Rev Med Chile 2019; 147: 1229-1238

Obtención automática de palabras clave en textos clínicos: una aplicación de procesamiento del lenguaje natural a datos masivos de sospecha diagnóstica en Chile

FABIÁN VILLENA[1,a], JOCELYN DUNSTAN[1,2,b]

Automatic keyword retrieval from clinical texts: an application of natural language processing to massive data of Chilean suspected diagnosis

Background: *Free-text imposes a challenge in health data analysis since the lack of structure makes the extraction and integration of information difficult, particularly in the case of massive data. An appropriate machine-interpretation of electronic health records in Chile can unleash knowledge contained in large volumes of clinical texts, expanding clinical management and national research capabilities.* ***Aim:*** *To illustrate the use of a weighted frequency algorithm to find keywords. This finding was carried out in the diagnostic suspicion field of the Chilean specialty consultation waiting list, for diseases not covered by the Chilean Explicit Health Guarantees plan.* ***Material and Methods****: The waiting lists for a first specialty consultation for the period 2008-2018 were obtained from 17 out of 29 Chilean health services, and total of 2,592,925 diagnostic suspicions were identified. A natural language processing technique called Term Frequency–Inverse Document Frequency was used for the retrieval of diagnostic suspicion keywords.* ***Results:*** *For each specialty, four key words with the highest weighted frequency were determined. Word clouds showing words weighted by their importance were created to obtain a visual representation. These are available at cimt.uchile.cl/lechile/.* ***Conclusions:*** *The algorithm allowed to summarize unstructured clinical free-text data, improving its usefulness and accessibility.*

(Rev Med Chile 2019; 147: 1229-1238)

Key words: *Data Mining; Information Storage and Retrieval; Machine Learning; Medical Informatics; Natural Language Processing.*

[1]Centro de Informática Médica y Telemedicina, Facultad de Medicina, Universidad de Chile. Santiago, Chile.
[2]Centro de Modelamiento Matemático, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile. Santiago, Chile.
[a]Cirujano Dentista.
[b]Física, PhD en Matemática Aplicada y Física Teórica.

FV y JD reciben apoyo del centro de costos 570111 - CIMT-CORFO. JD recibe además CMM-Basal AFB 170001.
Las organizaciones que financiaron este trabajo no tuvieron influencia en el diseño del estudio; en la recolección, análisis o interpretación de los datos ni en la preparación, revisión o aprobación del manuscrito.

Recibido el 28 de abril de 2019, aceptado el 13 de agosto de 2019.

Correspondencia a:
Jocelyn Dunstan
Avenida Independencia 1027, Independencia, Santiago, Chile.
jdunstan@uchile.cl

Los registros clínicos existen para documentar los síntomas iniciales del paciente, diagnósticos, medicamentos, tratamientos y resultados de estos tratamientos, teniendo además un carácter legal[1]. La información contenida en estos registros puede ser clasificada en estructurada y no estructurada. En el primer caso, se trata de datos que pueden ser categóricos o numéricos. En cambio, se dice que la información es no estructurada cuando no puede tabularse en una planilla de datos, como la información contenida en radiografías o historias clínicas donde no hay un lenguaje controlado.

En la práctica clínica, el texto libre no estruc-

turado constituye una proporción importante de la información de pacientes. Ejemplos de estos son anamnesis, informes de exámenes o notas diarias de pacientes hospitalizados, y es de gran valor el poder extraer información de estos datos y utilizarlos para mejorar la gestión e investigación clínica[2]. Actualmente, las historias clínicas electrónicas contienen una cantidad cada vez mayor de datos, y la información no estructurada es a menudo descartada en proyectos informáticos[1].

La codificación del texto clínico tiene como objetivo estructurar la información y que esta pueda ser fácilmente utilizada para realizar labores de gestión, clasificar enfermedades, realizar estadísticas o tomar decisiones[1]. Ejemplos de codificaciones son la Clasificación Internacional de Enfermedades (CIE) y la Nomeclatura Sistematizada de Términos de Medicina Clínica (SNOMED-CT). Sin embargo, es un proceso que requiere entrenamiento y que toma tiempo, ya sea del personal a cargo de la atención, como de quienes se dedican exclusivamente a codificar. Además, en un experimento realizado en Estados Unidos de Norteamérica, se estimó que solo 56% de un conjunto de diagnósticos estaba apropiadamente codificados en CIE-10. Asimismo, es común en la practica clínica utilizar códigos genéricos y luego expresar en texto libre un diagnóstico más preciso del paciente[3]. Es por lo anterior que creemos esencial desarrollar herramientas computacionales que apoyen el procesamiento de textos clínicos producidos en nuestro país.

El análisis automatizado de texto libre se distingue de aquel basado en reglas en que la intervención humana es menor, y se busca que el computador aprenda características del texto a partir de numerosos ejemplos[4]. En particular, el análisis automático de texto clínico tiene desafíos adicionales debido al uso extensivo de abreviaciones y acrónimos para describir el mismo concepto médico, presencia de negación, incertidumbre diagnóstica, jerga local en sus narrativas, disponibilidad restringida de textos por razones de privacidad y falta de herramientas computacionales para idiomas distintos del inglés[1,5].

La metodología computacional que permite el análisis del texto y discuso producido por humanos se conoce como procesamiento del lenguaje natural (PLN)[6]. Uno de los usos básicos del PLN es el de obtener palabras clave dentro de un conjunto de documentos, donde nos interesan vocablos que se repiten dentro de esa unidad, pero que no son comunes cuando se considera el total de los documentos. Ejemplos de palabras que deseamos ignorar son conectivos (ej. "a", "de") o expresiones comúnmente usadas en medicina que probablemente no desean ser identificadas como clave (ej. "paciente", "tratamiento"). Una forma de obtener conjunto de palabras relevantes sin tener que explícitamente eliminar palabras con bajo aporte semántico es mediante el cálculo de frecuencia ponderada[7], técnica central de este artículo.

El número de aplicaciones clínicas de PLN en inglés es mayor que en otros idiomas debido a su simplicidad como lenguaje, así como también a la gran cantidad de recursos computacionales y de texto anotado que existe en inglés[5]. Sin embargo, existen reportes de existosas aplicaciones de PLN a textos clínicos en español, tales como la identificación de enfermedades o medicamentos dentro de narrativas médicas[8,9] o la detección de expresiones que indican negación[10]. Sin embargo, la escasez de grandes volúmenes de texto clínico en español abiertos a la comunidad científica sigue siendo una dificultad.

En este artículo ejemplificaremos el uso del algoritmo de frecuencia ponderada para encontrar las palabras que mejor definen las razones de interconsulta por especialidad en hospitales públicos chilenos para patologías no cubiertas por las garantías explícitas de salud (no-GES). Hemos escogido este tema dado que la gestión de listas de espera es un desafío para el sistema sanitario. Asimismo, este tópico resulta adecuado para ejemplificar el método, puesto que los motivos de interconsulta se expresan como texto no estructurado y se organizan en especialidades médicas y odontológicas, en donde es útil determinar qué es lo frecuente dentro de estas agrupaciones y que no es común en la totalidad del texto, que en este caso sería la lista de espera completa.

Datos y Método

El sistema de gestión de tiempos de espera (SIGTE) es una base de datos, uniforme en el territorio nacional[11], que contiene interconsultas de pacientes[12]. Esta posee información personal de cada uno de los pacientes junto con los datos de la interconsulta, por ejemplo, la especialidad a la que se refiere y la sospecha diagnóstica en formato de texto libre no estructurado[11].

Se enviaron solicitudes por Ley de Transparencia a los 29 servicios de salud del país, pidiendo la lista de espera no cubierta por el plan GES para nueva consulta de especialidad. Diecisiete servicios (58,6%) respondieron a la solicitud, enviando información anonimizada que se resume en la Tabla 1.

Preprocesamiento

Ciertos servicios de salud realizan subconjuntos periódicos de la lista de espera y como muchas de estas no se resuelven en dicho período, puede que una interconsulta aprezca en más de un subconjunto. Para evitar la redundancia, se tomaron en cuenta solo una de las interconsultas que tenían la misma fecha de nacimiento, fecha de entrada, especialidad y sospecha diagnóstica. Además, se eliminaron las interconsultas que tenían datos faltantes en sus celdas.

Puesto que la norma técnica para el registro de lista de espera establece que existen 40 especialidades médicas y 11 odontológicas[11], se procedió a unificar los nombres de especialidades con el fin de agruparlas bajo un mismo concepto. Por ejemplo, Dermatología adulto y Dermatología infantil fueron agrupadas en Dermatología.

Para normalizar cada uno de los motivos de interconsulta, el primer paso fue transfomar el texto a letras minúsculas y eliminar caracteres no alfabéticos y tildes. Un cuerpo de texto se conoce con el nombre de *corpus*[4], y en este caso lo constituye el total de palabras en las 2,6 millones de interconsultas analizadas. Los motivos de interconsulta agrupados por especialidad se denominarán *documentos*. La Figura 1 muestra un diagrama del preprocesamiento realizado.

Se recopilaron 2.982.836 interconsultas, de las cuales se eliminaron 302.076 (10,4%) por estar duplicadas y 87.835 (3,0%) porque tenían celdas nulas, por lo que se trabajó con 2.592.925 (89,6%) interconsultas. Se obtuvieron consultas de 50 especialidades y los atributos del *corpus* se detallan en la Tabla 2.

Determinación de palabras clave

Para la representación de la relevancia de una

Tabla 1. Descripción de los atributos de las listas de espera no-GES recibidas desde los distintos servicios de salud

Servicio de Salud	Año mínimo	Año máximo	Interconsultas
Metropolitano Central	2013	2017	522.579
Atacama	2008	2017	385.124
Ñuble	2007	2018	376.868
Araucanía Sur	2008	2017	360.622
Aysén	2014	2017	218.017
Talcahuano	2009	2017	200.435
Chiloé	2008	2017	188.333
Osorno	2013	2018	183.360
Arica	2010	2018	134.402
Metropolitano Norte	2010	2017	124.852
Coquimbo	2011	2018	95.456
Metropolitano Occidente	2017	2018	88.863
Antofagasta	2016	2018	36.526
Iquique	2013	2018	21.323
Magallanes	2018	2018	21.204
Reloncaví	2015	2017	16.073
Aconcagua	2013	2018	8 799

Tabla 2. Atributos de cada una de las especialidades a analizar

Especialidad	Interconsultas	Palabras	Palabras Únicas
Especialidades médicas			
Anestesiología	1.370	5.633	938
Broncopulmonar	27.640	154.847	7.175
Cardio cirugía	232	973	318
Cardiología	88.770	452.317	14.031
Cirugía abdominal	24.828	72.590	2.805
Cirugía adulto	125.425	692.644	16.630
Cirugía de mamas	10.669	88.919	2.160
Cirugía infantil	40.501	236.385	7.251
Cirugía plástica	1.572	10.898	1.059
Cirugía proctológica	12.776	46.952	1.940
Cirugía tórax	437	1.975	621
Cirugía vascular periférica	45.331	250.078	6.399
Cirugía y traumatología maxilofacial	27.791	186.528	5.694
Dermatología	125.350	479.640	12.389
Endocrinología	37.179	168.487	7.805
Enfermedades de transmisión sexual	3.007	9.592	461
Gastroenterología	65.611	342.801	11.176
Genética	1.306	6.356	1.176
Geriatría	2.330	9.555	1.572
Ginecología	152.237	771.400	14.642
Hematología	8.802	40.345	2.527
Infectología	806	7.520	1.830
Medicina familiar	300	932	118
Medicina física y rehabilitación	21.393	83.946	4.865
Medicina interna	104.124	518.031	19.161
Nefrología	24.689	144.504	7.461
Neonatología	180	946	275
Neurocirugía	36.289	179.685	6.533
Neurología	139.737	668.118	19.986
Nutrición	1.938	7.372	520
Obstetricia	31.395	160.277	3.544
Oftalmología	416.761	1.962.614	15.149
Oncología	4.392	36.644	2.932
Otorrinolaringología	199.827	868.546	16.783
Pediatría	48.240	305.273	14.571
Psiquiatría	38.019	247.915	10.431
Reumatología	23.868	79.561	3.621
Traumatología	227.555	1.216.304	21.069
Urología	87.224	441.363	11.477
Especialidades Odontológicas			
Cirugía bucal	22.386	119.806	4.990
Cirugía maxilofacial	29.879	140.477	5.123
Endodoncia	98.694	404.067	5.336
Odontología indiferenciado	661	2.930	415
Odontopediatría	23.233	119.642	3.647
Operatoria	2.027	10.113	700
Ortodoncia	48.096	322.890	4.598
Periodoncia	31.339	140.046	4.127
Rehabilitación: prótesis fija	14.578	69.410	2.078
Rehabilitación: prótesis removible	108.081	776.881	4.592
Trastornos temporomandibulares y dolor orofacial	4.050	24.152	768

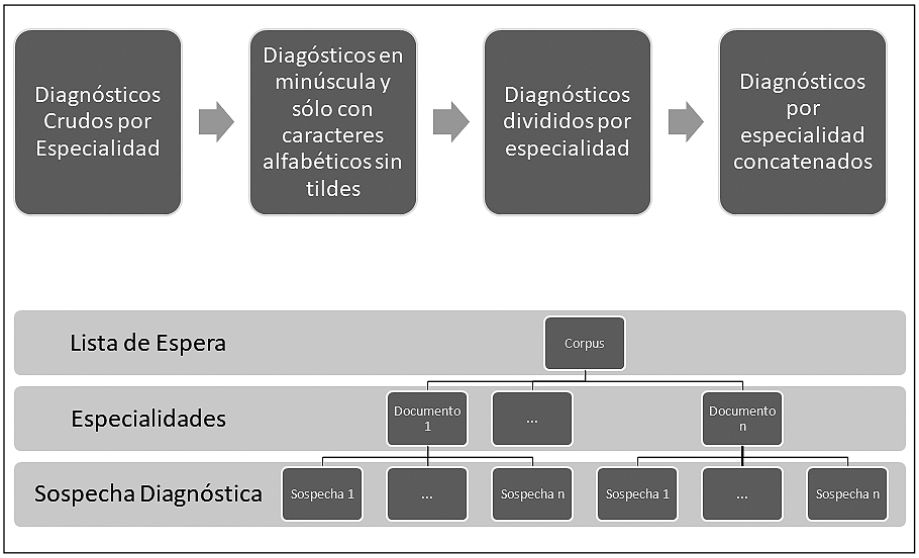

Figura 1. Arriba: Flujo de preprocesamiento de las sospechas diagnósticas. Abajo: Modelo de datos utilizado para el análisis del texto de las sospechas diagnósticas.

palabra dentro de una especialidad se buscó destacar aquellas palabras que aparecen frecuentemente en esa especialidad, pero no dar importancia a aquellas que aparecen frecuentemente en todas las sospechas diagnósticas. Por ejemplo, "paciente" en principio es una palabra que aparece frecuentemente en una especialidad, pero cuando se considera en el contexto del *corpus* es tan común que no aporta información clave o, dicho de otro modo, no es específica a esa especialidad. Distinta es la situación de "anemia", que sí es relevante para hematología, ya que es frecuente en esa especialidad y no lo es en la lista de espera en general.

Lo descrito anteriormente se logra cuantificar con la metodología de frecuencia de término (FT)-frecuencia inversa del documento (FID)[7], que consiste en ponderar positivamente las palabras en función de su frecuencia dentro del documento y negativamente en función de la frecuencia de los documentos que contienen la palabra.

Esta ponderación, FT-FID, se define como:

$$FT\text{ - }FID = FT(t,d)*FID(t),$$

en donde $FT(t,d)$ es la frecuencia bruta del término t en el documento d, y la frecuencia inversa del término $FID(t)$ es:

$$FID(t)=log\ \left(\frac{n_d}{FD(d,t)}\right)+1,$$

con n_d el número de documentos en el *corpus* y $FD(d,t)$ la frecuencia bruta de documentos d que contienen el término t. Esta definición se refiere a que deseamos contar el número de veces que aparece una palabra dentro de una especialidad, pero también queremos ponderar esta frecuencia por el número de veces que aparece en el total de las especialidades. Por ejemplo, si una palabra dada aparece 1.000 veces en una cierta especialidad, este conteo podría aumentar o disminuir dependiendo de si esa palabra está presente principalmente en esa especialidad o si es más bien común a toda la lista de espera.

Palabras dentro de un documento que tienen un FT-FID alto tienen una mayor relevancia dentro del documento[13]. Luego de obtener los valores de FT-FID para cada una de las palabras dentro de cada una de las especialidades, se determinó como palabras clave las 4 palabras con el mayor valor de FT-FID. Por otra parte, se identificaron como palabras sin contenido a aquellos vocablos frecuentes en el *corpus* (bajo FID), pero que en general no aportan especificidad en un documento[13]. Se estableció un umbral en el percentil 95 de la distribución de FID para decir que una palabra no aporta especificidad.

Nubes de palabras

Las nubes de palabras son colecciones de vocablos organizados en forma compacta, en donde el tamaño de la fuente codifica la relevancia de cada

palabra en el documento[14,15]. Para la generación de las nubes de palabras se mapeó el valor de FT-FID de la palabra al tamaño de esta. El tamaño de las palabras va descendiendo de manera logarítmica, con un punto de corte dado por el área disponible y en donde la ubicación de cada una de estas es determinada al azar.

Todos los procesamientos y análisis de los datos se realizaron con el lenguaje de programación Python 3[16], utilizando las bibliotecas NumPy[17], Pandas[18], Scikit-Learn[19] y NLTK[20].

Resultados

Para cada una de las 50 especialidades médicas y odontológicas se crearon nubes de palabras que están a libre disposición en cimt.uchile.cl/lechile, y que son actualizadas en la medida que nueva información se vuelve disponible[21]. Un ejemplo de nube de palabras se presenta en la Figura 2 para la especialidad de Genética.

Tal como se indicó anteriormente, se establecieron como palabras clave las 4 palabras con el mayor FT-FID por especialidad, las cuales se encuentran descritas en la Tabla 3.

En la Tabla 3 podemos ver las palabras con alta frecuencia ponderada, pero no sabemos la magnitud relativa de estas. Para no sobrecargar con información esta tabla, no se muestran los valores de FT-FID para cada palabra, pero sí se encuentran descritos en la Figura 3 para la especialidad de genética. Aquí se puede apreciar la importancia relativa entre las palabras y comprobar que en efecto todas ellas son atingentes a la especialidad médica de Genética.

Figura 2. Nube de palabras de la especialidad de Genética. El resto de las especialidades puede encontrarse en referencia 21.

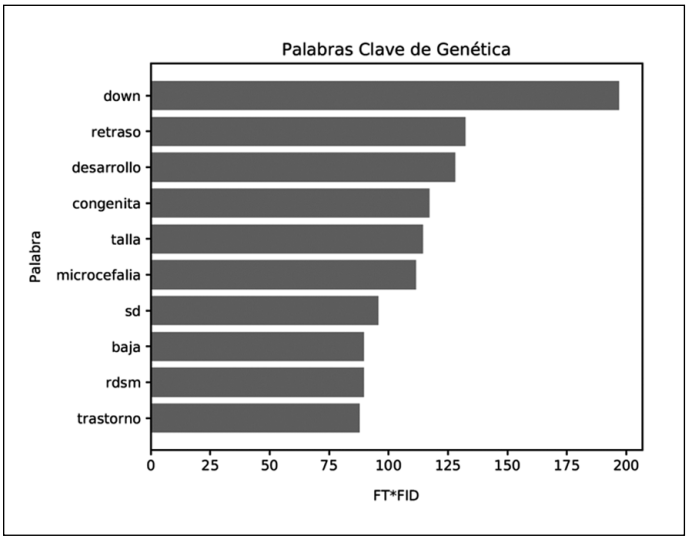

Figura 3. Palabras clave dentro de la especialidad de Genética junto con su valor de TF*IDF ordenadas de mayor a menor valor. En este ranking la palabra *sd* significa síndrome y *rdsm* significa retraso del desarrollo psicomotor.

Tabla 3. Palabras claves de diagnósticos por especialidad

	Palabras claves			
Especialidad	**1**	**2**	**3**	**4**
Anestesiología	pase	operatorio	colelitiasis	hernia
Broncopulmonar	asma	pulmonar	bronquial	pulmonares
Cardiocirugía	venosa	insuficiencia	varices	cardiaca
Cardiología	cardiaca	soplo	cardiaco	angina
Cirugía abdominal	colelitiasis	hernia	inguinal	umbilical
Cirugía adulto	hernia	colelitiasis	inguinal	abdominal
Cirugía bucal	incluidos	dientes	diente	pieza
Cirugía de mamas	mama	anormales	hallazgos	imagen
Cirugía infantil	fimosis	redundante	parafimosis	prepucio
Cirugía maxilofacial	dientes	incluidos	impactados	pericoronaritis
Cirugía plástica	mama	piel	hipertrofia	queloide
Cirugía proctológica	hemorroides	internas	externas	anal
Cirugía tórax	pulmonar	pulmonares	tórax	nódulos
Cirugía vascular periférica	varices	venosa	miembros	inferiores
Cirugía y traumatología maxilofacial	incluidos	dientes	impactados	diente
Dermatología	dermatitis	nevo	verrugas	psoriasis
Endocrinología	hipotiroidismo	tóxico	bocio	tiroideo
Endodoncia	caries	pulpitis	pulpa	necrosis
Enfermedades de transmisión sexual	anogenitales	venéreas	verrugas	sífilis
Gastroenterología	gastritis	reflujo	gastroesofágico	úlcera
Genética	Down	retraso	desarrollo	congénita
Geriatría	demencia	alzhéimer	quejas	cognitivo
Ginecología	útero	prolapso	esterilización	leiomioma
Hematología	anemia	trombocitopenia	especificado	anemias
Infectología	receta	evaluación	fundamento	atraves
Medicina familiar	consultas	depresión	ansiosa	distimia
Medicina física y rehabilitación	lumbago	pie	tendinitis	hombro
Medicina interna	insuficiencia	artritis	renal	hipotiroidismo
Nefrología	renal	insuficiencia	etapa	nefropatía
Neonatología	nacer	recién	peso	neonatal
Neurocirugía	lumbar	disco	pulposo	núcleo
Neurología	cefalea	epilepsia	demencia	migraña
Nutrición	obesidad	observación	afección	calorías
Obstetricia	embarazo	supervisión	riesgo	alto
Odontología indiferenciado	pase	ca	artrosis	operatorio
Odontopediatría	caries	dental	policaries	difícil
Oftalmología	refracción	vicio	especificado	trastorno
Oncología	maligno	tumor	órganos	digestivos
Operatoria	caries	dental	periapicales	dentales
Ortodoncia	anomalías	maloclusión	diente	dentofaciales
Otorrinolaringología	hipoacusia	hipertrofia	conductiva	adenoides
Pediatría	fundamento	aps	clínico	desarrollo
Periodoncia	periodontitis	gingivitis	periodontales	severa
Psiquiatría	trastorno	depresión	depresivo	personalidad
Rehabilitación: prótesis fija	endodonticamente	pieza	tratada	dientes
Rehabilitación: prótesis removible	desdentado	dientes	parcial	extracción
Reumatología	artritis	reumatoide	fibromialgia	artrosis
Trastornos temporomandibulares y dolor orofacial	temporomaxilar	articulación	trastornos	trast
Traumatología	artrosis	fractura	rodilla	pie
Urología	próstata	hiperplasia	renal	urinaria

Discusión

A través de la determinación de las palabras clave por especialidad fue posible detectar diferencias entre estas cuando el total de la lista de espera fue analizado. Se consideraron datos de todas las especialidades excepto Salud Ocupacional, la cual no registró pacientes en los servicios de salud analizados.

Se determinó que existen especialidades con palabras clave muy similares, como es el caso de Cirugía Maxilofacial, Cirugía y Traumatología Maxilofacial y Cirugía Bucal, que comparten 4 de 5 palabras clave. Esto podría tener origen en que sus campos de acción son similares y los profesionales refieren las mismas patologías a cada especialidad de manera indistinta. Esto se refleja en reportes de la comunidad odontológica que sugieren analizar las diferencias entre estas especialidades[22,23].

La utilización de la técnica de FT-FID nos permitió extraer de manera automática los términos clave del campo de sospecha diagnóstica agrupado por especialidad, existiendo también limitantes en su funcionamiento. Un ejemplo de ello son las especialidades de Infectología y Pediatría, en donde las palabras clave extraídas no comunican el contexto de sus diagnósticos. Esto se debe a que existía una gran cantidad de razones de interconsulta idénticas en Infectología: "evaluación y receta" y "SIC a través de citación espontánea Consulta, no especificada", además de "Fundamento Clínico APS" en Pediatría. Para robustecer el método proponemos un enfoque semiautomático para la determinación de las palabras sin sentido, donde expertos pudiesen revisar tanto las listas de palabras clave como las consideradas semánticamente pobres.

La presencia de términos distintos, pero semánticamente iguales, como es el caso de "diente" y "dientes" es un problema que el método no puede resolver. Esto podría ser solucionado mediante la técnica de lematización de términos, que consiste en reducir todas las inflexiones de una palabra a su forma base o *lema* ("dientes" se reduciría a "diente"). Desafortunadamente, la aplicación de lematización a nuestro *corpus* usando la herramienta con rendimiento del estado del arte en español[24], arrojó resultados inconsistentes. Otra reducción de dimensionalidad que podría implementarse sería considerar como sinónimos los términos "pieza" y "diente", lo cual podría obtenerse usando *word embeddings*[25], que es una técnica que permite reducir dimensionalidad a partir del uso de redes neuronales, y es algo en lo que nuestro grupo trabaja actualmente.

Las nubes de palabras se utilizan para la comunicación de palabras clave y ejemplos de su uso son el refuerzo a las familias de pacientes fallecidos que han sido tratados mediante un enfoque de medicina narrativa[14]; como forma de resumir los artículos publicados en revistas científicas en función de los años[26] y para el análisis de datos clínicos cualitativos[27].

En nuestro caso, este recurso permite comunicar de manera simple el contenido de una lista de espera masiva y que registra una alta mortalidad[28,29], y nuestro deseo es contribuir a los tomadores de decisiones y al público objetivo de la red pública de salud mediante la visualización de grandes volúmenes de texto no estructurado. Se usó la lista de espera como un ejemplo de aplicación del método, pero su uso y potencial se extiende a cualquier conjunto de datos en donde se encuentre una gran cantidad de texto no estructurado. Una posible aplicación podría ser la de encontrar, dentro de la ficha clínica de un paciente, las palabras clave de cada consulta médica, y de ese modo informar de manera ágil las consultas anteriores de esa persona. Otra aplicación de este método es la de visualizar palabras frecuentes dentro de un conjunto de reclamos, lo cual permitiría tener una idea rápida de cuales son los aspectos más deficientes dentro del sistema.

Conclusiones

Se logró una aplicación de la técnica FT-FID a texto no estructurado producido en el contexto clínico chileno, eligiéndose la lista de espera para nueva especialidad no-GES para ejemplificar su acción.

Este trabajo también entrega una forma gráfica de visualizar la lista de espera usando nubes de palabras por especialidad, la cual se encuentra disponible en línea y es actualizada en la medida que nuevos datos están disponibles. El método aquí expuesto permite obtener la importancia de palabras dentro de un conjunto de textos, lo cual en este caso en particular, permite informar a los

tomadores de decisiones acerca del contenido de las razones de interconsulta en Chile.

Agradecimientos: Los autores agradecen a Magdalena Bastías, Cristóbal Cuadrado, Manuel Durán, Claudio Olmos y César Parra por sus comentarios del manuscrito.

FV y JD reciben apoyo del centro de costos 570111 - CIMT-CORFO y DAAD 57220037 & 57168868.

Referencias

1. Dalianis H. *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer; 2018.
2. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform 2017; 26: 38-52.
3. Horsky J, Drucker EA, Ramelson HZ. Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. AMIA. Annu Symp proceedings AMIA Symp. 2017; 2017: 912-20.
4. Manning CD, Schütze H. *Foundations of Statistical Natural Language Processing*. The MIT press. Cambridge, MA, USA.; 1999.
5. Névéol A, Dalianis HK, Savova G, Zweigenbaum P. Clinical Natural Language Processing in Languages Other Than English: opportunities and challenfes. J Biomed Semantics. 2018; 9: 12: 1-13.
6. Hirschberg J, Manning CD. Advances in natural language processing. Science (80-). AAAS 2015; 349 (6245): 261-6.
7. Manning CD, Raghavan P, Schutze H. *An Introduction To Informational Retreival*. Cambridge University Press, Cambridge, MA, USA. 2009. 1-18 p.
8. Oronoz M, Casillas A, Pérez A, Gojenola K, Dalianis H, Weegar R. Semi-supervised medical entity recognition: A study on Spanish and Swedish clinical corpora. J Biomed Inform 2017; 71: 16-30.
9. Weegar R, Pérez A, Casillas A, Oronoz M. Deep Medical Entity Recognition for Swedish and Spanish. IEEE Int Conf Bioinforma Biomed. IEEE 2018; 1595-601.
10. Santiso S, Casillas A, Pérez A, Oronoz M. *Word embeddings for negation detection in health records written in Spanish*. Soft Comput [Internet]. Springer Berlin Heidelberg; 2018;1-7. Recuperado a partir de: https://doi.org/10.1007/s00500-018-3650-7.
11. Ministerio de Salud de Chile. Norma Técnica Para El Registro De Las Listas De Espera. 2011. Recuperado a partir de: https://www.minsal.cl/wp-content/uploads/2016/03/Norma-Tecnica-118.pdf.
12. Subsecretaría de Redes Asistenciales. Plan Nacional de Tiempos de Espera No GES en Chile. 2014-2018. Recuperado a partir de: https://www.minsal.cl/wp-content/uploads/2018/03/Plan-nacional-de-tiempos-de-espera-No-GES.pdf.
13. Robertson S. Understanding inverse document frequency: On theoretical arguments for IDF. J Doc 2004; 60 (5): 503-20.
14. Vanstone M, Toledo F, Clarke F, Boyle A, Giacomini M, Swinton M, et al. Narrative medicine and death in the ICU: word clouds as a visual legacy. BMJ Support Palliat Care 2016 Nov 24: 1-8.
15. Felix C, Franconeri S, Bertini E. Taking Word Clouds Apart: An Empirical Investigation of the Design Space for Keyword Summaries. IEEE Trans Vis Comput Graph. IEEE 2018; 24 (1): 657-66.
16. van Rossum G, Frake FL. The Python Language Reference Manual. Network Theory Ltd. 2003.
17. Oliphant TE. MIT Guide to NumPy. Methods 2010; 1: 378.
18. McKinney W. Data Structures for Statistical Computing in Python. En: van der Walt S, Millman J, editores. *Proceedings of the 9th Python in Science Conference*. 2010. p. 51-6.
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2012; 12: 2825-30.
20. Bird S, Klein E, Loper E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media; 2009.
21. Villena F, Dunstan J. Visualizador de Lista de Espera Chilena [Internet]. Recuperado a partir de: https://cimt.uchile.cl/lechile/
22. Zúñiga B, González M, González A, Gamonal J. Cirugía y traumatología bucomáxilofacial en la red hospitalaria chilena. Rev clínica periodoncia, Implantol y Rehabil oral 2017; 10 (1): 57-62.
23. Moscoso K. A Propósito de la Cirugía Bucal. Int J Odontostomat 2016; 10 (1): 5-6.
24. Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To Appear [Internet]. 2019; Recuperado a partir de: https://spacy.io/
25. Goldberg Y. A Primer on Neural Network Models for NLP. J Artif Intell Res [Internet] 2016; 57: 345-420. Recuperado a partir de: https://www.jair.org/index.php/jair/article/view/11030.
26. Atenstaedt R. Word cloud analysis of the BJGP. Br J Gen Pr 2012; 62 (596): 148.

27. Bridgett B, Sellars B, Sherrod DR, Chappel-aiken L. Using word clouds to analyze qualitative data in clinical settings. Nurs Manage 2018; 49 (10): 51-3.
28. MINSAL. Estado de situación personas fallecidas en listas de espera no-GES y garantías retrasadas GES. [Internet]. 2017. Recuperado a partir de: http://www.minsal.cl/wp-content/uploads/2018/01/Informe-Final-Comision-Asesora-LE-y-Garantias-Retrasadas-GES-17082017_.pdf.
29. Martínez DA, Zhang H, Bastias M, Feijoo F, Hinson J, Martínez R, et al. Prolonged wait time is associated with increased mortality for Chilean waiting list patients with non-prioritized conditions. BMC Public Health 2019; 19 (1).