ISSN 1870-4069

Extracción de signos vitales, medidas antropométricas y fechas en expedientes médicos

Rodrigo Diaz-Moreno, Helena Gómez-Adorno,
Alejandro Martínez-Torres

Universidad Nacional Autónoma de México,
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
México

{rodrigo.diaz, helena.gomez}@iimas.unam.mx,
alejandromartinezt@ciencias.unam.mx

Resumen. La extracción de información de notas médicas es una tarea importante en el área de procesamiento de lenguaje natural. En este artículo presentamos métodos de extracción de información basado en reglas para la identificación de datos relevantes en las notas médicas como son: los signos vitales, los datos antropométricos y las fechas. Para la evaluación de los métodos desarrollados utilizamos un corpus de 198 notas médicas de pacientes de la Secretaría de Salud de la Ciudad de México. Los métodos desarrollados para la extracción de signos vitales y antropométricos logran extraer datos relevantes en la mayoría de las notas médicas del corpus. El método de extracción de fechas obtiene una precisión del 99 % en el corpus de evaluación.

Palabras clave: Notas médicas, extracción de información, signos vitales, datos antropométricos, identificación de fechas.

Extraction of Vital Signs, Anthropometric Measurements and Dates in Medical Records

Abstract. The extraction of information from medical notes is an important task in the area of natural language processing. In this article we present rule-based information extraction methods for the identification of relevant data in medical notes such as: vital signs, anthropometric data, and dates. For the evaluation of the methods developed, we used a corpus of 198 medical notes from patients from the Ministry of Health of Mexico City. The methods developed for the extraction of vital and anthropometric signs manage to extract relevant data in most of the medical notes of the corpus. The date extraction method obtains an accuracy of 99% in the evaluation corpus.

Keywords: Medical notes, extraction of information, vital signs, anthropometric data, identification of dates.

1. Introducción

La extracción de información de las notas médicas es importante debido a que estas poseen una gran cantidad de datos valiosos para análisis estadísticos [6, 2]. Esta tarea resulta complicada debido a que la forma en como está escrito el texto puede ser ambiguo y sobre todo que los autores tienen distintos estilos de escritura, por lo que los datos pueden variar en donde se encuentran.

En este trabajo no enfocaremos en identificar los signos vitales y los datos antropométricos. Los signos vitales son parámetros clínicos que reflejan el estado fisiológico del organismo humano, los cuales son:

- Temperatura: grado de calor conservado por el equilibrio entre el calor generado (termogésis) y el calor perdido (termólisis) por el organismo.
- Frecuencia respiratoria: el número de veces que una persona respira por minuto.
- Frecuencia cardiaca: el número de latidos por minuto.
- Presión arterial: es la fuerza que la sangre ejerce contra las paredes arteriales.
- Saturación de oxígeno: mide el porcentaje de hemoglobina encuadernada del oxígeno en la sangre.

Los datos antropométricos son mediciones técnicas sistematizadas que expresan, cuantitativamente, las dimensiones del cuerpo humano, las cuales son:

- Edad.
- Peso.
- Talla.
- Índice de Masa Corporal (IMC).

Por otro lado, la detección de fechas en notas médicas es de utilidad para ubicar en el tiempo los elementos detectados en la misma. Por ejemplo, los datos antropométricos y signos vitales pueden cambiar respecto al tiempo. De la misma manera, la identificación de fechas en las notas médicas servirá para ubicar en el tiempo otros datos del historial médico como: enfermedades, síntomas y medicamentos recetados.

Existen varias herramientas para la detección de fechas, sobre todo en el idioma inglés [7], sin embargo, al trabajar con notas médicas hechas en México (escritas en español), muchas de ellas necesitan ajustarse para detectar los formatos que fueron usados. Otro problema que se tuvo durante la extracción de fechas fue diferenciar entre un año y una cifra de 4 dígitos en un contexto médico.

Este trabajo está estructurado de la siguiente manera. En la Sección 2, describimos trabajos relacionados. En la Sección 3, presentamos una breve descripción del corpus de notas médicas utilizado para evaluar los métodos desarrollados. En la Sección 4, introducimos los métodos de extracción de signos vitales, datos antropométricos y fechas; así como las métricas de evaluación. Finalmente, en la Sección 5 presentamos los resultados obtenidos y en la Sección 6 presentamos las conclusiones y direcciones de trabajo futuro.

2. Trabajo relacionado

En la extracción de información en el dominio del procesamiento de lenguaje natural, se ha trabajado sobre todo en métodos de aprendizaje automático, como las maquinas de soporte vectorial (SVM) [2] el cual es un método que básicamente construye un clasificador binario utilizando muestras de entrenamiento intentando encontrar el hiperplano óptimo, que maximiza la distancia entre clases de la muestra de entrenamiento (llamados vectores de soporte) y de esta manera poder predecir si una palabra es la entidad de interés.

Así como redes neuronales convolucionales [4], que ocupan filtros para extraer características de los vectores de palabra y de esa manera poder realizar un reconocimiento de entidades nombradas. Con la ventaja de que procesan efectivamente grandes cantidades de datos de entrenamiento y ocupan menos parámetros para la tarea por lo que tienden a tener un menor sobreajuste. Consiguiendo resultados en el estado del arte.

A pesar del buen rendimiento que tienen estos sistemas tienen el inconveniente de que estos modelos no son interpretables, por lo que es muy difícil que se puedan mejorar con un humano con su conocimiento de ese dominio. El otro inconveniente es que se requiere una gran cantidad de recursos computacionales y datos etiquetados, es debido a eso que nuestro sistema considera expresiones regulares [1], los cuales son modelos basados en reglas, que tienen la ventaja de no requerir tantos datos y al ser un modelo más simple, es más fácil de ajustar para obtener un mejor rendimiento.

Adicionalmente, existen varias herramientas en la literatura que permiten la extracción de fechas en documentos escritos en lenguaje natural. Se han desarrollado bibliotecas, como date-detector basado en expresiones regulares, y otras herramientas en el lenguaje de programación Python, sin embargo fueron desarrolladas para textos en el idioma inglés [3].

También existen herramientas para el reconocimiento de entidades nombradas que pueden ser utilizadas para la extracción de fechas en textos escritos en español, como Stanza desarrollado por la universidad de Stanford [5].

El inconveniente de herramientas como Stanza es que no reconocen fechas en una variedad de formatos, muy probablemente debido a que estos formatos no se encontraban entre los datos con los que se realizó el entrenamiento de la herramienta. Es por esto que se desarrolló el presente trabajo basado en expresiones regulares que captura una amplia variedad de fechas en distintos formatos usados en el idioma español.

3. Corpus

Para el presente trabajo, SEDESA nos proporcionó un corpus de 98 expedientes médicos electrónicos de pacientes diagnosticados con el nuevo coronavirus SARS-CoV-2 (COVID). Por otro lado, se nos proporcionaron 100 expedientes médicos adicionales de pacientes con distintas enfermedades, en el mismo formato al de los expedientes médicos de pacientes con SARS-CoV-2. Dichos datos fueron proporcionados en un formato XML el cual venía organizado por secciones de las cuales se describen a continuación:

Table 1. Corpus de notas médicas COVID vs No COVID.

Tipo de Nota Médica	Número de Notas
COVID	98
No COVID	100

– Nombre y apellidos del paciente.
– Edad del paciente.
– Sexo del paciente.
– Estado y alcaldía.
– Fecha de ingreso.
– Fecha alta.
– Fecha hora registro nota.
– Nota médica (XML).
  Signos vitales: contiene el resumen de los signos vitales del paciente.
  Objetivo: contiene la descripción del estado actual del paciente y motivo de la consulta o revisión hospitalaria.
  Análisis: contiene la descripción del hallazgo del médico.
  Diagnóstico: describe el diagnóstico de la enfermedad del paciente.
  Plan de manejo: describe el tratamiento recetado al paciente, tanto de medicamentos como dieta, estudios necesarios, etc.

Es importante destacar que el objeto de estudio de este trabajo es el análisis de la nota médica, por lo tanto, cada sección del XML de la nota médica fue extraído para formar un solo documento por paciente. La tabla 1 muestra la cantidad de notas médicas existentes en el corpus por tipo de pacientes (COVID y No COVID).

Inicialmente el texto de las notas médicas no contenía ningún tipo de etiquetado, la única etiqueta que se tenía son las relacionadas con el paciente y el diagnóstico. Con la colaboración de tres expertos de SEDESA, se etiquetó de manera manual cada nota médica del corpus de pacientes COVID mediante la interfaz de una plataforma web de anotación de datos Dataturks[1]. A continuación se enuncian las características etiquetadas:

1. Síntomas, se identifican las palabras que contienen referencia a síntomas presentados por el paciente.
2. Comorbilidades, se identifican las palabras que hacen referencia a enfermedades previas del paciente.
3. Medicamentos, se identifican los medicamentos recetados al paciente.
4. Medicamentos previos, se identifican los medicamentos de base que el paciente está tomando actualmente .
5. Dosis, se identifica la dosis de los medicamentos (recetados y previos).

[1] https://docs.dataturks.com/

Fig. 1. Ejemplo de una nota médica etiquetada con características que se muestran en la figura.

6. Medidas (alternativas), identifica tratamientos alternativos como ozonoterapia, dieta especial, etc.
7. Signos vitales, se identifican los signos vitales como frecuencia respiratoria (FR), frecuencia. cardiaca (FC), saturación de oxígeno (SATO2), tensión arterial sistólica (TS) y diastólica (TD) y temperatura.
8. Datos antropométricos, se marcan el peso y la altura del paciente.

La Figura 1 muestra el ejemplo de una nota médica etiquetada con algunas de las características descritas previamente. Es importante destacar que no todos los expediente contaban con todas las características.

4. Metodología

Para resolver la extracción de signos vitales, datos antropométricos y fechas, se desarrolla un sistema basado en expresiones regulares implementado en el lenguaje de programación python, en donde en cada signo vital, en cada dato antropométrico y en cada formato específico de fecha se busca una expresión regular que contemple todas sus posibles variantes.

4.1. Pre-procesamiento

Los datos de las notas clínicas vienen en un formato XML, el cual tiene algunas inconsistencias en las etiquetas, por lo que se tuvo que corregir una vez identificando la estructura del XML y de esta manera evitar en lo posible pérdidas de información.

Nota Médica <paciente> 7000000000486834 06/09/1962 515351 <paciente> Mujer <doctor> HOSPITAL ABC <doctor> Signos Vitales 21/07/2020 06:52: Temperatura: 36.4 / Frecuencia cardiaca - ADL: 82.0 / Frecuencia respiratoria - ADL: 20.0 / SaO2: 93.0 / Otras constantes de hoy:Tensión Arterial Sistólica - ADL: 125.0 / Tensión Arterial Diastólica - ADL: 80.0 / Tensión Arterial Media - ADL: 95.0 / Síntomas Se trata de <paciente> de 57 años de edad, con Obesidad grado I, sin otros antecedentes de importancia para la enfermedad actual. Niega dolor torácico, Negó sintomatología urinaria digestiva. La paciente niega la presencia de disnea. Objetivo Mujer de edad aparente igual la cronológica, orientada en tiempo, persona, lugar circunstancia, alerta. Coloración normal de mucosas. Estado de hidratación adecuado, con aporte de oxígeno suplementario por puntas nasales 0.5lpm . Saturando 96%.

Fig. 2. Ejemplo de nota médica preprocesada.

Una vez hecho esto se procedió a estructurar el XML a un formato de renglones y columnas para la facilidad de la extracción de datos antropométricos y signos vitales.

Posteriormente, se hizo una limpieza de los datos, ya que estos tenían errores de codificación por lo que acentos o caracteres especiales en las notas médicas venían representados de manera distinta por lo que generaban ruido en el texto. Se extrajeron los datos de la creación de la nota médica para el uso posterior de detección de fechas y el resto de la información de cada nota se agrupo en un solo texto.

Finalmente, el texto de la nota médica fue anonimizado para el etiquetado mencionado previamente en la sección Corpus, el cual se usa para evaluar la extracción de signos vitales y medidas antropométricas. En la Figura 2 se muestra un ejemplo del texto que con el que se trabaja para la extracción de signos vitales, medidas antropométricas y fechas.

4.2. Signos vitales

Temperatura: Se extrae la temperatura de la nota médica a partir de una expresión regular que busca una palabra que empiece por 'temp' seguida de hasta 7 caracteres que puede seguirle de la palabra 'axilar', o la palabra fiebre, seguida de un número que contempla decimales de hasta dos cifras. Quedando la expresión regular siguiente:

”(?:temp\w{,7}(?: axilar)?|fiebre \w{,2}) (\d{,2}(?:\.\d{,2})?)”.

Frecuencia cardiaca: Para la frecuencia cardiaca se ocupa una expresión regular que busca el patrón: Una palabra que empiece por 'cardiaca' seguida de hasta 7 caracteres o la palabra 'fc', seguida de un número que contempla decimales de hasta dos cifras. Quedando la expresión regular siguiente:

”(?:cardiaca.{,6}|fc) (\d{,3}(?:\.\d{,2})?)”.

Frecuencia respiratoria: Se extrae la frecuencia respiratoria de la nota médica a partir de una expresión regular que busca una palabra que empiece por 'respiratoria' seguida de hasta 6 caracteres o la palabra 'fr', seguida de un número que contempla decimales de hasta dos cifras. Quedando la expresión regular siguiente:

”(?:fr|respiratoria.{,6}) (\d{,3}(?:\.\d{,2})?)”.

Saturación de oxígeno: Para el caso de la extracción de la saturación de oxígeno se busca una palabra que empiece por 'sat' seguida de hasta 6 caracteres o la palabra 'sao2', seguida de un número que contempla decimales de hasta dos cifras. Este es el único caso que se presenta en las notas médicas en donde hay más de una ocurrencia por nota, por lo que se consideran todas las coincidencias y se guarda en una lista. Quedando la expresión regular siguiente:

"(?:sao2|sat\w{0,7}) (\d{,3}(?:\.\d{,2})?)".

Presión sistólica: Se extrae la presión sistólica de la nota médica a partir de una expresión regular que busca el patrón: Una palabra que empiece por 'sist' seguida de hasta 7 caracteres o la palabra 'ta', seguida de un número que contempla decimales de hasta dos cifras. Quedando la expresión regular siguiente:

"(?:sist.{,7} .{,6}|ta) (\d{,3}(?:\.\d{,2})?)".

Presión diastólica: Para la extracción de la presión diastólica se busca el patrón: Una palabra que empiece por 'diast' seguida de hasta 7 caracteres, seguida de un número o una palabra que empiece por 'ta' seguida de un número y una diagonal seguida de otro número para quedarnos con este segundo número que corresponderá a la presión diastólica. Quedando la expresión regular siguiente:

"diast.{,7} .{,6} (\d{,3}(?:\.\d{,2})?)|ta \d{,3}(?:\d{,2})?/(\d{,3}(?:\.\d{,2})?)".

4.3. Datos antropométricos

Peso: Se extrae el peso de la nota médica a partir de una expresión regular que busca el patrón: Un número que contempla decimales de hasta dos cifras seguidas de la palabra 'kg', o una palabra que empiece por 'kg' con o sin paréntesis seguida de un número que contempla decimales con hasta dos cifras. Quedando la expresión regular siguiente:

"(\d{,3}(?:\.\d{,2})?) kg|\(?kg\)? (\d{,3}(?:\.\d{,2})?)".

Talla: Para la talla se busca a partir de una expresión regular el patrón: Un número que contempla decimales de hasta dos cifras seguidas de la palabra 'cm' o una palabra que empiece por 'cm' con o sin paréntesis seguida de un número que contempla decimales con hasta dos cifras. Cuidando que cuando haya una ocurrencia que sea menor a 2, esta se multiplique por 100 debido a los casos escritos en metros, dejando todo en centímetros. Quedando la expresión regular siguiente:

"(\d{,3}(?:\.\d{,2})?) cm|\(?cm\)? (\d{,3}(?:\.\d{,2})?)".

IMC: Para el IMC se extrae con una expresión regular el patrón: Una palabra que empiece por 'imc' seguida de un número que contempla decimales de hasta 2 dígitos. Para las notas médicas donde no haya una ocurrencia y se tenga la información tanto de la talla como del peso, se calcula manualmente el índice de masa corporal. Siendo la fórmula:

$$IMC = \frac{Peso}{Estatura^2}.$$

Quedando la expresión regular siguiente:

"imc (\d{,3}(?:\.\d{,2})?)".

4.4. Fechas

Primero, para obtener un conjunto de fechas como objetivo a extraer de las notas médicas, se extrajo todas las posibles fechas detectadas con herramientas previamente existentes como date-detector y Stanza, además de expresiones regulares que atrapan toda expresión que contienen el nombre de un mes.

Una vez obtenido el conjunto de posibles fechas se descartaron manualmente todas las que no lo eran. Teniendo un total de 428 fechas que se deseaban detectar. Posteriormente se agruparon las fechas respecto a los diferentes formatos que presentan.

Los formatos que se detectaron fueron:

- DD/MM/AAAA (con variaciones de longitud y separadores).
- Mes Día Año (ej. Septiembre 24, 2020).
- Día Mes Año (ej. 24 de septiembre del 2020).
- Mes Año (ej. septiembre 2020).
- Año (ej. 2020).
- Referencias al año (ej. el año pasado).

Definiciones y separadores: Dado los formatos mencionados, hay espacio para variaciones en la escritura de los días, mese y años. Un día, al igual que un mes, puede variar entre uno y dos dígitos. Un año entre 2 y 4 dígitos. Y un mes, además de ser un número, puede ser una palabra, una palabra con errores ortográficos o una palabra abreviada.

De igual forma estos datos son distinguidos unos de otros por separadores, conformados por una variedad de símbolos y palabras, espacios o signos de puntuación. Las expresiones regulares deben ser lo suficientemente robustas para cubrir todos estos casos.

Diferenciación entre un año y un número de 4 dígitos: Obtener un número de 4 dígitos no es suficiente para juzgar si el número en cuestión es un año o no. Verificar si un número se encuentra en cierto rango no es suficiente, ya que se encontraron ejemplos, dentro de las notas médicas, donde un mismo número (2009) es usado como una cantidad y como un año.

Por ello se identificaron palabras, expresiones o signos de puntuación que preceden a la mención de un año. Se encontró en el contexto de notas médicas que seguido de un signo de puntuación se puede presentar un año, a diferencia de una cantidad. También se encontró que las palabras que anteceden a un año, pero no una cantidad son: en, el, del, de o año.

Referencias al año: Dentro del corpus presentado se encontraron fechas que no presentaban el año, ya que se asumía que era el 2020, año de escritura de la nota. A su vez, se encontraron referencias informales al año, tales como "este año" o "el año pasado". Para poder obtener la fecha de referencias informales se capturaron las expresiones usadas y se relacionaron con la fecha de escritura de la nota médica.

Extracción de fechas y sus componentes: Se crean las expresiones regulares para definir los distintos componentes de una fecha: día, mes dígito, mes palabra, año, separadores símbolos, separadores palabras, expresión informal para año y antecedentes al año. Posteriormente, se crean las expresiones regulares que juntan los componentes en los distintos formatos.

Al capturar la fecha, también se obtienen los datos del día, el mes y el año. En caso que la fecha no cuente con el dato del día, se asume que ocurrió el primero del mes. Si la fecha no cuenta con mes, se asume el primer mes del año, enero. Y dado el caso en que no se cuente con año, se usa el año de la creación de la nota médica como se mencionó anteriormente. Esta información se guarda para un futuro uso y también se crea una fecha en un formato estándar.

5. Resultados

5.1. Evaluación de los métodos

En la Tabla 2 presentamos un resumen de la extracción en notas médicas de pacientes COVID y no COVID, siendo la primera columna la entidad que extrajimos, la segunda y tercera columna su correspondiente número de notas médicas en donde hubo una ocurrencia para pacientes COVID y no COVID. Es importante destacar que las notas médicas en las que no hubo una ocurrencia de una entidad es debido a que esta no contaba con la entidad a extraer.

Se puede observar en la tercera columna que la cantidad de datos faltantes es mucho mayor a la de pacientes COVID, esto debido a que las notas son más variadas y no todas cuentan con todas las entidades. Los casos más extremos son la presión sistólica y diastólica, en donde no hubo una sola ocurrencia en estas notas.

En la Tabla 3 se presenta el puntaje de la extracción de los signos vitales, obtenido a partir de un etiquetado manual hecha por expertos, notando que todas las entidades tienen un buen puntaje.

Siendo solo el caso de la saturación de oxígeno el más bajo, esto debido a que en las notas médicas se hace una mención de que se tiene que monitorear que la saturación de oxígeno no tiene que bajar de un valor y en el etiquetado no consideran ese valor debido a que no es el que presenta el paciente. No pudiéndose evaluar los datos antropométricos debido a que no se contaba con datos etiquetado para ello.

Table 2. Resumen de la extracción notas médicas.

Dato a extraer	Datos pacientes COVID	Datos pacientes no COVID
Temperatura	96	54
F. Cardíaca	95	52
F. Respiratoria	95	50
Saturación de Oxígeno	91	53
P. Sistólica	81	0
P. Diastólica	81	0
Peso	20	64
Talla	13	54
IMC	14	54

Table 3. Puntaje de la extracción de los signos vitales.

Entidad	Precisión	Recall
Temperatura	1.0	1.0
F. Cardíaca	1.0	1.0
F. Respiratoria	1.0	1.0
Saturación de Oxígeno	0.9565	1.0
P. Sistólica	1.0	1.0
P. Diastólica	1.0	1.0

5.2. Signos vitales

La Figura 3 presenta los datos que se extrajeron de los signos vitales en forma de diagrama de caja, esto se hace para poder visualizar que los datos estas dentro del rango que es de esperarse y que se encuentran en la misma escala.

Notando que la temperatura tuvo un mínimo de 36°C y máximo de 37.75°C, lo cual se encuentra dentro del rango de la temperatura corporal que va de 36 a 40°C.

De igual forma se observa que la frecuencia cardiaca tuvo un mínimo de 28 y un máximo de 143 latidos por minuto.

Para la frecuencia respiratoria se observa que su rango va de 14 a un máximo de 40 latidos por minuto, notandose que hay una gran cantidad de datos atípicos debido a que las notas médicas de pacientes no COVID son en su mayoría niños, los cuales tienen una frecuencia respiratoria mayor.

Continuamos notando que la saturación de oxígeno se encuentra dentro del rango que es de esperarse, siendo el mínimo 83% y el máximo de 100%. Para la presión arterial se observa que tanto la presión sistólica como la diastólica se encuentran dentro del rango que es de esperarse siendo el mínimo de 81 y 50 mm Hg y máximo de 142 y 91 mm Hg respectivamente.

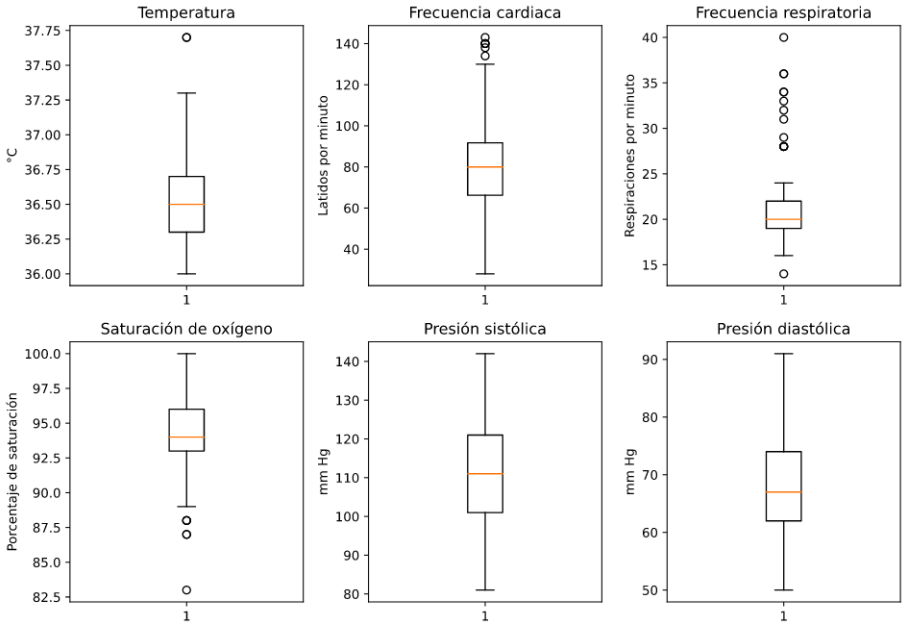

Fig. 3. Signos vitales.

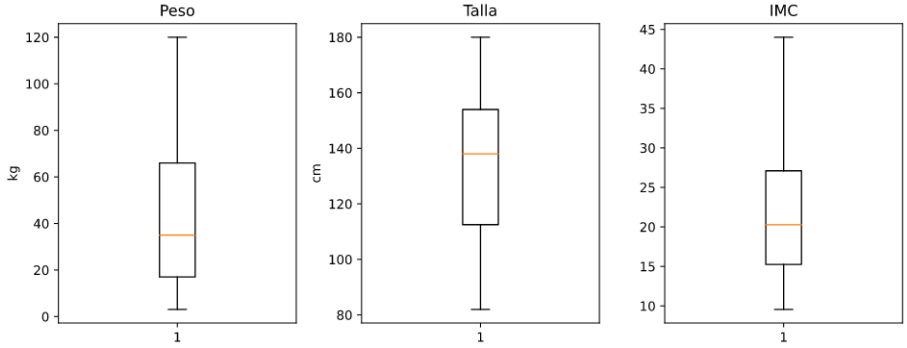

Fig. 4. Datos antropométricos.

5.3. Datos antropométricos

En la Figura 4, se hace un diagrama de caja para los datos antropométricos extraídos, de igual manera para visualizar que los datos extraídos estén en el rango que es de esperarse, observando que el peso tiene un mínimo de 3 kg y máximo de 120 kg, siendo importante mencionar que el mínimo se debe a que las notas médicas de pacientes no COVID son en su mayoría niños.

De igual manera para la talla se observa un mínimo de 82 y un máximo de 180 cm. Por ultimo el IMC tuvo un mínimo de 9.56 y un máximo de 44.

Table 4. Ejemplo de diferentes formatos de fechas extraídos.

Ejemplos de fechas extraídas
17.07.2020
01/03/1976
19.07.20
18 de junio de este año
19 de julio de 2020
abril 2019
julio de 2020
en 1994
: 2009

5.4. Fechas

Con el método propuesto se lograron extraer diferentes tipos de formatos de fechas, en la Tabla 4 se pueden apreciar algunos ejemplos de fechas extraídas.

Las fechas detectadas por nuestro método cubren por completo el conjunto objetivo de 428 fechas, además detecta otras fechas que no estaban contempladas.

Logramos extraer un total de 852 posibles fechas, de las cuales solo hubo 7 falsos positivos, es decir 7 elementos que no eran fechas. Esto se traduce a una precisión del 99.17% y un puntaje F1 de 99.49%.

6. Conclusiones

Los signos vitales y datos antropométricos tienen ciertos patrones en las notas médicas que pueden ser encontrados con una correcta exploración y teniendo la suficiente variedad de notas médicas, por lo que nuestro sistema basado en reglas funciona adecuadamente. Pudiendo extraer los datos correspondientes y validar que efectivamente están correctos.

Este sistema al estructurar la información de los signos vitales y datos antropométricos de las notas médicas, puede ser ocupado de base para una posterior explotación de la información encontrada.

Para la correcta extracción de fechas es necesario detectar todo posible formato utilizado para denotar una fecha.

La variedad de fechas presentadas y su clasificación posterior fue lo que hizo posible la creación de expresiones regulares capaces de realizar la tarea.

Como trabajo futuro se puede probar este sistema con un etiquetado que permita automatizar los resultados de los signos vitales y datos antropométricos, así como con nuevas notas médicas para volver más robusta las expresiones regulares.

En un futuro, el trabajo realizado puede ser utilizado para definir el alcance de cada fecha en la nota médica. De tal forma que se pueda ubicar en el tiempo otros datos tales como signos vitales y datos antropométricos.

References

1. Aguirre-Ojea, F., Manzotti, M., Díaz-Maffini, M.: Extracción automática de signos vitales en las evoluciones. In: Proceedings of IX Reunión Red Latinoamericana y del Caribe para el Fortalecimiento de los Sistemas de Información de Salud, pp. 1 (2019)
2. Ananiadou, S., Kell, D. B., Tsujii, J. I.: Text mining and its potential applications in systems biology. Trends in Biotechnology, vol. 24, no. 12, pp. 571–579 (2006) doi: 10.1016/j.tibtech.2006.10.002
3. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: Analyzing text with the natural language toolkit (2009)
4. Gavrilov, D., Gusev, A., Korsakov, I., Novitsky, R., Serova, L.: Feature extraction method from electronic health records in Russia. In: Proceedings of Conference of Open Innovations Association, FRUCT, pp. 497–500. No. 26 (2020)
5. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C. D.: Stanza: A Python natural language processing toolkit for many human languages (2020) doi: 10.48550/arXiv.2003.07082
6. Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., Liu, H.: Clinical information extraction applications: A literature review. Journal of biomedical informatics, vol. 77, pp. 34–49 (2018) doi: 10.1016/j.jbi.2017.11.011
7. Zhou, G., Su, J.: Named entity recognition using an HMM-based chunk tagger. In: Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pp. 473–480 (2002)