

Un algoritmo lingüístico-estadístico para resumen automático de textos especializados

Iria da Cunha,
IULA-Universitat Pompeu Fabra y
LIA-Université d'Avignon
iria.dacunha@upf.edu

Patricia Velázquez-Morales,
patricia_velazquez@yahoo.com

Juan-Manuel Torres-Moreno,
LIA-Université d'Avignon y
Ecole Polytechnique de Montréal
juan-manuel.torres@univ-avignon.fr

Jorge Vivaldi
IULA-Universitat Pompeu Fabra
jorge.vivaldi@upf.edu

Resumen

En este trabajo se presenta un nuevo algoritmo de resumen automático de textos especializados, en concreto del dominio médico, que aúna estrategias lingüísticas y estadísticas. La novedad del artículo radica en la correcta combinación de dichas estrategias de cara a demostrar que los sistemas híbridos pueden obtener mejores resultados que los sistemas estadísticos o lingüísticos por sí solos. Se aplica el algoritmo sobre un corpus de textos médicos y se evalúa siguiendo el protocolo de NIST y utilizando el paquete ROUGE. Se obtienen excelentes resultados en comparación con otros sistemas y se observa que los resúmenes realizados son muy similares a los de los especialistas del dominio.

1. Introducción

El resumen automático es actualmente un tema de investigación muy relevante. La investigación en esta área se inició en los años sesenta, empleando técnicas basadas en frecuencias de palabras (Luhn, 1959) o frases clave (Edmundson, 1969). Con el tiempo, estas técnicas han ido evolucionando y volviéndose más complejas. Podemos hacer una división general de estas técnicas en dos grupos principales: las técnicas estadísticas y las técnicas lingüísticas. En el primer grupo¹, encontramos, entre otros, trabajos que emplean modelos bayesianos (Kupiec, Pedersen, and Chen, 1995), la *Maximal Marginal Relevance* (Goldstein et al., 1999), técnicas de *clustering* (Radev, Jing, and Budzikowska, 2000), grafos (Radev et al., 2004; Vanderwende, Banko, and Menezes, 2004; Leskovec, Milic-Frayling, and Grobelnik, 2005) o aprendizaje automático (Kupiec, Pedersen, and Chen, 1995; Berger and Mittal, 2000; Marcu and Echiabi, 2002; Leskovec, Milic-Frayling, and Grobelnik, 2005; Barzilay and Lapata, 2005). En el segundo grupo, destacamos trabajos que explotan las posiciones textuales (Brandow, Mitze, and Rau, 1995; Lin and Hovy, 1997), la estructura del discurso (Ono, Sumita,

and Miike, 1994; Marcu, 1998; Marcu, 2000; Teufel and Moens, 2002; Polanyi et al., 2004; Thione et al., 2004) o las cadenas léxicas (Barzilay and Elhadad, 1997; Silber and McCoy, 2000; Fuentes, 2008). Todos estos sistemas de resumen automático emplean estrategias que conllevan algún tipo de criterio lingüístico y estadístico, pero por lo general siempre hay una mayor proporción de uno u otro. Hay pocos trabajos en los que se combinan ambos criterios de una manera más igualitaria, por ejemplo Nomoto, T. and Nitta, Y. (1994), Aretoulaki (1996), Hovy, E. and Lin, C.Y. (1999), Alonso and Fuentes (2003) y Lacatusu, Parker, and Harabagiu (2003). En nuestro trabajo intentamos combinar técnicas estadísticas y lingüísticas de una manera equitativa y adecuada, para aprovechar las ventajas de ambas en la tarea de resumen automático. En concreto, hemos diseñado un algoritmo de resumen que combina los sistemas estadísticos CORTEX (Torres-Moreno, Velázquez-Morales, and Meunier, 2001) y ENERTEX (Fernández, SanJuan, and Torres-Moreno, 2007) y los sistemas lingüísticos YATE (Vivaldi, 2001) y DISICOSUM (da Cunha, 2008). En da Cunha et al. (2007) se realizó una primera aproximación, diseñándose un algoritmo que combinaba estos sistemas. En este nuevo trabajo se desarrolla y se refina ese algoritmo inicial, se realizan nuevos experimentos y se evalúan mediante el paquete ROUGE (Lin, 2004), usando un protocolo similar al de las evaluaciones del

¹No pretendemos hacer aquí una revisión exhaustiva del estado de la cuestión en resumen automático. Para más información sobre técnicas y/o sistemas de resumen remitimos a los trabajos de (Spärck Jones, 2007; Mani, 2001; Mani and Maybury, 1999).

NIST². La decisión de llevar a cabo estas mejoras se tomó porque el algoritmo inicial fue diseñado para resumir textos largos, de cuatro o cinco páginas y con diversos apartados, mientras que ahora nos planteamos obtener resúmenes de textos de contenido específico y más cortos, de una página aproximadamente.

Nuestra tarea consiste en resumir textos médicos especializados. Tal como se indica en Afantenos, Karkaletsis, and Stamatopoulos (2005), en el ámbito médico debe gestionarse una gran cantidad de textos y el resumen automático puede ayudar a procesar esta masa de documentos. Nosotros trabajamos con artículos médicos de investigación, ya que este tipo de textos se publican con sus correspondientes resúmenes escritos por los autores de los textos y esto nos permite compararlos con los resúmenes obtenidos por nuestro sistema y facilitar así su evaluación. En el futuro podría adaptarse el componente lingüístico de nuestro algoritmo para ser empleado en otras áreas similares, como la biología, la genómica, la química, etc., así como en otras lenguas próximas.

En la sección 2 detallamos los diversos componentes del algoritmo. En la sección 3 explicamos su arquitectura. En la sección 4 mostramos los experimentos realizados y su evaluación. Por último, en la sección 5, exponemos las conclusiones y algunas perspectivas.

2. Componentes del algoritmo

A continuación explicamos los cuatro sistemas que se han empleado como componentes del algoritmo.

2.1. CORTEX

CORTEX (Torres-Moreno, Velázquez-Morales, and Meunier, 2001; Torres-Moreno, Velázquez-Morales, and Meunier, 2002) es un sistema de resumen automático basado en el Modelo de Espacio Vectorial (VSM) (Salton and McGill, 1983). Se trata de un sistema de resumen por extracción mono-documento que combina varias métricas sin aprendizaje. Estas métricas resultan de algoritmos de procesamiento estadísticos y de información sobre la representación vectorial del documento. La idea principal es la de representar un texto en un espacio vectorial adecuado y aplicar procesamiento estadístico. Con el fin de reducir la complejidad del espacio, se realiza un preprocesamiento del documento: se filtran y

se lematizan las palabras del texto. La representación de *bolsa-de-palabras* produce una matriz de frecuencias/ausencias $S_{[P \times N]}$ de $\mu = 1, \dots, P$ frases u oraciones³ (filas) y un vocabulario de $i = 1, \dots, N$ términos (columnas). CORTEX puede emplear hasta $\Gamma = 11$ métricas para evaluar la pertinencia de las frases. Algunas métricas utilizan el ángulo entre el título y cada una de las frases, la matriz de Hamming (matriz donde cada valor representa el número de frases en las que uno de los términos i o j está presente), la suma de pesos Hamming de palabras por segmento, la entropía, la frecuencia o las interacciones, entre otras. El sistema asigna una puntuación a cada frase con el algoritmo de decisión que combina las métricas normalizadas. Se calculan dos promedios: una tendencia positiva $\lambda_s > 0,5$ y otra negativa $\lambda_s < 0,5$ (el caso $\lambda_s = 0,5$ es ignorado). El algoritmo de decisión que permite combinar el voto de Γ métricas es el siguiente:

$$\sum \alpha = \sum_{\nu=1}^{\Gamma} (|\lambda_s^\nu| - 0,5); |\lambda_s^\nu| > 0,5 \quad (1)$$

$$\sum \beta = \sum_{\nu=1}^{\Gamma} (0,5 - |\lambda_s^\nu|); |\lambda_s^\nu| < 0,5 \quad (2)$$

Γ es el número de métricas y ν es el índice de las métricas. El valor de λ fue normalizado en el rango $[0 - 1]$ para evitar diferencias de magnitud entre las métricas. El valor dado a cada frase s se calcula de la siguiente manera:

$$\text{IF } \left(\sum \alpha > \sum \beta \right) \text{ THEN Score}_s = 0,5 + \frac{\sum \alpha}{\Gamma} \\ \text{ELSE Score}_s = 0,5 - \frac{\sum \beta}{\Gamma}$$

La figura 1 muestra el esquema del resumidor CORTEX. En los experimentos se usaron las métricas FAX (F=Frecuencia de términos, A=Ángulo entre el título y cada una de las frases y X=Posición de la frase en el documento, modelada con una función cuadrática que otorga una ponderación mayor a las primeras y últimas frases del texto).

2.2. ENERTEX

ENERTEX (Fernández, SanJuan, and Torres-Moreno, 2007; Fernández, SanJuan, and Torres-Moreno, 2008; Fernández, 2009) también es un sistema de resumen automático basado en VSM, pero en este caso se trata de un enfoque de redes de neuronas inspirado en la física estadística. El algoritmo modela los documentos como una red

²Más información sobre las campañas *Document Understand Conference* y *Text Analysis Conference* (DUC/TAC) puede encontrarse en el sitio web del NIST: <http://www.nist.gov/tac/>

³En función del segmentador empleado.

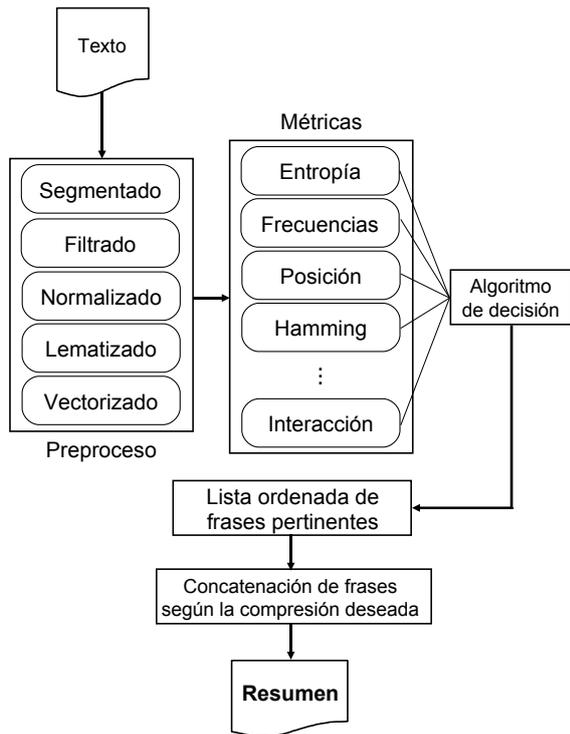


Figura 1: Arquitectura de CORTEX.

de neuronas de la que se estudia su energía textual. La idea principal es que un documento puede ser procesado como un conjunto de unidades interactivas (las palabras), donde cada unidad se ve afectada por el campo creado por las demás (véase figura 2).

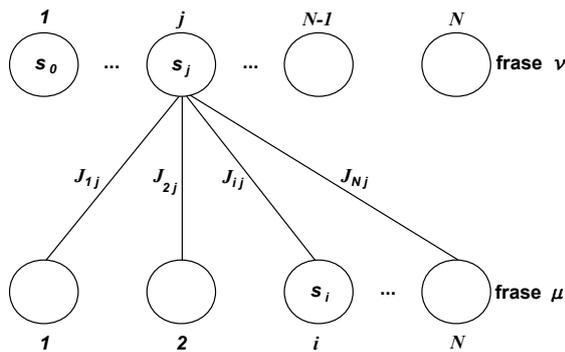


Figura 2: Campo creado por los términos de la frase μ que afecta cada término j de la frase ν .

La memoria asociativa de Hopfield (Hopfield, 1982) está basada en sistemas físicos, como el modelo magnético de Ising (formalismo que describe un sistema con dos estados, llamados *spines*, para construir una red de neuronas capaz de almacenar/recuperar patrones). El aprendizaje se realiza usando la regla de Hebb (Hertz, Krogh, and Palmer, 1991):

$$J_{i,j} = \sum_{P \text{ frases}} s_i s_j \quad (3)$$

y la recuperación por minimización de la energía del modelo de Ising:

$$E^{\mu,\nu} = \sum s_i^\mu J_{i,j} s_j^\nu \quad (4)$$

La principal limitación de la red de Hopfield es su capacidad de almacenamiento: los patrones no deben estar correlacionados para poder obtener sin problemas el error de recuperación. Esto restringe en gran medida sus aplicaciones, pero ENERTEX se beneficia de este comportamiento. El VSM representa las frases del documento en vectores (partiendo de la vectorización producida por CORTEX) y estos vectores pueden estudiarse como una red neuronal de Hopfield. Las frases son representadas como una cadena (patrón) de N neuronas activas (término presente) o inactivas (término ausente) con un vocabulario de N términos por documento. Un documento de P frases está formado por P cadenas en un espacio vectorial de N dimensiones. Estos vectores están relacionados en función de las palabras compartidas. Si los temas están semánticamente próximos, es razonable suponer un alto grado de correlación. ENERTEX calcula la interacción entre los términos (3) y la energía textual entre las frases (4). La ponderación de las frases se obtiene utilizando sus valores absolutos de energía. El resumen está formado por las frases más importantes, es decir las que obtienen los valores más altos.

2.3. DISICOSUM

DISICOSUM (da Cunha and Wanner, 2005; da Cunha, Wanner, and Cabré, 2007; da Cunha, 2008) es un modelo de resumen automático de textos médicos que parte de la idea de que los profesionales de un dominio especializado emplean técnicas concretas para resumir los textos de su ámbito. Para diseñar este sistema se analizó un corpus de artículos médicos con sus correspondientes resúmenes para determinar cuáles son las estrategias que usan los médicos para resumir y cuál es la información que debe seleccionarse para realizar un resumen de este tipo. Otra de las aportaciones del modelo es la combinación de diversos criterios lingüísticos. Por lo general (ver sección 1) los sistemas lingüísticos de resumen automático suelen emplear un solo tipo de criterio (por ejemplo, frecuencias de palabras, oraciones clave, posiciones textuales, estructura discursiva, etc.). Sin embargo, en DISICOSUM se integran criterios basados en la estructura textual, en las unidades léxicas y en la estructura discursiva y sintáctico-comunicativa del texto. El modelo

está formado por reglas que se relacionan con estos criterios lingüísticos.

Con respecto a la estructura textual, DISICOSUM incluye una regla que asigna un peso adicional a las oraciones que se encuentren en las siguientes posiciones del texto: las tres primeras oraciones de la sección de *Fundamento*, las dos primeras oraciones de las secciones de *Pacientes y métodos* y *Resultados*, y las tres primeras y las tres últimas oraciones de la sección de *Discusión*.

En cuanto a las unidades léxicas, DISICOSUM incluye reglas de dos tipos:

- Reglas que otorgan más peso a las oraciones que contienen: 1) palabras del título principal del artículo (excepto *stopwords*), 2) formas verbales en primera persona del plural, 3) palabras incluidas en una lista que contiene verbos y sustantivos del dominio médico que pueden ser pertinentes para el resumen (por ejemplo, *analizar*, *evaluar*, *objetivo*, *estudio*, etc.) y 4) cualquier información numérica en las secciones de *Pacientes y métodos* y *Resultados*.
- Reglas de eliminación de oraciones que contienen unidades que se refieren a: 1) tablas o figuras, 2) aspectos estadísticos o computacionales, 3) trabajos anteriores y 4) definiciones.

Por último, el modelo incluye reglas discursivas y reglas que combinan algunos aspectos de la estructura discursiva con la estructura sintáctica y comunicativa (reglas DISICO). Para formalizar dichas reglas el modelo emplea la *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1988) y la *Meaning-Text Theory* (MTT) (Mel'cuk, 1988; Mel'cuk, 2001). La RST es una teoría de organización del texto que caracteriza su estructura como un árbol jerárquico que contiene elementos (núcleos [N] y satélites [S]) ligados mediante relaciones discursivas (como, por ejemplo, *Elaboración*, *Concesión*, *Antítesis*, *Condición*, *Contraste*, *Background*, etc.). En la figura 3 se muestra un ejemplo de la representación discursiva de la RST en forma de árbol (con dos relaciones: *Elaboración* y *Background*).

La MTT es una teoría que integra diversos aspectos del lenguaje. DISICOSUM emplea, por un lado, elementos de la sintaxis de dependencias para representar una oración como un árbol donde las unidades léxicas son los nodos y las relaciones entre ellas son actanciales (ACT), atributivas (ATTR), apenditivas (APPEND) y coordinativas (COORD). Por otro lado, DISICOSUM emplea la distinción entre Tema y Rema, que es parte de

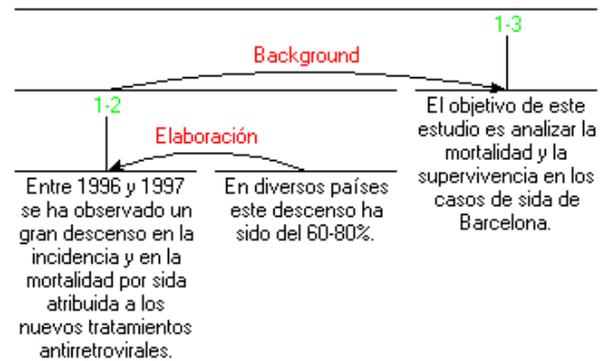


Figura 3: Ejemplo de árbol discursivo con relaciones de la RST.

la estructura comunicativa de la MTT. Algunos ejemplos de las reglas DISICOSUM son:^{4,5}

- IF *S* is satellite_{REFORMULATION} R
THEN ELIMINATE *S*
Ej. [Se incluyeron sólo pacientes estables.]_N [Es decir, se consideraron pacientes que no habían precisado cambiar su medicación habitual en los últimos 15 días y clínicamente no referían un empeoramiento importante.]_S
- IF *S* is satellite_{BACKGROUND} B
THEN ELIMINATE *S*
Ej. [La quimioprofilaxis (QP) antituberculosa es una de las principales intervenciones en la cadena de actuaciones para la prevención de la tuberculosis (TBC).]_S [El objetivo de este estudio es conocer el grado de cumplimiento y la tolerancia terapéutica de la QP antituberculosa en nuestro medio, así como describir y analizar sus factores condicionantes.]_N
- IF *S* is satellite_{ELABORATION} El
AND *S* elaborates on the Theme of the nucleus of El
THEN ELIMINATE *S*
Ej. [Como grupo de control se empleó el formado por 377 mujeres sanas.]_N [Este grupo se obtuvo mediante selección aleatoria entre mujeres que entre 1989 y 1991 habían dado a luz en nuestro hospital.]_S
- IF *N* is nucleus_{LIST} L
THEN KEEP *N*
Ej. [La primera prueba de marcha (PM) se

⁴La lista de todas las reglas que conforman DISICOSUM puede observarse en (da Cunha, 2008).

⁵Los fragmentos de texto en itálica son eliminados por las reglas.

efectuó respirando aire sintético a 2 litros por minuto a través de gafas nasales.]_N [La segunda PM se realizó con oxígeno continuo a 2 litros por minuto.]_N [La tercera se llevó a cabo acoplado a la misma fuente una VAO a un flujo de 2 litros por minuto.]_N

DISICOSUM es un modelo de resumen semiautomático, debido a la carencia actual de analizadores automáticos discursivos para el castellano⁶. Así pues, los textos de entrada deben estar previamente anotados con etiquetas que reflejen las relaciones de la RST. En la figura 4 se muestra la arquitectura del modelo.

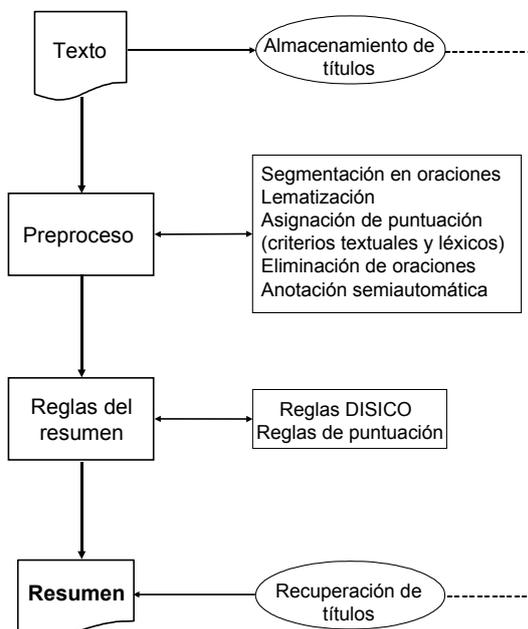


Figura 4: Arquitectura de DISICOSUM.

2.4. YATE

Otro de los sistemas empleados en este trabajo es el sistema YATE (Vivaldi, 2001; Vivaldi and Rodríguez, 2001; Vivaldi and Rodríguez, 2002). YATE es un extractor de términos híbrido cuyas características más relevantes son: el uso intensivo de información semántica junto con el uso de técnicas de combinación de los resultados obtenidos a partir de diferentes técnicas de extracción.

⁶Actualmente hay analizadores discursivos para el japonés (Sumita et al., 1992), el inglés (Marcu, 2000) y el portugués de Brasil (Pardo, T. and Nunes, M. and Rino, M., 2004; Pardo, T. and Nunes, M., 2008). También existe un proyecto en curso en el *Laboratoire Informatique d'Avignon* para desarrollar un analizador discursivo para el castellano.

Al igual que otros extractores de similares características ha sido desarrollado para el ámbito médico en castellano aunque está siendo adaptado con éxito a otros dominios (genómica, derecho, economía, informática y medio ambiente) y otras lenguas (catalán).

La figura 5 muestra el esquema general de YATE y permite apreciar los diferentes módulos de análisis que forman este extractor:

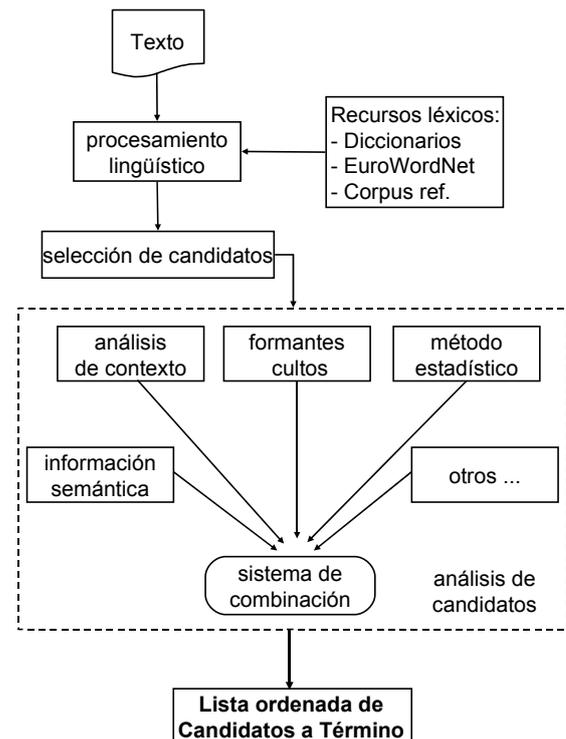


Figura 5: Arquitectura de YATE.

1. Información semántica: utiliza información obtenida a partir de EuroWordNet⁷.
2. Análisis de contexto: evalúa cada candidato a término (CAT) en función de otros candidatos que aparecen en su contexto oracional.
3. Formantes cultos: descompone cada CAT en sus formantes aprovechando las características formales de algunos términos (sobre todo en el dominio médico).
4. Método estadístico: evalúa los CAT poliléxicos según su información mutua (u otra medida de asociación).

YATE puede ser eficazmente integrado como un componente de ponderación de palabras en un

⁷<http://www.i11c.uva.nl/EuroWordNet>

sistema de resumen, como CORTEX o ENERTEX por ejemplo, de manera simple. En nuestro trabajo, hemos usado la capacidad de detección de términos médicos de YATE para ponderar cada término del modelo vectorial proporcionalmente a su terminologicidad (valor comprendido entre 0 y 1). Así, las oraciones que contienen términos del dominio médico serán bonificadas en relación a las otras.

3. Arquitectura del algoritmo

El algoritmo desarrollado, que como hemos mencionado anteriormente parte del trabajo de (da Cunha et al., 2007), combina los cuatro sistemas detallados en la sección 2: CORTEX, YATE, ENERTEX y DISICOSUM.

El sistema propuesto tiene la arquitectura que se indica en la figura 6 y consta de varios resumidores autónomos que se combinan de manera equilibrada para formar un único resumidor híbrido. Algunos de los resumidores utilizan métodos numéricos (CORTEX y ENERTEX), otro resumidor tiene un carácter estrictamente lingüístico (DISICOSUM) y en los dos sistemas restantes las métricas estadísticas (de CORTEX y ENERTEX) se combinan con la información lingüística procedente del extractor de términos (YATE).

Tomando como entrada los resúmenes realizados por estos sistemas, el algoritmo de decisión para la selección de las oraciones del resumen incluye hasta cuatro fases (depende de los resultados de cada fase que sea o no necesaria la siguiente fase), que se aplican en el siguiente orden:

- **Fase 1. ACUERDO:** si una oración del texto es seleccionada por todos los sistemas, el algoritmo la mantiene.
- **Fase 2. MAYORÍA:** si una oración del texto es seleccionada por la mayoría de los sistemas, el algoritmo la mantiene.
- **Fase 3. SCORE:** si una oración es seleccionada solo por uno o dos sistemas, el algoritmo elige la que tenga asignado un mayor *score*.
- **Fase 4. SCORE + ORDEN DE LAS ORACIONES EN EL TEXTO ORIGINAL:** si se necesita una cantidad determinada de oraciones para el resumen y varias oraciones coinciden en su *score*, el algoritmo prioriza la que aparece en primer lugar en el texto original.

El algoritmo puede ser construido siguiendo diversas combinaciones de los sistemas. Sin embargo, buscando contar siempre con la presencia del sistema lingüístico, decidimos evaluar las

combinaciones posibles de DISICOSUM con los sistemas estadísticos (modificando o no el peso de los términos por YATE en CORTEX o ENERTEX). Las pruebas mostraron que la combinación de CORTEX, CORTEX+YATE, ENERTEX y DISICOSUM obtiene los mejores resultados. Por esto nos referiremos a esta combinación como el sistema o algoritmo híbrido lingüístico-estadístico⁸.

Veamos un ejemplo para ilustrar el funcionamiento de este algoritmo. Imaginemos que necesitamos un resumen de 5 oraciones, a partir de un texto que contiene 10, como el mostrado en el cuadro 1. Las oraciones son numeradas de acuerdo al orden de aparición en el texto original.

Nº	Oración
1	En Ferrol las resistencias de M. tuberculosis son bajas, por lo que no está justificado un tratamiento inicial con 4 fármacos antituberculosos.
2	Los factores asociados a la presencia de resistencias son la edad y la existencia de tratamiento previo.
3	Nuestros valores de resistencias son similares o inferiores a las existentes en otras áreas geográficas.
4	No existieron diferencias en las resistencias primarias o secundarias según la presencia o no de infección por el VIH como en otros estudios, aunque algunos autores comunicaron mayor frecuencia de resistencias primarias y secundarias en pacientes positivos para el VIH.
5	Al comparar los resultados con otros estudios debe indicarse que no siempre utilizan la misma metodología, a veces no especifican el tipo de resistencias y algunos no aportan resultados según la presencia o no de infección por el VIH, o cuando lo hacen, el porcentaje de coinfección por el VIH en los pacientes con tuberculosis es muy variable.
6	La frecuencia de las resistencias de M. tuberculosis a los fármacos antituberculosos y su evolución en el tiempo debe ser un elemento de seguimiento epidemiológico de los programas de prevención y control.
7	Para que los resultados sean fiables y reflejen el funcionamiento de un programa, la muestra de pacientes estudiados debe ser representativa del área y se han de tener en cuenta los movimientos migratorios o la posibilidad de contaminaciones en el laboratorio.
8	El porcentaje de resistencias primarias nos ofrece una idea del funcionamiento de las medidas de prevención por cuanto reflejarían la mutación espontánea de M. tuberculosis y la transmisión de cepas resistentes desde pacientes bacilíferos.
9	Las resistencias secundarias reflejan el funcionamiento de las medidas de control, ya que indican la idoneidad y el buen cumplimiento de las pautas terapéuticas.
10	En nuestro medio se incumple el tratamiento en el 8,5% de los casos y se emplean combinaciones farmacológicas desde hace años, lo que podría justificar la menor frecuencia de resistencias secundarias respecto a la de otras zonas geográficas con incumplimientos de hasta el 50% de los casos.

Cuadro 1: Texto a modo de ejemplo con las oraciones divididas según su orden de aparición en el texto original.

El cuadro 2 muestra las oraciones seleccionadas por cada sistema para un resumen de 5 oraciones teniendo en cuenta el mayor *score* obtenido (indicamos solo el número de la oración en el texto original con su respectivo *score* normalizado).

A partir de estos datos de entrada el sistema pasa (en este caso) por todas las fases del algo-

⁸Constatamos que YATE apenas mejora el desempeño de ENERTEX por ejemplo, y resulta ligeramente contraproducente incluir YATE+ENERTEX en el algoritmo híbrido.

CORTEX		CORTEX+ YATE		ENERTEX		DISICOSUM	
Oración	Score	Oración	Score	Oración	Score	Oración	Score
4	0.9	4	0.9	1	0.9	1	0.9
2	0.4	5	0.9	10	0.8	10	0.8
1	0.3	2	0.8	5	0.7	2	0.6
3	0.1	1	0.7	6	0.6	8	0.5
6	0.1	6	0.5	7	0.3	9	0.4

Cuadro 2: Oraciones seleccionadas por los cuatro sistemas para un resumen de 5 oraciones del texto del cuadro 1 de acuerdo a los mayores *scores* obtenidos.

ritmo de decisión para seleccionar las oraciones que producirán el resumen final. El algoritmo incluirá en el resumen las cinco oraciones siguientes:

- oración 1 (seleccionada en la Fase 1)
- oración 2 (seleccionada en la Fase 2)
- oración 6 (seleccionada en la Fase 2)
- oración 4 (seleccionada en la Fase 3)
- oración 5 (seleccionada en la Fase 4)

Finalmente se reordenan estas oraciones en el orden del texto original para obtener el resumen final, que se presenta en el cuadro 3.

Nº	Oración
1	En Ferrol las resistencias de M. tuberculosis son bajas, por lo que no está justificado un tratamiento inicial con 4 fármacos antituberculosos.
2	Los factores asociados a la presencia de resistencias son la edad y la existencia de tratamiento previo.
4	No existieron diferencias en las resistencias primarias o secundarias según la presencia o no de infección por el VIH como en otros estudios, aunque algunos autores comunicaron mayor frecuencia de resistencias primarias y secundarias en pacientes positivos para el VIH.
5	Al comparar los resultados con otros estudios debe indicarse que no siempre utilizan la misma metodología, a veces no especifican el tipo de resistencias y algunos no aportan resultados según la presencia o no de infección por el VIH, o cuando lo hacen, el porcentaje de coinfección por el VIH en los pacientes con tuberculosis es muy variable.
6	La frecuencia de las resistencias de M. tuberculosis a los fármacos antituberculosos y su evolución en el tiempo debe ser un elemento de seguimiento epidemiológico de los programas de prevención y control.

Cuadro 3: Oraciones del resumen final obtenido mediante el algoritmo para el ejemplo presentado.

4. Experimentos y evaluación

El algoritmo de resumen automático desarrollado se aplicó sobre un corpus formado por 40 textos en castellano extraídos de la revista *Medicina Clínica*. Cada texto es un apartado de un artículo médico, de aproximadamente una página. Dependiendo del tipo de apartado del artículo se realizaron resúmenes de diferentes longitudes: 2 oraciones del apartado de *Fundamento*, 3 oraciones del apartado de *Pacientes y métodos*, 4 oraciones del apartado de *Resultados* y 2 oraciones

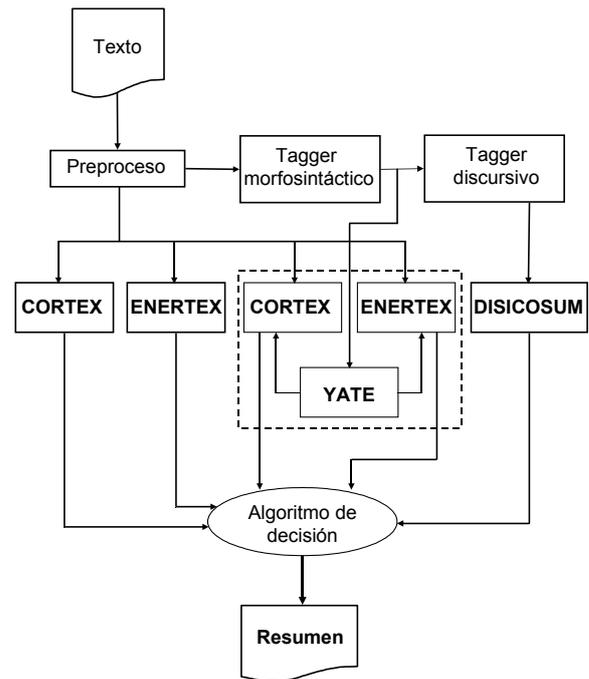


Figura 6: Arquitectura del sistema de resumen híbrido lingüístico-estadístico.

del apartado de *Discusión*. Para determinar este número, se calculó el promedio de oraciones incluidas en cada apartado de los *abstracts* de los autores. A continuación, se decidió incluir una oración adicional, ya que notamos que en muchas ocasiones en estos *abstracts* se fusionan en una los contenidos de dos o más oraciones de los artículos (veremos un ejemplo al final de este apartado, en el cuadro 5). En definitiva, fue una decisión empírica para no perder información.

Para evaluar los resúmenes se empleó un protocolo similar al usado por el NIST durante las campañas TAC/DUC. Este protocolo involucra el uso de resúmenes modelo o de referencia (escritos por personas) y el paquete ROUGE, un sistema de evaluación de resúmenes que se basa en la co-ocurrencia de n -gramas entre resúmenes candidatos (los que se quiere evaluar) y resúmenes modelo. ROUGE mide los máximos, los mínimos y el valor medio (reportado en este artículo) de la intersección de los n -gramas en los resúmenes candidatos y de referencia. Las campañas de evaluación del NIST han adoptado este test para medir la relevancia de los resúmenes. Para ser consistentes con la metodología del NIST, adoptamos el mismo protocolo en la evaluación de los resúmenes producidos por nuestro sistema. Sin embargo, hay otras pruebas estadísticas (como el test de Wilcoxon (Wilcoxon, 1945)) que podrían ser utilizadas para la evaluación. Deben llevarse

a cabo estudios más profundos en este sentido. Con el objetivo de guardar las mismas condiciones del protocolo NIST, se realizaron dos series de evaluaciones: la primera usando los resúmenes candidatos completos y la segunda usando los candidatos truncados a 100 palabras. En nuestro caso, analizamos bigramas (ROUGE-2) y bigramas separados por hasta 4 palabras (ROUGE-SU4). Para poder emplear ROUGE con textos en castellano, utilizamos un lematizador y una lista de *stopwords* en esta lengua.

Evaluamos el nuevo algoritmo lingüístico-estadístico utilizando diferentes combinaciones de los sistemas que lo forman (CORTEX, YATE, ENERTEX y DISICOSUM), así como cada uno de ellos individualmente para comparar sus resultados. Como comentamos en la sección 3, la mejor combinación fue la de CORTEX, CORTEX+YATE, ENERTEX y DISICOSUM.

Asimismo, realizamos resúmenes a modo de *baseline*. La primera (*Baseline1*) contiene oraciones del texto original seleccionadas de manera aleatoria. La segunda (*Baseline2*) incluye las primeras oraciones del texto original. El número de oraciones que incluyen estos resúmenes es el mismo que en todos los resúmenes candidatos.

Para fines de comparación se evaluaron resúmenes obtenidos con otros sistemas de resumen automático disponibles⁹: Microsoft Word, Pertinence¹⁰, Swesum¹¹ y Open Text Summarizer (OTS)¹². En principio la unidad lingüística para medir el tamaño de los resúmenes fue la oración (de nuevo, el número de oraciones que incluyen estos resúmenes es el mismo que el del resto de resúmenes candidatos). Sin embargo los fragmentos segmentados automáticamente por estos sistemas no se corresponden necesariamente con una oración. Además, al no tener acceso a ciertos parámetros de estos sistemas, es imposible obtener exactamente la misma segmentación. Por ejemplo, algunos sistemas deciden arbitrariamente que los dos puntos son un separador de oraciones. Las abreviaturas o cifras decimales pueden causar separaciones que producirán oraciones incompletas o, en el caso contrario, oraciones unidas.

Finalmente, también incluimos como resúmenes candidatos resúmenes por extracción realizados por tres médicos que colaboraron en el experimento. Estos resúmenes se incluyeron en la eva-

luación para observar en qué medida los resúmenes producidos por nuestro algoritmo y por los demás sistemas se asemejan a resúmenes producidos por humanos especialistas del dominio. Los resúmenes de los médicos no tienen un número concreto de oraciones, como los demás candidatos, sino que este varía entre 1 y 5 oraciones del texto original (sin importar la sección), ya que estas fueron las instrucciones que se les dieron a los médicos.

Los resúmenes modelo para la evaluación con ROUGE son los *abstracts* de los propios autores de los artículos. Como ya hemos comentado, en este trabajo se resumen apartados de artículos médicos de la revista *Medicina Clínica*. Esta revista solicita a los autores que realicen sus resúmenes siguiendo la misma estructura del artículo (ya postulada por Swales (1990)) y, por tanto, estos también se dividen en los cuatro apartados antes mencionados. El siguiente fragmento refleja un *abstract* realizado por un médico:

Evaluación de las vías de acceso venoso innecesarias en un servicio de urgencias

FUNDAMENTO. Determinar la prevalencia de catéteres venosos periféricos innecesarios en un servicio de urgencias.

PACIENTES Y MÉTODOS. Estudio retrospectivo sobre una muestra de 1,113 pacientes del total (24,673) que acudieron a urgencias. Se revisaron las vías venosas canalizadas y si fueron o no utilizadas.

RESULTADOS. Se practicó acceso venoso en 202 pacientes (18.15%). En 84 (41.6%) no fue utilizado. El coste económico de las vías innecesarias ascendió a 10,264.33 euros.

DISCUSIÓN. Los accesos venosos periféricos frecuentemente son utilizados innecesariamente, generando un coste de “mala calidad”.

Así, si necesitamos evaluar el resumen de un apartado de *Resultados*, por ejemplo, seleccionamos como resumen modelo únicamente el fragmento referido a este apartado en el *abstract* del autor. Al interpretar los resultados debe tenerse en cuenta que los resúmenes candidatos son resúmenes por extracción (incluidos los resúmenes de los médicos) mientras que los resúmenes modelo son resúmenes por abstracción.

Una vez obtenidos todos los resúmenes candidatos y modelo (tanto con oraciones completas como con oraciones truncadas), aplicamos ROUGE. Los resultados se muestran en el cuadro 4. Observamos que los resúmenes del algoritmo

⁹Otros sistemas disponibles, como Copernic Summarizer por ejemplo, no procesan textos en castellano y fueron descartados de este estudio.

¹⁰<http://www.pertinence.net/index.html>

¹¹<http://swesum.nada.kth.se/index-eng.html>

¹²<http://libots.sourceforge.net/>

lingüístico-estadístico diseñado obtienen los mejores resultados (en negrita y subrayados), tanto con ROUGE-2 con oraciones completas (0.3307) y con oraciones truncadas (0.3163), como con ROUGE-SU4 con oraciones completas (0.3457) y con oraciones truncadas (0.3288). Así, el algoritmo obtiene mayores puntuaciones que cualquiera de los sistemas que incluye por separado, ya sean lingüísticos o estadísticos. Esto es una muestra clara de que el algoritmo se ha realizado de manera adecuada y de que la combinación de técnicas estadísticas y lingüísticas favorece los resultados de los sistemas de resumen automático. Es de destacar que los sistemas que obtienen los siguientes mejores resultados (en negrita) son CORTEX con ROUGE-2 con oraciones completas (0.3193) y truncadas (0.2927) y con ROUGE-SU4 con oraciones completas (0.3386), y la combinación de CORTEX+YATE con ROUGE-SU4 con oraciones truncadas (0.3152). Se observa, pues, que la combinación del extractor de términos YATE mejora los resultados de CORTEX por separado al evaluar las oraciones truncadas, lo cual es una evidencia más de las ventajas de la combinación de sistemas estadísticos y lingüísticos. El sistema lingüístico DISICOSUM está algo por debajo del sistema estadístico CORTEX, pero en cambio por encima del sistema estadístico ENERTEX. De todas maneras, el algoritmo propuesto así como cualquiera de los cuatro sistemas que lo componen (ya sean estadísticos o lingüísticos) obtienen resultados mejores tanto de las dos *baselines* y todos los otros sistemas de resumen automático evaluados (Microsoft Word, Pertinence, Swesum y OTS).

Es muy destacable el hecho de que los resúmenes de los médicos obtienen una puntuación muy similar a la de nuestro algoritmo, siendo incluso menor en ocasiones. Esto quiere decir que las oraciones seleccionadas por nuestro algoritmo son prácticamente las mismas que las seleccionadas por los especialistas del dominio. Ha de tenerse en cuenta, de todas maneras, que los resúmenes de los médicos son resúmenes por extracción, por lo que es imposible que su contenido (y por tanto sus n -gramas, que es lo que compara ROUGE) coincida totalmente con el de los *abstracts* de los autores.

Las figuras 7 y 8 reflejan de forma gráfica (medidas ROUGE-2 contra ROUGE-SU4) los resultados numéricos del cuadro 4.

Finalmente, a modo de ejemplo, mostramos en el cuadro 5 un resumen (por extracción) producido por el algoritmo lingüístico-estadístico, el resumen del autor del artículo (por abstracción) y el resumen de uno de los tres médicos que colaboraron en el experimento (por extracción). Co-

Sistema de resumen	ROUGE-2		ROUGE-SU4	
	Frases	100 Palabras	Frases	100 Palabras
Híbrido	0.3307	0.3163	0.3457	0.3288
CORTEX	0.3193	0.2927	0.3386	0.3105
CORTEX+YATE	0.3038	0.2913	0.3281	0.3152
ENERTEX	0.2552	0.2314	0.2830	0.2589
DISICOSUM	0.2851	0.2750	0.3053	0.2945
Baseline ₁	0.1454	0.1428	0.1835	0.1808
Baseline ₂	0.1931	0.1861	0.2333	0.2260
Word	0.1873	0.1857	0.2273	0.2245
Pertinence	0.1768	0.1606	0.2266	0.2095
Swesum	0.2026	0.1939	0.2382	0.2289
OTS	0.2337	0.2176	0.2675	0.2533
Médico 1	0.3329	0.3030	0.3414	0.3130
Médico 2	0.3230	0.2993	0.3374	0.3130
Médico 3	0.3099	0.2721	0.3201	0.2898

Cuadro 4: Comparación de los valores medios de ROUGE entre los diferentes resúmenes.

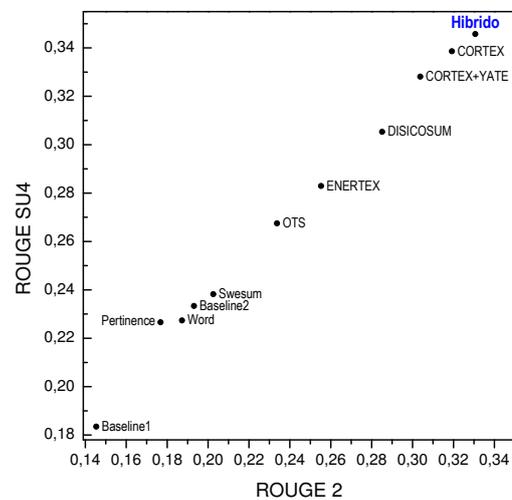


Figura 7: Resultados de la evaluación con ROUGE (resúmenes con oraciones completas).

mo puede observarse, los resúmenes tienen diferentes longitudes. El resumen producido por el algoritmo incluye dos oraciones, ya que se trata del resumen del apartado de *Discusión* de un artículo médico de nuestro corpus. El resumen del autor, en cambio, solo contiene una oración y el resumen de uno de los médicos, cuatro. Es de destacar que las dos oraciones seleccionadas por el algoritmo (1a y 2a) coinciden con dos de las oraciones incluidas en el resumen del médico (3c y 4c). También es relevante el hecho de que, aunque el autor solo ofrezca una oración a modo de resumen (1b), esta refleje una fusión de las ideas expresadas en las dos oraciones que incluye el resumen del algoritmo (1a y 2a).

5. Conclusiones

El método de resumen automático propuesto en este trabajo da como resultado un sistema híbrido de resumen que integra de manera equi-

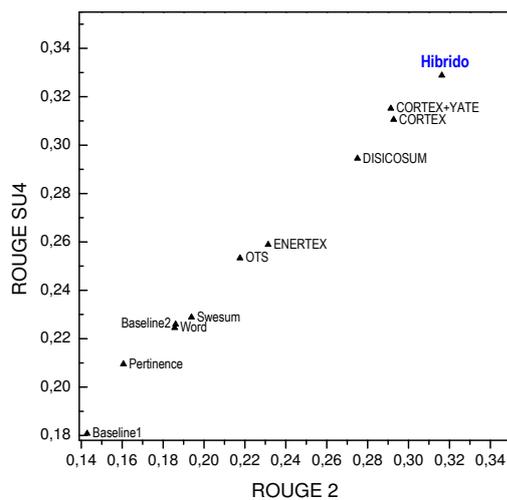


Figura 8: Resultados de la evaluación con ROUGE (resúmenes truncados a 100 palabras).

Oración	Resumen producido por el algoritmo
1a	El estudio genotípico de la región promotora del gen UGT-1 facilita la identificación de individuos con variantes alélicas asociadas a hiperbilirrubinemias no conjugadas y confirma el diagnóstico de síndrome de Gilbert que en la actualidad se realiza por exclusión de otros procesos patológicos.
2a	Estos resultados nos permiten plantear la idoneidad del escrutinio molecular para el síndrome de Gilbert como una prueba adicional, tanto en el protocolo diagnóstico de hiperbilirrubinemias no conjugadas, crónicas, de intensidad moderada, en ausencia de enfermedad hepática en el adulto, como en recién nacidos con ictericia neonatal prolongada.
Oración	Resumen del autor
1b	El porcentaje de alelos mutados detectados en la población analizada, similar al hallado en otras poblaciones caucásicas, plantea la inclusión del análisis genotípico del gen UGT-1 en el protocolo diagnóstico de hiperbilirrubinemias no conjugadas, crónicas, de intensidad moderada, en ausencia de hemólisis y de enfermedad hepática.
Oración	Resumen del médico 1
1c	Ante un paciente con episodios de hiperbilirrubinemia indirecta, habitualmente se efectúa el diagnóstico de síndrome de Gilbert después de excluir la existencia de una hepatopatía o de un síndrome hemolítico.
2c	Estas pruebas, además de ser incómodas para el paciente, son poco específicas.
3c	El estudio genotípico de la región promotora del gen UGT-1 facilita la identificación de individuos con variantes alélicas asociadas a hiperbilirrubinemias no conjugadas y confirma el diagnóstico de síndrome de Gilbert que en la actualidad se realiza por exclusión de otros procesos patológicos.
4c	Estos resultados nos permiten plantear la idoneidad del escrutinio molecular para el síndrome de Gilbert como una prueba adicional, tanto en el protocolo diagnóstico de hiperbilirrubinemias no conjugadas, crónicas, de intensidad moderada, en ausencia de enfermedad hepática en el adulto, como en recién nacidos con ictericia neonatal prolongada.

Cuadro 5: Ejemplo de resúmenes producidos por el algoritmo, el autor y un médico.

librada diversos sistemas lingüísticos y estadísticos ya existentes. En este trabajo hemos mostrado que, a diferencia de los resultados obtenidos en otras áreas del procesamiento del lenguaje, la combinación de métodos numéricos y simbólicos contribuye a mejorar los resultados

de la tarea propuesta. Más concretamente, los resúmenes automáticos producidos por métodos estadísticos (CORTEX y ENERTEX) son similares a los producidos por métodos lingüísticos (DISICOSUM) y que un extractor de términos especializado como YATE contribuye a mejorar ligeramente la salida de un resumidor estadístico, al menos en los resúmenes truncados a 100 palabras. Sin embargo, la principal conclusión a la que llegamos es que la combinación de técnicas estadísticas y lingüísticas en el problema del resumen automático obtiene excelentes resultados, mejores que cualquiera de los sistemas evaluados por separado. Nuestro algoritmo lingüístico-estadístico consigue superar a los otros sistemas de resumen y a las *baselines* diseñadas. Además, los resúmenes que ofrece son muy similares a los realizados por médicos especialistas, lo cual es una evidencia clara de la calidad de los mismos.

En el futuro prevemos realizar experimentos con textos de otros dominios (genómica) y con otras lenguas (francés y catalán). Además, realizaremos un experimento en el que se use el sistema YATE para detectar automáticamente los términos del título del texto y posteriormente seleccionar para el resumen las oraciones que contengan términos relacionados semánticamente con ellos.

Agradecimientos

Parte de este trabajo ha sido financiado mediante una ayuda de movilidad posdoctoral otorgada por el Ministerio de Ciencia e Innovación de España (Programa Nacional de Movilidad de Recursos Humanos de Investigación; Plan Nacional de Investigación Científica, Desarrollo e Innovación 2008-2011) a Iria da Cunha.

References

- Afantenos, S.D., V. Karkaletsis, and P. Stamato-poulos. 2005. Summarization of medical documents: A survey. *Artificial Intelligence in Medicine*, 2(33):157–177.
- Alonso, L. and M. Fuentes. 2003. Integrating cohesion and coherence for Automatic Summarization. In *EACL'03 Student Session*, pages 1–8. ACL, Budapest.
- Aretoulaki, M. 1996. *COSY-MATS: A Hybrid Connectionist-Symbolic Approach To The Pragmatic Analysis Of Texts For Their Automatic Smmarization*. Ph.D. thesis, University of Manchester, Institute of Science and Technology, Manchester.
- Barzilay, R. and M. Elhadad. 1997. Using lexical chains for text summarization. In *Intelli-*

- gent Scalable Text Summarization Workshop, ACL, Madrid, Spain.
- Barzilay, R. and M. Lapata. 2005. Modelling local coherence: An entity-based approach. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pages 318–325.
- Berger, A. and V. Mittal. 2000. A system for summarizing Web Pages. In *23rd Annual Conference on Research and Development in Information Retrieval*, pages 144–151. Atenas.
- Brandow, R., K. Mitze, and L. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Inf. Proc. and Management*, 31:675–685.
- da Cunha, I. 2008. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español*. Ph.D. thesis, IULA, Barcelona, España.
- da Cunha, I., S. Fernández, P. Velázquez, J. Vivaldi, E. SanJuan, and J.M. Torres-Moreno. 2007. A new hybrid summarizer based on Vector Space Model, Statistical Physics and Linguistics. In *MICAI 2007: Advances in Artificial Intelligence. Lecture Notes in Computer Science*, pages 872–882. Gelbukh, A. and Kuri Morales, A. F. (eds.), Berlín: Springer.
- da Cunha, I. and L. Wanner. 2005. Towards the Automatic Summarization of Medical Articles in Spanish: Integration of textual, lexical, discursive and syntactic criteria. In *Crossing Barriers in Text Summarization Research (RANLP-2005)*, pages 46–51. Saggion, H. and Minel, J. (eds.), Borovets (Bulgaria): INCOMA Ltd.
- da Cunha, I., L. Wanner, and M. T. Cabré. 2007. Summarization of specialized discourse: The case of medical articles in Spanish. *Terminology*, 13(2):249–286.
- Edmundson, H. P. 1969. New Methods in Automatic Extraction. *Journal of the Association for Computing Machinery*, 16:264–285.
- Fernández, S. 2009. *Applications exploratoires des modèles de spins au Traitement Automatique de la Langue*. Ph.D. thesis, Université Henri Poincaré Nancy 2, France.
- Fernández, S., E. SanJuan, and J. M. Torres-Moreno. 2007. Énergie textuelle de mémoires associatives. In *Traitement Automatique des Langues Naturelles*, pages 25–34. Toulouse, France.
- Fernández, S., E. SanJuan, and J. M. Torres-Moreno. 2008. Energetex : un système basé sur l'énergie textuelle. In *Traitement Automatique des Langues Naturelles*, pages 99–108. Avignon, France.
- Fuentes, M. 2008. *A Flexible Multitask Summarizer for Documents from Different Media, Domain, and Language*. Ph.D. thesis, UPC, Barcelona.
- Goldstein, J., J. Carbonell, M. Kantrowitz, and V. Mittal. 1999. Summarizing text documents: sentence selection and evaluation metrics. In *22nd Int. ACM SIGIR Research and development in information retrieval*, pages 121–128. Berkeley.
- Hertz, J., A. Krogh, and G. Palmer. 1991. *Introduction to the theory of Neural Computation*. Redwood City, CA : Addison-Wesley.
- Hopfield, J. 1982. Neural networks and physical systems with emergent collective computational abilities. *National Academy of Sciences*, 9:2554–2558.
- Hovy, E. and Lin, C.Y. 1999. Automated text summarisation in SUMMARIST. In *Advances in automatic text summarisation (Ed. I. Mani and M. Maybury)*, pages 81–94. Cambridge, MA: MIT Press.
- Kupiec, J., J. O. Pedersen, and F. Chen. 1995. A trainable document summarizer. In *SIGIR-95*, pages 68–73. New York.
- Lacatusu, V.F., P. Parker, and S.M. Harabagiu. 2003. Lite-GISTexter: Generating short summaries with minimal resources. In *Proceedings of the DUC 2003*, pages 122–128.
- Leskovec, J., N. Milic-Frayling, and M. Grobelnik. 2005. Impact of linguistic analysis on the semantic graph coverage and learning of document extracts. In *Proceedings of the AAAI 2005*, volume 3, pages 1069–1074.
- Lin, C. and E. Hovy. 1997. Identifying Topics by Position. In *ACL Applied Natural Language Processing Conference*, pages 283–290. Washington.
- Lin, C.Y. 2004. Rouge: A Package for Automatic Evaluation of Summaries. In *Workshop on Text Summarization Branches Out (WAS 2004)*, pages 25–26.
- Luhn, H. P. 1959. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2).
- Mani, I. 2001. *Automatic summarization*. Amsterdam: John Benjamins Publishing.

- Mani, I. and M.T. Maybury. 1999. *Advances in automatic text summarization*. Cambridge, MA: MIT Press.
- Mann, W. C. and S. A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.
- Marcu, D. 1998. *The rhetorical parsing, summarization, and generation of natural language texts*. Ph.D. thesis, Dep. of Computer Science, University of Toronto.
- Marcu, D. 2000. *The Theory and Practice of Discourse Parsing Summarization*. Institute of Technology, Massachusetts.
- Marcu, D. and A. Echiabi. 2002. An unsupervised approach to recognising discourse relations. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 368–375.
- Mel’cuk, I. 1988. *Dependency Syntax: Theory and Practice*. Albany: State University Press of New York.
- Mel’cuk, I. 2001. *Communicative Organization in Natural Language. The semantic-communicative structure of sentences*. John Benjamins, Amsterdam.
- Nomoto, T. and Nitta, Y. 1994. A Grammatico-Statistical Approach to Discourse Partitioning. In *15th Int. Conf. on Comp. Linguistics*, pages 1145–1150. Kyoto.
- Ono, K., K. Sumita, and S. Miike. 1994. Abstract generation based on rhetorical structure extraction. In *15th Int. Conf. on Comp. Linguistics*, pages 344–348. Kyoto.
- Pardo, T. and Nunes, M. 2008. On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *Journal of Theoretical and Applied Computing*, 15(2):43–64.
- Pardo, T. and Nunes, M. and Rino, M. 2004. DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. In *XVII Brazilian Symposium on Artificial Intelligence - SBIA2004*, pages 224–234. São Luís.
- Polanyi, L., C. Chris, M. van den Berg, G.L. Thione, and D. Ahn. 2004. A rule-based approach to discourse parsing. In *Proceedings of the fifth SIGdial workshop on discourse and dialogue*, pages 108–117.
- Radev, D., T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Celebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang. 2004. MEAD - a platform for multilingual summarisation. In *Proceedings of LREC 2004*.
- Radev, D.R., H. Jing, and M. Budzikowska. 2000. Centroid-based summarisation of multiple documents: Sentence extraction, utility-based evaluation, and user Studies. In *Proceedings of the ANLP/NAACL-00*, pages 21–30.
- Salton, G. and M. McGill. 1983. *Introduction to modern information retrieval*. Computer Science Series McGraw Hill Publishing Company.
- Silber, H. Gregory and Kathleen F. McCoy. 2000. Efficient text summarization using lexical chains. In *Intelligent User Interfaces*, pages 252–255.
- Spärck Jones, K. 2007. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481.
- Sumita, K., K. Ono, T. Chino, T. Ukita, and S. Amano. 1992. A discourse structure analyzer for Japanese text. In *International Conference on Fifth Generation Computer Systems*, pages 1133–1140. Tokyo, Japan.
- Swales, J. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, Cambridge.
- Teufel, S. and M. Moens. 2002. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28(4):409–445.
- Thione, G.L., M. van den Berg, L. Polanyi, and C. Culy. 2004. Hybrid text summarisation: Combining external relevance measures with structural analysis. In *Proceedings of the ACL-04*, pages 51–55.
- Torres-Moreno, J. M., P. Velázquez-Morales, and J. G. Meunier. 2001. Cortex : un algorithme pour la condensation automatique des textes. In *ARCo 2001*, pages 65–75. Lyon, France.
- Torres-Moreno, J. M., P. Velázquez-Morales, and J. G. Meunier. 2002. Condensés de textes par des méthodes numériques. In *JADT 2002*, pages 723–734. St. Malo, France.
- Vanderwende, L., M. Banko, and A. Menezes. 2004. Event-centric summary generation. In *Proceedings of the DUC 2004*, pages 76–81.
- Vivaldi, J. 2001. *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona.

- Vivaldi, J. and H. Rodríguez. 2001. Improving term extraction by combining different techniques. *Terminology*, 7(1):31–47.
- Vivaldi, J. and H. Rodríguez. 2002. Medical term extraction using the EWN ontology. In *Terminology and Knowledge Engineering*, pages 137–142. Nancy.
- Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics*, 1:80–83.