2017). Endangered data for endangered languages: Digitizing print dictionaries. *ComputEL-2, page 85*.

Mayer, T. & Cysouw, M. (2014). Creating a massively parallel bible corpus. *Oceania, 135(273):40*.

Medina-Urrea, A. (2007). Affix discovery by means of corpora: Experiments for spanish, czech, ralámuli and chuj. *Aspects of Automatic Text Analysis, pages 277–299*.

Medina-Urrea, A. (2008). Affix discovery based on entropy and economy measurements. *Texas Linguistics Society, pages 99–112*.

Medina-Urrea, A. & Alvarado García, M. (2006). Un experimento de reconocimiento automático de la derivación léxica en el ralámuli. *La lengua y la antropología para un conocimiento global del hombre*.

Medina-Urrea, A., Herrera Camacho, J. A., & Alvarado Garcıa, M. (2009). Towards the speech synthesis of raramuri: A unit selection approach based on unsupervised extraction of suffix sequences. *Advances in Computational Linguistics, page 243*.

Melis, G., Dyer, C., & Blunsom, P. (2017). On the state of the art of evaluation in neural language models. *CoRR, abs/1707.05589*.

Microsoft. Microsoft translator. `https://www.microsoft.com/en-us/translator/business/languages/`. Online; visto el 18 de diciembre del 2018.

Mikolov, T., Karafiat, M., Burget, L., Cernock, J. & Khudanpur, S. (2010). Recurrent neural network based language model. *INTERSPEECH 2010*.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *CoRR, abs/1309.4168*.

Ministerio de cultura, Perú. Base de datos oficial de pueblos indígenas u originarios. `http://bdpi.cultura.gob.pe/caracteristicas`. Online; visto el 16 de marzo del 2019.

Muñoz Basols, J. & Gironzetti, E. Portal de lingüística híspanica. `http://hispaniclinguistics.com/glosario/segmentacion-morfologica/`. Online; visto el 29 de enero del 2019.

Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. *Artificial and human intelligence*.

Ney, H., Essen, U., & Kneser, R. (1994). On structuring probabilistic dependences in stochastic language modeling. *Computer, Speech, and Language, 8:1–38*.

Nguyen, T. Q. & Chiang, D. (2017). Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, (pp. 296–301). Asian Federation of Natural Language Processing.

Nießen, S. & Ney, H. (2004). Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, *30*(2).

Och, F. J. & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics, volume 29, number 1, pp. 19-51*.

Olah, C. (2015). Understanding lstm networks. `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`. Online; visto el 29 de enero del 2019.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Popovic, M. & Ney, H. (2005). Exploiting phrasal lexica and additional morpho-syntactic language resources for statistical machine translation with scarce training data. *Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT), pages 212–218.*

Ramírez Celestino, A. & Herrera Meza, M. d. C. Náhuatl. `https://linguistica.inah.gob.mx/index.php/leng/92-nahuatl`. Online; visto el 20 de diciembre del 2018.

Santiago, H. (2017). Método para la alineación automática de textos entre los idiomas mixteco y español. *Master's thesis, Centro Nacional de Investigación y Desarrollo Tecnológico.*

Secretaría de cultura, México (2018). ¿Sabías que en México hay 68 lenguas indígenas, además del español? `https://www.gob.mx/cultura/articulos/lenguas-indigenas?idiom=es`. Online; visto el 16 de diciembre del 2018.

Sierra Martínez, G. E. (2017). *Introducción a los corpus lingüísticos.* Instituto de Ingeniería, UNAM.

Singh, A. K. & Husain, S. (2005). Comparison, selection and use of sentence alignment algorithms for new language pairs. *Proceedings of the ACL Workshop on Building and Using Parallel Texts.*

Smit, P., Virpioja, S., Gronroos, S.-A., & Kurimo, M. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 21–24.*

Soriano, M. (2018). Traductor automático español–purépecha mediante opennmt. *Master's thesis, Universidad de Guadalajara, Mexico.*

Stolcke, A. (2002). SRILM - An extensible language modeling toolkit. *Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado.*

Sullivan, T. D. (2014). *Compendio de la gramática náhuatl.* Universidad Nacional Autónoma de México, Instituto de Investigaciones Históricas.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *CoRR, abs/1409.3215.*

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Cornell University library.*

Terborg, R. & García, L. (2011). *Muerte y vitalidad de las lenguas indígenas y las presiones sobre sus hablantes.* Universidad Nacional Autónoma de México.

Thouvenot, M. (2011). Chachalaca en cen, juntamente. Compendio Enciclopedico del Nahuatl, DVD.

Torrey, L. & Shavlik, J. (2009). *Handbook of Research on Machine Learning Applications*. IGI Global.

Torruella, J. & Llisterri, J. (1999). Diseño de corpus textuales y orales. *Filología e informática. Nuevas tecnologías en los estudios filológicos.*

UNAM (2005). GDN - Gran Diccionario Náhuatl. `http://www.gdn.unam.mx/termino/search`. Online; visto el 16 de diciembre del 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *CoRR, abs/1706.03762*.

Virpioja, S., Smit, P., Grönroos, S.-A., & Kurimo, M. (2013). Morfessor2.0:python implementation and extensionsformorfessor baseline. *Aalto University publication series.*

Virpioja, S., Smit, P., Grönroos, S.-A., Kurimo, M., & Glass, J. (2013). Morfessor2.0: Python implementation and extensions for morfessor baseline. *Aalto Univesity, technical report.*

Wang, L., Cao, Z., Xia, Y., & de Melo, G. (2016). Morphological segmentation with window lstm neural networks. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence.*

Way, A. (2003). *Recent Advances in Example-Based Machine Translation.* Springer, Dordrecht.

Wu, D. (1994). Aligning a parallel english-chinese corpus statistically with lexical criteria. *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics, 80-87, Las Cruces, NM.*

Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1568–1575.*

A. Anexo: Corpus anotado de variantes ortográficas

[['acuemmitl'], ['ayocmitla']]

[['amaqueme', 'amaquemeo']]

[['amatzin'], ['amihiyotzin'], ['amihuatzin']]

[['amelel'], ['amolhil']]

[['amonahuan', 'amonehuan', 'amoneuan', 'amonouan']]

[['amotehtayotzin', 'amotetayotzin']]

[['amotlahtoltzin', 'amotlahtultzin', 'amotlatoltzin']]

[['anahuacatl'], ['anauacatialli'], ['anauacatl'], ['anauacayotl'], ['anecuyotl']]

[['anmexica'], ['anmocochi'], ['anmohuicazque']]

[['b'], ['ba'], ['bei'], ['beiy']]

[['campa'], ['campo'], ['compoa'], ['compouh']]

[['cenones'], ['conanazque'], ['conanque']]

[['chilpoztectzintli', 'chilpuztectzintli']]

[['christoval', 'cristoval']]

[['conaquitiuh'], ['concauhtia'], ['concuito'], ['conyectia']]

[['icelteotl', 'icelteutl']]

[['iehuantzitzin'], ['intocatzin']]

[['imalacayo'], ['imelahuaca', 'imellahuaca', 'immelahuac'], ['immohuayolque']]

[['imaxtlayacayo'], ['imoztlayoc', 'imuztlayoc']]

[['inintin'], ['inontin']]

[['intlamanitiliz', 'intlamanitilliz']]

[['ipan', 'ipayn'], ['ipeuhyan']]

[['iquizayampa', 'izquicampa']]

[['itlachaliliz'], ['itlaocoliloca'], ['itlaxeloliz']]

[['itlanahuatil', 'itlanaoatil']]

[['karl'], ['karol']]

[['mitzpinahuizque'], ['motecpanque']]

[['mocnelilique', 'mocneliloca']]

[['momauizotitoque', 'momauizzotitoque']]

[['mopehualtiz'], ['mopilhuatique']]

[['namonan', 'nimonan']]

[['nicnequizquiani'], ['noconcacon']]

[['nictlapiez'], ['nictlapihuiz']]

[['nimitznotlapololtiliz', 'nimitzonnotlapololtiliz']]

[['ninopechteca', 'nonopechteca'], ['nonpactica'], ['nonpictihuiz']]

[['noconelnamiqui', 'noconilnamiqui']]

[['nontlatlachia'], ['nontlatlayocoya']]

[['nopaltica'], ['nopiltze']]

[['ocamazon'], ['oximiquini']]

[['oconaxitito', 'oquinextito']]

[['ocontlalli', 'oquintlali']]

[['omitzapan'], ['omoteixpanhui']]

[['omitzitlecahui'], ['omotequitilihqueh']]

[['omonamicti'], ['omonamiquito']]

[['omotlanahoatili', 'omotlanahuatili', 'omotlanahuatilli', 'omotlanaoatili']]

[['onechtocayoti'], ['onictequito']]

[['onmotlahtocatlalli', 'onmotlatocatlali', 'onmotlatocatlalli']]

[['ontlaxtlaoaz', 'ontlaxtlauaz', 'ontlaxtlauazque']]

[['oquimmaquixti', 'oquimoquixti'], 'oquimocaquiti']

[['oquipehualtihqueh', 'oquipehualtique']]

[['oquiteteque'], ['oquitoeayotihqueh']]

[['otechmottili'], ['oticmottili']]

[['oteyacantiaque'], ['oticniuhtihqueh'], ['otzintic']]

[['oticmihiohuilti', 'oticmihiyohuilti', 'oticmiyiohuilti']]

[['otinechhuallapoihui'], ['otinechilpi']]

[['otitlacat'], ['otitlachiato']]

[['otitochiuhque'], ['otitocoque']]

[['palehuilia', 'paleuilo'], ['pololli', 'pololo']]

[['pedronilla', 'petronila']]

[['quatzontli', 'quauhtzintli']]

[['quetzalteuh'], ['quitquilitiui']]

[['quicenteca', 'quicentoca']]

[['quimatin'], ['quimotenehuia', 'quimotenehuiaya']]

[['quinchihuiliah', 'quincuilia'], ['quinyaochihuili']]

[['quineltilique', 'quineltillique'], 'quineltilizque']

[['quitlamique'], ['quitlamiz']]

[['tamoanchan', 'tamooanchan']]

[['techoncozcatia', 'teconquixtia', 'ticcenquixtia', 'tzonquixti']]

[['techteopoah', 'ticteopuhua', 'tictopehua']]

[['teicnoyttalliztli'], ['ticnotlacatl']]

[['teixamicauan'], ['teucsomazin']]

[['teopixcatzintli'], ['tepuztzintli']]

[['tequixquipantlaca'], ['tiquixpantiliz']]

[['tetetica'], ['teuhyotiuitz'], ['titeutiz'], ['titotizque'], ['ttotecuiyo']]

[['teteuhtepec', 'tototepec']]

[['texedor'], ['tocdor']]

[['ticmopalehuilia', 'tiquimmopalehuilia']]

[['ticneltilia', 'ticneltillia']]

[['tictzatzaqua'], ['tzatzitiquiza']]

[['timecapalli'], ['tomecapal']]

[['timoquetztehuaz'], ['timoquetztiaz']]

[['timotepexihui'], ['timotopehuaz']]

[['titopalhuizquiani'], ['totepalehuicahuan']]

[['tlacateccati'], ['tlecoticate']]

[['tlacuicuilitzintli'], ['tlahcohcoualtzintli'], ['tolxacaltzintli']]

[['tlahmachuipilli', 'tlamachuipilli']]

[['tlaltecatzin', 'tliltecatzin']]

[['tlapallan'], ['tlapiloni']]

[['tlateneuhtli', 'tlatentli']]

[['tlatetotoc'], ['tlatuitiuitz']]

[['xacalcatl'], ['xochihualcuahitl']]

[['xichualcui', 'xiquallihuaca', 'xiquiualuica'], ['xoxollochauh']]

[['xicmoeuitlahui'], ['xicmotla'], ['xicmottili']]

[['ximenes', 'ximenez']]

[['ximoquetza', 'xommoquetza']]

[['xivia'], ['xv'], ['xvi'], ['xvii'], ['xviii']]

[['yequexquichcauh'], ['yzquixochiqu']]

[['yoleatl', 'yollotli', 'yolotli', 'yoyolitl']]

[['ytlacahuiz', 'ytlacauiz']]

[['zihuatzintli', 'zouatzintli']]