

Detección de entidades en informes clínicos mediante el uso de LLMs

1. INTRODUCCIÓN

El reconocimiento de entidades nombradas (**NER**) es una tarea clave dentro del procesamiento del lenguaje natural (**NLP**), cuyo objetivo es identificar y clasificar menciones de entidades específicas en un texto, como nombres de personas, organizaciones, ubicaciones, fechas y otros conceptos relevantes. Por otro lado, NLP es un campo de la inteligencia artificial que se centra en la interacción entre las computadoras y el lenguaje humano, permitiendo a las máquinas procesar y analizar grandes cantidades de texto de manera automática. NER, en particular, se utiliza en aplicaciones como la extracción de información y el análisis de contenido, siendo esencial para estructurar el texto en datos más manejables y comprensibles.

Los modelos de lenguaje de gran escala (**LLMs**), como GPT-4 o BERT, han revolucionado el panorama del NLP al ser capaces de comprender y generar lenguaje natural con un alto grado de precisión. Estos modelos, entrenados con enormes volúmenes de datos textuales, capturan complejos patrones lingüísticos y contextuales. Una técnica emergente en este contexto es el **prompt engineering**, que consiste en diseñar de manera efectiva las instrucciones o entradas proporcionadas a los LLMs para obtener resultados precisos y relevantes. Con el ajuste adecuado de los **prompts**, los LLMs pueden resolver una amplia gama de tareas sin necesidad de entrenamientos adicionales, lo que los convierte en herramientas flexibles y potentes.

La **integración de LLMs en tareas de NER** ofrece nuevas oportunidades y desafíos. Por un lado, los LLMs permiten mejorar la precisión y generalización de los sistemas de NER, ya que comprenden mejor el contexto en el que aparecen las entidades. Sin embargo, para aprovechar al máximo su potencial, es necesario desarrollar estrategias de *prompt engineering* que optimicen las respuestas del modelo. Esto introduce el problema de cómo diseñar *prompts* eficientes para lograr que un LLM identifique y clasifique entidades de manera precisa en textos de distintos dominios. Resolver este desafío es clave para avanzar en el uso práctico de LLMs en NER y mejorar las aplicaciones que dependen de esta tecnología.

En el desafío planteado por GMV se propone el uso de LLMs junto con técnicas de *prompt engineering* para llevar a cabo tareas de reconocimiento de entidades nombradas (NER) en textos clínicos, con el objetivo de identificar y clasificar entidades relevantes dentro de dichos documentos.

2. EL PROBLEMA

El problema planteado consta de dos objetivos principales, que se dividen en varias fases bien definidas:

1. **Estado del arte:** El primer paso será realizar un estudio exhaustivo del estado del arte y las principales técnicas detrás de la detección de entidades mediante modelos de lenguaje (LLMs) y *prompt engineering*. En esta fase, se investigan las metodologías más avanzadas y las innovaciones recientes en el procesamiento del lenguaje natural (NLP) aplicado a textos médicos. Se espera también un análisis, desde el punto de vista matemático, de los fundamentos teóricos que hay detrás de estas técnicas.
2. **Generación de textos clínicos:** Se trata de crear textos clínicos que sean adecuados para lograr un buen rendimiento en el reconocimiento de entidades. Estos textos pueden ser generados utilizando herramientas como ChatGPT u otras alternativas que se consideren apropiadas, como por ejemplo data sets públicos. En esta fase, se espera que los informes clínicos contemplen diversas casuísticas, incluyendo diferentes formatos de fechas, nombres, idiomas (incluso varios idiomas en el mismo informe) y otros elementos típicos del lenguaje médico. La diversidad y realismo de los textos serán clave para el éxito de las fases posteriores.
3. **Implementación de técnicas de *prompt engineering*:** El siguiente paso implica aplicar técnicas de *prompt engineering* a modelos de lenguaje de gran escala (LLMs) para realizar el

reconocimiento de entidades nombradas (NER) en los textos clínicos generados. El objetivo es diseñar *prompts* que permitan al modelo identificar de manera efectiva entidades relevantes dentro de los textos, como:

- Nombres de pacientes
- Edad, sexo, altura y peso de los pacientes
- Diagnósticos y procedimientos médicos
- Enfermedades
- Fechas (de ingreso/alta)
- Fármacos y medidas (unidades y valor)

En la sección 4. Referencias, se proporcionará documentación útil que puede apoyar esta fase.

4. **Validación de los resultados:** Finalmente, los resultados obtenidos en la fase de NER deberán ser validados. Para ello, se propone la obtención de las métricas: Accuracy, Recall, Precision y F1. Se piden estas métricas tanto para el conjunto entero de entidades como para cada tipo de entidad. Se valorará positivamente la automatización de este procedimiento en la medida de lo posible. La validación adecuada permitirá extraer conclusiones sólidas sobre el rendimiento del sistema de NER en este contexto clínico.

Un ejemplo del paso 1 usando ChatGPT sería el siguiente:

Genérame un ejemplo de un informe clínico en español



A continuación, te presento un ejemplo de un informe clínico en español, considerando un formato típico que incluye antecedentes, evaluación y diagnóstico:

INFORME CLÍNICO

Fecha: 10 de octubre de 2024

Paciente: Juan Pérez García

Fecha de nacimiento: 15 de marzo de 1975

Número de historia clínica: 123456789

Motivo de consulta:

El paciente acude a consulta por presentar dolor torácico persistente desde hace tres días, acompañado de disnea leve y fatiga al realizar esfuerzos moderados.

Antecedentes personales:

- Hipertensión arterial diagnosticada en 2015, en tratamiento con enalapril 10 mg/día.
- Dislipidemia desde 2018, en tratamiento con atorvastatina 20 mg/día.
- Tabaquismo: 10 cigarrillos/día desde hace 20 años.
- No alergias conocidas a medicamentos.

Exploración física:

- Tensión arterial: 150/95 mmHg
- Frecuencia cardíaca: 88 lpm
- Frecuencia respiratoria: 18 rpm
- Saturación de oxígeno: 96%
- Auscultación cardíaca: Rítmica, sin soplos.
- Auscultación pulmonar: Normal, sin ruidos adventicios.
- Abdomen: Blando, depresible, sin masas ni dolor a la palpación.

Pruebas complementarias:

1. **Electrocardiograma (ECG):** Sin alteraciones significativas. Ritmo sinusal a 85 lpm.
2. **Radiografía de tórax:** Sin evidencia de consolidaciones, tamaño cardíaco normal.

3. Análisis de sangre:

Hemoglobina: 13.8 g/dL

Glucosa: 95 mg/dL

Colesterol total: 220 mg/dL

LDL: 140 mg/dL

HDL: 50 mg/dL

Triglicéridos: 180 mg/dL

Troponina I: 0.02 ng/mL (dentro de rangos normales).

Diagnóstico provisional:

Dolor torácico de origen no cardíaco, posiblemente relacionado con un cuadro de ansiedad.

Plan de tratamiento:

1. Continuar con el tratamiento antihipertensivo y para la dislipidemia.
2. Se prescribe ibuprofeno 600 mg cada 8 horas por 5 días para el manejo del dolor.
3. Realizar seguimiento con prueba de esfuerzo para descartar patología coronaria si los síntomas persisten.
4. Se recomienda dejar de fumar y adoptar un estilo de vida más saludable (dieta y ejercicio regular).
5. Nueva consulta en dos semanas para reevaluación y posibles ajustes en el tratamiento.

Médico responsable:

Dra. María López Fernández

Número de colegiado: 7654321

3. PLAN DE TRABAJO Y OBJETIVOS DE APRENDIZAJE

Una vez que se hayan creado los equipos de trabajo, el procedimiento es el siguiente:

1. Al comenzar el evento, se realizará una reunión inicial con los equipos inscritos en el problema, orientándoles sobre el camino a seguir.
2. A mitad del evento se contactará con los equipos para revisar los avances y resolver las posibles dudas.
3. Al finalizar, se sintetizan y preparan los resultados antes de enviar la solución.

Los objetivos de aprendizaje son:

- Familiarizarse con los desafíos que enfrentan las organizaciones en el procesamiento de textos clínicos y cómo las técnicas de NLP pueden ayudar a resolverlos.
- Conocer y aplicar modelos de lenguaje de gran escala (LLMs) y técnicas de *prompt engineering* para el reconocimiento de entidades nombradas (NER).
- Plantear y formular una solución a un problema basado en la generación y procesamiento de informes clínicos.
- Realizar el análisis, validación, discusión y presentación de los resultados obtenidos.

4. REFERENCIAS

- Named Entity Recognition — Clinical Data Extraction, from <https://vishal-aiml164.medium.com/named-entity-recognition-clinical-data-extraction-9b089d91b27b>
- Spacy-llm: Integrating LLMs into structured NLP pipelines, from https://medium.com/@pankaj_pandey/spacy-llm-integrating-llms-into-structured-nlp-pipelines-7134dd05ebc2
- LLMs in Biomedical: A Study on Named Entity Recognition, from <https://arxiv.org/html/2404.07376v2>
- Amazon Comprehend Medical (Developer guide), from <https://docs.aws.amazon.com/comprehend-medical/latest/dev/textanalysis-entitiesv2.html>