(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 716 634**

(21) Número de solicitud: 201831222

(51) Int. CI.:

G06N 3/04 (2006.01)

(12) PATENTE DE INVENCIÓN CON EXAMEN B2

(22) Fecha de presentación:

14.12.2018

(43) Fecha de publicación de la solicitud:

13.06.2019

Fecha de modificación de las reivindicaciones:

27.07.2020

Fecha de concesión:

19.11.2020

(45) Fecha de publicación de la concesión:

26.11.2020

(73) Titular/es:

UNIVERSIDAD DE LEÓN (100.0%)
Avenida de La Facultad, 25
24071 León (León) ES

(72) Inventor/es:

JOSHI , Akanksha ;
FIDALGO FERNÁNDEZ, Eduardo;
ALEGRE GUTIÉRREZ , Enrique y
FERNÁNDEZ ROBLES , Laura

(74) Agente/Representante:

CARVAJAL Y URQUIJO, Isabel

(54) Título: **PROCEDIMIENTO Y SISTEMA DE GENERACIÓN DE RESÚMENES DE TEXTO EXTRACTIVOS UTILIZANDO APRENDIZAJE PROFUNDO NO SUPERVISADO Y AUTOCODIFICADORES**

(57) Resumen:

Procedimiento y sistema para la generación de resúmenes de texto extractivo utilizando aprendizaje profundo no supervisado y autocodificadores.

Se describe un procedimiento y sistema automatizado para realizar resúmenes de texto extractivos utilizando aprendizaje profundo no supervisado y autocodificadores. Dicho procedimiento hace uso del aprendizaje automático profundo para realizar la codificación del texto contenido en los documentos a través de técnicas de incrustación de frases y su posterior codificación en una representación vectorial de menor dimensión utilizando una red profunda de un autocodificador. Del texto original, las frases incrustadas resultantes y de la representación vectorial de menor dimensión se calculan una medida de relevancia, una medida de novedad y una medida de posición por frase respectivamente. A partir de estas tres medidas se realiza una ordenación y selección de frases según su puntuación final o frecuencia de aparición en el documento original, las cuales formarán parte del documento final resumido.

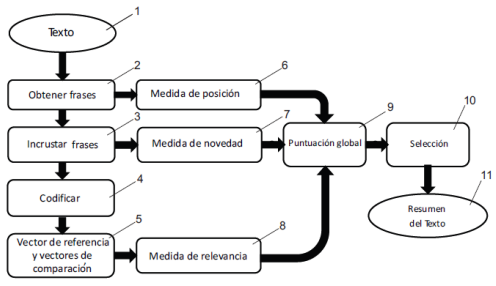

Fig. 1

ES 2 716 634 B2

DESCRIPCIÓN

PROCEDIMIENTO Y SISTEMA DE GENERACIÓN DE RESÚMENES DE TEXTO EXTRACTIVOS UTILIZANDO APRENDIZAJE PROFUNDO NO SUPERVISADO Y AUTOCODIFICADORES

OBJETO DE LA INVENCIÓN

El objeto de la presente invención es un procedimiento y sistema automatizado para realizar resúmenes de texto extractivos utilizando aprendizaje profundo no supervisado y autocodificadores. La invención permite resumir un documento de un modo extractivo, es decir, seleccionar los fragmentos más relevantes del documento y formar un documento de menor tamaño y que permita identificar el contenido textual del mismo. Dicho documento de menor tamaño permitiría a un usuario conocer la temática o contenido de un documento de texto extenso sin efectuar una lectura completa del documento.

ANTECEDENTES DE LA INVENCIÓN

Con la llegada de Internet y la gran cantidad de datos disponibles, el número de textos y documentos con contenido textual ha experimentado un aumento notable. Para poder gestionar esta la información contenida en dichos documentos, surge la necesidad de buscar una representación más pequeña de los mismos que recoja la información fundamental, es decir, un resumen. El resumen de textos automático es una rama importante del procesamiento del lenguaje natural que pretende representar los documentos de texto largo en una forma comprimida para que la información más relevante pueda ser comprendida e identificada rápidamente por los usuarios finales.

Se distinguen dos tipos de resúmenes de textos, resumen de texto extractivo y resumen de texto abstractivo (Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. Artificial Intelligence Review, 47(1)). El resumen de texto extractivo concatena las oraciones más relevantes del documento para producir el resumen. Como alternativa al resumen extractivo, se puede realizar un resumen abstractivo, donde no se utilizan frases exactas del propio documento, sino que se genera un resumen parafraseando los contenidos principales del documento usando técnicas de generación de lenguaje natural.

Existen técnicas tradicionales de resúmenes de textos, que se basan en la combinación de características estadísticas y lingüísticas, como la frecuencia de los términos (Luhn, H. P.

(1958). The automatic creation of literature abstracts. IBM Journal of Research Development.; Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. Technical report, Microsoft Research) o la longitud y posición de la oración entre otros. En estos métodos, se asigna un puntaje a cada oración en función de sus características. A continuación, dichas oraciones se eligen para formar parte del resumen final utilizando enfoques basados en gráficos (Radev, D., and Erkan, G. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research 457–479.) o enfoques basados en la optimización (Mcdonald, R. (2007). A Study of Global Inference Algorithms in Multi-Document Summarization. In Proceedings of the 29th European conference on IR research, 557–564) entre otros.

En la actualidad, las técnicas de resúmenes de textos han evolucionado al uso de algoritmos de aprendizaje profundo, dada la potencia y los buenos resultados de los mismos en múltiples problemas de Procesamiento de Lenguaje Natural (Natural Language Processing - NLP). A pesar de ello, existe una necesidad de tener grandes cantidades de datos para obtener un entrenamiento adecuado de la red, lo que supone un inconveniente en el uso de aprendizaje profundo supervisado para la generación resúmenes de documentos de texto.

La presente invención soluciona los problemas que presentan los métodos de la técnica anterior tales como, por ejemplo, la necesidad de grandes cantidades de documentos para el entrenamiento de los algoritmos, mediante la explotación de técnicas que no requieren datos etiquetados para el entrenamiento, especialmente el enfoque de aprendizaje profundo no supervisado basado en autocodificadores y las incrustaciones de oraciones, a través de redes de aprendizaje profundo entrenadas previamente utilizando un conjunto de datos predefinido.

La obtención de grandes cantidades de datos para entrenamiento de un algoritmo de aprendizaje profundo para resumir documentos de texto presenta una serie de inconvenientes. En primer lugar, es necesario disponer de un elevado número de documentos resumidos de un modo extractivo y manualmente por una persona. Segundo, es habitual en conjuntos de datos que contienen resúmenes de texto que cada documento original tenga asociado varios resúmenes, cada uno realizado por un operador humano. Además, el resumen de un documento depende en gran medida de la persona que lo realiza, aportando subjetividad, que genera una disparidad del contenido entre los diferentes resúmenes, que serán los utilizados para entrenar el modelo. Por último, es un proceso costoso por los elevados costes asociados al tiempo de la persona que realiza los resúmenes.

Debido a los anteriores problemas para disponer de datos necesarios para entrenar un modelo de resúmenes de texto automáticos utilizando aprendizaje profundo supervisado, se recurre a la realización de resúmenes de texto automático utilizando aprendizaje profundo no supervisado.

Son conocidas diversas aplicaciones de Procesamiento de Lenguaje Natural que pretenden mejorar la tarea de resumen de texto explotando las capacidades del aprendizaje automático profundo (Rush, A. M., Chopra, S., & Weston, J. (2015). A Neural Attention Model for Abstractive Sentence Summarization, (September), In Proceedings of Empirical Methods on Natural Language Processing, 379–389; Nallapati, R., Zhou, B., Santos, C. N. dos, Gulcehre, C., & Xiang, B. (2016). Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. Proceedings of The SIGNLL Conference on Computational Natural Language Learning, 280–290; Nallapati, R., Zhou, B., Santos, C. N. dos, Gulcehre, C., & Xiang, B. (2016). Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. Proceedings of The SIGNLL Conference on Computational Natural Language Learning, 280–290.).

DESCRIPCIÓN DE LA INVENCIÓN

El objeto de la presente invención es un procedimiento y sistema automatizado para realizar resúmenes de texto extractivos utilizando aprendizaje profundo no supervisado y autocodificadores.

Los autocodificadores se han aplicado anteriormente para la realización de resúmenes de texto en documentos únicos, pero en la presente invención, dichos autocodificadores se entrenan representando el documento de texto de entrada usando vectores de Termino de Frecuencia – Frecuencia Inversa de Documento (TF-IDF), que ignoran por completo el orden de las palabras de resumen de texto genérico. Una de las principales ventajas de utilizar autocodificadores es que se puede aprender, en este caso cuál es el conjunto de sentencias que mejor resumen el documento, de forma no supervisada.

El procedimiento y sistema automatizado para realizar resúmenes de texto extractivos utilizando aprendizaje profundo no supervisado y, preferentemente, autocodificadores según la presente invención permite realizar el resumen extractivo de un documento de texto, tanto

si es obtenido de la red a través de una conexión a internet, como transferido a un ordenador a través de un dispositivo extraíble de medios o de almacenamiento masivo.

El resumen automático de documentos, frente al resumen manual por un experto anula la subjetividad, los errores por cansancio y falta de atención, la disparidad de criterio entre expertos, los costes asociados al tiempo del experto y disminuye el tiempo necesario para la realización del resumen. Por este motivo, este procedimiento puede ser implementado en herramientas utilizadas por empresas y FFCCSSEE (Fuerzas y Cuerpos de Seguridad del Estado) para realizar resúmenes de cualquier tipo de documento con contenido textual conectado a la red, o de manera aislada, accediendo a los documentos a través de medios extraíbles de almacenamiento masivo.

La presente invención puede ser también aplicada a la generación de conjuntos de datos de una manera no supervisada, es decir, resúmenes de documentos de texto que se podrían utilizar posteriormente en el entrenamiento de algoritmos de aprendizaje supervisado y profundo. La disposición de grandes conjuntos de resúmenes de documentos permitiría el entrenamiento de sistemas de resúmenes de documentación más robustos y fiables, que permitirían la obtención de resúmenes más precisos de documentos.

En un ejemplo de realización, el sistema realiza una fase de incrustación de frases, cuya salida se puede utilizar para entrenar un codificador que convierta dichas incrustaciones en vectores incrustados, por ejemplo, mediante la metodología “Skip-Thoughts” (Kiros, M, et al, Skip-thought vectors, arXiv:1506.06726v1, June 22nd, 2015). Esto permite mapear las frases que son semánticamente y sintácticamente similares en representaciones de vectores similares. Dada una frase cualquiera, el vector representativo de la misma se construye usando las frases cercanas a la primera, debido a que se considera que suministran una gran información semántica y contextual. La representación de las frases en el espacio de incrustación hace que las frases con un significado similar estén representadas por vectores similares.

Dado que el problema de generación resúmenes automáticos se puede considerar como un problema de ordenación o selección de frases. La presente invención contempla un procedimiento para generar el resumen de un documento de texto obtenido, por ejemplo, de internet que comprende las etapas de:

- obtención del documento de texto mediante un procesador
- obtención de una serie de frases a partir del documento de texto;

- codificar la serie de frases mediante autocodificadores, obteniendo una serie de frases codificadas;
- asignar una medición de relevancia a cada una de las frases codificadas;
- asignar una medición de novedad a cada una de las frases codificadas;
- asignar una medición de posición de cada una de las frases codificadas;
- a partir de una combinación de las medidas de relevancia, novedad y posición, asignar una puntuación global a cada una de las frases codificadas;
- seleccionar las frases a disponer en el resumen a partir de la puntuación global de las frases codificadas;

En un ejemplo de realización, la obtención de una serie de frases a partir del documento de texto se realiza mediante un modelo construido mediante un algoritmo no supervisado.

Por otra parte, las frases codificadas pueden corresponder, por ejemplo, a una serie de vectores incrustados que se obtiene utilizando redes neuronales recurrentes. Preferentemente, la codificación de las frases se realiza mediante la metodología Skip-Thought.

En una realización preferente, el método comprende obtener una representación latente original del documento mediante la concatenación de las frases codificadas.

En cuando a la medida de relevancia de cada frase, dicha medida se puede obtener mediante varios métodos, por ejemplo, en base a la medida de similitud coseno existente entre una representación latente original del documento de texto y una representación latente modificada del documento de texto, siendo la representación latente modificada obtenida mediante la eliminación de la frase de la que se quiere obtener su relevancia.

Por otra parte, la medición de novedad se puede realizar, preferentemente, en base a calcular la similitud coseno de la serie de vectores incrustados obteniendo un valor intermedio de similitud y, en función del valor intermedio de similitud, asignar la medición de novedad. En una realización, el valor intermedio de similitud se calcula a partir del valor máximo de similitud coseno entre los vectores incrustados. En otra realización, la medición de novedad es 1 si el valor intermedio es inferior a un valor umbral predeterminado. En definitiva, la medición de novedad se puede definir como que es igual a 1-V, donde V es el valor intermedio si el valor intermedio es superior al valor umbral.

La medida de posición de cada frase se puede realizar, por ejemplo, teniendo en cuenta la posición de la frase dentro del documento de texto, así como el número de frases del documento de texto.

Preferentemente, la medición de relevancia comprende: generar un vector de referencia basado en la serie de frases, generar un vector de comparación de cada frase en el que el vector de comparación de cada frase corresponde al vector de referencia eliminando las partes del vector de referencia que corresponden a la frase y calcular la medición de relevancia en función de un cálculo de similitud coseno entre el vector de referencia y cada vector de comparación. Más preferentemente, el vector de referencia se obtiene a partir de la adición de elementos de los vectores incrustados, en concreto, el vector de referencia se puede obtener a partir de un autocodificador entrenado con la serie de vectores incrustados.

En una realización particular, la selección de las frases a disponer en el resumen comprende: organizar las frases en función de la puntuación global y seleccionar las frases que están por encima de una puntuación umbral predeterminada. Preferentemente, la selección de las frases a disponer en el resumen comprende: organizar las frases en función de la puntuación global y seleccionar las primeras X frases, siendo X un valor predeterminado de frases.

En una realización de la presente invención, la obtención del documento de texto se realiza a partir de un medio de almacenamiento externo seleccionado de entre: una memoria ROM, una memoria CD ROM o una memoria ROM de semiconductor, una memoria flash USB, SD, mini-SD o micro-SD, un soporte de grabación magnética, un disco duro o una memoria de estado sólido.

Además, la presente invención da a conocer un sistema de generación de un resumen a partir de un documento de texto que comprende medios de acceso a un documento de texto y un procesador configurado para:

- obtener del documento de texto mediante un procesador
- obtener de una serie de frases a partir del documento de texto;
- asignar una medición de novedad a cada una de las frases;
- asignar una medición de relevancia a cada una de las frases;
- asignar una medición de posición a cada una de las frases;
- a partir de las mediciones de novedad, relevancia y posición, asignar una

puntuación global a cada una de las frases; y

- seleccionar las frases a disponer en el resumen a partir de la puntuación global de las frases;

en el que la medición de novedad comprende codificar, mediante el procesador, las frases para obtener una serie de vectores incrustados; calcular la similitud coseno de la serie de vectores incrustados obteniendo un valor intermedio de similitud y, en función del valor intermedio de similitud, asignar la medición de novedad.

Preferentemente, la codificación de las frases para obtener la serie de vectores incrustados se realiza mediante la metodología Skip-Thought.

Además, el procesador puede estar configurado, por ejemplo, para:

- asignar una medición de relevancia a cada una de las frases; y
- asignar la puntuación global en función de la medición de relevancia

en el que la medición de relevancia comprende: generar un vector de referencia basado en la serie de frases, generar un vector de comparación de cada frase en el que el vector de comparación de cada frase corresponde al vector de referencia eliminando las partes del vector de referencia que corresponden a la frase y calcular la medición de relevancia en función de un cálculo de similitud coseno entre el vector de referencia y cada vector de comparación.

El procesador puede estar preferentemente configurado para:

- asignar una medición de posición; y
- asignar la posición global en función de la medición de posición;

en el que la medición de posición se calcula en función de la posición relativa de la frase respecto al documento.

Además, la presente invención contempla un producto de programa que comprende medios de instrucciones de programa para llevar a cabo los procedimientos anteriormente descritos cuando el programa se ejecuta en un procesador y, de igual manera, contempla un producto de programa almacenado en un medio de soporte de programas.

En una realización especialmente preferente, el procedimiento de la invención realiza el cálculo de la (i) medida de posición de la frase respecto al texto, (ii) medida de novedad de la frase en función de la detección de similitud entre vectores incrustados y (iii) medida de

relevancia de la frase. Dichas medidas se pueden combinan para dar una puntuación final a cada frase utilizando una fusión ponderada de dichas medidas.

Una vez se han obtenido todas las puntuaciones de todas las frases del documento de entrada, el procedimiento de la invención selecciona las frases con las mayores puntuaciones para representar el resumen del documento. Dicha selección se puede realizar de dos modos diferentes, (a) ordenando las frases del resumen en orden descendente con respecto a sus puntuaciones relativas y seleccionando las primeras frases hasta llegar a un número de frases preestablecido, (b) seleccionando todas las frases cuya puntuación global esté por encima de un umbral predeterminado u (c) ordenando las frases en base a su frecuencia de aparición en el documento de entrada.

En una realización preferente de la invención este procedimiento se aplica a cualquier tipo de documento textual, tanto descargado de la Web, como suministrado al sistema a través de un dispositivo de almacenamiento externo de cualquier tipo.

Un ejemplo del procedimiento y sistema automatizado para realizar resúmenes de texto extractivos utilizando aprendizaje profundo no supervisado y autocodificadores de la presente invención comprende las siguientes etapas:

1. Obtención de documento de texto. Esta obtención se puede realizar de un modo en línea, a través de un ordenador con conexión a internet, o en un modo sin línea, obteniendo el documento de texto a través de un dispositivo de almacenamiento externo.

2. Incrustación de las frases del documento de texto: Consiste en realizar la incrustación de las frases del documento de texto. Es decir, las frases del documento de entrada se transforman en una serie de vectores incrustados de longitud fija, de modo que frases con significados parecidos van a tener representaciones vectoriales similares, y viceversa, frases con diferentes significados van a tener representaciones vectoriales diferentes. En una realización preferente de la invención, se utiliza la metodología vectorial “skip-thought” para realizar esta incrustación.

3. Codificación de las frases incrustadas: En una realización preferente de la invención, se realiza un proceso de codificación de las frases incrustadas, convirtiéndose los vectores incrustados en una representación vectorial de menor dimensión denominada vector de

referencia. En una realización preferente, se diseña una red de autocodificadores para obtener el vector de referencia alimentándola con los vectores incrustados resultantes de la anterior fase. Dichos vectores incrustados se combinan en unidades textuales incrustadas que se utilizan para entrenar una red de autocodificadores. En una realización preferente de la invención, una vez entrenada la red, se utiliza solo su parte de codificador para generar representaciones de unidades textuales, cuya combinación dará lugar a la representación latente original del documento.

4. Cálculo de la medida de la relevancia de la frase: una vez obtenido el vector de referencia, correspondiente a la representación latente original del documento, se calculan una serie de vectores de comparación que serían representaciones latentes modificadas del documento, uno por cada frase contenida en el documento. Para ello, se elimina del vector de referencia la información correspondiente a una frase del documento generando así el vector comparación correspondiente a dicha frase, es decir, su representación latente modificada. Entonces, para calcular la medida de la relevancia de dicha frase, se calcula la similitud coseno entre la representación latente original (el vector de referencia) y la representación latente modificada (el vector de comparación). En una realización preferente, la medida de la relevancia toma valores entre 0 y 1, siendo la frase más relevante cuanto mayor sea el valor de esta medida próximo a uno.

5. Cálculo de la medida de la novedad de la frase: para realizar el cálculo de la medida de la novedad de una frase se calcula la similitud coseno entre los vectores incrustados correspondientes a dos frases. En una realización preferente, el valor resultante estará entre 0 y 1, siendo la frase más novedosa cuanto mayor sea el valor de esta medida próximo a uno.

6. Cálculo de la medida de posición de la frase: para realizar el cálculo de la medida de la posición de una frase con respecto a un documento, se tiene en cuenta la posición que ocupa dicha frase dentro del documento original, así como el número de frases del mismo. En una realización preferente, el valor resultante estará entre 1 y 0.5, siendo el valor de la posición de la primera frase 1 y decreciendo dicho valor en frases sucesivas.

6. Calculo de la puntuación final de cada frase: para realizar la ordenación de las frases de un documento según los valores de las medidas de relevancia, novedad y posición, se realiza el cálculo de la puntuación final de cada frase del documento original. En una realización preferente de la invención, dicho valor resulta de la suma ponderada de las

medidas de relevancia, novedad y posición.

7. Selección de frases que formarán el resumen final del documento: una vez calculada la puntuación final de cada frase del documento, se procede a la selección de las frases que constituirán el resumen del documento original. En una realización preferente, se puede realizar esta selección de dos modos diferentes: (i) realizando una ordenación de las frases según la puntuación final de cada frase por orden descendente y eligiendo un número específico de frases con la mayor puntuación final o (ii) realizando una ordenación de las frases según su frecuencia de aparición en el documento original.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

A continuación, se describen una serie de figuras que ayudan a comprender mejor la invención y que se relacionan expresamente con una realización de dicha invención que se presenta como un ejemplo no limitativo de ésta.

La Fig. 1 muestra un esquema simplificado de un sistema configurado para llevar a cabo el procedimiento de la invención.

La Fig. 2 muestra un ejemplo de la conversión de las frases $\mathcal{S}$ de un documento $\mathcal{D}$ a frases incrustadas $\vec{\mathcal{S}}$ a través de espacio vectorial “skip-thoughts”.

La Fig. 3 muestra un ejemplo de la conversión de las frases incrustadas $\vec{\mathcal{S}}$ en unidades textuales incrustadas $\vec{\mathcal{T}}$, posteriormente reducidas a unidades textuales latentes $\hat{\mathcal{T}}$, las cuales se unirán en la representación latente del documento $\hat{\mathcal{D}}$.

REALIZACIÓN PREFERENTE DE LA INVENCIÓN

Se describe a continuación un ejemplo de procedimiento de acuerdo con la invención, haciendo referencia a las figuras adjuntas. La **Figura 1** muestra un esquema simplificado de un ejemplo de sistema de generación de resúmenes automático de un texto (1) dispuesto en un documento. Dicho sistema pude implementarse en un ordenador o cualquier otro medio de procesamiento de datos, por ejemplo, un equipo de sobremesa o portátil con un núcleo, al menos 8Gb de RAM y al menos 16Gb de disco duro. El ordenador podría obtener el texto (1) de la red, para lo cual necesitaría conexión a internet, pero también se podría realizar la tarea

de resumen automático del texto (1) sin conexión a internet sobre documentos que se copien directamente al ordenador o estén almacenados en una memoria accesible al ordenador.

En primer lugar, el sistema está configurado para dividir el texto obtenido en una serie de frases (2). A continuación, se realiza una incrustación de las frases (3) obteniendo una serie de vectores incrustados, estos vectores se transfieren a un codificador (4) que, a su vez, genera un vector de referencia y una serie de vectores de comparación (5) estando cada uno de los vectores de comparación asociado a una de las frases de la serie de frases (2) obtenida anteriormente. Dicho codificador (4) puede, en un ejemplo de realización, estar configurado para generar una representación vectorial de menor dimensión, por ejemplo, mediante la utilización de autocodificadores entrenados utilizando los vectores de comparación (5).

En este punto se calculan las tres medidas que forman parte de una realización especialmente preferente del algoritmo programado utilizando un procedimiento del tipo dado a conocer mediante la presente invención, i.e., una medida de la posición de la frase (6), una medida de la novedad de la frase (7) y una medida de la relevancia de la frase (8). Con estas tres medidas, se puede realizar un cálculo de una puntuación global (9) de cada frase y se obtiene una selección de frases (10) en función de dicha puntuación global (9) que dará lugar al texto resumido (11) a partir del texto 1. En este ejemplo de procedimiento según la invención, se genera un texto resumido (11) por cada texto (1) analizado. A continuación, se describe cada paso de un ejemplo de procedimiento según la presente invención.

Para obtener el texto (1) a resumir automáticamente, el ordenador puede estar conectado a internet a través de una conexión inalámbrica o a través de un cable de red Ethernet. Alternativamente, el texto (1) se puede obtener a través de un medio de soporte, que puede ser cualquier entidad o dispositivo capaz de almacenar documentos de texto. Por ejemplo, el soporte podría incluir un medio de almacenamiento, como una memoria ROM, una memoria CD ROM o una memoria ROM de semiconductor, una memoria flash USB, SD, mini-SD o micro-SD, un soporte de grabación magnética, por ejemplo, un disco duro o una memoria de estado sólido (SSD, del inglés *solid-state drive*). El objeto de esta conexión y configuración a la red, o de la disponibilidad de soportes de medio de cualquier tipo, es la obtención del texto en bruto necesario para poder obtener el texto (1) sobre el que se va a realizar el resumen de texto extractivo utilizando aprendizaje profundo no supervisado y autocodificadores de la presente invención.

Una vez obtenido el texto (1), este se separa en frases (2) utilizando un clasificador (por ejemplo, una función similar a las conocidas en los diferentes lenguajes de programación como “tokenizer”) que utiliza un algoritmo no supervisado para construir un modelo para palabras abreviadas, frases hechas y palabras que se utilizan para iniciar frases. Antes de poder utilizarse, el modelo debe entrenarse usando una colección grande de texto en el lenguaje sobre el que se vaya a realizar la separación de frases.

A continuación, se procede a realizar la incrustación de las frases (3) para cada texto (1) que se pretenda resumir. En esta realización preferente de la invención, cada frase del documento $\mathcal{S}$ de entrada se incrusta en un vector $\vec{S}$ de 2400 dimensiones utilizando la metodología vectorial “skip-thought” para realizar esta incrustación. En esta realización preferente de la invención, el modelo está basado en una red codificador-decodificador, donde el codificador está formado por una red neuronal recurrente (RNN) codificada con unidades recurrentes cerradas (en inglés Gated Recurrent Units - GRUs) (Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, In Proceedings of Deep Learning and Representation Learning Workshop: NIPS 2014, 1–9) y el decodificador está formado por una red neuronal recurrente (RNN) con unidades recurrentes cerradas condicionales. En esta realización preferente de la invención, el modelo es entrenado en el conjunto de datos no etiquetado denominado BookCorpus (Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. Proceedings of the IEEE International Conference on Computer Vision, 19–27)

La **Figura 2** muestra un ejemplo de frases originales $\mathcal{S}$ (2) obtenidas de un documento de texto $\mathcal{D}$ (1) sin incrustar. Como se puede observar, tras aplicar el método “skip-thought” dichas frases quedan incrustadas en vectores incrustados $\vec{S}$ (3) que pueden ser de longitud fija, por ejemplo, de 2400 elementos. Dichos vectores incrustados son un ejemplo de representación vectorial numérica de las frases del texto (1) que permiten la realización de funciones matemáticas y, en consecuencia, automatizar el proceso de generación del resumen.

Una vez obtenidos los vectores incrustados (3), se pueden utilizar para obtener una medición de novedad (7) mediante la realización de cálculos de similitud entre ellos tal y como se explicará en mayor detalle más adelante.

En la siguiente etapa se procede a la codificación (4) de las frases incrustadas $\vec{S}$, por ejemplo, en una representación vectorial de menor dimensión denominada unidades textuales

incrustadas $\vec{\mathcal{T}}$. En una realización preferente, se diseña una red de autocodificadores para obtener esta representación vectorial de menor dimensión alimentándola con los vectores incrustados $\vec{\mathcal{S}}$ resultantes de la anterior fase. Dichos vectores incrustados se combinan en unidades textuales incrustadas $\vec{\mathcal{T}}$ que se utilizan para entrenar la red de autocodificadores. En una realización preferente de la invención, una vez entrenada la red, se utiliza solo su parte de codificador para generar unidades textuales latentes $\hat{\mathcal{T}}$, cuya combinación dará lugar a la representación latente original del documento $\widehat{\mathcal{D}}$. La **Figura 3** muestra un ejemplo de cómo se convierten las frases incrustadas $\vec{\mathcal{S}}$ en unidades textuales incrustadas $\vec{\mathcal{T}}$, posteriormente reducidas a unidades textuales latentes $\hat{\mathcal{T}}$, las cuales se unirán en la representación latente del documento $\widehat{\mathcal{D}}$.

En concreto, en la figura 3 se muestra cómo, a partir de los vectores incrustados (3) obtenidos en etapas anteriores, se genera un vector de unidades textuales incrustadas (40) que es, básicamente, la combinación de todos los elementos correspondientes a los vectores incrustados (3) en un único vector auxiliar. Dicho vector de unidades textuales incrustadas (40) se codifica, por ejemplo, mediante autocodificadores para obtener un vector de menor tamaño (41) y reducir el coste computacional del procedimiento. Una vez obtenido, se crea un vector de referencia (42) que contiene información correspondiente a cada una de las frases obtenidas (2). Finalmente, dicho vector de referencia (42) puede ser utilizado para calcular una medición de relevancia (8) como se explicará a continuación.

Tras la obtención de la representación latente del documento o vector de referencia (42), se procede a realizar el cálculo de la medida de la relevancia (8) de la frase. En una realización preferente de la invención, tras obtener la representación latente original del documento $\widehat{\mathcal{D}}$, se utiliza dicha representación latente como vector de referencia (42), posteriormente se calculan las representaciones latentes modificadas del documento $mod\widehat{\mathcal{D}}_{\mathcal{S}_i}$, una por cada frase contenida en el documento y se utilizan como vectores de comparación. Para ello, se elimina una frase del documento y se genera una representación latente del mismo $mod\widehat{\mathcal{D}}_{\mathcal{S}_i}$, pero donde dicha frase no se incluye. Entonces, para calcular la medida de la relevancia (8) de dicha frase, se calcula la similitud coseno $\text{score}^{\text{ContR}}(\mathcal{D}, \mathcal{S}_i)$ entre la representación latente original $\widehat{\mathcal{D}}$ y la representación latente modificada $mod\widehat{\mathcal{D}}_{\mathcal{S}_i}$.

$$\text{score}^{\text{ContR}}\left(\widehat{\mathcal{D}}, mod\widehat{\mathcal{D}}_{\mathcal{S}_i}\right) = 1 - \frac{\widehat{\mathcal{D}}.mod\widehat{\mathcal{D}}_{\mathcal{S}_i}}{\left\|\widehat{\mathcal{D}}\right\|\left\|mod\widehat{\mathcal{D}}_{\mathcal{S}_i}\right\|}$$

En una realización preferente de la invención, la medida de la relevancia de una frase toma valores entre 0 y 1, siendo la frase más relevante cuanto más cercano sea el valor de esta medida a uno.

La medida de la novedad (7) de la frase $\vec{s}_i$ es una medida de novedad, preferentemente, con un valor bajo si la frase es redundante o repetitiva, y un valor alto si la frase es nueva. En una realización especialmente preferente, se calcula la similitud coseno entre los vectores incrustados (3) correspondientes a cada dos frases $\vec{s}_i$ y $\vec{s}_j$.(30):

$$\mathrm{Sim}(\vec{s}_i, \vec{s}_j) = \frac{\vec{s}_i . \vec{s}_j}{\|\vec{s}_i\| \|\vec{s}_j\|}$$

En una realización preferente de la invención, la medida de novedad de una frase en un documento se calcula en base a la similitud coseno $\mathrm{Sim}(\vec{s}_i, \vec{s}_j)$ y la medida de la relevancia $\mathrm{score}^{\mathrm{ContR}}(\widehat{\mathcal{D}}, mod\widehat{\mathcal{D}}_{\mathcal{S}_i})$ previamente calculada del siguiente modo:

$$\mathrm{score}^{\mathrm{Nov}}(\mathcal{D}, \mathcal{S}_\mathrm{i}) = \begin{cases} 1, \ \text{if} \ \max(\{\mathrm{Sim}(\vec{s}_\mathrm{i}, \vec{s}_\mathrm{j})\}) < th, & 1 \leq \mathrm{j} \leq \mathcal{N}, \mathrm{i} \neq \mathrm{j} \\ 1, \ \text{if} \ \max(\{\mathrm{Sim}(\vec{s}_\mathrm{i}, \vec{s}_\mathrm{j})\}) > th, \ \mathrm{score}^{\mathrm{ContR}}(\mathcal{D}, \mathcal{S}_\mathrm{i}) > \mathrm{score}^{\mathrm{ContR}}(\mathcal{D}, \mathcal{S}_\mathrm{k}), & \\ \qquad\qquad \mathrm{k} = \mathrm{argmax}\left(\mathrm{Sim}(\vec{s}_\mathrm{i}, \vec{s}_\mathrm{j})\right), 1 \leq \mathrm{j} \leq \mathcal{N}, \mathrm{i} \neq \mathrm{j} & \\ 1 - \max(\{\mathrm{Sim}(\vec{s}_\mathrm{i}, \vec{s}_\mathrm{j})\}), \quad \mathrm{Otherwise}, & \end{cases}$$

Siendo $\mathcal{N}$ el número de frases de un documento y $\vec{s}_\mathrm{i}, \vec{s}_\mathrm{j}$ dos frases del documento. En una realización preferente de la invención, “th” es un umbral definido empíricamente para encontrar frases similares.

En una realización preferente, el valor resultante $\mathrm{score}^{\mathrm{Nov}}(\mathcal{D}, \mathcal{S}_\mathrm{i})$ estará entre 0 y 1, siendo la frase más novedosa cuanto mayor sea el valor de esta medida próximo a uno.

El cálculo de la medida de posición (6) de la frase con respecto al texto (1) se realiza teniendo en cuenta la posición que ocupa dicha frase dentro del documento original, así como el número de frases del mismo. En una realización preferente de la invención, la medida de la posición de una frase $\mathcal{S}_\mathrm{i}$ con respecto a un documento $\mathcal{D}$ se calcula del siguiente modo:

$$\mathrm{score}^{\mathrm{PosR}}(\mathcal{D}, \mathcal{S}_\mathrm{i}) = \max\left(0.5, \exp\left(\frac{-2 \cdot \mathcal{P}(\mathcal{S}_\mathrm{i})}{\mathcal{N}}\right)\right)$$

Donde $\max$ representa el valor máximo entre 0.5 y la expresión contigua, en la que $\exp.$ representa la función exponencial, $\mathcal{N}$ el número de frases del documento y $\mathcal{P}(\mathcal{S}_\mathrm{i})$ es una función que suministra la posición relativa de la frase en el documento. En una realización

preferente de la invención, $\mathcal{P}(\mathcal{S}_\mathrm{i}) = 1$ para la primera frase. En una realización preferente, el valor resultante estará entre 1 y 0.5, siendo el valor de la posición de la primera frase 1 y decreciendo dicho valor en frases sucesivas.

Una vez obtenidas las medidas de posición, novedad y relevancia de las frases, se realiza el cálculo de la puntuación final de cada frase 9. En una realización preferente de la invención, dicho valor resulta de la suma ponderada de las tres medidas anteriores, multiplicadas cada una de ellas por los correspondientes coeficientes de relevancia α, novedad β y posición γ.

$$\mathrm{score}^{\mathrm{f}}(\mathcal{D}, \mathcal{S}_\mathrm{i}) = \alpha \cdot \mathrm{score}^{\mathrm{ContR}}(\mathcal{D}, \mathcal{S}_\mathrm{i}) + \beta \cdot \mathrm{score}^{\mathrm{Nov}}(\mathcal{D}, \mathcal{S}_\mathrm{i}) + \gamma \cdot \mathrm{score}^{\mathrm{PosR}}(\mathcal{D}, \mathcal{S}_\mathrm{i})$$

En una realización preferente de la invención, los valores α, β, γ pueden tomar cualquier valor entre 0 y 1 y se determinan empíricamente. En una realización preferente de la invención, se identifican como valores especialmente preferentes: $\alpha = 0.45$, $\beta = 0.35$ y $\gamma = 0.20$.Sin embargo, en otras realizaciones de la presente invención se utiliza cualquier valor que cumpla con el requisito: $\alpha > \beta > \gamma$.

Finalmente, se realiza la selección de frases 10 que formarán parte del documento de texto resumido 11 del documento de texto original 1. En la siguiente ecuación, $SCORE(\mathcal{D})$ representa una lista ordenada de las medidas obtenidas para las frases de un documento:

$$SCORE(\mathcal{D}) = score^f(\mathcal{D}, \mathcal{S}_1),\ score^f(\mathcal{D}, \mathcal{S}_2), \ldots\ldots, score^f(\mathcal{D}, \mathcal{S}_\mathcal{N})$$

En una realización preferente de la invención, la ordenación relativa de cada frase $Rank(\mathcal{S}_i)$ dentro de un documento se puede obtener calculando la ordenación de su medición final según las siguientes ecuaciones:

$$Rank(\mathcal{S}_i) = 1 + \sum_{e=1}^{\mathcal{N}} \Psi(e, i), \qquad and$$

$$\Psi(e, i) = \begin{cases} 1, & if\ \ score^f(\mathcal{D}, \mathcal{S}_i) + \varepsilon \cdot i > \\ & \qquad score^f(\mathcal{D}, \mathcal{S}_e) + \varepsilon \cdot e \\ 0, & Otherwise \end{cases}$$

Donde ε(epsilon) es una constante muy pequeña. En una realización preferente, $\varepsilon \to 0+$ toma un valor muy pequeño y se utiliza para resolver la posible situación donde $score^f(\mathcal{D}, \mathcal{S}_e) == score^f(\mathcal{D}, \mathcal{S}_i)$ lo que permite dar prioridad a la posición de una frase.

En una realización preferente, el siguiente paso es elegir las frases con mayores ordenaciones

relativas para generar el resumen (11) del texto original (1). En una realización preferente, el documento de texto resumido $\mathrm{Summary}(\mathcal{D}, \mathcal{L})$ contendrá $\mathcal{L}$ frases. Esta selección de $\mathcal{L}$ frases se puede realizar de dos modos diferentes: (i) realizando una ordenación de las frases según la puntuación final de cada frase por orden descendente y eligiendo un número específico de frases con la mayor puntuación final

$$\mathrm{Summary}(\mathcal{D}, \mathcal{L}) = \sum_{\mathrm{a}=1}^{\mathcal{L}} \mathcal{S}_i(\mathcal{D}), \mid \mathrm{Rank}(\mathcal{S}_i) == \mathrm{a}\,, 1 \leq \mathrm{i} \leq \mathcal{N}$$

o bien (ii) realizando una ordenación de las frases según su frecuencia de aparición en el documento original.

$$\mathrm{Summary}(\mathcal{D}, \mathcal{L}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathcal{S}_i(\mathcal{D})\,, \mathrm{if}\ \mathrm{Rank}(\mathcal{S}_i) \leq \mathcal{L}$$

REIVINDICACIONES

1. Procedimiento de generación de un resumen a partir de un documento de texto que comprende las etapas de:

- obtención del documento de texto mediante un procesador;
- obtención de una serie de frases a partir del documento de texto;
- codificación de la serie de frases;

caracterizado por:

- obtener la serie de frases a partir del documento de texto mediante un algoritmo no supervisado;
- codificar la serie de frases mediante una red codificador-decodificador, obteniendo una serie de frases incrustadas;
- codificar las frases incrustadas obteniendo una representación vectorial de menor dimensión utilizando una red de autocodificadores, obteniendo frases codificadas;
- obtener una representación latente original del documento mediante la concatenación de las frases codificadas;
- asignar una medición de relevancia a cada una de las frases codificadas;
- asignar una medición de novedad a cada una de las frases codificadas;
- asignar una medición de posición de cada una de las frases codificadas;
- a partir de una combinación de las medidas de relevancia, novedad y posición, asignar una puntuación global a cada una de las frases codificadas;
- seleccionar las frases a disponer en el resumen a partir de la puntuación global de las frases codificadas.

2. Procedimiento, según la reivindicación 1, caracterizado por que las frases codificadas corresponden a una serie de vectores incrustados que se obtiene utilizando redes neuronales recurrentes.

3. Procedimiento, según cualquiera de las reivindicaciones 1 o 2, caracterizado por que la codificación de las frases se realiza mediante la metodología Skip-Thought.

4. Procedimiento, según la reivindicación 1, caracterizado por que la medida de relevancia de cada frase se obtiene en base a la medida de similitud coseno existente entre una representación latente original del documento de texto y una representación latente modificada del documento de texto, siendo la representación latente modificada obtenida

mediante la eliminación de la frase de la que se quiere obtener su relevancia.

5. Procedimiento, según la reivindicación 1, caracterizado por que la medición de novedad se basa en calcular la similitud coseno de la serie de vectores incrustados obteniendo un valor intermedio de similitud y, en función del valor intermedio de similitud, asignar la medición de novedad.

6. Procedimiento, según la reivindicación 5, caracterizado por que el valor intermedio de similitud se calcula a partir del valor máximo de similitud coseno entre los vectores incrustados.

7. Procedimiento, según la reivindicación 5, caracterizado por que la medición de novedad es 1 si el valor intermedio es inferior a un valor umbral predeterminado.

8. Procedimiento, según la reivindicación 5, caracterizado por que la medición de novedad es igual a 1-V, donde V es el valor intermedio si el valor intermedio es superior al valor umbral.

9. Procedimiento, según la reivindicación 1, caracterizado por que la medida de posición de cada frase se realiza teniendo en cuenta la posición de la frase dentro del documento de texto, así como el número de frases del documento de texto.

10. Procedimiento, según cualquiera de las reivindicaciones anteriores, caracterizado por que la medición de relevancia comprende: generar un vector de referencia basado en la serie de frases, generar un vector de comparación de cada frase en el que el vector de comparación de cada frase corresponde al vector de referencia eliminando las partes del vector de referencia que corresponden a la frase y calcular la medición de relevancia en función de un cálculo de similitud coseno entre el vector de referencia y cada vector de comparación.

11. Procedimiento, según la reivindicación 10, caracterizado por que el vector de referencia se obtiene a partir de la adición de elementos de los vectores incrustados.

12. Procedimiento, según la reivindicación 11, caracterizado por que el vector de referencia se obtiene a partir de un autocodificador entrenado con la serie de vectores incrustados.

13. Procedimiento, según cualquiera de las reivindicaciones anteriores, caracterizado por que la selección de las frases a disponer en el resumen comprende: organizar las frases en función

de la puntuación global y seleccionar las frases que están por encima de una puntuación umbral predeterminada.

14. Procedimiento, según cualquiera de las reivindicaciones anteriores, caracterizado por que la selección de las frases a disponer en el resumen comprende: organizar las frases en función de la puntuación global y seleccionar las primeras X frases, siendo X un valor predeterminado de frases.

15. Procedimiento según cualquiera de las reivindicaciones anteriores, caracterizado por que la obtención del documento de texto se realiza a través de internet.

16. Procedimiento, según cualquiera de las reivindicaciones 1 a 15, caracterizado por que la obtención del documento de texto se realiza a partir de un medio de almacenamiento externo seleccionado de entre: una memoria ROM, una memoria CD ROM o una memoria ROM de semiconductor, una memoria flash USB, SD, mini-SD o micro-SD, un soporte de grabación magnética, un disco duro o una memoria de estado sólido.

17. Sistema de generación de un resumen a partir de un documento de texto que comprende medios de acceso a un documento de texto y un procesador configurado para:

- obtener el documento de texto;
- obtener de una serie de frases a partir del documento de texto;
- codificar la serie de frases a partir del documento de texto;

caracterizado por que el procesador está configurado para codificar la serie de frases a partir del documento de texto mediante un algoritmo no supervisado;

- obtener una representación latente original del documento mediante la concatenación de un conjunto de frases codificadas;
- asignar una medición de novedad a cada una de las frases;
- asignar una medición de relevancia a cada una de las frases;
- asignar una medición de posición a cada una de las frases;
- a partir de las mediciones de novedad, relevancia y posición, asignar una puntuación global a cada una de las frases;
- seleccionar las frases a disponer en el resumen a partir de la puntuación global de las frases;

donde la medición de novedad comprende codificar, mediante el procesador, las frases para obtener una serie de vectores incrustados; calcular la similitud coseno de la serie de vectores

incrustados obteniendo un valor intermedio de similitud y, en función del valor intermedio de similitud, asignar la medición de novedad.

18. Sistema, según la reivindicación 17, caracterizado por que la codificación de las frases para obtener la serie de vectores incrustados se realiza mediante la metodología Skip-Thought.

19. Sistema, según cualquiera de las reivindicaciones 17 ó 18, caracterizado por que el procesador está configurado para:

- asignar una medición de relevancia a cada una de las frases; y
- asignar la puntuación global en función de la medición de relevancia

en el que la medición de relevancia comprende: generar un vector de referencia basado en la serie de frases, generar un vector de comparación de cada frase en el que el vector de comparación de cada frase corresponde al vector de referencia eliminando las partes del vector de referencia que corresponden a la frase y calcular la medición de relevancia en función de un cálculo de similitud coseno entre el vector de referencia y cada vector de comparación.

20. Sistema, según cualquiera de las reivindicaciones 17 a 19, caracterizado por que el procesador está configurado para:

- asignar una medición de posición; y
- asignar la posición global en función de la medición de posición;

en el que la medición de posición se calcula en función de la posición relativa de la frase respecto al documento.

21. Un producto de programa que comprende medios de instrucciones de programa para llevar a cabo el procedimiento definido en cualquiera de las reivindicaciones 1 a 16 cuando el programa se ejecuta en un procesador.

22. Un producto de programa según la reivindicación 21, almacenado en un medio de soporte de programas.

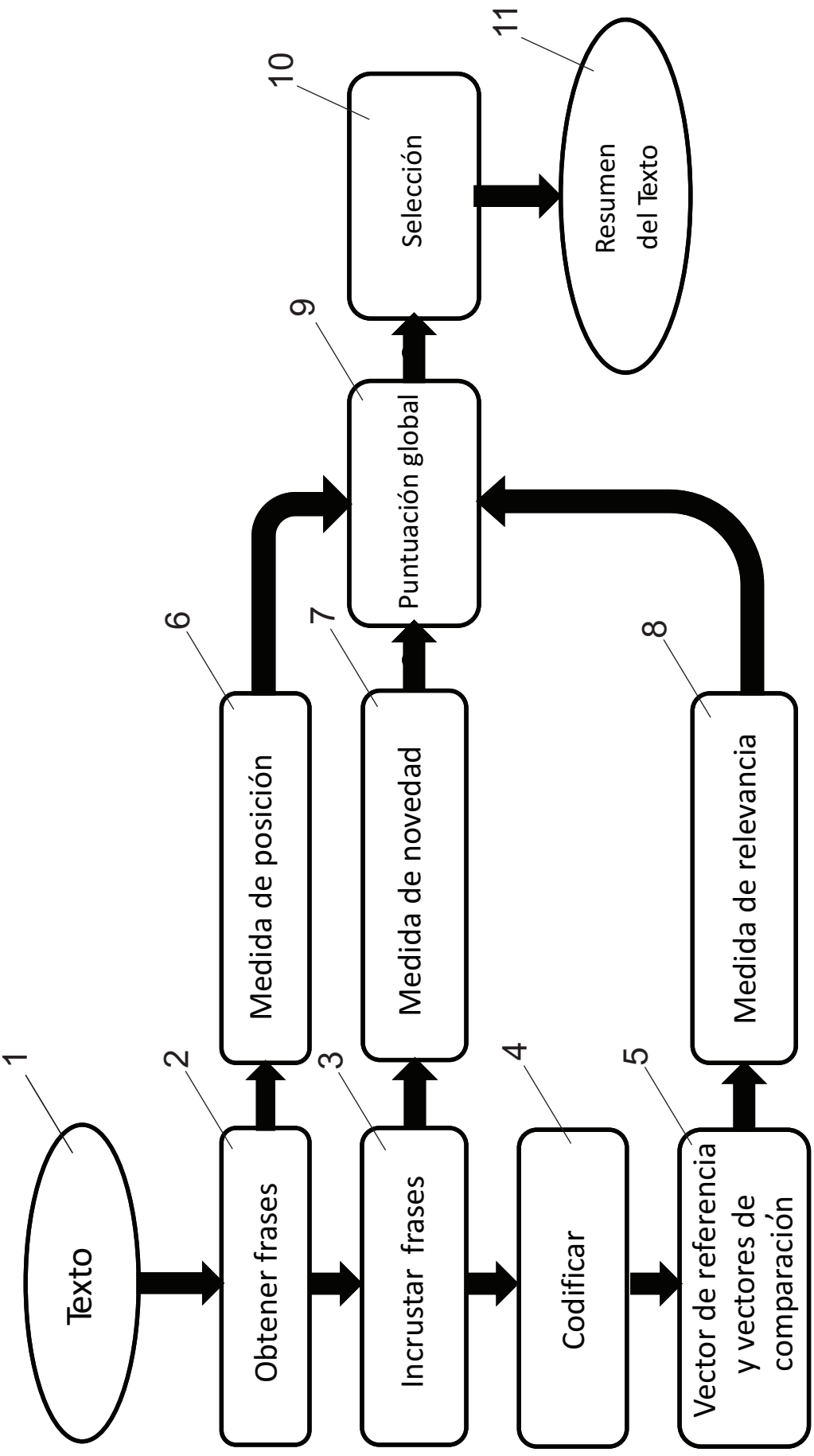
1
Texto
2
Obtener frases
3
Incrustar frases
4
Codificar
5
Vector de referencia y vectores de comparación
6
Medida de posición
7
Medida de novedad
8
Medida de relevancia
9
Puntuación global
10
Selección
11
Resumen del Texto

Fig. 1

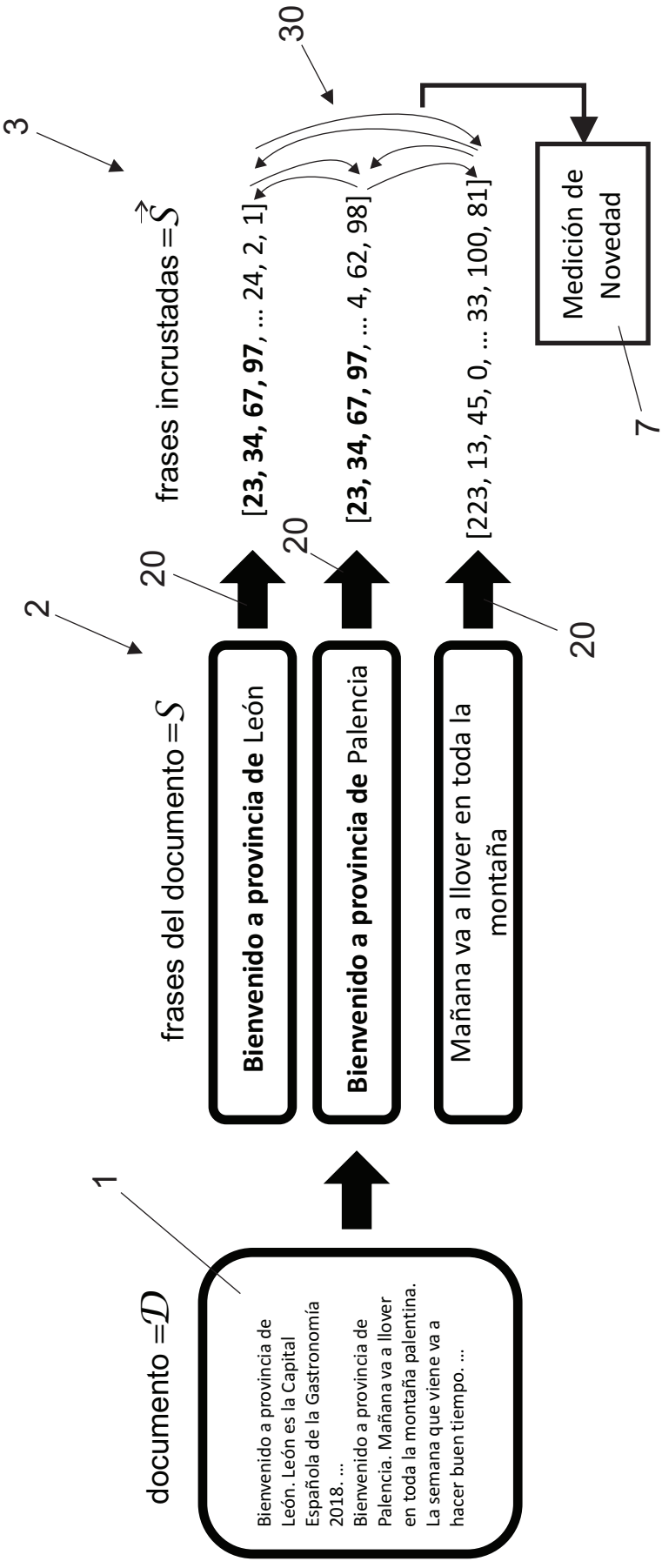
documento = $\mathcal{D}$
Bienvenido a provincia de León. León es la Capital Española de la Gastronomía 2018. ... Bienvenido a provincia de Palencia. Mañana va a llover en toda la montaña palentina. La semana que viene va a hacer buen tiempo. ...
1
frases del documento = $\mathcal{S}$
Bienvenido a provincia de León
Bienvenido a provincia de Palencia
Mañana va a llover en toda la montaña
2
20
20
20
frases incrustadas = $\vec{\mathcal{S}}$
[23, 34, 67, 97, ... 24, 2, 1]
[23, 34, 67, 97, ... 4, 62, 98]
[223, 13, 45, 0, ... 33, 100, 81]
3
30
Medición de Novedad
7

Fig. 2

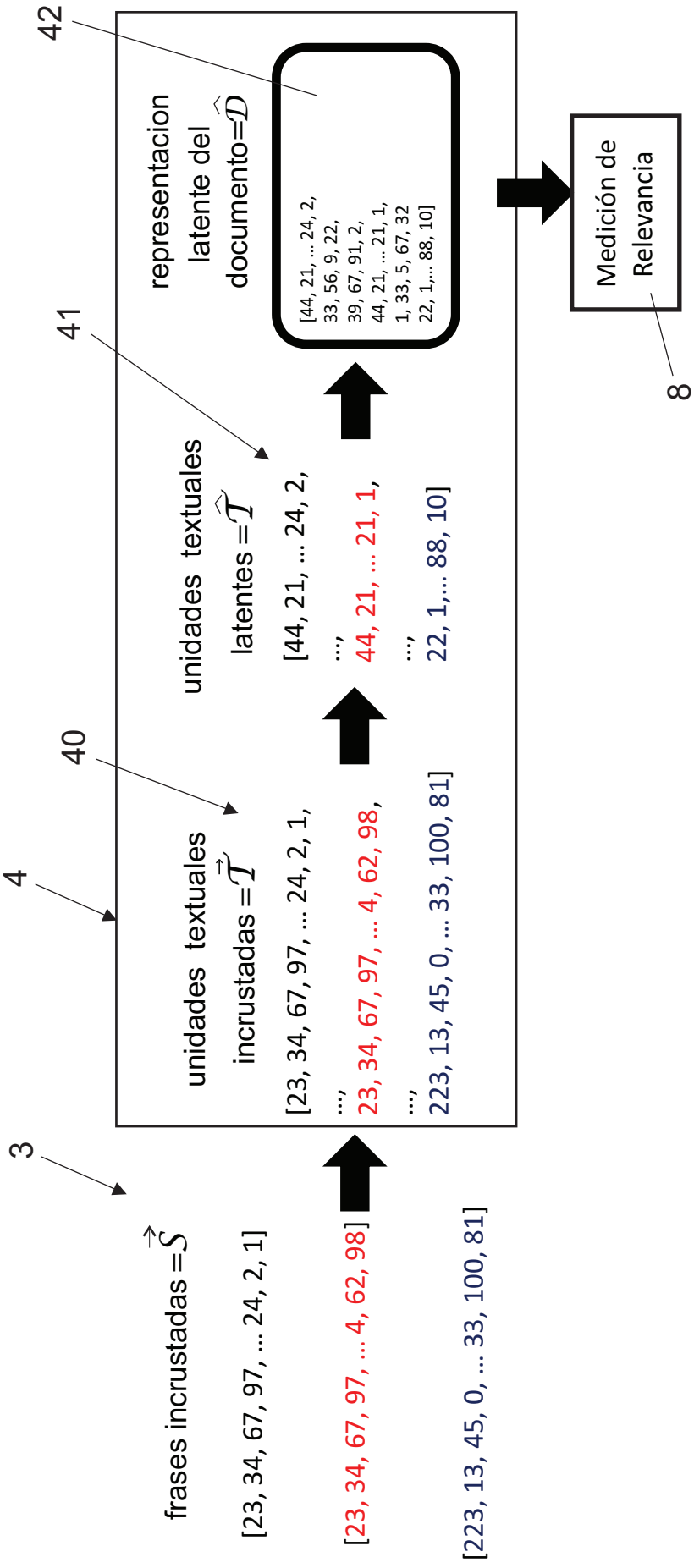
42
representacion latente del documento=$\hat{\mathcal{D}}$
[44, 21, ... 24, 2, 33, 56, 9, 22, 39, 67, 91, 2, 44, 21, ... 21, 1, 1, 33, 5, 67, 32 22, 1,... 88, 10]
41
unidades textuales latentes =$\hat{\mathcal{T}}$
[44, 21, ... 24, 2,
...,
44, 21, ... 21, 1,
...,
22, 1,... 88, 10]
40
unidades textuales incrustadas =$\vec{\mathcal{T}}$
[23, 34, 67, 97, ... 24, 2, 1,
...,
23, 34, 67, 97, ... 4, 62, 98,
...,
223, 13, 45, 0, ... 33, 100, 81]
4
3
frases incrustadas =$\vec{\mathcal{S}}$
[23, 34, 67, 97, ... 24, 2, 1]
[23, 34, 67, 97, ... 4, 62, 98]
[223, 13, 45, 0, ... 33, 100, 81]
Medición de Relevancia
8

Fig. 3