



Universidad
Nacional
de Loja

FACULTAD DE LA ENERGÍA, LAS INDUSTRIAS Y
LOS RECURSOS NATURALES NO RENOVABLES

Planificación de los instrumentos para obtener información

Autor: Edy Francisco Jiménez Merino

Cargo: Estudiante de la carrera de Computación

Email: edy.jimenez@unl.edu.ec

Teléfono: 0939943013



1. Resumen

Esta entrevista se realiza en el contexto del desarrollo de mi trabajo de integración curricular(TIC), enfocado específicamente en el modelo de búsqueda de respuestas o Question Answering (QA) para responder a preguntas sobre el contenido extraído de la Universidad Nacional de Loja(UNL). Apunta a un sistema innovador de accesibilidad para estudiantes y personal con discapacidad visual de la UNL en la carrera de Computación y se proyecta como una base sólida para ser implementada en toda la comunidad educativa; se busca precisar características técnicas, requerimientos y proyecciones del modelo QA mediante una interacción directa en lenguaje natural con respecto a los usuarios que ocupen este modelo.

2. Instrumento 1: Encuesta

Detalles básicos	
Actores involucrados	<ul style="list-style-type: none">● Investigador Principal: Edy Francisco Jimenez Merino Estudiante de la carrera de Computación UNL
Muestra seleccionada	<i>Director del proyecto de Vinculación con la sociedad de la carrera de Computación de la Universidad Nacional de Loja, Ing.Óscar Cumbicus</i>
Fecha de ejecución	29 de Enero de 2024
Duración	15-30 minutos
Objetivos	<ul style="list-style-type: none">● Precisar requerimientos técnicos y funcionales para el módulo búsqueda de respuestas a preguntas, tales como: tipos de documentos de entrada, formatos de salida esperados, métricas de rendimiento, infraestructura disponible, etc.● Determinar formas de ejecución o ajuste del modelo mediante técnicas para evaluar su reproducción



Plantilla de la Entrevista	
Título	<i>Accesibilidad a información sobre documentos o tareas académicas mediante un modelo de búsqueda de respuestas.</i>
Presentación	<p>Estimado Ing. Oscar Cumbicus,</p> <p>La presente entrevista se realiza en el marco de mi trabajo de integración curricular sobre el módulo de respuesta a consultas, que busca la accesibilidad a documentos académicos para estudiantes en la carrera de computación de la UNL. Necesito precisar detalles técnicos y requerimientos. Sus respuestas, serán tratadas bajo total confidencialidad además que resultan valiosas para conceptualizar y desarrollar el TIC.</p> <p>Sus experiencias previas en relación con mi área de trabajo aportarán significativamente a la concepción y desarrollo de mi iniciativa.</p> <p>Agradezco sinceramente su tiempo y retroalimentación ética para contribuir a mi TIC.</p>

Listado de preguntas

Las preguntas de la entrevista se ven enmarcadas en el proyecto de vinculación de la carrera de computación llamado “Inclusión lectora de los estudiantes con discapacidad visual de la Universidad Nacional de Loja mediante innovación tecnológica”.

INFORMACIÓN GENERAL DEL PROYECTO

1. ¿En qué consiste este proyecto de vinculación?
2. ¿Qué problemática busca resolver y cuáles son los objetivos generales?

MÓDULOS/COMPONENTES

1. ¿Cuáles son los módulos o componentes principales en el proyecto y que se implementará en cada uno de ellos?

TÉCNICAS Y TECNOLOGÍAS

El enfoque de las siguientes preguntas es en obtener conocimiento sobre: herramientas, lenguajes y técnicas que se podrían implementar específicamente en el módulo de búsqueda de respuesta a preguntas de documentos o tareas académicas en la carrera de Computación de la UNL, esto en base al enfoque del proyecto de vinculación:

1. ¿Qué tipo de algoritmo sugiere utilizar: redes neuronales convolucionales(CNN) , transformers, etc?
2. ¿Se recomienda crear modelo desde cero o usar técnicas como el fine-tuning con modelos existentes ?
3. Para implementar la mejora mediante técnicas de Fine Tuning o Transfer Learning ¿qué recomienda usar como base: modelos state-of-the-art grandes (GPT-3) o medianos (DistilBERT)?
4. ¿Recomienda alguna plataforma en específico para desarrollar el modelo: Tensor Flow, PyTorch, Keras u otra, además el procesamiento para estas tareas se podría hacer de forma local o con algún proveedor de cloud , por ejemplo en google colab?
5. ¿Qué desafíos ve en cuanto a uso/aprendizaje de las tecnologías?¿

EVALUACIÓN

1. Actualmente¿qué métricas se usan para evaluar el rendimiento de un modelo de este tipo? y ¿cómo medirán el cumplimiento de objetivos?
2. Al usar este tipo de técnicas en modelos Transformers, ¿que limitantes suelen existir respecto a pregunta-respuesta de acuerdo a un contexto ?

SOSTENIBILIDAD

1. ¿Se ha considerado el uso de tecnologías Cloud más "verdes" o con compensación para reducir el consumo energético?

CIERRE

1. ¿Desea agregar algo más sobre expectativas del proyecto?

Transcripción de la Entrevista

Recordemos que las preguntas de la entrevista se ven enmarcadas en el proyecto de vinculación de la carrera de computación llamado “Inclusión lectora de los estudiantes con discapacidad visual de la Universidad Nacional de Loja mediante innovación tecnológica”.

INFORMACIÓN GENERAL DEL PROYECTO

1. ¿En qué consiste este proyecto de vinculación?

El proyecto de vinculación trata de desarrollar un sistema tecnológico que les permita a los estudiantes con discapacidad visual poder realizar la lectura de documentos sin ningún inconveniente, con la mayor facilidad posible. En la actualidad no se cuenta con una herramienta de este tipo en la universidad y los estudiantes como alternativa, lo que hacen es pagar software con licencia, y muchas de las veces no está al alcance de su presupuesto, así que por la necesidad muchas de las veces los estudiantes optan por realizar prácticas no éticas como el crackeo de estos sistemas, entonces, lo que pretende el proyecto es darles una herramienta para que puedan realizar las lecturas de sus documentos dentro de la UNL.

2. ¿Qué problemática busca resolver y cuáles son los objetivos generales?

Los estudiantes con y sin discapacidad visual no cuentan con una herramienta para la lectura de documentos en la universidad, el objetivo general es crear este sistema e integrarlo en el laboratorio inteligente para que lo estudiantes lo usen, entonces el primer objetivo es el levantamiento de requerimientos, el segundo es el desarrollo del sistema que se pretende hacer con estos proyectos de integración curricular y finalmente la puesta en producción de este sistema para los estudiantes.

MÓDULOS/COMPONENTES

1. ¿Cuáles son los módulos o componentes principales en el proyecto y qué se implementará en cada uno de ellos?

Existen algunos módulos dentro de este sistema:

1. Transformar documentos con la extensión pdf, docs e imágenes a texto para que pueda ser traducido a voz
2. Realizar preguntas sobre el contenido extraído de documentos, donde los estudiante mediante los requerimientos solicitan hacer preguntas sobre los documentos.

Existen otros módulos más como: convertir o interpretar tablas e imágenes, además de integrar componentes de hardware como Jetson Nano.

TÉCNICAS Y TECNOLOGÍAS

- El enfoque de las siguientes preguntas es en obtener conocimiento sobre: herramientas, lenguajes y técnicas que se podrían implementar específicamente en el módulo de búsqueda de respuesta a preguntas de documentos o tareas académicas en la carrera de Computación de la UNL, esto en base al enfoque del proyecto de vinculación:

1. ¿Qué tipo de algoritmo sugiere utilizar: redes neuronales convolucionales(CNN) , transformers, etc?

Una de las mejores arquitecturas, algoritmos, etc que funcionan en este caso son las redes Transformers a través de los distintos modelos de lenguaje han demostrado en varios estudios que son muy buenas en el tema de preguntas respuestas acerca de un contexto, una recomendación es usar redes Transformers con cualquiera de los modelos de PLN como BERT, Bard, MT5, T5 y algunas distribuciones que han sido adaptadas, se toma en cuenta que esos modelos fueron creados en inglés, la mayoría han sido adaptados para trabajar en el idioma español, como DistilBert , BETO y otros .

2. ¿Se recomienda crear modelo desde cero o usar técnicas como el fine-tuning con modelos existentes ?

Aprovechar el conocimiento de modelos pre-entrenados como los mencionados , realizar técnicas de Fine Tuning o transfer Learning para aprovechar todos ese conocimiento que tienen esos modelos y ajustar a la tarea que nosotros necesitamos, en este caso usar un modelo pre entrenado y realizar un fine Tuning y hacer que este modelo de procesamiento de Lenguaje natural se adapte a nuestro requerimiento, que en este caso es responder a preguntas sobre tareas o texto académicos.

3. Para implementar la mejora mediante técnicas de Fine Tuning o Transfer Learning ¿qué recomienda usar como base: modelos state-of-the-art grandes (GPT-3) o medianos (DistilBERT)?

Por el tema del contexto en que se está haciendo el proyecto se recomienda usar modelos medianos o pequeños como DistilBERT, Electra, BertIN un lenguaje más pequeño que BETO, porque la infraestructura no nos va a permitir usar modelo grandes, la recomendación es usar modelo mediano que ya hayan sido probado en otras tareas en respuestas preguntas de otros contextos, por ejemplo que responda bien a preguntas sobre códigos, o leyes, o que ya haya sido probado en textos de medicina o similar, pero que funcione correctamente, la recomendación es usar modelos de lenguaje medianos.

4. ¿Recomienda alguna plataforma en específico para desarrollar el modelo: Tensor Flow, PyTorch, Keras u otra, además el procesamiento que se utiliza para estas tareas se podría hacer de forma local o con algún proveedor de cloud , por ejemplo en google colab?

Bueno, se ha trabajado dentro de la carrera en su gran mayoría con el framework pytorch, es un framework que se adapta muy bien al tema de procesamiento de lenguaje natural y a modelos como los que los que estábamos hablando, ya tiene algunas métricas incluso establecidas dentro

del mismo del mismo framework , que nos permite medir el rendimiento de cada uno de los modelos Entonces yo recomendaría por ejemplo pytorch sin descartar también otras framework Como por ejemplo tensor Flow Pero principalmente sería pytorch el lenguaje de programación que yo recomendaría sería python por ser un lenguaje que en la actualidad nos da grandes prestaciones dentro del Machine learning, como ya le habemos dicho anteriormente la red neuronal Transformers los modelos sec to sec , que son modelos que tienen encoder y decoder es decir codifican y decodifican las entradas de acuerdo a lo que nosotros necesitemos y si no estoy mal otro tipo de herramientas o plataformas que podríamos utilizar sería Google colab si es que nos da la capacidad en su versión libre o si no utilizar el hpc de cedia con el que cuenta la universidad para poder entrenar este tipo de modelos, de esas tecnologías y las GPU gpus que justamente es el tema del supercomputador que nosotros podemos utilizar tanto en colab como hpc de cedia.

5. ¿Qué desafíos ve en cuanto a uso/aprendizaje de las tecnologías?

Los desafíos más importantes es justamente entender cómo funcionan estos modelos de lenguaje, qué hiper-parámetros debemos modificar cuándo deberíamos parar el entrenamiento yo creo que esas son algunos desafíos que se Se podrían presentar, tratar de que las versiones sean compatibles entre sí, todos estos temas de de cómo funciona justamente la red neuronal Transformer cómo funcionan las redes sec to sec, cómo se codifica y decodifica cada una de estas entradas y salidas y cómo se realiza efectivamente un ajuste fino de estos modelos creo que eso es una de los desafíos que se tendría dentro del desarrollo del modelo y sobre todo el desafío más grande es el hardware, ya habíamos dicho que podemos solucionar con el hpc de cedia, que en este momento la universidad lo puede utilizar.

EVALUACIÓN

1. Actualmente ¿qué métricas se usan para evaluar el rendimiento de un modelo de este tipo? y ¿cómo medirán el cumplimiento de objetivos?

Bueno, ¿cómo mediremos?, una de las cosas importantes es , poder determinar cuán efectivo o cuál es el porcentaje global de acierto del modelo, ¿cómo se mide eso?, se mide a través de las entradas que nosotros le vamos a dar una vez entrenado ya el modelo, entonces a nuestro modelo que ya hemos hecho el Fine Tuning o Transfer learning; este dar un conjunto de entradas y medir cuántas de esas entradas han sido en este caso respondidas correctamente o, en este caso dado un conjunto de preguntas cuáles de esas preguntas en base a su contexto ha sido respondidas correctamente entonces una de las métricas principales que debemos utilizar dentro de la evaluación es el porcentaje global de acierto, ahora ¿cuándo ese porcentaje global de acierto es el correcto? o es el que nos va a

determinar el éxito, cuando el porcentaje global de acierto de preguntas-respuestas esté entre un 80% para arriba, es decir mayor al 80% si ese es el caso se considera exitoso ya que se estaría dando una buena respuesta de todo el contexto que se tiene de la pregunta que se está realizando, tomando en cuenta que nosotros no vamos a crear un modelo desde cero, vamos a aprovechar el transfer learning o perdón el conocimiento a través de transfer learning y de Fine tuning de esos modelos, entonces tener un acierto mayor al 80% se consideraría como un éxito para el proyecto.

2. Al usar este tipo de técnicas en modelos Transformers, ¿que limitantes suelen existir respecto a pregunta-respuesta de acuerdo a un contexto ?

Una de las limitantes que tienen este tipo de modelos es el tamaño del texto justamente, estos modelos procesan de entre 64 tokens a 1024 si no estoy mal o a 512, eso quiere decir que los contextos que vamos a enviar deberían ser de ese tamaño; entonces si tenemos contextos más grandes por ejemplo:unas cuatro hojas, unas cinco hojas, va a ser muy difícil que nosotros mandemos todo ese texto y se pueda responder sobre ese texto. Entonces es una de las limitantes ahora ¿cómo se podría solucionar esas limitantes?, enviando sección por sección, haciendo preguntas de una sección o escogiendo la sección sobre la cual se quiere hacer la pregunta,es decir, si es que ya se haría la aplicación dentro de aplicación un combobox donde escojas tú la sección sobre la cual quieras hacer ,entonces una de las limitantes justamente es eso,e incluso los modelos grandes como gpt existe esa limitante, si ustedes se dan cuenta cuando hacen una pregunta a chat se queda y ya no sigue generando, no sé si se han dado cuenta, es porque ya no puede generar más texto justamente por esta limitante, de hecho la más grande que va a existir es el tamaño del texto.Pero eso, lo podríamos solucionar luego que se haría la interfaz ,es decir, ya en interfaz.Primeramente crear el modelo que responda bien a las preguntas y respuestas dado un contexto y luego ya en una siguiente etapa cuando le hagamos interfaz al modelo que ese modelo escoja por ejemplo de la página uno de la página dos de la página tres o de la página cuatro para hacer preguntas. Entonces yo creo que este modelo debería funcionar con un mínimo de tres páginas a cinco páginas para que podamos hacer las consultas respectivas.

SOSTENIBILIDAD

1. ¿Se ha considerado el uso de tecnologías Cloud más "verdes" o con compensación para reducir el consumo energético?

Bueno el primer punto es considerar modelos entrenados que es una de las cosas que nos va a reducir, otra es justamente el tamaño de los textos como tú estabas diciendo, que sean pequeños no grandes, porque si hacemos para textos grandes igual vamos a consumir mucha energía y mucho procesamiento en los computadores, otra de las cosas es que podamos utilizar herramientas libres, que no está dentro del lado del tema de clima pero está dentro del lado de que nosotros vamos a ofrecer a la universidad algo que no va a tener que ser pagado



por los estudiantes. Usar GPUs porque hacerlo con un procesador normal no se va a lograr y también podemos utilizar los convenios que tiene la universidad que son el hpc de cedia para poder procesar este tipo de técnicas.

CIERRE

1. ¿Desea agregar algo más sobre expectativas del proyecto?

La expectativa que se tiene dentro de este módulo es que se pueda obtener respuestas considerablemente buenas, esto quiere decir que, no exactamente en un 100% buenas, pero sí entre un más de un 80% que tengan coherencia con lo que se está diciendo; y tomar en consideración que todo todos los modelos de procesamiento de lenguaje natural los modelos de lenguaje van a tener alucinaciones, una alucinación es que ellos pueden inventar cualquier cosa, si yo le pregunto de un tema específico y si no tiene una respuesta se puede inventar, empezar a escribir cosas que no son correctas y que son puntos que hay que analizar y no por eso va a estar mal el modelo, sino que los modelos ya vienen con eso, porque tienen muy poco texto o tienen muy pocos token con qué responder porque no tienen información sobre el tema, no hay una forma en que se les diga no contestes esto, estos modelos están entrenados para contestar cualquier cosa, eso es un punto muy importante que tenemos que tomar en consideración que la gran mayoría va a empezar a alucinar.

Atentamente,

Estudiante: Edy Francisco Jiménez Merino

C.I: 1950170389