ISSN 1870-4069

Distancia de Levenshtein para anonimización de notas médicas y detección de comorbilidades

Alejandro Martínez-Torres[1], Helena Gómez-Adorno[2]

[1] Universidad Nacional Autónoma de México,
Facultad de Ciencias,
México

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

alejandromartinezt@ciencias.unam.mx,
helena.gomez@iimas.unam.mx

Resumen. En este trabajo se anonimizaron notas médicas y posteriormente se realizó un proceso de extracción de información relacionado a las comorbilidades que presentan los pacientes. Para la evaluación de los métodos desarrollados se utilizó un corpus de 98 notas médicas de pacientes diagnosticados con el nuevo SARS-CoV-2, la información fue proporcionada por la Secretaría de la Salud de la Ciudad de México. Logramos anonimizar en su totalidad la notas médicas y el método que desarrollamos extrajo el 96 % de las comorbilidades presentes en las notas médicas. Ambos resultado se lograron haciendo uso de la distancia de Levenshtein y la metodología creada puede ser usada para distintas tareas de la misma índole.

Palabras clave: Notas médicas, extracción de información, distancia de Levenshtein, anonimización de documentos, detección de comorbilidades.

Levenshtein Distance for Anonymization of Medical Notes and Detection of Comorbidities

Abstract. In this work, medical notes were anonymized and subsequently an information extraction process related to the comorbidities presented by the patients was carried out. For the evaluation of the developed methods, a corpus of 98 medical notes of patients diagnosed with the new SARS-CoV-2 was used, the information was provided by the Ministry of Health of Mexico City. We managed to completely anonymize the medical notes and the method we developed extracted 96% of the comorbidities present in the medical notes. Both results were achieved using the Levenshtein distance and the created methodology can be used for different tasks of the same nature.

Keywords: Medical notes, information extraction, Levenshtein distance, anonymization of documents, detection of comorbidities.

1. Introducción

Para la realización de este trabajo se elaboraron métodos para anonimizar los datos del paciente y extraer un listado de las posibles comorbilidades mencionadas en una nota médica.

Las notas médicas, en las que se basó el presente trabajo, fueron proporcionadas por la Secretaría de Salud de la Ciudad de México, quien solicitó la completa anonimidad de los pacientes a los que se refiere en las notas médicas. Por ello, anonimizar los datos del paciente es esencial para el uso, experimentación y etiquetado manual en notas médicas.

También es importante remarcar que al momento de anonimizar la nota médica se desea conservar todo dato que no sea parte del nombre, ya que puede poseer información relevante para futuros procesamientos.

Los expedientes médicos se almacenan en formato textual, por lo cual datos de alta relevancia no siempre se encuentran disponibles para el cómputo estadístico. En el ámbito médico, las comorbilidades son altamente relevantes, ya que predisponen a más enfermedades y son un factor de riesgo importante.

Las comorbilidades se refieren a enfermedades o trastornos secundarios que afligen al paciente, además de la enfermedad primaria por la cual se realizó la consulta médica [4]. Para el caso de la detección de comorbilidades se usó como referencia el listado del CIE-10 [3] de enfermedades y problemas relacionados con la salud, tomándolos como las posibles comorbilidades.

El presente trabajo desarrolla un método para la detección de comorbilidades sin necesidad de datos etiquetados. Se hace uso de uno de los métodos más sencillos, pero efectivo, para el etiquetado de términos específicos, la búsqueda en un texto de las ocurrencias de los términos especificados con anterioridad en una lista.

Aunque existen una gran variedad de métodos para la búsqueda de palabras o frases en un texto [8], la mayoría de estos no permiten errores en la escritura de la palabra.

La distancia de Levenshtein [2] es un método que sirve para medir la distancia entre dos palabras respecto a su escritura y es comúnmente usado para corregir errores de ortografía [1]. Sin embargo, comparar la distancia de Levenshtein de un gran número de palabras es relativamente costoso computacionalmente hablando.

Este trabajo está estructurado de la siguiente manera. En la Sección 2, se describen trabajos relacionados. En la Sección 3, se presenta una breve descripción del corpus de notas médicas utilizado para evaluar los métodos desarrollados. En la Sección 4, se presentan los métodos desarrollados para la anonimización de las notas médicas y la detección de comorbilidades. En la Sección 5 se presentan los resultados obtenidos. Y finalmente, en la Sección 6 presentamos las conclusiones y trabajo futuro.

2. Trabajo relacionado

En el reconocimiento de entidades nombradas, específicamente, la detección y etiquetado de nombres en el idioma español, el proyecto de Stanza, realizado por la universidad de Stanford [5], es uno de los que se encuentra más a la vanguardia. Sin embargo, muchos nombres comunes en México no son identificados como entidades.

Esto en su mayoría se debe a errores en la escritura del nombre, nombres poco comunes en el entrono que se entrenó el modelo y palabras que cuentan con más de un significado. Lamentablemente, para este trabajo no se cuenta con la cantidad necesaria de datos para reentrenar un modelo de este tipo. Por otro lado, dado que se cuentan con los datos del paciente de cada nota médica, pudimos usar métodos más tradicionales para la detección del nombre del paciente en la nota médica y posteriormente su anonimización.

En el área de la detección de comorbilidades, los trabajos previos cuentan normalmente con una gran cantidad de datos etiquetados, lo cual permite una variedad más amplia a procesos a realizar como lo son el uso de los vectores de palabras (más conocidos como *embeddings*, en inglés) [10].

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). Dichos datos fueron proporcionados en un formato XML el cual venía organizado por secciones de las cuales se describen a continuación:

- 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. En la Figura 1 se muestra un ejemplo del una nota médica ya anonimizado por cuestiones de confidencialidad.

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. Después de animizar las notas médicas, 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*[3]. A continuación se enuncian las características etiquetadas:

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

Nota Médica <paciente> [redacted] 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. 1. Ejemplo de nota médica preprocesada.

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).
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 2 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.

Para definir las posibles comorbilidades se usó el listado del CIE-10 [3] de enfermedades y problemas relacionados a la salud. El listado se extrajo de la página oficial del CIE-10 usando técnicas de *webcrawling* [9], con el objetivo de extraer los más de 10 mil casos específicos que la CIE-10 ha especificado y cubrir toda posible comorbilidad.

En las siguientes secciones describimos los métodos desarrollados en este trabajo para la anonimización de los datos paciente y la detección de comorbilidades de la nota médica.

4. Anonimización del paciente

Como se mencionó con antelación se quiere anonimizar los datos del paciente conservando la mayor cantidad de información del texto original. El siguiente ejemplo muestra un registro del expediente médico donde se tiene los datos del paciente de forma tabular y en la nota médica se vuelve a especificar el nombre del paciente:

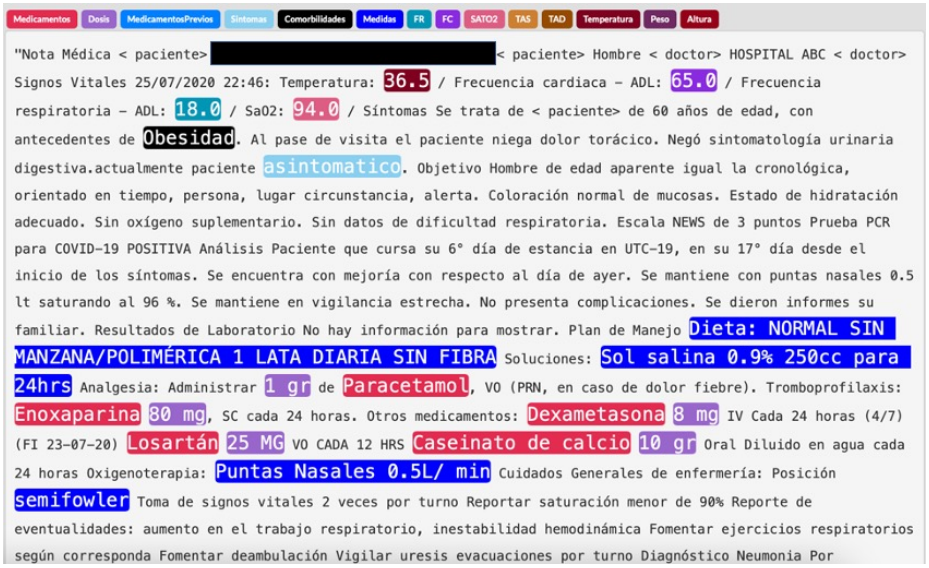

Fig. 2. Ejemplo de una nota médica etiquetada con características que se muestran en la figura.

- Nombre: José María.
- Apellido Paterno: Sánchez.
- Apellido Materno: de los Ángeles.
- Nota Médica: Se trata de José María Sanches de los Ángeles de 57 años de edad, con Obesidad grado I, sin otros antecedentes de importancia para la enfermedad actual. El señor José María Sanches de los Ángeles niega dolor torácico.

El objetivo de este proceso es eliminar las menciones del nombre del paciente para luego proceder a analizar la nota médica sin riesgo de violar la confidencialidad del paciente. El proceso de anonimización se describe de la siguiente manera:

1. Se realiza un preprocesamiento retirando signos de puntuación, acentos, números y mayúsculas. Se desea eliminar la mención del nombre del paciente del texto, pero conservar el resto del texto, lo cual incluye, entre otros elementos, signos de puntuación y acentos. Para no perder la información mencionada se creó una copia del texto. El resultado del preprocesamiento en el ejemplo dado es el siguiente:

   se trata de jose maria sanches de los angeles de años de edad con obesidad grado i sin otros antecedentes de importancia para la enfermedad actual el señor sanchez niega dolor toracico.

2. Una vez obtenido el texto preprocesado se utilizaron expresiones regulares [7] para reemplazar los elementos que conforman el nombre completo del paciente (nombres, apellido paterno y apellido materno) por un símbolo predefinido (#). Cada uno de los elementos del nombre puede estar constituido por una o más palabras, por ello se remplazó cada elemento del nombre por la misma cantidad de palabras marca. El resultado de este proceso en el ejemplo dado es el siguiente:

se trata de # # # # # sanches de años de edad con obesidad grado i sin otros antecedentes de importancia para la enfermedad actual el señor # niega dolor toracico.

Es importante notar aquí que no todos los elementos del nombre del paciente fueron reemplazados por el símbolo # ya que existen errores ortográficos. El error ortográfico en el apellido del paciente (*sanches* en lugar de *sanchez*) impide hacer un reemplazo directo ya que la palabra *sanches* no está en vocabulario de búsqueda.

3. Una vez remplazadas en el texto las ocurrencias bien escritas de los elementos del nombre, se buscaron y remplazaron las ocurrencias mal escritas. Para ello se separaron las palabras que conforman el nombre, se descartaron las palabras que tengan una longitud menor a tres caracteres y se calculó la distancia de Levenshtein con las palabras en el texto que tenían una longitud similar. Si la distancia entre dos palabras (ponderada con la longitud de la palabra) es menor a cierto rango, definido previamente, la palabra es remplazada con la palabra marca. El resultado de este proceso en el ejemplo dado es el siguiente:

   se trata de # # # # # # de años de edad con obesidad grado i sin otros antecedentes de importancia para la enfermedad actual el señor # niega dolor toracico.

4. Finalmente, al tener una correspondencia uno a uno entre las palabras del texto procesado y el texto original, se pudo recuperar todos los elementos eliminados en el preprocesamiento. Y a su vez se remplazó la ocurrencia de una o más palabras marca consecutivas por la etiqueta paciente. El resultado de este proceso en el ejemplo dado es el siguiente:

   Se trata de <paciente> de 57 años de edad, con Obesidad grado I, sin otros antecedentes de importancia para la enfermedad actual. El señor <paciente> niega dolor torácico.

5. Detección de comorbilidades

5.1. Preprocesamiento

Dadas las notas médicas anonimizadas se procedió a hacer un proceso de limpieza del texto en donde se removió elementos no relevantes para la tarea a realizar, tales como acentos y puntuación. El mismo proceso se realizó con el listado del CIE-10 de enfermedades y problemas relacionados a la salud.

Dado que en su mayoría cada elemento del listado del CIE-10 se conforma por más de una palabra se extrajeron las palabras más relevantes para su búsqueda en las notas médicas.

Para la extracción de palabras del listado del CIE-10 primeramente se retiraron todas las palabras vacías. Posteriormente, tomando cada elemento del listado como un documento diferente, se calculó la frecuencia inversa de documento de cada palabra. Obteniendo así una forma de valorar que tan buen discriminante es cada palabra [6].

Esta información fue usada para extraer las 3 palabras más relevantes de cada elemento del listado del CIE-10 y crear un conjunto de palabras (5,968) a buscar en las notas médicas. Para cada palabra a buscar se guardó con que elementos del listado está relacionada y la relevancia tiene en cada elemento.

5.2. Búsqueda de múltiples palabras con posibles errores en un texto

Dado un conjunto de palabras y un texto, se realizó un proceso para obtener en orden una lista de todas las palabras del conjunto que aparecen en el texto, considerando posibles errores en la escritura de la palabras en el texto. Para ello se iteró sobre cada una de las palabras que conforman el texto y se realizó lo que se presenta a continuación:

1. Filtrado de candidatos

La ocurrencia que dos palabras tengan el mismo número de letras, a lo cual se denomina como anagrama, es poco frecuente. Tal es el caso de nuestro conjunto de 5,968 palabras a buscar, de las cuales solo hubo 63 casos de anagramas.

Por ello, lo primero que se realizó fue un proceso para que dada una palabra objetivo y un conjunto de palabras, determinar un subconjunto de palabras que tenga aproximadamente las mismas letras, a lo cual denominamos como semi-anagrama.

Para realizar de forma eficiente se realizó un preprocesamiento en donde se crean subconjuntos de palabras que tienen el mismo número de veces la misma letra, a los cuales denominamos como filtros. Ejemplo creación de filtros:

Conjunto de palabras: {“ab”, “abc”, “aca”},
- (a, 0) : {},
- (a, 1) : {“ab”, “abc”},
- (a, 2) : {“aca”},
- (b, 0) : {“aca”},
- (b, 1) : {“ab”, “abc”},
- (c, 0) : {“ab”},
- (c, 1) : {“abc”, “aca”}.

Dada la información anteriormente procesada, para encontrar un subconjunto de las palabras que sean anagramas de una palabra objetivo, solo se necesita calcular el número que tiene la palabra objetivo de cada letra y sacar la intersección de los filtros correspondientes.

Sin embargo, seleccionar palabras que sean anagramas de otra solo permite considerar errores en el orden de la escritura de la palabra, no en las letras que la conforman. Para ampliar el número de candidatos es importante tener un umbral de cuantas letras se pueden errar.

Dado x un porcentaje que se puede errar en una palabra, previamente definido, y una palabra p de longitud l, se calcula u que corresponde a un número entero del número de letras en p que se pueden errar; donde errar es cambiar una letra por otra, agregar una letra o eliminar una letra. Una vez calculado u y teniendo los filtros correspondientes a p, se puede calcular un subconjunto de palabras que sean semi-anagramas.

Lo primero a denotar es que toda palabra que sea semi-anagrama de p, puede no estar en a lo más $2u$ filtros. Esto se debe a que un cambio del tipo agregación o eliminación de una letra retira la palabra de exactamente un filtro (en la letra agregada o eliminada), sin embargo un cambio tipo remplazo elimina la palabra en dos filtros (en la letra agregada y en la eliminada).

Tabla 1. Ejemplo de comparación de dos palabras con $u = 1$.

ejemplo	eje	jem	emp	mpl	plo
eejemplo	eej	eje	jem	emp	mpl
Elementos en común	{e,e,j}	{e,j}	{e,m}	{m,p}	{p,l}
Puntaje	3	2	2	2	2

Dadas estas observaciones, podemos eliminar, de un conjunto de palabras candidato, las palabras que no son semi-anagramas. Esto se realiza con un sistema de *strikes*, donde un *strike* a una palabra es no estar presente en un filtro.

Para calcular el conjunto de palabras candidato para p basta tomar la unión de cualesquiera $2u + 1$ filtros de p, ya que cualquier palabra que sea semi-anagrama de p está en al menos uno de esos filtros. Posteriormente, se calcula la intersección entre el conjunto de candidatos y de cada filtro.

Si un candidato no se encuentra en un filtro, acumula un *strike*. En el dado caso que un candidato acumule $2u + 1$ *strikes*, se retira del conjunto de candidatos.

Finalmente, si la suma de los *strikes* acumulados y la diferencia de longitud entre una palabra candidato y objetivo es mayor a $2u$, también se retira del conjunto de candidatos. Esto se debe a que los errores de tipo remplazo de letra no afectan en la longitud de la palabra, pero son los que afectan en dos filtros; y los errores de tipo eliminación y agregación de letra afectan en a lo más un filtro, pero también afectan en una unidad la longitud de la palabra.

2. Filtrado individual

Una vez obtenido un conjunto de candidatos que son semi-anagramas de la palabra objetivo, se realiza un proceso que hace una comparación lineal entre cada palabra candidato y la palabra objetivo. Primero se compara si la palabras son iguales, de no ser así se realiza otro proceso con el objetivo de descartar las palabras que tienen un orden considerablemente diferente al de la palabra objetivo.

Este proceso se realiza obteniendo todas las $2u+1$ secuencias de letras consecutivas de cada palabra y comparando, en orden de aparición, la intersección de sus elementos, un ejemplo se presenta en la Tabla 1.

3. Decisión final

Una vez obtenido un conjunto menor de palabras a comparar con la palabra objetivo, se calcula la distancia de Levenshtein de esta con cada una de las palabras del conjunto. Si la distancia (ponderando con la longitud de la palabra) de una palabra candidato a la objetivo es menor a cierto rango definido previamente, se establece que la palabra apareció en el texto y se puede agregar a la lista de palabras.

5.3. Selección de enfermedades por lista de palabras

Una vez obtenido una lista, en orden, de las palabras del conjunto que aparecen en el texto, se realiza un proceso de selección para ver que enfermedades hacen referencia a las palabras en la lista.

Usando la relación entre las palabras y las enfermedades que se guardo previamente, se implementó un algoritmo glotón. En el cual, para definir la primera enfermedad, se toma desde la primera palabra en la lista la mayor secuencia de palabras consecutivas que hagan referencia a una o más enfermedades (en el dado caso que las mismas palabras aparezcan en más de una enfermedad).

Dada esta secuencia de palabras se calcula la relación que tienen con la enfermedad en la que aparecen, sumando el valor individual de las relaciones de cada palabra. Si el valor calculado es mayor a cierto umbral definido previamente, se define que la enfermedad aparece en el texto. Se eliminan las palabras de la lista y se vuelve a repetir el proceso hasta que no queden elementos en la lista.

6. Resultados

A continuación se presentan los resultados de los métodos de anonimización y detección de comorbilidades.

6.1. Anonimización

La metodología usada para anonimizar las notas médicas fue bastante efectiva cubriendo todas las menciones del nombre del paciente, incluyendo las instancias en donde hay errores en la escritura de este, sin embargo a su vez puede presentar problemas. Un problema es la sustitución de palabras que tienen una similitud con alguno de los elementos que compone el nombre.

Si bien experimentando con el umbral de la distancia de Levenshtein se puede evitar el remplazo de varias palabras, hay casos en donde dos palabras son demasiado parecidas para evitar que sean confundidas por la misma palabra con un error en su escritura. Ejemplo: Daniel y Daniela. La Figura 1 muestra una nota médica que ya pasó por el proceso de anonimización, donde el nombre del paciente fue reemplazado por $<paciente>$.

6.2. Detección de comorbilidades

La metodología que se desarrolló fue pensada para buscar en notas médicas los elementos del listado del CIE-10 de enfermedades y problemas relacionados a la salud, sin embargo al realizarse el trabajo se procuró que pudiera funcionar para cualquier tipo de texto y cualquier lista con una cantidad extensa de elementos que consten de una o más palabras.

Para la evaluación de esta tarea se contó con la ayuda de personal médico que etiquetó las comorbilidades que aparecieron en las notas médicas, como se menciono en la Sección de Corpus.

La comparación uno a uno de términos médicos del listado del CIE-10 con los términos usados en las notas médicas haría complicada la evaluación. Por ello, para la evaluación se optó por usar una lista de todas las comorbilidades encontradas en las 98 notas médicas en lugar del listado del CIE-10.

Tabla 2. Evaluación detección de comorbilidades por umbral de similitud entre palabras.

Umbral	**70-100**	**85-100**	**99-100**
Recall	92.63	95.78	81.05
Precisión	20.37	32.15	33.62
F1	33.39	48.14	47.52

Tabla 3. Comparación de comorbilidades en una nota médica y de resultados con listado CIE-10 y total de comorbilidades usada para la evaluación.

Comorbilidades etiquetadas	***Detección listado CIE-10***	***Detección listado etiquetas***
DIABETES MELLITUS 2	DIABETES MELLITUS	DIABETES MELLITUS 2
HIPERTENSION ARTERIAL SISTEMICA		HIPERTENSION ARTERIAL SISTEMICA, HIPERTENSION ARTERIAL SISTEMATICA
EXFUMADOR	OTRAS ENFERMEDADES CARDIOPULMONARES	EXFUMADOR
		COLECISTECTOMIA, HIPOKALEMIA
	AISLAMIENTO, DISNEA, TOS, CEFALEA	

Se tomaron aleatoriamente 30 notas médicas para la evaluación y se usaron 3 distintos umbrales para la similitud entre palabras.

El umbral de 99 a 100 de similitud entre palabras es equivalente a no aceptar errores o alteraciones en una palabra (sin considerar acentos).

Se compararon los resultados entre el etiquetado y los resultado de la metodología presentada en este documento, los resultado se muestran en la Tabla 2. El mejor desempeño se logra con el umbral entre 85 a 100 de similitud entre palabras, equivalente a permitir errores de una o dos letras en las palabras.

La metodología desarrollada sirve para la búsqueda de múltiples elementos de más de una palabra en un texto, donde se considera que las palabras pueden tener errores en su escritura. Alta importancia tiene en su efectividad la lista de elementos a buscar.

En la Tabla 3 se presenta la variación en los resultado obtenidos al usar diferentes listas y se compara con el etiquetado real de la nota médica.

En el caso de comorbilidades y el uso del listado del CIE-10 cabe destacar dos problemas que se presentaron. El primero es que la distinción entre comorbilidades y enfermedades o problemas de la salud puede llegar a ser bastante considerable, abarcando estos últimos síntomas.

El segundo problema proviene de las particularidades que puede llegar a tener a cada uno de las enfermedades o problemas de la salud, no mencionadas en la descripción

directa. Ejemplo: I10 Hipertensión esencial (primaria) cubre el caso de Hipertensión arterial sistémica.

7. Conclusiones

Los métodos propuestos para la realización de este trabajo son bastante útiles cuando no se cuenta con una cantidad muy grande de datos para entrenar modelos de aprendizaje supervisado.

Particularmente, el uso de la distancia de Levenshtein para la anonimización puede cubrir múltiples omisiones de usarse otro método. Por el lado de la detección de comorbilidades, el trabajo realizado puede servir para múltiples tareas de la misma índole (detección de una colección amplia de términos en un texto).

Como trabajo a futuro se sugiere el uso de un mejor listado de posibles comorbilidades, como lo sería uno realizado por doctores, enfocado en las comorbilidades más relevantes. Al igual se propone en un futuro realizar las tareas realizadas con métodos de aprendizaje de máquina.

Referencias

1. Allen, J.: Natural Language Understanding. Benjamin Cummings (1987)
2. Backurs, A., Indyk, P.: Edit distance cannot be computed in strongly subquadratic time (unless seth is false). In: Proceedings of the forty-seventh annual ACM symposium on Theory of computing. pp. 51–58 (2015)
3. Martín-Vegue, A., Vázquez-Barquero, J., Castanedo, S. H.: Cie-10 (i): Introducción, historia y estructura general. Papeles medicos, vol. 11, no. 1, pp. 24–35 (2002)
4. Plasencia-Urizarri, T. M., Aguilera-Rodríguez, R., Almaguer-Mederos, L. E.: Comorbilidades y gravedad clínica de la COVID-19: Revisión sistemática y meta-análisis. Revista Habanera de Ciencias Médicas, vol. 19 (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)
6. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management, vol. 24, no. 5, pp. 513–523 (1988)
7. Stubblebine, T.: Regular expression pocket reference: Regular expressions for perl, ruby, PHP, python, C, java and .NET (2007)
8. Thabit, K., AL-Ghuribi, S. M.: A new search algorithm for documents using blocks and words prefixes. Scientific Research and Essays, vol. 8, no. 16, pp. 640–648 (2013)
9. Viveros-Jiménez, F., Sanchez-Perez, M. A., Gómez-Adorno, H., Posadas-Durán, J. P., Sidorov, G., Gelbukh, A.: Improving the boilerpipe algorithm for boilerplate removal in news articles using html tree structure. Computación y Sistemas, vol. 22, no. 2, pp. 483–489 (2018)
10. Zhang, Y., Ma, X., Song, G.: Chinese medical concept normalization by using text and comorbidity network embedding. In: 2018 IEEE International Conference on Data Mining (ICDM). pp. 777–786 (2018)