DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
CARRERA DE SOFTWARE

ARTICULO ACADÉMICO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SOFTWARE

TEMA: "HERRAMIENTA DE TRADUCCIÓN AUTOMÁTICA NEURONAL DEL ESPAÑOL AL INGLÉS EN EL ÁMBITO MÉDICO"

AUTOR:
GORDILLO LUCAS, ARIEL SANTIAGO

DIRECTOR:
ING. UYAGUARI UYAGUARI, ALVARO DANILO

LATACUNGA
AGOSTO, 2022

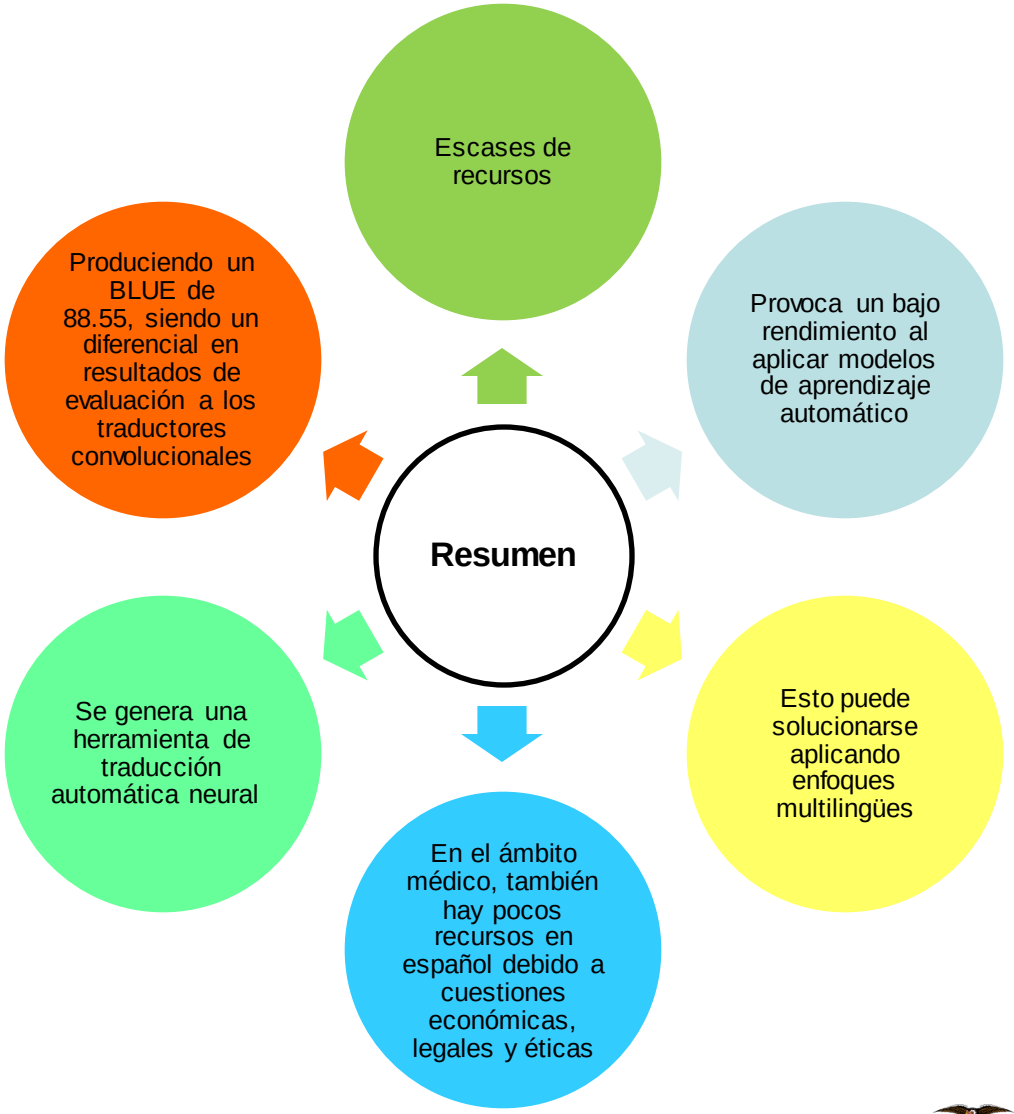
“Las cosas simples deben ser simples, las complejas deben ser posibles.”

Alan Kay

CONTENIDO

- Resumen
- Antecedentes investigativos
- Planteamiento del problema
- Objetivo general
- Objetivos específicos
- Hipótesis
- Implementación del modelo
- Implementación del sitio web
- Análisis de resultados
- Conclusiones
- Recomendaciones

RESUMEN


Escases de recursos
Provoca un bajo rendimiento al aplicar modelos de aprendizaje automático
Esto puede solucionarse aplicando enfoques multilingües
En el ámbito médico, también hay pocos recursos en español debido a cuestiones económicas, legales y éticas
Se genera una herramienta de traducción automática neural
Produciendo un BLUE de 88.55, siendo un diferencial en resultados de evaluación a los traductores convolucionales
Resumen

ANTECEDENTES INVESTIGATIVOS

En 2015 se desarrolla un método que combina el procesamiento del lenguaje natural (NLP) y la revisión manual asistida por computadora de notas clínicas para identificar evidencia de uso problemático de opioides en registros de salud. Dando como resultados una alta predicción en la detección de un consumo indebido de opioides.

En 2017 se presentó un sistema de predicción en la admisión hospitalaria de pacientes en un departamento de emergencias en el cual se evidencio una mejora en el traslado y el triaje de pacientes mejorando también la inclusión de texto libro a partir de las predicciones, esto, mediante y el análisis de componentes principales a partir del motivo de visita del paciente.

ANTECEDENTES INVESTIGATIVOS – Traducción

En el campo de la traducción automática de texto escrito de un lenguaje natural a otro, ha experimentado un importante cambio de paradigma en los últimos años.

2010

- Sistema de traducción donde se integra n-gramas conexionistas

2014

- Proponen una nueva arquitectura con una red neuronal recurrente codificador-decodificador

2016

- Se pretende construir una única red neuronal para tareas de traducción.

2017

- Proponen una solución simple para usar un solo modelo de TAN multilenguaje.

2019

- Proponen una nueva arquitectura de traducción seq2seq para destacar la importancia de las dependencias secuenciales

ANTECEDENTES INVESTIGATIVOS – Traducción - Score

BLEU es un algoritmo para evaluar la calidad del texto que ha sido traducido automáticamente de un lenguaje natural a otro.

2017
- Se implementa un traductor multilenguaje de español a inglés, generando un BLEU de 29.7

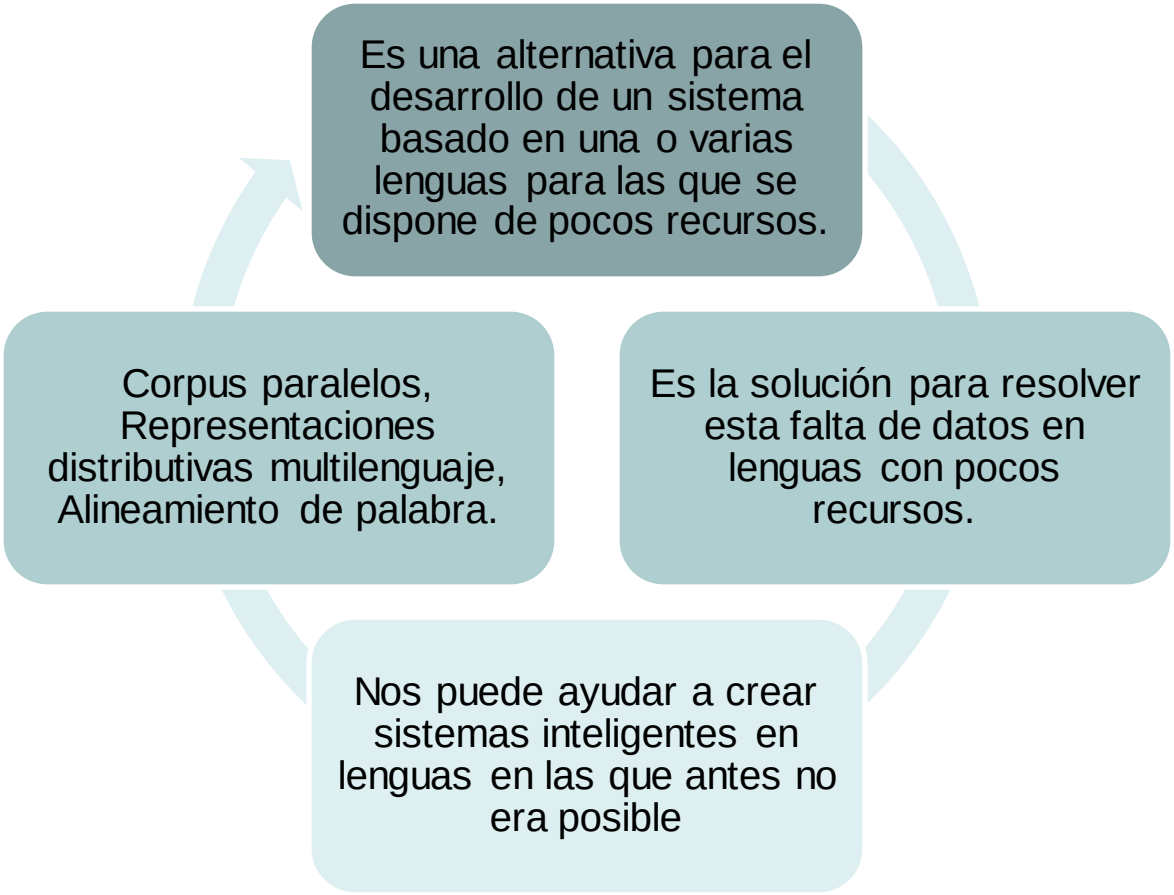
2019
- Evaluan modelos para la traducción donde genera un BLEU de 56.47

En el dominio médico

2020
- Se implementan tareas de traducción biomédicas con redes neuronales recurrentes utilizando el corpus EMA dando un BLEU 0.541

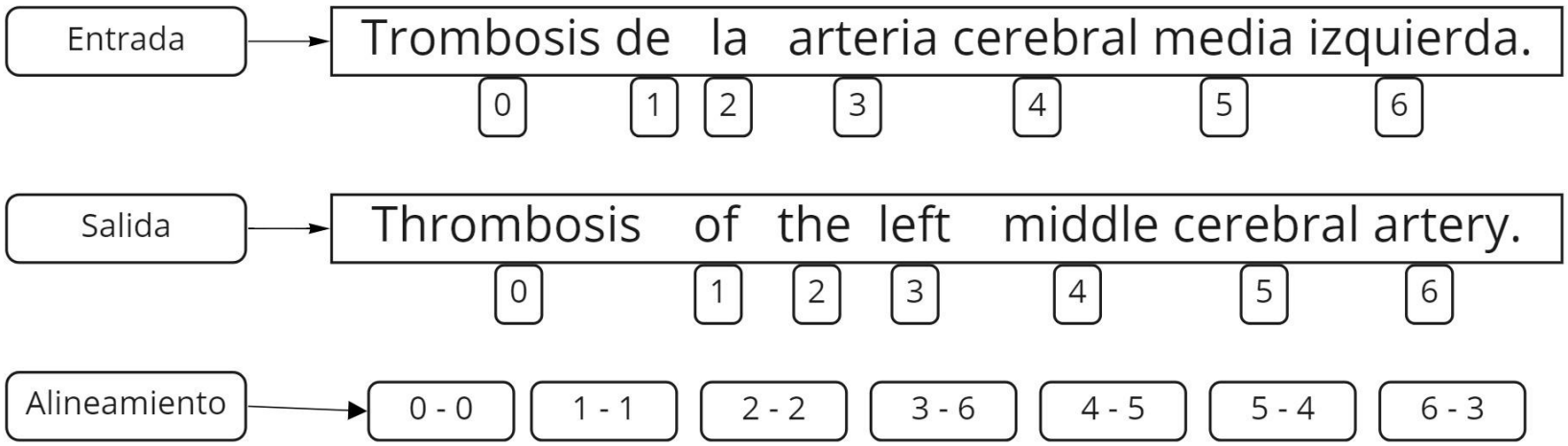
2021
- Un traductor biomédico con alineamiento de palabras del WMT 2021 de español a inglés, donde su mayor BLUE es de 0.538

ANTECEDENTES INVESTIGATIVOS – CrossLingual



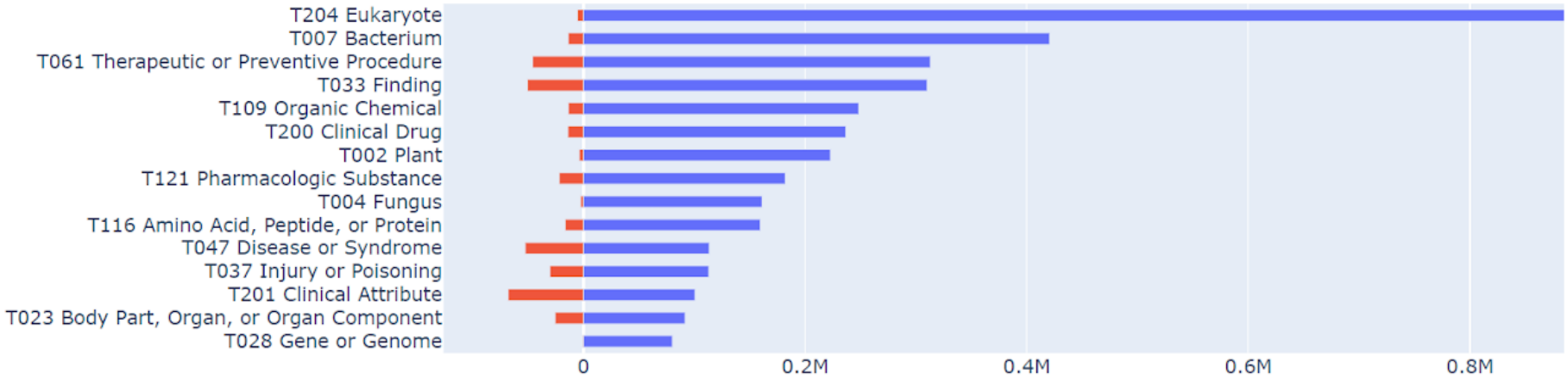
ANTECEDENTES INVESTIGATIVOS – Alineamiento de palabras

Se define como la detección de la alineación correspondiente entre las palabras de oraciones paralelas traducidas una de otra.



PLANTEAMIENTO DEL PROBLEMA

Unified Medical Language System (UMLS)



Tipos semánticos en español.

Tipos semánticos en inglés.

PLANTEAMIENTO DEL PROBLEMA

Generar un traductor que ayude a la generación de corpus etiquetados mediante el alineamiento de palabras.

Desarrollar un modelo que con contenido médico para la generación de traducciones automáticas.

"HERRAMIENTA DE TRADUCCIÓN AUTOMATICA NEURAL CON ALINEAMINTO INDUCIDO"

Hoy en día solo el 9% de UMLS está en español generando una escasez de recursos lingüísticos como corpus paralelos, modelos preentrenados o corpus paralelos.

En la actualidad no se disponen de muchos modelos de español a ingles en el campo médico para su implementación en tareas de NLP.

OBJETIVO GENERAL

- Desarrollar un traductor automático neuronal con alineamiento de palabras inducido para incrementar los recursos lingüísticos en español.

OBJETIVOS ESPECÍFICOS

- Fundamentar teóricamente las incidencias más notables de traductores automáticos neuronales a lo largo de estos años.
- Identificar y seleccionar corpus paralelos dentro del dominio médico.
- Determinar un modelo arquitectónico que asegure un buen rendimiento en el preprocesamiento y entrenamiento de los corpus.
- Implementar un modelo que contenga las mejores épocas del entrenamiento.
- Inducir e alineamiento de palabras en las etapas intermedias del edsarrollo del modelo.
- Asegurar un BLEU score lo suficientemente bueno a comparación del estado del arte.
- Implementar un sitio web que permita traducciones.

HIPÓTESIS

¿Si creamos un traductor automático neural del español al inglés en el dominio médico entonces incrementamos recursos lingüísticos en este idioma?

IMPLEMENTACIÓN DEL MODELO

- Fase 1: Selección de la herramienta para el modelado de secuencias
- Fase 2: Selección de Corpus paralelos
- Fase 3: Preprocesamiento
- Fase 4: Entrenamiento
- Fase 5: Inferencia
- Fase 6: Validación

Fase 1: Selección de la herramienta para el modelado de secuencias

Toolkit	Sponsor	Google Scholar citations	Last update	Language	Implement Word Alignment
OpenSeq2Seq	NVIDIA	31	2 years ago	Python	No
OpenNMT	Harvard NLP and SYSTRAN	1645 Version 2017 20 Version 2020	3 months ago	PyTorch inits latest version	Yes
Seq2SeqPy	IDEX Universite´ Grenoble Alpes	3	1 year ago	PyTorch	No
Sockeye	Amazon	201 Version 1 7 Version 2	2 months ago	PyTorch	Yes
Fairseq	Facebook	1439	A few days ago	PyTorch	Yes

Fase 2: Selección de Corpus paralelos

Scielo

Este corpus reúne publicaciones electrónicas de artículos y textos completos de revistas científicas de América Latina, Sudáfrica y España. Actualmente cuenta con el apoyo de la Fundación de Investigación de São Paulo (FAPESP) y el Consejo Nacional de Desarrollo Científico y Tecnológico de Brasil (BIREME)

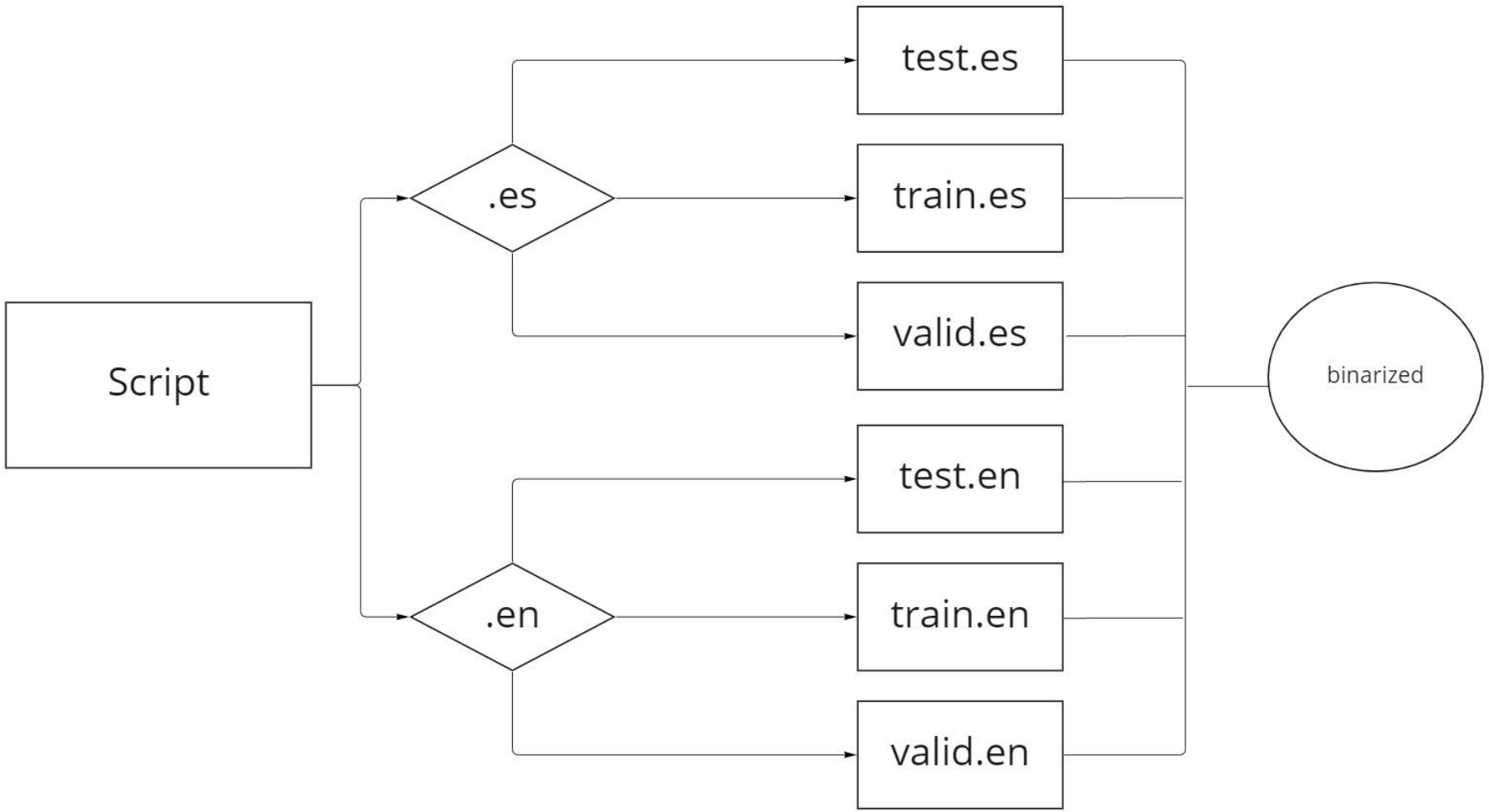
Cuenta con dos archivos .es y .en con 177.781 oraciones cada uno dentro del dominio médico

EMA

EMA es un corpus de documentos biomédicos recuperados de la Agencia Europea de Medicamentos (EMA). Incluye documentos relacionados con medicamentos y sus traducciones a 22 lenguas oficiales de la Unión Europea.

Cuenta con dos archivos .es y .en con 1.098.327 cada uno, con oraciones, frases y términos dentro del dominio médico

Fase 3: Preprocesamiento



Fase 4: Entrenamiento

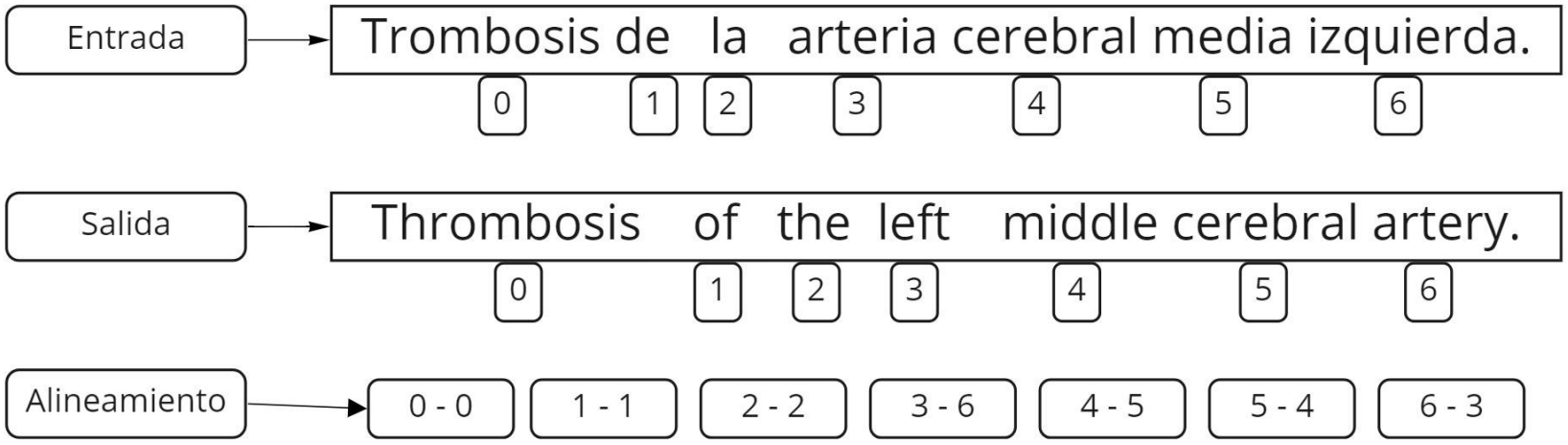
- Se implementa fairseq-train para el entrenamiento del modelo.
- Se aplica la tarea TranslationTask.
- Se implementa el modelo Transformer.
- Se especifica un modelo de guía.
- Se inicializa con un modelo de alineación de palabras TransformerAlignModel, y se toma como referencia transformer_wmt_en_de_big align.
- Cada época generada se almacena por defecto en un fichero llamado checkpoints.best, y guarda una copia de modelo en un fichero llamado checkpoints.last.

Fase 5: Inferencia

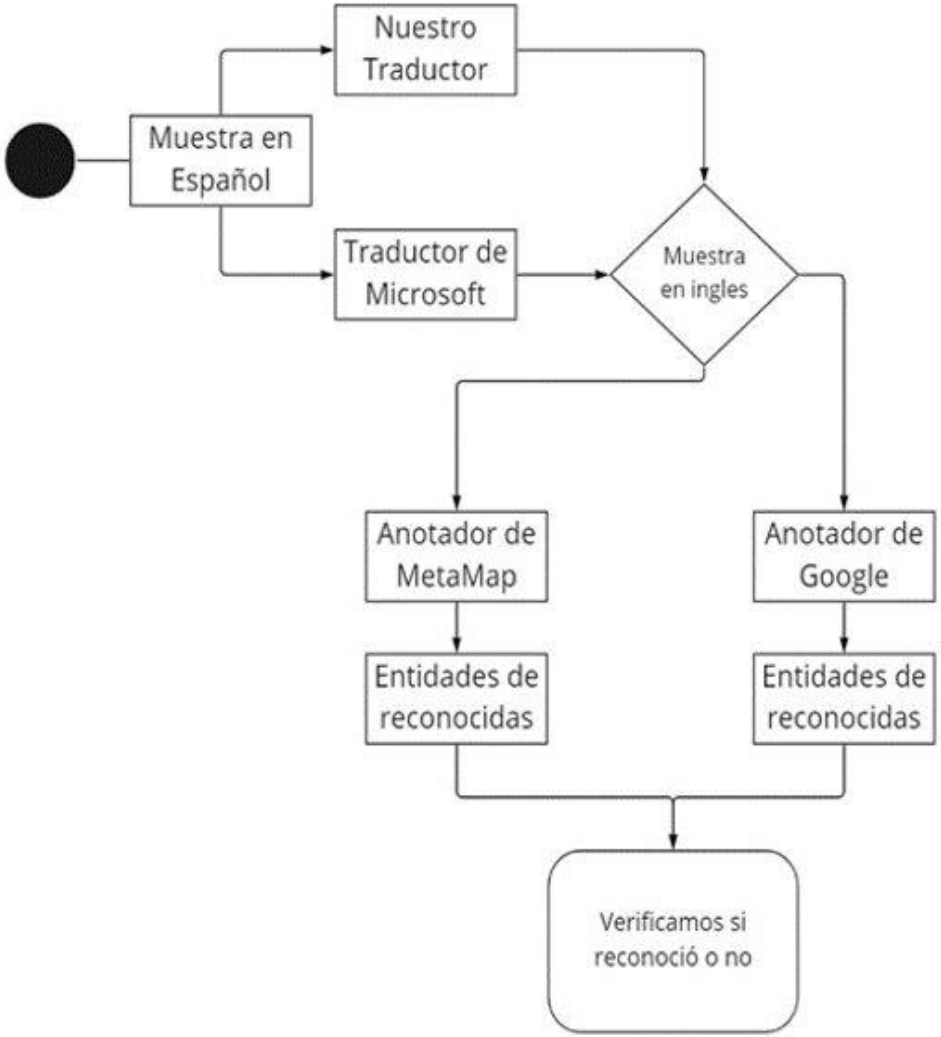
- Se utilizan los ficheros paralelos llamados test.
- Mediante fairseq-interactive, y con los ficheros binarizados, se especifica el lenguaje de entrada src=en y el lenguaje de salida trg=en.
- Se añade la salida de alineación mediante print-alignment.
- Se envía el modelo, checkpoint.best.

- **S**-> Oración de entrada
- **H**-> Hipotesis
- **D**-> Hipotesis no etiquetada
- **P->** Pesos
- **A->** Alineamiento

```
...  2022-06-24 20:53:34 | INFO | fairseq.tasks.text_to_speech | Please install tensorboardX: pip install tensorboardX
     2022-06-24 20:53:36 | INFO | fairseq_cli.interactive | {'_name': None, 'common': {'_name': None, 'no_progress_bar': False, 'log_interval': 100, 'log_format': None, '
     2022-06-24 20:53:36 | INFO | fairseq.tasks.translation | [es] dictionary: 30624 types
     2022-06-24 20:53:36 | INFO | fairseq.tasks.translation | [en] dictionary: 30624 types
     2022-06-24 20:53:36 | INFO | fairseq_cli.interactive | loading model(s) from /content/drive/MyDrive/Titulación/Corpus/check/checkpoint_best.pt
     2022-06-24 20:53:52 | INFO | fairseq_cli.interactive | NOTE: hypothesis and token scores are output in base 2
     2022-06-24 20:53:52 | INFO | fairseq_cli.interactive | Type the input sentence and press return:
     Se trata de una mujer de 29 años sometida a un estudio ecográfico pélvico de control tras una ligadura de trompas por vía laparoscópica.
     S-0     <unk> trata de una mujer de 29 años <unk> a un estudio <unk> pélvico de control tras una <unk> de <unk> por vía <unk> .
     W-0     0.367   seconds
     H-0     -1.4464961290359497     these included a 29 year female patient who underwent a pelvic control study after receiving a subcutaneous dose .
     D-0     -1.4464961290359497     these included a 29 year female patient who underwent a pelvic control study after receiving a subcutaneous dose.
     P-0     -4.6658 -0.2258 -0.2488 -1.5063 -0.0790 -0.7440 -0.1660 -2.5470 -2.5376 -0.1869 -1.1555 -1.6946 -0.3170 -0.7450 -3.4165 -1.0141 -3.7571 -3.6944 -0.1229 -0.10
     A-0     0-0 24-1 24-2 6-3 7-4 4-5 24-6 8-7 8-8 13-9 13-10 15-11 11-12 16-13 18-14 23-15 23-16 22-17 24-18
```


Entrada
Trombosis de la arteria cerebral media izquierda.
0 1 2 3 4 5 6
Salida
Thrombosis of the left middle cerebral artery.
0 1 2 3 4 5 6
Alineamiento
0 - 0
1 - 1
2 - 2
3 - 6
4 - 5
5 - 4
6 - 3

Fase 6: Validación



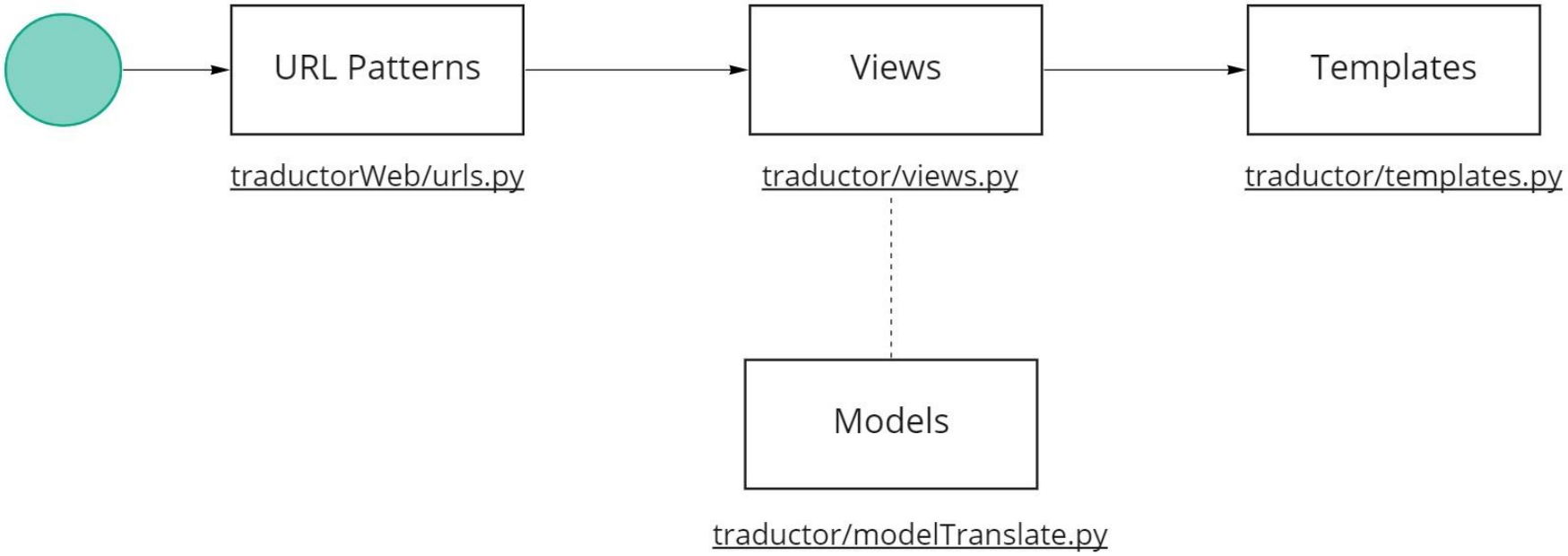
- Healthcare Natural Language API de Google
- MetaMap
- Reconocimiento de entidades biomédicas
- Corpus Gold del Clef

Entities recognized by Google Annotator		
Total entities	Microsoft Translator	Our Translator
83	27	**30**

Entities recognized by the MetaMap Annotator		
Total entities	Microsoft Translator	Our Translator
83	52	**52**

IMPLEMENTACIÓN DEL SITIO WEB

Implementación del sitio web



```python
from django.contrib import admin
from django.urls import path, include

urlpatterns = [
    path('admin/', admin.site.urls),
    path("", include('traductor.urls')),

]
```

```python
from django.urls import path
from . import views

urlpatterns=[
    path('index', views.index, name ='index'),
    path('models', views.models, name='models'),
    path('why', views.why, name='why'),
    path('ours', views.ours, name='ours'),

]
```

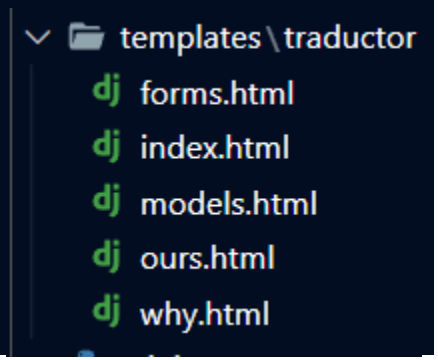
```python
def index(request):
    if not request:
          return render(request, "traductor/index.html")
    if request.method == 'POST':
        input_es = request.POST.get('input_esp')
        input_es = input_es.lower()
        dict = modelTranslate.Translate.Traductor(input_es)

        return render(request, "traductor/index.html", {"dict" : dict})
    return render(request, "traductor/index.html")

def ours(request):
    return render(request, 'traductor/ours.html')

def models(request):
    return render(request, 'traductor/models.html')

def why(request):
    return render(request, 'traductor/why.html')
```



```
class Translate ():
  def Traductor(input_esp):

    es_en = TransformerModel.from_pretrained(
    'traductor/static/utils',
    checkpoint_file='checkpoints/checkpoint_best.pt',
    data_name_or_path='traductor/static/utils/binarized',
    bpe='fastbpe',
    bpe_codes='traductor/static/utils/codes/codes'
    )
    output_ing = es_en.translate(input_esp)
    print(input_esp)
    print(output_ing)
    datos={}
    datos['dict'] = [" ",""]
    datos['dict'].append({'input_esp': input_esp, 'output_ing': output_ing})

    content=datos['dict']
    return(content)
```

ANÁLISIS DE RESULTADOS

ANÁLISIS DE RESULTADOS

Tras haber realizado la implementación con dos corpus paralelos y la aplicación de diferentes parámetros de optimización, se obtuvieron los resultados:

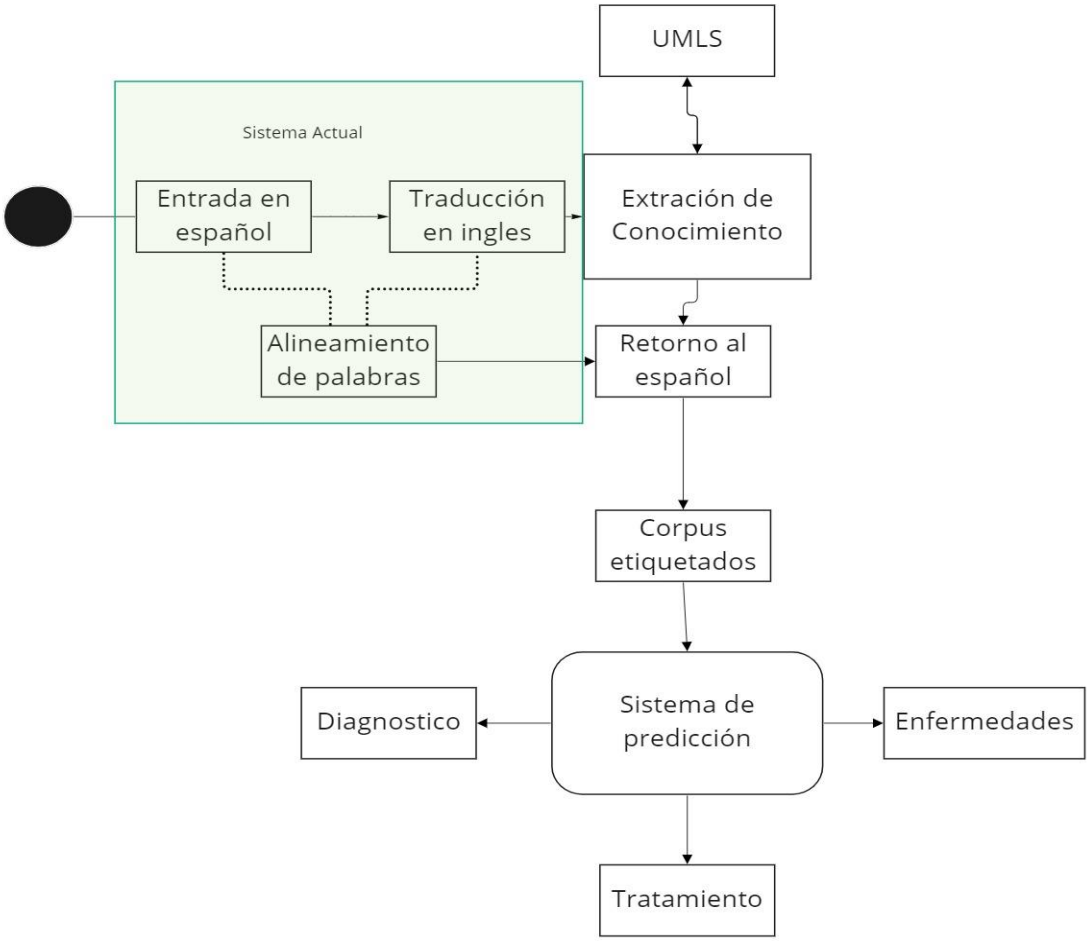
	EMEA			SCIELO		
Epoch	5	8	15	9	15	25
BLEU	78.21	83.61	**88.55**	33.87	45.85	**54.73**
batch-size	128	128	128	128	128	128
beam	5	5	5	5	5	5
train	929044	92904	929044	190351	190351	190351
test	40469	40439	40439	6984	6984	6984
valid	42229	42229	42229	7288	7288	7288

Análisis de Resultados

- El BLEU del corpus Scielo superó en varios puntos a los modelos preentrenados recomendados por Fairseq.
- Una vez que supimos que el rendimiento del NMP cumplía nuestras expectativas, comprobamos que, a diferencia de otros traductores similares, nuestro traductor superaba con creses un valor de 88,55.
- Generamos un BLEU eficiente con sólo 15 épocas de entrenamiento.
- En consecuencia, la eficiencia de nuestro traductor se puso a prueba frente a un traductor comercial como Microsoft Translator.

CONCLUSIONES

- Se puede concluir que el modelo que generó un BLEU de 88.55 evidenció una precisión alta en sintaxis y contexto en cada una de las traducciones.
- Se genero el score más alto en las tareas de traducción automática de español a inglés en el dominio médico.
- Además, mediante la comparación entre traductores se pudo detectar un número mayor de entidades médicas con respecto a las traducciones generadas por el traductor de Microsoft.
- Demostramos que nuestra herramienta se equipara y supera a un traductor comercial.
- Generamos un incremento de objetos biomédicos reconocidos mediante la traducción automática neuronal y el alineamiento de palabras del idioma español al inglés en el dominio médico.


UMLS
Sistema Actual
Entrada en español
Traducción en ingles
Extración de Conocimiento
Alineamiento de palabras
Retorno al español
Corpus etiquetados
Diagnostico
Sistema de predicción
Enfermedades
Tratamiento

RECOMENDACIONES

- Continuar investigando métodos que ayuden a las transferencias de conocimiento para lenguas con pocos recursos lingüísticos.
- Promover la documentación clara y precisa de los modelos desarrollados.
- Construir grupos de investigación para cubrir las necesidades en el ámbito medico dentro del procesamiento del lenguaje natural.

“Cada texto es único y, simultáneamente, es la traducción de otro texto. Ningún texto es enteramente original porque el lenguaje mismo, en su esencia, es ya una traducción: primero, del mundo no verbal y, después, porque cada signo y cada frase es la traducción de otro signo y de otra frase. Pero ese razonamiento puede invertirse sin perder validez: todos los textos son originales, porque cada traducción es distinta. Cada traducción es, hasta cierto punto, una invención y así constituye un texto único”

Octavio Paz