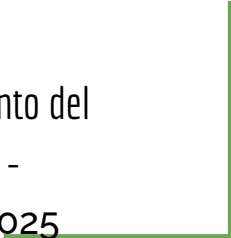




Universidad de Buenos Aires  
Departamento de Computación

# Construcción de recursos para NLP (anotación)

Temas de Procesamiento del  
Lenguaje Natural -  
2do trimestre 2025



Dra. Viviana Cotik

# Implementación de algoritmos de ML



# Aprendizaje supervisado

Necesitamos datos anotados

Para esto requerimos:

- obtener datos
  - disponibilidad
  - ¿se pueden publicar?, sensibilidad de la información

Anotar datos. **Ejercicio**

# Ejercicio de anotación

Dados los siguientes textos, anotar:

- María Vanina García es empleada de la empresa New York Company desde el 1 de enero de 2010.
  - Nombres, ciudades, nombres de empresa, fechas
- La resonancia magnética cerebral revela un cerebro sin alteraciones estructurales, con lóbulos cerebrales izquierdo y derecho bien diferenciados y sin evidencia de lesiones. Además, se observa una adecuada perfusión sanguínea en las arterias cerebrales principales. Los resultados son normales y no se detectan anomalías significativas en las entidades anatómicas evaluadas.
  - Entidades anatómicas, negaciones, hallazgos clínicos.

Anotar [aquí](#).

# Anotación

**Definición:** Una anotación es cualquier etiqueta o metadato utilizado para marcar elementos del conjunto de datos. Pustejovsky and Stubbs [PS2012].

## Necesaria para:

- evaluar algoritmos
- entrenar algoritmos de aprendizaje automático supervisado
- Muchas veces **cuello de botella** para avanzar con el estado del arte.
- Los datos anotados también se denominan ***ground truth***.
- Nos focalizamos en **anotación de textos**, aunque muchos de los conceptos también son válidos para la anotación de datos en otro tipo de formato

# Cómo obtener textos anotados

- Conseguir recursos disponibles
  - Competencias
  - Papers
  - Benchmarks
- Anotar el propio dataset

# Tipos de anotación

- **Linguística**
  - Morfológica (POS tagging, lematización, rasgos gramaticales)
  - Sintácticas (chunking -identificación de unidades sintácticas, como frases nominales, verbales-, árboles sintácticos .parsing-, anotación de dependencias sintácticas)
  - Semánticas (semantic role labeling -identificación de roles desempeñados por diferentes constituyentes de la oración- (agente, tema), word sense disambiguation, NER).
- **Discursivas**
  - Análisis del discurso (identificación de estructuras como anáforas)
  - Actos de habla (clasificación de fragmentos según actos de habla, por ej. Afirmación, pregunta)
- **Sentimiento y Emoción**
  - Reconocimiento de emociones y sentimientos
- **Anotaciones con conocimiento del dominio**
  - Médicas, jurídicas, finanzas
- **Otros**
  - Ej. topic modelling (identificación de temas principales de un conjunto de datos)

# Tareas de un proceso de anotación

Entre otras tareas, un proceso de anotación demanda [IP2017]

- Crear archivos en un formato de archivo estándar
- Escribir **criterios de anotación**
- Definir las **habilidades y conocimientos** necesarios para los **anotadores**
- **Entrenar a los anotadores** en el esquema de anotación hasta alcanzar un acuerdo entre anotadores (**IAA**) **razonable**
- Planificar el orden y las **asignaciones de anotación**
- Distribuir documentos a los anotadores
- Monitorear el progreso de los anotadores
- Recoger las anotaciones de los anotadores
- Rastrear el acuerdo entre anotadores para asegurar la calidad de las anotaciones
- Programar reuniones
- Rastrear las horas de trabajo y el presupuesto del proyecto



# Desafíos del proceso de anotación

- Costosa
  - tareas involucradas
  - conseguir y pagar anotadores
- Puede ser compleja (anotación y definición de criterio). Los textos a anotar pueden
  - carecer de estructura gramatical y de signos de puntuación.
  - tener abundancia de acrónimos y abreviaturas ambiguas
  - tener un vocabulario muy específico.

# Temario

- Introducción a Anotación
- **Definición de esquema y criterios de anotación**
- Anotación
- Data Statements
- Benchmarks

# Definición de esquema y criterios de anotación

- Depende de la tarea de interés:
  - clasificación, NER, RE, análisis sintáctico, etc.
- **Hay que definir:**
  - **Esquema de anotación.**
    - Objetivo: identificar entidades y relaciones de interés, negaciones, hedges, etc.
  - **Criterios de anotación.**
    - Documentación. Con ejemplos.
    - Lo más objetivo posible. Distintos anotadores tienen que poder leerlo y entender lo mismo.

# Esquema de anotación

**Definir qué se va a anotar.**

Se va puliendo en las distintas rondas de anotación. Por ej.

## **NER:**

- Describir cada entidad de interés, qué significa la entidad, ejemplos de la entidad.

## **Extracción de relaciones:**

- describir cada relación de interés, qué entidades relaciona, dar ejemplos

## **Clasificación:**

- en qué categorías

# Criterios de anotación

- Definir criterios.
- Se van **ajustando** en las distintas rondas de anotación
- Ejemplos de cuestiones a definir:
  - Se anota aún si tiene **errores de ortografía**
  - ¿Cómo se anotan **entidades anidadas**? "The MRI scan showed abnormalities in the [left ventricle of the [heart]] and the [frontal lobe of the [brain]]."
  - ¿Hay algún recurso, por ej, lexicon, que pueda utilizarse para resolver las dudas acerca de cómo anotar un término?
  - ¿Se pueden anotar entidades discontinuas? **Intra y extra hepáticas**
  - Si no se sabe a cuál de dos entidades nombradas pertenece un término, **cuál o con qué criterio priorizar a alguna de ellas**. Ej: Priorizar findings (FI) over locations (LO). Si un FI incluye una LO, anotar el término más grande correspondiente a FI, : pyloric stenosis refers to a FI, that includes a location. Therefore, [pyloric stenosis](FI) should be chosen over [pyloric](AE) [stenosis](FI)).

# Temario

- Introducción a Anotación
- Definición de esquema y criterios de anotación
- **Anotación**
  - **Selección de herramienta de anotación**
  - Selección de anotadores
  - Selección de datos a anotar
  - Proceso de anotación
  - Medidas de acuerdo entre anotadores (inter annotator agreement)
  - Alternativas de anotación
- Data Statements
- Benchmarks

# Herramientas de anotación

- Dependiendo de la tarea estudiar utilidad de utilizar herramientas de anotación:
  - desarrollo propio
  - planilla de cálculo, por ej., para clasificación de textos
  - herramienta existente
    - ej: para NER y relaciones [brat annotation toolkit](#) ([ejemplo](#)). Formato: [brat standoff-format](#)
    - existen surveys de herramientas, ej [NL2014] para textos biomédicos.
- Evaluar:
  - funcionalidades (por ej. permite anotar entidades discontinuas?)
  - existencia de funcionalidades de pre anotación
  - posibilidad de acceso fuentes de conocimiento externas, por ej, ontologías
  - visualización clara de las anotaciones
  - facilidad de anotar
  - formato de anotaciones

# Temario

- Introducción a Anotación
- Definición de esquema y criterios de anotación
- **Anotación**
  - Selección de herramienta de anotación
  - **Selección de anotadores**
  - Selección de datos a anotar
  - Proceso de anotación
  - Medidas de acuerdo entre anotadores (inter annotator agreement)
  - Alternativas de anotación
- Data Statements
- Benchmarks



# Selección de anotadores

- Evaluar:
  - conocimientos del dominio requeridos
  - idioma materno que habla el anotador
- Retribución:
  - ¿se puede ofrecer?
- Cálculo de tiempos de anotación

# Temario

- Introducción a Anotación
- Definición de esquema y criterios de anotación
- **Anotación**
  - Selección de herramienta de anotación
  - Selección de anotadores
  - **Selección de datos a anotar**
  - Proceso de anotación
  - Medidas de acuerdo entre anotadores (inter annotator agreement)
  - Alternativas de anotación
- Data Statements
- Benchmarks

# Selección de dataset a anotar

- Obtener el dataset
- Normalizarlo/estandarizarlo
- Eliminar duplicados
- Extraer un subconjunto de interés
- Anonimizar los textos, de ser necesario
- Definir qué porcentaje se anotará por más de un anotador
- Asignar los textos a los distintos anotadores

# Anonimización

- De tratarse de información sensible, el dataset debería anonimizarse previo a la anotación.
- Existen distintos tipos de anonimización:
  - **No perturbativa:** preserva demografía de pacientes (por ej.)
  - **Perturbativa:** no la preserva

# Temario

- Introducción a Anotación
- Definición de esquema y criterios de anotación
- **Anotación**
  - Selección de herramienta de anotación
  - Selección de anotadores
  - Selección de datos a anotar
  - **Proceso de anotación**
  - Medidas de acuerdo entre anotadores (inter annotator agreement)
  - Alternativas de anotación
- Data Statements
- Benchmarks

# Posibles formas de anotar textos

Posibilidades para anotar textos:

1. **anotación manual** basada en el conocimiento de los anotadores
2. **anotación asistida**, en donde los resultados de una herramienta de anotación son corregidos (**pre-anotación**)
3. de existir, una **anotación** (manual o asistida) **basada en una ontología**, en donde sólo los términos o relaciones presentes en la misma son anotados.

Fuente: Matthew S. Simpson and Dina Demner-Fushman. Mining Text Data, chapter 14. Biomedical Text Mining: A Survey of Recent Progress. Springer, 2012

# Pre- anotación

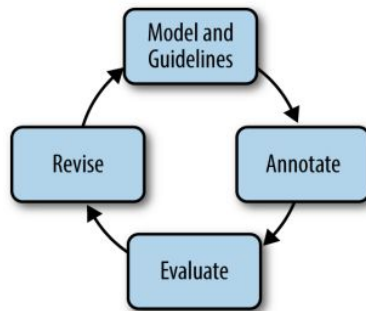
- Por ej. utilizando recursos propios y/o ajenos como ser **ontologías** o **diccionarios**, y **reglas**, por ej: expresiones regulares.
- Los recursos se pueden **retroalimentar** en un proceso iterativo, a partir de las anotaciones corregidas por los anotadores.
- Los anotadores tienen que revisar datos pre-anotados, corregirlos y agregar lo que falte.

|   |   |
|---|---|
|  |  |
| Es más rápido   | Sesga a los anotadores  |

# Proceso de anotación

- Definir el esquema y criterio y anotar pocos textos.
- Mirar anotaciones y medir desacuerdos
- Iterar hasta que se estabilizan el criterio y las anotaciones.

Model Annotate Model Annotate (MAMA Cycle)



Ciclo MAMA para el proceso de anotación. Tomado de Pustejovsky y Stubbs [PS2012].



# Temario

- Introducción a Anotación
- Definición de esquema y criterios de anotación
- **Anotación**
  - Selección de herramienta de anotación
  - Selección de anotadores
  - Selección de datos a anotar
  - Proceso de anotación
  - **Medidas de acuerdo entre anotadores (inter annotator agreement)**
  - Alternativas de anotación
- Data Statements
- Benchmarks

# Acuerdo entre anotadores (IAA)

Inter annotator agreement (IAA)

- Medida utilizada para evaluar cuán buena es la anotación
  - un subconjunto de textos/datos es anotado por más de un anotador
  - se evalúa el inter annotator agreement, que mide el acuerdo entre anotadores
- Métricas para cálculo de IAA
  - Kappa de Cohen [C1960]. Definir estrategia cuando hay más de dos anotadores.
  - Alfa de Krippendorff
  - Kappa de Fleiss
  - F1

# Dataset final

- Para aquellos datos que fueron anotados por más de un anotador hay que elegir con qué anotación quedarse. Algunos criterios:
  - aquella hecha por el anotador más experto.
  - tener un tercer anotador o un revisor experto que decida
- Utilidad de describir características del dataset final para conocerlo mejor.
  - Ej. cantidad de instancias anotadas de cada entidad

# Temario

- Introducción a Anotación
- Definición de esquema y criterios de anotación
- **Anotación**
  - Selección de herramienta de anotación
  - Selección de anotadores
  - Selección de datos a anotar
  - Proceso de anotación
  - Medidas de acuerdo entre anotadores (inter annotator agreement)
  - **Alternativas de anotación**
- Data Statements
- Benchmarks

# Alternativas de Anotación

- **Anotación manual** (eventualmente con pre-anotación automática)
- **Distant supervision**
  - Anotación automática usando base de conocimiento externa.
  - Reduce esfuerzo, pero introduce ruido.
    - Se puede mejorar, por ej., agregando un conjunto pequeño de datos anotados por humanos.
- **Anotación automática**
  - Por ej. con grandes modelos de lenguaje

¿A cuál podríamos llamar gold standard?

# Temario

- Introducción a Anotación
- Definición de esquema y criterios de anotación
- Anotación
- **Data Statements**
- Benchmarks

# Data statements

Propuestos por Bender y Friedman [BF2018] para **tratar el sesgo** y otras cuestiones críticas que surgen al trabajar con procesamiento del lenguaje natural. Tratan entre otros sobre:

- **Tipo de lenguaje** utilizado en los textos (formal, tipo de español, por ej)
- **Demografía de los autores** de los textos.
- **Demografía de anotadores** (por ej. hablantes nativos de español rioplatense, profesión, conocimientos de anotación de textos)
- **¿Hubo compensación** por la tarea?

# Data statements

- “to help alleviate issues related to exclusion and bias in language technology”
- “lead to better precision in claims about how natural language processing research can generalize and thus better engineering results”

Fuente: [BF2018]

“Data statements are part of an emerging landscape for toolkits about **documentation for transparency in artificial intelligence systems**, including: Datasheets for Datasets, Model Cards for Model Reporting, Dataset Nutrition Labels, Nutrition Labels for Data and Models, FactSheets, and Data Cards.”

Fuente: <https://techpolicylab.uw.edu/data-statements/>



# Temario

- Introducción a Anotación
- Definición de esquema y criterios de anotación
- Anotación
- Data Statements
- **Benchmarks**

# Benchmarks

Para evaluar y comparar la performance de modelos en ciertas tareas y evaluar progreso en el área.

Tareas, datos y métricas.

- **GLUE (General Language Understanding Evaluation)**
  - collection of resources for training, evaluating, and analyzing NLP models relative to a range of **language understanding tasks** such as sentiment analysis, textual entailment, and similarity assessment. It includes datasets like SST-2, MNLI, QNLI, QQP, and others.
- **SuperGLUE**
  - Designed as a successor to GLUE, SuperGLUE includes a set of **more challenging language understanding tasks**. Intended to push the performance of more powerful models, including larger-scale transformer models like BERT and RoBERTa. Tasks include BoolQ, CB, COPA, MultiRC, ReCoRD.
- **SQuAD (Stanford Question Answering Dataset)**
  - Benchmark for models to answer questions based on their understanding of Wikipedia articles.
- **CoNLL Shared Tasks**
  - Annual tasks that have included a **variety of NLP challenges over the years**, such as language-independent named entity recognition, dependency parsing, semantic role labeling, and more. Each year typically focuses on a different aspect of NLP.

# Benchmarks

Para evaluar la performance de modelos en ciertas tareas.

- **SemEval (International Workshop on Semantic Evaluation)**
  - This ongoing series of evaluations provides a range of tasks related to the semantic analysis of text. SemEval tasks have included sentiment analysis, semantic textual similarity, and more.
- **WNUT (Workshop on Noisy User-generated Text)**
  - WNUT provides benchmarks for tasks that involve user-generated text, which is often noisy and informal. Tasks have included named entity recognition, entity linking, and event detection in social media text.
- **Commonsense Reasoning Benchmarks**
  - These include datasets like CommonsenseQA and Winograd Schema Challenge, designed to test a model's ability to perform reasoning based on general world knowledge.

# Algunos corpora de GLUE

Los corpora de GLUE son predominantemente en **inglés**

**CoLA (Corpus of Linguistic Acceptability):** oraciones en inglés etiquetadas como gramaticalmente correctas o incorrectas. Se utiliza para evaluar la capacidad de los modelos para juzgar la aceptabilidad gramatical.

**SST-2 (Stanford Sentiment Treebank):** contiene críticas de películas etiquetadas con sentimientos positivos o negativos. Se utiliza para evaluar la capacidad de los modelos para realizar análisis de sentimiento.

**MRPC (Microsoft Research Paraphrase Corpus):** Conjunto de pares de oraciones extraídas de noticias en línea, etiquetadas como paráfrasis si ambas oraciones en el par expresan el mismo significado o no. Utilizado para evaluar la capacidad de los modelos para detectar paráfrasis.

**STS-B (Semantic Textual Similarity Benchmark):** Conjunto de pares de oraciones donde cada par está anotado con un puntaje de 0 a 5, representando el grado de similitud semántica entre ellas. Sirve para evaluar la capacidad de los modelos para determinar la similitud semántica.

**WNUT (Workshop on Noisy User-generated Text: )** Conjunto de datos que contiene pares de preguntas de Quora, etiquetadas para determinar si las preguntas en el par son semánticamente equivalentes. Ayuda a evaluar la capacidad de los modelos para identificar preguntas duplicadas.

# Algunos corpora de GLUE

**MNLI (Multi-Genre Natural Language Inference):** Conjunto de datos de inferencia textual en el que se presentan pares de oraciones (premisa e hipótesis) etiquetadas como contradicción, neutral o entailment (implicación). Utilizado para evaluar la comprensión de lectura y el razonamiento lógico.

**QNLI (Question Natural Language Inference):** Versión adaptada del dataset SQuAD para el formato de inferencia textual. Cada ejemplo consiste en una pregunta y un pasaje, y el objetivo es determinar si el pasaje contiene la respuesta a la pregunta.

**RTE (Recognizing Textual Entailment):** Conjunto de datos que consiste en pares de oraciones etiquetadas con la relación de implicación textual (si una oración implica lógicamente la otra). Utilizado para evaluar la capacidad de reconocimiento de implicación textual.

**WNLI (Winograd Schema Challenge):** Conjunto de datos basado en el desafío de Winograd que incluye oraciones que contienen ambigüedades que se resuelven mediante el conocimiento del mundo y el razonamiento. Evalúa la capacidad de los modelos para resolver estas ambigüedades de manera coherente.

# Referencias

**[BF2018]** Bender, Emily M and Friedman, Batya, "Data statements for natural language processing: Toward mitigating system bias and enabling better science" .Transactions of the Association for Computational Linguistics. 6, pp 587--604, 2018, MIT Press. [[Link](#)]

**[C1960]** Cohen, Jacob. "A coefficient of agreement for nominal scales." *Educational and psychological measurement* 20, no. 1 (1960): 37-46.

**[IP2017]** Nancy Ide and James Pustejovsky. Handbook of Linguistic Annotation. Springer, 2017.

**[NL2014]** Neves, Mariana, and Ulf Leser. "A survey on annotation tools for the biomedical literature." *Briefings in bioinformatics* 15, no. 2 (2014): 327-340.

**[PS2012]** James Pustejovsky and Amber Stubbs. Natural Language Annotation for Machine Learning. O'Reilly Media, Inc., 2012.

# Referencias

Otras:

- [Data and its \(dis\)contents: A survey of dataset development and use in machine learning research - PubMed \(nih.gov\)](#)