Generación automática inteligente de resúmenes de textos con técnicas de Soft Computing

Tesis doctoral realizada en cotutela por:

Augusto Villa Monte

Directores:

Laura Cristina Lanzarini
Universidad Nacional de La Plata
Buenos Aires, Argentina

José Angel Olivas
Universidad de Castilla-La Mancha
Ciudad Real, España

Fecha de exposición:
La Plata, 18 de Marzo de 2019

---

1. Introducción

Muchos años después de que el escritor y científico Toffler [1970] pronosticara que se produciría más información de la que sería posible procesar, la explotación de datos y el descubrimiento de conocimiento se volvieron fundamentalmente necesarios en todos los ámbitos. Así fue como en los últimos tiempos, los avances tecnológicos favorecieron la generación y almacenamiento de grandes volúmenes de datos y, como resultado, se volvió esencial el desarrollo de métodos inteligentes capaces de representar la información disponible [Fayyad et al., 1996].

Si bien los sistemas y aplicaciones actuales generan datos en diferentes formatos, la mayoría produce y almacena texto. Este formato resulta menos atractivo que otros como el sonido, las imágenes y el video, pero es, sin lugar a duda en la actualidad, el principal medio de comunicación entre seres humanos [Schreibman et al., 2016].

Desde la invención de la escritura, el ser humano ha almacenado el conocimiento en textos. Desde entonces, la cantidad de documentos disponibles ha aumentado exponencialmente mientras que el costo de generar, almacenar, duplicar y compartir dicha información fue disminuyendo [Saracco, 2017]. Al mismo tiempo, el consumo de información aumentó notablemente [Johnson, 2011].

Generalmente, los documentos de texto digitales no están estructurados ni mucho menos organizados en bases de datos tradicionales. El texto por sí mismo no tiene ningún tipo de estándar ni restricción al crearse y almacenarse y, por lo tanto, procesarlo se ha vuelto una tarea extremadamente difícil. El desarrollo de soluciones computacionales que permitan resumir texto pretende reducir los problemas generados por el crecimiento desmedido de información textual.

2. Motivación

A partir de la evidente explosión de información, sería ideal que el ser humano pudiera recordar toda la información disponible, pero en ese caso el cerebro colapsaría. El ser humano recuerda muchas cosas al mismo tiempo que olvida muchas otras. Inconscientemente capta la información esencial. Esta tarea, tratándose de texto, se conoce como "resumir".

Durante los últimos 60 años se han logrado grandes avances en lo que refiere a resúmenes automáticos. Sin embargo, desarrollar un programa por computadora que resuma requiere instrucciones precisas [Borko and Berniers, 1975]. Existen dos grandes enfoques para construir un resumen de forma automática, uno de los cuales tiene que ver con la interpretación del significado del contenido del documento (enfoque abstractivo) y el otro con el análisis más bien de su estructura (enfoque extractivo) [Hahn and Mani, 2000]. Esta tesis trata ambos enfoques a través de la propuesta de dos soluciones diferentes aplicadas a documentos médicos.

Contar con estrategias capaces de identificar automáticamente lo principal de un documento facilita su procesamiento, no sólo en lo que se refiere a lectores humanos, quienes son beneficiados agilizándoles la lectura, sino también en lo que se refiere a técnicas de aprendizaje automático y de recuperación de información. Por ejemplo, cuando se hace una búsqueda web resulta más eficiente trabajar sobre algunas partes de los documentos. Esto implica un menor esfuerzo en el proceso de búsqueda al operar únicamente con lo más importante de cada página web. Además, la búsqueda se concreta en menos tiempo y se resuelve más rápido al buscar el documento completo solo en caso de corresponder el criterio de búsqueda con el resumen de los documentos.

3. Objetivos y contribuciones

Esta tesis tiene por objetivo principal contribuir al área conformada por el *Procesamiento de Lenguaje Natural* y la *Minería de Texto* con dos estrategias distintas capaces de identificar a partir de un conjunto de documentos lo relevante y construir con eso un resumen en forma automática. Por un lado, identificando el criterio del usuario al seleccionar las partes principales de un documento y, por otro, extrayendo de los documentos patrones textuales específicos sumamente útiles en la toma de decisiones.

La primera de las soluciones, para documentos con cierta estructura, permite crear resúmenes extractivos utilizando una técnica de *Optimización mediante Cúmulo de Partículas* (PSO por sus siglas en inglés) a partir de la representación vectorial de los mismos basada en un conjunto amplio de métricas de puntuación de sentencias. Dicha técnica identifica el criterio del usuario al seleccionar las partes del documento que considera importante. En lugar de utilizar las distintas métricas en forma independiente para construir el resumen, la respuesta del método sugiere la combinación que mejor se ajusta a la valoración que el usuario realizó de cada parte de un documento en forma previa. El método propuesto, que combina una representación binaria y una continua utilizando una variante original de la técnica mencionada, no sólo permite identificar los coeficientes asociados a cada una de las métricas sino también cuáles son las métricas que permiten resumir lo más parecido al criterio del usuario.

La segunda solución extrae las sentencias causales con restricciones temporales existentes en un conjunto de documentos médicos y luego convierte dichas oraciones en un grafo causal equivalente. Para ello, se identifican y extraen patrones textuales específicos del texto. El modelo proporciona las relaciones que describen el contenido de los documentos originales mostrando únicamente todos los vínculos "causa-efecto" junto con las restricciones temporales que afecten su interpretación. Identificar estos patrones específicos en documentos médicos resulta sumamente útil para la toma de decisiones en el área de salud. El grafo, además, tiene la información necesaria para generar nuevas frases a partir del recorrido entre sus nodos y arcos.
Es importante aclarar que los términos de las afirmaciones que se obtienen del grafo están en los documentos originales pero las afirmaciones que se consiguen no necesariamente son extracciones textuales de los mismos, a diferencia de la primera solución en la cual se transcriben partes textuales de los documentos y que al conformar el resumen final no necesariamente tienen una coherencia narrativa. En las próximas dos secciones se describirán brevemente ambas soluciones.

4. Resumen utilizando una técnica de optimización mediante cúmulo de partículas

Un resumen extractivo está formado por un conjunto de porciones de texto (desde palabras sueltas hasta párrafos enteros) literalmente copiadas de la entrada de datos [Mani, 2001a]. Estas porciones de texto generalmente son llamadas "sentencias" haciendo referencia a las oraciones del mismo. Para extraer las partes del documento que formarán el resumen, este enfoque requiere asignar una puntuación a cada sentencia. Esto permite clasificar el contenido del documento ordenando las sentencias en una lista cuyas primeras posiciones son ocupadas las sentencias más relevantes que recibieron el puntaje más alto [Edmundson and Wyllys, 1961]. Hay muchas maneras de puntuar las sentencias. Cada una de ellas permite seleccionar las "mejores" sentencias para producir el resumen pudiendo variar el resumen resultante en cada caso. En las Secciones 2.10 y 2.11 de la tesis se describe detalladamente la representación vectorial de los documentos de texto y las métricas más utilizadas para puntuar las partes de un documento.
Es así como, los documentos se modelan como vectores n-dimensionales obtenidos de calcular n métricas. Por lo tanto, cada documento está representado por una matriz S de p filas (sentencias) y n columnas (métricas). Dicha vectorización transforma el documento en un conjunto de p vectores "sentencia" de la forma $S_i = [s_{i1}, s_{i2}, \ldots, s_{in}]$. Luego, estos vectores son utilizados para obtener resúmenes automáticos aplicando algoritmos más sofisticados [Nenkova and McKeown, 2012].
En general, obtener un resumen extractivo puede considerarse un problema de clasificación de dos clases en el que a cada parte del documento se la etiqueta como "correcta" si forma parte del resumen, o "incorrecta" caso contrario [Neto et al., 2002]. Trabajos recientes consideran la obtención de resúmenes extractivos un problema de optimización, donde una o más funciones objetivo se formulan para seleccionar las "mejores" oraciones del documento que forman el resumen [Vázquez et al., 2018]. Sin embargo, en este tipo de trabajos los documentos se representan a través de un conjunto de métricas establecido a priori, y su selección no forma parte del proceso de optimización, como sucede en [Meena and Gopalani, 2015] por ejemplo. Esta es la característica clave del método desarrollado en este capítulo de la tesis.
Cuando la solución exacta a un problema es difícil de obtener, las estrategias de búsqueda aproximadas mediante técnicas de optimización han demostrado ser sumamente efectivas. Este tipo

de técnicas mejoran, a través de procesos biológicos, un grupo de soluciones denominado población de individuos. Cada individuo representa una posible solución al problema y se desplaza por el espacio de soluciones observando su propio comportamiento y el de su entorno. En este principio se basa el primero de los métodos propuestos en esta tesis, el cual construye el resumen automático utilizando PSO. Dicha técnica fue propuesta por Kennedy and Eberhart [1995] y desde entonces se han desarrollado diferentes versiones. Originalmente se definió para operar sobre un espacio de búsqueda continuo pero años más tarde Kennedy and Eberhart [1997] definieron una versión discreta. Sin embargo, uno de los principales problemas que tuvo dicha versión fue la dificultad para cambiar de 0 a 1 y de 1 a 0 una vez estabilizado. Esto impulsó el desarrollo de diferentes versiones binarias del PSO que buscaron mejorar su capacidad exploratoria.

Usar PSO para generar, a partir de la representación vectorial de los textos, un resumen extractivo que combine adecuadamente varias métricas de puntuación de sentencias, requiere utilizar ambos tipos de PSO. Por un lado, se requiere seleccionar el subconjunto de métricas a utilizar (parte discreta) y, por el otro, se necesita establecer la relevancia de cada una de estas métricas (parte continua).

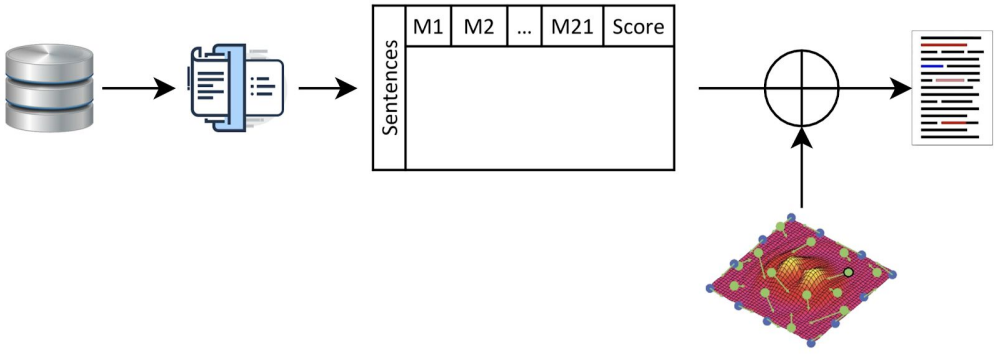

Fig. 1 - Construcción del resumen extractivo utilizando PSO

La Figura 1 muestra cómo se construye el resumen con la estrategia propuesta. En lugar de utilizar las distintas métricas en forma independiente para construir el resumen, la respuesta del método sugiere la combinación que mejor se ajusta a la valoración que el usuario realizó de cada una de las partes de un documento en forma previa. El método propuesto, que combina una representación binaria y una continua utilizando una variante original de la técnica mencionada, no sólo permite identificar los coeficientes asociados a cada una de las métricas sino también cuáles métricas permiten resumir lo más parecido al criterio del usuario.

El PSO inicia su población en forma aleatoria y mide iteración a iteración el desempeño de la solución de cada partícula para resolver el problema. Para eso se construye el resumen de los documentos de entrenamiento utilizando el criterio establecido por la partícula en cada iteración y el resumen obtenido se lo compara con el esperado midiendo cuán parecido ha resultado. Esto justifica la necesidad de preprocesar los documentos una única vez y tenerlos almacenados en forma de vectores, para que cada partícula pueda medir sobre los documentos su desempeño sin tener que volverlos a procesarlos. Una vez evaluadas todas las partículas, se mueve la parte continua de las partículas con un movimiento estándar y la parte binaria con algunas modificaciones propias para favorecer la estabilidad de las mismas una vez conseguido el óptimo. El detalle

completo del funcionamiento del método propuesto puede encontrarse en el Capítulo 3 de la tesis, así como el detalle de la representación y almacenamiento de los documentos en el Capítulo 2 y Anexo B respectivamente.

Para evaluar la calidad del resumen automático producido por el método propuesto, se utilizaron 3322 artículos publicados en PLOS Medicine entre 2004 y 2018. Se realizó el entrenamiento del modelo con los artículos de un mes y el testeo con los del mes siguiente. Se realizaron 30 ejecuciones independientes con 100 iteraciones como máximo y se resumió el 10% del documento. El PSO fue global y de población fija con 10 partículas inicializadas aleatoriamente. Además, para determinar el aporte del método propuesto se lo comparó con un método previo que no poseía la capacidad de seleccionar métricas.

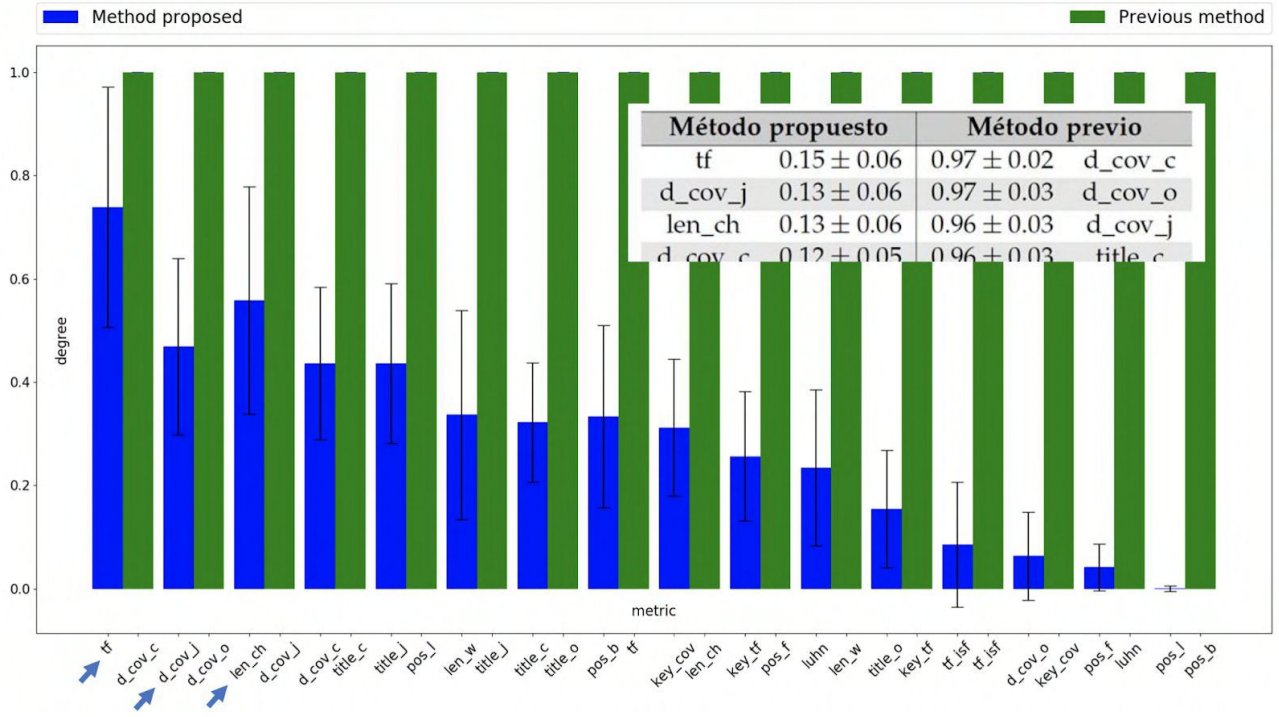

Fig. 2 - Participación de las métricas ordenadas descendente por valor de coeficiente

La Figura 2 muestra el nivel de participación de las métricas para los dos métodos evaluados, en orden decreciente según el valor de su coeficiente promedio indicado en la tabla que se encuentra superpuesta en la imagen. La tabla completa con la media y desviación de los coeficientes de todas las métricas utilizadas se encuentra en la página 86 de la tesis. Estos coeficientes son los que se usan para ponderar el valor de cada métrica al obtener el puntaje de cada sentencia. Por ejemplo, si se observan los tres primeros valores en la columna "Método propuesto" de dicha tabla, se puede observar que los coeficientes promedio de las métricas "tf", "d_cov_j" y "len_ch" son 0.15, 0.13 y 0.13, respectivamente. Por lo tanto, el criterio para resumir hace el mismo énfasis entre "d_cov_j" y "len_ch". Sin embargo, si se observa la Figura 2, se puede ver que el nivel de participación de "len_ch" es mayor que el de "d_cov_j". Esto se debe a que la primera ha sido seleccionada más veces por la técnica de optimización. En cambio, la métrica "tf" para el método propuesto tiene el coeficiente promedio más alto de la tabla y también el nivel más alto de participación. Por otro lado, la Figura 3 muestra cómo la precisión de cada uno de los métodos evoluciona a medida que se

incorporan métricas en la construcción final del resumen. Esto se hace en el orden indicado en la tabla de coeficientes promedio. Como se puede ver, el comportamiento del método propuesto aquí es más estable. Además, después de agregar la cuarta métrica, la precisión se vuelve notablemente mejor que la que se obtiene utilizando todas las métricas. Aunque el valor máximo se observa con la incorporación de la séptima métrica, cuatro métricas serían suficientes para obtener un buen rendimiento. También se debe tener en cuenta que, utilizando el método previo, incluso si la precisión resultante es mayor para las dos primeras métricas, las restantes arrojan un resultado inferior en comparación con el método propuesto, sin lograr superar el valor más alto.

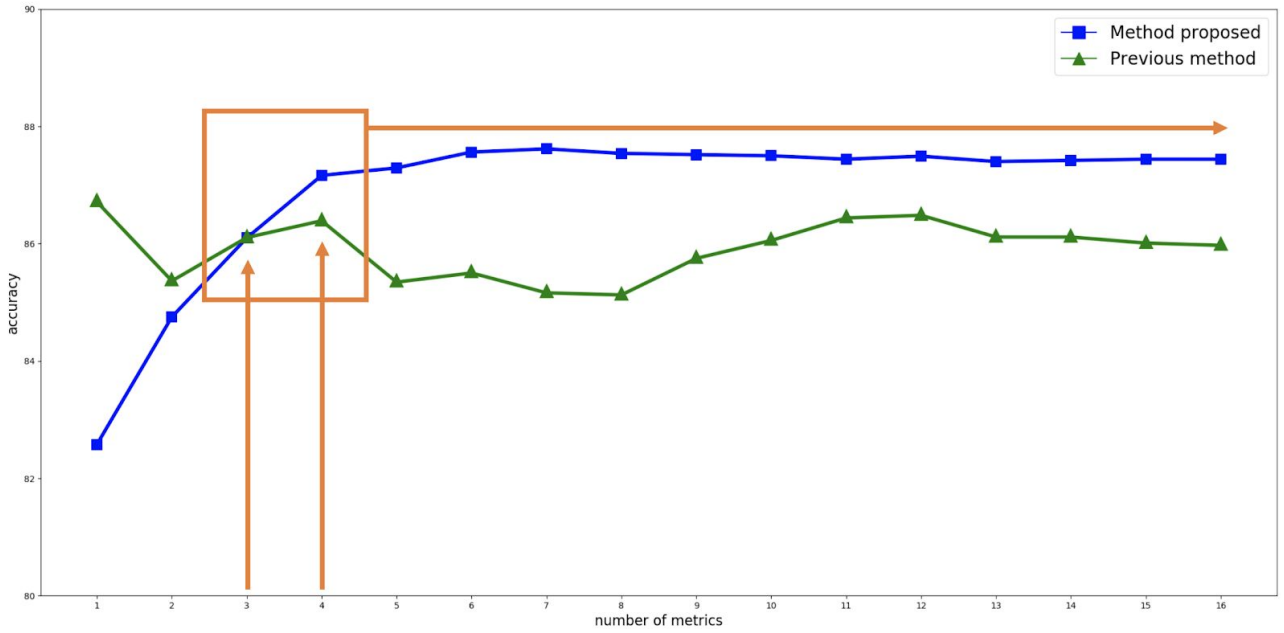

Fig. 3 - Evolución de la precisión a medida que se agregan métricas para obtener el resumen

5. Resumen mediante grafos causales y con componentes temporales

Así como existen métodos de construcción de resúmenes cuyo enfoque es extractivo, existen otros métodos que utilizan estructuras tipo grafo para representar el contenido de un conjunto de documentos a partir de la extracción de patrones textuales específicos. En esta dirección, se han realizado estudios utilizando relaciones conceptuales, centrándose principalmente en aspectos semánticos y no tanto en la causalidad presente en los textos. La causalidad cumple un rol importante en cualquier toma de decisiones ya que los efectos de una decisión pueden determinarse por sus causas. Desde ese punto de vista, la causalidad permite generar conocimiento científico objetivo [Kant et al., 1999; Mill, 1843] y describir fenómenos [Mach, 1976]. Si de Medicina se trata, este tipo de sentencias están presentes al detallar las causas de una enfermedad o al informar los efectos de un tratamiento. Con frecuencia, las Ciencias de la Salud muestran la causalidad como un proceso complejo que involucra la evolución de una causa anterior en causas intermedias a lo largo del tiempo antes de alcanzar el efecto final. La lógica causal ofrece un vocabulario y reglas específicas para explicar y predecir procesos complejos en términos de vínculos “causa-efecto”.
La causalidad es una relación en la cual la ocurrencia de una entidad B de cierta clase depende de la ocurrencia de una entidad A de otra clase. Generalmente por “entidad” se entiende un fenómeno, un

hecho, una característica, una situación o un evento, entre otras cosas. En este tipo de relación, "A" representa una causa y "B" un efecto. Entre ambos existe una clara relación de dependencia ya que la causa provoca un efecto, y el efecto se deriva de la causa considerándola su consecuencia. Se expresa no sólo utilizando el término "cause" sino también "produce", "bring about", "issue", "generate", "result", "effect" o "determine", entre otros [Kim, 1995].

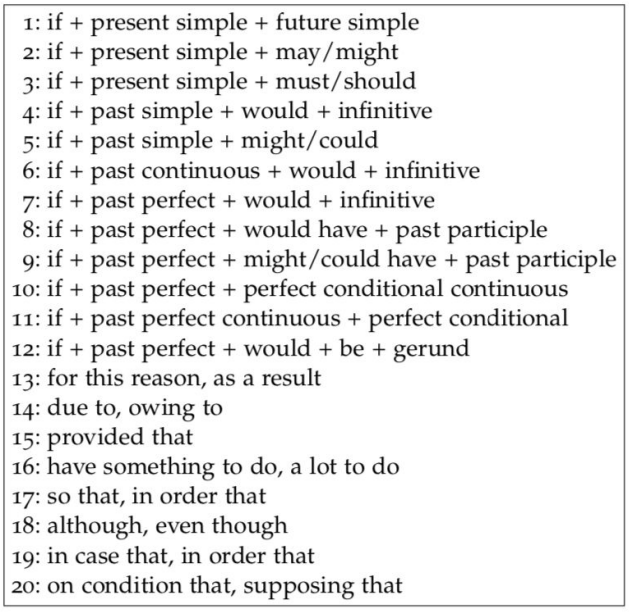

Fig. 4 - Estructuras condicionales y causales implementadas

La causalidad puede ser un proceso directo cuando A causa B y B es un efecto directo de A; o un proceso indirecto cuando A causa C a través de B, y C es un efecto indirecto de A. Tratándose de texto la fuente de información, para extraer las relaciones causales directas se necesita tener un amplio conocimiento del lenguaje que permita definir correctamente una serie de patrones de búsqueda específicos como los de la Figura 4. En cambio, no ocurre lo mismo con las relaciones causales indirectas. Este tipo de relaciones no se puede identificar a través de patrones. A pesar de ello, extrayéndose relaciones causales directas y representándolas adecuadamente en forma de grafo (como el de la Figura 5), a través del recorrido de sus caminos pueden establecerse relaciones indirectas que no eran obvias y que no podrían haberse obtenido de otra manera.

Como se dijo anteriormente, las explicaciones causales relacionan causas y efectos y, en Medicina, a menudo se las restringe temporalmente. El tiempo juega también un papel importante en diferentes escenarios médicos [Keravnou, 1996]:

- *Prevención* evaluando factores de riesgo en pacientes que se comportan de cierta manera ⇒ «Smoking for a long time causes lung cancer»
- *Pronóstico* prediciendo la probabilidad de supervivencia de una persona ⇒ «25 % of patients with septic shock will die within 15 days»
- *Tratamiento* administrando medicamentos durante ciertos períodos de tiempo o bajo ciertas restricciones temporales ⇒ «Omeprazole should be taken before aspirin»
- *Diagnóstico* identificando el origen o la causa de las enfermedades ⇒ «According to the symptoms that he presents, he has flu»

La toma de decisiones basada en el tiempo cubre áreas de la atención al paciente donde resulta fundamental la administración de medicamentos durante ciertos períodos de tiempo o bajo ciertas restricciones temporales.

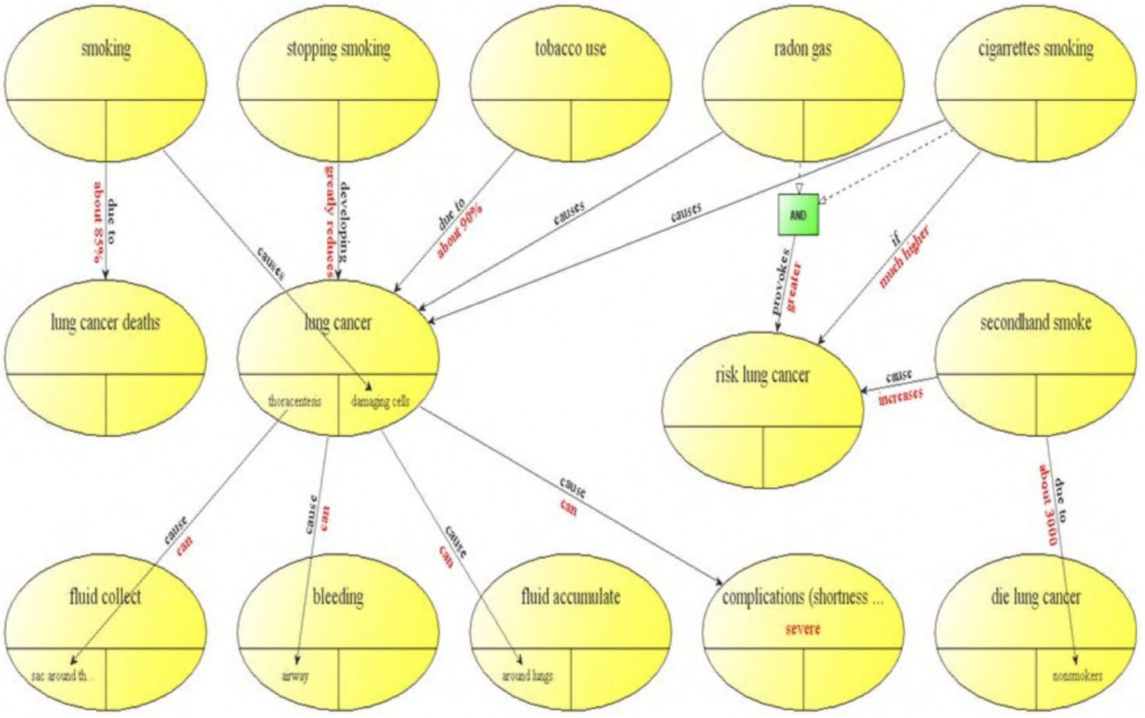

Fig. 5 - Grafo causal relacionado con «lung cancer» construido de forma automática

El tiempo utiliza un léxico específico. Las oraciones temporales incluyen con frecuencia referencias relacionadas con el calendario (año, mes, día) o momentos del día (mañana, tarde, noche). También utilizan conjunciones y preposiciones como “by”, “until”, “before”, “since”, “past”, “next”. Además, hay frases causales como “secondary to” o “because of” que también denotan influencia causal temporal. Para analizar el comportamiento de este tipo de sentencias en documentos de texto, se requiere contar con un proceso capaz de detectar, extraer y clasificar las sentencias con base en ciertos patrones estructurados. En el Capítulo 4 de la tesis podrá encontrarse la descripción completa del estudio realizado sobre las asociaciones causales que implican dependencias temporales y se podrá encontrar también más detalle sobre el proceso que permite detectarlas y extraerlas para construir el grafo causal.

Las sentencias finalmente extraídas por el proceso descrito en la tesis permiten luego crear una base de conocimiento causal sobre un determinado tema. Esto es posible siempre que se procesen documentos de una misma temática. Tratándose de páginas web, pueden conseguirse los documentos a través de una simple búsqueda en Internet y posterior recopilación del contenido de los sitios arrojados como resultado.

Una vez producida la base de conocimiento en un determinado ámbito y sobre un tema en particular, el usuario deberá introducir una pregunta. Dicha pregunta permitirá seleccionar aquellas oraciones que se encuentren directamente relacionadas con ella. Una vez que todas las frases hayan sido procesadas rastreando los conceptos asociados a causas o consecuencias, se construye el grafo

causal que las representará (como el de la Figura 5). En la Figura 6 se muestra el esquema del proceso completo que permite obtener el grafo.

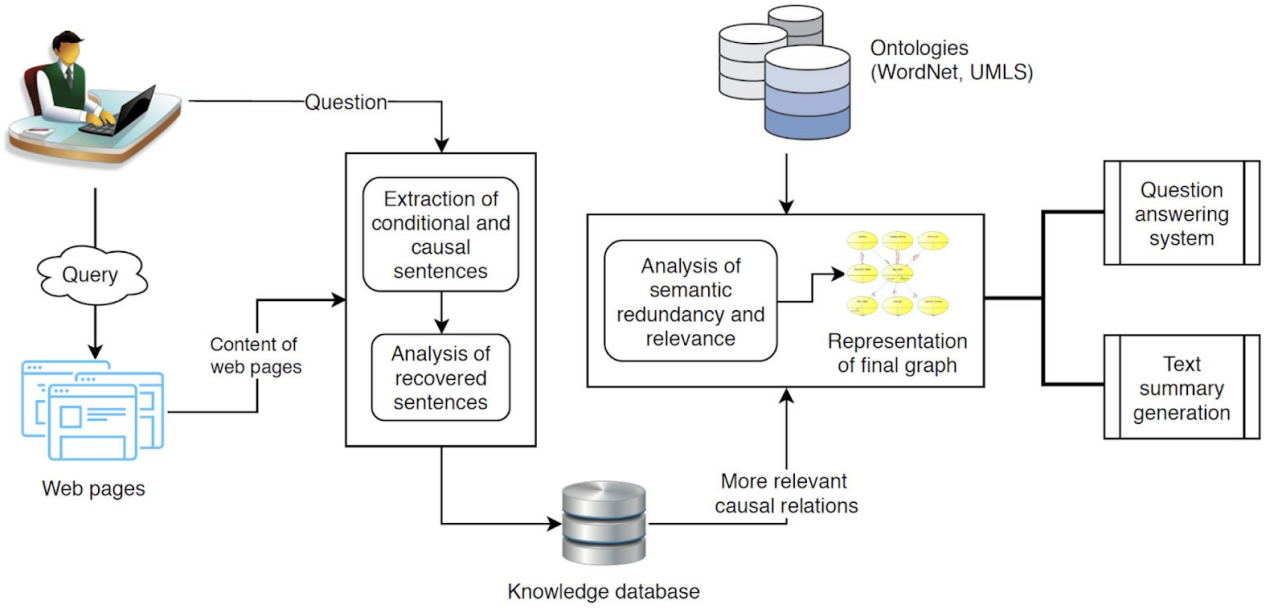

Fig. 6 - Proceso para obtener un grafo causal a partir de texto plano

Los grafos causales son una forma gráfica de mostrar las dependencias causales que posee la información no estructurada como lo es el texto. La causalidad implica una transferencia de la causa al efecto que en el grafo se denota mediante una flecha dirigida. Dicha flecha conecta el nodo "causa" con el nodo "efecto" en el grafo. Como podemos ver en la Figura 6, cada nodo representa un concepto mediante una elipse. Las relaciones entre nodos están representadas por arcos que llevan el tipo de conector causal. Entonces, de la lectura del grafo pueden obtenerse relaciones ocultas.

Para introducir restricciones de tiempo al grafo, se debe buscar por cada nodo aquellas oraciones que tengan modificadores temporales. El siguiente paso consiste en poder identificar si el indicador de tiempo viene asociado al nodo antecedente o al consecuente para introducirlo en el nodo correcto. A partir del grafo, se puede generar un resumen que incluya información causal con componentes temporales.

Aunque existen sistemas que recuperan sentencias causales de los textos, no hay muchos que organicen dichas sentencias para resolver un problema concreto y, mucho menos, haciendo uso de información temporal. Los diagnósticos de enfermedades o la prescripción de medicamentos frecuentemente implican dependencias temporales. En el caso de una enfermedad generalmente depende de cuándo haya aparecido el síntoma y, en el caso de los medicamentos, su uso está limitado a ciertas condiciones temporales.

Cada año la salud de los pacientes se ve perjudicada por la mala administración de medicamentos. Existen estudios que evidencian un alto porcentaje de medicamentos mal administrados en cuidados intensivos [Keers et al., 2013]. Dado que la correcta administración de medicamentos depende del tiempo y, además, se encuentra a cargo de seres humanos (enfermeras generalmente), desarrollar sistemas de alerta puede evitar errores, contribuir con la seguridad del paciente y mejorar la gestión

de los hospitales. Algunos medicamentos, por ejemplo, deben tomarse antes o después de las comidas, siendo “antes” y “después” restricciones temporales. Por eso en esta tesis, para contribuir al manejo del tiempo en tratamientos médicos que deben tener en cuenta el tiempo se propuso el desarrollo de una aplicación llamada “My Medicine”, que permite controlar las restricciones temporales indicadas por los médicos al prescribir tratamientos para ciertas enfermedades. La descripción detallada de la aplicación podrá encontrarla en el Capítulo 5.

6. Conclusiones y líneas de trabajo futuro

En esta tesis se han propuesto dos estrategias basadas en técnicas de Soft Computing para la generación automática de resúmenes de texto. Se las ha evaluado con procedimientos habituales en el área de estudio dando buenos resultados.

Por un lado se ha desarrollado un método para identificar el criterio del usuario al seleccionar las partes principales de un documento por medio de una variante original de PSO. Esta propuesta ha sido evaluada sobre una amplia colección de artículos científicos y los resultados reflejan que con pocas métricas es posible caracterizar el criterio del usuario para resumir. Además, la técnica utilizada no se limita a resúmenes, habiendo sido aplicada también en la obtención de reglas de clasificación.

Por el otro, se ha desarrollado una estrategia de extracción de relaciones causales en los textos a partir de las cuales se construye un grafo con anotaciones temporales que afectan su interpretación. Esto además dio lugar al desarrollo de una aplicación para dispositivos móviles que combina la causalidad y temporalidad para ayudar a controlar la administración de dosis de medicamentos.

También, se ha comparado la calidad de los resúmenes generados de forma extractiva a partir de métricas y aquellos creados a partir de las relaciones causales. El resultado ha demostrado que la calidad está vinculada al tipo de narrativa del documento. Los resúmenes extractivos basado en PSO son más adecuados para compactar el volumen de información textual y los resúmenes abstractivos basados en causales son mejores para generar resúmenes conceptualmente más ricos.

Como líneas de trabajo futuras se plantea ampliar el conjunto de métricas utilizado para caracterizar los documentos de entrada y así enriquecer su representación; incorporar conceptos de Lógica Borrosa o Difusa que permitan flexibilizar el criterio del usuario; continuar con el desarrollo de la aplicación My Medicine para utilizarla además en la gestión hospitalaria; e, incluir nuevas estrategias para verbalizar el grafo causal resultante.

Por último, cabe mencionar que la tesis desarrollada tiene asociadas varias publicaciones científicas en revistas y congresos internacionales, de las cuales el tesista es el autor principal. Por la tesis realizada, el autor fue reconocido como “Mejor Egresado de Postgrado” por la Universidad Nacional de La Plata y obtuvo la “Mención Cum Laude” en la Universidad de Castilla-La Mancha.

7. Referencias

Las referencias se encuentran en la tesis, publicada en http://sedici.unlp.edu.ar/handle/10915/74098.