Procesamiento del lenguaje natural y las lenguas mexicanas

Ximena Gutierrez-Vasques

Abril, 2019

$WhoAmI

- Doctora en Ciencias de la Computación (lingüística computacional/NLP)

- Intereses de investigación:

NLP/Machine learning, **Quantitative linguistics**, **Low-resource languages**, **Machine translation…**

- Actualmente soy parte de **Elotl.mx**

Índice de la charla

- Procesamiento del Lenguaje Natural (NLP)
- Traducción automática y tecnologías multilingües
- El caso de México, bajos recursos digitales

Natural **Language** Processing (NLP)

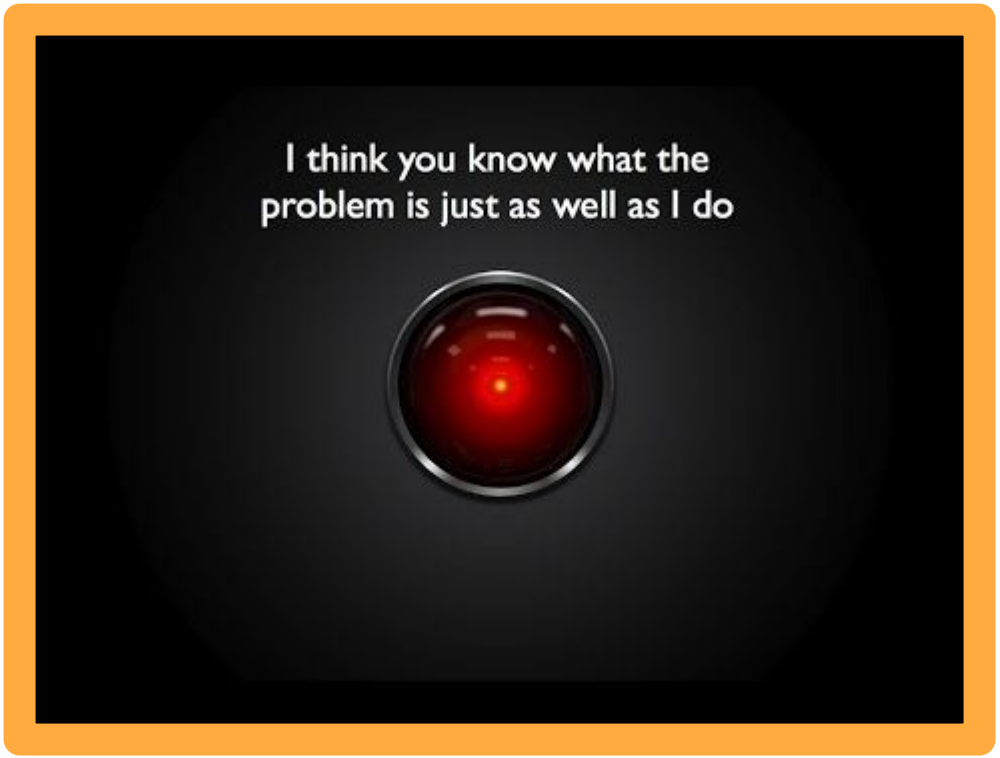
Ximenta Gutiérrez-Vasques 2017

Procesamiento del lenguaje natural (NLP)

Gran reto:

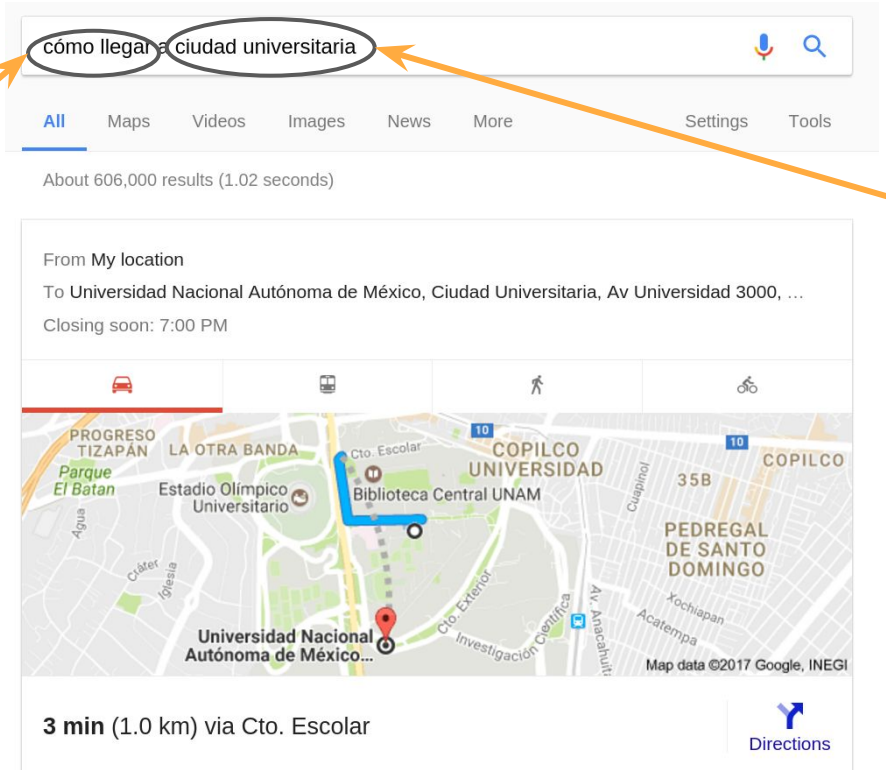
Modelar el lenguaje humano desde una perspectiva computacional

- Modelos capaces de “entender”, procesar/manipular, generar lenguaje humano
- Tarea ambiciosa, se necesita una perspectiva interdisciplinaria



Tecnologías del lenguaje

Resolución a una acción

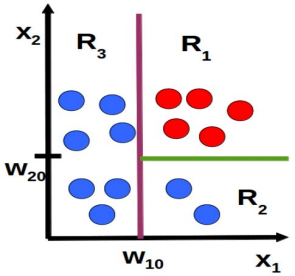


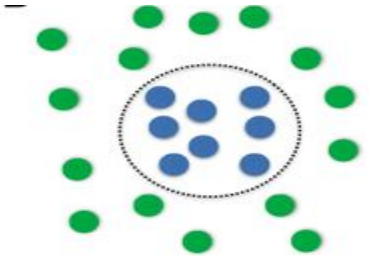
Detección de entidades nombradas

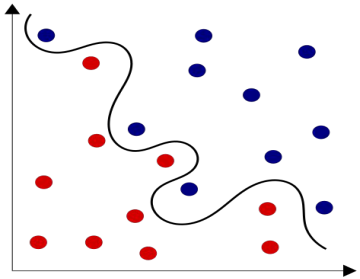
Tecnologías del lenguaje

NLP+ML

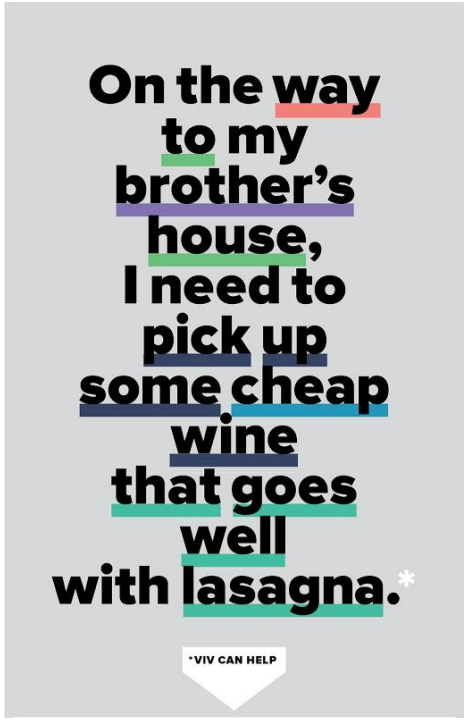
- Hoy en día, muchas de las tecnologías del lenguaje se basan en **aprendizaje de máquina**(generalmente supervisado),
- Muchas tareas esenciales de NLP pueden verse como un problema de **clasificación**...





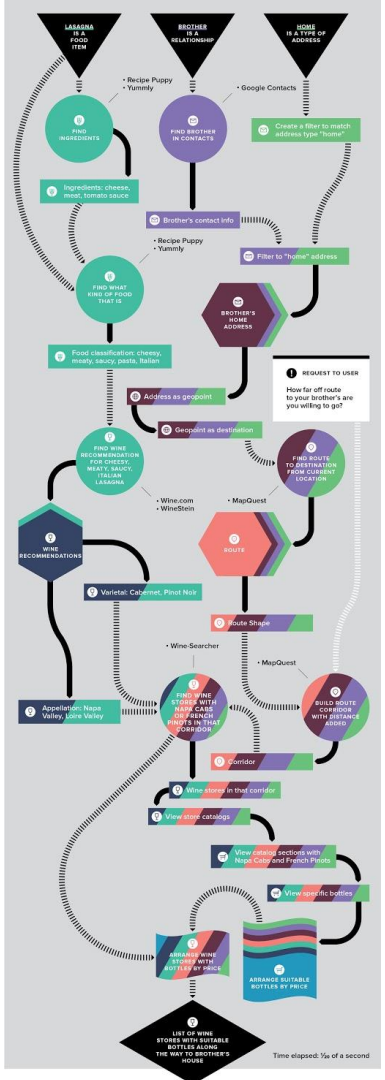


NLP. Limitaciones



Ejemplo sistema pregunta-respuesta / asistente de voz

Limitaciones



- Reconocimiento de entidades nombradas
- Análisis de correferencia
- Análisis sintáctico
- Ontologías
- Sistema de recuperación de la información
- ...

(Responder en menos de 1 segundo)

NLP. Limitaciones

- Ahora imaginen que el humano pide...

Quiero pasar por un pomo chido antes de jalar con mi carnal

NLP. Limitaciones

- Chatbot

¿No me enseñas tu encuesta?

Aún no tengo una respuesta para eso.

Sobre las elecciones me gustaría un panorama general

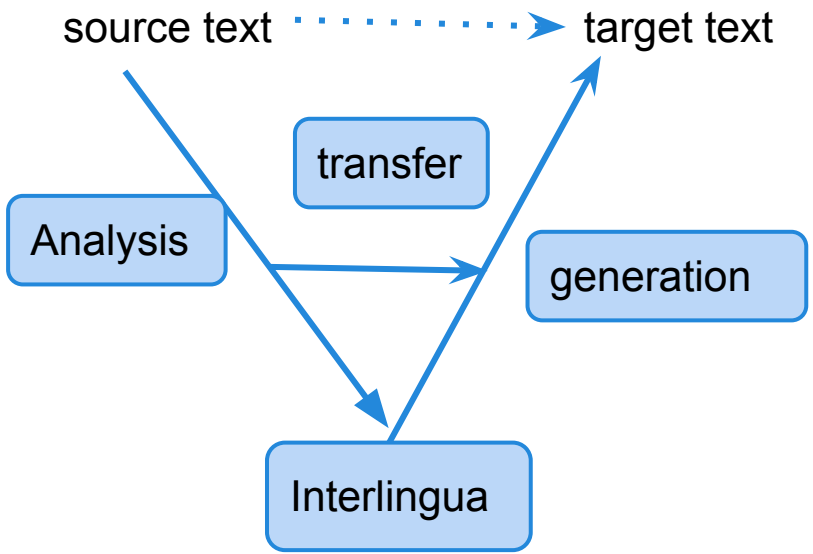
Aún no tengo una respuesta para eso. ¿Quieres que te contacte con una persona?

Tecnologías multilingües

Ximenta Gutiérrez-Vasques 2017

Traducción automática. Breve historia

- 1ra aplicación de la lingüística computacional
- Patentes desde 1933
- Gran atención e inversión durante la 2da Guerra mundial
- Reporte ALPAC (1966): los logros son muy pobres y se necesita investigación lingüística a largo plazo



Triángulo Vaquois

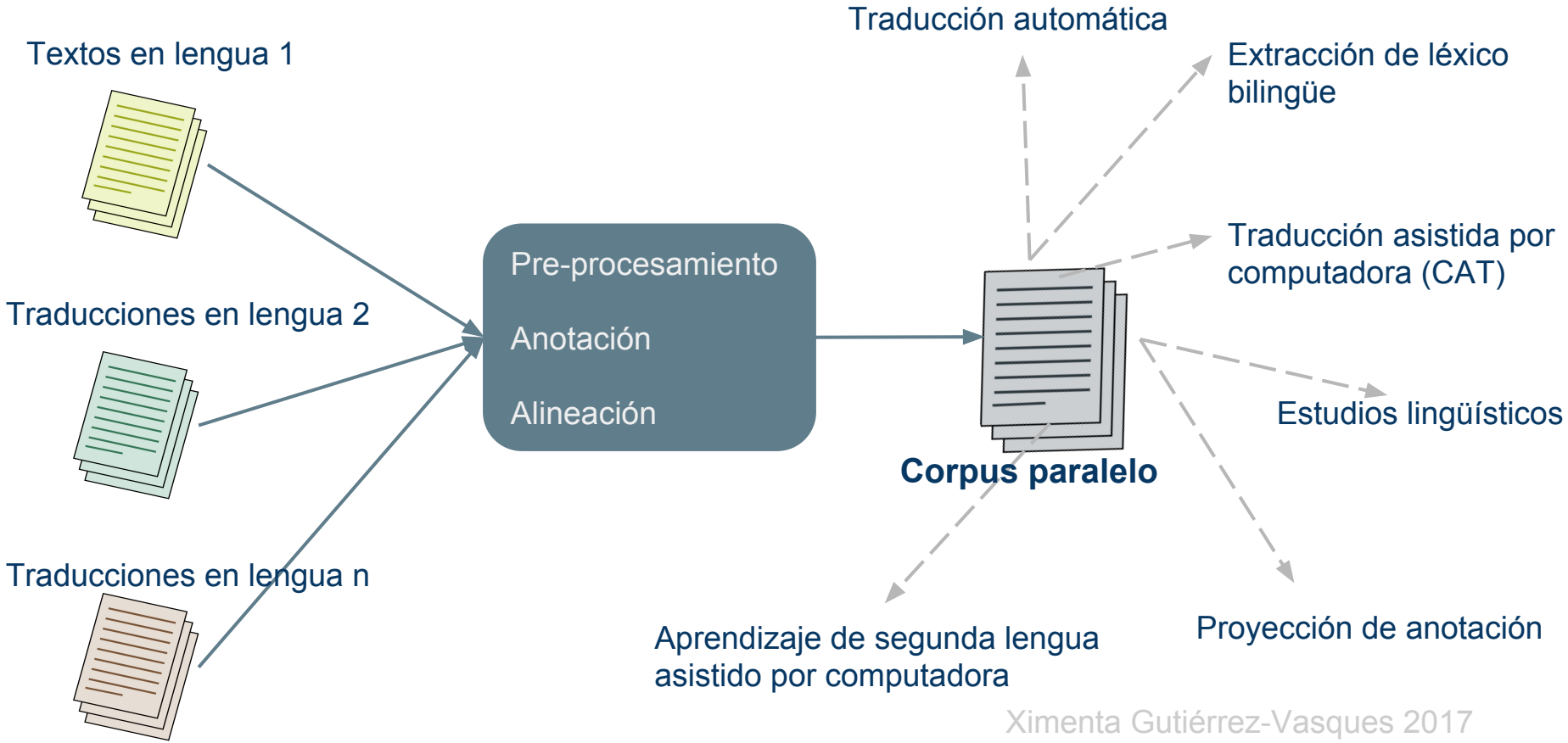
Traducción automática. Breve historia

- **Interés renovado** en la década de los 90's.

Tendencias actuales: **Métodos estadísticos**

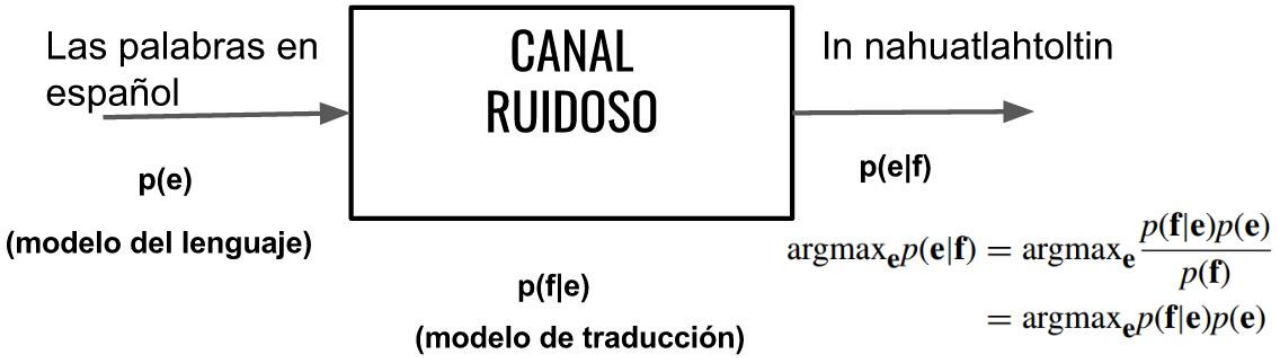
- En vez de diseñar reglas para traducir de una lengua a otra, podemos observar **miles de traducciones** y modelar automáticamente las relaciones bilingües

Traducción automática y corpus paralelos



Traducción automática estadísitica (SMT)

When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode. (Warren Weaver, 1949)



- p(f|e) se modela a partir de un corpus paralelo (alineado a nivel oración).
- Proceso generativo en donde se modelan las probabilidades de traducción a nivel palabra y, a partir de esto, se estima la traducción a nivel oración.

Traducción automática estadísitica (SMT)

- Tablas de traducción léxica

Tengo		flores		azules	
e	t(e|f)	e	t(e|f)	e	t(e|f)
nicpia	0.751	xochimeh	0.627	texohqueh	0.235
niquinpia	0.393	xochitl	0.590	yeloh	0.188
onicpiaya	0.265	cuiccaxochitl	0.374	texohtiqueh	0.183

Traducción automática estadística (SMT)

Se utiliza **esperanza-maximización** para estimar las tablas de traducción **léxica** t(e|f):

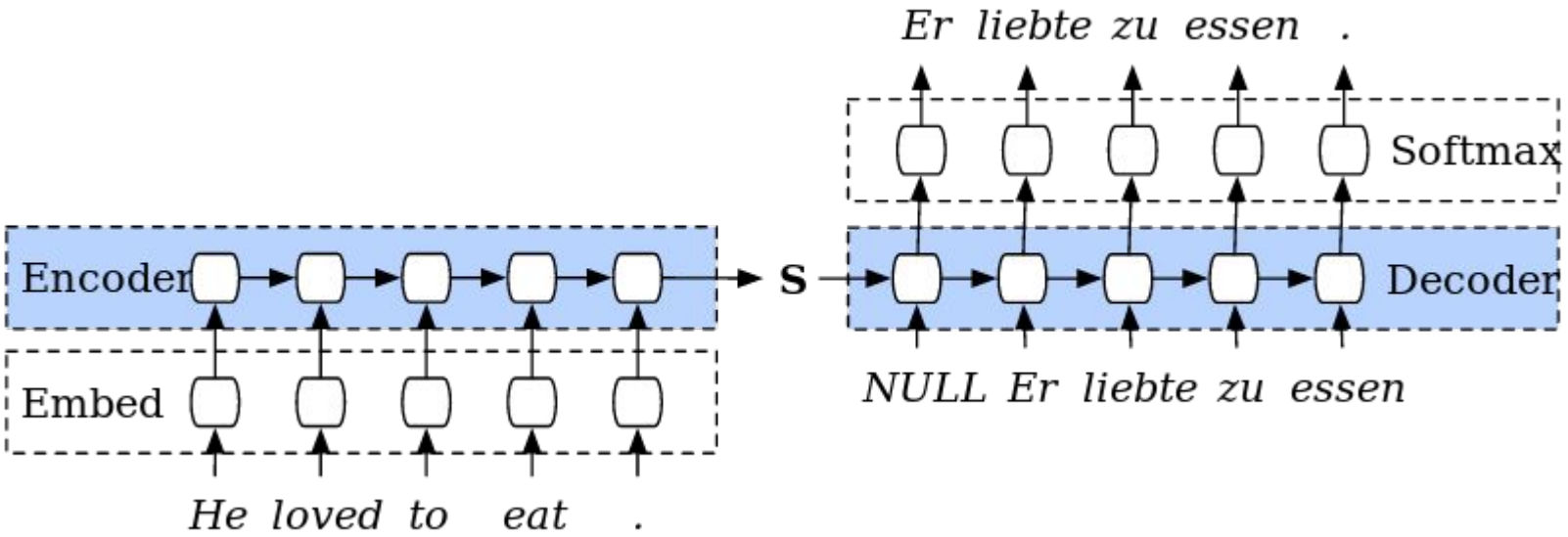
- El algoritmo se inicializa con una distribución probabilística uniforme (todas las alineaciones entre palabras son igualmente probables)
- Estas probabilidades se van refinando observando en el corpus paralelo qué pares de traducción co-ocurren en más oraciones paralelas

Este es el planteamiento básico del modelo **IBM-1** pero existen más modelos

Traducción automática neuronal (NMT)

- Noción similar: **Codificamos** en la lengua fuente, **decodificamos** en la lengua destino
- Haciendo uso de **redes neuronales profundas**, y corpus paralelos, podemos generar una representación abstracta para una oración en inglés, y para su equivalente en francés.
- Estas representaciones deben ser similares pues comparten lo esencial: **el significado**
- Las palabras se representan como **vectores** (embeddings)

Traducción automática neuronal (NMT)



Traducción automática, limitaciones

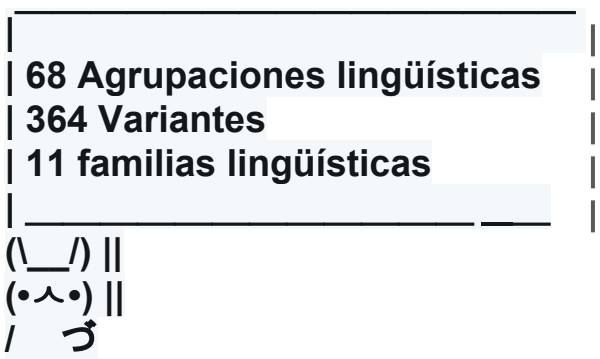
- Métodos altamente dependientes de la cantidad de datos
- Entre más distantes sean las lenguas, se necesita mayor corpus de entrenamiento.
- No todas las lenguas del mundo tienen grande corpus digitales, procesados, listos para utilizarse

Par de lenguas	Corpus de entrenamiento (palabras)
francés-inglés	**40 millones**
árabe-inglés	**200 millones**
chino-inglés	**200 millones**

*Philipp Koehn, Statistical Machine Translation, Cambridge University Press, 2010.

El caso de México

Ximenta Gutiérrez-Vasques 2017



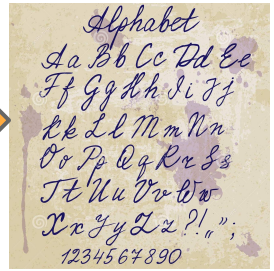
- México es un país con una enorme **diversidad lingüística**
- Sin embargo, prácticamente **ninguna tecnología** se ha desarrollado para estas lenguas
- A pesar de tener carácter nacional, rara vez es posible encontrar **contenido Web** gubernamental, turístico, educativo, etc. Así como **plataformas digitales**

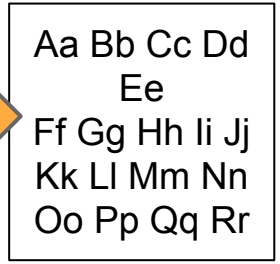
Caso 1. Español-Náhuatl

- Construcción de un **corpus paralelo digital** español-náhuatl

Recopilación bibliográfica de libros con traducciones

Digitalización

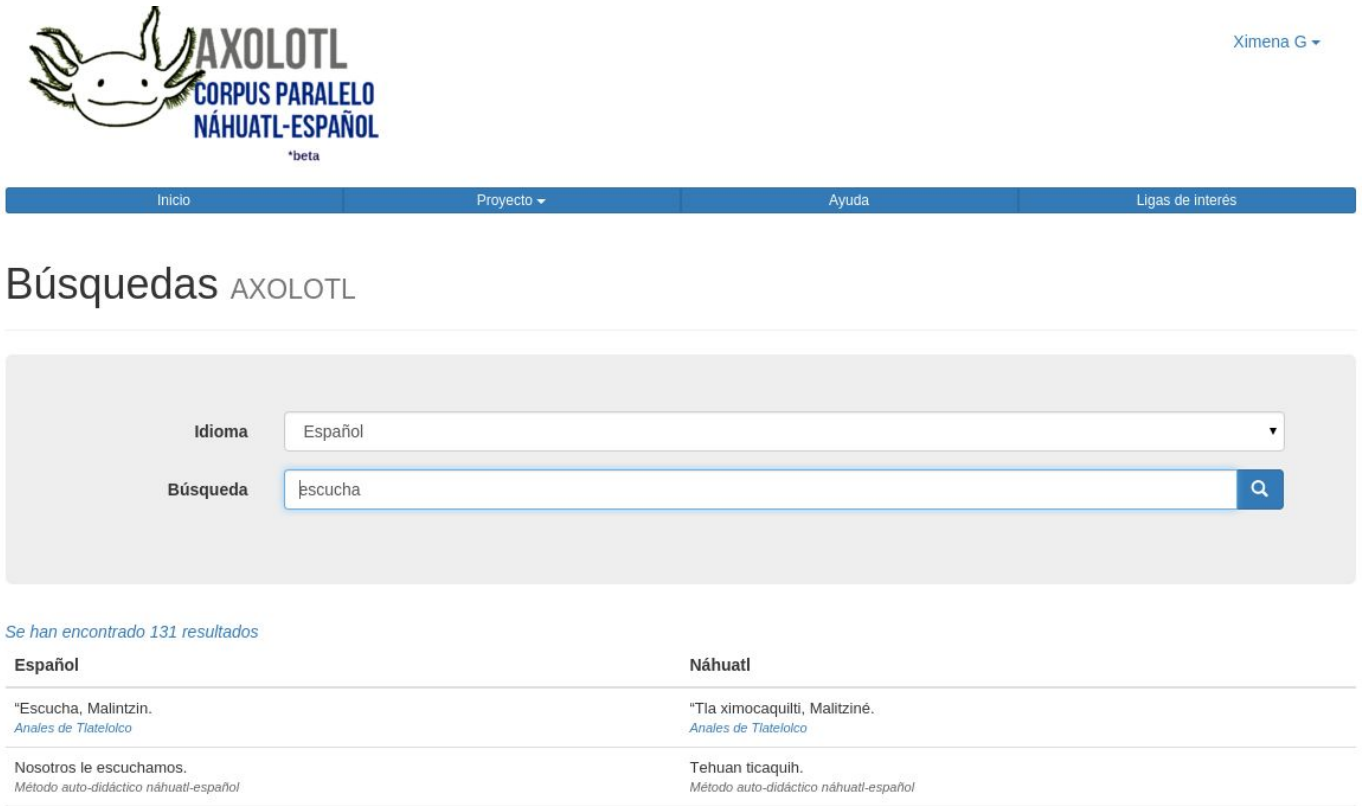




Corrección

Sistema Web

www.corpus.unam.mx/axolotl



Caso 1. Español-Náhuatl

- Metodología para extracción léxica bilingüe automática español-náhuatl. Condiciones experimentales: **par de lenguas distantes**, **corpus paralelo pequeño**, una de las lenguas (náhuatl) **carece de herramientas** y recursos digitales

ti - c - cohua - z ("lo comprarás")
2sg.s-3sg.o-'comprar'-fut
ni - c - cohua ("lo compro")
1sg.s-3sg.o-'comprar'
ni- c - cohua - tica ("lo estoy comprando")
1sg.s-3sg.o-'comprar'-prog

Correspondencia léxica buscada: comprar-cohua

Caso 1. Español-Náhuatl

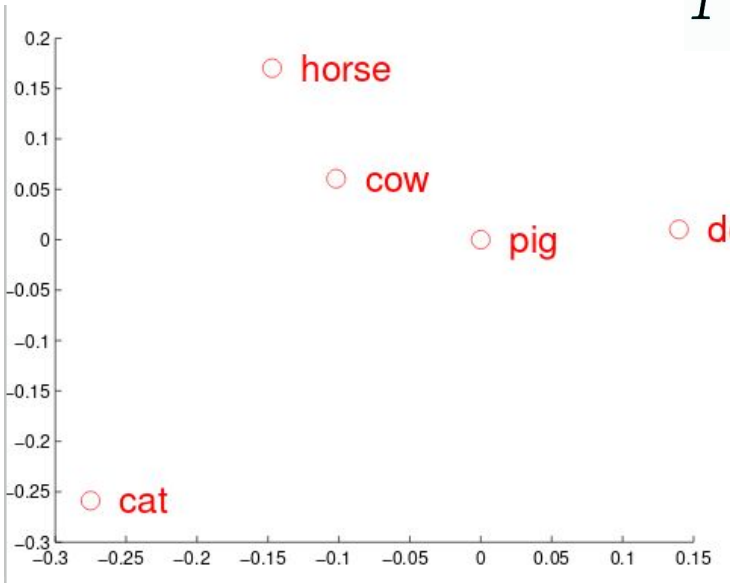
Estrategias:

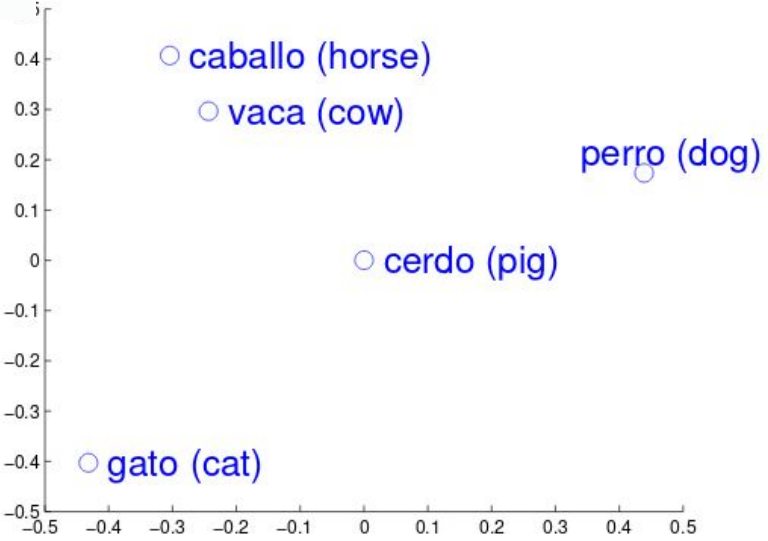
- **Segmentación morfológica semi-supervisada.** Trabajar con unidades sub-palabra (en vez de palabras completas) ayuda a mejorar las representaciones y tener más “repeticiones”
- **Aprender una transformación lineal para proyectar vectores de una lengua a otra*.** Representaciones vectoriales de las traducciones entre dos lenguas mantienen un arreglo geométrico similar (de hecho estas regularidades son lineales)

Exploiting Similarities among Languages for Machine Translation (Mikolov et. al 2013)

Caso 1. Español-Náhuatl

$$T : \mathbb{R}^d \rightarrow \mathbb{R}^{d_2}$$





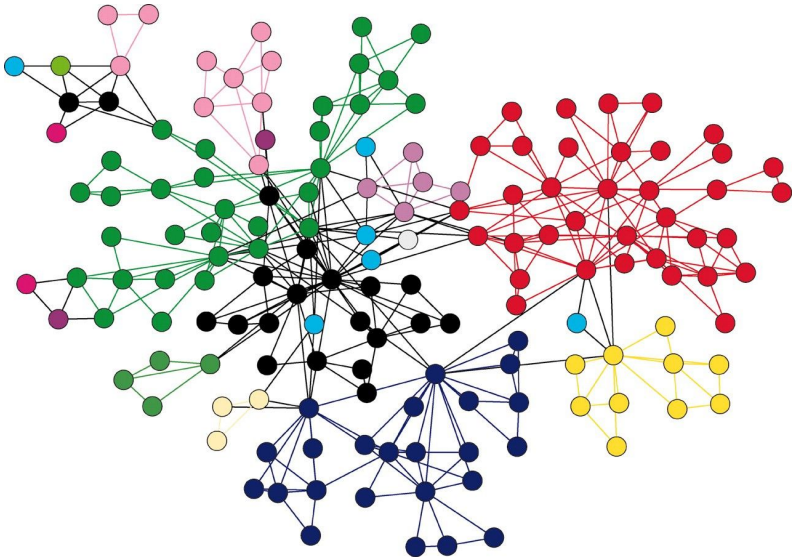
$$\min_T \sum_{i=1}^{N} ||T(x_i) - y_i||^2 + \lambda ||T||^2$$

Caso 1. Español-Náhuatl

- Las representaciones distribuidas como Word2Vec no funcionan cuando son inducidas con pocos recursos.

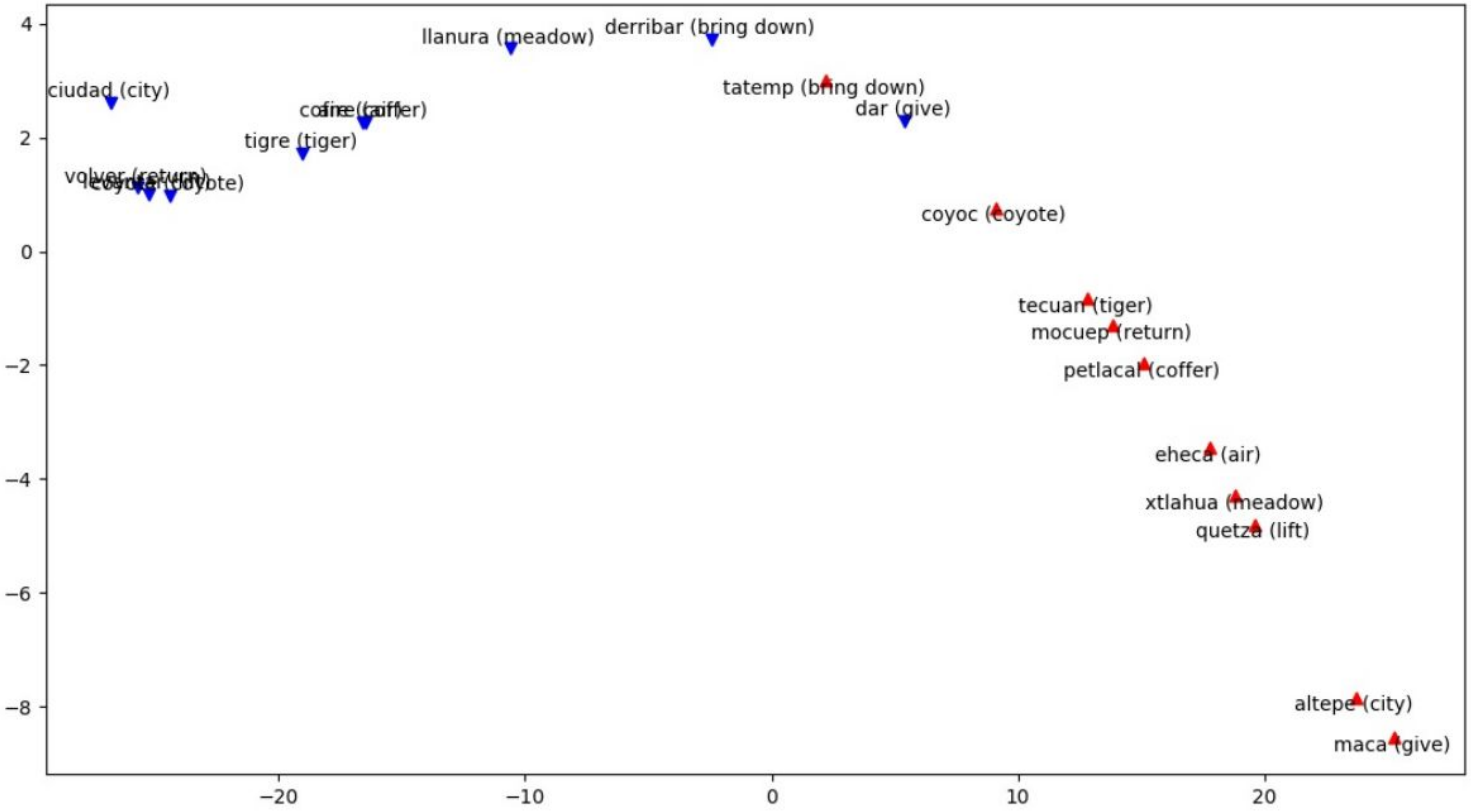
Proponemos nuestras **representaciones vectoriales multilingües basadas en grafos**:

1. Construimos un **grafo** (a partir de candidatos de traducción y score obtenido con un método estadístico)
2. Una vez obtenida esta estructura, convertimos cada nodo a una representación vectorial utilizando el algoritmo **Node2Vec**

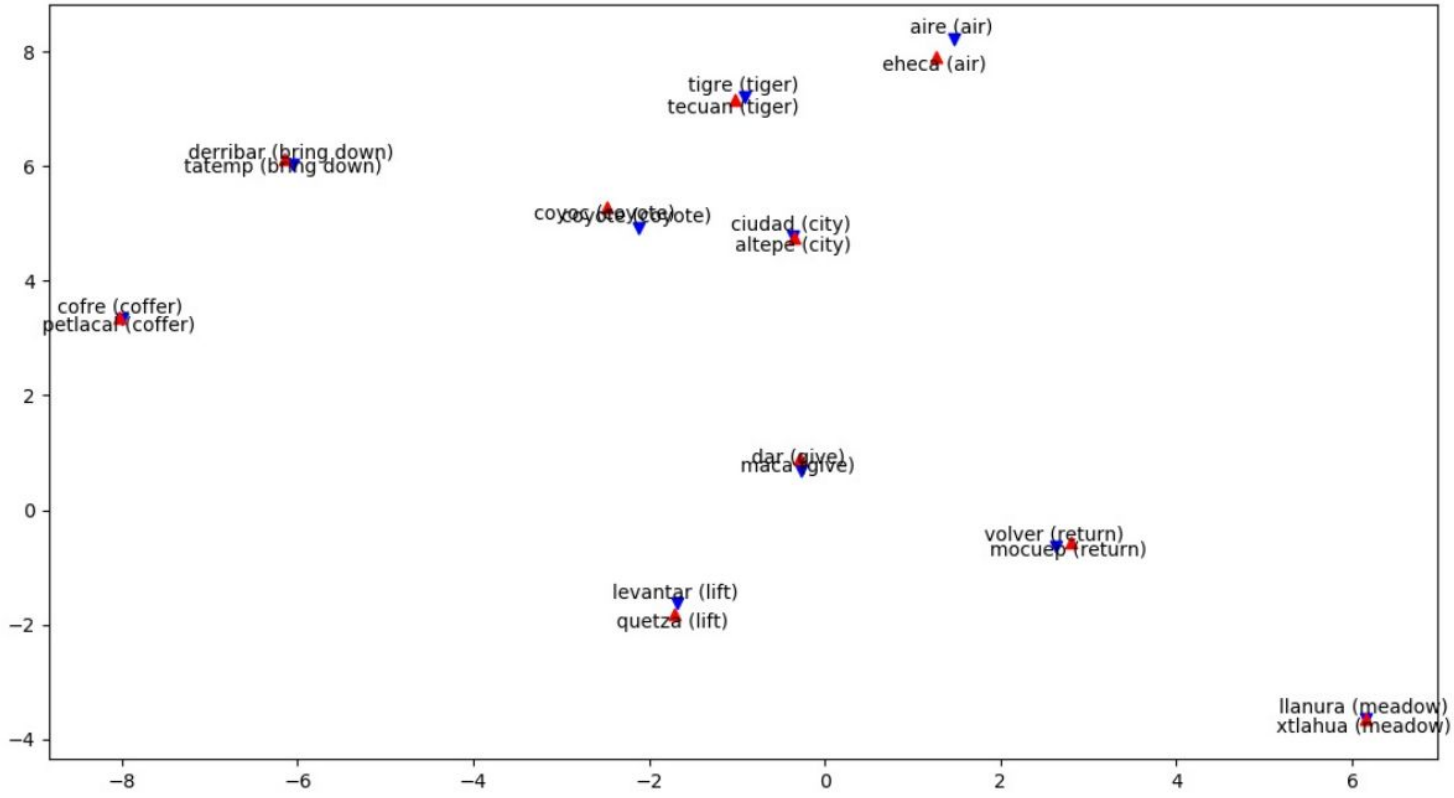


Ximenta Gutiérrez-Vasques 2017

Vectores word2vec español-náhuatl



Vectores node2vec español-náhuatl



Corpus paralelo español-otomí
https://tsunkua.elotl.mx

Idioma: Español

Búsqueda: México

Ayuda

México

Encontramos 26 resultados relacionados con «México»

Exportar resultados CSV

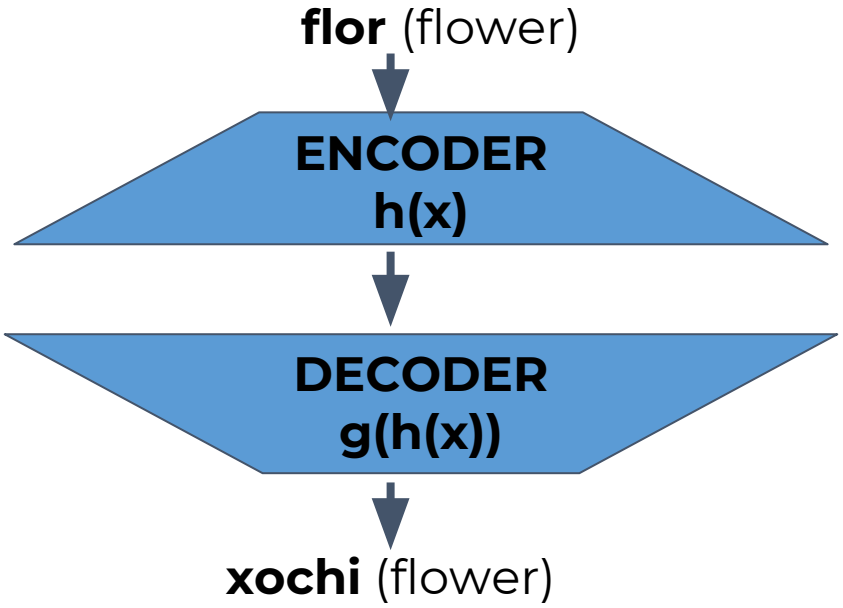
Show 10 entries

Español	Otomí (Hñahñu)
acerca de la Conquista y la ruina final de su metrópoli, México-Tenochtitlan. *Visión de los vencidos (hñahñu)*	getbu̱ ra ts'okat'o̱t'amfeni ne ra yo̱t'e ra hnini M'onda—Bondo. *Visión de los vencidos (hñahñu)*

Aprendiendo una transformación no lineal para el español-náhuatl...

Denoising autoencoder:

$$\arg\min_{h,g} \frac{1}{N} \sum_{i=1}^{N} ||x_i - g(h(\hat{x}_i))||^2$$



José Luis Olivares Castillo
@otrofama

Más información

- ***Low-resource bilingual lexicon extraction using graph based word embeddings***

Ximena Gutierrez-Vasques, Víctor Mijangos

Caso 2. Otomí

Etiquetado morfológico automático débilmente supervisado

- Uso de Conditional Random Fields (CRF's)

m=ba=tsuhu

PSD=3.ICP=venir

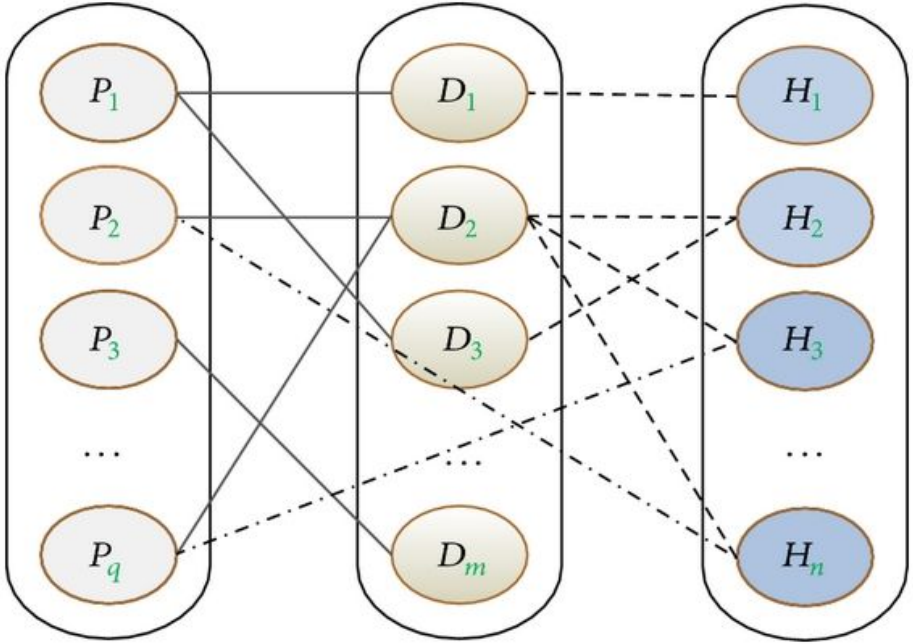
'Venía hacia aqui'

En marcha…

Diego Alberto Barriga

@umoqnier

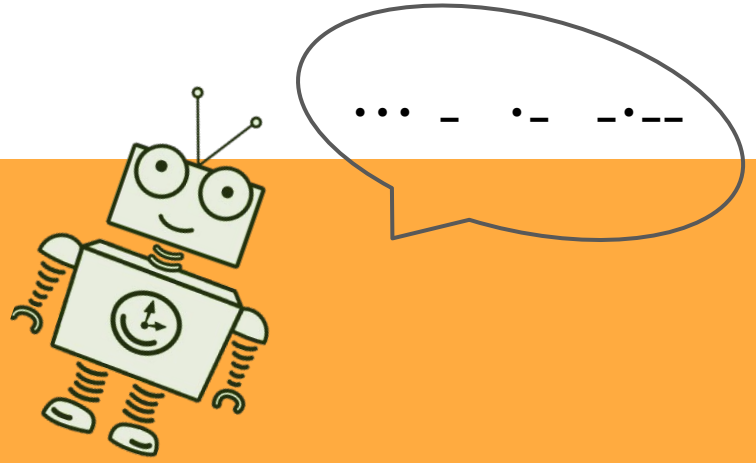
Caso 2. Otomí



- Grafos y representaciones vectoriales para obtener traducciones náhuatl-otomí
- Proyecto en desarrollo

Daniel Vargas Sánchez

Comentarios finales



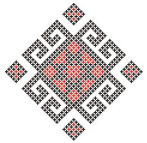
Consideraciones importantes

- Las lenguas originarias de México carecen de normalización **ortográfica**. Esto es un problema grande en NLP (muchas grafías asociadas a la misma palabra)
- **Gran variación dialectal**, textos con muchos diversos orígenes
- **Escasez de herramientas** para automatizar procedimientos

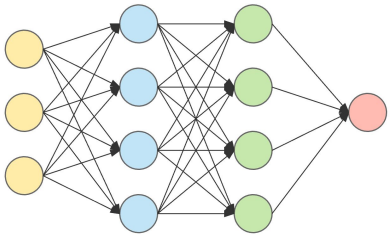
  Estas lenguas pueden beneficiarse de los avances en **teoría de ML** para entornos con pocos datos de entrenamiento (enfoques no supervisados, zero-shot learning, etc)

Consideraciones importantes

- Realizar investigación y desarrollo para estas lenguas no sólo puede tener **impacto social** positivo



- También representa un importante **reto tecnológico y científico** por las características que exhiben estas lenguas



Gracias
(tlasohkamati)
(jamädi)

xim@unam.mx

contacto@elotl.mx

@elotl_ **elotl.mx**