Hacia la traducción automática de las lenguas indígenas de México

Towards Machine Translation of the indigenous languages of Mexico

Jeús Manuel Mager Hois, Ivan Valdimir Meza Ruiz
Instituo de investigaciones en Matemáticas Aplicas
Universidad Nacional Autónoma de México

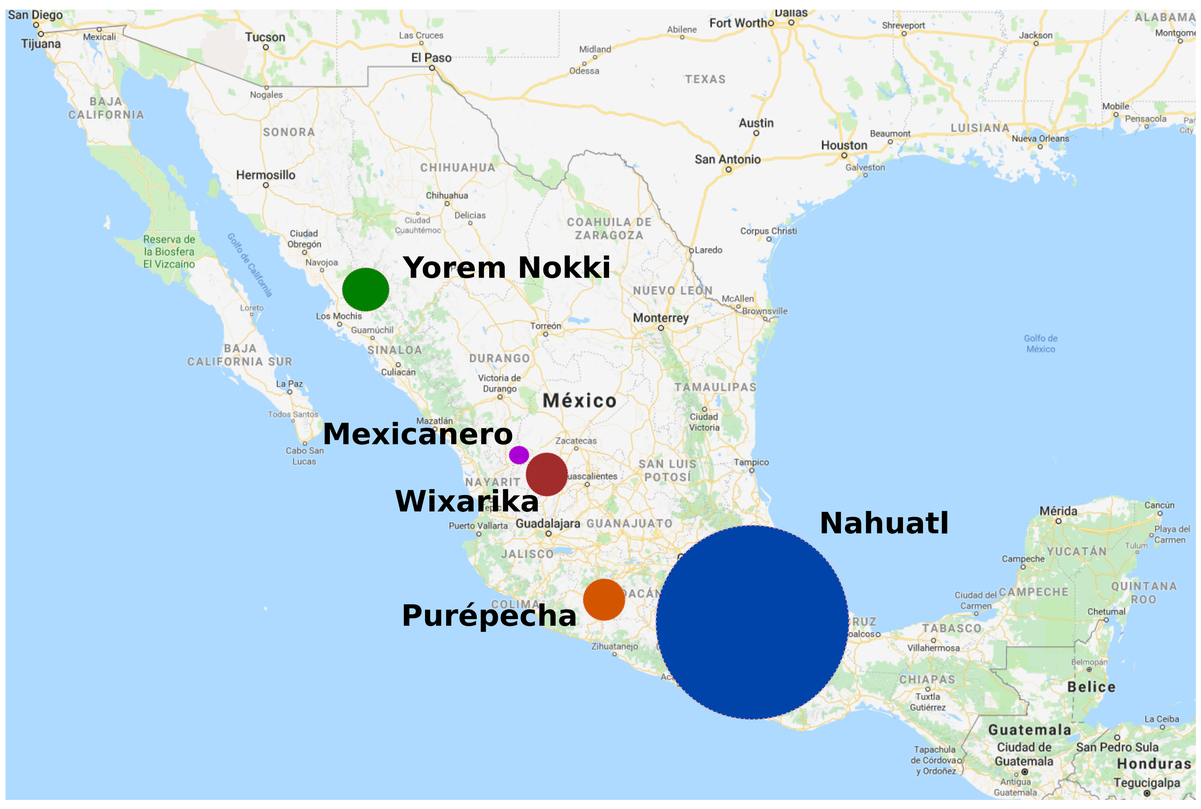
Introducción

En México se hablan 68 lenguas indígenas. Sin embargo, los esfuerzos en materia tecnológica enfocados en estas lenguas son casi nulos. En el presente trabajo mostramos primeros resultados en traducción automática entre cinco de estas lenguas y el español. Usamos metodos automáticos de traducción, los cuales nos permitirán ampliar nuestro estudio a un mayor número de lenguas en el futuro.

Introduction

In Mexico 68 indigenous languages are spoken. However, the technological efforts focused on these languages are almost nil. In the present work we show first results in Machine Translation between five of these languages and Spanish. We use data diven approaches, which will allow us to expand our study to a greater number of languages in the future.

Los idiomas



Estudiamos la traducción automática de cuatro lenguas de la familia yuto-nahua (mexicanero, nahuatl, wixárika y yorem nokki) y la lengua aislada purépecha, al español. Las cinco lenguas tienen una tipología polisintética.

Los datos usados

Para los cinco casos, usamos frases traducidas del **Archivo de Lenguas indígenas del Colegio de México**. Logramos recabar los siguientes datos:

	Train	Dev	Test
Mexicanero-Español	545	78	156
Nahuatl-Español	559	79	160
Purépecha-Español	559	80	159
Wixarika-Español	598	84	170
Yorem Nokki-Español	524	74	149

Tabla 1: Número de Frases paraleleas y su uso.

Modelos del experimento

Realizamos experimentos con SMT por frases y NMT a nivel morfológico. Para SMT utilizamos Moses y Giza. Para NMT utilizamos OpenNMT con un modelo Codificador-Decodificador (seq2seq) con atención y con celdas GRU. Cómo hiperparámetros usamos optimización Adam con un rate de aprendizaje de 0.00028, un tamaño de RNN de 1000 y un tamaño de encoding de 200.

Resultados

Para la evaluación utilizamos le métrica BLEU y comprobamos que los idiomas con menor razón morfema/palabra tienen mejor desempeño.

	NMT	SMT
Mexicanero-Español	2.95	23.5
Náhuatl-Español	3.04	10.1
Purépecha-Español	0	5.38
Wixarika-Español	0	0
Yorem Nokki-Español	0	2.44

Tabla 2: Resultados de la traducción automática usando evaluación BLEU (mas alto es mejor).

Conclusiones y Retos

Los resultados presentados son primeras aproximaciones a la traducción automática de estas lenguas. Con ello identidicamos los sigueintes retos para trabajo futuro:

Escasez de recursos.
Complejidad Morfológica
Traducción de lenguas distantes
Falta de estandarización ortográfica.
Amplio espectro dialectal interno en las lenguas.

Otros trabajos nuestros.

Sistema MT online Wixarika-Español:
`http://turing.iimas.unam.mx/wix/`
Segmentador morfológico para lenguas polisintéticas:
`http://turing.iimas.unam.mx/wix/mexseg`
Corpus Wixarika-Español
`https://github.com/pywirrarika/wixarikacorpora`
Lista de NLP sobre lenguas indígenas.
`https://pywirrarika.github.io/naki/`

Contacto

Jesús Manuel Mager Hois
`mmager@turing.iimas.unam.mx`
Ivan Vladimir Meza Ruiz
`ivanvladimir@turing.iimas.unam.mx`