Digit·HVM revista digital d'humanitats

Sistemas de resumen automático de documentos

Salvador Climent Roca

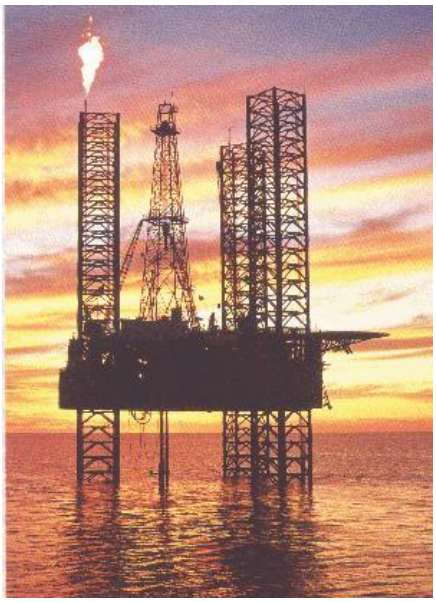
Profesor de los Estudios de Humanidades y Filología de la UOC
scliment@campus.uoc.es

Resumen: La aparición de la WWW y de Internet como red de telecomunicaciones, han cambiado el concepto de lo que se considera información: no está mejor informado quien más datos posee, sino quien dispone de mejores medios para obtener exclusivamente los datos que necesita y asimilarlos (digerirlos). Esta situación ha contribuido a impulsar la investigación y el desarrollo de aplicaciones en el campo de las tecnologías de recuperación y extracción de información. En este contexto, los sistemas de resumen automático de documentos suponen un nuevo paso hacia la optimización del tratamiento de la información en formato digital y su adaptación a las necesidades de los usuarios. Este artículo presenta las principales líneas de investigación en lo que respecta a la elaboración de resúmenes automáticos y sus relaciones con otras áreas de la ingeniería lingüística.

1. Introducción

Sin duda es innecesario describir aquí por enésima vez el problema que se plantea en la sociedad digital respecto a la obtención, la gestión y la digestión de la información. Todos hemos leído cifras mareantes sobre los miles de millones de páginas que contiene Internet, los centenares de millones de mensajes de correo electrónico que se envían anualmente o los miles de millones de gigabytes de información nueva que se genera en el mismo lapso de tiempo (podemos consultar por ejemplo la revista digital Infonomia).

Ante este panorama, normalmente la primera reacción consiste en un cierto deseo de tirar el ordenador por la ventana y marchar a una isla desierta. Pero aún siendo la más sensata, ésta solución no acostumbra a ser viable, por lo que, tras una breve reflexión, el siguiente paso suele ser buscar algún tipo de ayuda. “De entre todo este marasmo, ¿alguien o algo puede ayudarme a seleccionar únicamente aquello que me interesa?”.



Para llegar a hacer esta última selección, hay una cosa que nos sería tremendamente útil: que alguien o algo nos pudiera dar una idea un poco más precisa del contenido de los documentos. En pocas palabras, que nos respondiera a la pregunta: “¿este documento, de qué va?”.

Un amplio espectro de investigaciones en tecnologías de la información está dirigido a la solución de estos y parecidos problemas, muchos de los cuales implican en mayor o menor medida algún tipo de tratamiento del lenguaje humano ya que, como a nadie se le habrá escapado, la mayor parte de la información que circula por el mundo está codificada en alguna de estas curiosas lenguas naturales, que se nos antojan tan fáciles y transparentes cuando son precisamente la nuestra y tan complicadas cuando no lo son.

La obtención, el filtrado, la clasificación y la extracción de información, en sus dimensiones mono y multilingüe, son ámbitos prioritarios de la investigación en el campo de la tecnología lingüística. En este panorama, el resumen automático de documentos es posible que sea la guinda que coronará el pastel.

Nuestro buscador favorito será aquél al que le podamos pedir (en nuestra lengua): búscame documentos en cualquier idioma que hablen de tal asunto, selecciónname los más relevantes, clasifícalos de acuerdo con tal criterio, y dame de cada uno un resumen de cincuenta palabras (en mi lengua).

2. Analogías

El resumen automático de documentos comparte o puede compartir propiedades con la obtención y la extracción de información. De hecho las dos grandes líneas de investigación en el campo del resumen se corresponden por analogía con estas dos tecnologías.

Entendemos por obtención (o recuperación) de información lo que por ejemplo hacen los buscadores de Internet: de entre un conjunto (enorme) de documentos, obtener aquéllos que responden a unos determinados criterios; en general, ciertas palabras clave. Por otra parte, la extracción de información consiste en tratar uno o varios documentos para extraer de ellos una determinada información que nos interesa y generar a partir de ésta un nuevo documento (u otra estructura informativa, por ejemplo una plantilla) que contenga únicamente esa determinada información relevante. Las dos líneas básicas de investigación en la confección de resúmenes automáticos son análogas a éstas.

Los resúmenes por extracción actúan sobre uno (o varios) documentos, vistos como una colección de oraciones; y de estas oraciones extraen y presentan aquéllas consideradas más relevantes o que responden a unos determinados criterios. En este caso el resumen es un subconjunto de las oraciones del texto original.

Los resúmenes por abstracción utilizan técnicas más sofisticadas de tratamiento del lenguaje, ya que el resultado no consiste en determinadas oraciones entresacadas del texto original, sino en un documento de nueva redacción generado a partir del tratamiento de la información contenida el primero. Las técnicas necesarias para la aplicación de esta estrategia distan de haber obtenido resultados satisfactorios y pertenecen aún al campo de la investigación básica. Es por ello que, hasta el momento, los avances más significativos se hayan realizado en el campo de los resúmenes por extracción, y es a estos últimos a los que nos referiremos a partir de aquí.

En líneas generales, cuando hablamos de una tecnología tan relativamente nueva y prometedora, es preciso hacer una prevención para el no experto: no debe caerse en los errores de albergar falsas expectativas, como sucedió en los primeros tiempos de la traducción automática y en cierta forma sigue sucediendo con otras tecnologías como los sistemas de diálogo oral con máquinas. No debemos esperar que surjan rápidamente aplicaciones perfectas, que, por así decirlo, sustituyan a los agentes humanos. Mas bien debemos esperar aplicaciones realistas y operativas, que ayuden a los humanos y les liberen de algunas tareas redundantes y fatigosas.

Para exponerlo con ejemplos paralelos, del mismo modo que la traducción automática no nos proporciona textos directamente publicables sino documentos que deben ser invariablemente posteditados, o que sirven para hacerse una idea aproximada de su contenido, con los sistemas de resumen automático normalmente lo que podremos obtener serán extractos útiles para saber de qué va el documento original, posiblemente para clasificarlo en alguna estructura temática, y decidir si finalmente nos interesa o no acudir a la lectura del texto completo.

3. Tipologías

Cuando uno oye hablar por primera vez de un programa que resume documentos, algunas de las primeras preguntas que se le vienen a la cabeza son parecidas a las siguientes. (1) *"¿Y esos programas, cómo lo hacen? ¿De la misma manera que lo haríamos nosotros?". Y (2) "esos resúmenes automáticos... ¿son buenos? ¿Son tan buenos como los que pueda hacer una persona?"*.

Dejamos para después la respuesta a la primera pregunta, pero es pertinente avanzar que, como es norma habitual en la ingeniería lingüística y en general en la inteligencia artificial, los sistemas automáticos raramente actúan siguiendo las mismas estrategias que los humanos. Entre otras razones porque, en realidad, tampoco tenemos una idea muy clara de qué estrategias de razonamiento utilizamos. Ni tampoco los sistemas informáticos disponen ni por asomo del mismo tipo de información de qué disponemos las personas. En inteligencia artificial entendemos por sistema inteligente aquél que emula un comportamiento inteligente. O dicho en otras palabras, aquél que realiza una tarea que, si la hubiera realizado un humano, podríamos decir que ha requerido de un determinado grado de inteligencia.

Pero empecemos por la segunda de las preguntas. *Esos resúmenes automáticos, ¿son buenos?* Como muchas veces suele suceder, la primera respuesta es otra pregunta: *¿y qué es exactamente un buen resumen?* Dependerá de diversos factores, especialmente de qué sea lo que queramos encontrar en el resumen.

Puede que nos interese saber, de forma genérica, qué dice el autor en el texto; o quizá lo que queramos es saber qué dice el autor en el texto *respecto a algún tema en concreto*. Se trata de la diferencia entre "¿de qué habla este artículo?" y "¿qué dice este artículo sobre la Bolsa de Tokio?".

O quizá la situación es que ya tenemos información previa sobre el tema, y queremos enterarnos de qué hay de nuevo al respecto. En ese caso no queremos que nos cuenten toda la historia, nos basta con tener un resumen que contenga sólo los aspectos novedosos de la cuestión.

El tipo de resumen que queremos también puede tener que ver con la cantidad y calidad de la información que deseamos obtener. Puede que nos sea suficiente con saber más o menos de qué trata el texto para poder clasificarlo, o quizá lo que queramos es un nuevo documento que condense su contenido. Y en el caso de que nos baste con aquél resumen puramente indicativo, quizá queramos que nos sea ofrecido en forma de texto coherente, o a lo mejor será suficiente con una simple lista de fragmentos o palabras clave, o quizá prefiramos un indicador temático. En el segundo caso, dependiendo del volumen de documentos y tiempo que tengamos para leerlos, puede que queramos resúmenes muy concisos, por ejemplo, de cincuenta o cien palabras. O quizá vamos más holgados y podemos abordar resúmenes de hasta un par de folios.

Y ya en otra dimensión, quizá desearíamos que el sistema fuera lo suficientemente bueno como para ofrecernos un único resumen de un conjunto de documentos. Por ejemplo, de todas las noticias de todos los periódicos de hoy que tratan de la última crisis de gobierno.

Estos ejemplos corresponden a algunas de las diferentes tipologías y líneas de investigación actuales en este campo: resumen indicativo (para ser usado en clasificación y filtrado de documentos) frente a resumen informativo (para ser leído en sustitución o como avance del documento completo); resumen genérico frente a resumen guiado (en respuesta a una búsqueda o pregunta que define los intereses del lector); resumen genérico frente a resumen de actualización (que ofrezca únicamente los aspectos nuevos de una cuestión, obviando información ya conocida); resumen multidocumental (que condense en un documento varios textos que versan sobre una misma cuestión); resúmenes con

diferentes niveles de compresión del texto original, etc. Cada uno de estos géneros precisa del uso de técnicas diferentes de tratamiento del lenguaje.

Por otra parte, también debemos tener en cuenta que hay muchos tipos posibles de documentos a resumir: noticias, artículos, columnas de opinión, mensajes de correo, reportajes, *papers* científicos, webs, programas, listas... quizá hasta incluso libros.

Y en todos los casos el nivel de marcaje del texto original puede ser muy diverso, desde documentos en texto simple sin ningún tipo de marca estructural hasta documentos codificados al detalle en XML, pasando por páginas HTML o cualquier otro tipo de texto marcado de una u otra manera. El género y el nivel de estructuración del texto original son variables extremadamente importantes a tener en cuenta para abordar la tarea, de modo que muchas de las técnicas utilizadas en un caso concreto pueden no ser en absoluto reutilizables en otro.

Así que, en resumen, ¿qué es un buen resumen? Pues bien, ante todo deberemos plantearnos lo siguiente: un buen resumen, *de qué*, y *para qué*.

Pero también nos habíamos preguntado otra cosa: los resúmenes automáticos, ¿son tan buenos como los que pueda hacer una persona? Seguramente, en este caso es pertinente preguntarse lo siguiente: *como los que pueda hacer... ¿qué persona?*

Una de las tareas habituales y necesarias en cualquier tipo de investigación es la evaluación de los resultados. En el caso del resumen automático de textos la evaluación del sistema se suele realizar comparando los resúmenes obtenidos automáticamente con unos resúmenes tipo realizados por humanos, para lo cual se suele pedir a un conjunto de analistas que realicen resúmenes o extractos de un corpus de documentos de prueba.

Y el caso es que se ha demostrado que la propia obtención de un resumen tipo no es en absoluto una tarea trivial: dado un mismo documento, difícilmente existe acuerdo entre los analistas sobre cuál es la información que es pertinente seleccionar como resumen. Es más, en determinadas pruebas de evaluación se ha pedido a idénticos sujetos que resuman el contenido de idénticos documentos transcurrido un cierto tiempo entre la primera prueba y la segunda. Pues bien, el grado de acuerdo de los evaluadores consigo mismos resultó ser de tan sólo un 55%.

El problema se suele solucionar mediante métodos estadísticos, determinando de alguna manera un resumen tipo o promedio que pueda servir de banco de pruebas. Ésta parece ser una manera razonable de evaluar los sistemas de resumen automático. Sin embargo, sirvan simplemente estos comentarios para darnos cuenta que una pregunta como "¿los resúmenes automáticos, son tan buenos como los que pueda hacer una persona?", normalmente no suele tener una respuesta clara.

4. Técnicas

Y una vez expuestos los problemas, vayamos a por las soluciones. ¿Qué técnicas se utilizan para resumir automáticamente?

Habitualmente se clasifican en tres familias: técnicas basadas en la superficie del texto (no se realiza ningún tipo de análisis lingüístico); técnicas basadas en las entidades nombradas en el texto (se realiza algún tipo de reconocimiento y clasificación del léxico utilizado); y técnicas basadas en la estructura discursiva (requieren algún tipo de tratamiento estructural del documento, generalmente de tipo lingüístico).

Los tratamientos *superficiales* son los que se utilizan habitualmente en los productos comerciales. Abordan el texto simplemente como cadenas de caracteres —es decir, suele dar el caso de que tratan con letras pero también pueden ser números o cualquier otro tipo de símbolo— y realizan con ellos operaciones de cálculo para seleccionar algunos de ellos como representación del documento. Un método clásico es la selección de los términos estadísticamente frecuentes en el documento. P.e., seleccionamos como resumen del texto las oraciones (en realidad, aproximadamente, más que de oraciones debemos hablar de cadenas de caracteres que comienzan con un símbolo de mayúscula y acaban en un punto) que contienen el mayor número de términos más frecuentes en el documento —con diversas variantes o precisiones—.

Otro grupo de métodos se basa en la posición. Posición en el texto, en el párrafo, en la profundidad de la clasificación de secciones, etc. Se trata de seleccionar los fragmentos de texto que ocupan las posiciones que prometen ser más relevantes. Típicamente, para resumir noticias periodísticas, el mejor resumen corto puede ser elegir el primer párrafo de la noticia. En artículos científicos, es posible que sean especialmente relevantes el compendio inicial (o *abstract*), las conclusiones y la bibliografía.

Altres mètodes treuen profit de parts destacades del text: títols, subtítols, leads... Se suposa que les oracions que continguin les paraules del títol són millors candidates a resumir de forma adequada el document. O calculen la rellevància de les oracions recolzant-se en la presència de determinades cadenes bonificadores o penalitzadores (termes com "en conclusió", "és important", etc.). Otros métodos sacan provecho de partes destacadas del texto: títulos, subtítulos, *leads*. Se supone que las oraciones que contengan las palabras del título son mejores candidatas a resumir de forma adecuada el documento. O calculan la relevancia de las oraciones apoyándose en la presencia de determinadas cadenas *bonificadoras* o *penalizadoras* (términos como “en conclusión”, “es importante”...).

Una segunda familia de métodos, y entramos ya en el campo de la investigación más que el de los productos comerciales, es la basada en entidades. Aquí ya se da un mayor nivel de análisis lingüístico. Y digo “mayor” y no “algún” porque, aunque pueda no parecerlo, los métodos basados en la superficie normalmente sí requieren de algún tipo de tecnología lingüística: dividir un texto en oraciones no es una simple cuestión de mayúsculas y puntos, pues hay mayúsculas y puntos que no delimitan oración (piénsese p.e. en “N.A.T.O.”), y hay que aplicar un cierto conocimiento lingüístico para decidir qué caracteres (típicamente, puntos, pero también signos de admiración o interrogación)

delimitan o no una oración.

Los métodos basados en entidades parten de técnicas que permiten reconocer unidades lingüísticas de entre el marasmo de la cadena de caracteres alfanuméricos. Qué cadena separada por espacios en blanco es un nombre, cuál es un verbo... Para ello se precisa de analizadores morfológicos y desambiguadores léxicos, ya que una misma cadena puede ser nombre o verbo o pertenecer a otra categoría (pensemos en “casa”, como nombre o como forma del verbo “casar”). También es preciso detectar qué cadenas superficialmente diferentes pertenecen a un mismo concepto o categoría (pensemos en “era” y “son”, formas distintas del verbo “ser”, y también por cierto nombres comunes); para ello es necesario disponer de programas lematizadores. Y así hasta llegar a análisis más sofisticados de tipo sintáctico o semántico.

Este tipo de métodos pueden ser capaces de tomar las entidades (previamente analizadas lingüísticamente) y detectar diversos tipos de relaciones establecidas entre ellas: recurrencia de formas o lemas, relaciones semánticas (un coche, una moto y un autobús son vehículos), relaciones temáticas (un árbitro, un delantero centro y un penalty son términos del ámbito del fútbol), etc.

A partir de aquí es posible construir una representación de la conectividad del texto, en términos de grafos, de manera que se pueda determinar qué partes del texto (oraciones) son especialmente relevantes y por tanto son candidatas a formar parte del extracto o resumen.

Para implementar este tipo de métodos es especialmente importante contar con bases de conocimiento léxico (como EuroWordNet), reconocedores de entidades (p.e. de nombres propios, que por su propia naturaleza no forman parte de los diccionarios o repertorios léxicos) y sistemas de resolución de referencias anafóricas (p.e. anáforas pronominales).

Finalmente, los métodos basados en la estructura pueden sacar partido del propio armazón hipertextual de un documento HTML, o en otro ámbito de cosas pueden intentar, partiendo de la base de las técnicas y recursos de ingeniería lingüística antes apuntados y otros de tipo más sofisticado, construir una estructura de la argumentación contenida en el documento que permita detectar los fragmentos discursivamente más relevantes. Este último tipo de técnicas se basa en la detección de marcadores discursivos, básicamente conjunciones y adverbios, del tipo “en primer lugar”, “por el contrario”, “sin embargo”, “además”, etc. Cabe insistir en que este tipo de métodos pertenece aún al más estricto campo de la investigación.

Un aspecto importante a tener en cuenta para todos los métodos basados en extracción es el de recomposición del texto. Es evidente que al crear un documento entresacando oraciones del texto original es muy fácil que se produzcan disfunciones en las referencias y, en general, en la cohesión textual. Un buen sistema basado en extracción deberá ser capaz de, a partir de la comparación y el análisis del texto entresacado y el texto original, insertar material léxico que recomponga las cadenas de cohesión textual.

5. Referencias y productos

Si queremos conocer el estado actual de la cuestión en el campo del resumen de documentos deberemos atender por una parte a los sistemas comerciales existentes, y por otra al estado de la investigación.

En el primer caso podemos acudir a los resumidores de Inxight (Xerox), al más reciente de Copernic, o al típico y elemental sistema de autorresumen de Microsoft Word. Un listado bastante completo de sistemas actualmente accesibles lo podemos encontrar en la página de Mitre Org. En cuanto a sistemas en vías de desarrollo que operen online, podemos probar los muy interesantes SweSum y Extractor, que ofrecen resúmenes del español.

Por lo que respecta al estado de la investigación, son muy relevantes las páginas de la Universidad de Ottawa, la del investigador de la Universidad de ColumbiaDragomir Radev y, sobre todo, la

extensa bibliografía contenida en la página que mantiene el propio Radev en la Universidad de Michigan. En el Estado español, estamos construyendo un prototipo de sistema resumidor de noticias periodísticas en el marco del proyecto Hermes.

Pero los programas de mayor envergadura hay que localizarlos en el marco de las acciones impulsadas por el Departamento de Defensa de los EE.UU., DARPA, como el programa TIDES para la detección, extracción y resumen de información multilingüe y el congreso de comprensión de documentos, Document Understanting Conference, que propone una serie de experimentos a gran escala para la evaluación de técnicas y sistemas de resumen y comprensión de texto.

En todo caso, y para finalizar, es importante remarcar nuevamente que para superar los sistemas comerciales ya existentes y avanzar hacia sistemas relativamente*inteligentes* o poseedores de un cierto *conocimiento*, es crucial contar con una gran batería de recursos lingüísticos, tales como analizadores morfológicos, desambiguadores categoriales, bases de conocimiento léxico, analizadores sintácticos, reconocedores de entidades o sistemas de resolución de anáforas y otros tipos de coreferencia. Como suele suceder en la mayoría de tecnologías de la lengua, el crecimiento cualitativo es de tipo asintótico, de manera que es relativamente fácil alcanzar niveles de funcionamiento aceptables pero, de forma progresiva, cada vez es más costoso elevar los niveles de calidad de los sistemas.

A este respecto, es imprescindible el trabajo en equipos interdisciplinares, en los que lingüistas e ingenieros informáticos colaboren estrechamente. En el terreno de la tecnología lingüística, a diferencia de muchos otras áreas de investigación en lingüística y otras ciencias humanas y sociales, los esfuerzos individuales, por meritorios que resulten, raramente suelen conducir a resultados aceptables.

6. Apéndice

6.1. Resumen del artículo, de 150 palabras, realizado por Extractor.

Este resumen ha sido generado automáticamente por el sistema Extractor, y se presenta aquí sin ningún retoque.

- SISTEMAS DE RESUMEN AUTOMÁTICO DE **DOCUMENTOS Salvador Climent** (Universitat Oberta de Catalunya) scliment@campus.uoc.es Pero aún siendo la más sensata, ésta solución no acostumbra a ser viable, por lo que, tras una breve reflexión, el **siguiente paso suele** ser buscar algún tipo de ayuda.

- La obtención, el filtrado, la clasificación y la **extracción** de información, en sus dimensiones mono y multilingüe, son ámbitos prioritarios de la investigación en el campo de la **tecnología lingüística**.

- En este panorama, el **resumen automático** de documentos es posible que sea la guinda que coronará el pastel.

- Los resúmenes por abstracción utilizan técnicas más sofisticadas de tratamiento del lenguaje, ya que el resultado no consiste en **determinadas oraciones entresacadas del texto original**, sino en un documento de nueva redacción generado a partir del tratamiento de la información contenida el primero.

6.2. Resumen del artículo, de 300 palabras, realitzado por Microsoft Word

Este resumen ha sido generado automáticamente por el sistema Microsoft Word y se presenta aquí sin ningún retoque.

El resumen automático de documentos comparte o puede compartir propiedades con la

obtención y la extracción de información. Las dos líneas básicas de investigación en la confección de resúmenes automáticos son análogas a éstas. En este caso el resumen es un subconjunto de las oraciones del texto original. Y (2) "*esos resúmenes automáticos...*

Estos ejemplos corresponden a algunas de las diferentes tipologías y líneas de investigación actuales en este campo: resumen indicativo (para ser usado en clasificación y filtrado de documentos) frente a resumen informativo (para ser leído en sustitución o como avance del documento completo); resumen genérico frente a resumen guiado (en respuesta a una búsqueda o pregunta que define los intereses del lector); resumen genérico frente a resumen de actualización (que ofrezca únicamente los aspectos nuevos de una cuestión, obviando información ya conocida); resumen multidocumental (que condense en un documento varios textos que versan sobre una misma cuestión); resúmenes con diferentes niveles de compresión del texto original, etc.

Y en todos los casos el nivel de marcaje del texto original puede ser muy diverso, desde documentos en texto simple sin ningún tipo de marca estructural hasta documentos codificados al detalle en XML, pasando por páginas HTML o cualquier otro tipo de texto marcado de una u otra manera. Ésta parece ser una manera razonable de evaluar los sistemas de resumen automático. ¿Qué técnicas se utilizan para resumir automáticamente?

Habitualmente se clasifican en tres familias: técnicas basadas en la superficie del texto (no se realiza ningún tipo de análisis lingüístico); técnicas basadas en las entidades nombradas en el texto (se realiza algún tipo de reconocimiento y clasificación del léxico utilizado); y técnicas basadas en la estructura discursiva (requieren algún tipo de tratamiento estructural del documento, generalmente de tipo lingüístico).como el programa TIDES para la detección, extracción y resumen de información multilingüe y el congreso de comprensión de documentos, Document Understanting Conference, que propone una serie de experimentos a gran escala para la evaluación de técnicas y sistemas de resumen y comprensión de texto.

Enlaces relacionados:

- DARPA Information Technology Office: Translingual Information, Detection, Extraction and Summarization (TIDES): http://www.darpa.mil/ito/research/tides/
- The Text Summarization Project. Universitat de Ottawa: http://www.site.uottawa.ca/tanka/ts.html
- A Bibliography of Research in Text Summarization: http://www.si.umich.edu/~radev/summarization/large-bib.doc
- SweSum: http://www.nada.kth.se/~xmartin/swesum/index-eng.html
- Extractor: http://extractor.iit.nrc.ca/on_line_demo.html

[Fecha de publicación: septiembre de 2001]

Cita recomendada:

CLIMENT ROCA, Salvador (2001). "Sistemas de resumen automático de documentos". *Digithum*, n.º 3 [artículo en línea]. DOI: http://dx.doi.org/10.7238/d.v0i3.598