Large Language Models (LLMs) on Tabular Data: Prediction, Generation, and Understanding - A Survey

Los modelos de lenguaje grandes (LLM) son modelos de aprendizaje profundo entrenados con una gran cantidad de datos, lo que los dota de capacidades versátiles de resolución de problemas que se extienden mucho más allá del ámbito de las tareas de procesamiento del lenguaje natural (PLN). Investigaciones recientes han revelado capacidades emergentes de los LLM, como un mejor desempeño en tareas con pocas indicaciones. El desempeño notable de los LLM ha despertado interés tanto en el ámbito académico como en la industria, lo que ha generado la creencia de que podrían servir como base para la Inteligencia Artificial General (AGI) de esta era. Un ejemplo notable es ChatGPT, diseñado específicamente para participar en conversaciones humanas, que demuestra la capacidad de comprender y generar texto en lenguaje humano.

Antes de los LLM, los investigadores han estado investigando formas de integrar datos tabulares con redes neuronales para tareas de PLN y gestión de datos. Hoy, los investigadores están interesados ​​en investigar las capacidades de los LLM al trabajar con datos tabulares para diversas tareas, como predicción, comprensión de tablas, razonamiento cuantitativo y generación de datos. Los datos tabulares son uno de los formatos de datos más omnipresentes y esenciales en el aprendizaje automático (ML), con aplicaciones generalizadas en diversos dominios, como finanzas, medicina, negocios, agricultura, educación y otros sectores que dependen en gran medida de bases de datos relacionales .

Optimized Feature Generation for Tabular Data via LLMs with Decision Tree Reasoning

Generar características de columna adecuadas, incluso con conocimiento del dominio, puede ser desafiante y costoso. Por ejemplo, la validación manual para identificar características útiles es inviable debido a la cantidad exponencial de combinaciones posibles para explorar. Para abordar este problema, los métodos de ingeniería de características existentes utilizan esquemas de filtrado adicionales para evaluar y seleccionar características útiles de forma automática. Si bien estos enfoques reducen el esfuerzo manual y mejoran la calidad de las características, aún presentan varios desafíos. En primer lugar, los profesionales a menudo dependen de espacios de búsqueda definidos manualmente para generar características candidatas debido a la ambigüedad inherente de lo que constituye características informativas particularmente a medida que el número de características y la complejidad del espacio de búsqueda crecen. Además, descuidan los diseños experimentales más efectivos, confiando únicamente en las puntuaciones de validación para seleccionar buenas características, a pesar del valor de los datos de experimentos anteriores para mejorar la selección. Motivados por esto, proponemos abordar este problema desde una perspectiva novedosa: la optimización para descubrir reglas de generación efectivas, aprovechando la comprensión del lenguaje y las capacidades de razonamiento de los modelos de lenguaje grandes (LLM). Investigaciones recientes han demostrado que los LLM pueden optimizar varios problemas no diferenciables utilizando indicaciones que describen la tarea de optimización en lenguaje natural. Esto sugiere el potencial de los LLM para generar automáticamente y refinar iterativamente generadores de características sin la necesidad de especificar manualmente el espacio de reglas. Por ejemplo, las capacidades de razonamiento de los LLM permiten incorporar retroalimentación sobre sus resultados anteriores en el proceso de refinamiento iterativo. Además, los contextos lingüísticos, como los nombres de las columnas (por ejemplo, “Género” y “Edad”) y los valores categóricos (por ejemplo, “Mujer” y “Hombre”), podrían integrarse naturalmente en la optimización, lo que es difícil, si no imposible, con los métodos convencionales.

Exploring the Impact of Table-to-Text Methods on Augmenting LLM-based Question Answering with Domain Hybrid Data

Introducción

La mejora del rendimiento de los modelos de lenguaje grandes (LLM) en la formulación de preguntas y respuestas (QA) específicas de un dominio ha sido un foco de investigación, empleando predominantemente dos enfoques: el ajuste fino específico del dominio (DSFT), que implica el entrenamiento de los LLM en el corpus específico del dominio, y la generación aumentada por recuperación (RAG), que utiliza un corpus específico del dominio como base de conocimiento externa. Estos enfoques, que aprovechan las fortalezas inherentes del procesamiento de texto de los LLM, se han adoptado ampliamente en escenarios de solo texto, lo que ha producido mejoras significativas.

Sin embargo, los datos del mundo real en muchos dominios suelen existir en un formato híbrido, que comprende no solo texto sino también volúmenes sustanciales de tablas semiestructuradas, como se observa, por ejemplo, en la literatura científica, los informes médicos y en nuestro caso específico en los anuarios estadísticos . Estas tablas aparecen con frecuencia junto con el texto dentro del mismo documento, lo que proporciona información semánticamente suplementaria o complementaria crucial para una comprensión integral del contenido. Al explorar el potencial de aprovechar los datos híbridos para mejorar el rendimiento de los LLM, es crucial integrar eficazmente estos datos, asegurando la coexistencia de texto y tablas. Los métodos actuales para manejar la heterogeneidad de textos y tablas tienen inconvenientes significativos:

1. Aplanar directamente las tablas mediante la concatenación de celdas fila por fila que no solo da como resultado la pérdida de información estructural incorporada en la tabla original, sino que también corta los vínculos informativos al asignar texto y tablas a diferentes espacios vectoriales por separado y luego integrarlos, no solo aumenta la complejidad, sino que también altera la conexión semántica entre los dos tipos de datos.
2. Una solución prometedora es la generación de tablas a texto que tiene como objetivo generar declaraciones en lenguaje natural que describen fielmente la información en la tabla proporcionada. A través de esto, podemos transformar datos híbridos en una representación unificada en lenguaje natural que es más adecuada para su uso por parte de los LLM, al mismo tiempo que preserva la información importante de las tablas y las conexiones semánticas entre los datos.

Aunque la generación de tablas a texto ha sido ampliamente estudiada por la comunidad de NLP, actualmente no existe un análisis comparativo sobre cómo los corpus generados por diferentes métodos de tabla a texto afectan el desempeño de los sistemas de control de calidad específicos del dominio.

TableLLM: Enabling Tabular Data Manipulation by LLMs in Real Office Usage Scenarios

Una cantidad sustancial de datos se estructura rutinariamente en formatos tabulares, un formato ampliamente adoptado en varias industrias para diferentes propósitos.

…

Si bien los datos tabulares son omnipresentes, las tareas específicas relacionadas con las tablas pueden ser laboriosas, propensas a errores y requerir habilidades especializadas. La automatización de estas tareas ofrece beneficios significativos tanto para los sectores académicos como industriales, atrayendo un interés considerable. Los métodos convencionales para procesar datos tabulares se enfocan predominantemente en adaptar arquitecturas de modelos de lenguaje, incorporando elementos como incrustaciones de posición, mecanismos de atención y objetivos de aprendizaje para codificar los atributos estructurales inherentes de los datos tabulares. Sin embargo, se ha producido un cambio de paradigma con el surgimiento de grandes modelos de lenguaje (LLM) como GPT-4, GPT-3.5 y PaLM2. Las investigaciones recientes enfatizan la creación de indicaciones precisas que integren información parcial crucial a partir de datos tabulares proporcionados y el aprovechamiento de lenguajes de programación externos como SQL y Python.

Este enfoque facilita un proceso de inferencia de cadena de pensamiento (COT) paso a paso mediante LLM de código cerrado. La disponibilidad de LLM de código abierto, ejemplificada por Llama [34], permite el ajuste fino de los modelos para el procesamiento de datos tabulares, como lo demuestra TableLlama. Numerosos estudios previos se han centrado en mejorar las capacidades de razonamiento de un modelo para responder preguntas de tabla (tableQA). Más allá del control de calidad de las tablas, algunos de estos esfuerzos también han abordado diversas tareas relacionadas con las tablas, incluida la verificación de hechos, la anotación de tipos de columnas, la correspondencia de tablas, la ampliación de esquemas y más. Sin embargo, muchos de estos esfuerzos, si bien son valiosos en un contexto académico, pueden no reflejar el realismo completo de las personas que trabajan con datos tabulares en escenarios de oficina del mundo real. Para capturar información auténtica de los usuarios de la oficina, realizamos un extenso estudio de usuarios utilizando un cuestionario centrado en varias tareas relacionadas con las tablas.

En particular, las tablas en formatos Excel/CSV y Word/PDF, así como las tablas largas, surgieron como las opciones preferidas entre los participantes. Desafíos. En comparación con las tareas de tablas centradas en lo académico, el uso real de datos tabulares en la oficina presenta dos desafíos principales. (1) Operaciones diversas: las tareas preferidas por los usuarios involucran una amplia gama de operaciones, incluidas consultas, actualizaciones, fusiones y gráficos, que van más allá de las operaciones de consulta en el control de calidad de tablas. (2) Enfoques de procesamiento únicos para diferentes formatos: los documentos Word/PDF a menudo contienen datos textuales contextuales junto con información tabular, lo que permite consultas híbridas. Las hojas de cálculo Excel/CSV, por otro lado, contienen tablas regulares y largas, lo que permite operaciones más complejas como la actualización y la fusión. Si bien los trabajos existentes se centran en aprovechar la capacidad de los LLM para obtener respuestas directamente de sus parámetros internos, particularmente adecuados para datos tabulares integrados en documentos, o se especializan en escribir y ejecutar código para obtener respuestas de datos tabulares de hojas de cálculo, cada uno tiene limitaciones. El primero tiene dificultades con tablas largas y operaciones diversas en hojas de cálculo, mientras que el segundo no logra manejar consultas híbridas que involucran tanto datos de texto como tabulares. En resumen, los trabajos existentes aún no han abordado de manera efectiva ambos tipos de datos tabulares simultáneamente, cumpliendo con los requisitos del uso real en la oficina.

TAT-LLM: A Specialized Language Model for Discrete Reasoning over Tabular and Textual Data

Los documentos que contienen tanto tablas como texto, por ejemplo, presentaciones ante la SEC, artículos académicos e informes médicos, constituyen una categoría de contenido muy frecuente en el mundo real. A menudo presentan datos numéricos extensos tanto en el contenido tabular como en el textual, lo que requiere capacidades de razonamiento discreto para que las máquinas los comprendan. Investigaciones recientes investigan la comprensión inteligente de dichos documentos a través de tareas de preguntas y respuestas (QA). El modelo, provisto de una tabla y el texto relevante como contexto necesita realizar varios tipos de razonamiento discreto, como cálculos aritméticos, hacer comparaciones y contar, para responder la pregunta. Para realizar el control de calidad sobre datos textuales y tabulares híbridos, un enfoque sencillo implica tomar la tabla, el texto y la pregunta como entrada y generar la respuesta directamente. Este enfoque puede ser ineficaz debido al complejo proceso de razonamiento involucrado. Para abordar este problema, algunos trabajos descomponen la tarea en múltiples pasos, produciendo resultados intermedios que sirven como referencias para la respuesta final. Estos enfoques de múltiples pasos generalmente diseñan módulos distintos en cada paso y, a menudo, optimizan estos módulos simultáneamente a través del aprendizaje de múltiples tareas. Hasta la fecha, no ha habido consenso sobre cómo descomponer el proceso de respuesta en la literatura existente.

Recientemente, los modelos de lenguaje grandes (LLM) como GPT-4 y FLAN han mostrado fuertes capacidades de razonamiento de múltiples pasos con instrucciones adecuadas como la cadena de pensamiento (CoT) y de menor a mayor por lo tanto, consideramos aprovechar este asombroso poder de los LLM para un mejor razonamiento discreto sobre datos híbridos tabulares y textuales. Para lograrlo, primero identificamos tres pasos clave en el proceso de control de calidad tabular y textual de métodos de varios pasos anteriores, y abstraemos una secuencia de pasos, Específicamente, 1) el Extractor identifica la información o evidencia relevante para la pregunta a partir del contexto dado; 2) el Razonador genera una ecuación matemática o una regla lógica con la información obtenida; y 3) el Ejecutor deriva la respuesta final al ejecutar la ecuación matemática o la regla lógica con la información asociada. Los tres pasos enfatizan diferentes capacidades del modelo de control de calidad tabular y textual: comprender la pregunta y el contexto, inferir la lógica para responder la pregunta y calcular la respuesta con precisión. Estos pasos producen una secuencia de resultados intermedios, lo que significa que podemos modelar y mejorar específicamente uno (o más) de ellos dado un escenario de aplicación específico. Siguiendo el proceso de diseño por pasos, inicialmente diseñamos una instrucción de tarea y la validamos en GPT 4 (OpenAI, 2023), logrando resultados prometedores en múltiples puntos de referencia. Sin embargo, utilizar un LLM en línea presenta desafíos en términos de costo, latencia y riesgo de seguridad de datos. Por el contrario, ajustar un modelo de lenguaje más pequeño, específicamente para problemas de palabras matemáticas, ha demostrado ser bastante atractivo. Luego, estamos motivados para explorar la especialización de modelos de lenguaje más pequeños para abordar este desafío siguiendo el proceso de diseño por pasos.

Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study4

1 INTRODUCCIÓN

Los datos estructurados consisten en bloques de texto sin formato organizados por estructuras predefinidas para comprimir información recurrente. Hace que los datos sean más manejables y facilita el análisis y procesamiento de datos por parte de las máquinas. La tabla es uno de esos tipos de datos estructurados con muchas aplicaciones, como la respuesta a preguntas basada en tablas (TQA), la verificación de hechos basada en tablas (TFV), la conversión de tabla a texto y la clasificación de tipo de columna y relación. La adopción de datos estructurados ha contribuido significativamente al avance de la recuperación de información y la extracción de conocimiento en la minería web y el análisis de contenido.

…

Sin embargo, el trabajo previo no ha proporcionado estudios exhaustivos que examinen si los LLM pueden realmente comprender datos tabulares o han dado una discusión detallada de hasta qué punto los LLM ya han logrado capacidades de comprensión estructural. Además, a pesar del notable éxito de los LLM en el manejo de lenguajes naturales, su aplicación a la modalidad de datos tabulares presenta desafíos únicos: dado que las diferentes tablas definen la estructura y las características de maneras distintas y a menudo carecen de una transformación directa en texto secuencial (serialización de tablas). Con base en nuestra encuesta, creemos que el proceso de serialización de tablas, junto con el contexto y consultas correspondientes, es altamente flexible. Es decir, hay una falta de consenso fundamentado o investigación integral sobre lo que constituye un diseño de entrada exhaustivo o de sentido común para LLM en tareas tabulares.

El complejo panorama de diseño de entrada variado complica aún más los desafíos

que enfrentan los investigadores y desarrolladores en este campo. Por lo tanto, en este documento, nuestro objetivo es abordar la pregunta: ¿Qué diseños y opciones de entrada son los más efectivos para permitir que los LLM comprendan las tablas?

**Contexto Histórico/Social**

El procesamiento de datos ha evolucionado significativamente en las últimas décadas, impulsado por el crecimiento exponencial de la información disponible y la necesidad de análisis precisos en un mundo cada vez más complejo, donde la capacidad de extraer, analizar y utilizar datos de manera efectiva se ha convertido en un pilar para la toma de decisiones en múltiples disciplinas. En este contexto, los anuarios estadísticos juegan un papel crucial al proporcionar datos sistemáticos sobre diversos aspectos económicos, sociales y demográficos, son un compendio valioso que reflejan la evolución y el estado actual. En Cuba, la ***Oficina Nacional de Estadística e Información (ONEI)*** es responsable de compilar y publicar estos anuarios, que son fundamentales para la planificación económica y la formulación de políticas públicas, como un recurso para investigadores y académicos y también son esenciales para los responsables políticos que buscan formular estrategias basadas en evidencia.

Históricamente, Cuba ha enfrentado desafíos económicos significativos que han influido en su desarrollo social y político. La recopilación y análisis de datos a través de anuarios estadísticos permiten a los investigadores y responsables políticos comprender mejor las dinámicas económicas del país y tomar decisiones informadas basadas en evidencias empíricas. Sin embargo, el acceso a esta información no siempre se traduce en una comprensión inmediata. La complejidad y el volumen de datos presentados en estos anuarios pueden dificultar su interpretación.

**Antecedentes del Problema Científico**

A pesar de la disponibilidad de datos en los anuarios estadísticos, existe una brecha significativa entre la recopilación de estos datos y su utilización efectiva para responder preguntas específicas. Los ***Modelos de Lenguaje de Gran Escala(LLM)*** han demostrado ser herramientas poderosas para el procesamiento del lenguaje natural, pero su rendimiento puede verse limitado cuando se enfrentan a datos tabulares y textuales sin un contexto adecuado. Esto se debe a que los LLM generalmente se entrenan con texto no estructurado. Esto plantea una necesidad crítica: desarrollar un sistema que no solo procese estos datos, sino que también facilite la interacción con un LLM para generar respuestas precisas y relevantes.

La justificación para esta investigación radica en la necesidad de mejorar la capacidad de los LLM para responder preguntas específicas sobre datos tabulares y textuales al proporcionar un marco estructurado que facilite la interacción entre el modelo y los datos. Este enfoque no solo beneficiará a los investigadores que buscan respuestas rápidas y precisas, sino que también mejorará la accesibilidad de la información para los responsables políticos y usuarios en general.

**Breve Presentación de la Problemática**

La problemática central se relaciona con la dificultad que enfrentan los modelos de lenguaje al intentar interpretar y responder preguntas basadas en datos tabulares extraídos de anuarios estadísticos. Aunque estos modelos son capaces de procesar texto no estructurado con gran eficacia, su rendimiento disminuye cuando se trata de datos organizados en tablas. Esto limita su utilidad en contextos donde se requiere información precisa y contextualizada sobre indicadores económicos.

Además, el volumen y la complejidad de los datos presentados en los anuarios pueden resultar abrumadores para los usuarios que no están familiarizados con el análisis estadístico. Por lo tanto, es esencial desarrollar un sistema que no solo procese estos datos, sino que también proporcione un contexto adecuado para que el LLM pueda generar respuestas relevantes.

**Actualidad, Novedad Científica e Importancia Teórica y Práctica**

En el contexto actual, donde los LLM están ganando popularidad en diversas aplicaciones, desde asistentes virtuales hasta análisis predictivos, la necesidad de integrar estos modelos con datos estructurados es más relevante que nunca. La novedad científica de esta investigación radica en su enfoque interdisciplinario que combina técnicas de procesamiento del lenguaje natural con análisis estadístico.

Desde una perspectiva teórica, este estudio contribuirá al campo del procesamiento del lenguaje natural al explorar cómo los LLM pueden ser mejorados mediante el uso de datos tabulares y textuales estructurados. Desde una perspectiva práctica, facilitará el acceso a información económica crítica para investigadores y responsables políticos, permitiendo una toma de decisiones más informada basada en evidencia empírica.

**Diseño Teórico**

1. **Problema Científico**: ¿Cómo puede un LLM ser optimizado para responder preguntas sobre datos tabulares y textuales extraídos de anuarios estadísticos?

2. **Objeto de Estudio**: Datos tabulares y textuales extraídos de anuarios estadísticos publicados por la ONEI.

3. **Objetivos**:

- **Objetivo General**: Desarrollar un sistema que procese tanto los datos tabulares como los textos extraídos de anuarios estadísticos para mejorar las respuestas generadas por un LLM.

- **Objetivos Específicos**:

- Extraer y limpiar tanto los datos tabulares como el texto desde anuarios estadísticos.

- Estructurar estos datos en una base de datos accesible.

- Enriquecer el contexto del LLM mediante metadatos.

- Evaluar la precisión del modelo al responder preguntas específicas basadas en estos datos.

4. **Campo de Acción**: Este estudio se centrará en el ámbito del procesamiento del lenguaje natural aplicado a análisis económicos utilizando herramientas tecnológicas contemporáneas.

**Hipótesis o Pregunta Científica**

La hipótesis central es que "la integración efectiva de datos tabulares y textuales extraídos de anuarios estadísticos con un modelo de lenguaje mejorará significativamente la precisión y relevancia de las respuestas generadas por dicho modelo ante consultas específicas relacionadas con indicadores económicos".