UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

Modelado de las características extraídas de los pares de nombres confusos de medicamentos por su parecido ortográfico y fonético

Tesis

Para obtener el grado de

Doctor en Ciencias de la Computación

Que presenta:

M.C.C. Christian Eduardo Millán Hernández

Tutor académico:

Dr. René Arnulfo García Hernández

Tutor adjunto:

Dra. Yulia Nikolaevna Ledeneva

Dr. Ángel Hernández Castañeda

Tianguistenco, México. Febrero 2020

¡Y demos por perdido el día en que no hayamos bailado al menos una vez!

– Friedrich Nietzsche.

Resumen

Los nombres de medicamentos que se parecen, por como se ven o como suenan, son la causa principal de los errores de medicación por confusión. Por esta razón, la Administración de Alimentos y Medicamentos en Estados Unidos ha implementado estrategias para revisar el nombre propuesto de un nuevo medicamento. El objetivo es evitar que se formen pares confusos a partir del nombre propuesto con los que existen en el mercado.

Las herramientas utilizadas para identificar pares confusos calculan un valor de similitud entre el nombre propuesto y una base de datos de los nombres de medicamentos previamente registrados. El valor de similitud obtenido es utilizado para identificar nombres potencialmente confusos. En específico, algoritmos de similitud léxica se implementan en esta tesis de manera individual o en combinación para predecir el grado de confusión de dos nombres de medicamentos por su parecido en medios de comunicación escritos o verbales.

La presente investigación se enfoca en medir el parecido ortográfico y fonético entre dos nombres a evaluar, a partir de las características presentes en los nombres de medicamentos indicados en los reportes de errores de medicación, con la finalidad de mejorar el proceso de identificación de pares confusos por su parecido ortográfico y fonético.

En esta tesis se realiza una revisión de las medidas individuales que consideran los aspectos ortográficos y fonéticos. Además, se estudian las medidas combinadas que consideran simultáneamente ambos aspectos, con el fin de detectar la potencial confusión entre nombres de medicamentos. Asimismo, también se muestra una discusión de los problemas presentes en cada una de las soluciones del estado del arte.

En los resultados de esta investigación se presenta un modelo que combina de manera eficaz las características ortográficas y fonéticas que están presentes en los nombres confusos de medicamentos. Para este objetivo, se utilizó un método de regresión logística con un proceso de entrenamiento evolutivo que supera los resultados obtenidos del método tradicional de entrenamiento. Este modelo se ha publicado en una revista especializa arbitrada e indexada de reconocimiento internacional.

Del mismo modo, se presenta bajo el mismo principio un modelo que considera al características ortográficas y fonéticas presentes en los pares confusos. Una nueva medida individual de similitud ortográfica para la identificación de pares confusos. También, este resultado ha sido publicado en una revista especializada arbitrada e indexada de reconocimiento internacional.

Los resultados publicados en ambos artículos prueban que el modelo de las características extraídas de los pares confusos permite ajustar o diseñar medidas eficaces para el problema de identificación de pares potenciales. El modelado ha sido obtenido de manera automática y adaptado mediante un enfoque evolutivo o con uso de técnicas de aprendizaje automático.

Índice general

Índice de Figuras

Índice de Tablas

Capítulo I. Protocolo

En este capítulo se presenta el protocolo de tesis actualizado con los antecedentes, teorías y estado del arte que dan fundamento a la propuesta y a los resultados mostrados en los siguientes capítulos.

1.1. Antecedentes

Los *profesionales de la salud* frecuentemente utilizan la *farmacoterapia* para prevenir o curar alguna enfermedad. No obstante, por la falta del monitoreo adecuado o como resultado de un error de comunicación la administración de *medicamentos* incorrectos puede causar daños al paciente [1]–[3].

Los incidentes que ocurren sin intención durante el uso de un medicamento son conocidos como errores de medicación[1] y representan un riesgo para la *seguridad*

[1] El consejo *National Coordinating Council for Medication Error Reporting and Prevention (NCC MERP)* recomienda a investigadores de errores de medicación, desarrolladores de software y a las instituciones utilizar la siguiente definición estándar: ***error de medicación*** *es cualquier incidente prevenible que puede causar daño al*

del paciente [4]. La prescripción y administración de un fármaco en el lugar de otro es un error frecuente causado por el parecido que se puede presentar entre sus nombres [1], [2], [5].

Es decir, los nombres de medicamentos que se ven o suenan igual (LASA, por las siglas en inglés de *Look-Alike and Sound-Alike*) son la causa principal de errores de medicación por confusión [6]–[8]. La formación de pares LASA es común entre *nombres de propietario o de marca*, así como en, los *nombres genéricos* de medicamentos. La probabilidad de error aumenta debido a la cantidad y al gran número de marcas de medicamentos que existen en el mercado. En consecuencia, doctores, enfermeras, farmacéuticos o cualquier profesional de la salud, incluso el mismo paciente; están en riesgo de cometer errores de confusión [6].

Los errores de medicación por confusión ocasionalmente producen daños leves al paciente, pero en otras circunstancias producen daños severos e incluso la muerte. En Estados Unidos, las cifras indican que hay 1.3 millones de afectados al año por que estos errores que incluso causan la perdida de vida de una persona al día [9]. Estos errores genera costos aproximados de 3.5 billones de dólares adicionales por atender al paciente afectado [2], [9].

Aunque a nivel mundial resulta complicado enunciar una cifra exacta que dimensione el problema. La *Organización Mundial de la Salud* (WHO, por sus siglas en inglés de *World Health Organization*) estima que uno de cada diez pacientes es afectado por errores de medicación. Por tal razón, el programa Tercer Desafío Mundial de Seguridad del Paciente: Medicamentos sin daño (*Third Global Patient Safety Challenge: Medication Without Harm*) fue lanzado para mejorar los sistemas

---

paciente o dar lugar a el uso inapropiado de medicamentos, mientras la administración de estos está bajo el control del profesional de la salud, el mismo paciente o simplemente como consumidor. Dichos incidentes pueden estar relacionados con la práctica profesional, uso de productos para el cuidado de la salud, los procedimientos y con los sistemas, incluyendo la prescripción; comunicación al ordenar la medicación; la etiquetas o logotipo del producto, el envasado y denominación (distintiva o genérica); compuestos; dispensación; distribución; administración; educación; seguimiento y utilización.

de salud con el fin de evitar y disminuir los daños que ocasionan los errores de medicación. Una objetivo en particular de esta estrategia es reducir la aparición de nuevos pares LASA [2], [9].

Para identificar errores de medicación por pares LASA se han diseñado procedimientos con el objetivo de realizar la observación directa de los procesos de uso de medicamentos y detectar los errores en tiempo, con un alto grado de validez y confianza. Sin embargo, este enfoque resulta demasiado costo [1].

Otra alternativa es utilizar sistemas de reportes de errores de medicación, donde los profesionales de la salud de manera voluntaria informen sobre cualquier incidente o error de medicación, con el propósito de analizar qué sucedió, dónde sucedió y por qué sucedió. Es decir, examinar todos los detalles de los errores para realizar las correcciones necesarias y evitar que suceda nuevamente, sin importar quién lo cometió.

Los errores de medicación, de acuerdo con los datos obtenidos de los reportes, ocurren en todas las etapas de administración del medicamento, como son: la prescripción u ordenación, la transcripción, la dispensación, la administración o el monitoreo del producto. Además, se estima que el 25 porciento de los casos están relacionados con pares LASA [10].

Según la Administración de Alimentos y Medicamentos (FDA, por sus siglas en inglés de *Food and Drug Administration*), el proceso de nombrar un medicamento es una tarea que debe ser guiada por un conjunto de buenas prácticas y estándares que brinden seguridad para reducir el potencial riesgo de confusión. Para prevenir estos riesgos es necesario la distinción inequívoca del nombre del medicamento para que el paciente reciba el medicamento correcto. Por tal razón, en varios países se han implementado estrategias para limitar la aprobación de un medicamento para su comercialización.

Las estrategias se dividen en: previas y posteriores a su aprobación. Las estrategias *a posteriori* se implementan, por ejemplo, en situaciones cuando un nombre de medicamento a incurrido en errores de confusión frecuentes, causando severos daños. En este caso, se ha cambiado el nombre para terminar con la aparición de futuros incidentes de confusión. Sin embargo, las estrategias más comunes, posteriores a cuando el medicamento fue puesto en el mercado, buscan prevenir la reincidencia de los errores reportados [1], [10]–[12]. Entre las principales acciones se sugiere utilizar advertencias y alertas en los sistemas electrónicos y en áreas donde se utilicen medicamentos, escribir las recetas incluyendo el nombre comercial y el nombre genérico, indicar en las recetas la dosis, vía de administración y uso. Además, se recomienda escribir el propósito de la prescripción, almacenar por separado los medicamentos de riesgo, colocar etiquetas, mejorar los condiciones ambientales como la iluminación, reducción de ruidos o evitar que se desarrollen distintas tareas simultáneamente por los profesionales de la salud [1], [10]–[12].

Adicionalmente, en las estrategias de prevención *a posteriori* se proponen identificar los casos de confusión en una lista de nombres LASA para que los profesionales de la salud estén alertas, de los pares de alto riesgo indicados, mientras estos medicamentos son utilizados [1]. El *Instituto para las Prácticas Seguras de Medicación* (ISMP, por las siglas en inglés de *Institute For Safe Medication Practices*) en Estados Unidos, publica una lista de nombres de medicamentos confusos ISMP[2]. La lista contiene pares LASA identificados en errores de medicación registrados en el *Programa Nacional de Reporte de Errores de Medicación* de la ISMP (ISMP MERP, por sus siglas en inglés de *ISMP National Medication Errors Reporting Program*) [13].

Además, ya se han propuesto estrategias *a posteriori* que aumentan la distinción entre nombres confusos (ver Tabla I-1) [1], [10], como utilizar letras

[2] La lista de la ISMP de pares LASA es publicada en: https://www.ismp.org/tools/confuseddrugnames.pdf

mayúsculas, negritas, colores o contraste en las letras de la parte del nombres donde son diferentes [14], [15]. El ISMP publica una lista de nombres confusos con la implementación de letras mayúsculas para aumentar su fácil identificación[3].

Tabla I-1. Ejemplos de estrategias que aumentan la distinción entre los nombres confusos rifampin - rifaximin.

Estrategia	**Ejemplo**
Mayúsculas	rifaMPin – rifaXIMin
Negritas	rifa**mp**in – rifa**xim**in
Negritas – Mayúsculas	rifa**MP**in – rifa**XIM**in
Texto en color	rifampin – rifaximin
Contraste	rifampin – rifaximin

En cambio, una estrategia de prevención en la fase de aprobación del medicamento, previa a que salga a la venta, evita que nuevos pares LASA se formen a partir del nombre propuesto [1], [5], [16], [17]. Para detectar *a priori* los potenciales pares LASA se utilizan distintas pruebas y herramientas. Por ejemplo, la implementación de herramientas computacionales para realizar búsquedas de nombres similares en bases de datos de medicamentos existentes, la evaluación por expertos en errores de confusión, la aplicación de pruebas psicolingüísticas de memoria y percepción, así como la observación directa de la frecuencia de errores en la tarea simulada de prescribir, dispensar y administrar medicamentos.

La confusión de un medicamento ocurre durante el proceso cognitivo al identificarlos [1]. La similitud entre nombres aumenta la posibilidad de que ocurran errores humanos [8]. Por ejemplo, un error de percepción visual ocurre cuando en

[3] Lista de la ISMP con pares LASA con uso de letras mayúsculas: https://www.ismp.org/recommendations/tall-man-letters-list

una prescripción en formato impreso es leído el medicamento hydralazine como hydroxyzine, ver Figura I-1b. En ocasiones, el error visual se genera en medios manuscritos (ver Figura I-1a) bajo circunstancias donde el medicamento es ordenado de manera verbal y se confunde la pronunciación, por ejemplo, de Xanax con Zantac (ver Figura I-1c). Del mismo modo, los errores de memoria de corto plazo ocurren cuando el farmacéutico lee una prescripción de hydroxizine, mientras se dirige al estante donde se almacenan los medicamentos y tiene un *lapsus de memoria* o por error toma el medicamento hydralazine. Los errores de control motor ocurren, por ejemplo, cuando se selecciona el medicamento incorrecto de una lista desplegable computarizada. Sin embargo, es más importante reconocer que la causa de la confusión es el parecido en la escritura o en la pronunciación que ocurre entre los nombres de medicamentos.

En Estados Unidos, la Administración de Alimentos y Medicamentos (FDA, por sus siglas en inglés de *Food and Drug Administration*) evalúa el nombre de marca propuesto para un medicamento propietario, al identificar una posible confusión ortográfica o fonética con los nombres de otros productos existentes en el mercado [18]. En el proceso de revisión se realizan distintas pruebas, donde el humano experto en "errores de medicación" dictamina si el nombre del fármaco resulta confuso con otros.

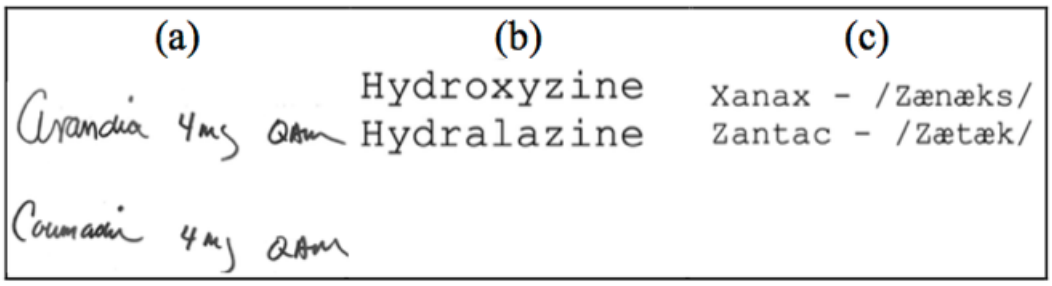

Figura I-1. Ejemplos de nombres de medicamentos confusos que pueden participar en errores de percepción visual y auditiva. (a) Parecido ortográfico en medios manuscritos. (b) Parecido ortográfico en medios impresos. (c) Parecido fonético en la pronunciación [19].

La FDA indica una serie de sugerencias dirigidas a quienes proponen un nuevo nombre de medicamento [18], [20]. Estas sugerencias incluyen un conjunto de criterios necesarios a considerar en la revisión del parecido ortográfico o fonético de un nombre propietario de medicamento, ver Tabla I-2.

Tabla I-2. Criterios para identificar pares LASA.

Tipo de similitud	**Consideraciones cuando se realiza la búsqueda en las bases de datos**		
	Causa de similitud potencial en nombres de medicamentos	**Atributos a examinar para identificar nombres de productos similares**	**Efectos potenciales de los nombres**
Look-alike	Similitud de escritura (grupo de letras que representan una palabra)	Prefijos idénticos Infijos idénticos Sufijos idénticos Tamaño del nombre Traslape de características.	Pueden parecer similares impresos o medios electrónicos, causando una confusión en la comunicación impresa o electrónica. Pueden verse similares cuando se escriben y causar una confusión en la comunicación escrita
	Similitud ortográfica (escritura de palabras con las letras apropiadas)	Grupo de letras similar Tamaño del nombre Movimiento ascendente Movimiento descendente Cruces Acentuaciones Ambigüedad de la letra mano escrita Traslape de características	Pueden verse similares cuando se escriben y causar una confusión en la comunicación escrita
Sound-alike	Similitud fonética	Prefijos idénticos Infijos idénticos Sufijos idénticos Número de silabas Énfasis Colocación de sonidos vocálicos Colocación de sonidos de consonantes Traslape de características	Pueden sonar de manera similar cuando se pronuncian y llevar a la confusión de un medicamento en la comunicación oral.

Como anteriormente se menciona, en el proceso de revisión de un nombre propuesto se utilizan herramientas computacionales que implementan técnicas de inteligencia artificial para realizar la búsqueda de posibles pares confusos entre una base de datos de nombres de medicamentos existentes. Como resultado de esta búsqueda se obtiene una lista ordena por un valor de similitud, la cual tiene como finalidad facilitar la evaluación para el humano experto, ya que solo ha de centrarse en aquellos casos de mayor riesgo. La FDA desarrolló el software POCA (*Phonetic and Orthographic Computer Analysis*) que utiliza un conjunto de algoritmos para determinar la similitud entre dos nombres de medicamentos por su parecido ortográfico o fonético con el objetivo de comparar el nombre propuesto contra la base de datos de medicamentos existentes [21].

En una investigación [22] se realizó un análisis estadístico de tres características en los pares confusos de la lista publicada por la ISMP. Los resultados muestran que los pares presentan coincidencias en la primera letra, cadenas compartidas de al menos tres letras y la misma longitud entre los nombres. De igual forma, se reportó que el 99% de las confusiones presentan al menos una de las características buscadas, por lo que se considera que un par de nombres es confuso si presenta estos rasgos distintivos. En la conclusión de esta investigación se menciona que estas características comunes en pares confusos deben ser consideradas en los procesos de revisión de nombres propuestos. Además, se realizó el cálculo de la similitud entre los pares LASA mediante el software POCA y solo el 75% de las confusiones muestra una similitud superior a 50 %. La discusión en el artículo indica que se esperaría un valor de similitud alto para toda la lista, ya que se refiere a casos confusos reportados en el ISMP MERP.

La medidas léxicas o también referidos como algoritmos de *string matching* han sido implementadas originalmente en otras áreas como [23]–[28]. Existen evidencias de su efectividad en la identificación de pares LASA. Incluso, nuevas medidas han sido propuestas para este problema en particular [29]–[32]. Además, se han propuesto

métodos que combinan los valores de obtenidos de distintas medidas individuales han comprobado que mejoran la exactitud de los resultados en comparación de utilizarlas de manera aislada [30], [33]. Por ejemplo, Lambert [33] evaluó 22 medidas de similitud ortográficas y fonéticas. El mejor resultado fue alcanzado por la medida *Trigram-2B*. Esta medida utiliza trigramas y repite la primera letra en los primeros tres trigramas. Para combinar los valores obtenidos de distintas medidas individuales, Lambert propuso un modelo de regresión logística que utiliza las medidas *Editex*, *NED* y *Trigram*-2B para medir la similitud de un par de medicamentos sin lograr resultados significativos ante Trigram-2B [33].

Kondrak [30] propuso dos medidas de similitud Bisim y Aline. La primera mide la similitud ortográfica en pares de medicamentos, mientras que la segunda es un algoritmo de alineamiento de cognados adaptada para medir la similitud fonética [34], [35]. La medida Bisim utiliza bigramas y repite una vez la primera letra. Los resultados muestran como Bisim supera a las medidas empleadas por Lambert. Además, Kondrak evaluó un método que combina el resultado de cuatro medidas (Aline, Bisim, NED, Prefix), mediante el promedio aritmético. Este método mejora el resultado de todas las medidas individuales que componen la combinación. Puesto que, en un resultado combinado las debilidades de cada medida particular son compensadas por el resto de las medidas que componen en el ensamble al momento de determinar el valor de similitud de un par LASA.

La FDA ha sugerido constantemente que el uso de una medida individual no puede evaluar la similitud en todas sus dimensiones y es necesario utilizar un ensamble que de manera inteligente combine medidas ortográficas, fonéticas [18], [20] para cubrir todas los criterios a considerar en la identificación de pares LASA (ver Tabla I-2).

La presente investigación se centra en el problema de identificar la potencial confusión que puede existir entre los nombres de medicamentos. Considerando el

aspecto ortográfico y fonético a partir de la extracción de las características presentes en los pares previamente reportadas como casos de confusión.

1.2. Planteamiento del problema

Las medidas individuales son ajustadas o diseñas para medir la similitud en un problema en particular [23], [24], [26]–[28], [36]–[42]. Desde las primeras investigaciones relacionadas con la confusión de nombres de medicamentos se ha comprobado y verificado su eficacia para identificar la similitud en estos errores [7], [43]. No obstante, se destaca Bisim [29]–[32] por ser una medida propuesta para considerar las características ortográficas reconocidas en los pares previamente reportados como confusos [7], [30].

El funcionamiento interno de Bisim [30] evita problemas comunes de otras medidas que están basadas en ngramas (bigramas, trigramas); por ejemplo, registrar como pares confusos aquellos que no lo son (falsos positivos) o, indicar como un caso de no confución aquellos pares que si lo son (falsos negativos). Estos ajustes no muestran evidencia distinta de que fueron realizados de manera manual, todo indica que después de analizar las características comunes presentes en una colección en específico de pares LASA, fueron generalizadas al diseñar el algoritmo de Bisim con el fin de aumentar su valor de similitud y ser recuperados como pares LASA.

Aunque la necesidad de combinar varias medidas tiene como fin aumentar la precisión de recuperación de pares LASA. En el estado del arte no se muestran evidencias exactas del por qué seleccionar solo un subconjunto medidas para participar en una combinación o cuáles deben ser las características que debe tener una medida para participar en un ensamble.

En la sección anterior se menciona que Lambert [33] seleccionó las tres mejores medidas después de evaluar de manera individual. Se formuló un modelo de regresión logística que combina a: Trigram-2B (similitud ortográfica), ED (distancia ortográfica) y Editex (distancia fonética). Sin embargo, el modelo propuesto combina las medidas para obtener un predictor dicotómico que indica si un par es confuso o no y no una medida que determine un valor de similitud. Además, concluyó que el

resultado obtenido por el modelo no difería sustancialmente, en términos de significancia estadística, del desempeño obtenido por la mejor medida individual (Trigram-2B).

Por otra parte, a pesar de que Kondrak reconoce que el desempeño de una medida combinada nunca será superado por las medidas individuales que la componen y de que ocurre por el hecho de que las medidas presentes en la composición cubren entre si a las otras que tiene alguna falla para identificar pares LASA. El método combinado que promedia los valores asume que la participación de cada una de las medidas es equitativa. Además, en el desarrollo de esta medida combinada no se presenta evidencia del proceso de selección o de la justificación de utilizar solo algunas mediadas para participar en la combinación.

Por lo que la discusión central del uso de medidas individuales o combinadas debe estar orientada a conocer cuáles deber ser las características que posean cada medida para dimensionar la similitud; y cómo deben de ser seleccionas y ensambladas para aumentar la precisión en la identificación de nombres confusos a partir de los pares LASA previamente reportados en los reportes de errores de medicación.

1.2.1. Pregunta de investigación

De este modo, considerando las problemáticas tratadas en la sección anterior se plantea las siguientes preguntas de investigación:

¿Cómo identificar pares potenciales de nombres confusos de medicamentos, por su parecido ortográfico y fonético, considerando las relaciones que existe entre las características de los pares confusos previamente reportados en los errores de medicación?

1.2.2. Preguntas de apoyo

- ¿Cómo mejorar el ajuste de un método de regresión logística para incrementar la precisión en la identificación de pares potenciales de nombres confusos de medicamentos considerando las características presentes en los pares confusos previamente reportado en los errores de medicación?
- ¿Cómo ajustar una escala de similitud entre bigramas en una medida basada en Bisim para incrementar la precisión en la identificación de pares potenciales de nombres confusos de medicamentos considerando las características presentes en los pares confusos previamente reportado en los errores de medicación?

1.3. Objetivo general

El objetivo general de la investigación es realizar un modelo que permita distinguir cuándo un par de nombres de medicamentos es potencialmente confuso, por su parecido ortográfico o fonético, a partir de las características presentes en los pares confusos previamente reportados en los errores de medicación.

1.3.1. Objetivos específicos

- Ajustar una regresión logística mediante un enfoque evolutivo para incrementar la precisión en la identificación de pares potenciales de nombres confusos de medicamentos considerando las características presentes en los pares confusos previamente reportado en los errores de medicación.
- Proponer una nueva medida de similitud suavizada basada en Bisim para incrementar la precisión en la identificación de pares potenciales de nombres confusos de medicamentos considerando las características presentes en los pares confusos previamente reportado en los errores de medicación

1.4. Hipótesis

Las hipótesis de investigación formuladas a partir de pregunta de investigación y las preguntas de apoyo se listan a continuación:

H_1: Considerando que en un par de nombres de medicamentos confusos existe un conjunto de características presentes, por ejemplo: la frecuencia de la posición de la coincidencia de caracteres. Se puede afirmar que, si se genera un modelo, que represente la relación de los pares confusos y se implementa en los métodos actuales, entonces se mejorará la identificación de pares LASA.

H_2: Además, un enfoque evolutivo puede ajustar mejor los parámetros de un modelo de regresión logística basado en la métrica de la medida F (*F-measure*, en inglés) en comparación con el algoritmo de entrenamiento tradicionalmente utilizado.

H_3: Un enfoque evolutivo puede ajustar mejor los pesos de una escala de similitud entre bigramas implementada en una medida para la identificación de nombres confusos de medicamentos.

1.5. Organización de la tesis

En este capítulo se presentan los antecedentes del problema. El resto de la tesis está organizado de la siguiente forma. En el capítulo dos, se presentan los conceptos sobre medidas de similitud léxicas que sirven como base para comprender los métodos del estado del arte. En seguida, se revisan las investigaciones más recientes sobre la identificación de pares LASA y los esfuerzos de resolver el problema. En el capítulo tres, se presenta un artículo publicado en una revista especializa arbitrada e indexada de reconocimiento internacional donde se propone el uso de un algoritmo de entrenamiento evolutivo para un modelo de regresión logística que mejorar la exactitud en la identificación de pares LASA. En el capítulo cuatro se presenta otro artículo que muestra una medida individual que utiliza bigramas y una escala de similitud suavizada que mejora los resultados en el problema de la identificación de pares LASA, también publicado en una revista especializa arbitrada e indexada de reconocimiento internacional. En el capítulo cinco se presentan la discusión general y las conclusiones.

1.6. Resumen

El uso la farmacoterapia puede causar daños como resultado de un error, accidente o problema de comunicación. Los errores de medicación por la confusión de medicamentos es un error frecuente por el parecido entre los nombres por como se ven o como suenan (LASA, Look-Alike and Sound-Alike). Para reducir los daños a la seguridad del paciente se deben evitar la aparición de pares LASA. Los sistemas de reportes de errores de medicación ayudan a identificar cuando suceden por causas de

confusión. Las estrategias para asegurar que el paciente reciba el medicamento correcto se dividen en previas o posteriores a su comercialización. La mayoría de los esfuerzos se centran en utilizar herramientas computacionales para realizar la búsqueda de nombres similares previas que el nombre propuesto salga al mercado. Es el caso de la FDA que utiliza el software POCA que utiliza medidas de similitud. En el estado del arte un primer trabajo muestra las mejores medidas individuales y el uso de una regresión logística que combina los resultados de las tres mejores medidas. Otro trabajo más reciente propone dos medidas individuales que mejorar la exactitud de los resultados previos, así como un método que la combinación de cuatro medidas mediante el promedio. La presente investigación se centra en descubrir el parecido ortográfico o fonético a partir de la extracción de las características presentes en los pares LASA previamente reportados.

Al final de este capítulo se presenta el problema y la pregunta de investigación: ¿Cómo identificar pares potenciales de nombres confusos de medicamentos, por su parecido ortográfico y fonético, considerando las relaciones que existe entre las características de los pares confusos previamente reportados en los errores de medicación? El objetivo general de la investigación es: realizar un modelo que permita distinguir cuando un par de nombres de medicamentos es potencialmente confuso, por su parecido ortográfico o fonético, a partir de las características presentes en los pares confusos previamente reportados en los errores de medicación.

Capítulo II. Marco teórico

En el capítulo anterior se puntualiza la relación que existe entre el uso de métodos computacionales basados en medidas numéricas de la similitud que comparan los nombres de los medicamentos para predecir de manera automática el grado potencial confusión. En esta sección se define cada una de las medidas de similitud y distancia de edición utilizados en las investigaciones previas.

Medir la similitud o la distancia entre dos objetos es un requerimiento básico en el área de reconocimiento de patrones. La similitud presente entre textos o palabras tiene un papel importante en investigaciones y tareas automatizadas como: recuperación de información, clasificación y agrupación de texto, detección de temas, generación automática de preguntas o de respuestas, traducción, generación de resúmenes, entre otras. En este capítulo se abordan las medidas que han sido implementadas en el problema de la identificación de nombres confusos de medicamentos.

2.1. Medidas de similitud léxica

Las similitudes entre dos cadenas se miden de manera léxica y semántica. En el caso de la similitud léxica o basada en cadenas se enfoca en buscar una secuencia de caracteres comunes entre ambas [41].

La similitud entre dos objetos está relacionada con las características que comparten. Mientas más aspectos presenten en común más similares resultan ser. Aunque también se puede considerar sus disimilitudes para alcanza el mismo objetivo. La similitud máxima entre dos cadenas *A* y *B* es alcanzada cuando ambas son idénticas, sin importar la cantidad de características en común que se comparten [44].

En el caso de par de nombres de medicamentos el objetivo es medir secuencias de caracteres o una cadena. Las medidas de similitud léxicas utilizadas en nombres de medicamentos operan por lo tanto entre secuencias de cadenas y se mide su composición, entre la correspondencia de cadenas (*string matching*) o por comparación.

2.2. Clasificación de las medidas para el problema LASA.

Las medidas se pueden clasificar como de similitud (mientras más cerca el valor a cero más relacionados están los nombres) o de distancia (mientras mayor es el valor obtenido más relacionados existe entre los nombres) entre dos cadenas de texto [41], [45].

Para las medidas de similitud el valor obtenido depende de la escala de las mediciones y también del tipo de dato. Mientras que las de distancia miden las diferencias entre los nombres de medicamentos. Además, las medidas de similitud normalmente son normalizadas para tener una escala entre diferentes similitudes.

En el caso de las medidas de distancia cada modelo es definido por una función d, para cualquier conjunto de cadenas $\{s_1$ y s_2 y $s_3\}$ y satisfacen las siguientes propiedades:

$d(s_1, s_2) \geq 0$; la distancia entre dos cadenas debe ser siempre mayor o igual a cero.

$d(s_1, s_2) = 0$; la distancia solo es igual a cero si la cadena se mide con respecto de si misma.

$d(s_1, s_2) = d(s_2, s_1)$; la distancia es simétrica.

$d(s1, s3) \leq d(s1, s2) + d(s2, s3)$; la distancia debe satisfacer la inequidad triangular.

Una medida de distancia es considera una métrica si cumple las condiciones previas. De manera que, no todas las distancias son métricas pero si todas las métricas son distancias [45].

Tanto para la medida de similitud y de distancia, dependiendo de las características que evalúen al comparar ambas cadenas se pueden definir como ortográficas o fonéticas. Las medidas ortográficas se centran en la escritura de los nombres (como se ven) las fonéticas en como se pronuncian (como suenan). En la Tabla II-1 se resumen las medias encontradas en el estado del arte utilizadas para la identificación de pares LASA.

Tabla II-1. Medidas de similitud y distancia para la identificación ortográfica y fonética de pares LASA.

	Similitud	**Distancia**
Ortográfico	Prefix [30]	NED [23], [40], [46]
	NLCS [29]–[32]	NTED [23], [47], [48]
	Nsim (Bisim, and Trisim) [29]–[32]	OmissionKey [38], [49]
	Ngram (Bigram and Trigram) [30], [33]	SkeletonKey [38], [49]
	Bigram-(1B, 1B1A, and 1A)) [30], [33]	
	Trigram-(1B, 1A, 1B1A, 2B, 2A, 2B2A, 1B2A, and 2B1A) [33]	
Fonético	Aline [30], [34], [35]	Soundex [30], [33]
		Phonix [23], [28], [38]
		Editex [23]

2.3. Medidas para identificación de nombres confusos de medicamentos

En los trabajos relacionados con la identificación de pares LASA se proponen y evalúan distintas medidas individuales tanto de similitud y de distancias. En un primer intento, Lambert [7] evaluó tres medidas ortográficas (Bigram, Trigram y Edit Distance) entre nombres de medicamentos confusos y grupo de casos de control. Los valores de similitud obtenidos mostraron relación con respecto al riesgo de ocurrencia de los errores LASA. En un segundo trabajo [33] expandió la investigación anterior incluyendo 22 medidas de similitud y de edición tanto para medir el aspecto ortográfico y fonético. En una investigación posterior, Kondrak [29]–[32] propone dos nuevas medidas individuales para identificar la confusión: Nsim una medida ortográfica y Aline una medida fonética.

Una desventaja de utilizar y evaluar medidas individuales para el potencial riesgo de confusión entre un par de medicamentos es que no se sabe *a priori* si la

confusión fue ortográfica o fonética. Por lo que trabajos previos evalúan la eficacia de métodos que combinan distintas medidas individuales [30], [33].

En las siguientes secciones se describen de una manera original medidas utilizadas por el estado del arte de manera individual y en combinación.

2.3.1. Medidas ortográficas

El parecido de los nombres de medicamentos, durante una comunicación escrita principalmente en medios impresos o electrónico, da como resultado que los nombres resulten similares y ocurran confusiones. Por consiguiente, medias de distancia ortográficas ha sido implementadas para determinar identificar la disimilitud.

2.3.1.1. Distancia de edición

Dados dos nombres de medicamentos (X, Y) como secuencias de tamaño n y m, respectivamente. La distancia de edición (ED, por las siglas en inglés de *Edit Distance*) también llamada *Levenshtein* se refiere al costo de operaciones de edición mínimo (inserción, borrado y sustitución) para convertir la secuencia X en Y [23], [40], [46]. Algunas aplicaciones [23], [26] consideran que solo son necesarias dos operaciones (inserción y borrado) para realizar la sustitución, en ese caso la sustitución sería la suma de los costos de ambas operaciones (inserción y borrado). En esta investigación, todas las operaciones de edición tienen el costo de uno. El costo para sustituir la letra x_i por la letra y_i, se denota como $cs(x_i, y_i)$ donde i es igual a uno cuando son diferentes, e igual a cero en caso contrario. Por lo que la distancia de edición entre X y Y esta dada por $edit(n, m)$ y se calcula de la siguiente forma:

$$edit(i,j) = \begin{cases} max(i,j) & \begin{matrix} i = 0 \\ or \\ j = 0 \end{matrix} \\ edit(i-1,j-1) & x_i = y_j \\ min\begin{cases} edit(i-1,j)+1 \\ edit(i,j-1)+1 \\ edit(i-1,j-1)+cs(x_i,y_i) \end{cases} & x_i \neq y_j \end{cases}$$

Figura II-1. Función de recurrencia de la distancia de edición, donde $cs(x_i, y_i)$ donde i es igual a uno cuando son diferentes, e igual a cero en caso contrario.

Por ejemplo, la distancia de edición entre Zantac y Xanax es 3 porque la transformación mínima requiere dos sustituciones (Z→X y c→x) y una operación de borrado (letra t).

La distancia de edición normalizada (NED, por siglas en inglés de *Normalized Edit Distance*) es calculada por la división de la distancia de edición entre el tamaño de la secuencia mas larga [26], [30]–[33], [50], [51]. Por lo tanto, en el ejemplo anterior la NED es $3/6 = 0.5$.

2.3.1.2. Tapered Edit Distance

La medida *Tapered Edit Distance* (TED) es utilizada para encontrar nombres similares en bases de datos basados en el deletreo del nombre. La medida TED da mayor relevancia a las primeras coincidencias que a las últimas. Consecuentemente, el máximo costo de penalización de una sustitución y borrado al principio es mayor que el mínimo costo de penalización al final de las secuencias [23], [48]. La medida *Normalized Tapered Edit Distance* es la medida normalizada de TED.

2.3.1.3. SkeletonKey

El algoritmo *Skeleton* convierte una secuencia de letras en una clave. Esta clave consiste en la primera letra de la secuencia seguida de las restantes consonantes únicas en orden de aparición, al final se agregan las vocales únicas restantes en orden de aparición [25], [49]. Por ejemplo, los nombres de medicamentos Zantac y Contac

tienen la clave *Skeleton:* Zntca y Cntoa respectivamente. La distancia nombrada *SkeletonKey* calcula la distancia de edición entre las dos claves correspondientes a los dos nombres [25], [33].

2.3.1.4. OmissionKey

Análogo al algoritmo de las claves de *Skeleton*, la clase de *Omission* utiliza un orden invertido de la mayor frecuencia de consonantes omitidas envueltas en errores de escritura. El ranking de las consonantes que son frecuentemente omitidas tiene el siguiente orden: *RSTNLCHDPGMFBYWVZXQKJ*, donde *R* es la consonante más frecuente y *J* es la menos frecuente. La clave de *Omission* de una secuencia consiste en las consonantes en ordenadas en una frecuencia inversa seguidas de las vocales en orden de aparición en la secuencia [25], [49]. Por ejemplo, los nombres Zantac y Contac tienen las claves de Omission Zcnta y Cntoa, respectivamente. La distancia llamada *OmissionKey* utiliza la distancia de edición entre las claves correspondientes de los nombres de medicamentos.

2.3.1.5. Prefix

Dados los nombres de medicamentos X y Y como secuencias de tamaño m y n, respectivamente. Prefix representa la proporción de letras iniciales comunes continuas más largas [30]. Ver figura II-2, donde $max(|X|, |Y|)$ devuelve el tamaño de cadena más largo de X o Y. El prefijo común para Accutane y Accolate es Acc ($|Acc| = 3$) y le prefijo normalizado es 0.375.

$$Prefix(X, Y) = \frac{|x_1 = y_1, x_2 = y_2, \dots, x_i = y_i|}{max(|X|, |Y|)}$$

Figura II-2.Función de recurrencia de Prefix.

2.3.1.6. N-gram

La media de similitud N-gram representa una secuencia del conjunto de todas sus continuas subsecuencias (gramas) de tamaño *N* [25]. Por ejemplo, si $|X| = m$ y $N = 2$ (bigramas), entonces $X' = \{x_1x_2, x_2x_3, \dots, x_{m-1}x_m\}$. Dadas las secuencias X y Y, la similitud N-gram se define como la similitud de Dice entre los dos conjuntos X' y Y' de la siguiente manera:

$$Dice(X', Y') = \frac{2|X' \cup Y'|}{|X'| + |Y'|}$$

Figura II-3. Fórmula del coeficiente de Dice

Si se consideran los nombres de medicamentos $X = Zantac$ y $Y = Contac$ para una representación de bigramas entonces $X' = \{Za, an, nt, ta, ac\}$ y $Y' = \{Co, on, nt, ta, ac\}$. En el ejemplo anterior, $Dice(X', Y') = 6/10$. Lambert [7], [33], [51] usa similitudes en bigramas y trigramas.

La similitud entre N-gramas presenta una debilidad al identificar pares LASA porque es bien conocido que los prefijos y sufijos de los nombres de medicamentos están relacionados con la confusión [22], [30]. Para incrementar la sensibilidad de las medidas de similitud con N-gramas se han implementado variaciones con respecto a introducir letras al inicio o al final. Lambert propone agregar espacios (o una letra no incluida en el nombre) antes y después en ambos nombres de medicamentos para hacer que el principio y el final aparezca en uno o mas bigramas [33]. Por ejemplo, la medida Trigram-2B usa la similitud *Trigram* agregando dos espacios en ambos nombres al principio de las secuencias. Siguiendo el ejemplo, el nombre de medicamento Zantac es representado como:

$$X = --Zantac$$
$$X' = \{--Z, -Za, Zan, ant, nta, tac\}.$$

Figura II-4. Representación en trigramas del medicamento *Zantac*

Lambert utiliza variaciones de Bigramas: *Bigram-1B, Bigram-1B1A, y Bigram-1A) y Trigramas: Trigram-1B, Trigram-1A, Trigram-1B1A, Trigram-2B, Trigram-2A, Trigram-2B2A, Trigram-1B2A y Trigram-2B1A*.

2.3.1.7. Normalized Longest Common Subsequence

Una desventaja de la similitud de *Dice* (por lo tanto, también de la similitud de N-gramas) es que es aplicada a conjuntos pierde el orden de los elementos. La medida normalizada de la subsecuencia común más larga (*NLCS*, por sus siglas en inglés de *Normalized Longest Common Subsequence*) permite mantener un orden de la correspondencia entre letras comunes. Dada las secuencias X y Y de tamaño n y m respectivamente. La medida *NLCS* es definida como el promedio de el tamaño de la longitud común más larga entre X y Y, $NLCS = |lcs(n,m|/max(|X|,|Y|)$, donde $lcs(n,m)$ puede ser calculado con la recurrencia de la Ecuación X [30]–[33], [51]. Por ejemplo, con los nombres de medicamentos Zantac y Contac el tamaño de la longitud común más larga es $|ntac| = 4$, por lo tanto $NLCS = 4/6$. La NLCS es usada por [30]–[33], [51].

$$lcs(i,j) = \begin{cases} 0, & \begin{matrix} i = 0 \\ or \\ j = 0 \end{matrix} \\ lcs(i-1,j-1)+1, & x_i = y_j \\ max\big(lcs(i,j-1), lcs(i-1,j)\big) & x_i \neq y_j \end{cases}$$

$$NLCS = |lcs(n,m|/max(|X|,|Y|)$$

Figura II-5. Función de recurrencia de la *NLCS*

Por otra parte, la medida de similitud con N-gramas permite manejar pequeñas subsecuencias, pero pierde el orden de las posiciones donde coinciden. Mientras que la similitud NLCS mantiene el orden de las posiciones que coinciden, pero solo por letras y no da relevancias a las primeras letras del nombre del medicamento.

2.3.1.8. Nsim

Es una media de similitud ortográfica propuesta por Kondrak [29]–[32]. Esta medida combina características implementadas por los *ngramas* de tamaño. restricción de no enlaces cruzados y la primera letra es repetida al principio del nombre del medicamento. Un caso particular de *Nsim* es la mediad *Bisim*. Dadas dos secuencias de caracteres (con la primera letra repetida) X y Y que representan los nombres de medicamentos de tamaño n y m, respectivamente. La medida *Bisim* se define como:

$$Bisim(X,Y) = \frac{nsim(n,m)}{max(n,m)}$$

$$nsim(i,j) = \begin{cases} 0, & i = 0 \vee \mathrm{j} = 0 \\ max \begin{cases} nsim(i,j-1), \\ nsim(i-1,j), \\ nsim(i-1,i-1) + \\ s\left(x_i x_{i+1}, y_j y_{j+1}\right), \end{cases} & \text{en otro caso} \end{cases}$$

$$s\left(x_i x_{i+1}, y_j y_{j+1}\right) = \frac{1}{2}\sum_{k=0}^{1} id(x_{i+k}, y_{j+k})$$

$$id(a,b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases}$$

Figura II-6. Función de recurrencia de Bisim.

2.3.2. Medidas fonéticas

La confusión fonética ocurre como resultado por la pronunciación que tienen los nombres de medicamento donde se utiliza una comunicación verbal.

2.3.2.1. Soundex

Es un algoritmo de indexación desarrollado por Russell [26] que agrupa letras con sonido similar con la finalidad de crear un código que representa un nombre. Debido a que *Soundex* no es una medida, Lambert [33] y Kondrak [30] usan el

algoritmo de *Soundex* para codificar un par de nombres de medicamentos, pero la distancia entre los nombres es obtenida por la distancia de edición. En la presente investigación se implementa la distancia *Soundex* de acuerdo con lo que describen en sus artículos Lambert [33] y Kondrak [33]. Se comienza con la segunda letra de un nombre de medicamento, *Soundex* reemplaza cada letra de el nombre por un código numérico, después todos los ceros son suprimidos (*sup*) y la secuencia resultante es truncado a cuatro símbolos [30], [33], [52]. Dado un par de medicamentos como una secuencia X y Y de tamaño n y m, respectivamente. La distancia *Soundex* es definida en la ecuación descrita de la Figura II-7.

Por ejemplo, los nombres de medicamentos Zantac y Xanax son codificadas en Z532 y X520, respectivamente, estos códigos tienen una distancia de 3.

$$Soundex(X,Y) = Edit(Code(x_{1..n})_{1..4}, Code(y_{1..m})_{1..4})$$

$$Code(\alpha_i) = \begin{cases} \alpha_1, & \alpha_{i=1} \\ sup, & \alpha_{i>1} \in \{a,e,h,i,o,u,w,y\} \\ 1, & \alpha_{i>1} \in \{b,f,p,v\} \\ 2, & \alpha_{i>1} \in \{c,g,j,k,q,s,x,z\} \\ 3, & \alpha_{i>1} \in \{d,t\} \\ 4, & \alpha_{i>1} \in \{l\} \\ 5, & \alpha_{i>1} \in \{m,n\} \\ 6, & \alpha_{i>1} \in \{r\} \end{cases}$$

Figura II-7. Función de recurrencia de la medida fonética de distancia Soundex

2.3.2.2. Distancia Phonix

Es una medida similar a *Soundex*. Es un algoritmo de indexación que usa 160 grupos de letras para codificar un nombre de medicamento. Debido a que Phonix no es una medida, la distancia de edición es calculada para obtener la distancia Phonix entre un par de nombres de medicamentos X y Y de tamaño n y m, respectivamente. La distancia Phonix se define como:

$$Phonix(X,Y) = Edit(\alpha, \beta)$$

Donde $\alpha = PhonixCode(X)_{1..8}$ y $\beta = PhonixCode(Y)_{1..8}$. Dado un nombre Z, el código Phonix sigue el siguiente algoritmo:

1. *Se realiza la sustitución fonética al reemplaza ciertos grupos ortográficos de letra por otro grupo de letras.*
2. *Se reemplaza la primera letra por V si es una $vocal$ o la consonante Y*
3. *Quitar el sonido final del nombre (la ultima parte después de la $vocal$ o la Y)*
4. *Eliminar todas las vocales, las consonantes H, W, Y y todas las letras repetidas consecutivas.*
5. *Codificar el prefijo del nombre reemplazado la letra α_i con la siguiente función (ver Ecuación X). El máximo tamaño de un código Phonix esta restringido a ocho caracteres.*

$$CodePrefix(\alpha_i) = \begin{cases} \alpha_1, & \alpha_{i=1} \\ 1, & \alpha_{9>i>1} \in \{b, p\} \\ 2, & \alpha_{9>i>1} \in \{c, g, j, k, q\} \\ 3, & \alpha_{9>i>1} \in \{d, t\} \\ 4, & \alpha_{9>i>1} \in \{l\} \\ 5, & \alpha_{9>i>1} \in \{m, n\} \\ 6, & \alpha_{9>i>1} \in \{r\} \\ 7, & \alpha_{9>i>1} \in \{f, b\} \\ 8, & \alpha_{9>i>1} \in \{s, x, z\} \\ sup, & in\ other\ case \end{cases}$$

6. *Codificar el sonido final reemplazando cada letra α_i de acuerdo con su valor numérico definido en la última función CodePrefix. La longitud máxima para un código fónico de un sonido final esta restringida a 8 caracteres.*

2.3.2.3. Editex

La distancia Editex calcula la distancia de edición entre grupos fonéticos de letras de nombres de medicamentos [23]. En este caso, el costo de las operaciones de edición depende del grupo de letras que son comparadas (ver Tabla II-2) si dos letras son iguales el costo de sustitución es cero. Sin embargo, si son diferentes y están en el mismo grupo entonces el costo de sustitución es uno. En otro caso, todas las otras operaciones de edición tienen un costo de dos [30]. Por ejemplo, los nombres de medicamentos Zantac y Xanax tienen una distancia Editex de 5, por que el costo de sustitución de $Z \rightarrow X$ es uno debido a que están en el mismo grupo, y el costo de

sustitución para $c \rightarrow x$ es dos debido a que están en diferentes grupos, y el costo de borrado de t es dos.

Tabla II-2. Grupo de letras para Editex

Código	EDITEX
0	a, e, i, o, u, y
1	b, p
2	c, k, q
3	d, t
4	l, r
5	m, n
6	g, j
7	f, p, v
8	s, x, z
9	c, s, z

2.3.2.4. Aline

Identifica los cognados en palabras de lenguajes relacionados. Aline mide el mejor alineamiento entre secuencias de fonéticas de nombres de medicamentos. La similitud Aline es normalizada por el tamaño de la secuencia mas larga multiplicado por el valor de similitud máximo entre segmentos [34], [35].

2.3.3. Medidas combinadas

Debido a que no es posible conocer la causa de la confusión (ortográfica o fonética) *a priori*, distintas medidas individuales son combinadas. Una medida combinada usada para tomar ventaja de las fortalezas particulares de cada medida mediante un ensamble para dar como resultado una medida final entre dos nombres de medicamentos.

Después de que Lambert evaluará 22 medidas individuales [33] para la identificación de pares LASA fueron seleccionadas las tres mejores medidas Trigram-2B, NED y Editex para participar en un Método de Regresión Logística (LR, por las siglas en inglés de *Logistic Regression*) y combinar sus cualidades en una nueva medida que obtuviera mejores resultados. La Regresión Logística es un algoritmo de predicción de clases binario. Como otros algoritmos de aprendizaje automático supervisado (ML por sus siglas en inglés de *Machine Learning*), La regresión logística implementa un algoritmo de aprendizaje estándar para ajustar el modelo de predicción a partir de un conjunto de datos de entrenamiento [53]. En la RL de Lambert los parámetros ajustados $\theta = \{\theta_0, \theta_1, \dots, \theta_n\}$ para la hipótesis $0 \leq h_\theta(x) \leq 1$ para clasificar un par (X, Y) como confuso o no confuso con respecto a:

$$h_\theta(x) = g(y(x))$$

Donde:

$y(x) = (\theta^T x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$ representa el modelo lineal,

$x_1 = Editex(X, Y)$,

$x_2 = NED(X, Y)$,

$x_3 = Trigram2B\ (X, Y)$ son los resultados obtenidos de cada medida individual,

$g\big(y(x)\big) = \frac{1}{1+ e^{-y(x)}}$ es la función sigmoide.

Figura II-8. Modelo de regresión logística.

Otro modelo que combina medidas individuales para la identificación de pares LASA fue propuesto por Kondrak [30]. Este modelo utiliza el promedio para combinar n medidas. Además, se probó de manera manual diferentes combinaciones de medidas, pero el método reportado que presentó mejores resultados solo utiliza cuatro medidas (AVG-4). La medida AVG-4 se describe en la Figura II-9, donde $m_1 = Prefix$, $m_2 = NED$, $m_3 = Aline$, and $m_4 = Bisim$.

$$AVG - n(\mathrm{X},\mathrm{Y}) = \frac{\sum_{k=1}^{n} m_k(X,Y)}{n}$$

Figura II-9. Medida combinada AVERAGE-4

2.4. Resumen

Para determinar un valor de similitud entre dos nombres de medicamentos se utilizan medidas individuales que se clasifican con base a las características que miden: similitud o distancia. Además, se pueden clasificar por el enfoque ortográfico o fonético. En este capítulo se describe todas las medidas del estado del arte que han sido implementadas individualmente o en combinación. Las medidas individuales son descritas de acuerdo con su capacidad de medir el aspecto ortográfico o fonético. En cada descripción se indica el detalle del funcionamiento, el algoritmo o la función de recurrencia de cada medida. Al final se explican las medidas combinadas. Las cuales son el resultado de un ensamble de medida individuales que consideran el aspecto ortográfico y fonético debido a que no se puede saber de manera *a priori* si la confusión será resultado del parecido visual de los nombres en medios escritos o por la pronunciación que presenta en una comunicación verbal. Un primer modelo combina tres medidas individuales (Trigram2B, NED y Editex). Este modelo utiliza la regresión logística para predecir si un par de nombres es confuso. En otra medida combinada se utiliza el promedio aritmético para obtener un solo valor de similitud de las medidas Aline, Bisim, NED y Prefix. Estas propuestas no muestran evidencia del proceso de selección específico de solo un subgrupo de medidas a ser combinadas, a pesar de identificar varias medidas individuales que se distinguen por medir diferentes aspectos de similitud.

Capítulo III. Metodología general

De acuerdo con las estrategias para reducir la formación de nuevos pares confusos de medicamentos la Administración de Alimentos y Medicamentos (FDA, *Food and Drug Administration*) sugiere que es necesario contar con herramientas computacionales que sirvan como apoyo en el proceso de revisión de nombres propuestos de medicamentos. Estas herramientas deben tener un grado de validez y confianza que permita identificar *a priori* los potenciales pares de nombres de medicamentos confusos con base en los casos de confusión previamente reportados en los errores de medicación. A partir de este enfoque se ha propuesto una metodología que permite medir y comparar la exactitud para recuperar los pares de nombres de medicamentos confusos. Por lo que, en primer lugar, se presenta una formalización del problema de la identificación de pares LASA como un proceso de recuperar un conjunto de pares de nombres que pertenecen a el conjunto de pares LASA previamente identificados en los reportes de errores de medicación. Después se presenta la lista de pares LASA utilizada en la experimentación y los principios

utilizados para validar y estimar el grado de confianza de los métodos propuestos en esta tesis.

El enfoque presentado en esta sección ha sido utilizado para dar respuesta a las hipótesis planteas en el capítulo I. Además, representa la base de cada uno de los enfoques metodológicos presentados en las soluciones expuestas en esta tesis.

3.1. Formalización del problema de la identificación de pares LASA

Un nombre de medicamento confuso X con n letras sea considerado como una secuencia de elementos, es decir, $X = < x_1, x_2, \dots, x_n >$ donde $|X|$ denota el largo (tamaño) de X.

Dado un conjunto $D = \{d_1, d_2, \dots, d_m\}$ de potenciales nombres de medicamentos confusos y un conjunto de pares LASA, se define como:

$$T = \left\{(d_i, d_j) \middle| \left((d_i, d_j) \in \lambda \Longrightarrow (d_j, d_i) \in \lambda\right) \wedge i \neq j\right\}$$

Figura III-1. Conjunto de pares LASA previamente encontrados en los reportes de errores de medicación

Donde λ es una lista de pares LASA, y $T \subseteq DXD$. El problema de la identificación de nombres confusos de medicamentos del conjunto D consiste en recuperar un subconjunto H de todos los pares que pertenecen a T, que es definido como:

$$H = \left\{(d_i, d_j) \middle| \left((d_i, d_j) \in T \vee (d_j, d_i) \in T\right) \wedge i \neq j\right\}$$

Figura III-2. Conjunto recuperado por un método para identificar potenciales pares confusos

De acuerdo con los trabajos relacionados con el estado del arte, para identificar un par de nombres confusos de medicamentos no es suficiente utilizar medidas individuales para captura patrones en el parecido de cómo se ven (ortográfico) o como

suenan (fonético) entre los nombres, si no que también es necesario combinar los resultados de similitud de las medidas individuales ya que no se puede conocer *a priori* la causa de la confusión (error ortográfico o fonético).

3.2. Lista de medicamentos confusos

En esta tesis se ha utilizado una lista de pares de nombres confusos obtenidos de los reportes de errores de medicación en Estados Unidos [54]. Por lo que la similitud ortográficas y fonéticas de estos casos están relacionados con el idioma inglés. La lista contiene 858 pares de nombres de medicamentos confusos, de los cuales 630 son nombres únicos de medicamentos. A partir de la lista de nombres únicos se han formado 396,900 pares de nombres. Solo el 0.3 por ciento son casos de confusión lo que representa una relación de 1:460 entre pares confusos y aquellos que no lo son.

3.3. Evaluación de eficacia de los métodos de recuperación de nombres confusos de medicamentos.

Para medir la eficacia de recuperación de pares confusos de nombres de medicamentos se utiliza la métrica de medida F (*F-measure*) [55], [56]. Dado un par LASA $(d_i, d_j) \in T$, el resultado de *F-measure* de la consulta d_i evalua el tamaño de el conjunto recuperado de nombres recuperados en el ranking 1 (nombres similares muy cercanos a la consulta d_i), pero si d_j no aparece en el último conjunto, el *F-measure* es sumado al tamaño de los nombres de medicamentos recuperados en el siguiente ranking, hasta que aparezca d_j. De este modo, *F-measure* evalúa la habilidad de recuperar los nombres de medicamentos relevantes (confusos) de una consulta.

Por último, para calcular un valor que represente el *F-measure* de diferentes consultas de un conjunto de nombres D se define como:

$$F - measure(D) = \frac{2RP}{R + P}$$

Donde, R representa la exhaustividad y es calculada de la siguiente forma: $R(D) = \frac{|T \cap H|}{|H|}$ y P es la precisión $P(D) = \frac{|T \cap H|}{|H|}$.

El F-measure puede ser calculado en distintos rankings. Incluso es preferible mejorar el F-measure en los primeros rankings. Por lo tanto, también se calcula mediante *macro-averaging* [55] un valor de *F-measure* de todas las consultas de los diferentes nombres de medicamentos (conjunto D) basa en la suma de los primeros rankings.

En otras palabras, esta propuesta asigna una relevancia mayor a las medidas individuales o combinadas que calculen valores de similitud que permitan recuperar todos los pares confusos de un conjunto de nombres de medicamentos únicos.

3.4. Resumen

En este capítulo se muestran los elementos claves que definen los métodos propuestos en la investigación y que guiaron el desarrollo de la metodología para verificar las hipótesis planteadas al inicio. Primero, el problema es define de manera formal mediante un enfoque basado en la recuperación de información relevante. Se describe la lista de pares confusos utilizada en la experimentación y las métricas utilizadas para validar los resultados obtenidos.

Capítulo IV. Identificación de pares de nombres confusos de medicamentos mediante combinación de medidas

Es este capítulo se incluye el artículo titulado: *An Evolutionary Logistic Regression Method to Identify Confused Drug Names*. El artículo fue aceptado y publicado por la revista especializada arbitrada e indexada de reconocimiento internacional titulada: *Journal of Intelligent & Fuzzy Systems*, de la editorial *IOS Press*, con *índice JCR* y un factor de impacto de 1.637 (2020). Publicado el 14 de mayo de 2019. Se anexa la carta de aceptación (ver Figura III-1).

4.2. Ejemplar de autor del artículo publicado

Journal of Intelligent & Fuzzy Systems 36 (2019) 4609–4619
DOI:10.3233/JIFS-179012
IOS Press

An evolutionary logistic regression method to identify confused drug names

Christian Eduardo Millán-Hernández*, René Arnulfo García-Hernández and Yulia Ledeneva
Autonomous University of the State of Mexico, Instituto Literario, col. Centro, Toluca, Mexico

Abstract. Confused drug names are a common cause of medication errors, and are related to look-alike and sound-alike drug names. For the problem of identifying confused drug name pairs, individual similarity measures are used between the drug names. In the state-of-art, a logistic regression with the standard learning algorithm has been used to combine individual similarity measures. However, only three similarity measures have been combined but the results of previous research do not outperform with a statistical significance to any individual measure. In addition, the problem of potential confused drug names pairs presents a high unbalanced distribution of dataset that it is a hard problem to supervised machine learning models. In this paper, an improved combined logistic regression measure based on 21 individual measures is presented with the standard learning algorithm. Also, we present an evolutionary learning method for a combined logistic regression measure that allows to learn an unbalanced dataset. According to the experimentation with a gold standard dataset, our proposed combined measures outperform previous research with a statistical significance to identify pairs of confused drug names. In addition, the rankings of individual and combined similarity measures are presented.

Keywords: Look-alike sound-alike drug names, patient safety, logistic regression, genetic algorithm, imbalanced dataset.

1. Introduction

Medication names that sound and look similar to others are related to medication errors. Drug names mix-up is a risk to patient safety that causes at least one death per day and harming approximately 1.5 million people per year in the United States. According to the World Health Organization (WHO), the estimated annual cost of medication errors is around $42 billion USD, an additional annual cost of $3.5 billion USD for patients who suffer harm [46].

Look-Alike/Sound-Alike drug names (LASA) are the most common cause of medication errors worldwide. LASA leads pharmacists, nurses, patients, doctors, and others health care to the unintended interchange of medicine brand names at different stages of the administration process that can result in patient injury or death [1, 5, 38, 46]. Name confusion occurs as the result of a weak medication system and human errors-related factors. These undesired medication errors are potentially preventable events [42].

Medication errors are voluntarily and anonymously reported in national reporting programs by the health practitioners and consumers. The National Medication Errors Reporting Program (MERP) in the United States is an internationally recognized program [11]. The MERP and similar programs are used for determining the causes about the medication errors to obtain stronger medication systems [6, 16, 17].

Regulatory authorities in the United States, Canada, and like in other countries are improving the reports systems to identify the factors and potential causes of LASA. Strategies to decrease potential LASA medication errors are to educate the healthcare community and work with others regulatory

*Corresponding author. Christian Eduardo Millán-Hernández. E-mail: ceduardo.millan@gmail.com.

ISSN 1064-1246/19/$35.00

organizations, professionals, manufacturers, and patients; to improve the safe medication practices.

About 25 percent of the errors reported in the Institute for Safe Medication Practices (ISMP) corresponds to LASA confusion problem [7]. In 2002, according to the U.S. Pharmacopeia (USP) [20] there were reported 192,477 medication errors. Fortunately, in 67,707 cases were possible to intercept the medication before it was administrated to the patient. In 91,446 cases, the medication was administrated but it did not damage the patient. However, in 2,600 cases there were required an intervention to keep the patient alive, but in 20 cases the patient dead. Since, the errors were detected before writing the prescription of the medication the rest of them are classified as potential.

The ISMP publishes an updated *List of Confused Drug Names* [28]. The ISMP list is used to know which medications need special attention and safeguard by healthcare practitioners to reduce LASA errors and patient harm.

LASA problem is growing continuously [7, 8, 21]. Since the first LASA list was published in 1973, LASA lists has been updated frequently with new pairs [41]. In 1995, USP published the *Quality Review 49* with 200 confusable pairs [21], in 2001 the list grows to 850 pairs in the *Quality Review 76* [43], in 2004 to 1,950 pairs in the *Quality Review 79* [44] and in 2006 to 3,170 pairs [21]. Recently in 2017, the ISMP reports that 36 drug names have been added because they are involved in 17,133 medication errors. A drug name may also participate in several pairs of LASA [21], for example, *Serentil* participates in four pairs related to other drug names [43].

Multiple human errors-related factors are involved in the LASA confusion problem [21, 30]. For example, although LASA *Avandia* and *Coumadin* pair does not have a similar spelling, it is classified as a visual perception error when is prescribed in a poor handwritten letter. Confusion occurs when the first capital letter "A" (*A*) looks like "C" (*C*) and the last letters "ia" (*ia*) looks like "in" (*in*) [10, 30]. Another visual perception error takes place when the LASA *Hydroxyzine* and *Hydralazine* pair is typewritten communicated because they have a similar spelling, it means they share identical prefixes, suffixes, and lengths [10]. For example, the LASA *Xanax* and *Zantac* pair does not have a similar spelling, but it is classified as an auditory perception error because it sounds similar [21]. Sometimes in this kind of auditory errors, the pair shares some features related to similar spelling.

Also, short-term memory errors are related to memory lapses, for example, when the pharmacist, after reading the name *Hydroxyzine,* writes *Hydralazin*e [21, 30].

A motor control error occurs when an incorrect medicine is selected from a computerized down-list [21, 30].

While the type of error is not easy to be identified, the root-cause of similarity could be detected or avoided. As part of the strategy to reduce the risk of registering new confused drug names, the U.S. Food and Drug Administration (FDA) needs to identify potentially confused drug names *a priori* [5].

In the approval process for a new drug, the FDA encourage the implementation of computerized methods and algorithms to evaluate the similarity to the proposed name [10]. First, FDA recommends that the industry follow a series of best practices in the selection of drug proprietary name for new medications [12]. In the review process of the proposed name, it is evaluated by the Phonetic and Orthographic Computer Analysis (POCA) software for comparing the name against different drug databases [10, 12, 13]. Once the drug name is reviewed, FDA could reject the name if it is too similar to existing previously registered drug names [12].

A preventive action is taken when a confusable drug name is responsible of serious medication errors, in these cases, the proprietary name has been changed in order to avoid the error with another drug names [5].

Recent research proposes focusing on homologating the review process and implementing automated processes based on machine learning techniques to solve the challenge [15, 39, 42].

In this paper, we demonstrate how a logistic regression model outperforms a unique similarity measure model when the model is appropriately trained for identifying confusing names.

The problem of identifying LASA pairs is defined as, **definition 1.** Let a *confused drug name X* with n letters be considered a sequence of elements, it means $X =< x_1, x_2, \ldots x_n >$, where $|X|$ denotes the length (size) of X.

Given a set $D = \{d_1, d_2, \ldots, d_m\}$ of potentially confused drug names, and a set of look-alike-and-sound-alike pairs, defined as:

$$T = \{(d_i, d_j)|((d_i, d_j) \in \lambda \Rightarrow (d_j, d_i) \in \lambda) \wedge i \neq j\} \quad (1)$$

where λ is the LASA list, and $T \subseteq D \times D$. The problem of identifying confused drug names from

D consists in retrieve a subset H of all the pairs that belongs to T, that is defined as:

$$H = \{(d_i, d_j) | ((d_i, d_j) \in T \vee (d_j, d_i) \in T) \wedge i \neq j\} \tag{2}$$

According to related works [4, 24–27, 29–32], for identifying a pair of confusable drug names not only it is needed to use individual measures for capturing particular look-alike (orthographic cause) and sound-alike (phonetic cause) patterns between the names, but also it is necessary to combine the previous results because it is not possible to know *a priori* the cause of the confusion (*i.e.* an orthographic or phonetic). In both cases, the measures are classified as distance (as closer to zero as more related are the names) and similarity (as greater is the value as more related are the names). Normally, similarity measures are normalized to have a scale between different similarity values.

Lambert [18] presents a wide compilation of 22 measures for LASA problem, [30–32] where the classical string-matching and based-distance measures (and some variants of them), are used. After evaluating individual measures, Lambert concludes Trigram-2B is the best orthographic similarity measure, Normalized Edit Distance (NED) is the best orthographic distance measure, and Editex is the best phonetic distance measure. However, only a subset of selected measures is used in a Logistic Regression Model (LRM) to combine the strengths of them.

The individual measures found in related work to this problem are present below.

1.1. Orthographic distance measures

Given the drug names X and Y as sequences of size n and m, respectively, *Edit distance* (also called *Levenshtein*) refers to the minimum cost of editing operations (insertion, deletion and substitution) to convert the sequence X into Y [33, 45, 47]. Some applications [9, 47] consider that there is needed two operations (insertion and deletion) for a substitution, in this case the cost of substitution is the sum of the cost of insertion and deletion. In this paper, all editing operations have a cost of 1, it is, the cost for substituting the letter x_i by the letter y_i, denoted as: $cs(x_i, y_i)$ is 1 when they are different, or 0 in other case. In this case, the edit distance between X and Y is given by $edit\,(n, m)$ computed by the following recurrence:

$$edit(i, j) = \begin{cases} \max(i, j) & \begin{array}{l} i = 0 \\ or \\ j = 0 \end{array} \\ edit(i-1, j-1) & x_i = y_i \\ \min \begin{cases} edit(i-1)+1 \\ edit(i, j-1)+1 \\ edit(i-1, j-1)+cs(x_i, y_i) \end{cases} & x_i \neq y_j \end{cases} \tag{3}$$

For example, the edit distance between *Zantac* and *Xanax* is 3 because the minimum transformation involves two substitutions ($Z \rightarrow X$ and $c \rightarrow x$) and one deletion (letter t).

A Normalized Edit Distance (NED) is computed by dividing the total *edit distance* between the length of the longer sequence [4, 9, 24–26, 31, 32]. For the above example, the NED is $3/6 = 0.5$.

A Tapered Edit Distance (TED) is used for finding similar names on databases based on the pronunciation of the names. TED gives more relevance to the first coincidences than the last ones. For this, the maximum cost of penalization for a substitution and deletion at the beginning is greater than the minimum cost for a penalization at the ending of the sequences [3, 19, 47]. *Normalized Tapered Edit Distance* (NTED) is the normalized distance of TED.

Skeleton is an algorithm to convert a sequence of letters in a key. This key consists of the first letter of the sequence followed by its remaining unique consonants (in order of appearance) followed by its remaining unique vowels (in order of appearance) [36, 37]. For example, the drug names *Zantac* and *Contac* have the Skeleton keys *Zntca* and *Cntoa*, respectively. The distance named *SkeletonKey* uses the edit distance between corresponding keys of the drug names [31–36].

Analogous to Skeleton key algorithm, *Omission* key uses the inverse order of the most frequently omitted consonants involve in spelling errors to build a key from a name for spelling correction. The ranking of the consonants presents the following order: *RSTNLCHDPGMFBYWVZXQKJ*, where the *R* is the most frequent consonant and *J* the less frequent consonant. The omission key of a sequence consists of the unique consonants in the above inverse frequency order followed by the vowels in appearance order [36, 37]. For example, the drug names *Zantac* and *Contac* have the omission keys *Zcnta* and *Cntoa*, respectively. The distance called *OmissionKey* uses the edit distance between the corresponding keys of the drug names [31–36].

1.2. Orthographic similarity measures

N-gram similarity represents a sequence of the set of all its contiguous subsequences (grams) of size N [36]. For example, if $|X| = m$ and $N = 2$ (bigrams), then $X' = \{x_1x_2, x_2x_3, \ldots, x_m - 1x_m\}$ [4, 26, 31]. Given the sequences X and Y, the *N-gram similarity* is defined as the *Dice similarity* [2] between the sets X' and Y' in the next way:

$$Dice(X'Y') = \frac{2|X' \cup Y'|}{|X'| + |Y'|} \tag{4}$$

Considering the drug names $X = Zantac$ and $Y = Contac$ for a bigram representation, then $X' = \{Za, an, nt, ta, ac\}$ and $Y' = \{Co, on, nt, ta, ac\}$. In the above example, $Dice(X', Y') = 6/10$. Lambert [29, 31, 32] uses bigram and trigram similarity.

N-gram similarity presents a weakness with the LASA problem because it is well-known that the prefixes and suffixes of the drug names are involved in their confusion [26, 40]. For increasing the sensitivity of the *N-gram similarity* variations with respect to initial and final letters are introduced. Lambert [31] proposes to add spaces (or a letter not included in the names) (B)efore and (A)fter in both drug names to make that the initial or final letters appear in one or more n-grams [31]. For example, the Trigram-2B measure uses the t*rigram similarity* adding two spaces in both drug names. Following the example, the drug name *Zantac* is represented as:

$$\begin{aligned} X &= --Zantac \\ X' &= \{--Z, -Za, Zan, ant, nta, tac\}. \end{aligned} \tag{5}$$

Lambert uses the variants of Bigram (1B, 1B1A and 1A) and Trigram (1B, 1A, 1B1A, 2B, 2A, 2B2A, 1B2A and 2B1A).

One disadvantage of the Dice similarity (therefore of the N-gram similarity, too) is that it is applied in sets; losing the order of the elements. The *Normalized Longest Common Subsequence (NLCS) similarity* lets to maintain an order in the common matching letters. Given the sequences X and Y of size n and m, respectively, the *NLCS similarity* is defined as the ratio of the length of the longest common subsequences between X and Y, $NLCS = |lcs(n, m)|/\max(|X|, ||Y|)$, where $lcs(n, m)$ can be calculated by the recurrence in Equation (7) [24–27, 31, 32]. For example, with the drug names *Zantac* and *Contac* the length of the longest common subsequence is $|ntac| =$, therefore the NLCS = 4/6. The NLCS is used by [4, 24–27, 31, 32].

$$lcs(i, j) = \begin{cases} 0 & \begin{array}{r} i = 0 \\ or \\ j = 0 \end{array} \\ lcs(i-1, j-1) + 1, & x_i = y_i \\ \max(lcs(i, j-1), lcs(i-1, j)) & x_i \neq y_j \end{cases} \tag{6}$$

On the one hand, N-gram similarity lets to manage small subsequences, but it loses the order of the matching positions. On the other hand, NLCS similarity maintains an order in the matching positions but only for letters and it does not give relevance to the first and initial position.

1.3. Phonetic distance measures

Soundex is an indexing algorithm developed by Russell [9] that groups letters with similar sounds in order to get a coded representation from a name. Since the Soundex algorithm is not a measure, Lambert [31] and Kondrak [26] use the Soundex algorithm for coding a pair of drug names, but the distance between the names is obtained using the edit distance. In this paper, the *Soundex distance* implemented follows the description of Lambert [31] and Kondrak [26]. Beginning with the second letter of a drug name, *Soundex distance* replaces each letter of a drug name by a numeric code, then all zeros are suppressed (*sup*) and the resulting sequence is truncated to four symbols [26, 31–35]. Given a pair of drug names as sequences X and Y of size n and m, respectively, *Soundex distance* is defined in Equation (7).

For instance, drug names *Zantac* and *Xanax* are coded as *Z532* and *X520*, respectively; and these codes have an edit distance of 3.

$$Soundex(X, Y) = Edit(Code(x_{1..n})_{1..4}, Code(y_{1..m})_{1..4})$$

$$Code(\alpha_i) = \begin{cases} \alpha_1, & \alpha_i = 1 \\ \text{sup}, & a_{i>1} \in \{a, e, h, i, o, u, w, y\} \\ 1, & \alpha_{i>1} \in \{b, f, p, v\} \\ 2, & \alpha_{i>1} \in \{c, g, j, k, q, s, x, z\} \\ 3, & \alpha_{i>1} \in \{d, t\} \\ 4, & \alpha_{i>1} \in \{l\} \\ 5, & \alpha_{i>1} \in \{m, n\} \\ 6, & \alpha_{i>1} \in \{r\} \end{cases} \tag{7}$$

Phonix is similar to *Soundex*, it is just an indexing algorithm that uses 160 groups of letters for coding a drug name [14, 36, 47]. Since *Phonix* is not a measure, the edit distance is computed to obtain the *Phonix distance* between a pair of coded drug names. Given a pair of drug names as sequences X and Y of size n and m, respectively, *Phonix distance* is defined as:

$$Phonix(X, Y) = Edit(\alpha, \beta) \tag{8}$$

where $\alpha = PhonixCode(X)_{1..8}$,
and $\beta = PhonixCode(Y)_{1..8}$.

Given a name Z, *PhonixCode* follows the next algorithm [36]:

1. Replace groups of orthographic letters by letters representing certain phonetic groups.
2. Replace the first letter by V if it is a vowel or the consonant Y.
3. Split the ending-sound from the name (roughly the part after the last vowel or Y).
4. Removed all vowels, the consonant H, W, Y and all duplicated consecutive letters.
5. Code the prefix of the name by replacing the letter α_i with the next *CodePrefix* function, see Equation (9). The maximum length of a *Phonix* code is restricted to 8 characters.

$$CodePrefix(\alpha_i) = \begin{cases} \alpha_1, & a_{i=1} \\ 1, & \alpha_{9>i>1} \in \{b, p\} \\ 2, & \alpha_{9>i>1} \in \{c, g, j, k, q\} \\ 3, & \alpha_{9>i>1} \in \{d, t\} \\ 4, & \alpha_{9>i>1} \in \{l\} \\ 5, & \alpha_{9>i>1} \in \{m, n\} \\ 6, & \alpha_{9>i>1} \in \{r\} \\ 7, & \alpha_{9>i>1} \in \{f, b\} \\ 8, & \alpha_{9>i>1} \in \{s, x, z\} \\ \text{sup}, & \textit{in other case} \end{cases} \tag{9}$$

6. Code the ending-sound by replacing every letter according to its numerical value defined in the last *CodePrefix* function. The maximum length for a Phonic code of an ending-sound is restricted to 8 characters.

Editex distance computes the edit distance between the phonetic groups of the letters of the drug names [47]. In this case, the cost of the editing operations depends on the group of the letters (see Table 1) that are compared. If two letters are equal then the cost for a substitution is zero. However, if they are different but they belong to the same group then the substitution cost is one. Otherwise, all other editing operations have a cost of two [26]. For example, the drug names *Zantac* and *Xanax* have an *Editex distance* of 5, because the substitution cost for $Z \rightarrow X$ is one due to both are in the same group, and the substitution cost for $c \rightarrow x$ is two due to they are in distinct groups; and the deletion cost for t is two.

Table 1
Group of letters for Editex

Code	EDITEX
0	a, e, i, o, u, y
1	b, p
2	c, k, q
3	d, t
4	l, r
5	m, n
6	g, j
7	f, p, v
8	s, x, z
9	c, s, z

1.4. Combined measures

Since, it is not possible to known the cause of the confusion (orthographic or phonologic) *a priori*, several orthographical and phonological measures are combined [4, 26, 31]. A combined measure is used to take advantage of strengthens of the individual measures as an ensemble for giving the final similarity result between two drug names.

First, Lambert [31] evaluated 22 measures for the LASA problem. After that, the three best measures Trigram-2B, NED and Editex were selected for participating in a Logistic Regression Method (LRM-3) for combining its strengths in a new measure to get better results. The Logistic Regression (LR) is an algorithm to predict a binary classification. Like others machine learning algorithms, LR implements a standard learning algorithm to fit a model predictor [23]. In LRM-3 the set of parameters to fit are $\theta = \{\theta_0, \theta_1, \ldots, \theta_n\}$ for the hypothesis $0 \leqslant h_\theta(x) \leqslant$ to classify a pair (X, Y) of drug names as confusable is respect to:

$$h_\theta(x) = g(y(x)) \tag{10}$$

where $y(x) = (\theta^T x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$ and $x_1 = Editex(X, Y)$, $x_2 = NED(X, Y)$, $x_3 = Trigram2B(X, Y)$ are the result obtained

from each individual measure, and $g(y(x)) = 1/(1 + e\hat{}(-y(x)))$ is the sigmoid function [23].

2. Proposed method

In this paper, a logistic regression is adjusted by an evolutionary approach to increase the accuracy on the process of identifying confusable drug names. The hypothesis is that an evolutionary can adjust better the parameters at the logistic regression model based on the F-measure in comparison to the traditional learning algorithm. Therefore, if we could determine the ranking similarity with a better accuracy between a set of drug names, then the problem of identify confused drug names is reduced to a subset with the highest score of similarity. In other words, we treat this problem as an optimization problem using an evolutionary approach.

2.1. Optimization problem

Given a set $M = m_1, m_2, \ldots, m$ of normalized individual measures (between 0 and 1) and a set $D = d_1, d_2, \ldots, d_m$ of confused drug names, the problem of combining all individual measures of M consists in finding the associated weights to each measure that in Equation (11) maximizes the f-measure evaluation of the set H from the query result to retrieve all T between all pairs in $D \times D$, where the constraint is $\sum_{i=1}^{n} w_i = 1$. Therefore, the proposed Optimized Logistic Regression Model (OLRM) is defined as:

$$OLRM - n(d_i, d_j) = \frac{1}{1+e - \sum_{k=1}^{n} w_k m_k(d_i, d_j)} \tag{11}$$

2.2. Proposed genetic algorithm

Genetic Algorithm (GA) is an Evolutionary Algorithm (EA) inspired in the theory of natural selection mechanism proposed by Darwin that has proved to be an alternative solution for global optimization problems in large search spaces [18–22].

In the first step, the GA proposes a population of random solutions (*initial population step*) that are evaluated according to the objective function to optimize (*fitness function step*). In this sense, a solution for one problem is not absolute, it means, there is a set of possible solutions where some are better than others. Considering mostly the best solutions (*parents selection step*), the GA proposes a new population mixing (*crossover step*) some parts from a canonical codification (*chromosome encoding step*) of these good solutions in order to get better solutions (*evolution principle*). Eventually, the way of mixing some parts from the canonical codification could produce repeated solutions. Therefore, the GA applies a small variation (*mutation step*) to the canonical codification in the new population in order to explore new solutions. The new population is evaluated again and the process is repeated until a satisfactory solution is reached or until some arbitrary stop-criteria is reached (*stop condition*) [34].

2.3. Proposed genetic operators

Chromosome Encoding. The associated weights $W = \{w_1, w_2, \ldots, w_n\}$ to each measure are represented by a binary chromosome with five precision decimals. Each weight has a value between 0 and 1.

Initial Population. All chromosomes in the initial Population (P_o) are created in random way.

Fitness Function. The key step of a GA is the Fitness function. Here, the aptitude of each chromosome must be evaluated. It is worth mentioning, that the objective of the FDA is to recover the closest LASA from a proposed drug name. Hence, the information-retrieval f-measure evaluation is used. Given a LASA pair $(d_i, d_j) \in T$, the f-measure for the query d_i evaluates the size of the set of retrieved drug names in ranking 1 (closest similar drug names to the query d_i), but if d_j does not appears in the last set, the f-measure add the size of the retrieved drug names in the next ranking, until appears d_j. In this way, f-measure evaluates the ability to find a relevant drug name from a query. F-measure is a harmonic balance between recall R and precision P. Precision P is defined as the number of correctly recovered units (LASA pairs) divided by the number of recovered units; D, T and H are defined in **definition 1**.

$$P(D) = \frac{|T \cap H}{|H|} \tag{12}$$

Recall R is defined as the number of correctly recovered units divided by the number of correctly units. In this sense, Precision measures the fraction of retrieved units that are relevant, while Recall measures the fraction of relevant instances that are

retrieved.

$$R(D) = \frac{|T \cap H|}{|H|} \tag{13}$$

The f-measure for the queries of all different drug names (set D) is defined as:

$$F - measure(D) = \frac{2RP}{R + P} \tag{14}$$

The f-measure could be obtained at every ranking. In fact, it is desired to improve the f-measure in the top four rankings. Therefore, the fitness function computes a macro-averaging f-measure for the queries of all different drug names (set D) based on the sum of the first four rankings.

$$fitness(D) = \sum_{r=1}^{4} f - measure(D, r) \tag{15}$$

In other words, the fitness function gives more relevance to the combination of weights in W that, after retrieving the queries of all different drug names with the combined $OLRM - n$ measure, produces the best sum of the first four f-measure evaluation.

Parent Selection. Once each chromosome has an associated fitness value, those stronger chromosomes have more probability of being selected as parents (natural selection mechanism). Natural selection mechanism establishes that two good solutions (chromosomes) could produce better solutions; nevertheless, in some cases the solution could be worse. In this step, the classical tournament selection is employed where the strongest chromosome from a small random subsample is selected as a parent. The smaller the subsample, the greater the possibility of select weaker chromosomes is.

Crossover. *n-point crossover* is used for mixing the genetic information of the parents. In this case, n random points between the genes of the parent chromosomes are selected, and then two offspring chromosomes are created swapping everything between the selected points.

Mutation. According to the evolution scheme, the mutation slightly happens in nature, with a low probability of 0.1%. However, it is one of the fundamental mechanisms to preserve the evolution. Since the chromosome has a binary codification, the classical inverse mutation operator is used.

Elite selection. It helps to keep the best solution of the previous generation.

3. Results and discussion

In the first section, with the objective of having a baseline result for the LASA pairs, the comparison between the individual similarity measures for identifying LASA pairs is presented. In the second section, with the objective to prove how our proposed similarity measure increase the accuracy when more individual measure are added, a comparison of the LRM-3 method to the proposed LRM-21 method is included in this section [26]. In this section, a statistical significance experiment is achieved with the aim to compare the proposed method to previous research.

In the third section, the proposed optimized logistic regression similarity measures OLRM-3 and OLRM-21 are evaluated and compared to the Lambert proposed measure (LRM-3). Additional to that, the ranking of individual measures and the ranking of combined measures (OLRM-21) are compared.

For the experimentation, the dataset USP-858 is used, that it is the same list of confused drug names reported by the USP [43]. Such list contains 858 pairs of confused LASA, with 630 unique drug names. It is worth mentioning that one drug name can be involved in more than one confusion pair. With this list is possible to generate 396,900 pairs of drug names, but only 0.3% of them are LASA pairs. This situation represents an unbalance distribution between the LASA pairs and no-LASA pairs with approximately 1:460.

In the same way, for comparing all approaches the macro-average f-measure, described in section 2.3, is used. It should be noted that in previous works different evaluation measures have been used. Since the LASA problem is treated as an information recovery system, the well-known macro-average f-measure is used in all the experiments.

In fact, it is desired that any system could recover only the 858 LASA pairs in the first ranking position. Nevertheless, it will be impossible because a drug name can be involved until four LASA pairs. Therefore, the fitness function is configured to improve the first fourth ranking positions.

3.1. Evaluation of individual similarity measures

All the orthographic and phonetic similarity measures previously described are individually evaluated with the USP-858 collection. The evaluation of each individual measure is computed using the macro-averaging f-measure accumulated in the first four ranking.

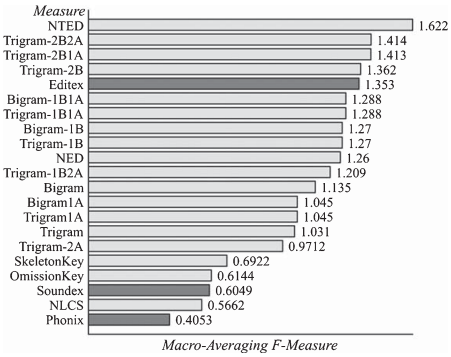

Fig. 1. Ranking obtained for each individual measure according to macro-averaging f-measure. Orthographic and phonetic measures are showed in light gray and dark gray, respectively.

Figure 1 shows that the similarity measures NTED (orthographic distance measure), Trigram-2B2A (orthographic similarity measure), and Editex (phonetic distance measure) are the best ones to identify LASA pairs. It is worth noting, that the best similarity measures reported by Lambert, NED and Trigram-2B, fall into the ranking position and Phonix, Soundex and NLCS are the worst ones.

3.2. Evaluation of combined similarity measures

Firstly, the similarity measure proposed by Lambert based on Logistic Regression of three individual measures (LRM-3) and our proposed similarity measure based on Logistic Regression of 21 individual measures (LRM-21) are compared using the standard learning algorithm. For the LRM-3 measure, the original individual measures proposed by Lambert, Editex, NED, and Trigram-2B (LRM-3); are used. For the LRM-21 measure, all the normalized individual measures presented before are used.

As initial evaluation, LRM-3 and LRM-21 measures are evaluated with ten-fold cross-validation using F-measure evaluation used by Lambert [31]. In this case, LRM-3 measure obtains 98.67 and LRM-21 measure obtains 98.64. These high results agree to Lambert [31] results. However, to get a relevant recovery LASA pairs in top positions of a ranking, we propose to use a macro-averaging F-measure evaluation.

In Table 2, the macro-averaging F-measure evaluations of LRM-3 and LRM-21 measures are shown. As it is possible to observe in Table 2, our proposed

Table 2
Macro-averaging f-measure evaluation obtained after the learning process over the training set.

Ranking	LRM-3		LRM-21	
	F-Meas.	ΣF-Meas.	F-Meas.	ΣF-Meas.
1	0.511192	0.511192	0.547041	0.547041
2	0.449191	0.960383	0.465676	1.012717
3	0.386266	1.346649	0.403605	1.416322
4	0.335319	1.681968	0.353058	1.76938
5	0.296816	1.978784	0.314903	2.084283
6	0.265614	2.244398	0.28319	2.367473
7	0.238906	2.483304	0.256329	2.623802
8	0.216782	2.700086	0.234056	2.857858
9	0.198077	2.898163	0.213918	3.071776
10	0.181176	3.079339	0.197864	3.26964

LRM-21 measure outperforms to LRM-3 measure in all the ranking positions for the F-measure and the accumulated macro-averaging F-measure.

LRM-21 outperforms to LRM-3 with a statistical significance of 95% of confidence in the learning step over the training set. The fitness function in LRM-21 reaches 1.7847 in comparison to 1.6812 of LRM-3. Even though we are interested in improving only the first four positions, LRM-21 measure is better in all ranking positions.

3.3. Evaluation of optimized logistic regression measures

In this section, we evaluate our proposed genetic-algorithm optimized logistic regresion measures OLRM-3 and OLRM-21. The same individual similarity measures used for LRM-3 and LRM-21 are used for OLRM-3 and OLRM-21, respectively.

It should be mentioned that in the next experiments, the same tuning in the operators of the proposed GA is used. In the chromosome, each value of (see section 2.1) has a precision of five decimals. The initial population used is of $W = \{w_1, w_2, \ldots, w_n\}$ chromosomes. As we explain above, the fitness function evaluates the first four positions on the ranking result of f-measure. Only two competitors ($k = 2$) are selected for the selection operator tournament. The crossover operator used is one-point crossover. A mutation probability of 0.17% is used. Also, an elite strategy preserves only the best chromosome for the next generation.

In Fig. 2, the average weights that are found by the GA for the OLRM-3 are showed. These weights show that there are not a directly relation to the ranking position of individual measures (see Fig. 1).

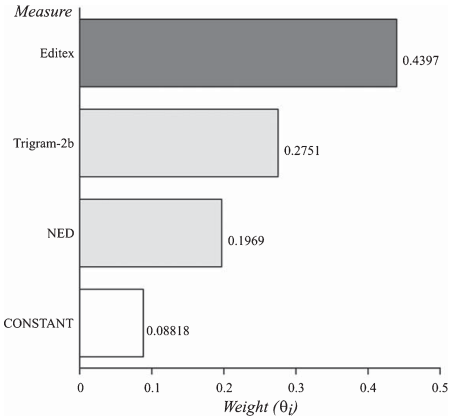

Fig. 2. Ranking obtained for the measure OLRM-3 based on the weights. The orthographic measures are showed in light gray and phonetic in dark gray.

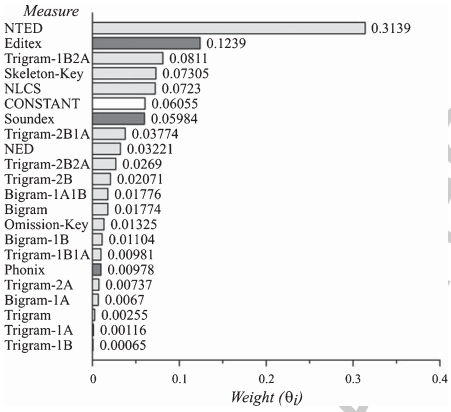

Fig. 3. Ranking obtained for the measure OLRM-21 based on the weights. The orthographic measures are showed in light gray and phonetic in dark gray.

In Fig. 3, the average weights that are found by the GA for the OLRM-21 are showed. Also, in this case, the finding weights show that there is not a trivial relation to the ranking position of individual measures (Fig. 1). It is worth noting that the phonetic measures for OLRM-21 hold upper positions, for example, the phonetic measure Soundex, which is one of the worst individual-evaluated measures, now holds the seventh position.

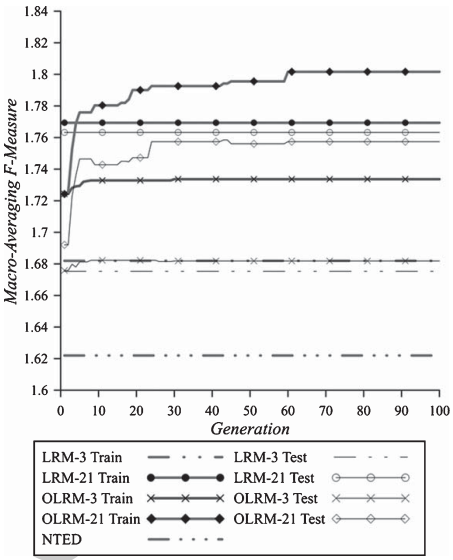

Fig. 4. Training evolution of the macro-averaging f-measure obtained by OLRM-3 and OLRM-21 against LRM-3 and LRM-21 results. The values of the training lines (for the measure OLRM-3 and OLRM-21) correspond to the average of the top-four positions of the macro-averaging f-measure of the training sets.

Figure 4 shows how the learning of our proposed OLRM-3 measure evolves through the generations of the GA. In the first generation, the learning of the OLRM-3 measure outperforms to the standard learning of LRM-3 measure. Analogously, the learning of the proposed OLRM-21 measure outperforms to the standard learning of our proposed LRM-21 measure. In this case, with 21 features the OLRM-21 finds a better learning step than OLRM-3 measure.

Also, Fig. 4 shows the correlation, in the evolution of the GA, between the training dataset and the test dataset. In this case, it is clear that the test results are related to the training results. In this case, our proposed OLRM-3 outperforms to all the measures but OLRM-21 could not outperform LRM-21.

The measures proposed in this work (LRM-21, OLRM-3, and OLRM-21) outperform in the learning and the test process to LRM-3 previously presented by Lambert. After compute the Wilcoxon Signed-Rank Test the proposed measures are statistically significant according to of 95% of confidence, see Table 3.

Table 3
Comparison of our proposed evolutionarily learning measures with the previous standard learning measures (* denotes the individual measure considered as baseline, and ** denotes the method proposed by Lambert)

	Train	$\sum$ F − Meas.	Test	$\sum$ F − Meas.	p-val
1	**OLRM-21**	1.80165	**LRM-21**	1.76326	0.005
2	**LRM-21**	1.76938	**OLRM-21**	1.75739	0.005
3	**OLRM-3**	1.73365	**OLRM-3**	1.68182	0.005
4	LRM-3	1.68196	LRM-3	1.67528	**
5	NTED	1.62178	NTED	1.62178	*

4. Conclusion

LASA is a preventable harmful-health problem that is still growing with more than two decades of research. In this paper, an evolutionary learning method for a logistic regression model to improve the training process is proposed, despite unbalanced dataset of potential LASA pairs. For this, a genetic algorithm with a fitness function based on the sum of the top fourth macro-averaging f-measure is proposed. In specific, the sum of the top four macro-averaging f-measure achieves a greater amount of the accuracy in top positions.

According to the experimentation, our 21-combined measure base on a standard learning logistic regression method (LRM-21) outperforms the 3-combined measure base on a standard learning logistic regression method (LRM-3) for the train and test datasets. Also, our proposed 3-combined measure based on an evolutionarily-adjusted logistic regression model (OLRM-3) outperforms to LRM-3 in the train and test dataset. The same behavior is preserved with our proposed 21-combined measure based on an evolutionarily-adjusted logistic regression method (OLRM-21) that outperforms to LRM-3 with the train and test datasets; and to LRM-21 with the train dataset. However, our proposed LRM-21 obtains the best result in the test dataset. According to the results of the evolution, the training results of the optimized models and the test results are related.

The ranking of the finding weights of the proposed measures do not show a similar relation to the ranking of individual similarity measures. Thus, the novel method to train a logistic regression model is an option to outperform the learning algorithms of the machine learning models. As opposed to the traditional learning algorithm, in the proposed combined measure shows that the greater is the number of individual measures, the better are the results.

In future work, for improving the accuracy of the LASA problem more individual measures must be tested. Also, it will be interesting to select a different machine learning model in order to apply an evolutionary process for tuning its internal parameters.

Acknowledgments

Work done under partial support of Mexican Government CONACyT. We also thank UAEMex for their assistance.

References

[1] ASHP guidelines on preventing medication errors in hospitals, *American Journal of Health-System Pharmacy* **50** (1993), 305–314.
[2] G.W. Adamson and J. Boreham, The use of an association measure based on character structure to identify semantically related pairs of words and document titles, *Information storage and retrieval* **10** (1974), 253–260.
[3] A. Aneja, A.R. Patki and R. Kumbhalwar, *Approximate proper name matching*, 2007.
[4] L.-C. Chen, C.-H. Chen, H.-M. Chen and V.S. Tseng, Hybrid data mining approaches for prevention of drug dispensing errors, *Journal of Intelligent Information Systems* **36** (2011), 305–327.
[5] M.R. Cohen, G.D. Domizio and R.E. Lee, The role of drug names in medication errors, *Medication errors. Wahihgton, DC: The American Pharmacists Association* (2007), 87–110.
[6] V. Craigle, MedWatch: The FDA safety information and adverse event reporting program, *Journal of the Medical Library Association* **95** (2007), 224–225.
[7] E. de Andrade-Azevedo, T. Azevedo-Anacleto and M. Borges-Rosa, Nomes de medicamentos com grafia ou som semelhantes: Como evitar erros, *Bol ISMP-Brasil* **3** (2014).
[8] B.K. Dixon, Similar drug names a growing cause of errors, *Internal Medicine News* **41** (2008), 51–51.
[9] A.K. Elmagarmid, P.G. Ipeirotis and V.S. Verykios, Duplicate record detection: A survey, *IEEE Transactions on Knowledge and Data Engineering* **19** (2007).
[10] FDA, PDUFA Pilot Project - Propietary Name Concept Paper, 2008.
[11] FDA, FDA and ISMP Work to Prevent Medication Errors, 2012.
[13] FDA, Guidance for industry. Contents of a complete submission for the evaluation of proprietary names, 2014.
[13] FDA, Phonetic and Orthographic Computer Analysis (POCA) program, 2017.
[14] T. Gadd, PHONIX: The algorithm, *Program* **24** (1990), 363–366.
[15] B.H. Garcia, R. Elenjord, C. Bjornstad, K.H. Halvorsen, S. Hortemo and S. Madsen, Safety and efficiency of a new generic package labelling: A before and after study in a simulated setting, *BMJ Quality & Safety* **26** (2017), 817–823.
[16] J.A. Gershman and A.D. Fass, Medication safety and pharmacovigilance resources for the ambulatory care setting: Enhancing patient safety, *Hospital Pharmacy* **49** (2014), 363–368.

[17] K.A. Getz, S. Stergiopoulos and K.I. Kaitin, Evaluating the completeness and accuracy of MedWatch data, *American Journal of Therapeutics* **21** (2014), 442–446.
[18] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison Wesley, Reading, Massachusetts, 1989.
[19] S. Gupta, A.P. Srivastava and S. Awasthi, Fast and Effective Searches of Personal Names in an International Environment, *International Journal of Innovative Research in Engineerinf and Managemenr* **1** (2014).
[20] R. Hicks, D.D. Cousins and R.L. Williams, *Summary of information submitted to MEDMARX in the year 2002: The quest for quality*, US Pharmacopeia, 2003.
[21] R.W. Hicks, S.C. Becker and D.D. Cousins, MEDMARX data report. A report on the relationship of drug names and medication errors in response to the Institute of Medicine's call for action, in *Center for the Advancement of Patient Safety*, US Pharmacopeia, Rockville, MD, 2008.
[22] J.H. Holland, *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence,* MIT Press, 1992.
[23] D.W. Hosmer Jr, S. Lemeshow and R.X. Sturdivant, *Applied logistic regression*, John Wiley & Sons, 2013.
[24] G. Kondrak, N-gram similarity and distance, in: *String processing and information retrieval*, Springer, 2005, pp. 115–126.
[25] G. Kondrak and B. Dorr, Identification of confusable drug names: A new approach and evaluation methodology, in: *Proceedings of the 20th International Conference on Computational Linguistics*, Association for Computational Linguistics, Geneva, Switzerland, 2004, p. 952.
[26] G. Kondrak and B. Dorr, Automatic identification of confusable drug names, *Artificial Intelligence in Medicine* **36** (2006), 29–42.
[27] G. Kondrak and B.J. Dorr, A similarity-based approach and evaluation methodology for reduction of drug name confusion, in *Albetta Univ Edmonton*, 2003.
[28] L. Kovacic and C. Chambers, Look-alike, sound-alike drugs in oncology, *Journal of Oncology Pharmacy Practice* **17** (2011), 104–118.
[29] B.L. Lambert, Predicting look-alike and sound-alike medication errors, *American Journal of Health-System Pharmacy* **54** (1997), 1161–1171.
[30] B.L. Lambert, K.-Y. Chang and S.-J. Lin, Effect of orthographic and phonological similarity on false recognition of drug names, *Social Science and Medicine* **52** (2001), 1843–1857.
[31] B.L. Lambert, S.-J. Lin, K.-Y. Chang and S.K. Gandhi, Similarity as a risk factor in drug-name confusion errors: The look-alike (orthographic) and sound-alike (phonetic) model, *Medical Care* **37** (1999), 1214–1225.
[32] B.L. Lambert, C. Yu and M. Thirumalai, A system for multiattribute drug product comparison, *Journal of Medical Systems* **28** (2004), 31–56.
[33] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, in: *Soviet physics doklady*, 1966, pp. 707–710.
[34] M. Mitchell, *An introduction to genetic algorithms*, Cambridge, Massachusetts London, England, Fifth printing, 1999.
[35] T. Nagata, M. Kimura and F. Tsuchiya, Similarity index for sound-alikeness of drug names with pitch accents, *Procedia Computer Science* **35** (2014), 1519–1528.
[36] U. Pfeifer, T. Poersch, N. Fuhr and L. Vi, Searching Proper Names in Databases, in: *HIM*, Citeseer, 1995, pp. 259–275.
[37] J.J. Pollock and A. Zamora, Automatic spelling correction in scientific and scholarly text, *Communications of the ACM* **27** (1984), 358–368.
[38] Z. Rahman and R. Parvin, Medication errors associated with look-alike/sound-alike drugs: A brief review, *Journal of Enam Medical College* **5** (2015), 110–117.
[39] S.R. Schroeder, M.M. Salomon, W.L. Galanter, G.D. Schiff, A.J. Vaida, M.J. Gaunt, M. Bryson, C. Rash, S. Falck and B.L. Lambert, Cognitive tests predict real-world errors: The relationship between drug name confusion rates in laboratory-based memory and perception tests and corresponding error rates in large pharmacy chains, *BMJ Quality & Safety* **26** (2016), 395–407.
[40] M.B. Shah, L. Merchant, I.Z. Chan and K. Taylor, Characteristics that may help in the identification of potentially confusing proprietary drug names, *Therapeutic Innovation & Regulatory Science* (2016), 2168479016667161.
[41] B. Teplitsky, Hazards of sound-alike, look-alike drug names, *California Medicine* **119** (1973), 62.
[42] P.L. Trbovich and S. Hyland, Responding to the challenge of look-alike, sound-alike drug names, *BMJ Quality & Safety* **26** (2017), 357–359.
[43] USP, USP quality review (76). US Pharmacopeia, 2001.
[44] USP, USP Quality Review (79). US Pharmacopeia, 2004.
[45] R.A. Wagner and M.J. Fischer, The string-to-string correction problem, *J ACM* **21** (1974), 168–173.
[46] WHO, Medication Without Harm *Global Patient Safety Challenge on Medication Safety*, 2017.
[47] J. Zobel and P. Dart, Phonetic string matching: Lessons from information retrieval, in: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 1996, pp. 166–172.

Página intencionalmente dejada en blanco

Capítulo V. Identificación de pares de nombres confusos de medicamentos mediante una medida suavizada

Es este capítulo se incluye el artículo titulado: *Soft Bigram Similarity to Identify Confusable Drug Names*. El artículo fue aceptado y publicado por la revista especializada arbitrada e indexada de reconocimiento internacional titulada: *Lectures Notes in Computer Sciences 11524*, de la editorial *springer*, con *índice Scopus*. Publicado el 14 de mayo de 2019. Se anexa la carta de aceptación.

5.2. Ejemplar de autor del artículo

Soft Bigram Similarity to Identify Confusable Drug Names

Christian Eduardo Millán-Hernández, René Arnulfo García-Hernández(✉), Yulia Ledeneva, and Ángel Hernández-Castañeda

Autonomous University of State of Mexico, 50000 Toluca, Mexico
ceduardo.millan@gmail.com, renearnulfo@hotmail.com, yledeneva@yahoo.com, angelhc2305@gmail.com

Abstract. Look-alike and Sound-alike drug names are related to medication errors where doctors, nurses, and pharmacists prescribe and administer the wrong medication. Bisim similarity is reported as the best orthographic measure to identifying confusable drug names, but it lacks from a similarity scale between the bigrams of a drug name. In this paper, we propose a Soft-Bisim similarity measure that extends to the Bisim to soften the comparison scale between the Bigrams of a drug name for improving the detection of confusable drug names. In the experimentation, Soft-Bisim outperforms others 17 similarity measures for 396,900 pairs of drug names. In addition, the average of four measures is outperformed when Bisim is replaced by Soft-Bisim similarity.

AQ1

AQ2

Keywords: LASA drug names · N-gram similarity · Orthographic similarity

1 Introduction

A medication error that involves confusable drug names occurs as result of weak medication system and human errors-related factors [1–3]. Many human factors are related to the Look-Alike and Sound-Alike drug names (LASA) problem like visual perception error, auditory perception error, short term memory error, and motor control are errors. However, the similarity between confusable drug names is a detectable root-cause. Drug names like *cycloserine* and *cyclosporine* are involved in LASA errors. LASA pairs normally sound similar and have a similar spelling [4]. Sometimes the confusion happens when the names are communicated in prescriptions handwritten, for example, the drugs *Avandia* and *Coumadin* [5]. In other cases, the confusion occurs in verbal communication when the pronunciation sounds similar. For example, *Zantac* and *Xanax* [6].

Nowadays, the Institute for Safe Medication Practice (ISMP) publishes a list that contains LASA pairs that were previously reported [7–10]. Regulatory agencies, including the Food and Drug Administration (FDA), the World Health Organization (WHO), and the Joint Commission are implementing strategies to identify and to prevent a LASA error.

String-matching algorithms are used to measure the distance or the similarity between two drug names and to identify a priori potential confused drug names. For

J. A. Carrasco-Ochoa et al. (Eds.): MCPR 2019, LNCS 11524, pp. 1–10, 2019.
https://doi.org/10.1007/978-3-030-21077-9_40

example, *Edit Distance* (ED) measures the minimum of the insertion, elimination and substitution operations to transform a string to another [11]. For example, the LASA pair *cycloserine* and *cyclosporine* has a distance of two because there are needed at least two edit operations (a substitution *p*→*e* and an elimination of letter *o*) to transform *cycloserine* in *cyclosporine*.

Longest Common Subsequence (LCS) measures the maximum possible length of the longest common subsequences between two drug names. NLCS represents the Normalization of LCS that is obtained by dividing the maximum length of the longest drug name. In the previous example, */cyclos-rine/* is the LCS and the NLCS is 0.833. NLCS presents a weakness to ignore subsequences that does not represent a similarity between drug names. For example, the no-LASA pair *Benadryl* and *Cardura* have the LCS /adr/ [6].

Ngram similarity represents a drug name as the set of all its contiguous subsequences (grams) of size *n* [12, 13]. For example, the bigrams for the LASA pair *cycloserine* and *cyclosporine* are {*cy*, *yc*, *cl*, *lo*, *os*, *se*, *er*, *ri*, *in*, *ne*} and {*cy*, *yc*, *cl*, *lo*, *os*, *sp*, *po*, *or*, *ri*, *in*, *ne*}, respectively. In this case, eight bigrams are shared, and the number of bigrams is 10 and 11, respectively. Therefore, the similarity is $(2 \times 8)/(10 + 11) = 0.762$. However, the Ngram similarity of the LASA pair *Verelan* and *Virilon* is zero [6].

Nsim similarity [14] extends to NLCS but it manages the *n*-grams of a drug name with a scale of similarity. The predefined scale of similarity between a pair of *n*-grams is computed by counting the identical matching letters in each position and normalized by *n*. *Bisim* is a specific case of *Nsim* with a predefined scale of similarity. For example, the bigrams ***cy*** and ***cy*** have a similarity of 2/2 = 1 and the bigrams ***se*** and ***sp*** of 1/2 = 0.5. The similarity scale presents a weakness when computes values for bigrams like *sp* and *ps*, or *sp* and *es*; because it misplaces completely the common letters in previous or next positions. This issue is a common root-cause when a visual or auditory perception error happens [15–17]. Even the pairs of bigrams {*aa*}{*aa*} and {*ac*}{*ac*} computes the same similarity, it is clear that in the first example the letter *a* match all the letters of the bigrams showing a higher similarity. In this manner, commonalities characteristics that are presented in LASA pairs [18] needs to be considered to adjust a softened similarity scale.

In this paper, we propose a new softened similarity measure based on Bisim that increase the accuracy to identify LASA pairs. For this, different cases that form the scale of bigrams are identified, and a proposed methodology based on an evolutionary algorithm to soften the scale of the similarity is described. Therefore, this paper is based on the hypothesis that an evolutionary approach can adjust better the weights of the scale of similarity between *n*-grams.

2 Definitions

String matching algorithms recover the common correspondences between the drug names that are used to determinate a similarity or a distance measure. Measures are classified as distance (as closer to zero as more related are the names) or similarity (as

greater is the value as more related are the names). A normalized similarity/distance measure keeps a scale between different similarity values.

Similarity and distance measures detect the particular look-alike (orthographic cause) and sound-alike (phonetic case) issue. In this sense, the measures are classified as orthographic or phonetic in relation to the used approach to detect the confusion.

2.1 Orthographic Distance Measures

Edit distance (ED). Given the drug names X and Y as sequences of size n and m, respectively, ED (also called *Levenshtein*) refers to the minimum cost of editing operations (insertion, deletion and substitution) to convert the sequence X into Y [11, 19–21]. In this paper, all editing operations have a cost of 1. In this case, the edit distance between X and Y is given by edit (n, m) computed by the following recurrence:

$$edit(i,j) = \begin{cases} max(i,j) & i = 0 \vee \mathrm{j} = 0 \\ edit(i-1,j-1) & x_i = y_j \\ min \begin{cases} edit(i-1,j)+1 \\ edit(i,j-1)+1 \\ edit(i-1,j-1)+cs(x_i,y_i) \end{cases} & x_i \neq y_j \end{cases} \quad (1)$$

A Normalized ED (NED). NED is computed by dividing the ED between the length of the longer sequence [6, 21–25].

2.2 Orthographic Similarity Measures

Prefix Similarity. Given the drug names X and Y as sequences of size m and n respectively, *Prefix* represents the ratio of the longest contiguous common initial letters [6], see Eq. 2. The common prefix for drug names *Accutane* and *Accolate* is *Acc* ($|Acc| = 3$), and the normalized prefix similarity is 0.375.

$$Prefix(X,Y) = \frac{|x_1 = y_1, x_2 = y_2, \ldots, x_i = y_i|}{\max(X,Y)} \quad (2)$$

N-gram Similarity. Represents a sequence of the set of all its contiguous subsequences (grams) of size n [12]. For example, if $|X| = m$ and $n = 2$ (bigrams), then $X' = \{x_1x_2, x_2x_3, \ldots, x_{m-1}x_m\}$ [6, 14, 26]. Given the sequences X and Y, the *n-gram similarity* is defined as the *Dice similarity* [27] between the sets X' and Y' in the next way:

$$Dice\left(X', Y'\right) = \frac{2\left|X' \cap Y'\right|}{|X'| + |Y'|} \quad (3)$$

N-gram similarity presents a weakness because it is well-known that the prefixes and suffixes of the drug names are involved in their confusion [6, 18]. For increasing the sensitivity of the *N-gram similarity* some variations with respect to initial and final letters area applied. Lambert [14] proposes to add spaces (or a letter not included in the names) (B)efore and (A)fter in both drug names to make that the initial or final letters appear in one or more *n*-grams. Lambert experimented with the variants of Bigram-(1B, 1A, 1B1A and 1A) and Trigram-(1B, 1A, 1B1A, 2B, 2A, 2B2A, 1B2A and 2B1A) [14, 17, 28].

Normalized LCS (NLCS). *NLCS* similarity lets to maintain an order in the common matching letters. Given the sequences X and Y of size n and m, respectively, the *NLCS similarity* is defined as the ratio of the length of the longest common subsequences between X and Y, $NLCS = |lcs(n,m)|/max(m,n)$, where $lcs(n,m)$ can be calculated by the recurrence in Eq. (4) [6, 14, 23–25, 29].

$$lcs(i,j) = \begin{cases} 0, & i = 0 \vee \mathrm{j} = 0 \\ lcs(i-1,j-1)+1, & x_i = y_j \\ max(lcs(i,j-1), lcs(i-1,j)) & x_i \neq y_j \end{cases} \tag{4}$$

Nsim Similarity. It is proposed by Kondrak [6, 23, 24] and it combines features implemented by grams of size β, non-crossing-links constraints and the first letter it is repeated at the begging of the drug name. A specific case of *Nsim* is the measure *Bisim* [6]. Given the sequences (with the first repeated letter) X and Y representing the drug names of size n and m, respectively, *Bisim similarity* is defined as:

$$Bisim(X,Y) = \frac{nsim(n,m)}{max(n,m)}$$

$$nsim(i,j) = \begin{cases} 0, & i = 0 \vee j = 0 \\ max \begin{cases} nsim(i,j-1), \\ nsim(i-1,j), \\ nsim(i-1,i-1) + \\ s\left(x_i x_{i+1}, y_j y_{j+1}\right), \end{cases} & \begin{matrix} in \\ other \\ case \end{matrix} \end{cases} \tag{5}$$

$$s\left(x_i x_{i+1}, y_j y_{j+1}\right) = \frac{1}{2}\sum_{k=0}^{1} id(x_{i+k}, y_{j+k}) \tag{6}$$

$$id(a,b) = \begin{cases} 1, a = b \\ 0, a \neq b \end{cases} \tag{7}$$

2.3 Related Work

Using a list of 1,127 LASA pairs and 1,127 non-LASA pairs, Lambert [14] evaluates 22 measures with ten-fold cross-validation technique and concludes that Trigram2B, NED and Editex [20] are the best measures to identify LASA pairs.

Kondrak [6, 23–25] proposes the orthographic Nsim similarity and the phonetic Aline similarity [30, 31] where the recall metric is used to evaluate the results of 12 measures with the USP LASA list [32] of 360 unique drug names. Kondrak [6] concludes that Bisim is the best orthographic measure. Bisim is used to create automated warning systems to identify potential LASA errors in prescription electronic systems [4, 33] and in the software POCA by the FDA [6]. Furthermore, the average of Bisim, Aline, Prefix, and NED measures outperform to Bisim [6].

3 Proposed Method

In this paper, a Soften Bigram Similarity measure (Soft-Bisim) is proposed. First, the cases of bigrams involved in the scale of similarity in Soft-Bisim are described. After that, the fitness function used to find the weights in the scale of similarity by a genetic algorithm is described. Our hypothesis is that an evolutionary approach defines better the levels in the scale of similarity compared to the original similarity scale proposed by Kondrak in Bisim (cf. Eqs. 7 and 8). In other words, we consider this problem as an evolutionary approach for optimizing the internal parameters of the similarity scale.

3.1 Definition of Soft-Bisim Similarity

Given the drug names X and Y as sequences of size n and m, respectively, Soft-Bisim is defined as:

$$Soft - Bisim(X,Y) = \frac{Bisim(n,m)}{max(n,m)} \tag{8}$$

$$Bisim(i,j) = \begin{cases} 0, & i = 0 \vee j = 0 \\ max \begin{cases} Bisim(i,j-1), \\ Bisim(i-1,j), \\ Bisim(i-1,i-1) + \\ s\left(x_i x_{i+1}, y_j y_{j+1}\right), \end{cases} & \begin{matrix} in \\ other \\ case \end{matrix} \end{cases} \tag{9}$$

Where the proposed scale of similarity for Soft-Bisim is defined as:

$$s\left(a_i a_{i+1}, b_j b_{j+1}\right) = \begin{cases} w_1, a_{i+1} = b_{j+1} \neq a_i \neq b_j \\ w_2, a_i \neq a_{i+1} \neq b_j \neq b_{j+1} \\ w_3, a_i = a_{i+1} = b_j \neq b_{j+1} \vee a_i = b_j = b_{j+1} \neq a_{i+1} \\ w_4, a_i = b_{j+1} \wedge a_{i+1} = b_j \\ w_5, a_i = b_{j+1} \neq a_{i+1} \neq b_j \vee a_{i+1} = b_j \neq a_i \neq b_{j+1} \\ w_6, a_i = a_{i+1} = b_{j+1} \neq b_j \vee a_{i+1} = b_j = b_{j+1} \neq a_i \\ w_7, a_i = b_j \neq a_{i+1} \neq b_{j+1} \\ w_8, a_i = a_{i+1} = b_j = b_{j+1} \\ w_9, a_i = b_j \wedge a_{i+1} = b_{j+1} \end{cases} \tag{10}$$

For increasing accuracy to identify confusable drug names it is needed to find the set of weights $W = \{w_1, w_2, \ldots, w_9\}$ of the scale of similarity of Soft-Bisim. For this, a Genetic Algorithm is used [34–36].

3.2 Finding the Scale of Similarity for Soft-Bisim

The fitness function of the Genetic Algorithm is designed to evaluate each individual in relation to the objective to optimize.

The FDA reviews the similarity of a new drug name with all drug names that were previously registered. Therefore, the f-measure evaluation widely used in the information retrieval is used as the fitness function [37]. Given a LASA pair $(d_i, d_j) \in$ *List of LASA pairs*, the f-measure for the query d_i evaluates the size of the set of retrieved drug names in ranking one (most similar drug names to the query d_i), but if d_j does not appears in the last set, the f-measure add the size of the retrieved drug names in the next ranking, until appears d_j. In this way, f-measure evaluates the ability to find a relevant drug name from a query. The f-measure could be obtained at every ranking (r). In fact, we desire to improve the f-measure in the top four rankings. Therefore, the fitness function computes a macro-averaging f-measure for the queries of all different drug names (set D) based on the sum of the first four rankings, see Eq. 11. In other words, the fitness function gives more relevance to the combination of weights in W (Eq. 10) that, after retrieving the queries of all different drug names with the Soft-Bisim measure, produces the best sum of the first four f-measure evaluation.

$$fitness(D) = \sum_{r=1}^{4} f - measure(D, r) \tag{11}$$

4 Results and Discussion

In all the experiments, the ground truth USP-858 collection with 858 LASA pairs is used. The USP-858 contains 630 unique drug names, and it can generate 36,900 pairs of drug names. That means that 0.3% of LASA pairs must be recovered.

4.1 Calculating the Scale of Similarity for Soft-Bisim

Although, the genetic algorithm only optimizes the macro-averaging f-measure for the top four positions, the comparison in Table 1 shows an improvement, with respect to Bisim, in all positions of ranking to retrieve LASA pairs. As it is possible to observe, the weight w_9 for Soft-Bisim maintains a higher relevance than w_8 while in Bisim w_8 and w_9 are the same. On the contrary, the case when all the letters are different the weight is not zero.

Table 1. Comparison of the macro-averaging f-measure evaluation for Bisim and Soft-Bisim with the USP-858 collection where the resulting weights for Soft-Bisim are: $w_1 = 0$, $w_2 = 0.1$, $w_3 = 0.4$, $w_4 = 0$, $w_5 = 0$, $w_6 = 0.2$, $w_7 = 0.4$, $w_8 = 0.6$ and $w_9 = 0.8$; and the implicit weights for Bisim are: $w_{1...3} = 0$, $w_{4...7} = 0.5$, $w_7 = 0$, and $w_8 = w_9 = 1$. *In the last row the ten-fold cross-validation results are showed.

Ranking		1	2	3	4	5	6	7	8	9	10
Bisim	F-Meas.	48.86	39.78	29.74	22.45	17.18	13.44	10.64	8.75	7.24	5.96
	ΣF-Meas.	48.86	88.65	118.40	140.86	158.04	171.49	182.14	190.90	198.14	204.11
Soft-Bisim	F-Meas.	51.07	45.20	39.03	33.45	28.89	25.75	23.08	20.93	19.07	17.52
	ΣF-Meas.	51.07	96.27	135.31	168.76	197.65	223.41	246.49	267.42	286.50	304.03
Soft-Bisim*	F-Meas.	51	44.75	38.79	33.51	29.18	25.97	23.33	21.21	19.38	17.85
	ΣF-Meas.	51	95.75	134.53	168.05	197.23	223.2	246.53	267.74	287.12	304.97

4.2 Evaluation of Orthographic Measures

In Fig. 1, Soft-Bisim is compared to all orthographic measures presented in Sects. 2.1 and 2.2. In this case, Trigram-2B maintains the relevance indicated by Lambert but Bisim is more relevant that Trigram-2B. It is worth to mention that Trigram-2B2A and Trigram-2B1A are more relevant that Bisim. However, Soft-Bisim obtains the best performance with the adjusted similarity scale.

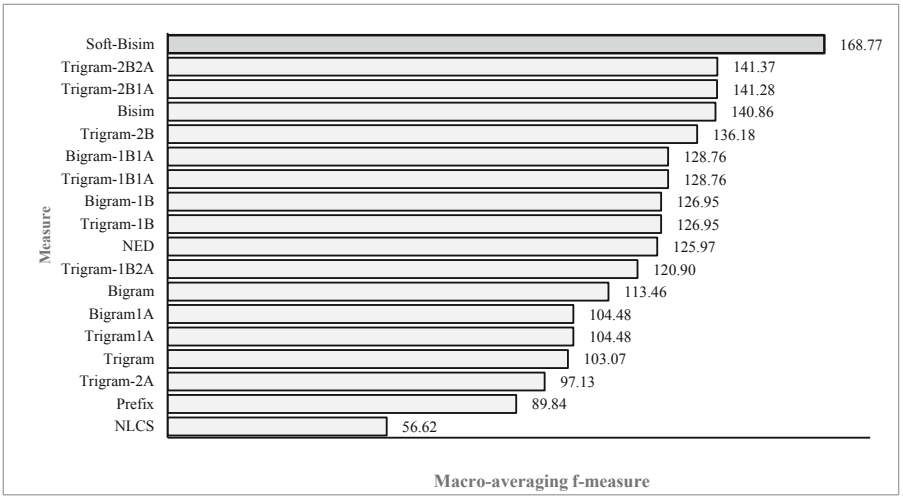

Fig. 1. Ranking obtained for each orthographic measure according to sum of top four positions of macro-averaging f-measure.

4.3 A Combined Measure with Soft-Bisim

Using the Average of Prefix, NED, Bisim and Aline, Kondrak [6] proposes a combined measure that outperform to Bisim: $\text{Avg}_{\text{Bisim}}$(Prefix, NED, Bisim, Aline). In this paper,

we propose two combined measures, in the first one, Soft-Bisim is added to the average Avg_{all} (Prefix, NED, Bisim, Aline, Soft-Bisim), and the second one, Bisim is replaced by Soft-Bisim in the average, $Avg_{SoftBisim}$(Prefix, NED, Aline, Soft-Bisim). In Table 2, the comparison of original combined proposed by Kondrak and our proposed combined measures are presented.

Table 2. Macro-averaging f-measure evaluation for Avg_{Bisim}, Avg_{All} and $Avg_{SoftBisim}$.

Ranking	Avg_{Bisim}		Avg_{All}		$Avg_{SoftBisim}$	
	F-Measure	ΣF-Measure	F-Measure	ΣF-Measure	F-Measure	ΣF-Measure
1	51.36	51.36	51.63	51.63	51.70	51.70
2	44.69	96.05	45.56	97.20	45.42	97.12
3	39.63	135.68	39.78	136.98	40.01	137.13
4	35.11	**170.80**	34.87	**171.85**	35.13	**172.27**
5	30.79	201.59	30.72	202.58	30.89	203.16
6	27.55	229.15	27.76	230.34	27.85	231.02
7	25.04	254.19	25.05	255.39	25.10	256.12
8	22.81	277.00	22.87	278.27	22.86	278.98
9	20.84	297.85	20.92	299.19	21.06	300.04
10	19.31	317.17	19.45	318.64	19.47	319.52

Table 3. Comparison of Soft-Bisim with the best previous measures.

Rank	1	2	3	4	5	6	7
Measure	$Avg_{SoftBisim}$	Avg_{all}	Avg_{Bisim}	Soft-Bisim	Trigram-2B2A	Trigram-2B1A	Bisim
F-Meas.	172.27	171.85	170.80	168.76	141.37	141.27	140.86
p-value	0.005	0.005	0.005	0.005	0.005	0.005	Baseline

In Table 3, using Bisim as a Baseline measure the best measures to identify confuse drug names are showed. In this case, all the combined measures outperform to the individual measures. However, the best individual measure is Soft-Bisim that it is involved in the first two combined measures. Moreover, the best performance is achieved when Bisim is replaced by Soft-Bisim.

5 Conclusion

The problem of confusion of drug names needs attention because it is still growing. All measures presented in this paper (except by Nsim) are designed or adjusted to different application or domain. In this sense, Nsim takes into consideration characteristics that take part on confusable drug names like the fact that the initial letters are frequently involved in a confused drug name. In this paper we propose to Soft-Bisim measure that it is a new orthographic measure for identifying LASA pairs based on Nsim similarity

with an extension to soften the scale of similarity between the bi-grams that conforms a drug name. In specific, nine combinations of weights were calculated. For this, the sum the first-four macro-averaging f-measure of the retrieved pairs is proposed as the fitness function in a genetic algorithm.

According to the experimentation, Soft-Bisim increases the accuracy with respect to Bisim in a retrieved list of potential LASA pairs in all the ranking positions. Furthermore, Soft-Bisim outperforms significantly to the others 17 orthographic measures used in this problem. In this paper, we found that the measures Trigram-2B2A and Trigram-2B1A are good measures since outperform to the Bisim measure.

In addition, a new average combination of four measures using Soft-Bisim is proposed. This new average combination outperforms to the previous average that use Bisim measure. Even thought, we only use a list of drug names Soft-Bisim can be used to retrieve other cases of confusions like in proper names or brand names.

References

1. Billstein-Leber, M., Carrillo, C.J.D., Cassano, A.T., Moline, K., Robertson, J.J.: ASHP guidelines on preventing medication errors in hospitals (2018). https://www.ashp.org/Pharmacy-Practice/Policy
2. Cohen, M.R., Domizio, G.D., Lee, R.E.: The role of drug names in medication errors. In: Medication Errors, pp. 87–110. American Pharmacists Association, Washington, DC (2007)
3. Medication Without Harm.: World Health Organization, Geneva (2017) AQ3
4. Rash-Foanio, C., et al.: Automated detection of look-alike/sound-alike medication errors. Am. J. Heal. Pharm. **74**, 521–527 (2017)
5. Tittemore, L.M.: The name game (2017). https://sunsteinlaw.com/l-tittemore/
6. Kondrak, G., Dorr, B.: Automatic identification of confusable drug names. Artif. Intell. Med. **36**, 29–42 (2006)
7. FDA: FDA and ISMP Work to Prevent Medication Errors 2017 (2012)
8. Craigle, V.: MedWatch: the FDA safety information and adverse event reporting program. J. Med. Libr. Assoc. **95**, 224–225 (2007)
9. Gershman, J.A., Fass, A.D.: Medication safety and pharmacovigilance resources for the ambulatory care setting: enhancing patient safety. Hosp. Pharm. **49**, 363–368 (2014)
10. Getz, K.A., Stergiopoulos, S., Kaitin, K.I.: Evaluating the completeness and accuracy of MedWatch data. Am. J. Ther. **21**, 442–446 (2014)
11. Wagner, R.A., Fischer, M.J.: The string-to-string correction problem. J. ACM **21**, 168–173 (1974)
12. Pfeifer, U., Poersch, T., Fuhr, N., Vi, L.I.: Searching proper names in databases. In: HIM, pp. 259–275. Citeseer (1995)
13. Pfeifer, U., Vi, L.I.: Searching proper names in databases, vol. 20, pp. 1–13, October 1994 AQ4
14. Lambert, B.L., Lin, S.J., Chang, K.Y., Gandhi, S.K.: Similarity as a risk factor in drug-name confusion errors: The look-alike (orthographic) and sound-alike (phonetic) model. Med. Care **37**, 1214–1225 (1999)
15. Schroeder, S.R., et al.: Cognitive tests predict real-world errors: the relationship between drug name confusion rates in laboratory-based memory and perception tests and corresponding error rates in large pharmacy chains. BMJ Qual. Saf. **26**, 395–407 (2017)
16. Lambert, B.L., et al.: Listen carefully: the risk of error in spoken medication orders. Soc. Sci. Med. **70**, 1599–1608 (2010)

17. Lambert, B.L.: Predicting look-alike and sound-alike medication errors. Am. J. Heal. Pharm. **54**, 1161–1171 (1997)
18. Shah, M.B., Merchant, L., Chan, I.Z., Taylor, K.: Characteristics that may help in the identification of potentially confusing proprietary drug names. Ther. Innov. Regul. Sci. **51**, 232–236 (2017)
19. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet Physics Doklady, pp. 707–710 (1966)
20. Zobel, J., Box, G.P.O., Dart, P.: Phonetic string matching : lessons from information retrieval. In: Proceedings of 19th Annual International ACM SIGIR Conference Research and Development in Information Retrieval, pp. 166–172 (1996)
21. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: a survey. IEEE Trans. Knowl. Data Eng. **19**, 1–16 (2007)
22. Chen, S., Liu, Y., Wei, L., Guan, B.: PS-FW: a hybrid algorithm based on particle swarm and fireworks for global optimization. Comput. Intell. Neurosci. **2018**, 1–27 (2018)
23. Kondrak, G., Dorr, B.: Identification of confusable drug names: a new approach and evaluation methodology (2004)
24. Kondrak, G., Dorr, B.J.: A similarity-based approach and evaluation methodology for reduction of drug name confusion. Alberta University, Edmonton (2003)
25. Kondrak, G.: *N*-Gram similarity and distance. In: Consens, M., Navarro, G. (eds.) SPIRE 2005. LNCS, vol. 3772, pp. 115–126. Springer, Heidelberg (2005). https://doi.org/10.1007/11575832_13
26. Chen, L.-C., Chen, C.-H., Chen, H.-M., Tseng, V.S.: Hybrid data mining approaches for prevention of drug dispensing errors. J. Intell. Inf. Syst. **36**, 305–327 (2011)
27. Adamson, G.W., Boreham, J.: The use of an association measure based on character structure to identify semantically related pairs of words and document titles. Inf. Storage Retr. **10**, 253–260 (1974)
28. Lambert, B.L., Chang, K.-Y., Lin, S.-J.: Effect of orthographic and phonological similarity on false recognition of drug names. Soc. Sci. Med. **52**, 1843–1857 (2001)
29. Lambert, B.L., Yu, C., Thirumalai, M.: A system for multiattribute drug product comparison. J. Med. Syst. **28**, 31–56 (2004)
30. Kondrak, G.: Phonetic alignment and similarity. Comput. Hum. **37**, 273–291 (2003)
31. Kondrak, G.: Algorithms for language reconstruction (2002)
32. USP: USP quality review (76). US Pharmacopeia. (2001)
33. Or, C.K.L., Wang, H.H.L.: Examining text enhancement methods to improve look-alike drug name differentiation accuracy. In: Proceedings of the Human Factors and Ergonomics Society, pp. 645–649 (2013)
34. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison Wesley, Reading (1989)
35. Holland, J.H.: Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. MIT Press, Cambridge (1992)
36. Mitchell, M.: An Introduction to Genetic Algorithms. Cambridge, Massachusetts, London, England, Fifth Printing (1999)
37. Croft, B., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company, Boston (2009)

Página intencionalmente dejada en blanco

Capítulo VI. Discusión general y conclusiones

A pesar de tres décadas de esfuerzos por reducir nombres confusos de medicamentos se continúan registrando casos donde los nombres similares contribuyeron al error de medicación. La principal causa es la complejidad de los procesos de uso de medicamentos y la dificultad por disminuir los factores relacionados con el error humano implicados en la confusión. Esto a su vez, limita los intentos de prevención y reducción de formación de pares LASA. En particular, las estrategias de evaluación de nombres propuesto para nuevos medicamentos no resultan completamente efectivas en la predicción de nombres confusos. Las medidas individuales y combinadas utilizadas para determinar un valor de similitud deben considerar las características presentes en los nombres de medicamentos indicados en los reportes de errores de medicación, con la finalidad de mejorar el proceso de identificación de pares confusos por su parecido ortográfico y fonético. Con la

finalidad de ofrecer resultados válidos para la toma de decisiones en los procesos evaluación por los expertos en errores de medicación por confusión.

Proponer medidas ajustadas o diseñas específicamente para el problema de identificación de pares LASA resulta conveniente, sin embargo, se requiere tener conocimientos especializados en cuestiones en computación lingüística, reconocimientos de patrones, aprendizaje automático y por su puesto en errores de medicación por confusión de nombres de medicamentos. Por lo que resulta viable contar con un modelo de las características extraídas de los pares de nombres confusos de medicamentos. El cual de manera automática y adaptable utilice enfoques evolutivos o técnicas de aprendizaje automático.

En esta tesis se presentan un modelo que combina de manera eficaz las características ortográficas y fonéticas que están presentes en los pares nombres confusos de medicamentos para identificar de manera *a priori* pares potenciales de nombres confusos de medicamentos por como se ven o como suena. Los resultados obtenidos han sido publicados en revistas especializadas arbitradas e indexadas de reconocimiento internacional. Los métodos probados en ambos artículos utilizan las características presentes en las listas de pares LASA previamente reportados como causas de confusión en errores de medicación y muestran mejoras con respecto al las medidas individuales y combinadas presentadas en el estado del arte. En consecuencia, el objetivo de esta investigación se ha cumplido, puesto que los artículos han sido publicados bajo una revisión arbitrada que corrobora la originalidad de ambos escritos.

6.1. Aportaciones

En el primer artículo se presenta un método de regresión logística con un proceso de entrenamiento evolutivo que en primera instancia resuelve el problema de sesgo que representa una muestra desbalanceada. Además, se logra ajustar los parámetros del modelo logístico con base a una métrica que no solo considera la

precisión, si no que al mismo tiempo considera la exhaustividad a través del promedio armónico de ambos (*F measure*). Por último, en un proceso habitual de entrenamiento estándar de la regresión logística se recurre a la reducción de características con el fin de mejorar la eficiencia del modelo obtenido, en la experimentación se muestra el caso contrario, ya que la propuesta incrementa su eficacia mientras más medidas se agreguen a la combinación. En conclusión, la medida combinada propuesta supera el estado del arte mediante una solución original.

En el segundo articulo se plantea una nueva medida de similitud ortográfico para la identificación de pares LASA. Aunque esta medida se basa en Bisim, la utilización de un enfoque evolutivo permite proponer una función de escala de similitud suavizada entre bigramas. La evaluación del resultado obtenido por la nueva medida supera a las medidas individuales del estado del arte. En una experimentación adicional se agrega la medida propuesta (SoftBisim) a la combinación promediada original de Kondrak (Prefix, NED, Aline y Bisim), logrando mejorar el resultado de la original. Por ultimo, se evalúa una modificación donde se elimina Bisim de la combinación (Prefix, NED, Aline, SofBisim). Esta última combinación supera a todas las medidas anteriores.

6.2. Trabajo futuro

En esta sección se presentan algunas recomendaciones para trabajos futuros relacionados con el problema de la confusión de nombres de medicamentos.

Se recomienda implementar un aprendizaje evolutivo para otros modelos de aprendizaje automático, así como, otras muestras de entrenamiento para comprobar la efectividad del enfoque.

Se sugiere implementar nuevas medidas léxicas al modelo de regresión logístico evolutivo y comprobar si eficacia, así como, su desempeño en un ensamble.

A partir de los resultados en Soft Bisim resulta interesante realizar la optimización de los parámetros internos de mas medidas, para conocer su eficacia ante el problema de la identificación de pares LASA.

Durante el estudio del estado del arte se detectó la falta de medidas fonéticas por lo que se sugiere realizar la implementación de mas medidas o realizar propuestas de nuevas medidas.

Resulta interesante considerar poner a prueba los métodos propuestos de esta tesis en otras tareas relacionas con la obtención de un valor de similitud.

Referencias

[1] R. W. Hicks, S. C. Becker, and D. D. Cousins, "MEDMARX data report. A report on the relationship of drug names and medication errors in response to the Institute of Medicine's call for action," Center for the Advancement of Patient Safety, US Pharmacopeia., Rockville, MD, 2008.

[2] L. J. Donaldson, E. T. Kelley, N. Dhingra-Kumar, M. P. Kieny, and A. Sheikh, "Medication Without Harm: WHO's Third Global Patient Safety Challenge," *The Lancet*, vol. 389, no. 10080. Lancet Publishing Group, pp. 1680–1681, 29-Apr-2017.

[3] World Health Organization (WHO), *Medication Errors: Technical Series on Safer Primary Care*. Geneva, 2016.

[4] NCCMERP, "About Medication Errors," 2017. [Online]. Available: http://www.nccmerp.org/about-medication-errors.

[5] B. L. Lambert, R. Bhaumik, W. Zhao, and D. K. Bhaumik, "Detection and prediction limits for identifying highly confusable drug names from experimental data," *J. Biopharm. Stat.*, vol. 26, no. 2, pp. 365–385, Mar. 2016.

[6] A. Kawano, Q. (Kathy) Li, and C. Ho, "Preventable Medication Errors – Look-alike/Sound-alike Drug Names," *Pharm. Connect.*, pp. 28–33, 2014.

[7] B. L. Lambert, "Predicting look-alike and sound-alike medication errors," *Am. J. Heal. Pharm.*, vol. 54, no. 10, pp. 1161–1171, 1997.

[8] B. L. Lambert, K.-Y. Chang, and S.-J. Lin, "Effect of orthographic and phonological similarity on false recognition of drug names," *Soc. Sci. Med.*, vol.

52, no. 12, pp. 1843–1857, 2001.

[9] WHO, "WHO launches global effort to halve medication-related errors in 5 years," 2017. [Online]. Available: http://www.who.int/mediacentre/news/releases/2017/medication-related-errors/en/.

[10] A. Berman, "Reducing Medication Errors Through Naming, Labeling, and Packaging," *J. Med. Syst.*, vol. 28, no. 1, pp. 9–29, Feb. 2004.

[11] "This is how generic drugs get their names | American Medical Association." [Online]. Available: https://www.ama-assn.org/delivering-care/patient-support-advocacy/how-generic-drugs-get-their-names. [Accessed: 18-Jan-2020].

[12] E. Seoane-Vazquez, R. Rodriguez-Monguio, S. Alqahtani, and G. Schiff, "Exploring the potential for using drug indications to prevent look-alike and sound-alike drug errors," *Expert Opin. Drug Saf.*, vol. 16, no. 10, pp. 1103–1109, Oct. 2017.

[13] FDA, "FDA and ISMP Work to Prevent Medication Errors," vol. 2017, no. Jun 21 2017, 2012.

[14] C. K. L. Or and H. H. L. Wang, "Examining text enhancement methods to improve look-alike drug name differentiation accuracy," in *Proceedings of the Human Factors and Ergonomics Society*, 2013, pp. 645–649.

[15] R. Filik, J. Price, I. Darker, D. Gerrett, K. Purdy, and A. Gale, "The Influence of Tall Man Lettering on Drug Name Confusion," *Drug Saf.*, vol. 33, no. 8, pp. 677–687, Aug. 2010.

[16] P. L. Trbovich and S. Hyland, "Responding to the challenge of look-alike, sound-alike drug names," *BMJ Qual. Saf.*, vol. 26, no. 5, pp. 357–359, 2017.

[17] B. L. Lambert, S. J. Lin, and H. Tan, "Designing safe drug names," *Drug Safety*, vol. 28, no. 6. pp. 495–512, 2005.

[18] FDA, "PDUFA PILOT PROJECT - PROPIETARY NAME CONCEPT PAPER," 2008.

[19] L. M. Tittemore, "The name game." 2017.

[20] FDA, "Guidance for industry. Contents of a complete submission for the evaluation of proprietary names," vol. 2015, 2014.

[21] FDA, "Phonetic and Orthographic Computer Analysis (POCA) program," 2017.

[22] M. B. Shah, L. Merchant, I. Z. Chan, and K. Taylor, "Characteristics that may help in the identification of potentially confusing proprietary drug names," *Ther. Innov. Regul. Sci.*, vol. 51, no. 2, pp. 232–236, 2017.

[23] J. Zobel, G. P. O. Box, and P. Dart, "Phonetic String Matching : Lessons from Information Retrieval," *Proc. 19th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 166–172, 1996.

[24] J. Zobel and P. Dart, "Finding approximate matches in large lexicons," *Softw. Pract. Exp.*, vol. 25, no. 3, pp. 331–345, Mar. 1995.

[25] U. Pfeifer and L. I. VI, "Searching Proper Names in Databases," *October*, vol. 20, pp. 1–13, 1994.

[26] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 1, pp. 1–16, 2007.

[27] D. Bharambe, S. K. Jain, and A. K. Jain, "A Survey: Detection of Duplicate Record." 2012.

[28] T. N. Gadd, "PHONIX: The algorithm," *Progr. Electron. Libr. Inf. Syst.*, vol. 24, no. 4, pp. 363–366, Apr. 1990.

[29] G. Kondrak and B. J. Dorr, "A similarity-based approach and evaluation methodology for reduction of drug name confusion," ALBERTA UNIV EDMONTON, 2003.

[30] G. Kondrak and B. Dorr, "Automatic identification of confusable drug names," *Artif. Intell. Med.*, vol. 36, no. 1, pp. 29–42, Jan. 2006.

[31] G. Kondrak and B. Dorr, "Identification of confusable drug names: a new approach and evaluation methodology," *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, Geneva, Switzerland, p. 952, 2004.

[32] G. Kondrak, "N-gram similarity and distance," *Lect. Notes Comput. Sci.*, vol. 3772, pp. 115–126, 2005.

[33] B. L. Lambert, S. J. Lin, K. Y. Chang, and S. K. Gandhi, "Similarity as a risk factor in drug-name confusion errors: The look-alike (orthographic) and sound-alike (phonetic) model," *Med. Care*, vol. 37, no. 12, pp. 1214–1225, Dec. 1999.

[34] G. Kondrak, "Algorithms for Language Reconstruction," University of Toronto, 2002.

[35] G. Kondrak, "Phonetic alignment and similarity," *Comput. Hum.*, vol. 37, no. 3, pp. 273–291, 2003.

[36] A. Kotal, "A New Algorithm to Find Longest Common Sub- sequence's," *Int. J. Sci. Eng. Res.*, vol. 4, no. 5, pp. 664–669, 2013.

[37] M.-A. M. Patsaraporn Somboonsak, "A New Edit Distance Method for Finding Similarity in Dna Sequence," *Int. J. Biol. Biomol. Agric. Food Biotechnol. Eng.*

Vol5, No10, 2011, vol. 5, no. 10, pp. 622–626, 2011.

[38] U. Pfeifer, T. Poersch, N. Fuhr, and L. I. Vi, "Searching Proper Names in Databases," in *HIM*, 1995, pp. 259–275.

[39] U. Pfeifer, T. Poersch, and N. Fuhr, "Retrieval Effectiveness of Proper Name Search Methods."

[40] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, 1966, vol. 10, no. 8, pp. 707–710.

[41] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, 2013.

[42] D. Pinto, D. Vilariño, Y. Alemán, H. Gómez, N. Loya, and H. Jiménez-Salazar, "The Soundex Phonetic Algorithm Revisited for SMS Text Representation," 2012.

[43] B. L. Lambert *et al.*, "Listen carefully: the risk of error in spoken medication orders," *Soc. Sci. Med.*, vol. 70, no. 10, pp. 1599–1608, 2010.

[44] D. Lin, "An Information-Theoretic Definition of Similarity," pp. 1–3, 1842.

[45] S. Bandyopadhyay and S. Saha, "Similarity Measures," in *Unsupervised Classification*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 59–73.

[46] R. A. Wagner and M. J. Fischer, "The String-to-String Correction Problem," *J. ACM*, vol. 21, no. 1, pp. 168–173, 1974.

[47] A. Aneja, A. R. Patki, and R. Kumbhalwar, "Approximate proper name matching," 2007.

[48] S. Gupta, A. P. Srivastava, and S. Awasthi, "Fast and Effective Searches of

Personal Names in an International Environment," *Int. J. Innov. Res. Eng. Manag.*, vol. 1, no. 1, 2014.

[49] J. J. Pollock and A. Zamora, "Automatic spelling correction in scientific and scholarly text," *Commun. ACM*, vol. 27, no. 4, pp. 358–368, 1984.

[50] L.-C. Chen, C.-H. Chen, H.-M. Chen, and V. S. Tseng, "Hybrid data mining approaches for prevention of drug dispensing errors," *J. Intell. Inf. Syst.*, vol. 36, no. 3, pp. 305–327, 2011.

[51] B. L. Lambert, C. Yu, and M. Thirumalai, "A system for multiattribute drug product comparison," *J. Med. Syst.*, vol. 28, no. 1, pp. 31–56, 2004.

[52] T. Nagata, M. Kimura, and F. Tsuchiya, "Similarity index for sound-alikeness of drug names with pitch accents," *Procedia Comput. Sci.*, vol. 35, pp. 1519–1528, 2014.

[53] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, vol. 398. John Wiley & Sons, 2013.

[54] USP, "USP quality review (76). US Pharmacopeia," 2001.

[55] B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, 2009.

[56] M. Consens and G. Navarro, Eds., *String Processing and Information Retrieval*, vol. 3772. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.