

Tareas en NLP Clínico

Profesora: Tamara Quiroga
t.quiroga@uc.cl

Departamento de Ciencia de la Computación
Pontificia Universidad Católica de Chile

May 9, 2025



Contenidos

Presentacion

Conceptos Previos

Tareas en NLP Clínico

Resolución de Tareas NLP Clínico

Contenidos

Presentacion

Conceptos Previos

Tareas en NLP Clínico

Resolución de Tareas NLP Clínico

Acerca de mí



- Soy de Santiago
- Soy Ingeniera Civil Matemática y también hice el Magister en Economía Aplicada ambos en la Universidad de Chile
- Soy estudiante del Doctorado de Ingeniería en Computación UC

Contenidos

Presentacion

Conceptos Previos

Tareas en NLP Clínico

Resolución de Tareas NLP Clínico

Conceptos Previos

- Tarea
- Dataset
- NLP Clínico
- NLP Biomédico

Tareas y Datasets

Tareas y Datasets

- **Tarea:** Es un **tipo específico de problema** se busca resolver usando un modelo de NLP.

Tareas y Datasets

- **Tarea:** Es un **tipo específico de problema** se busca resolver usando un modelo de NLP. Ej: clasificar texto, clasificar palabras.

Tareas y Datasets

- **Tarea:** Es un **tipo específico de problema** se busca resolver usando un modelo de NLP. Ej: clasificar texto, clasificar palabras.
- **Dataset:** Es un **conjunto de datos** que usamos para **entrenar, validar, evaluar una tarea específica**.

Tareas y Datasets

- **Tarea:** Es un **tipo específico de problema** se busca resolver usando un modelo de NLP. Ej: clasificar texto, clasificar palabras.
- **Dataset:** Es un **conjunto de datos** que usamos para **entrenar, validar, evaluar una tarea específica**.
- Es usual que una tarea pueda evaluarse con diferentes datasets.

Ejemplos Tareas

Tarea: Analisis de Sentimiento

Ejemplos Datasets

Dataset: Stanford Sentiment Treebank (SST)

- **Contenido:** Comentarios sobre películas.
- **Etiquetas:** Positivo, Negativo, Neutral.

comentario	etiqueta
"A touching and insightful story."	Positivo
"The plot was predictable and boring."	Negativo

Table 1: ejemplos del dataset

Ejemplos Datasets

Dataset: STS-Gold Dataset

- **Contenido:** Comentarios de Tweets.
- **Etiquetas:** Positivo, Negativo.

comentario	etiqueta
"@SeattleWillow Have a nice time work!"	Positivo
"Well, that sucked."	Negativo

Table 2: ejemplos del dataset

¿Cómo se construyen los datasets?

¿Cómo se construyen los datasets?

Etiqueta los siguientes textos	
• <i>texto 1</i> • <i>texto 2</i> • <i>texto 3</i> • <i>texto 4</i> • <i>texto 5</i> • <i>texto 6</i>	+, -, 0



¿Cómo se construyen los datasets?

Etiqueta los siguientes textos	
• <i>texto 1</i> • <i>texto 2</i> • <i>texto 3</i> • <i>texto 4</i> • <i>texto 5</i> • <i>texto 6</i>	+, -, 0



¿Cómo se construyen los datasets?

Etiqueta los siguientes textos



+
-
0
+
+

+
+
0
+
-

0
-
+
-
+

0
-
0
+
+

$$\begin{matrix} + \\ + \\ + \\ + \\ 0 \end{matrix}$$

MEDIR AGREEMENT

- Trabaja con **artículos científicos, abstracts, guías clínicas.**

Ejemplos: papers, ensayos clínicos, bases de datos públicas.



Contenidos

Presentacion

Conceptos Previos

Tareas en NLP Clínico

Resolución de Tareas NLP Clínico

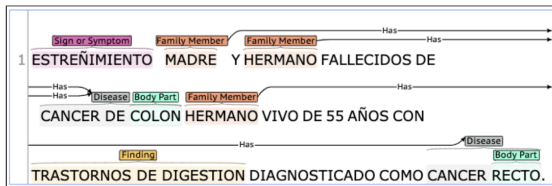
Tareas en NLP Clínico

Tareas que veremos hoy.

- Reconocimiento de entidades nombradas (NER)
- Preguntas a respuestas:

El Reconocimiento de entidades nombradas (NER por sus siglas en inglés) es una tarea de extracción de información que busca **localizar y clasificar en categorías entidades nombradas encontradas en un texto.**

En el contexto clínico estas entidades podrán ser **entidades médicas** como: **enfermedades, síntomas, medicamentos, procedimientos y términos anatómicos.**



Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (VAERS) using deep learning

[Jingcheng Du](#)^{1,1}, [Yang Xiang](#)^{1,1}, [Madhuri Sankaranarayanapillai](#)¹, [Meng Zhang](#)¹, [Jingqi Wang](#)¹, [Yuqi Si](#)¹, [Huy. Anh Pham](#)¹, [Hua Xu](#)¹, [Yong Chen](#)², [Cui Tao](#)^{1,8}

Figure 2: Paper que crean dataset para tarea NER

VAERS: Contiene registros individuales que notifican eventos adversos que ocurren después de la vacunación de EE.UU.

Cada reporte incluye:

- **Datos del paciente:** edad, sexo, ubicación.
- **Vacuna administrada:** tipo, marca, fecha de aplicación.
- **Descripción del evento adverso:** qué síntomas o problemas ocurrieron, cuándo empezaron, su duración y gravedad.
- **Resultado:** si la persona se recuperó, fue hospitalizada, tuvo complicaciones graves o falleció.
- **Datos del reportante:** puede ser un profesional de salud, el propio paciente o un familiar.

- Selección 91 reportes para etiquetar.
- Identifican 6 entidades: *investigation*, *nervous_AE*, *other_AE*, *procedure*, *social_circumstances*, and *temporal_expression*.
- 5 médicos etiquetan este dataset.

Entity Type	Definition	Example	Annotation Agreement
<i>investigation</i>	Lab tests and examinations	“neurological exam,” “lumbar puncture”	0.7436
<i>nervous_AE</i>	Symptoms and diseases related to nervous system disorders	“tremors,” “Guillain-Barré syndrome”	0.7885
<i>other_AE</i>	Other symptoms and diseases	“complete bowel incontinence,” “diarrhea”	0.8017
<i>procedure</i>	Clinical interventions to the patient, including vaccination, treatment and therapy, intensive care, etc.	“flu shot,” “hospitalized”	0.7492
<i>social_circumstance</i>	Events associated with the social environment of a patient	“smoking,” “alcohol abuse”	0.4706
<i>temporal_expression</i>	Temporal expressions with prepositions	“for 3 days,” “on Friday morning”	0.8448

Figure 3: vaers Du et al. [2021]

The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish

Pablo Báez¹, Fabián Villena^{1,2}, Matías Rojas³, Manuel Durán¹, and Jocelyn Dunstan^{1,2}

¹Center for Medical Informatics and Telemedicine, University of Chile.

²Center for Mathematical Modeling, University of Chile.

³Department of Computer Sciences, University of Chile.

`{pablobaez,manuel.duran,matias.rojas.g}@ug.uchile.cl`

`{fabian.villena,jdunstan}@uchile.cl`

- PubMedQA (Jin et al. [2019]) es un dataset diseñado específicamente para tareas de pregunta-respuesta (QA) en el ámbito biomédico. Se basa en artículos científicos de PubMed, que es una base de datos muy conocida de literatura biomédica

PubMedQA

- PubMedQA contiene preguntas tipo sí/no/quizás relacionadas con temas médicos.
- Cada pregunta está asociada a un resumen (abstract) de un artículo científico y a una respuesta correcta (yes, no, o maybe).
- Su objetivo es evaluar si los modelos de lenguaje o sistemas automáticos pueden leer un texto biomédico y responder preguntas de manera precisa, como lo haría un experto humano.

Question:

Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?

Context:

(Objective) Recent studies have demonstrated that statins have pleiotropic effects, including anti-inflammatory effects and atrial fibrillation (AF) preventive effects [...]

(Methods) 221 patients underwent CABG in our hospital from 2004 to 2007. 14 patients with preoperative AF and 4 patients with concomitant valve surgery [...]

(Results) The overall incidence of postoperative AF was 26%. *Postoperative AF was significantly lower in the Statin group compared with the Non-statin group (16% versus 33%, $p=0.005$).* Multivariate analysis demonstrated that independent predictors of AF [...]

Long Answer:

(Conclusion) Our study indicated that preoperative statin therapy seems to reduce AF development after CABG.

Answer: yes

Figure 4: Ejemplo

PQA-Artificial (211.3k)	PQA-Unlabeled (61.2k)	PQA-Labeled (1k)
Ori. Title -> Question	Ori. Question Title	Ori. Question Title
Structured Context (Ori. Abstract w/o conclusion)	Structured Context (Ori. Abstract w/o conclusion)	Structured Context (Ori. Abstract w/o conclusion)
Long Answer (Conclusion)	Long Answer (Conclusion)	Long Answer (Conclusion)
Generated yes/no	Unlabeled	Yes/no/maybe

Figure 2: Architecture of PubMedQA dataset. PubMedQA is split into three subsets, PQA-A(rtificial), PQA-U(nlabeled) and PQA-L(abeled).

Figure 5: Estructura PubMedQA Jin et al. [2019]

Otros dataset QA

Otros dataset QA Médico/Biomédico:

- MedQA
- MedMCQA

Limitaciones de los dataset para tareas clínicas

- Requieren conocimiento experto.
- Su elaboración es usualmente costosa.
- Tienden a ser dataset pequeños.
- Principalmente se desarrollan en inglés.

Contenidos

Presentacion

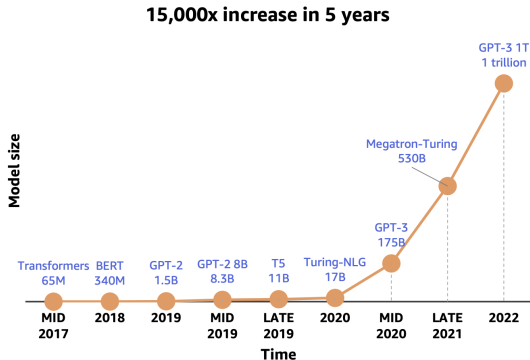
Conceptos Previos

Tareas en NLP Clínico

Resolución de Tareas NLP Clínico

Modelos de lenguaje y tamaño

- Los modelos a **mayor número de parámetros mejoran su desempeño**.
- A estos modelos se les denomina **grandes modelos de lenguaje (LLM)**.



Modelos de Lenguaje

Los modelos de lenguaje pueden **resolver tareas a través de diferentes metodologías** usualmente las metodologías factibles se definen según

- **Arquitectura**
- **Open source o cerrado**

Modelos de Lenguaje

Caso I: modelos enmascarados

Fill-Mask

Examples

Mask token: [MASK]

Paris is the [MASK] of France.

Compute

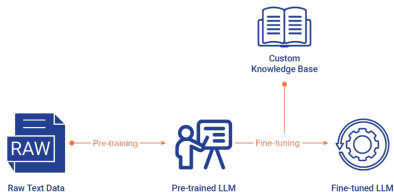
Computation time on cpu: cached

capital	0.997
heart	0.001

Usualmente este tipo de modelos son descargables de HuggingFace.

Métodos de resolución de tareas

La metodología **más usual** de resolución es a través de **finetuning**.



Finetuning

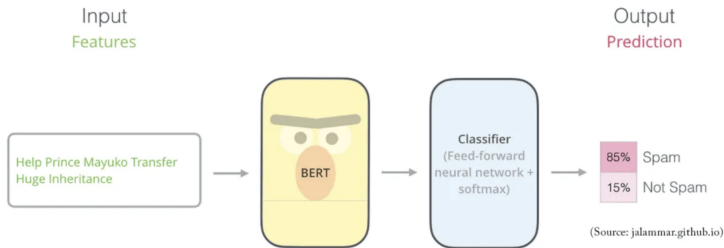


Figure 6: Finetuning Bert

Finetuning

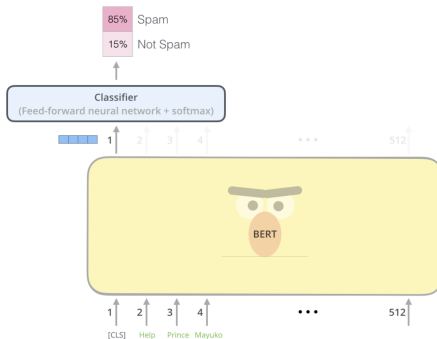
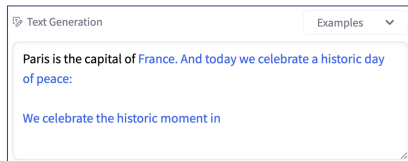


Figure 7: Finetuning Bert

Modelos de Lenguaje

Por ejemplos modelos **generativos**:



Métodos de resolución de tareas

- Finetuning
- Prompting

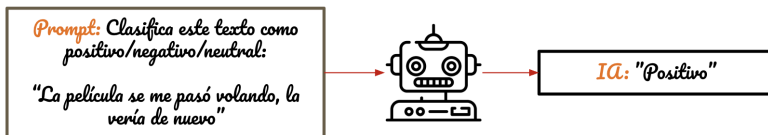


Figure 8: Prompt

Referencias I

- Pablo Báez, Fabián Villena, Matías Rojas, Manuel Durán, and Jocelyn Dunstan. The chilean waiting list corpus: a new resource for clinical named entity recognition in spanish. In *Proceedings of the 3rd clinical natural language processing workshop*, pages 291–300, 2020.
- Jingcheng Du, Yang Xiang, Madhuri Sankaranarayananapillai, Meng Zhang, Jingqi Wang, Yuqi Si, Huy Anh Pham, Hua Xu, Yong Chen, and Cui Tao. Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (vaers) using deep learning. *Journal of the American Medical Informatics Association*, 28(7):1393–1400, 2021.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.