INFORME DE CONTROL DE CALIDAD

Plan de impulso de las Tecnologías del Lenguaje

Obdulia Rabal

Ander Intxaurrondo

Martin Krallinger

2018

Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital, que no comparte necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores.

1 ÍNDICE

RESUMEN

Este documento presenta el control de calidad que se ha realizado sobre las anotaciones de menciones químicas, fármacos y biosimilares con relevancia terapéutica en informes clínicos mediante un estudio de consistencia inter-anotado, detallando estadísticas e menciones y entidades.

2 ESTADÍSTICA DEL CORPUS

Del total de 1000 casos clínicos que componen el corpus, se encontraron un total de 16504 sentencias, con un promedio de sentencias por caso clínico de 16.5 (mínimo y máximo número de sentencias en un documento de 3 y 66, respectivamente). En palabras, el corpus contiene un total de 396988 palabras, con un promedio de 396.2 palabras por caso clínico (mínimo y máximo número de palabras de 76 y 1280, respectivamente).

El número total de menciones etiquetadas es de 7704, distribuyéndose éstas en las siguientes cuatro clases, según el formato de normalización:

PROTEINAS: 3040

NO_NORMALIZABLES: 64

UNCLEAR: 182

NORMALIZABLES: 4418

Finalmente, la Figura 1 muestra la distribución de las 2439 menciones únicas atendiendo a su frecuencia:

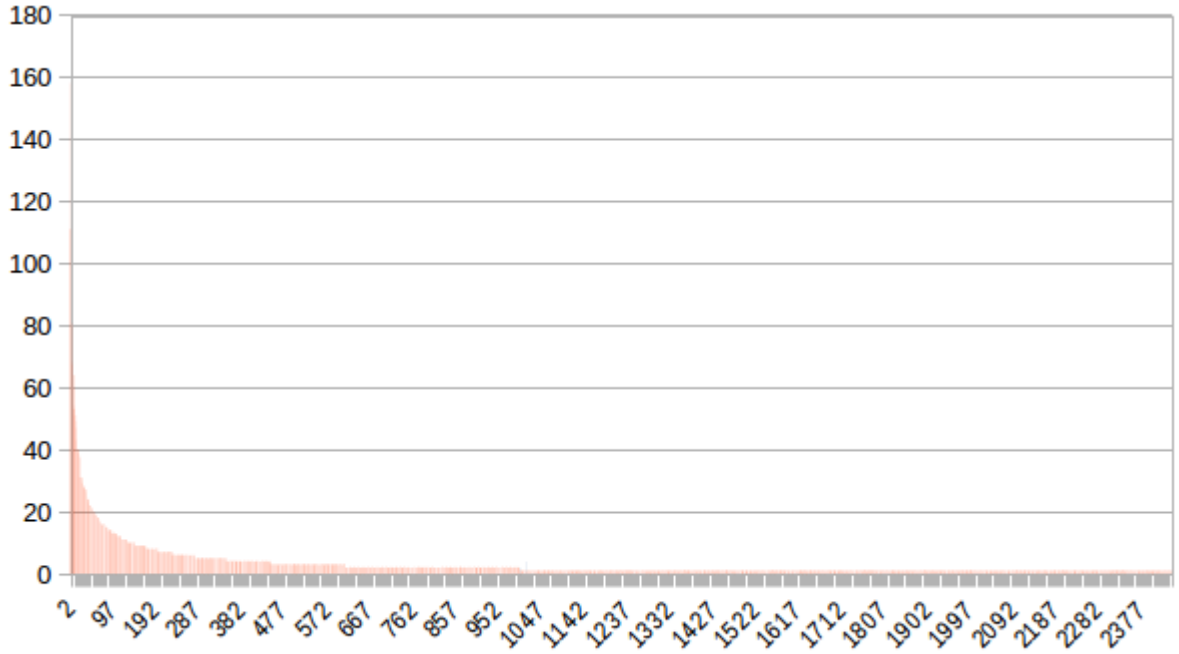

Figura 1: Distribución de las 2439 menciones únicas del corpus según su frecuencia

Entre los términos más comunes aparecen: creatinina, hemoglobina, corticoides y prednisona

3 CONSISTENCIA DE LAS ANOTACIONES

Los resultados de la consistencia de la anotación para este segundo set de documentos es:

Número total de anotaciones:	251
Anotaciones realizadas por el anotador 1:	217 (86%)
Anotaciones realizadas por el anotador 2:	201 (80%)
Coincidencia en las anotaciones:	235 (93%)
Coincidencia en las anotaciones (mismo tipo):	191 (76%)

Siendo un criterio aceptable (93%).

Tras revisar la consistencia de este segundo set y el origen de las discrepancias (7%), se realizó la revisión manual de todas las anotaciones de los 1000 textos clínicos para minimizar errores.