Propuesta

Título

The "SpanishMedicaLLM " a Large Language Model (LLM) on medical domain for Spanish language

Motivación

Actualmente, en plataformas como hugginface[1] no existen modelos pre-entrenados exclusivamente en español sobre el dominio médico [1], [2].

La existencia de más de 600 millones de personas que son hablantes del idioma español sugiere la necesidad de crear recursos como los LLMs que permita la consulta en diversas formas para la obtención de información médica de forma libre y segura, cumpliendo con varios objetivos del milenio propuestos por la ONU [3], [4].

El proceso de pre-entrenamiento o autoajuste de un LLM necesita de información como materia prima para el aprendizaje. En la actualidad existen al menos 5 conjuntos de datos en plataformas como hugginface (p.ej: LenguajeNaturalAI/casos_clinicos_tratamiento,) para ser utilizados de alguna forma en el entrenamiento de LLMs. Estos no tienen igual objetivo o formato y la cantidad de información no es suficiente para el preentrenamiento de un LLM.

Otros investigadores han creado diferentes iniciativas con distintos formatos (p.ej.; MedLexSp[2]) y objetivos [2], [5], [6], [7] que pueden ser empleados en la creación de un LLM para el dominio médico en español, pero no han sido agrupados y organizados para permitir ser utilizados como fuente de conocimiento de LLMs.

Para enfrentar estos retos: La existencia de pocos recursos o conjuntos de datos del dominio médico para el entrenamiento o autoajuste de un LLM en idioma español y la no existencia de un LLM en este idioma para este dominio proponemos:

- Generar un conjunto de datos para la plataforma hugginface que permita agrupar varios de los corpus, resultados de investigaciones, diccionarios o plataformas en internet (p.ej.; Wikipedia, MedlinePlus) sobre el dominio médico en idioma español para el uso posterior de la comunidad de investigadores en el entrenamiento de LLMs.
- Proponer la creación de un LLM mediante pre-entrenamiento o autoajuste para el dominio médico. Ajustando esta tarea o reto a que tiene un alto costo computacional, lo que resulta en un costo económico y medioambiental elevado y el tiempo reducido

[1] https://huggingface.co/
[2] https://digital.csic.es/handle/10261/270429

para el desarrollo de una propuesta en el marco del Hackathon 2024 de la organización Somos NLP, en el cual surge la propuesta.

El conjunto de datos se categorizará por temas dentro de la medicina, e incluirán el número de tokens por cada tema, el tipo de tarea para la que podrían ser usada la información, y si están anotados o son texto libre. Este corpus será donado al hackathon para que los investigadores y organizaciones puedan hacer uso de ello y el código para la creación como código abierto (open source).

Organización y Tareas

Cada persona se le asignará una o varias fuentes de información con el objetivo de extraer todo su contenido y organizar la información propuesta en cada fuente.

A tener en cuenta:

- ✓ La fuente de donde procede la información y el tipo de licencia que esta posee para copiar o usar la misma.
- ✓ Contar los tokens totales presente en esta fuente de información.
- ✓ Penso en Mbyte o Gigabytes.
- ✓ Si se encuentra dividida por diferentes tópicos (p.ej.; temas, tratamientos, diagnósticos, Q&A, texto libre) o es una única fuente como pudiera ser un libro.
- ✓ En caso que la fuente de información asignada no se ajuste a la organización propuesta para el conjunto de datos, consultar con Dionis para revisar si hace falta extender la estructura propuesta.

Se propone por cada entrada o documento en la fuente de información organizarla en un dataset de hugginface de la siguiente forma:

raw_text: Texto asociado al documento, pregunta, caso clínico u otro tipo de información.

topic: (puede ser tratamiento, diagnostico, tema, respuesta a pregunta, o estar vacío p.ej en el texto abierto)

speciality: (especialidad médica a la que se relaciona el raw_text p.ej: cardiología, cirugía, otros)

raw_text_type: (puede ser caso clínico, open_text, question)

topic_type: (puede ser medical_topic, medical_diagnostic,answer,natural_medicine_topic, other, o vacio)

source: Identificador de la fuente asociada al documento que aparece en el README y descripción del dataset.

country: Identificador del país de procedencia de la fuente (p.ej.; ch, es) usando el estándar ISO 3166-1 alfa-2 (Códigos de país de dos letras.).

Al inicio de este proceso de construcción se debe actualizar en una tabla del README y descripción de la fuente de información los siguientes datos:

Identificador: Este será un numero para que la fuente de información pueda ser referenciada en cada entrada del conjunto de datos.

Nombre de la Fuente:

Cantidad de tokens:

Licencia de uso: En este caso si es solo para investigación o si posee otra licencia como MIT, Apache 2

Dirección: URL de donde se puede descargar o consultar la información.

NOTA IMPORTANTE:

Al procesar y estructurar los datos de las fuentes de información esta debe ser actualizada (commit) en **somosnlp/spanish_medica_llm** que es el conjunto de datos creados para almacenar el corpus creado por nuestro equipo.

Referencias

[1] G. G. Subies, Á. B. Jiménez, y P. M. Fernández, «A Survey of Spanish Clinical Language Models», *ArXiv Prepr. ArXiv230802199*, 2023.

[2] J. Harkawat y T. Vaidhya, «Spanish Pre-Trained Language Models for HealthCare Industry.», en *IberLEF@ SEPLN*, 2021, pp. 796-802.

[3] Á. M. B. Pulgarín, «De los Objetivos de Desarrollo del Milenio (ODM) a los Objetivos de Desarrollo Sostenible (ODS): una oportunidad para la educación sostenible con perspectiva de géneros», *Rev. En-Contexto*, vol. 8, n.º 12, pp. 69-91, 2020.

[4] A. G. Caja y M. G. de las Heras González, «Latin America in Spanish press: A geopolitical analysis about Ibero-America Summits held in Spain», *methaodos.*, vol. 60, p. 77.

[5] P. Báez, F. Villena, M. Rojas, M. Durán, y J. Dunstan, «The Chilean Waiting List Corpus: a new resource for clinical named entity recognition in Spanish», en *Proceedings of the 3rd clinical natural language processing workshop*, 2020, pp. 291-300.

[6] L. Campillos-Llanos, «MedLexSp–a medical lexicon for Spanish medical natural language processing», *J. Biomed. Semant.*, vol. 14, n.º 1, p. 2, 2023.

[7] C. P. Carrino *et al.*, «Pretrained biomedical language models for clinical NLP in Spanish», en *Proceedings of the 21st Workshop on Biomedical Language Processing*, 2022, pp. 193-199.

Ojo cambiar el Tokenizador que use datos de español y catalán como en el de Aguila en hugginface.

Conjunto de datos	Tipo	Responsable	Tokens	Fuente
Wikipedia	**Para español**	**Alvaro**		**https://huggingface.co/datasets/wikipedia**
CodiEsp corpus: gold standard Spanish clinical cases coded in ICD10 (CIE10) - eHealth CLEF2020	**In** [1]	**Alvaro**		**https://zenodo.org/records/3837305#.XsZFoXUzZpg**
Cantemist corpus: gold standard of oncology clinical cases annotated with CIE-O 3 terminology	**In** [1]	**Dionis**		**https://huggingface.co/datasets/PlanTL-GOB-ES/cantemist-ner** **https://zenodo.org/records/3978041**
The Chilean Waiting List Corpus	**In** [1]	**Daniel**		**https://zenodo.org/records/7555181**
CT-EBM-SP - Corpus of Clinical Trials for	**In** [1]	**Daniel**	**292 173**	**https://zenodo.org/records/6059737**

Evidence-Based-Medicine in Spanish				
MedlinePlus Spanish (National Library of Medicine, NLM)		**Dionis**		**https://medlineplus.gov/spanish/**
PlanTL-GOB-ES/pharmaconer (Estudiar como llevar a formato y esta relacionado con el de mas abajo)	**In** [1]	**Dionis**		**https://huggingface.co/datasets/PlanTL-GOB-ES/pharmaconer**
The Spanish Clinical Case Corpus (SPACCC)	**In** [1]	**Dylan**	**396,988**	**https://zenodo.org/records/2560316** **https://github.com/PlanTL-GOB-ES/SPACCC**
ORDO	**Estudiar como construir un recurso que se parezca a la base de datos**			**https://www.orphadata.com/ordo/** **https://www.orphadata.com/docs/WhatIsORDO.pdf**
DisTEMIST corpus:	**In** [1]	**Dionis**		

detection and normalization of disease mentions in spanish clinical cases				
Spanish Biomedical Crawled Corpus	**En el texto tiene la descripción y la clasificación como enfermedad esto suena como un recurso a prentrenar**	**Dionis**	**Grande**	**https://zenodo.org/records/5513237#.Yp7lU_exWV4**
MeSpEn_Parallel-Corpora		**Dylan**		**https://zenodo.org/records/3562536#.YlP1UshBwio**
eHealthKD	**In** [1] **No muy sencillo**	**Dionis**		**https://github.com/ehealthkd/corpora/tree/master**
European Clinical Case Corpus	**In** [1]	**Dylan**		**https://live.european-language-grid.eu/catalogue/corpus/7618/download/**
Biomedical Abbreviation Recognition and	**In** [1]	**Daniel**		

Resolution 2nd Edition (BARR2)				
CARES	**In** [1]	**Dionis**		**https://huggingface.co/datasets/chizhikchi/CARES**
IULA-SCRC	**In** [1]	**Daniel**		**https://github.com/Vicomtech/NUBes-negation-uncertainty-biomedical-corpus**
LivingNER	**In** [1]	**Daniel**		**https://zenodo.org/record/7614764**
MEDDOCAN	**In** [1]	**Dionis**	**MEDDOCAN (Medical Document Anonymization Track) [60] is a corpus of clinical cases sampled from the SPACCC corpus and enriched with synthetic personal information.**	**https://huggingface.co/datasets/bigbio/meddocan**