Tecnológico de Monterrey | Posgrados

Maestría en Inteligencia Artificial Aplicada (MNA)

Traductor Español-Mayo

Hub de Ciencia de Datos

Presenta:

- Rafael Sergio Garcia Martinez A00529676

Asesor: Ricardo Ambrocio Ramírez Mendoza

Patrocinador(es): Dr. Juan Arturo Nolazco Flores

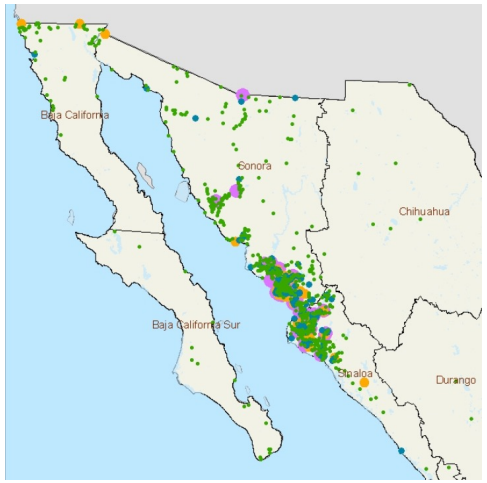
Fecha 18 Junio 2024

Presentación del Proyecto Integrador

Problemática

Establecer un Traductor Español – Mayo – Español basado en Inteligencia Artificial para Texto y Voz.

En las regiones de pueblos originarios de México, sobre todo en áreas no densamente pobladas existe riesgo de desaparecer las lenguas originarias, así como su cultura pre-hispánica.
La Región Mayo Yoreme como se les conoce a las comunidades del Sur del Estado de Sonora están en esta situación de riesgo.
En esta Región existen alrededor de 42,000 habitantes de esta Etnia, la mayoría hablan español que es la lengua dominante, el monolingüismo es bastante bajo donde solo niños y ancianos hablan el mayo.



Agrupación lingüística mayo							
Hablantes de lengua indígena		Habla también español					
		Sí		No		No especificado	
total	%	total	%	total	%	total	%
42,601	100	40,454	95.0	108	0.3	2,039	4.8
FUENTE: Microdatos de la Encuesta Intercensal, INEGI 2015.							

Problemas y Objetivos Generales

Problemas en la comunidad:

- Uso del dilecto Mayo:
  La falta de recursos educativos para hablar Mayo entre los jóvenes y niños han propiciado que cada día existan menos usuarios de la lengua en actividades comunitarias, dejando el uso solo en ambiente familiar.
- Acceso a la Tecnología:
  La tecnología en las comunidades indígenas no se ha dispersado lo suficiente como para promover el uso de aplicaciones o traductores que accedan al Mayo.
- Cultura:
  Las tradiciones o festivales de las comunidades ayudan a difundir y dar a conocer su dialecto, es necesario promover los espacios culturales hacia las comunidades mayoritarias.
- Comunicación:
  Las comunidades se aíslan por temor al rechazo o discriminación al no tener la confianza de usar su lenguaje de una forma abierta en cualquier situación en público.

Propósito en la Sociedad:

- Mejorar la Comunicación:
  Contribuir con material de apoyo para promover el uso de tecnologías en las comunidades Mayo con el fin de incrementar el uso del lenguaje entre sus miembros.
- Educativo:
  Colaborar con las instituciones educativas para enriquecer los procesos de traducción lingüística con usuarios finales de la comunidad.
- Cultura Comunitaria:
  Participar en el crecimiento del lenguaje Mayo con la integración de la tecnología en las comunidades sobre todo en la vida cotidiana de sus habitantes.

Situacion Actual

Disponibilidad de Datos:

Para el Lenguaje Español es fácil encontrar base de datos con millones de palabras y frases que facilitan el entrenamiento en los modelos dedicados a Traducción o Reconocimiento de voz

European Parliament

TCSE TED Corpus Search Engine

Hugging Face

opensubtitles.org

Wikipedia The Free Encyclopedia

TATOEBA.ORG

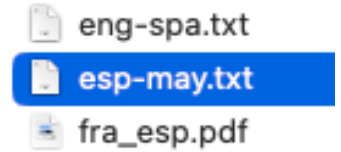
Dilecto Mayo:
Una de las principales desventajas en los dialectos pre-hispánicos es su poca documentación, y el lenguaje Mayo no es la excepción, ya que los mismos integrantes reconocen la lengua como una tradición oral de comunicación y no escrita.

Estructura de Datos

Se propone alinear las palabras o frases en forma paralela, es decir, deberá estar la palabra en español y enseguida el valor correspondiente a la traducción Mayo, esto facilita el análisis del modelo de procesamiento de lenguaje que se quiera utilizar.

Español	Mayo	
abajo, debajo	bétucu	
abanico	tapichaléero	
abeja	muúmum, culiichi	
abejita	cuta cúmelera	
abejón	cucusebora	
ablandar, ablandecer		buálcote
abogado	yore nocríame	

Los archivos para procesar se recomiendan utilizar en formato de texto separados por un tabulador (\t) para no agregar otro símbolo que pueda interferir a la hora de leer los datos.



Se mantuvo un archivo genérico en Excel para captura donde según el modelo se puede generar el archivo de texto necesario, si se requiere en un futuro relacionar Ids con audio o análisis de sentimientos.

ID	Frase Ingles	Frase Español	Frase en Mayo	Estado
0011_000001	The nine the eggs, I keep.	Los nueve huevos, me los quedo.	ume totori cahugmane thwunkake	Neutral
0011_000002	I did go, and made many prisoners.	Fui, e hice muchos prisioneros.	aman sika yunm pa'tinteuok	Neutral
0011_000003	That I owe my thanks to you.	Que te debo mi agradecimiento.	amanne jmuk'te may lamadniacay	Neutral

Modelos Adecuados de IA

Modelo NLP utilizando Transformers con el modelo *Atencion*
Proceso utilizando la librería MLearner:

- Lectura y limpieza de datos con expresiones regulares
- Tokenizar Texto con Tensorflow
- Limpieza de frases o palabras
- Se crean las entradas y salidas
- Entrenamiento del Transformer con keras
- Evaluación del modelo con frases de los datos

Modelo Transformer - Entrenamiento

```
[ ]  tf.keras.backend.clear_session()

     # Hiper Parámetros
     D_MODEL = 128 # 512
     NB_LAYERS = 4 # 6
     FFN_UNITS = 512 # 2048
     NB_PROJ = 8 # 8
     DROPOUT_RATE = 0.1 # 0.1

     model_Transformer = Transformer(vocab_size_enc=VOCAB_SIZE_EN,
                                     vocab_size_dec=VOCAB_SIZE_ES,
                                     d_model=D_MODEL,
                                     nb_layers=NB_LAYERS,
                                     FFN_units=FFN_UNITS,
                                     nb_proj=NB_PROJ,
                                     dropout_rate=DROPOUT_RATE)
```

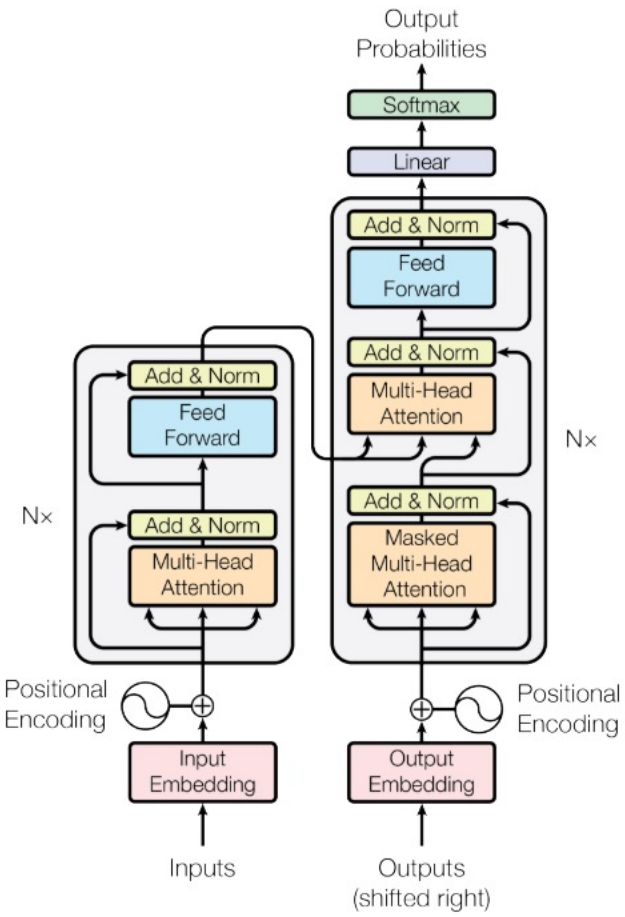


Figure 1: The Transformer - model architecture.

Predicciones

```
[ ]  translate("I did go")

     Entrada: I did go
     Traducción predicha: Y lo dijo con dijo .
```

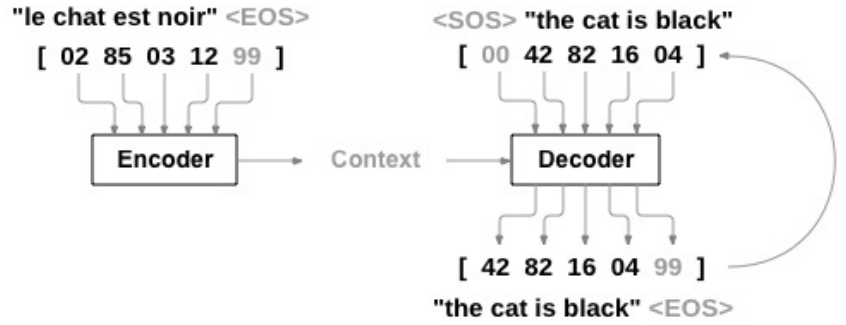
```
[ ]  translate("This is a problem we have to solve.")

     Entrada: This is a problem we have to solve.
     Traducción predicha: Y su su su su su su su su su
```

Se utilizaron las frases disponibles en los datos recolectados y el resultado no fue favorable, este modelo requiere de una cantidad muy grande de datos para alcanzar una buena eficiencia por eso no se vio aceptable.

Modelos Adecuados de IA

Modelo NLP Seq2Seq

Este modelo se basa en dos RNN para transformar una secuencia a otra, la red encoder condensa en un vector la secuencia de entrada y la red decoder despliega el vector en una nueva secuencia, además agrega el modelo de *Atencion* para agregar contexto a la secuencia.



Lectura y Limpieza de Datos

```
# Se normaliza el texto para manejar
# minusculas y remover espcios y caracteres no-letras

def normalizeString(s):
    s = unicodeToAscii(s.lower().strip())
    s = re.sub(r"([.!?])", r" \1", s)
    s = re.sub(r"[^a-zA-Z.!?]+", r" ", s)
    s = re.sub(r"[^A-Za-zñÑáéíóúÁÉÍÓÚ ]", r" ", s)
    return s
```

Clase para RNN Encoder

```
class EncoderRNN(nn.Module):
    def __init__(self, input_size, hidden_size):
        super(EncoderRNN, self).__init__()
        self.hidden_size = hidden_size
        print('input_size=', input_size)
        self.embedding = nn.Embedding(input_size, hidden_size)
        self.gru = nn.GRU(hidden_size, hidden_size)
```

Clase para RNN Decoder

```
class DecoderRNN(nn.Module):
    def __init__(self, hidden_size, output_size):
        super(DecoderRNN, self).__init__()
        self.hidden_size = hidden_size

        self.embedding = nn.Embedding(output_size, hidden_size)
        self.gru = nn.GRU(hidden_size, hidden_size)
        self.out = nn.Linear(hidden_size, output_size)
        self.softmax = nn.LogSoftmax(dim=1)
```

Clase para aplicar Atencion

```
class AttnDecoderRNN(nn.Module):
    def __init__(self, hidden_size, output_size, dropout_p=0.1, max_length=MAX_LENGTH):
        super(AttnDecoderRNN, self).__init__()
        self.hidden_size = hidden_size
        self.output_size = output_size
        self.dropout_p = dropout_p
        self.max_length = max_length
```

Metodo de Evaluacion

```
def evaluateAndShowAttention(input_sentence):
    output_words, attentions = evaluate(
        encoder1, attn_decoder1, input_sentence)
    print('input =', input_sentence)
    print('output =', ' '.join(output_words))
    showAttention(input_sentence, output_words, attentions)
```

En este Modelo se obtuvo una respuesta más aceptable aun con la limitante de datos, además es sencillo el escalamiento para cuando crezcan los datos de entrada

Retos

•Captura de datos en comunidades indígenas donde se requiere la convivencia con la comunidad para lograr su colaboración.

• Evaluación de Modelos con entrenamientos adecuados a los datos disponibles para optimizar recursos de cómputo.

• Asegurar la validación de los modelos con datos nuevos de las personas originarias.

Beneficios:

• Implementar una herramienta de fácil uso para las personas en las comunidades indígenas con acceso ya sea a móvil o web para su beneficio.

• Enfocar la implementación en una causa de conservación de la lengua, así como motivar su uso y conocimiento atreves de las tecnologías disponibles de comunicación.

El utilizar las tecnologías de Inteligencia Artificial en conjunto con los sistemas de comunicación web o móvil puede incrementar el uso y esparcimiento de estas lenguas, el desafío es grande al proponerse obtener una buena base de datos que puedan manejar los existentes o nuevos modelos por venir de procesamiento de lenguaje natural.

Indígena Mayo

Trabajos Futuros

- Continuar con el crecimiento de la Base de Datos textual, así como la de voz.
- Participar en la comunidad de Etchojoa con la etnia Mayo para implementar un Traductor.
- Colaborar con el Hub de Ciencia de Datos para estandarizar la forma de capturar datos de lengua pre-hispánicas.
- Probar Algoritmos de reconocimiento de voz en lengua Mayo.
- Crear el vínculo con la etnia Mayo para futuros trabajos del Hub.

GRACIAS