MDS Master of Data Science
Universidad de Chile

PRESENTACIÓN 3 MDS7201

DEFINICIÓN DE PROYECTO

ENTIDADES MINSAL

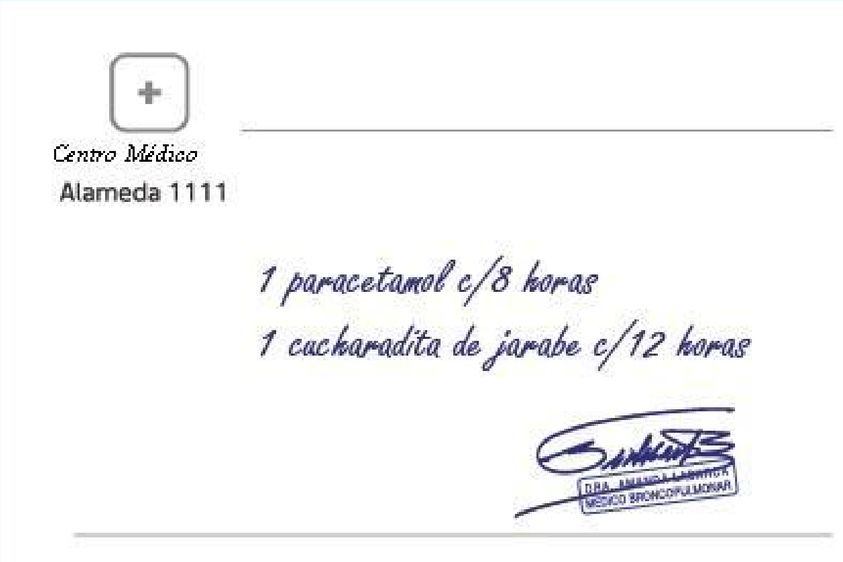
DANIEL CARMONA, MARTÍN SEPÚLVEDA,
MONSERRAT PRADO, CAMILO CARVAJAL

ENTIDADES MINSAL

TABLA DE CONTENIDO

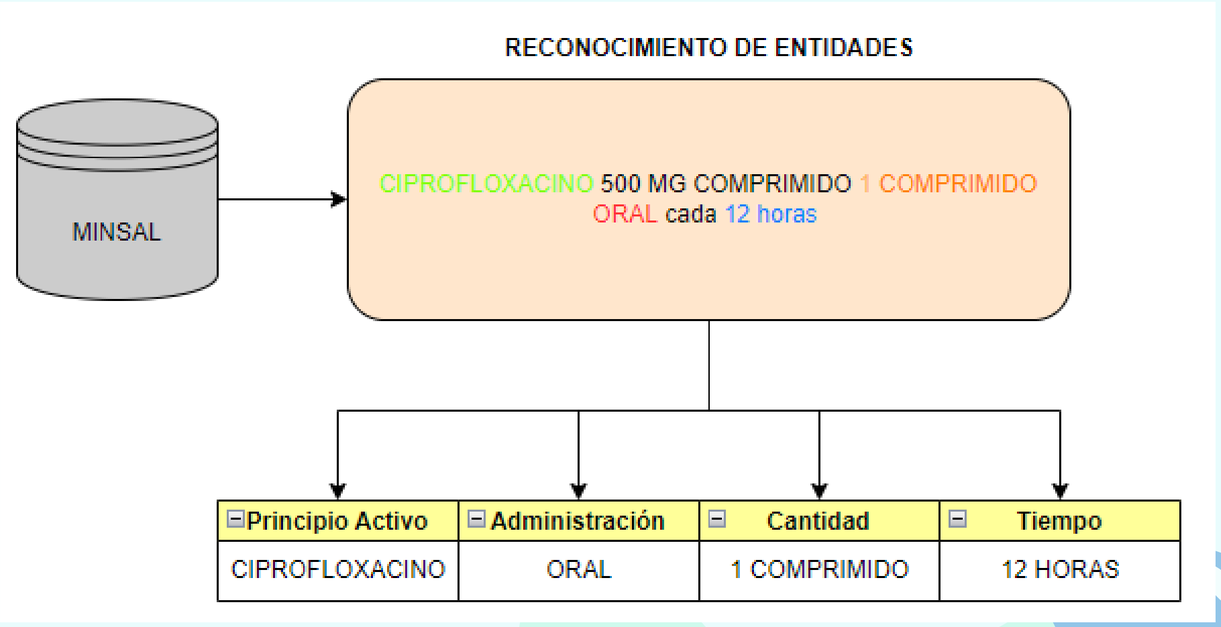
RESUMEN

- Existen recetas médicas que pueden carecer de cierta información importante, llevando a errores de medicación y a un empeoramiento en el estado del paciente.
- Las recetas electrónicas pueden contener campos de texto libre.
- Esto dificulta la verificación de la completitud de la prescripción.



RESUMEN
DESCRIPCIÓN DEL PROYECTO

- Dado un campo de texto libre, utilizar algoritmos de NLP para reconocer entidades y de esta manera completar columnas de manera automática en los datos de un paciente.
- Detectar errores de completitud o gramática en las indicaciones.
- Refraseo de la información para evitar errores de administración de medicamentos.



RESUMEN DATOS

- 1.5 [M] de prescripciones, con un total de 20 atributos por cada una

	CODIGO_MEDICAMENTO	PRES_DENOMINACION	RESUMEN	IND_ADMINISTRACION_1	IND_ADMINISTRACION_2
1526553	FACC09001	CAPTOPRIL 25 MG COMPRIMIDO	1 COMPRIMIDO ORAL cada 8 horas	NaN	NaN
1526554	FANN02016	PARACETAMOL 500 MG COMPRIMIDO	2 COMPRIMIDO ORAL cada 8 horas	NaN	NaN
1526555	FAAA10002	INSULINA CRISTALINA HUMANA 100 U.I./ML SOLUCIO...	2 UNIDAD INTRAVENOSA cada 6 horas	NaN	NaN

- En ciertos atributos se cuenta con un gran porcentaje de valores vacíos o NaN.
- Estos no se consideran relevantes para el entrenamiento del modelo.

RESUMEN LITERATURA

NER: Named Entity Recognition

Sang | Meulder — 2003 — ↳ 2419

Introduction to the CoNLL-2003 shared task: language-independent named entity recognition

CoNLL

Texto Clínico

Bose | ... | Ghosh — 2021 — ↳ 4

A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts

Applied Sciences

Báez | ... | Villena — 2022 — ↳ 0

Automatic Extraction of Nested Entities in Clinical Referrals in Spanish

ACM transactions on computing for healthcare

Báez | ... | Dunstan — 2020 — ↳

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

Association for Computational Linguistics

Dunstan | Villena — 2022 — ↳ 0

Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing

Association for Computational Linguistics

Contexto Chileno

RESUMEN LITERATURA

Texto Clínico

Báez ... Dunstan — 2020

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

Association for Computational Linguistics

Báez ... Villena — 2022 — ↳ 0

Automatic Extraction of Nested Entities in Clinical Referrals in Spanish

ACM transactions on computing for healthcare

Dunstan Villena — 2022 — ↳ 0

Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing

Association for Computational Linguistics

Conocimiento previo

Jiang ... Liu — 2019

Combining Contextualized Embeddings and Prior Knowledge for Clinical Named Entity Recognition: Evaluation Study.

JMIR medical informatics

Akbik ... Vollgraf — 2019

FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP

Association for Computational Linguistics

Kazama Torisawa — 2007 — ↳ 266

Exploiting Wikipedia as External Knowledge for Named Entity Recognition

EMNLP

Modelos de lenguaje

Devlin ... Toutanova — 2019

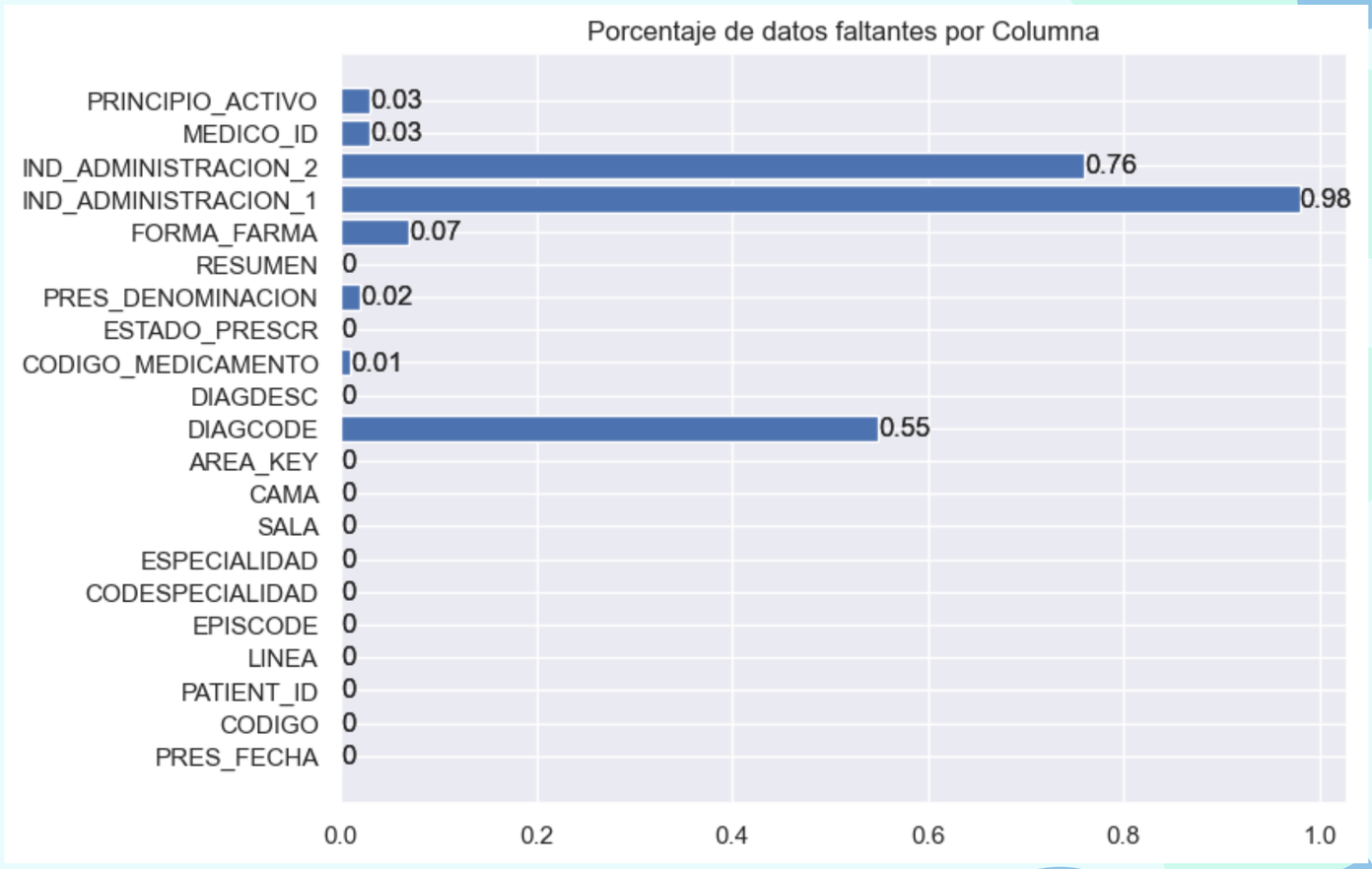
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Association for Computational Linguistics

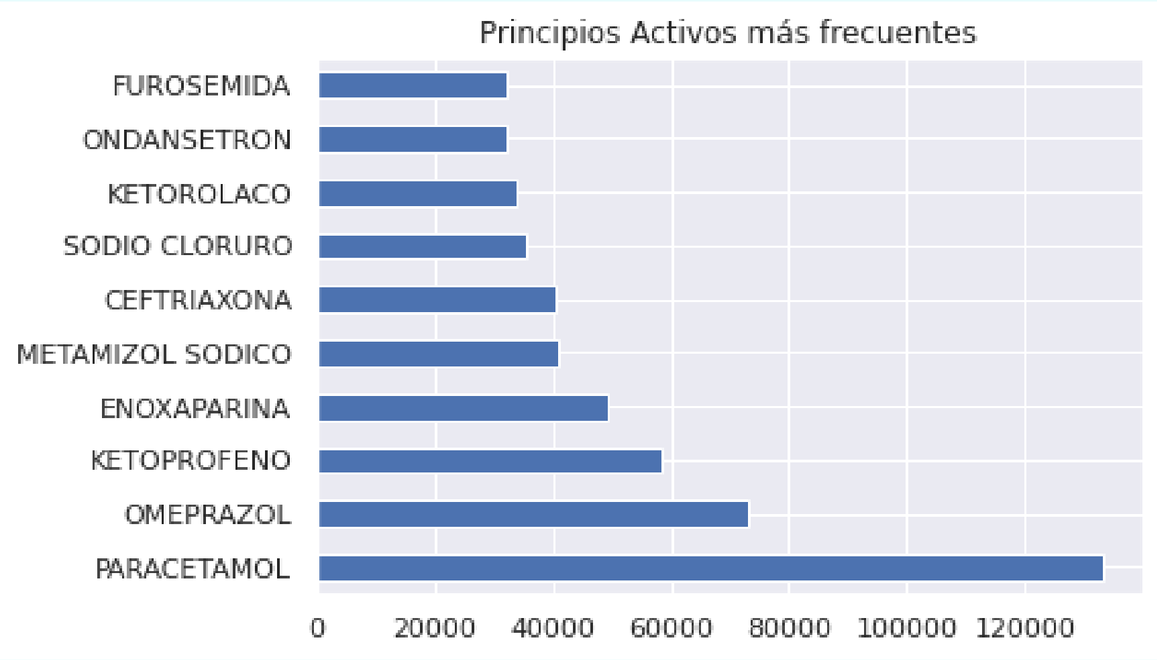
DATA FALTANTE Y ADICIONAL

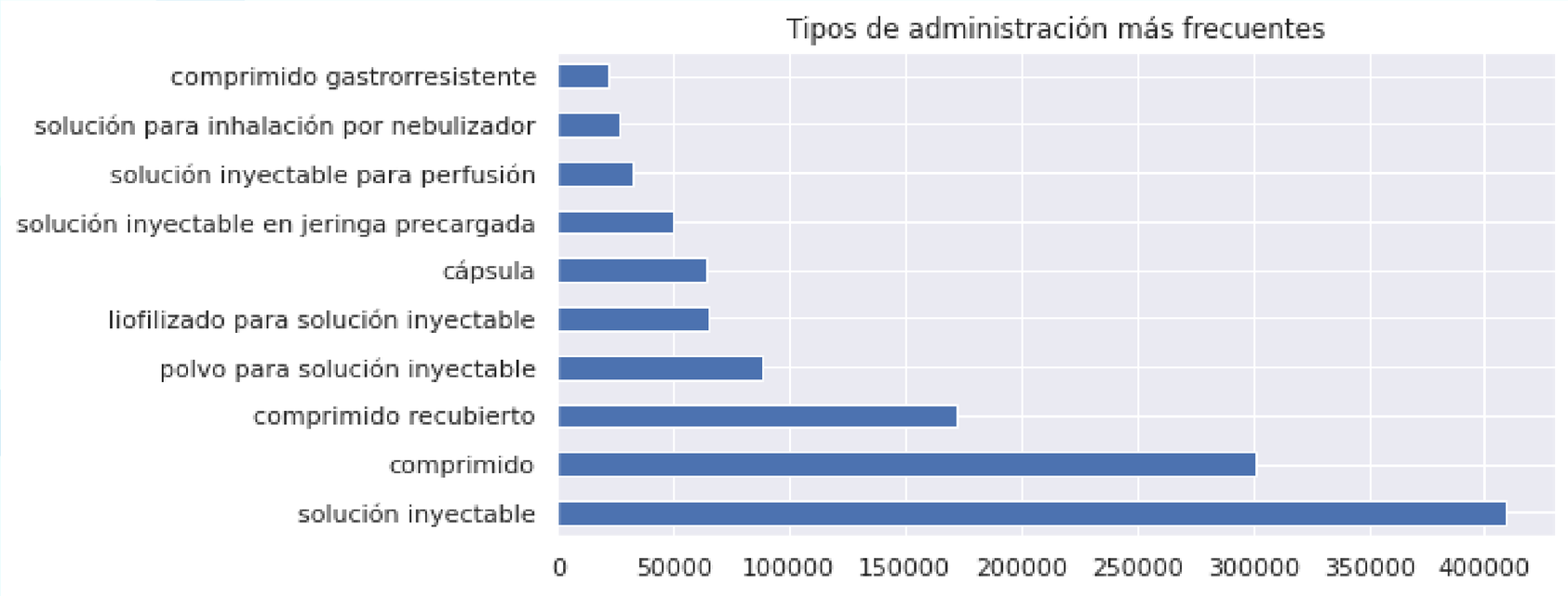
- Se agregó una nueva columna de Principios activos siguiendo el código HLF.
- Gran parte de los datos faltantes corresponden a los atributos Indicación de Administración 1 y 2, los cuales son utilizados para casos especiales de administración.
- Datos faltantes en códigos de medicamentos: **18379**
- Datos faltantes en Principio activo: **50437**

No todos los códigos de medicamento en las prescripciones tienen código HLF asociado.



VISUALIZACIÓN DE DATOS



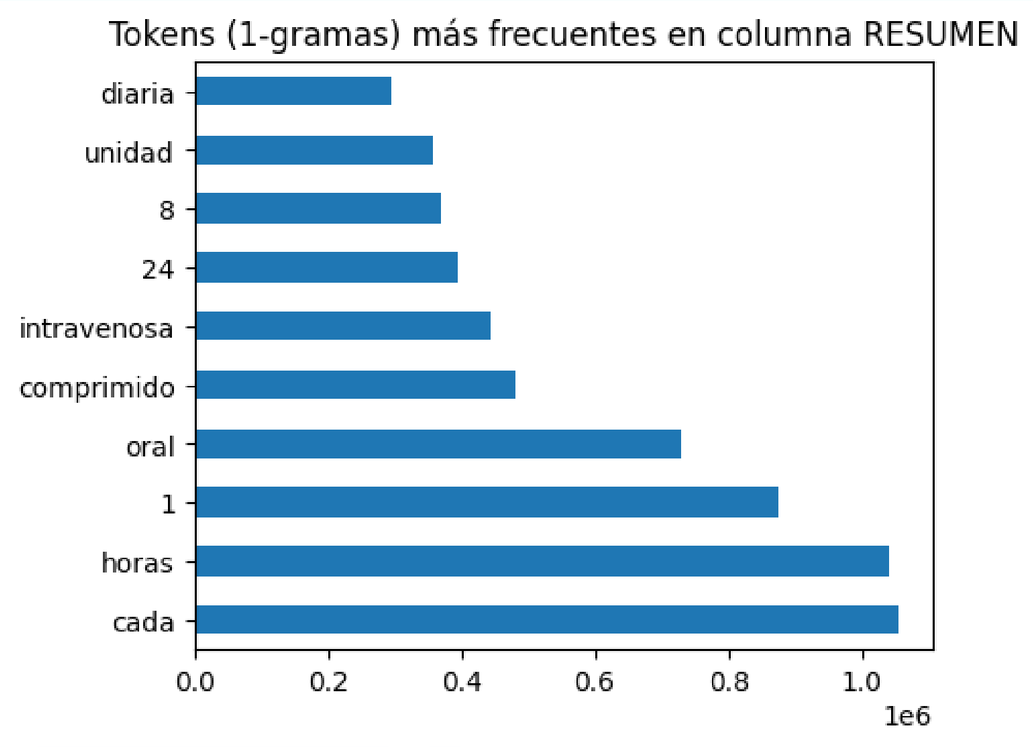


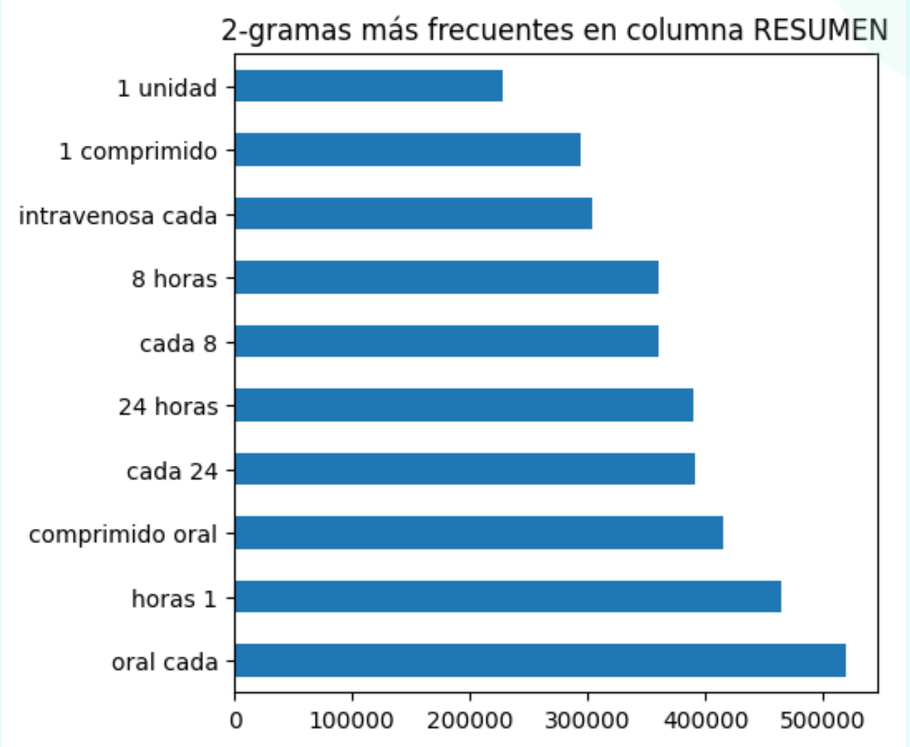
PALABRAS COMUNES EN TEXTO LIBRE: RESUMEN DE LA PRESCRIPCIÓN

Cantidad de valores nulos en columna: 1

Total de 1473779 filas duplicadas (96.5%)

Cantidad de valores únicos : 52777 (3.5%)





PLANTEAMIENTO DE SOLUCIÓN

Consideremos una entrada de receta médica

RANITIDINA 50 MG/2 ML SOLUCIÓN INYECTABLE

AMPOLLA 2 1 UNIDAD INTRAVENOSA cada 8 horas

PLANTEAMIENTO DE SOLUCIÓN

Esto bajo la forma que tenemos en el dataset se vería:

CODIGO MEDICAMENTO	PRES_DENOMINACION	RESUMEN
FAAA02007	RANITIDINA 50 MG SOLUCIÓN INYECTABLE	SOLUCIÓN INYECTABLE AMPOLLA 2 1 UNIDAD INTRAVENOSA cada 8 horas

PLANTEAMIENTO DE SOLUCIÓN

El problema se convierte en una clasificación para cada token en la secuencia.

RANITIDINA 50 MG/2 ML SOLUCIÓN INYECTABLE

AMPOLLA 2 1 UNIDAD INTRAVENOSA cada 8 horas

PLANTEAMIENTO DE SOLUCIÓN

El problema se convierte en una clasificación para cada token en la secuencia.

B-ACTVPRNCP	B-ADMIN	B-ADMIN
RANITIDINA	**50 MG/2 ML**	**SOLUCIÓN INYECTABLE**

I-ADMIN	I-ADMIN	I-PERIODICITY
AMPOLLA 2	**1 UNIDAD INTRAVENOSA**	**cada 8 horas**

PLANTEAMIENTO DE SOLUCIÓN

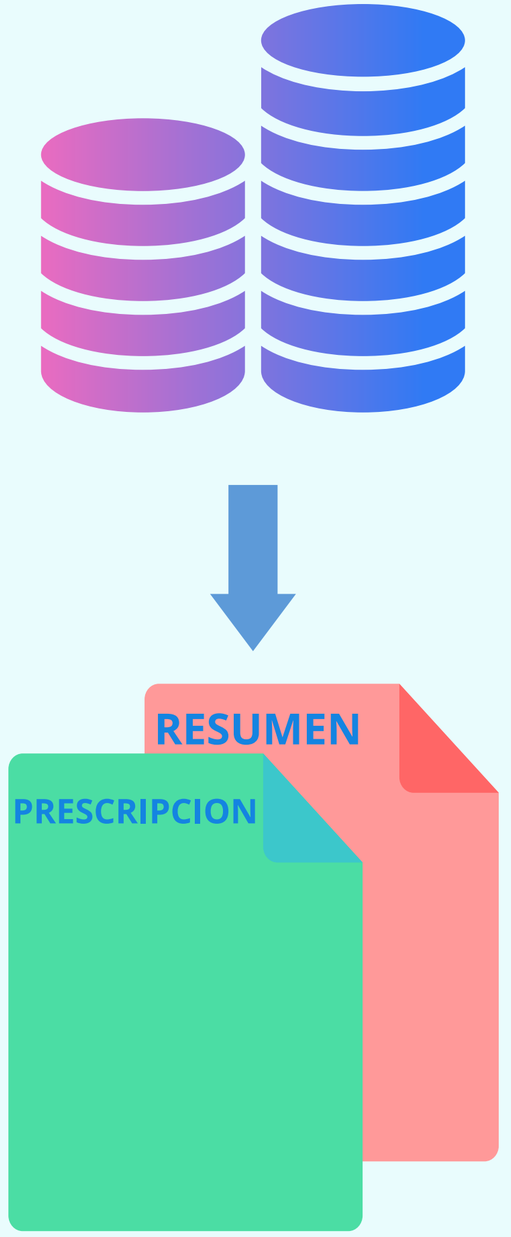
Es necesario ejecutar un proceso de tagging manual, para esto se definen entidades identificables dentro del dataset.

CODIGO MEDICAMENTO	PRES_DENOMINACION	RESUMEN
FAAA02007	RANITIDINA 50 MG SOLUCIÓN INYECTABLE	SOLUCIÓN INYECTABLE AMPOLLA 2 1 UNIDAD INTRAVENOSA cada 8 horas

RANITIDINA 50 MG/2 ML SOLUCIÓN INYECTABLE
AMPOLLA 2 1 UNIDAD INTRAVENOSA cada 8 horas

PREPARACIÓN DE LOS DATOS

- Se define el texto libre como las columnas: PRES_DENOMINACION y RESUMEN.
- Otras columnas útiles PRINCIPIO_ACTIVO y FORMA_FARMA
- Número de Ejemplos etiquetados se reduce a 108.049



PREPARACIÓN DE LOS DATOS

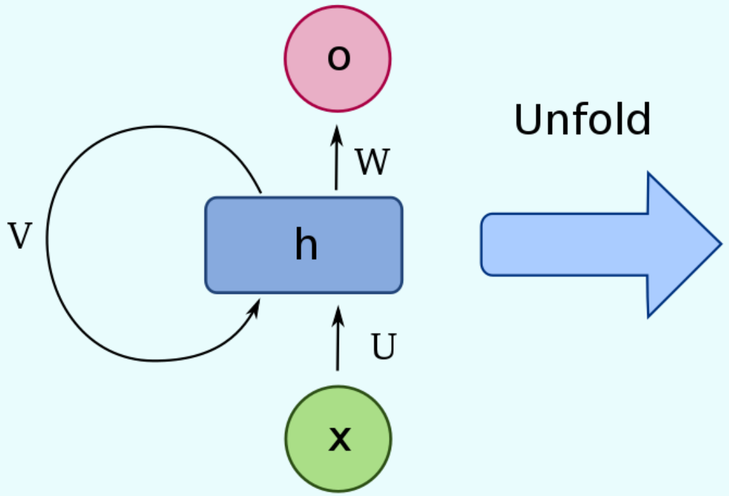
- Se etiquetan los datos a través de reglas.
- Definen 4 entidades:
  - ACTVPRNCP
  - ADMIN
  - PERIODICITY
  - DURATION.

PARACETAMOL	B-ACTVPRNCP
500	O
MG	O
COMPRIMIDO	B-ADMIN
1	O
COMPRIMIDO	B-ADMIN
ORAL	I-ADMIN
CADA	B-PERIODICITY
6	I-PERIODICITY
HORAS	I-PERIODICITY
DURANTE	B-DURATION
3	I-DURATION
DIAS	I-DURATION

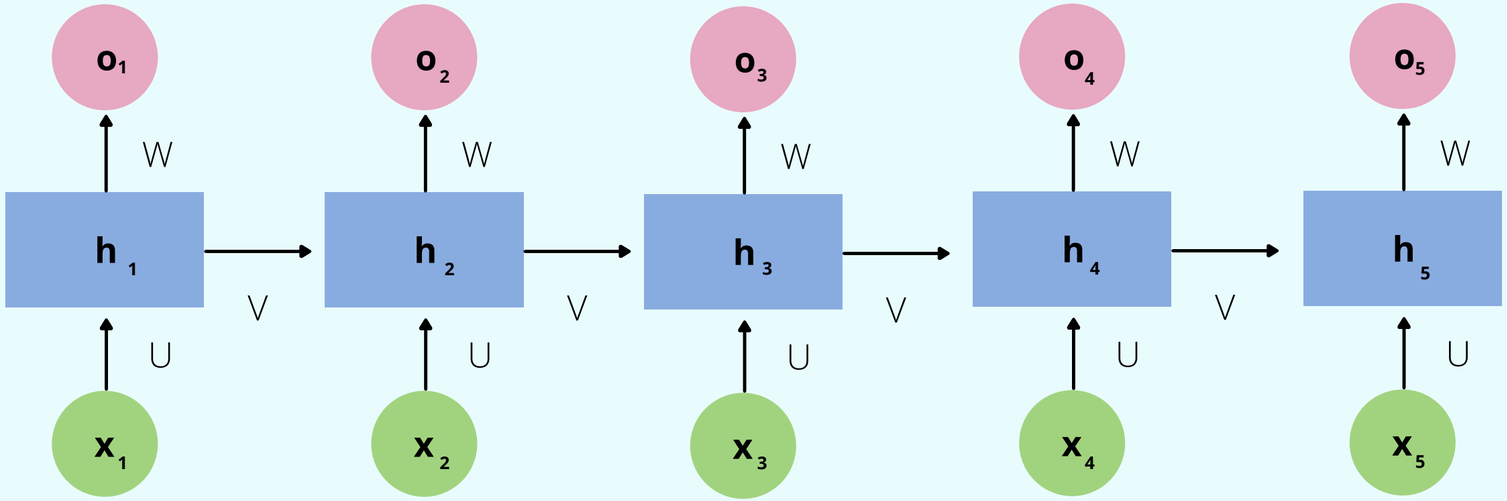
MODELO DE REFERENCIA

- Recurrent Neural Network (RNN)
- Una capa de embedding, 3 capas de LSTM y una capa lineal.
- A todas se les aplica dropout de 0.5.
- Entrada: vectores one-hot.
- Métricas a utilizar: recall, precision y puntuación F1.
- Debido a lo costoso del etiquetado, queda pendiente obtener los resultados.

MODELO DE REFERENCIA



Unfold



0.1	1.8	-0.6	-0.2	-1.3
-1.5	0.5	1.4	1.5	-0.5
...	...	...	...	...
-0.3	-0.9	0.8	0.3	0.2
1.2	-0.5	0.1	1.8	-0.6

RANITIDINA 50 MG SOLUCIÓN INYECTABLE

RANITIDINA 50 MG SOLUCIÓN INYECTABLE

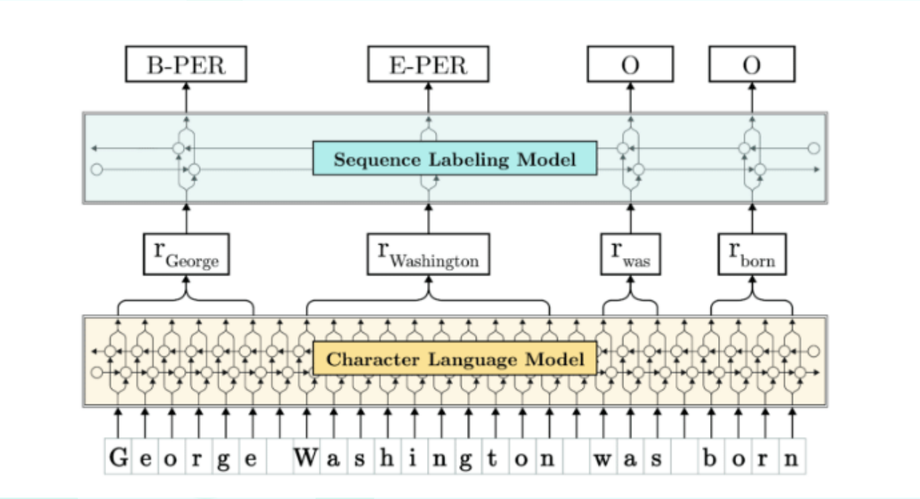
/4 CONCLUSIONES Y TRABAJO FUTURO

- Sería posible utilizar reglas para la detección de entidades.
- Se determinaron las columnas necesarias para el modelo de solución y las reglas con las cuales deben ser etiquetados los datos.
- El proceso de etiquetado conlleva una tarea ardua que requiere constante comprobación, actualización y tiempo.
- Una vez etiquetados los datos, se procede a la implementación de la solución.
- Luego, se comparará el modelo implementado con los resultados obtenidos con el baseline escogido según métricas escogidas.

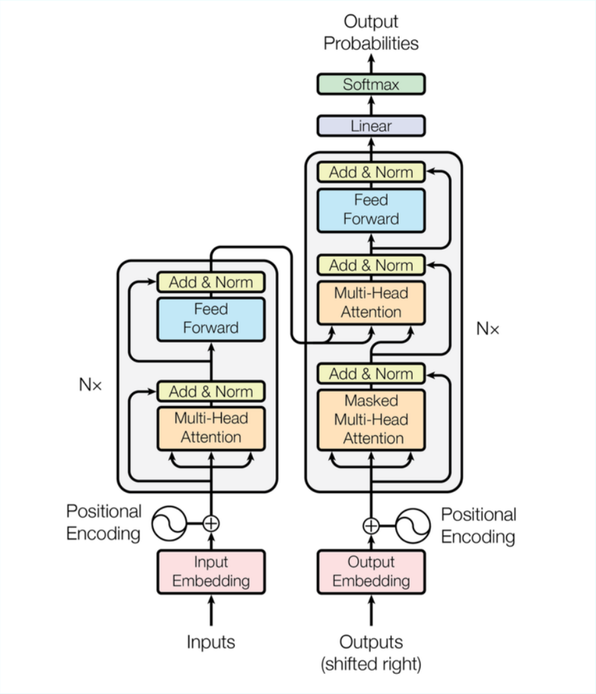
/4 CONCLUSIONES Y TRABAJO FUTURO

Otros modelos a considerar:

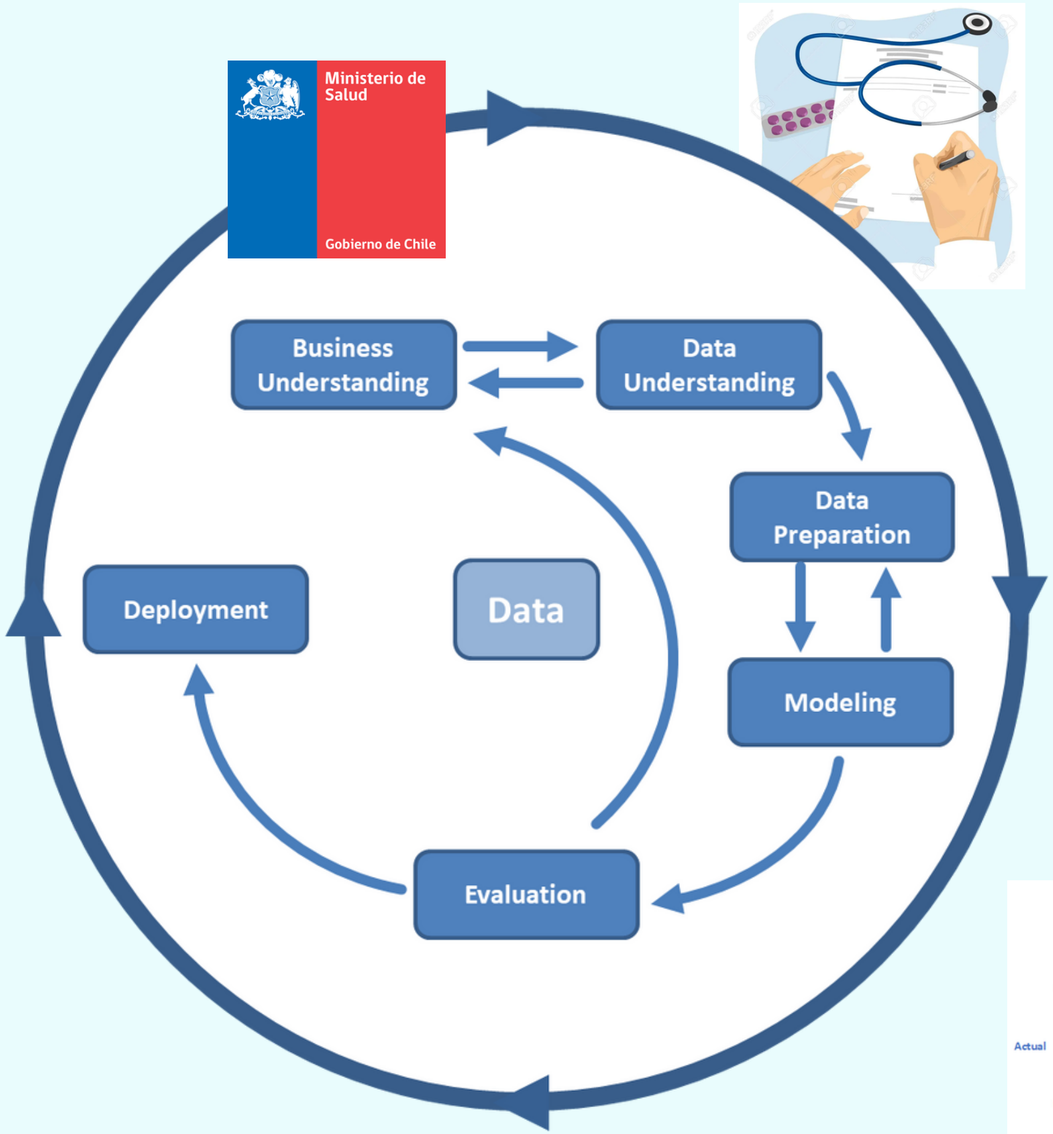
- **FLAIR embeddings**
  Usado por CMM-PLN

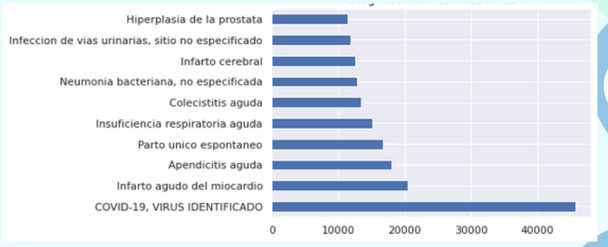


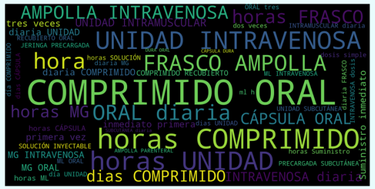
Transformers

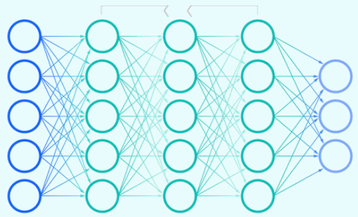


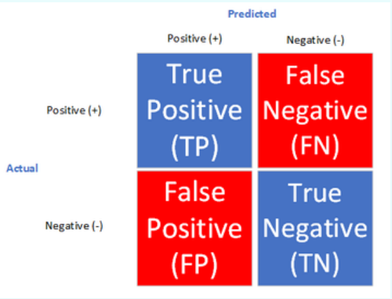
/4 PLAN DE TRABAJO

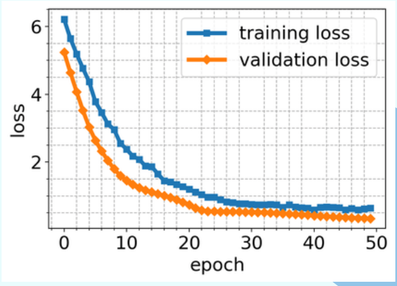












MDS Master of Data Science
Universidad de Chile

PRESENTACIÓN 3 MDS7201

PROYECTO DEFINIDO

ENTIDADES MINSAL

DANIEL CARMONA, MARTÍN SEPÚLVEDA,
MONSERRAT PRADO, CAMILO CARVAJAL

/5 REFERENCIAS

- Sang, E. F., De Meulder, F.

Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition.
In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 (142–147), 2003.

- Bose, P., Srinivasan, S., Sleeman IV, W. C., Palta, J., Kapoor, R., Ghosh, P.

A survey on recent named entity recognition and relationship extraction techniques on clinical texts.
In Applied Sciences (11(18), 8319.), 2021.

- Báez, P., Villena, F., Rojas, M., Durán, M., Dunstan, J. (2020, November).

The Chilean Waiting List Corpus: a new resource for clinical named entity recognition in Spanish.
InProceedings of the 3rd clinical natural language processing workshop (pp. 291-300)., 2020.

- Báez, P., Bravo-Marquez, F., Dunstan, J., Rojas, M., Villena, F.

Automatic Extraction of Nested Entities in Clinical Referrals in Spanish.
In ACM Transactions on Computing for Healthcare, (3(3), 1-22.) - 2022.

- Rojas, M., Dunstan, J., Villena, F.

Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing.
In Proceedings of the 4th Clinical Natural Language Processing Workshop, (pp. 87-92)., 2022.

- Jiang, M., Sanger, T., Liu, X.

Combining contextualized embeddings and prior knowledge for clinical named entity recognition: evaluation study.
In JMIR medical informatics, (7(4), e14850.) - 2019