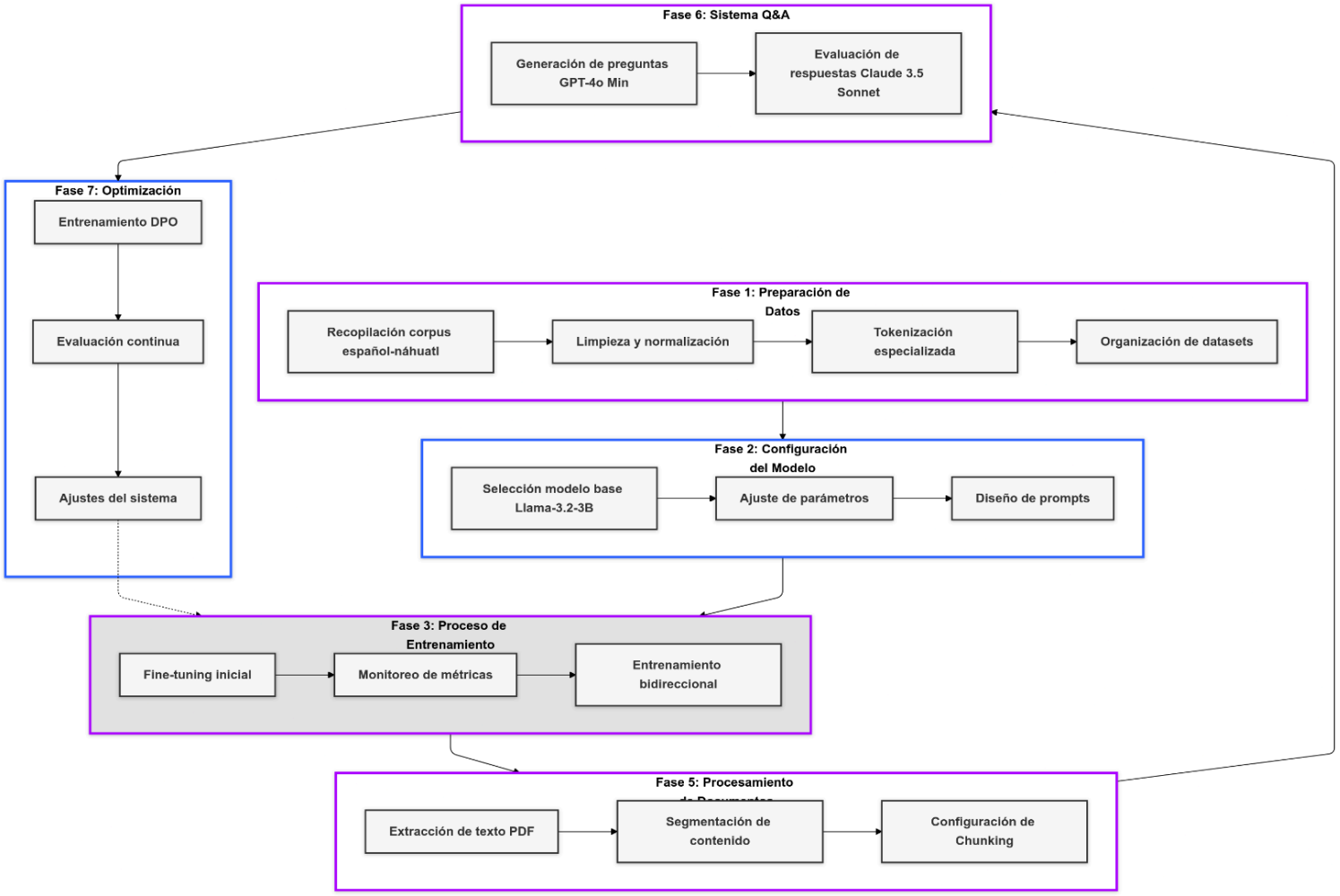
Metodología para AjakatlAI

Introducción

Este documento presenta una guía detallada para implementar un sistema de traducción automática entre español y náhuatl, combinado con capacidades de análisis de documentos administrativos. La metodología se ha diseñado para ser seguida de manera secuencial, donde cada paso construye sobre los anteriores para crear un sistema completo y funcional.



Fase 1: Preparación del Entorno de Trabajo

El primer paso consiste en establecer un entorno de desarrollo robusto en Google Colab. Comience accediendo a Google Colab y creando un nuevo notebook. Es fundamental montar Google Drive desde el inicio, ya que será el espacio donde almacenaremos todos

nuestros modelos, datos y resultados. Para esto, ejecute el comando de montaje en la primera celda del notebook y verifique que tiene acceso a su Drive.

Una vez montado el Drive, necesitará crear una estructura de directorios organizada. Cree carpetas separadas para modelos, datos de entrenamiento, resultados y documentos. Esta organización es crucial para mantener un flujo de trabajo eficiente y permitir un seguimiento adecuado del progreso.

La instalación de dependencias es el siguiente paso crítico. Deberá instalar todas las bibliotecas necesarias utilizando pip. Es importante mantener un registro de las versiones específicas de cada biblioteca para garantizar la reproducibilidad del proyecto. Las bibliotecas fundamentales incluyen Transformers de Hugging Face, Unsloth para optimización de memoria, y Langchain para el procesamiento de documentos.

Fase 2: Preparación de Datos

La preparación de datos comienza con la recopilación de un corpus paralelo español-náhuatl. Este corpus debe ser limpiado y normalizado cuidadosamente. El proceso de limpieza incluye la eliminación de caracteres especiales no deseados, la normalización de espacios y puntuación, y la verificación de la alineación correcta entre las oraciones en ambos idiomas.

Para el procesamiento de los datos, implemente un sistema de tokenización que sea sensible a las características específicas del náhuatl. El náhuatl tiene características morfológicas únicas que deben ser consideradas durante la tokenización. Utilice la biblioteca de Transformers para crear un tokenizador personalizado que maneje adecuadamente estas características.

Los datos deben organizarse en un formato que facilite el entrenamiento. Cree un conjunto de datos estructurado donde cada entrada contenga el texto fuente, la traducción objetivo y cualquier metadato relevante. Es recomendable dividir los datos en conjuntos de entrenamiento, validación y prueba con una proporción aproximada de 80-10-10.

Fase 3: Configuración del Modelo

La selección y configuración del modelo base es un paso crucial. Comenzamos con el modelo Llama-3.2-3B-Instruct, que debe ser cargado y configurado específicamente para nuestra tarea. La configuración incluye ajustar parámetros como la longitud máxima de secuencia, el tipo de datos (preferiblemente float16 para optimizar el uso de memoria), y la configuración del dispositivo para utilizar la GPU disponible.

Los prompts de entrenamiento deben diseñarse cuidadosamente para guiar al modelo en la tarea de traducción. Cree templates de prompts que incluyan instrucciones claras y ejemplos de contexto. Estos prompts deben seguir un formato consistente que el modelo pueda aprender a interpretar efectivamente.

Fase 4: Proceso de Entrenamiento

El entrenamiento del modelo se realiza en varias etapas. Comience con un proceso de fine-tuning inicial utilizando el SFTTrainer de la biblioteca Transformers. Configure los parámetros de entrenamiento cuidadosamente: la tasa de aprendizaje debe ser lo suficientemente pequeña para permitir un ajuste fino efectivo (típicamente alrededor de 2e-5), y el tamaño del batch debe ajustarse según la memoria disponible en la GPU.

Implemente un sistema de monitoreo durante el entrenamiento para seguir métricas clave como la pérdida de entrenamiento, la pérdida de validación y las métricas de traducción como BLEU score. Este monitoreo le permitirá detectar problemas temprano y ajustar los parámetros según sea necesario.

El entrenamiento debe realizarse en ambas direcciones: español a náhuatl y náhuatl a español. Cada dirección puede requerir ajustes específicos en los parámetros de entrenamiento debido a las diferentes características de los idiomas.

Fase 5: Procesamiento de Documentos

Para el procesamiento de documentos PDF, implemente un pipeline utilizando Langchain. El proceso comienza con la extracción de texto de los PDFs, seguido de una segmentación cuidadosa del contenido. La segmentación debe realizarse de manera que preserve el contexto y la coherencia del texto.

Configure el RecursiveCharacterTextSplitter con parámetros apropiados para su caso de uso. Un tamaño de chunk de 1000 caracteres con un solapamiento de 200 caracteres suele funcionar bien para la mayoría de los documentos administrativos. Asegúrese de que los separadores estén configurados para respetar la estructura natural del documento.

Fase 6: Sistema de Preguntas y Respuestas

El sistema de preguntas y respuestas se construye sobre la base del procesamiento de documentos. Utilice la API de OpenAI para generar preguntas relevantes sobre el contenido de los documentos. Las preguntas generadas deben ser diversas y cubrir diferentes aspectos del contenido.

Implemente un sistema de evaluación de respuestas que califique la calidad de las respuestas generadas. El sistema debe considerar factores como la precisión, la relevancia y la completitud de las respuestas. Utilice una combinación de métricas automáticas y evaluación humana para validar la calidad de las respuestas.

Fase 7: Optimización y Mejora Continua

La optimización del sistema es un proceso continuo que involucra varios aspectos. Implemente el entrenamiento DPO (Direct Preference Optimization) para mejorar la calidad de las respuestas generadas. Este proceso implica recopilar datos de preferencias de los usuarios y utilizarlos para ajustar el modelo.

Realice evaluaciones periódicas del rendimiento del sistema utilizando un conjunto diverso de métricas. Estas evaluaciones deben incluir tanto métricas automáticas como evaluaciones humanas. Utilice los resultados de estas evaluaciones para identificar áreas de mejora y ajustar el sistema según sea necesario.