Análisis de textos (vectoriales) o el arte de escribir resúmenes (automáticos)

Juan Manuel Torres Moreno
Laboratoire d'Informatique d'Avignon / Université d'Avignon
BP 1228, 84911 Avignon cedex 9, FRANCE.
torres@univ-avignon.fr

Resumen. Este artículo trata sobre la manera en como las computadoras pueden hacer un resúmen automático a partir de un texto. Un enfoque vectorial de representación de textos y un tratamiento algorítmico serán presentados.

Si usted ha comenzado a leer esta frase, probablemente está interesado en el contenido completo del resto del texto. Y este interés se ha despertado quizás, por la lectura de las palabras del resúmen (*abstract)*. Estas frases son a veces tan importantes que significan tomar la decisión de leer o no el texto completo. Como se habrá dado cuenta, este artículo trata sobre la manera en cómo las computadoras pueden hacer resúmenes automáticos a partir de un texto. Usted está quizás interesado también en temas tales que: ¿Cómo hacen resúmenes las personas? ¿Es tan dificil hacer resúmenes? ¿Puede automatizarse este proceso? ¿Pueden las computadoras hacer resúmenes?... o más simple: ¿Qué es un resúmen? Estas preguntas tienen probablemente más de una respuesta. El Procesamiento de Lenguaje Natural (PLN) tiene ciertas respuestas a estas preguntas y sobre todo, ha creado herramientas que hacen cada vez más posible el sueño (o la tentación de lo imposible) de crear máquinas capaces de procesar, abstraer y generar textos humanamente creíbles.

Cuando se habla de textos y de su procesamiento automático, casi de manera intuitiva viene a la mente el uso obligado de herramientas lingüisticas computacionales: el análisis sintáctico, semántico, los árboles de representación lexica, la manipulación simbólica y otras técnicas bien conocidas que resultan pertinentes al tratar pequeños textos. Pero uno de los problemas al que nos enfrentamos actualmente es la existencia de volumenes verdaderamente gigantescos de información disponible. Otro problema no menos importante lo constituye el babel mundial de lenguas en Internet: el depósito más grande y heterogéneo de información no estructurada jamás construído por el hombre. De seguir la tendencia actual, el inglés no representará más del 40% de los idiomas presentes en Internet. En otras palabras, habrá cada vez más documentos en español, alemán, francés, chino, etc. Y no debe olvidarse, por una parte, que las herramientas lingüísticas computacionales mencionadas son todas *dependientes* de cada idioma, y por otra, que una representación sintáctica de un texto medianamente grande es impráctica. Ambas limitaciones complican su utilización. ¿Qué hacer? La solución debe quizás buscarse en otro lado. ¿Dónde? Afortunadamente, gentes como Andrei Markov comenzaron a interesarse en el análisis de símbolos del idioma utilizando medidas de probabilidad y de transición entre los estados de un modelo matématico. Salton por otra parte (SAL83), propuso un modelo vectorial de representación de textos. La idea es seductora, pues en términos simples sugiere que es posible transformar un texto como el que usted está leyendo, en un espacio matématico. Será necesario mostrar algunos detalles en el ejemplo que se presentará. Imagínese que desea transformar un pequeño texto en un ente matématico. Tomemos como ejemplo el título de este artículo y una frase del texto, que vamos a concatenar para obtener: **"Análisis de textos (vectoriales) o el arte de escribir resúmenes (automáticos). Como se habrá dado cuenta, este artículo trata sobre la manera en cómo las computadoras pueden escribir un resúmen automático a partir de un texto"**, que contiene 41 símbolos o palabras. ¿Cómo representarlo en forma vectorial como lo indica el título? Varias fases deben tener lugar. Para que una representación matemática sea facilmente

manipulable, debería guardar únicamente la esencia del texto. Un estudio estadístico del idioma español (o de cualquier otro idioma) muestra la cantidad sorprendentemente grande de palabras que no aportan verdaderamente información, pero que sierven para vehicular las ideas y hacer fluida la lectura. ¿Son acaso indispensables para comprender un texto? Supongamos que podemos prescindir de ellas. Comenzar entonces por eliminar las palabras huecas de aquellas que son portadoras de información. ¿Cuáles son las palabras huecas? Los pronombres, artículos, conjunciones, símbolos de puntuación, números, expresiones ("**darse cuenta, por qué, cómo,...**"). Las palabras portadoras de información son por tanto los verbos, sustantivos y adjetivos. Un caso particular son los verbos *poder, ser, estar* y *tener,* que dado su carácter de auxiliares, serán considerados como palabras huecas. Así pues, el texto original se verá reducido a la secuencia siguiente: "**análisis textos vectoriales arte escribir resúmenes automáticos. artículo trata manera computadoras escribir resúmen automático partir texto**". Que es aún bastante comprensible y donde hemos reducido el número de palabras a 16. ¿De qué manera se podría reducir todavía más esta secuencia? Una mirada rápida indica que "**resúmen"** y "**resúmenes"** o "**texto"** y "**textos"** deberían contar como una sola palabra diferente. La lematización permite llevar las formas plural/femenino y/o verbos conjugados al infinitivo. Lematizando la secuencia anterior obtendremos: **"analizar texto vector arte escribir resúmen automático. artículo tratar manera computadora escribir resúmen automático partir texto"**. Con un ligero esfuerzo, todavía resulta comprensible. Ahora salta a la vista que hay palabras repetidas o que representan conceptos más generales. Por ejemplo, "**artículo**" y "**resúmen**" podrían estar contenidos en el concepto "**texto**". Una detección de sinónimos o conceptos dará lugar entonces a la secuencia: **"analizar texto vector arte escribir texto automático. texto tratar manera computadora escribir texto automático partir texto"**. Si contamos las palabras diferentes (léxico) en esta última secuencia, encontraremos 10. Dado el número enorme de palabras de un idioma, y de sus variaciones, esta reducción de lexico es esencial, pues permite reducir la dimensión en la que seran representados los textos. Nos queda la fase final para convertir todo a números. Podemos enumerar las palabras del léxico en su orden de aparición: **analizar=1, texto=2, vector=3, arte=4, escribir=5, automático=6, tratar=7, manera=8, computadora=9, partir=10.** Adicionalmente puede verse que la palabra **texto** tiene una frecuencia de 5, **automático** y **escribir** 2 y el resto del léxico frecuencias de 1. Ahora daremos un paso definitivo: asignemos un 1 si la palabra está presente en la frase o un 0 si no lo está, es decir: una representación binaria. Una representación en tabla o matriz es necesaria, pues tenemos dos frases (o vectores) a representar:

Frase / Palabra	analizar	texto	vector	arte	escribir	automático	tratar	manera	computadora	partir	partir
analizar texto vector arte escribir texto automático	1	1	1	1	1	1	0	0	0	0	0
texto tratar manera computadora hacer texto automático partir texto	0	1	0	1	0	1	0	1	1	1	1

El texto original está ahora representado por los vectores binarios **a=(1,1,1,1,1,1,0,0,0,0)** y **b=(0,1,0,0,0,1,0,1,1,1)** contenidos en un espacio o hipercubo binario de 10 dimensiones que resulta absolutamente incomprensible para una persona. Su carácter textual se ha

transformado.¿Y qué se ha ganado con todo ello? Pues habremos dado un paso formidable, ya que ahora el texto estará representado por medio de vectores, entidades matématicas a las que se les pueden aplicar todos los tratamientos matématicos permitidos. Se pueden sumar, multiplicar, normalisar, comparar,... Por ejemplo, se puede calcular el ángulo entre los vectores **a** y **b** por una operación simple: $\cos \alpha = \mathbf{a} \cdot \mathbf{b} / |\mathbf{a}| |\mathbf{b}|$; siendo $\cos \alpha$ el coseno del ángulo entre **a** y **b.** Esto daría un número entre 0 y 1. Si los vectores están "próximos" entre sí un valor cercano a 0 será obtenido, si no, estará cercano a 1. ¿Y qué relación tiene un ángulo vectorial con las frases originales? Pues simplemente que dos vectores cercanos en el *espacio vectorial* representan dos frases semánticamente próximas entre sí en el *espacio textual*. En particular, en nuestro ejemplo**, cos(a,b)=0.50**, lo que quiere decir que el titulo **a** y la frase **b** están relativamente próximas. Por otra parte, una nueva frase como **c = "Tratado del cómo y por qué del arte por computadora"** estaría representada por el vector **c=(0,0,0,1,0,0,1,0,1,0)** El coseno del ángulo entre el título **a** y la frase **c** sería sólo de **cos(a,c)=0.23**, lo que indica que *a pesar* de tener palabras comunes, están más alejadas entre si. Una representación vectorial en el espacio de dimensiones definido por el léxico de un texto lleva implícita la representación semántica de dicho texto.

Un sistema real de resúmenes automáticos: matemáticas de textos

Ahora sabemos que los textos pueden ser matematizados. Queda ver cómo obtener resultados interesantes a partir de ese proceso. Cuando se habla de escribir el resúmen de un documento, todo mundo está de acuerdo en que es necesario un esfuerzo cognitivo para obtenerlo (HUO00). Y puede no resultar un esfuerzo pequeño, pues la comprensión del texto en cuestión resulta indispensable. ¿Qué es entonces el resúmen de un texto? Según el estándar ANSI, un *abstract* o resúmen de un documento es la forma más concreta y reconocida de los *condensados* de textos (ANS79). Como la producción de verdaderos resúmenes es un proceso muy difícil de realizar con la tecnología disponible, hemos adoptado un método más simple, llamado *extracción* de frases. Esencialmente la extracción busca producir una versión reducida de un texto seleccionando las frases más relevantes del documento original, que serán únicamente juxtapuestas sin mayor modificación. El resultado final es un condensado del texto. En el *Laboratoire d'Informatique d'Avignon* (LIA) (Francia), el sistema **Cortex** (COndensación y Resúmenes de TEXtos) continúa produciendo resultados a partir de un proyecto iniciado en 2000 en Québec (Canadá). La idea es simple: retener las frases que aportan el máximo de información de un texto y eliminar el resto. Cortex es un sistema compuesto de módulos de tratamiento automático de la lengua que operan en un espacio vectorial. Esto significa que Cortex es capaz de transformar un texto en forma vectorial, después aplica operaciones en ese espacio para comprimirlo, que representan el condensado del texto en el espacio textual. Primero, un procesamiento inteligente es aplicado: filtrado de símbolos de puntuación, lematización, identificación del léxico y generación de la representación vectorial. A partir de los vectores contenidos en la matriz, diversas métricas detectan la proximidad semántica de las frases, su *peso* relativo o contenido informacional, sus relaciones o grado de interacción a través de palabras comunes con el resto del documento, sus distancias, la entropía,... que calculan la pertinencia de cada frase (vector). Las métricas son consideradas como votantes estadísticamente independientes. Los dictamenes de cada votante son combinados por un algoritmo de voto que entrega una probabilidad de retener o no un vector (una frase). Las frases retenidas son ordenadas y producen el condensado del texto original. Retomemos el pequeño ejemplo de las frases **"Como se habrá dado cuenta, este artículo trata sobre la manera en cómo las computadoras pueden hacer un resúmen automático a partir de un texto."** y **"Tratado del cómo y por qué del arte por computadora",** representadas respectivamente por los vectores **b=(0,1,0,0,0,1,0,1,1,1)** y **c=(0,0,0,1,0,0,1,0,1,0)**. Es evidente para un lector humano

que la primera es relevante y la segunda no lo es *en el contexto definido* por el título. El sistema debe comprender claramente esta situación. Para simplificar un ejemplo del proceso de extracción, consideremos las métricas *peso de la frase* (número de 1's en el vector) y *ángulo* respecto al título (vector **a).** De esta forma, tendremos **peso(b) = 6; peso(c) = 3.** Los ángulos ya habían sido calculados: **cos(a,b)=0.50; cos(a,c)=0.23**. Un voto simple que multiplique el peso y el ángulo, daría una ponderación (o *score*) para cada frase: ***score*(b) =** 6 x 0.50 **= 3.0**; ***score*(c)** = 3 x 0.23 **= 0.71.** Claramente la frase **b** sería retenida y la frase **c** eliminada por este algoritmo. En la realidad, nueve métricas diferentes (incluyendo frecuencias de palabras que permite usar un hipercubo no binario) son utilizadas y el algoritmo de voto es más complicado. ¿Y qué hay respecto a la calidad de los condensados? Responder a esta pregunta no resulta fácil, pues no existen estándares precisos de medida de la calidad de un resúmen (producido por una persona o por una máquina). La medida directa, que consiste en que un jurado humano lea los resúmenes generados por un sistema y compararlo con aquellos escritos por las personas es bastante objetiva, pero dificil de llevar a cabo en una gran masa de documentos. Mas allá de corpus pequeños de unas cuantas decenas de páginas, la decisión del jurado sobre la pertinencia de los resúmenes resulta muy subjetiva. No obstante, realizamos estas medidas y el sistema Cortex produjo resúmenes de calidad comparable o superior a otros sistemas existentes, numéricos o lingüísticos (TOR01, TOR02, SAG00). Otra forma de medir la pertinencia de un resúmen consiste en usar una medida indirecta de su calidad. Una manera de hacerlo es acoplar la salida del sistema de resúmen automático a la entrada de otro sistema de tratamiento de textos y medir su influencia sobre este último. Un sistema interesante para realizar esta medida es un sistema de pregunta-respuesta genérico, como el sistema **LIA-QR** desarrollado en el LIA. Un corpus textual es presentado a LIA-QR para entrenamiento. Una vez el corpus analizado, sistema es capaz de responder correctamente a un número N_C de preguntas abiertas formuladas en lenguaje natural. Un resúmen (a un cierto porcentaje de entre 10% y 30% del tamaño original del corpus) producido por el sistema Cortex fue presentado a LIA-QR, y un número N_R del mismo conjunto de preguntas fue respondido. Nuestra hipótesis era: si el número N_R está suficientemente próximo a N_C, entonces los resúmenes habrán conservado el carácter informativo de las frases, eliminando las que son irrelevantes. Esto mostraría la calidad del condensado. Sorprendentemente, los resúmenes al 20% mostraron que N_R es superior a N_C; es decir, el número de respuestas correctas encontradas ¡fue superior al buscar en los resúmenes en lugar de buscar en el corpus completo! ¿Cómo es posible? La explicación reside en el hecho que el corpus es verdaderamente grande: alrededor de 20,000 artículos periodísticos (de *Le Monde*) de diversos temas que representan aproximadamente 60 Mbytes de texto bruto. Las preguntas son abiertas y las verdaderas respuestas a veces inexistentes. Comprobamos que el sistema LIA-QR (y en general cualquier sistema de búsqueda de información) se desempeña mejor en un volumen reducido de información que en uno inmenso. Cortex proporciona un volumen reducido y *pertinente* donde LIA-QR encuentra un mayor número de respuestas correctas. Por supuesto problemas de cohesión o consistencia pueden presentarse en los condensados producidos por Cortex, pero hay que considerar que el texto ha sido procesado llevando el análisis numérico al extremo: ningún tratamiento simbólico, semántico o lingüístico fue utilizado. Prácticamente ninguna herramienta dependiente del idioma es necesaria, lo que lo hace particularmente atractivo para tratar el babel de Internet. La independencia de la temática de los corpus es también otra ventaja importante en los resúmenes producidos por Cortex. Conviene destacar que en los proceso descritos, una comprensión del texto se realiza por la máquina usando mecanismos muy distintos a los mecanismos cognitivos de una persona. Un proceso completamente inhumano permite de comprender y condensar un texto. Un proceso que no tiene ninguna relación de cómo una persona escribe resúmenes, y que sin embargo, produce condensados pertinentes. Pero no

hace falta demostrar que un auto no tiene músculos para rendirse ante la evidencia de que se mueve.

Del condensado al resúmen: generación de texto

Un verdadero resúmen es mucho más que un condensado del texto. En un resúmen se reescriben algunas frases para eliminar redundancias o se fusionan otras aumentando la compresión y la consistencia. Tareas así necesitan de herramientas lingüísticas apropiadas. Hemos pensado en la idea de hacer trabajar un sistema de generación de texto sobre los condensados obtenidos por extracción. El resultado debe ser interesante, dado que los condensados contienen únicamente información pertinente. En un sistema hibrido numérico-simbólico, las técnicas algorítmicas vectoriales permitirían obtener los condensados y las técnicas simbólicas retrabajarían dichos condensados de manera fina. Esta es un área de investigación actualmente en estudio en nuestro laboratorio.

Una última reflexión: ¿Y qué hay de la literatura? ¿Qué futuro le espera con estos sistemas automáticos de análisis, abstracción y generación automática de textos? Siendo yo mismo escritor de textos literarios, no puedo eludir responder a esta pregunta que podría parecer inquietante, y que sin embargo, no debería serlo. Disfrutar un texto literario (narrativa o verso), implica un placer estético ante una creación. El placer estético puede no ser facilmente explicable (subjetivo, diferente en cada persona, relacionado con experiencias personales, emotivo, ¿cómo medirlo?...) y sin embargo es real. Un cuento o una novela pueden simplemente gustarnos y ser disfrutados sin necesidad de explicación alguna si son lo *suficientemente* buenos. Esto es generalizable a cualquier creación artística. ¿Qué importancia tiene entonces que el autor de una creación sea tal o cual persona? Una obra debe tener una independencia propia para ser disfrutada sin necesidad de ser explicada o justificada por alguien, ni siquiera por su autor. La fotografía no acabó con la pintura: simplemente la transformó. Desde esta perspectiva, si usted disfruta leyendo un buen texto ¿importa acaso que haya sido escrito por un humano?

La tentación de lo imposible está suficientemente próxima para dejarla escapar. Los sistemas de análisis, comprensión procesamiento y generación automatica de textos en lenguaje natural son un gran desafio, pero las herramientas matématicas permiten hacer esta sueño cada vez más real. Está cada vez menos lejos el momento en que usted leerá artículos como éste, que habrán sido producidos por un módulo automático de generación de textos, apoyado en un módulo de resúmen automatico conectado a Internet, que habrá analizado y comprendido la petición de un usuario: “Redacta por favor, un artículo sobre el Analisis de textos (vectoriales) o el arte de escribir resúmenes (automáticos)”.

Referencias

ANS79. American National Standards for Writing Abstracts. ANSI Inc., USA.

HUO00. Huot F. (2000). Copernic summarizer ou la tentation de l'impossible. Québec Micro, 6.12(12):61--64.

SAG00. Saggion H. and Lapalme G. (2000). Concept identification and presentation in the context of technical text summarization. In Automatic Summarization Workshop, pages 1--10, Seattle. ANLP/NAACL.

SAL83. Salton G. and McGill M. (1983). Introduction to Modern Information Retrieval. McGraw-Hill.

TOR02. Torres-Moreno, J.M, Velázquez-Morales, P. et Meunier, J.G. (2002). Condensés de textes par des méthodes numériques. In JADT 2002, 2:723-734, A. Morin & P. Sébillot éditeurs, IRISA/INRIA, France.

TOR01. Torres-Moreno, J.M, Velázquez-Morales, P. et Meunier, J.G. (2001). Cortex : un algorithme pour la condensation automatique des textes. In la cognition entre individu et société ARCo 2001. Coord. Hélène Paugam-Moissy, Vincent Nyckees, Josiane Caron-Pargue Lyon, Hermès Science, pages 365 vol. 2, ISC-Lyon, France.

Newsletter Lania. Vol. 45-46 (12), p 1-2 Otoño Invierno 2004, Mexico.