A3: Portugués a Ingles

Ana Valentina López Chacón

Traducción Automática
MIARFID, UPV

Enero, 2025

1. Descripción de la Práctica

Se buscó realizar diferentes experimentos sobre un *dataset* de Common Voice [1], específicamente para la tarea de traducción de Portugués a Ingles, se exploraron métodos de transcripción, modelos pre-entrenados y métricas de evaluación que permiten medir la calidad tanto de la transcripción como de la traducción.

1.1. Objetivos

Se plantearon los siguientes objetivos:

- Emplear diferentes modelos de Whisper para la transcripción de los audios en el idioma de origen y evaluar su error a nivel de palabra.
- Evaluar métricas de traducción empleando los resultados al utilizar diferentes modelos de Whisper.
- Realizar un experimento de traducción de habla en cascada sobre el *dataset* escogido y probar con diferentes modelos pre-entrenados.
- Ajustar los resultados del objetivo anterior y realizar una exploración de parámetros.

2. Análisis de los Datos

El *dataset* venía dividido en tres particiones: train, validation y test, cada una de estas contiene su correspondiente traducción de portugués a ingles con frases tomadas de situaciones reales, expuestas a ruido y que requieren un preprocesamiento. Su distribución de muestras venía definida de la siguiente forma:

Partición	Muestras	Palabras Totales	Palabras por Frase
Train	9,158	67,300	7.35
Validation	3,318	25,295	7.62
Test	4,023	31,742	7.89

Cuadro 1: Distribución de los Datos.

Se observó que las frases no eran muy extensas y no estaban relacionadas entre sí, por lo que, muchas de estas palabras eran conectores que dependiendo de su uso pueden alterar el significado de la frase.

3. Experimentación

Todos los experimentos se realizaron usando *google collab*, debido a los recursos limitados se optó por tomar subconjuntos de las muestras de train y test, de 2000 y 1000 respectivamente. Estos valores se tomaron en el orden en el que venían en el *dataset* para asegurar que los resultados fueran comparables entre sí.

3.1. Baseline de Reconocimiento de Habla

Como herramienta para transcribir las entradas de audio se empleo *Whisper* [2], un modelo general en reconocimiento de habla que también se puede usar en tareas de traducción e identificación de idiomas. Existen seis tamaños de modelo disponibles, para esta tarea se utilizaron el base, small y medium cuya descripción se da en el cuadro 2:

Tamaño	Parámetros	Modelo multilingüe	VRAM requerida	Velocidad relativa
base	74 M	base	1 GB	7x
small	244 M	small	2 GB	4x
medium	769 M	medium	5 GB	2x

Cuadro 2: Tamaños de Modelos y Rendimiento.

Ahora bien, para cada uno de estos tres modelos se realizó el mismo experimento con el subconjunto extraído de los datos de test, se concluyó que el mejor funcionamiento se daba con tamaño medium, luego se repitió el experimento para train con este modelo y se reportaron las métricas de *Word Error Rate* (WER) correspondientes (ver Cuadro 3)

Partición	Modelo Whisper	Word Error Rate (WER)
test	base	23.72 %
test	small	10.71 %
test	medium	7 %
train	medium	6.35 %

Cuadro 3: WER para tamaños de Whisper.

3.2. Baseline de Traducción de Habla

Se repitió el mismo experimento pero realizando la tarea de traducir y se probaron con los tres modelos de *Whisper* disponibles. En este caso se reportaron métricas de BLEU y COMET (ver Cuadro 4).

Partición	Modelo Whisper	BLEU	COMET
test	base	22.2 %	62.36 %
test	small	41 %	79.15 %
test	medium	51.8 %	86.38 %

Cuadro 4: BLEU y COMET para Traducción de Habla.

3.3. Traducción de Habla en Cascada

En esta parte se cargaron dos modelos pre-entrenados y se usaron para realizar las predicciones que posteriormente fueron evaluadas con el modelo de *Whisper* medium. NLLB (No Language Left Behind) y M2M100 (Many-to-Many Multilingual Translation) son modelos de traducción desarrollados por Meta. NLLB se centra en idiomas de bajos recursos, ofreciendo alta calidad para más de 200 idiomas y dialectos, mientras que M2M100 es ideal para traducción muchos a muchos en 100 idiomas, destacando en idiomas de medios y altos recursos. Sus resultados para la tarea asignada se dan en el siguiente cuadro 5:

Partición	Modelo	BLEU	COMET
test	NLLB	49.9 %	86.28 %
test	M2M100	41.4 %	82.09 %

Cuadro 5: BLEU y COMET para Traducción de Habla en Cascada.

3.4. Finetuning de Traducción de Habla en Cascada

Se realizaron cuatro experimentos de ajuste o finetuning con el fin de mejorar los resultados obtenidos en la etapa anterior, logrando adaptar los modelos pre-entrenados a la tarea de traducción escogida. Las primeras dos pruebas siguieron la misma secuencia de tareas y los mismos parámetros para NLLB y M2M100, se concluyó que los mejores resultados se dan para NLLB. Adicionalmente, se realizó una búsqueda de hiperparámetros con la librería *optuna* de `python` y de forma paralela se verificó el cambio en las métricas unicamente al darle al modelo datos limpios, es decir, al emplear el método `BasicTextNormalizer()` de *Whisper*.

Partición	Modelo	BLEU	COMET
test	NLLB	53.1 %	86.9 %
test	M2M100	44.8 %	83.05 %
test	NLLB (clean)	54 %	87.3 %
test	NLLB (parameter search)	53.3 %	86.97 %

Cuadro 6: BLEU y COMET para Finetuning de Traducción de Habla en Cascada.

A partir de estos resultados (ver Cuadro 6) se determinó que el mejor resultado se da al usar los datos limpios y que los mejores valores de hiperparámetros son: {'learning_rate': 0.00031379463096667376, 'batch_size': 2, 'lora_r': 32, 'lora_alpha': 16, 'lora_dropout': 0.09819107823091107}.

4. Conclusiones

Para finalizar este informe podemos destacar las siguientes conclusiones:

- Las muestras de audio que componen el *dataset* son tomadas de contextos muy diferentes y no siguen una linealidad, esto le aporta gran variedad y da pie a generar un modelo más robusto y consistente. Los mejores resultados para *Whisper* se dan para el tamaño medium, en este caso era lo suficientemente ligero para probarse en diferentes entornos pero manteniendo un nivel de error bajo.
- A partir de las métricas observadas se puede afirmar que NLLB es mejor para lenguas con poca representación, y M2M100 se adapta a tareas generales de traducción multilingüe sin depender del inglés como intermediario, ya que está entrenado de forma mucho más general y puede que destaque en entornos multilingües. Otros modelos pre-entrenados como XLM-RoBERTa, GPT o MarianMT requerian una cantidad de recursos mucho mayor e incluso no contenían el idioma de origen necesitado.
- En general se obtuvieron muy buenos resultados para esta tarea de traducción resaltando la gran aplicabilidad de herramientas como *Whisper* y el uso de modelos pre-entrenados al momento de generar predicciones. De igual forma se resalta la importancia de la limpieza de texto a la hora de ajustar y evaluar un modelo, especialmente si se miden métricas que son sensibles a errores de similitud.

Referencias

[1] *Mozilla Common Voice*. (n.d.). https://commonvoice.mozilla.org/en/datasets

[2] Openai. (s. f.). *GitHub - openai/whisper: Robust Speech Recognition via Large-Scale Weak Supervision*. GitHub. https://github.com/openai/whisper